



GENE REGULATION AS A DRIVER OF ADAPTATION AND SPECIATION

EDITED BY: Ekaterina Shelest, Katja Nowick and Deborah A. Triant
PUBLISHED IN: *Frontiers in Genetics*



frontiers

Frontiers eBook Copyright Statement

The copyright in the text of individual articles in this eBook is the property of their respective authors or their respective institutions or funders. The copyright in graphics and images within each article may be subject to copyright of other parties. In both cases this is subject to a license granted to Frontiers.

The compilation of articles constituting this eBook is the property of Frontiers.

Each article within this eBook, and the eBook itself, are published under the most recent version of the Creative Commons CC-BY licence.

The version current at the date of publication of this eBook is CC-BY 4.0. If the CC-BY licence is updated, the licence granted by Frontiers is automatically updated to the new version.

When exercising any right under the CC-BY licence, Frontiers must be attributed as the original publisher of the article or eBook, as applicable.

Authors have the responsibility of ensuring that any graphics or other materials which are the property of others may be included in the CC-BY licence, but this should be checked before relying on the CC-BY licence to reproduce those materials. Any copyright notices relating to those materials must be complied with.

Copyright and source acknowledgement notices may not be removed and must be displayed in any copy, derivative work or partial copy which includes the elements in question.

All copyright, and all rights therein, are protected by national and international copyright laws. The above represents a summary only. For further information please read Frontiers' Conditions for Website Use and Copyright Statement, and the applicable CC-BY licence.

ISSN 1664-8714

ISBN 978-2-88971-981-5

DOI 10.3389/978-2-88971-981-5

About Frontiers

Frontiers is more than just an open-access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers Journal Series

The Frontiers Journal Series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the Frontiers Journal Series operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to Quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews.

Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the Frontiers Journals Series: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area! Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers Editorial Office: frontiersin.org/about/contact

GENE REGULATION AS A DRIVER OF ADAPTATION AND SPECIATION

Topic Editors:

Ekaterina Shelest, University of Portsmouth, United Kingdom

Katja Nowick, Freie Universität Berlin, Germany

Deborah A. Triant, University of Virginia, United States

Citation: Shelest, E., Nowick, K., Triant, D. A., eds. (2021). Gene Regulation as a Driver of Adaptation and Speciation. Lausanne: Frontiers Media SA.
doi: 10.3389/978-2-88971-981-5

Table of Contents

- 04 Editorial: Gene Regulation as a Driver of Adaptation and Speciation**
Deborah A. Triant, Katja Nowick and Ekaterina Shelest
- 06 Integrative Omics Analysis Reveals a Limited Transcriptional Shock After Yeast Interspecies Hybridization**
Hrant Hovhannisyan, Ester Saus, Ewa Ksiezopolska, Alex J. Hinks Roberts, Edward J. Louis and Toni Gabaldón
- 20 Contributions of Adaptive Plant Architecture to Transgressive Salinity Tolerance in Recombinant Inbred Lines of Rice: Molecular Mechanisms Based on Transcriptional Networks**
Isaiah Catalino M. Pabuayon, Ai Kitazumi, Glenn B. Gregorio, Rakesh Kumar Singh and Benildo G. de los Reyes
- 43 The Effects of Gene Duplication Modes on the Evolution of Regulatory Divergence in Wild and Cultivated Soybean**
Na Zhao, Xiaoyang Ding, Taotao Lian, Meng Wang, Yan Tong, Di Liang, Qi An, Siwen Sun, Scott A. Jackson, Bao Liu and Chunming Xu
- 52 Epigenetics as Driver of Adaptation and Diversification in Microbial Eukaryotes**
Agnes K. M. Weiner and Laura A. Katz
- 58 Promoter Architecture and Transcriptional Regulation of Genes Upregulated in Germination and Coleoptile Elongation of Diverse Rice Genotypes Tolerant to Submergence**
Bijayalaxmi Mohanty
- 84 KLF4, a Key Regulator of a Transitive Triplet, Acts on the TGF- β Signaling Pathway and Contributes to High-Altitude Adaptation of Tibetan Pigs**
Tao Wang, Yuanyuan Guo, Shengwei Liu, Chaoxin Zhang, Tongyan Cui, Kun Ding, Peng Wang, Xibiao Wang and Zhipeng Wang
- 101 Positive Selection in Gene Regulatory Factors Suggests Adaptive Pleiotropic Changes During Human Evolution**
Vladimir M. Jovanovic, Melanie Sarfert, Carlos S. Reyna-Blanco, Henrike Indrischek, Dulce I. Valdivia, Ekaterina Shelest and Katja Nowick
- 116 Gene Expression Modification by an Autosomal Inversion Associated With Three Male Mating Morphs**
Jasmine L. Loveland, David B. Lank and Clemens Küpper
- 132 Impact of Genetic Variation in Gene Regulatory Sequences: A Population Genomics Perspective**
Manas Joshi, Adamandia Kapopoulou and Stefan Laurent
- 142 Evolutionary Perspective and Expression Analysis of Intronless Genes Highlight the Conservation of Their Regulatory Role**
Katia Aviña-Padilla, José Antonio Ramírez-Rafael, Gabriel Emilio Herrera-Oropeza, Vijaykumar Yogesh Muley, Dulce I. Valdivia, Erik Díaz-Valenzuela, Andrés García-García, Alfredo Varela-Echavarría and Maribel Hernández-Rosales



Editorial: Gene Regulation as a Driver of Adaptation and Speciation

Deborah A. Triant^{1*}, Katja Nowick² and Ekaterina Shelest³

¹Department of Biochemistry & Molecular Genetics, University of Virginia, Charlottesville, VA, United States, ²Institute for Biology, Freie Universität Berlin, Berlin, Germany, ³Centre for Enzyme Innovation, University of Portsmouth, Portsmouth, United Kingdom

Keywords: gene regulatory factors, adaptation, speciation, transcription factors, biodiversity, evolution

Editorial on the Research Topic

Gene Regulation as a Driver of Adaptation and Speciation

One fundamental goal in evolutionary biology is to understand the interaction between adaptation and speciation and how it can generate and maintain biodiversity (Wolf et al., 2010). Speciation leads to the diversification of lineages, whereas adaptation maximizes the survival and reproductive success of organisms in an ever-changing environment, thereby further increasing diversification. Genetic approaches have been critical for examining the molecular basis of adaptation and speciation. The availability of an increasing number of genome assemblies allows for the identification of genomic components that underlie these processes, providing new insights into the proximate mechanisms of diversification and deeper understanding of biodiversity evolution.

Recent discoveries suggest that gene regulation plays an important role in speciation and adaptive diversification (Jones et al., 2012; Mack et al., 2018). We present this collection of papers, eight of original research, one opinion and one review, that explore genomic features that foster divergence in gene regulation and how evolutionary changes of regulation systems impact adaptation and speciation. The incorporation of gene regulation in evolutionary processes is multifaceted, and so are the approaches to its understanding. This multiplicity is reflected by the scope of the articles in our collection, ranging from the interplay between regulation and evolutionary patterns on a molecular level (e.g., Aviña-Padilla et al., Mohanty) to the investigation of adaptations of organisms to certain conditions (Wang et al.). Changes caused by epigenetics, structural variants, transcription factors (TFs) and regulatory sequences that elicit modifications in gene expression and drive adaptation and diversification are also explored. Here, we briefly summarize the authors' findings.

We start with contributions that investigated genome architecture. In interspecific hybridization in *Saccharomyces* yeast described by Hovhannisyan et al., the coexistence of two distinct genomes after hybridization is supported by the limited regulatory interference on broad transcriptional changes. This may explain the unusually strong ability of these organisms to successfully hybridize. In instances of genome duplications, the genomic history can play a crucial role in adaptation. As shown by Zhao et al., genes can be regulated differently depending on whether or not they have been duplicated. The authors evaluated transcriptional changes in cultivated and wild *Glycine* soybeans with highly duplicated genomes after whole genome duplications. They found more *trans*-only regulation in duplicated genes than in singletons, especially in tandem duplications. Their results suggest that genes with different duplication modes accumulate different types of regulatory divergence.

Two papers show how changes on a chromosomal level can be instrumental in adaptation and speciation. Pabuayan et al. presented plant architectural modifications created through transgressive segregation in Pokkali rice. The authors examined morpho-developmental and physiological profiles and used transcriptomic data of recombinant inbred lines varying in salt tolerance to build genetic networks. The network analysis indicated novel adaptive architectures correlating with levels of

OPEN ACCESS

Edited and reviewed by:

Denis Baurain,
University of Liège, Belgium

*Correspondence:

Deborah A. Triant
dtriant@virginia.edu

Specialty section:

This article was submitted to
Evolutionary and Population Genetics,
a section of the journal
Frontiers in Genetics

Received: 12 October 2021

Accepted: 18 October 2021

Published: 12 November 2021

Citation:

Triant DA, Nowick K and Shelest E
(2021) Editorial: Gene Regulation as a
Driver of Adaptation and Speciation.
Front. Genet. 12:793933.
doi: 10.3389/fgene.2021.793933

salinity stress. The study demonstrates how genetic recombination creates novel morphology that has important implications in plant defenses. Loveland et al. investigated the impact of structural variants on gene expression. They tested whether genes located in chromosomal inversions creating three morphs of the ruff *Philomachus pugnax* differ in gene expression. They demonstrated a clear expression difference between morphs for CENPN, a gene disrupted by the inversion. Genes located inside the inverted genomic area, including genes coding for TFs and sex hormones, displayed tissue specific allelic imbalance, suggesting complex and wide-ranging changes in gene expression caused by structural variants.

Aviña-Padilla et al. inferred the conservation and functions of intronless genes (IGs), most of them originating from retrotransposition, in vertebrate genomes. Interestingly, many IGs code for TFs and other molecules involved in gene expression regulation. Comparative analysis confirmed high conservation of IGs, but revealed also a large proportion of evolutionary young IGs. These findings strengthen previous observations that IGs and multi-exon genes are under different evolutionary constraints.

Direct involvement of TFs in evolutionary processes was tested in two studies. The first, by Wang et al., used RNAseq-based analysis of genetic networks in high and low-altitude adapted pigs. A particular regulatory circuit triggered by a TF, KLF4, and involving TGF- β signaling active in the lungs of high-altitude pigs may have helped them to adapt to hypoxic environments. Jovanovic et al. investigated the evolution of more than 3000 gene regulatory factors (GRF) genes across 27 primate genomes and reported five candidate GRFs that have been positively selected on the human branch. Without finding common patterns of co-expression, the authors concluded that positively selected sites could have pleiotropic effects on phenotypic adaptation.

GRFs are one side of the coin, while their binding sites present the other. Mohanty explored the promoter architecture associated with transcriptional and hormonal regulation in

diverse rice genotypes that provide varying tolerance to submergence. Analyzing transcriptome data, he identified putative *cis*-elements in promoters of genes upregulated in each tolerance group along with the TFs from each genotype. He surmised that phenotypic differences are due to the differences in transcriptional regulation across the groups. Evolution of *cis*-regulatory elements is also in the focus of the review from Joshi et al. In particular, they discuss studies on genetic variation in these elements within the context of population genetics. They introduce common approaches to identify regulatory regions and some statistical tools for inferring selection on non-coding functional regions. The authors highlight studies on the analyses of selective forces acting on non-coding genomic elements.

Finally, Weiner and Katz propose in their opinion paper that epigenetic processes are involved in generating the tremendous phenotypic diversity of protists by driving plasticity, differential adaptation and diversification. Their argument is based upon findings demonstrating that epigenetic processes are widespread across eukaryotes and that some epigenetic marks can be inherited.

Together, these articles represent a broad perspective on the role of gene regulation in speciation and adaptation. They cover a variety of scientific questions and approaches, both methodological and conceptual, over diverse taxonomic groups. Most importantly, they highlight that although an abundance of work on regulatory drivers of speciation and adaptation already exists, the ever-expanding amount of genomic data and techniques will continue to push for new questions and discoveries in this compelling field.

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct and intellectual contribution to the work, and approved it for publication.

REFERENCES

- Jones, F. C., Grabherr, M. G., Grabherr, M. G., Chan, Y. F., Russell, P., Mauceli, E., et al. (2012). The Genomic Basis of Adaptive Evolution in Threespine Sticklebacks. *Nature* 484 (7392), 55–61. doi:10.1038/nature10944
- Mack, K. L., Ballinger, M. A., Phifer-Rixey, M., and Nachman, M. W. (2018). Gene Regulation Underlies Environmental Adaptation in House Mice. *Genome Res.* 28 (11), 1636–1645. doi:10.1101/gr.238998.118
- Wolf, J. B. W., Lindell, J., and Backström, N. (2010). Speciation Genetics: Current Status and Evolving Approaches. *Philos. Trans. R. Soc. B* 365 (1547), 1717–1733. doi:10.1098/rstb.2010.0023

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2021 Triant, Nowick and Shelest. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



OPEN ACCESS

Edited by:

Ekaterina Shelest,
German Centre for Integrative
Biodiversity Research (iDiv), Germany

Reviewed by:

Amparo Querol,
Consejo Superior de Investigaciones
Científicas (CSIC), Spain
Matthias Sipiczki,
University of Debrecen, Hungary

*Correspondence:

Toni Gabaldón
toni.gabaldon@bsc.es

† Present address:

Hrant Hovhannisyan,
Barcelona Supercomputing Centre
(BSC-CNS) and Institute for Research
in Biomedicine (IRB), Barcelona,
Spain
Ester Saus,
Barcelona Supercomputing Centre
(BSC-CNS) and Institute for Research
in Biomedicine (IRB), Barcelona,
Spain
Ewa Ksiezopolska,
Barcelona Supercomputing Centre
(BSC-CNS) and Institute for Research
in Biomedicine (IRB), Barcelona,
Spain
Toni Gabaldón,
Barcelona Supercomputing Centre
(BSC-CNS) and Institute for Research
in Biomedicine (IRB), Barcelona,
Spain

Specialty section:

This article was submitted to
Evolutionary and Population Genetics,
a section of the journal
Frontiers in Genetics

Received: 05 February 2020

Accepted: 30 March 2020

Published: 07 May 2020

Citation:

Hovhannisyan H, Saus E,
Ksiezopolska E, Hinks Roberts AJ,
Louis EJ and Gabaldón T (2020)
Integrative Omics Analysis Reveals
a Limited Transcriptional Shock After
Yeast Interspecies Hybridization.
Front. Genet. 11:404.
doi: 10.3389/fgene.2020.00404

Integrative Omics Analysis Reveals a Limited Transcriptional Shock After Yeast Interspecies Hybridization

Hrant Hovhannisyan^{1,2†}, Ester Saus^{1,2†}, Ewa Ksiezopolska^{1,2†}, Alex J. Hinks Roberts³,
Edward J. Louis³ and Toni Gabaldón^{1,2,4*†}

¹ Centre for Genomic Regulation, Barcelona Institute of Science and Technology, Barcelona, Spain, ² Department of Health and Life Sciences, Universitat Pompeu Fabra, Barcelona, Spain, ³ Centre for Genetic Architecture of Complex Traits, University of Leicester, Leicester, United Kingdom, ⁴ Institució Catalana de Recerca i Estudis Avançats, Barcelona, Spain

The formation of interspecific hybrids results in the coexistence of two diverged genomes within the same nucleus. It has been hypothesized that negative epistatic interactions and regulatory interferences between the two sub-genomes may elicit a so-called genomic shock involving, among other alterations, broad transcriptional changes. To assess the magnitude of this shock in hybrid yeasts, we investigated the transcriptomic differences between a newly formed *Saccharomyces cerevisiae* × *Saccharomyces uvarum* diploid hybrid and its diploid parentals, which diverged ~20 mya. RNA sequencing (RNA-Seq) based allele-specific expression (ASE) analysis indicated that gene expression changes in the hybrid genome are limited, with only ~1–2% of genes significantly altering their expression with respect to a non-hybrid context. In comparison, a thermal shock altered six times more genes. Furthermore, differences in the expression between orthologous genes in the two parental species tended to be diminished for the corresponding homeologous genes in the hybrid. Finally, and consistent with the RNA-Seq results, we show a limited impact of hybridization on chromatin accessibility patterns, as assessed with assay for transposase-accessible chromatin using sequencing (ATAC-Seq). Overall, our results suggest a limited genomic shock in a newly formed yeast hybrid, which may explain the high frequency of successful hybridization in these organisms.

Keywords: hybridization, yeast hybrid, transcriptome shock, allele-specific expression, buffering

INTRODUCTION

Interspecific hybridization, meaning the mating of two different species to produce viable offspring, has been observed across a wide range of eukaryotic taxa and is considered a major mechanism driving adaptation to new environmental niches (Gladieux et al., 2014; Depotter et al., 2016; Session et al., 2016). Hybridization in animals (Schwenk et al., 2008) and plants (Rieseberg, 1997) has long been recognized, and these organisms have focused the attention of most of the studies on addressing the mechanisms and consequences of hybridization. In contrast, hybridization in microbial eukaryotes has been historically neglected, given the difficulty to detect morphological or physiological differences between species and their hybrids. It was the deep physiological and genetic characterization of the model yeast species *Saccharomyces cerevisiae* that allowed

the discovery that several strains, initially classified as independent species, were in fact hybrids (Dujon, 2010). More recently, advances in next-generation sequencing (Goodwin et al., 2016) have facilitated the discovery of hybrids, demonstrating that hybridization is more frequent than previously anticipated, particularly in some microbial groups such as fungi (Albertin and Marullo, 2012). *Saccharomycotina* yeasts seem particularly prone to hybridization (Morales and Dujon, 2012), and there are numerous examples of yeast hybrid lineages of clinical (Pryszcz et al., 2014, 2015; Schröder et al., 2016; Mixão et al., 2019) or industrial (Le Jeune et al., 2007; Baker et al., 2015; Krogerus et al., 2017) relevance. Furthermore, a hybridization event has been proposed to have led to an ancient whole-genome duplication in the lineage leading to *S. cerevisiae* and related yeasts (Marcet-Houben and Gabaldón, 2015).

An immediate outcome of interspecies hybridization is the coexistence of divergent genetic material within the same nucleus. This has been proposed to lead to a state called “genomic shock” (McClintock, 1984), in which negative epistatic interactions between the two coexisting sub-genomes, including interference between their gene regulatory networks, result in large physiological alterations.

Recent research has studied the effects of this “shock” on different layers of cellular organization, including, among others, the genome (Dutta et al., 2017; Smukowski Heil et al., 2017), the transcriptome (Cox et al., 2014; Hu et al., 2016; Lopez-Maestre et al., 2017), the epigenome (Groszmann et al., 2011; Greaves et al., 2015), and the proteome (Guo et al., 2013; Hu et al., 2017). Specifically, the assessment of transcriptomic changes in hybrids has been used for exploring *cis*- and *trans*-regulation of gene expression (Tirosh et al., 2009; Graze et al., 2012; Li and Fay, 2017; Metzger et al., 2017; Waters et al., 2017). The comparison of gene expression levels in hybrid lineages *versus* their respective parents constitutes a versatile model for assessing gene regulation (Wittkopp et al., 2004). Considering that parental genomes in a hybrid are exposed to the same cellular environment, and thus *trans*-regulatory elements, differences in the gene expression levels within a hybrid can be attributed to *cis*-regulation, while the differences observed between parental organisms are due to a combination of *cis* and *trans* effects (Wittkopp et al., 2004). Using this concept, *cis*- and *trans*-regulatory effects on gene expression have been studied in numerous taxa, including fungi (Thompson and Regev, 2009), flies (McManus et al., 2010), and plants (Guo et al., 2008; Combes et al., 2015). Most transcriptomic studies of fungal hybrids have been performed in that particular context. For instance, Tirosh et al. (2009) investigated the impact of *cis* and *trans* effects on gene expression divergence in closely related *S. cerevisiae* and *Saccharomyces paradoxus* and their interspecific hybrid at four different growth conditions. By performing within-hybrid (*cis* effects) comparisons and subtracting those from between-parent comparisons (*trans* effects), the authors demonstrated that the majority of the regulatory divergence was the result of *cis* effects, attributed to differences in the promoter and regulatory regions that were independent of the environmental condition. On the other hand, *trans* effects were related to the transcription and chromatin regulators and were mostly condition-specific.

Using a similar approach, Metzger et al. (2017) used publicly available RNA sequencing (RNA-Seq) datasets of two *S. cerevisiae* strains and their hybrid (Schaefer et al., 2013) and data of *S. cerevisiae*, *S. paradoxus*, *Saccharomyces mikatae*, and *Saccharomyces bayanus* and their respective hybrids (Schraiber et al., 2013) to assess the dynamics of the regulatory changes throughout long evolutionary distances. They concluded that, as sequence divergence increases, *cis*-regulatory divergence becomes the dominant regulatory mechanism and that both differences in the gene expression and regulatory sequences increase with genetic distance, reaching a plateau for distantly related species.

Another study (Li and Fay, 2017) used *S. cerevisiae* × *Saccharomyces uvarum* hybrid, resulting from the mating of two thermally divergent species, to investigate the effect of temperature on allele-specific expression (ASE). Using RNA-Seq, the authors assessed the ASE patterns in the hybrids grown at different temperatures and showed that most of the *cis* divergence is temperature-independent, with only a small fraction of the ASE genes influenced by thermal condition. Overall, most previous studies used the transcriptomics of hybrids as a means to investigate *cis* and *trans* effects on gene regulation at various conditions and evolutionary distances, but they did not directly assess the impact of hybridization on gene expression and how this compares with the regulatory impact of other stresses. Given their different focus, these studies do not measure gene expression in matched parental pairs and their hybrids across different conditions, preventing the reanalysis of their data for the purpose of assessing the impact of hybridization and how it compares with environmental effects.

The direct consequences of hybridization on the gene expression profiles of parental species have been mostly studied in plants and animals (McManus et al., 2010; Yoo et al., 2013; Li et al., 2014; Wu et al., 2018; Zhang et al., 2018). Though using different methodologies, all these studies report widespread transcriptomic changes following hybridization, 10–30% of the genes being significantly affected. In this context, fungal studies are more limited. Cox et al. (2014) did address this issue in the natural fungal diploid hybrid (allopolyploid) *Epichloë* Lp1 by comparing its expression patterns with those in its haploid parental species. The authors found that this natural hybrid retained most gene copies of the two parental species and, most importantly, that these genes generally retained the gene expression levels from the parental counterparts. In addition, differences in expression between homeologous genes tended to be lower than the corresponding differences between the orthologous genes in the parental species. Based on these findings, the authors concluded that the transcriptional response to hybridization was largely buffered. However, being based on a natural hybrid, this study does not allow discarding the possibility that the lack of strong differences in the gene expression is due to amelioration through compensatory mutations subsequent to the hybridization. In addition, by comparing a diploid hybrid to haploid parentals, that study could not disentangle the effects of ploidy change from those of hybridization.

We here set out to directly assess the immediate transcriptional impact of hybridization and compare it with the

effect of an environmental stress. To this end, we conducted an integrative multi-omics study comparing two distantly related fungal species—*S. cerevisiae* (SC) and *S. uvarum* (SU)—and their newly made hybrid at two thermal conditions. Using RNA-Seq, we assessed the transcriptional differences between orthologous genes in the parental species, between genes in the parental and the hybrid genetic background, and between homeologous genes coexisting in the hybrid. To compare the relative impact of hybridization with an environmental stress, we performed these experiments at two different temperatures, of which one affects the two parental species differently. We further investigated the consequences of hybridization on chromatin states by performing an assay for transposase-accessible chromatin using sequencing (ATAC-Seq) and integrated its results with our RNA-Seq data to obtain mechanistic insights behind the transcriptomic alterations caused by interspecific hybridization.

MATERIALS AND METHODS

Strains

The diploid hybrids of *S. cerevisiae* and *S. uvarum* were generated as follows: genetically tractable isogenic *MATa* and *MATα* haploids of the North American *S. cerevisiae* strain YPS128, isolated from the bark of an oak tree, were previously generated (Cubillos et al., 2009; Liti et al., 2009) by first isolating a single meiotic spore from the wild-type homothallic strain, resulting in complete homozygosity across the genome except for the *MAT* locus. The *HO* gene was then replaced by a hygromycin resistance cassette, resulting in a diploid heterozygous for *HO*. Haploid spores (*ho:HYG MATa* or *MATα*) were then isolated from this and *URA3* was replaced in these by the G418 resistance cassette *KANMX*. Similarly, the *S. uvarum* strain UWOPS99-807.1.1, isolated from Argentina, was dealt with in the same way, resulting in isogenic haploids of both mating types (Wimalasena et al., 2014). Diploid hybrids were formed by mating the *MATa* *S. cerevisiae* to the *MATα* *S. uvarum* and vice versa.

Experimental Conditions and RNA Extraction

Samples for RNA extraction were collected during the mid-exponential growth phase in rich medium [yeast extract peptone dextrose (YPD)] at two different temperatures: 30°C (normal growing temperature for those species; Salvadó et al., 2011) and 12°C (“cold-shock” condition).

Experiments were performed as follows: first, to delimit the timing of the mid-exponential growth phase, growth curves were obtained for each considered strain individually. For this, each strain was streaked from our glycerol stock collection onto a YPD agar plate and grown for 3 days at 30°C. Single colonies were cultivated in 15 ml YPD medium in an orbital shaker (30°C, 200 r/min, overnight). Then, each sample was diluted to an optical density (OD) at 600 nm of 0.2 in 50 ml of YPD and grown for 3 h in the same conditions (30°C, 200 r/min). Then, dilutions were made again to reach an OD of 0.1 in 50 ml of YPD in order to start all experiments with approximately the same amount of cells. The increasing growth was investigated

in parallel with manual measurement of the OD from the 50-ml samples and in 100-μl samples by a microplate reader (TECAN Infinite M200). For manual measurements, we inspected the absorbance in 1 ml every 1 h. For automated measurements, the samples were centrifuged for 2 min at 3000 × *g*, washed with 1 ml of sterile water, and centrifuged again for 2 min at 3000 × *g*. The pellet was resuspended in 1 ml of sterile water. Finally, 5 μl of each sample was inoculated in 95 μl of YPD in a 96-well plate. All experiments were run in triplicate. Cultures were grown in 96-well plates at 30°C for 24 h and monitored to determine the OD every 10 min with the microplate reader. Both manual and automated OD readouts showed similar growth patterns.

Once the mid-exponential phase was determined at around 5 h for all three species, the above-mentioned protocol was repeated until all samples were growing at the exponential phase and then the cultures were centrifuged at a maximum speed of 16,000 × *g* to harvest 3×10^8 cells per sample. For the cold shock experiments, when samples reached the mid-exponential phase, they were grown for 5 h more at 12°C and then the cells were harvested as described above. Total RNA from all samples was extracted using the RiboPure RNA Yeast Purification Kit (ThermoFisher Scientific) according to the manufacturer's instructions. Total RNA integrity and the quantity of the samples were assessed using the Agilent 2100 Bioanalyzer with the RNA 6000 Nano LabChip Kit (Agilent) and NanoDrop 1000 Spectrophotometer (Thermo Scientific).

RNA-Seq Library Preparation and Sequencing

Libraries were prepared using the TruSeq Stranded mRNA Sample Prep Kit v2 (ref. RS-122-2101/2, Illumina) according to the manufacturer's protocol. All reagents subsequently mentioned are from this kit, unless specified otherwise. Of the total RNA, 1 μg was used for poly(A)-mRNA selection using streptavidin-coated magnetic beads. Subsequently, the samples were fragmented to approximately 300 bp. Complementary DNA (cDNA) was synthesized using reverse transcriptase (SuperScript II, Invitrogen) and random primers. The second strand of the cDNA incorporated dUTP in place of dTTP. Double-stranded DNA (dsDNA) was further used for library preparation. dsDNA was subjected to A-tailing and ligation of the barcoded TruSeq adapters. All purification steps were performed using AMPure XP Beads (Agencourt). Library amplification was performed by PCR on the size-selected fragments using the primer cocktail supplied in the kit. Final libraries were analyzed using Agilent DNA 1000 chip (Agilent) to estimate the quantity and check fragment size distribution, which were then quantified by qPCR using the KAPA Library Quantification Kit (KapaBiosystems) prior to amplification with Illumina's cBot. Libraries were loaded and sequenced 2 × 50 or 2 × 75 on Illumina's HiSeq 2500.

Genome-Wide Chromatin Accessibility Profiling by ATAC-Seq

The two studied strains, namely, SC and the hybrid, were grown to the mid-exponential phase in YPD at 30°C, as described above,

and subjected to an assay for transposase-accessible chromatin with high-throughput sequencing (ATAC-Seq). This procedure was performed as described in Buenrostro et al. (2015) and Schep et al. (2015), with slight modifications.

For the cell nuclei preparation, approximately five million cells (counted with a hemocytometer) were harvested (centrifugation at $500 \times g$ for 5 min, 4°C) and washed twice (centrifugations at $500 \times g$ for 5 min, 4°C) with 50 μ l of cold sorbitol buffer (1.4 M sorbitol, 40 mM HEPES-KOH, pH 7.5, 0.5 mM $MgCl_2$). We used Zymolyase 100-T (ZymoResearch) to remove the cell wall, and three different concentrations were tested before proceeding with the final experiments: 1, 3, and 5 μ l of Zymolyase 5 U/ μ l. We incubated the cells with the corresponding amount of zymolyase in 50 μ l of sorbitol buffer at 30°C for 30-min shaking at 300 r/min. Then, the cells were pelleted ($500 \times g$ for 10 min at 4°C) and washed twice with 50 μ l of cold sorbitol buffer (centrifugations at $500 \times g$ for 10 min, 4°C). Fresh pellets of fungal spheroplasts were brought to the Genomics Unit at the Centre for Genomic Regulation (CRG) for further transposase reactions and library preparations for ATAC-Seq. Briefly, the nuclei were resuspended in 50 μ l $1 \times$ TD buffer containing 2.5 μ l transposase (Nextera, Illumina). The transposase reaction was conducted for 30 min at 37°C. Library amplification and barcoding were performed with NEBNext Q5 Hot Start HiFi PCR Master Mix (New England Biolabs) using index primers, designed according to Buenrostro et al. (2015), at a final concentration of 1.25 μ M. PCR was conducted for 12–13 cycles. Library purification was performed with Agencourt AMPure XP beads (Beckman Coulter) and library size distribution was assessed using the Fragment Analyzer (AATI, Agilent) or the Bioanalyzer High-Sensitivity DNA Kit (Agilent). The use of 3 μ l of Zymolyase 5 U/ μ l was chosen as the optimal concentration for the experiments based on the visual inspection of the obtained profiles. ATAC-Seq libraries were quantified before pooling and sequencing using the real-time library quantification kit (KAPA Biosystems). Paired-end sequencing was performed on a HiSeq 2500 (Illumina) with 50 cycles for each read.

All experiments were performed in three biological replicates. All library preparation and sequencing steps were performed at the Genomics Unit of the Centre for Genomic Regulation (CRG), Barcelona, Spain.

Sequencing Reads Quality Control and Visualization

We used FastQC v0.11.6¹ and Multiqc v1.0 (Ewels et al., 2016) to perform quality control of raw sequencing data. Adapter trimming was performed by Trimmomatic v0.36 (Bolger et al., 2014) with TruSeq3 and Nextera adapters (for RNA-Seq and ATAC-Seq, respectively) using 2:30:10 parameters and the minimum read length of 30 bp. To visualize genomic/transcriptomic alignments and coverages, we used the Integrative Genomic Viewer v2.3.97 (IGV) (Robinson et al., 2011).

¹<http://www.bioinformatics.babraham.ac.uk/projects/fastqc>

RNA-Seq Analysis

RNA-Seq read mapping and summarization was performed using the splice-junction aware mapper STAR v2.5.2b (Dobin et al., 2012) with default parameters. For parental species, we mapped RNA-Seq data to the corresponding reference genomes, while for the hybrid strain, we mapped raw data to the combined *S. cerevisiae* \times *S. uvarum* reference genomes. Further, to assess the rates of reads originated from one species while mapped to another (i.e., cross-mapping, which possibly can bias the inference of the gene expression levels), we employed two approaches: (i) mapping the reads of each parental to the concatenated reference genome and then calculating the proportion of wrongly mapped reads to a different parental genome and (ii) using the tool Crossmapper v1.1.0 (Hovhannisyan et al., 2019), which simulates the data from both parental species, maps the reads to the concatenated genome, and calculates the cross-mapping statistics. The reference genomes and genome annotations were obtained from Ensembl (release 93; Zerbino et al., 2018) and www.saccharomycessensustricto.org (Scannell et al., 2011) for SC and SU, respectively. For SU, we merged ultrascaffolds and unplaced regions in one reference and converted GFF to GTF format using gffread v0.9.8 (Trapnell et al., 2012) utility.

One-to-one orthologs between SC and SU were retrieved from www.saccharomycessensustricto.org (Scannell et al., 2011). Differential gene expression and ASE were assessed using DESeq2 v1.18.0 (Love et al., 2014). For between-species comparisons, we included the matrix of gene lengths to DESeq2 object to account for their differences. Additionally, for ASE analysis (within-hybrid comparison), we supplied the DESeq2 object with the matrix of gene lengths using `normalizationFactors(dds) <- lengths/exp(rowMeans(log(lengths)))`, allowing DESeq2 to account only for the differences in gene lengths when calculating sizeFactors and ignoring the library size since the read counts for alleles come from the same library. For a gene to be considered differentially/allele-specifically expressed, we used a threshold of $|\log_2 \text{fold change}|$ (L2FC) > 1.5 and padj (adjusted p-value) < 0.01 , unless specified otherwise.

Differentially expressed (DE) genes were used in Gene Ontology (GO) enrichment analysis as implemented in *Saccharomyces* Genome Database (Cherry et al., 2012) to find functional enrichments in biological process, molecular function, and cellular component GO categories. GO enrichment analysis for SU was done based on SC orthologous genes. To visualize the gene expression data, we utilized ggplot2 v2.3.0.0 R library (Wickham, 2016).

ATAC-Seq Analysis

Data generated by ATAC-Seq were mapped to the corresponding reference genomes using BWA v0.7.17-r1188 (Li and Durbin, 2010) with the MEM algorithm.

Initial mapping showed that ~15–18% of reads mapped to two regions of chromosome XII (450915–469179 and 489349–490611), which contain highly repetitive ribosomal RNA (rRNA) genes of SC. Thus, to remove the adverse effects in further

analysis, we have masked these two regions with bedtools maskfasta v.2.27.1 (Quinlan, 2014).

PCR duplicates were marked using Picard MarkDuplicates v.2.9.2 function². We used MACS2 v.2.1.1 (Zhang et al., 2008; Li and Durbin, 2010) to perform peak calling and the bedtools genomecov to generate bedgraph files of genome coverage by ATAC-Seq reads.

Bioconductor package DiffBind v.2.4.8 (Ross-Innes et al., 2012) was used to perform general quality control and occupancy and affinity analysis of the ATAC-Seq peaks. By occupancy analysis, DiffBind finds the overall peak set between replicates of a given biological condition and/or identifies the consensus peaks between different biological conditions (i.e., parental peak set and peak set of the hybrid), while in affinity analysis it performs differential accessibility analysis of corresponding peaks, which is based on the DESeq2 workflow.

For comparing the peak sets between parentals and the corresponding homeologous chromosomes in the hybrid, we split the bam files of the hybrid into separate files for the SC and SU chromosomes using SAMtools v.1.3.1 (Li et al., 2009).

To perform differential accessibility (affinity) analysis within the hybrid, we first defined the orthologous/homeologous promoter regions as upstream, non-coding, and genomic regions up to 1 kb of length at each one-to-one orthologous locus. Defined promoter regions were usually shorter than 1 kb since the neighboring genes or chromosome borders were often encountered within that distance. We obtained bed files of the promoter regions for each species using custom python scripts. Based on the bed files, we quantified the overlapping ATAC-Seq reads within the promoter regions using bedtools multicov function. Further, differential accessibility analysis was performed using DESeq2 controlling for the length of regions.

We used bedtools closest and custom python scripts to define, for each peak, the closest upstream and downstream genes within 1-kb distance and for which the ATAC-Seq peak falls within the promoter region (Supplementary Figure S1).

Transcription Factor Footprinting

Besides defining open chromatin regions, we used ATAC-Seq data to perform transcription factor (TF) footprinting in order to identify potential differences in TF binding site occupancy between the parental and the hybrid. Position weight matrices for the available *S. cerevisiae* TFs ($n = 176$) were retrieved from the Jasp database (Khan et al., 2018). Footprinting was performed using the HINT software of Regulatory Genomics Toolbox v.0.11.4 (RGT) package (Gusmao et al., 2014; Khan et al., 2018; Li et al., 2019). Fungal organisms were added to HINT following the recommendations of the package developers. The *Motifanalysis* function of the RGT package was used to match the motifs of fungal TFs with the ATAC-Seq footprints. We used the *differential* function of HINT to carry out differential TF binding site occupancy analyses and generate footprinting plots. The potential targets of differentially active TFs were identified using Yeasttract platform (Teixeira et al., 2018) by setting the

Regulation filters to “DNA binding and expression evidence” to account only for target genes with strong experimental evidence.

All custom scripts used in this study are available at https://github.com/Gabaldonlab/Hybrid_project. Raw sequencing data of the RNA-Seq and ATAC-Seq experiments were deposited in the Sequence Read Archive under the accession numbers SRR10246851-SRR10246868 and SRR10261591-SRR10261596, respectively.

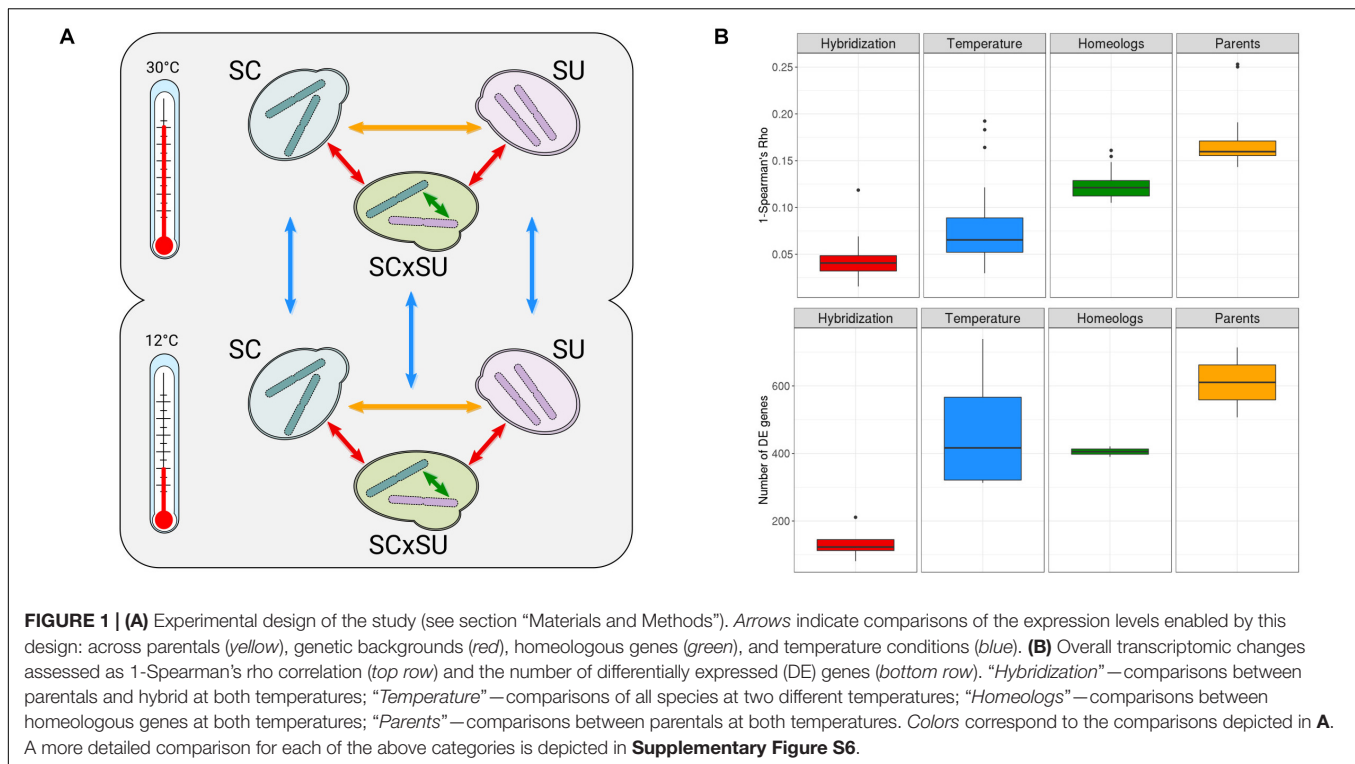
RESULTS

Limited Transcriptional Impact of Hybridization

To assess the impact of hybridization on gene expression, we used an RNA-Seq approach to profile the transcriptomes of diploid strains of *S. cerevisiae*, *S. uvarum*, and a *de novo* reconstructed diploid hybrid strain between these two species (see section “Materials and Methods”). We repeated the experiment at 30°C, a temperature within the optimal growth range of both species, and at 12°C, which represents a cold shock, particularly for the non-cryotolerant *S. cerevisiae* (Salvadó et al., 2011). This experimental design allowed us to directly compare transcriptional differences across genetic backgrounds (homozygous parentals and the hybrid), species (orthologous genes), homeologous chromosomes, and temperatures (see Figure 1A) and therefore assess the relative impact of these factors on gene expression levels.

As recommended for robust inference of the transcriptional levels (Liu et al., 2014), we performed all experiments in three biological replicates and sequenced over 30 million reads per replicate (see section “Materials and Methods”). The negligible level of cross-mapping between reads of the two species, as assessed by Crossmapper (Hovhannisyan et al., 2019) (Supplementary File S1), and the independent mapping of parental RNA-Seq reads to both reference genomes (Supplementary Table S1B) allowed us to accurately assign reads to each parental sub-genome in the hybrid and, thus, infer the relative expression of each of the two homeologous alleles. To additionally test whether these negligible cross-mapping rates can influence downstream results, we compared the read counts obtained from mapping parental data to the combined reference and the counts obtained by mapping parental data to corresponding parental genomes. In the case of both species, we observed Spearman’s correlations > 0.99 and that differential expression analysis with relaxed filters ($|L2FC| > 1$, $p_{adj} > 0.05$) did not show any gene affected by cross-mapping, verifying the accuracy of read assignments to corresponding species. Mapping statistics are shown in Supplementary Table S1, and quality control and reproducibility metrics for all samples are available in Supplementary Figures S2–S4. Overall, we observed lower mapping rates of SU as compared to SC, which likely reflect the lower quality of the reference assembly for the former species. For each of the 11 pairwise comparisons depicted in Figure 1A, we performed differential expression analyses (Supplementary Tables S2–S13), tested for enrichment of functional GO terms among the DE genes (Supplementary

²<http://broadinstitute.github.io/picard>



Tables S2–S13), and assessed the correlation between the levels of expression (**Figure 1B**).

Although not the focus of our research, we found that the cryotolerant *S. uvarum* species had a more significant rewiring of its transcriptional landscape than did *S. cerevisiae*, especially upon exposure to the lower temperature, likely reflecting an adaptive response. The observed functional enrichments among the DE genes upon change in temperature were largely consistent with previous analyses on *S. cerevisiae* and *S. uvarum* (**Supplementary Tables S10, S11**), such as the upregulation of chaperone activity and trehalose catabolism in low temperatures for *S. uvarum* (Li and Fay, 2017). Upregulation of trehalose metabolism in the *S. uvarum* sub-genome was also observed in the hybrid when it was exposed to 12°C (**Supplementary Table S12**), which might be associated with the adaptation of the hybrid to low temperatures (Pérez-Torrado et al., 2015). Most importantly, a comparison of the relative level of transcriptional differences across genetic backgrounds, temperatures, homeologous genomes, and species (**Figure 2B**) shows that hybridization has a rather reduced transcriptional impact, being significantly lower than that observed for the temperature shift.

Overall, only 81 and 123 genes are DE when comparing the hybrid and parental genetic backgrounds for *S. cerevisiae* and *S. uvarum*, respectively (**Supplementary Tables S2, S3**) at 30°C, which represents between 1 and 2% of the total gene repertoire of each species. In comparison, a temperature shift significantly alters the expression of 509 (7.1%) and 739 (11.5%) genes in these species, respectively. Additionally, we observed that the differences in expression between orthologous genes in the two

species were significantly larger than those observed between homeologous genes in the hybrid. This indicates that interspecies differences in terms of transcriptional landscape are attenuated rather than increased in the hybrid.

Low Levels of Allelic Imbalance in Yeast Hybrid

We next explored ASE in the hybrid. Consistent with the largely conserved transcriptional landscape after hybridization, we found relatively low levels of allelic imbalance (i.e., significantly different expression levels of the two homeologous genes) (**Figure 1B**). Specifically, at 30°C, 390 (~7.4% of the homeologous pairs; **Supplementary Table S5**), homeologous genes in the hybrid show allelic imbalance, from which 180 genes show higher expression of the SU allele while 210 show higher expression of the SC allele. Thus, there is no strong preference for the hybrid to express one of the two sub-genomes. To identify whether the genes with allelic imbalance in the hybrid were a consequence of hybridization or were already DE when comparing the parental species (ASE inheritance), we compared the list of DE genes between parental species (**Supplementary Table S4**) with the list of imbalanced homeologous genes (**Figure 2**).

At 30°C (**Figure 2A**), 68% (143/210) and 51% (92/180) of the genes preferentially expressing the SC or SU alleles, respectively, were also found to have differential expression (with the same direction) in comparisons across species. Hence, this result indicates that the majority of genes with allele-specifically expressed genes in the hybrid are driven by inheritance of expression levels from parental species rather than resulting

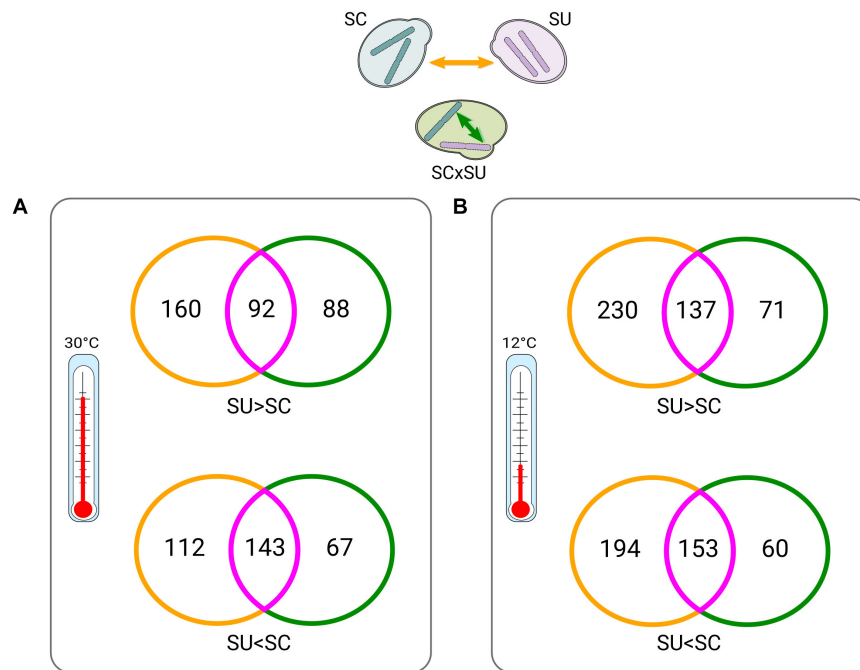


FIGURE 2 | Venn diagrams of between-parent (yellow) versus within-hybrid (green) comparisons (depicted on the top) at 30°C (**A**) and at 12°C (**B**). Intersections (violet) indicate differentially expressed (DE) genes that appear in both conditions. Numbers indicate DE genes. Colors of the Venn diagrams correspond to the types of comparisons, as indicated by the arrows in the top scheme and consistent with **Figure 1A** (except for intersections). “>” and “<” symbols denote which homeologs or orthologs of a given species are expressed at significantly higher and lower levels, respectively (see Supplementary Tables 4, 5, 8, 9 for lists of the DE genes).

from the hybridization. Additionally, we found fewer genes that acquired ASE in the hybrid without being DE across species (88 and 67) as compared to genes that show no ASE despite being DE across the two species (160 and 112).

Overall, similar trends were found at 12°C (**Figure 2B**). Collectively, these observations suggest that hybridization tends to attenuate, rather than exacerbate, differences in the expression levels of parental orthologous genes. Finally, we also found that there is a small set ($n = 40$) of temperature-dependent allele-specifically expressed genes (**Supplementary Table S13**), which is congruent with an early study (Li and Fay, 2017).

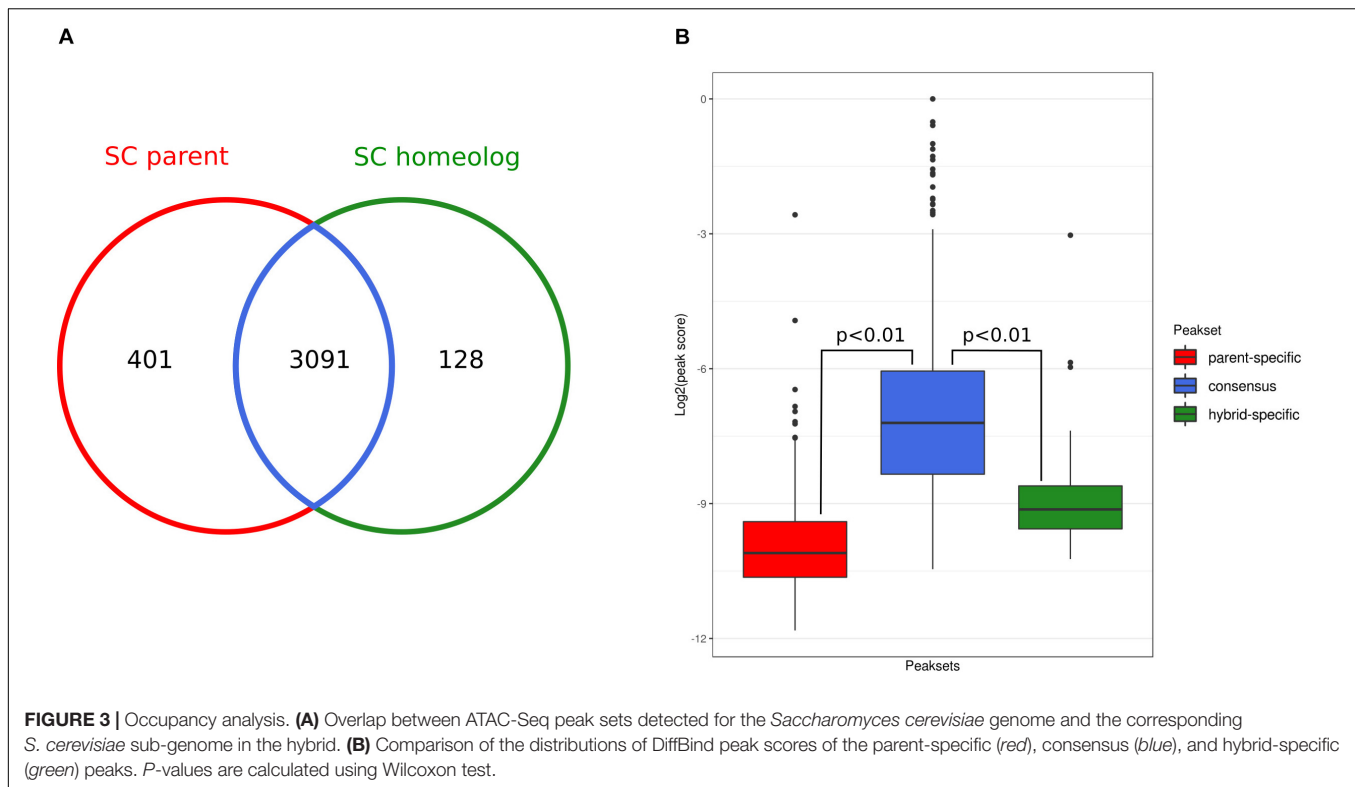
Overall Conservation of Genome-Wide Chromatin Accessibility Patterns After Hybridization

We further investigated gene regulation differences upon hybridization by performing genome-wide chromatin accessibility analysis based on ATAC-Seq at 30°C of the hybrid and the *S. cerevisiae* parental (see section “Materials and Methods” and **Supplementary Table S14**). We compared the ATAC-Seq profiles by performing peak calling and comparing the overlap between called peaks (i.e., inferred open chromatin regions) in the parental and the SC sub-genome of the hybrid (**Supplementary Figure S5**). After removing one outlier (see section “Materials and Methods”), we found that replicate experiments showed a large overlap of the called

peaks (83% for both parent and hybrid replicates). Then, we compared peak sets of the SC parental with the corresponding sub-genome in the hybrid.

This analysis showed that the state of chromatin accessibility is largely similar between the SC parental genome and the corresponding sub-genome of the hybrid (**Figure 3A**). From 3492 parental open chromatin regions, 3,091 (88%, consensus peak set) are present in the hybrid, and conversely, 96% of the hybrid SC sub-genome peaks are shared with the SC parental. Although they represent a small fraction, we did observe parent-specific ($n = 401$) and hybrid-specific ($n = 128$) accessible chromatin regions. However, we found that these specific peaks have significantly lower scores than did the shared peaks (**Figure 3B**), suggesting that some of these differences might represent false-positive peak calls.

Nevertheless, to assess whether parent- and hybrid-specific peaks in chromatin accessibility were driving the observed transcriptional changes, we integrated the ATAC-Seq and RNA-Seq data. For each parental- and hybrid-specific ATAC-Seq peak, we identified the nearest downstream gene for each strand, which potentially could be regulated by the open chromatin region identified by the peak. Only one gene near a parental-specific peak was found to be also overexpressed with respect to the hybrid context: the gene encoding pyridoxal-5'-phosphate synthase (YFL059W). Conversely, the gene coding for the NADH-dependent aldehyde reductase (YKL071W) was overexpressed in the hybrid and was sitting downstream of a



hybrid-specific chromatin accessibility peak. The low fraction of peaks that are specific to each genetic background, their low scores, and the very low number of downstream genes that actually show differential expression suggest that changes in chromatin accessibility upon hybridization have a very limited impact at the transcriptional level.

Further, we performed differential chromatin accessibility analysis (i.e., affinity analysis) within the consensus peak set, shared between the parental and the hybrid. Even when using liberal thresholds [L2FC > 1 and false discovery rate (FDR) < 0.01], we found only two differentially accessible peaks in this consensus set: namely, regions at chromosomes XII 682106-682709 (more open in the SC parent, L2FC = 1.3) and IX 391660-392121 (more open in the hybrid, L2FC = 1.32). We combined this result with RNA-Seq and visualized the integrated data (Figure 4).

In the first region (Figure 4A), the gene *YLR271W* that is downstream of the peak is not DE between the parent and hybrid. In contrast, the second peak (Figure 4B) coincides with the significantly higher expressed gene *FLO11* (YIR019C; L2FC = 3.44, padj < 0.01) in the hybrid as compared to the parent. However, in this case, the peak entirely overlaps the gene, and therefore it is unlikely that it regulates its expression. Thus, differential levels of chromatin accessibility do not seem to drive the few differences observed between hybrids and parental genetic backgrounds.

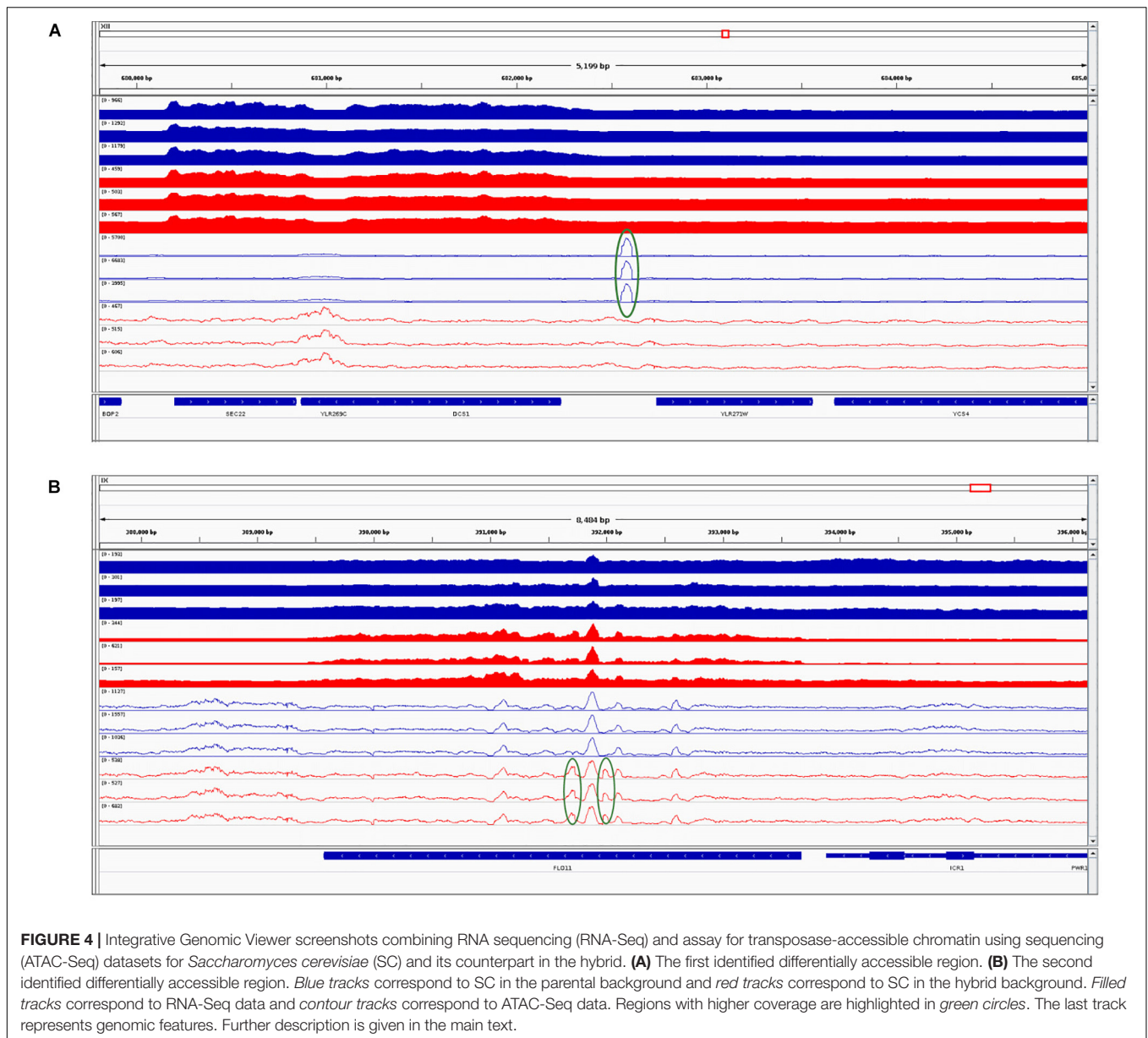
Next, taking advantage of the high sequencing depth of our ATAC-Seq data, we also assessed the changes in TF activity (as defined by Li et al., 2019) in a genome-wide manner. The

results show (Figure 5) that only eight out of 176 *S. cerevisiae* TFs have significantly ($p < 0.05$) changed their activity levels upon hybridization. In most cases, the differences in activity are moderate, with the notable exception of ARG81, which mediates the arginine-dependent repression of arginine biosynthesis genes and the activation of arginine catabolic genes.

We further identified the potential target genes for each of the deregulated TFs (Supplementary Table S15) and compared the target genes with the DE genes upon hybridization. From the 189 target genes of these eight TFs, only two genes corresponded to the genes DE in the same direction as their TF—*YHL040C* (gene encoding for siderochrome iron transporter) and *YLR346C* (encodes for a protein of unknown function), were upregulated in the hybrid genetic background. Both of these genes are regulated by BAS1, a TF involved in regulating the expression of genes of the purine and histidine biosynthesis pathways. Overall, our results suggest that, upon hybridization, the changes in TF activities are subtle and largely do not correlate with the patterns of differential gene expression observed by RNA-Seq.

Chromatin Accessibility Patterns Within the Hybrid Are Weakly Correlated With Allele-Specific Gene Expression

Finally, we compared the chromatin accessibility profiles of the SC and SU homeologous regions within the hybrid. First, we defined the homeologous regions between two sub-genomes as maximum of 1 kb upstream regions of each one-to-one orthologous gene. Further, we quantified the ATAC-Seq coverage for these regions and performed differential



accessibility analysis using DESeq2, controlling for the length differences of the regions. This analysis identified 59 and 75 genomic regions that were significantly more open or less open, respectively, in the SU sub-genome as compared to the SC sub-genome ($|L2FC| > 1$, $p_{adj} < 0.01$). By comparing these data with the results of ASE, we found that eight out of 180 preferentially expressed SU homeologs coincided with the significantly more open SU regions and that nine of 210 preferentially lower expressed SU homeologs corresponded to the significantly less open SU regions. These results are in agreement with a previously published study which identified a low correlation between changes in nucleosome positioning and gene expression levels in yeasts (Tirosh et al., 2010).

DISCUSSION

Fungi, and in particular *Saccharomycotina* yeasts, have been shown to be prone to hybridization, with an increasing number of hybrid species that are highly successful in certain niches and have industrial or clinical relevance (Pryszcz et al., 2014; Krogerus et al., 2017; Mixão et al., 2019). Hybridization has also been shown to be at the root of entire clades, e.g., the post-whole-genome duplication clade comprising *Saccharomyces* and related genera has been shown to result from a hybridization event (Marcet-Houben and Gabaldón, 2015). Thus, rather than representing evolutionary dead-ends, fungal hybrids might be highly successful and long-lived. This implies that fungal species which form a hybrid organism must overcome

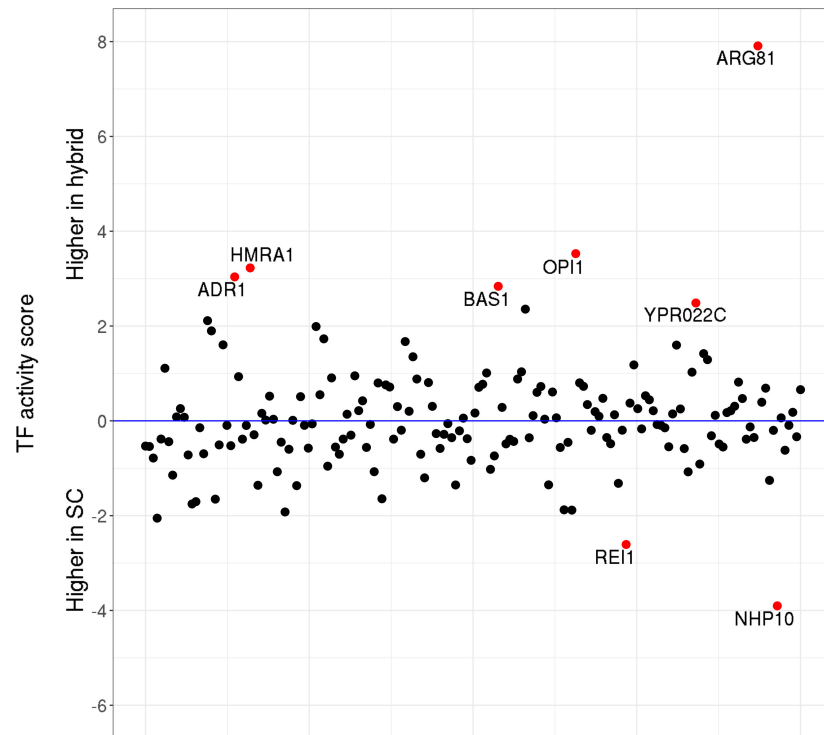


FIGURE 5 | Transcription factor (TF) activity scores. Relative activity levels between *Saccharomyces cerevisiae* parental and the hybrid counterparts. Red dots highlight the TFs which significantly ($p < 0.05$) changed their activity levels upon hybridization.

molecular differences and potential incompatibilities which evolved through the evolutionary history of the parents. On the relatively well-studied genomic level, fungal hybrids, and in particular those from the *Saccharomyces* genus, tend to undergo genomic rearrangements, including loss of heterozygosity, gene conversion, and partial or full chromosome loss, among others, that help to overcome incompatibilities and stabilize genomes, which results in genome mosaicism.

In our study, we assessed how the two distantly related *Saccharomyces* species cope with hybridization at the levels of the transcriptome and chromatin landscapes.

Collectively, our results show that, despite genome merging of extremely diverged species, hybridization has a comparatively smaller effect on the transcriptome than a shift from temperate to cold temperature. Moreover, we found that most loci express the two homeologous alleles in similar proportions, and those genes that show ASE largely overlap with those whose orthologs in the two parental species already show different levels of expression, suggesting that most of the ASE derives from already existing interspecies differences. Furthermore, homeologous genes within the hybrid showed fewer differences than did orthologous genes in the two parental species, indicating that, rather than being exacerbated, interspecies differences in expression are attenuated upon hybridization. This is consistent with an earlier study on a natural hybrid of the genus *Epichloë* (Cox et al., 2014). However, as our model involves a newly formed hybrid, our study clarifies that the attenuation of the differences is an immediate effect of

hybridization and not the result of adaptation through evolution of the hybrid lineage. Additionally, Cox et al., proposed that with an increase in genome divergence between the parents, which $\sim 5\%$ in *Epichloë* (Campbell et al., 2017), the magnitude of transcriptome shock will increase accordingly, while our results demonstrate that despite a large evolutionary distance of 20% of nucleotide divergence in coding regions (Kellis et al., 2003) between SC and SU the consequences of hybridization are still buffered. Moreover, in contrast to the *Epichloë* study that compares haploid parents with a diploid hybrid, our study compares diploid parental strains and diploid hybrids and thus avoids any potential misleading effect resulting from a ploidy change.

The absence of a large impact of hybridization in gene expression in fungal hybrids is in stark contrast with what has been reported in animals or plant studies (McManus et al., 2010; Yoo et al., 2013; Li et al., 2014; Wu et al., 2018; Zhang et al., 2018), where widespread changes in expression following hybridization have been observed. For instance, in newly resynthesized allotetraploid *Brassica napus* (Wu et al., 2018), 30.4% of the genes showed expression changes upon hybridization compared to its diploid parents *Brassica rapa* and *Brassica oleracea*, and over 90% of the deregulated genes were downregulated in the allotetraploid compared to the parents. Additionally, 36.5% of homeologous pairs within the hybrid *B. napus* displayed differential expression toward either of the alleles, with a slight preference to the *B. rapa* parental. Similarly, in allopolyploid

cotton *Gossypium arboreum* (A2) × *Gossypium raimondii* (D5), 22–30% of parental genes showed differential expression in the hybrid as compared to the parentals (Yoo et al., 2013). The study of the allelic imbalance of a synthetic hexaploid wheat showed that 24.1% of the identified homeologous genes were imbalanced in the hybrid, and this difference in expression could not be attributed to preexisting expression divergence between the parentals (Li et al., 2014). Finally, a recent study on the diatom microalgae *Fistulifera solaris* showed that ~61% of homeologous genes displayed allelic imbalance (Nomaguchi et al., 2018).

For *Drosophila* hybrids, different amplitudes of transcriptome misexpression have been previously reported, which depended on the level of genetic divergence of the parental species. For instance, a recent transcriptomic study of *Drosophila mojavensis* and *Drosophila arizonae* (diverged 0.6–1 mya) and their hybrid showed that 12% of genes in the hybrid are DE as compared to the parents (Lopez-Maestre et al., 2017). This is much larger than the fraction of DE genes in this study despite the much lower genetic distance in the *Drosophila* species. The same study showed that 8% of genes between parentals have diverged expression. That is, in that case, differential expression between homeologous genes in the hybrid was more abundant than between orthologous genes in the two parental species, which is the contrary of what we have observed for the *Saccharomyces* hybrid in this study. On the other hand, the comparison between more diverged fly species, namely, *Drosophila melanogaster* and *Drosophila sechellia* (diverged ~1.2 mya), identified 78% of DE genes between parents (McManus et al., 2010). Interestingly, a recent study of hybrid chicken breeds (intraspecies hybrids) showed tissue-dependent rates of gene expression divergence: while it was ~15% of genes in the liver, gene expression divergence between parental breeds in the brain was as low as 0.8% of genes (Gu et al., 2019).

It must be noted that comparing results across species and studies is difficult. These studies were performed using different technologies and data analysis methods, which makes direct comparisons problematic. For instance, Wu et al. (2018) used fragments per kilobase of transcript per million mapped reads (FPKM) expression values and applied a filter of FDR ≤ 0.05 and absolute log₂ fold change ≥ 1 for a gene to be considered as DE. Yoo et al. (2013) used raw read counts for differential expression analysis and a filter of adjusted *p*-value < 0.05 with no filtering on fold change. Gu et al. (2019) applied an absolute fold change ≥ 1.25 and FDR < 0.5 for assigning DE genes in chicken hybrid breeds, and Lopez-Maestre et al. (2017) used FDR < 0.01 and log₂ fold change > 1.5 for assigning DE genes in *Drosophila*, which are the filters also used in our study. In order to assess how data filtering can influence our results, we applied a set of more liberal filters for finding DE genes upon parental–hybrid transition. With padj (FDR) < 0.05, |log₂ fold change| > 1, and mean normalized expression levels > 10 (which, in fact, take into account only the expressed part of the transcriptome, limiting transcriptome-wide inferences), we obtained, on average, ~4.6 and ~9% of DE genes for SC and SU, respectively, across temperatures. This shows that, even with relaxed filters, transcriptome shock in our yeast hybrid is lower

than in plants and animals. Thus, methodological differences notwithstanding, the different animal and plant studies seem to agree in reporting a large transcriptomic impact of hybridization, as well as large levels of allele-specific imbalance, whereas the fungal studies consistently report more moderate effects.

We further assessed the impact of hybridization on another level: that of chromatin accessibility. Here, consistent with the low level of differences in gene expression, we found minor differences in terms of chromatin accessibility and TF activity between the hybrid and the *S. cerevisiae* parental. Admittedly, subtle differences in TF expression might have significant biological effects. However, our results suggest that the few observed differences in chromatin accessibility and TF activity are not driving the few observed differences in gene expression levels.

Based on our results and those from other previous studies, we hypothesize that unicellular fungi and multicellular plants and metazoans respond fundamentally differently to hybridization in terms of transcriptional response, which may explain why hybridization is so common in fungi, and, as compared to plants and animals, it can encompass larger genetic distances (Morales and Dujon, 2012). What molecular phenomena govern these different responses to hybridization? One could argue that differences in the magnitudes of transcriptomic shock in fungi and metazoans and plants can be attributed to differences in the mechanisms regulating gene expression. Though the general and fundamental principles of transcriptional regulation are largely conserved across eukaryotes, the complexity of gene regulation in plants and animals is more sophisticated than that in fungi (Rando and Chang, 2009; Hahn and Young, 2011; Lelli et al., 2012). For example, plants and animals possess a richer repertoire of chromatin modification regulators as compared to yeasts, which provide them with additional layers of regulation and a more sophisticated fine-tuning of the expression levels (Rando and Chang, 2009).

Additionally, fungi and yeasts in particular are prone to genomic rearrangements, ranging from small indels to large-scale copy number variations, inversions, translocations, and duplications (Albertin and Marullo, 2012; Plissonneau et al., 2016; Möller and Stukenbrock, 2017; Möller et al., 2018; Steenwyk and Rokas, 2018). Not only are these genomic alterations compatible with fungal viability but also, inversely, can promote fitness and adaptability to different niches (Selmecki et al., 2005, 2009; Croll et al., 2013; de Jonge et al., 2013; Gabaldón and Carreté, 2016; Mixão and Gabaldón, 2018). Hence, one could expect that, following hybridization, two complex gene regulatory systems, such as that of animals and plants, are more likely to experience larger levels of incompatibilities and perturbations as compared to simpler and more versatile regulatory systems, such as those of yeasts.

In this context, Cox et al. (2014) introduced the concept of “modulon” which encompasses all gene regulatory mechanisms, including *cis*- and *trans*-regulation, posttranscriptional regulation, TFs, epigenetics, etc. Differences in the levels of expression of orthologous genes in different species arise due to differences in the species’ modulons, which have evolved independently for some time. Upon hybridization, several regulatory scenarios can take place: (i) modulons of the two

species have no or little crosstalk with each other because they are too divergent; (ii) modulons are largely similar and compatible with each other, resulting in a so-called homeolog expression blending; or (iii) modulons of the two species preferentially target one of the alleles. Importantly, these regulatory outcomes can coexist, affecting different portions of the transcriptome, which can be quantitatively assessed. In the first scenario, the genes from the parental species would inherit their expression levels in the hybrid with no subsequent expression alterations. This outcome accounts for the majority of genes in our study: ~92 and ~89.3% of orthologous genes at 30 and 12°C, respectively (non-DE orthologs + violet parts in **Figure 2**). The second scenario will result in diminished differences in homeologous expression levels as compared to differences across species. In our study, this could account for ~5.24 and ~8.2% of orthologous genes at 30 and 12°C, respectively (yellow parts in **Figure 2**). In the third case, homeologous genes will acquire divergence in gene expression that was not observed in parentals, which represents the transcriptomic shock caused by hybridization. In our study, this accounts for ~2.9 and ~2.52% of homeologous genes at 30 and 12°C, respectively (green parts in **Figure 2**).

Altogether, our study suggests a conservative and restricted impact of hybridization at the transcriptomic and chromatin profiles in hybrid yeast, which can be largely attributed to the absence of a regulatory crosstalk between highly diverged fungal modulons. We hypothesize that the moderate impact that hybridization has on the levels of chromatin accessibility and gene expression is at the root of the strong ability for successful hybridization in yeasts and other fungi. Further research involving diverse taxonomic groups of fungi is required to address this hypothesis in order to disentangle the role of transcriptome and chromatin profile buffering in fungal hybridization.

DATA AVAILABILITY STATEMENT

Raw sequencing data of RNA-Seq and ATAC-Seq experiments were deposited in the Sequence Read Archive under the accession numbers SRR10246851-SRR10246868 and SRR10261591-SRR10261596, respectively.

REFERENCES

- Albertin, W., and Marullo, P. (2012). Polyploidy in fungi: evolution after whole-genome duplication. *Proc. Biol. Sci.* 279, 2497–2509. doi: 10.1098/rspb.2012.0434
- Baker, E., Wang, B., Bellora, N., Peris, D., Hulfachor, A. B., Koshalek, J. A., et al. (2015). The genome sequence of *Saccharomyces eubayanus* and the domestication of lager-brewing yeasts. *Mol. Biol. Evol.* 32, 2818–2831. doi: 10.1093/molbev/msv168
- Bolger, A. M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30, 2114–2120. doi: 10.1093/bioinformatics/btu170
- Buenrostro, J. D., Wu, B., Chang, H. Y., and Greenleaf, W. J. (2015). ATAC-seq: a method for assaying chromatin accessibility genome-wide. *Curr. Protoc. Mol. Biol.* 109, 21.29.1–21.29.9.
- Campbell, M. A., Tapper, B. A., Simpson, W. R., Johnson, R. D., Mace, W., Ram, A., et al. (2017). *Epichloë hybrida*, sp. nov., an emerging model system for investigating fungal allopolyploidy. *Mycologia* 109, 715–729. doi: 10.1080/00275514.2017.1406174
- Cherry, J. M., Hong, E. L., Amundsen, C., Balakrishnan, R., Binkley, G., Chan, E. T., et al. (2012). *Saccharomyces* genome database: the genomics resource of budding yeast. *Nucleic Acids Res.* 40, D700–D705. doi: 10.1093/nar/gkr1029
- Combes, M.-C., Hueber, Y., Dereeper, A., Rialle, S., Herrera, J.-C., and Lashermes, P. (2015). Regulatory divergence between parental alleles determines gene expression patterns in hybrids. *Genome Biol. Evol.* 7, 1110–1121. doi: 10.1093/gbe/evv057
- Cox, M. P., Dong, T., Shen, G., Dalvi, Y., Scott, D. B., and Ganley, A. R. D. (2014). An interspecific fungal hybrid reveals cross-kingdom rules for allopolyploid gene expression patterns. *PLoS Genet.* 10:e1004180. doi: 10.1371/journal.pgen.1004180
- Croll, D., Zala, M., and McDonald, B. A. (2013). Breakage-fusion-bridge cycles and large insertions contribute to the rapid evolution of accessory chromosomes in a fungal pathogen. *PLoS Genet.* 9:e1003567. doi: 10.1371/journal.pgen.1003567

AUTHOR CONTRIBUTIONS

TG, ES, and HH designed the study. ES and EK performed the experiments. HH analyzed the data. HH and TG interpreted the results and prepared the manuscript. AH and EL constructed the strains used in the study. All the authors have read and approved the final manuscript.

FUNDING

This work was funded in part by the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement no. H2020-MSCA-ITN-2014-642095. TG group also acknowledges support from the Spanish Ministry of Economy, Industry, and Competitiveness (MEIC) for the EMBL partnership, and grants "Centro de Excelencia Severo Ochoa" SEV-2012-0208, and BFU2015-67107 co-funded by European Regional Development Fund (ERDF); from the CERCA Program/Generalitat de Catalunya; from the Catalan Research Agency (AGAUR) SGR857; and grants from the European Union's Horizon 2020 Research and Innovation Program under the grant agreement ERC-2016-724173. TG also receives support from an INB grant (PT17/0009/0023—ISCIII-SGEFI/ERDF).

ACKNOWLEDGMENTS

We thank Dr. Susana Iraola (Comparative Genomics group, BSC/IRB, Barcelona, Spain) for her support in the laboratory experiments. We also thank the team of rgt-hint software developers for the technical support and valuable recommendations in ATAC-Seq footprinting analysis.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fgene.2020.00404/full#supplementary-material>

- Cubillos, F. A., Louis, E. J., and Liti, G. (2009). Generation of a large set of genetically tractable haploid and diploid *Saccharomyces* strains. *FEMS Yeast Res.* 9, 1217–1225. doi: 10.1111/j.1567-1364.2009.00583.x
- de Jonge, R., Bolton, M. D., Kombrink, A., van den Berg, G. C., Yadeta, K. A., and Thomma, B. P. (2013). Extensive chromosomal reshuffling drives evolution of virulence in an asexual pathogen. *Genome Res.* 23, 1271–1282. doi: 10.1101/gr.152660.112
- Depotter, J. R., Seidl, M. F., Wood, T. A., and Thomma, B. P. (2016). Interspecific hybridization impacts host range and pathogenicity of filamentous microbes. *Curr. Opin. Microbiol.* 32, 7–13. doi: 10.1016/j.mib.2016.04.005
- Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., et al. (2012). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 29, 15–21. doi: 10.1093/bioinformatics/bts635
- Dujon, B. (2010). Yeast evolutionary genomics. *Nat. Rev. Genet.* 11, 512–524. doi: 10.1038/nrg2811
- Dutta, A., Lin, G., Pankajam, A. V., Chakraborty, P., Bhat, N., Steinmetz, L. M., et al. (2017). Genome dynamics of hybrid during vegetative and meiotic divisions. *G3*, 7, 3669–3679. doi: 10.1534/g3.117.1135
- Ewels, P., Magnusson, M., Lundin, S., and Käller, M. (2016). MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* 32, 3047–3048. doi: 10.1093/bioinformatics/btw354
- Gabaldón, T., and Carreté, L. (2016). The birth of a deadly yeast: tracing the evolutionary emergence of virulence traits in *Candida glabrata*. *FEMS Yeast Res.* 16:fov110. doi: 10.1093/femsyr/fov110
- Gladieux, P., Ropars, J., Badouin, H., Branca, A., Aguilera, G., de Vienne, D. M., et al. (2014). Fungal evolutionary genomics provides insight into the mechanisms of adaptive divergence in eukaryotes. *Mol. Ecol.* 23, 753–773. doi: 10.1111/mec.12631
- Goodwin, S., McPherson, J. D., and McCombie, W. R. (2016). Coming of age: ten years of next-generation sequencing technologies. *Nat. Rev. Genet.* 17, 333–351. doi: 10.1038/nrg.2016.49
- Graze, R. M., Novelo, L. L., Amin, V., Fear, J. M., Casella, G., Nuzhdin, S. V., et al. (2012). Allelic imbalance in *Drosophila* hybrid heads: exons, isoforms, and evolution. *Mol. Biol. Evol.* 29, 1521–1532. doi: 10.1093/molbev/msr318
- Greaves, I. K., Gonzalez-Bayon, R., Wang, L., Zhu, A., Liu, P.-C., Groszmann, M., et al. (2015). Epigenetic changes in hybrids. *Plant Physiol.* 168, 1197–1205. doi: 10.1104/pp.15.00231
- Groszmann, M., Greaves, I. K., Albertyn, Z. I., Scofield, G. N., Peacock, W. J., and Dennis, E. S. (2011). Changes in 24-nt siRNA levels in *Arabidopsis* hybrids suggest an epigenetic contribution to hybrid vigor. *Proc. Natl. Acad. Sci. U.S.A.* 108, 2617–2622. doi: 10.1073/pnas.1019217108
- Gu, H., Qi, X., Jia, Y., Zhang, Z., Nie, C., Li, X., et al. (2019). Inheritance patterns of the transcriptome in hybrid chickens and their parents revealed by expression analysis. *Sci. Rep.* 9:5750. doi: 10.1038/s41598-019-42019-x
- Guo, B., Chen, Y., Zhang, G., Xing, J., Hu, Z., Feng, W., et al. (2013). Comparative proteomic analysis of embryos between a maize hybrid and its parental lines during early stages of seed germination. *PLoS One* 8:e65867. doi: 10.1371/journal.pone.0065867
- Guo, M., Yang, S., Rupe, M., Hu, B., Bickel, D. R., Arthur, L., et al. (2008). Genome-wide allele-specific expression analysis using Massively Parallel Signature Sequencing (MPSSSTM) Reveals cis- and trans-effects on gene expression in maize hybrid meristem tissue. *Plant Mol. Biol.* 66, 551–563. doi: 10.1007/s11103-008-9290-z
- Gusmao, E. G., Dieterich, C., Zenke, M., and Costa, I. G. (2014). Detection of active transcription factor binding sites with the combination of DNase hypersensitivity and histone modifications. *Bioinformatics* 30, 3143–3151. doi: 10.1093/bioinformatics/btu519
- Hahn, S., and Young, E. T. (2011). Transcriptional regulation in *Saccharomyces cerevisiae*: transcription factor regulation and function, mechanisms of initiation, and roles of activators and coactivators. *Genetics* 189, 705–736. doi: 10.1534/genetics.111.127019
- Hovhannisyán, H., Hafez, A., Llorens, C., and Gabaldón, T. (2019). CROSSMAPPER: estimating cross-mapping rates and optimizing experimental design in multi-species sequencing studies. *Bioinformatics* 36, 925–927. doi: 10.1093/bioinformatics/btz626
- Hu, X., Wang, H., Diao, X., Liu, Z., Li, K., Wu, Y., et al. (2016). Transcriptome profiling and comparison of maize ear heterosis during the spikelet and floret differentiation stages. *BMC Genomics* 17:959. doi: 10.1186/s12864-016-3296-8
- Hu, X., Wang, H., Li, K., Wu, Y., Liu, Z., and Huang, C. (2017). Genome-wide proteomic profiling reveals the role of dominance protein expression in heterosis in immature maize ears. *Sci. Rep.* 7:16130. doi: 10.1038/s41598-017-15985-3
- Kellis, M., Patterson, N., Endrizzi, M., Birren, B., and Lander, E. S. (2003). Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature* 423, 241–254. doi: 10.1038/nature01644
- Khan, A., Fornes, O., Stigliani, A., Gheorghe, M., Castro-Mondragon, J. A., van der Lee, R., et al. (2018). JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework. *Nucleic Acids Res.* 46:D1284.
- Krogerus, K., Magalhães, F., Vidgren, V., and Gibson, B. (2017). Novel brewing yeast hybrids: creation and application. *Appl. Microbiol. Biotechnol.* 101, 65–78. doi: 10.1007/s00253-016-8007-5
- Le Jeune, C., Lollier, M., Demuyter, C., Erny, C., Legras, J.-L., Aigle, M., et al. (2007). Characterization of natural hybrids of *Saccharomyces cerevisiae* and *Saccharomyces bayanus* var. *uvarum*. *FEMS Yeast Res.* 7, 540–549.
- Lelli, K. M., Slattery, M., and Mann, R. S. (2012). Disentangling the many layers of eukaryotic transcriptional regulation. *Annu. Rev. Genet.* 46, 43–68. doi: 10.1146/annurev-genet-110711-155437
- Li, A., Liu, D., Wu, J., Zhao, X., Hao, M., Geng, S., et al. (2014). mRNA and small RNA transcriptomes reveal insights into dynamic homeolog regulation of allopolyploid heterosis in nascent hexaploid wheat. *Plant Cell* 26, 1878–1900.
- Li, H., and Durbin, R. (2010). Fast and accurate long-read alignment with burrows-wheeler transform. *Bioinformatics* 26, 589–595. doi: 10.1093/bioinformatics/btp698
- Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009). The sequence alignment/Map format and SAMtools. *Bioinformatics* 25, 2078–2079. doi: 10.1093/bioinformatics/btp352
- Li, X. C., and Fay, J. C. (2017). Cis-Regulatory divergence in gene expression between two thermally divergent yeast species. *Genome Biol. Evol.* 9, 1120–1129. doi: 10.1093/gbe/evx072
- Li, Z., Schulz, M. H., Look, T., Begemann, M., Zenke, M., and Costa, I. G. (2019). Identification of transcription factor binding sites using ATAC-seq. *Genome Biol.* 20:45. doi: 10.1186/s13059-019-1642-2
- Liti, G., Carter, D. M., Moses, A. M., Warringer, J., Parts, L., James, S. A., et al. (2009). Population genomics of domestic and wild yeasts. *Nature* 458, 337–341. doi: 10.1038/nature07743
- Liu, Y., Zhou, J., and White, K. P. (2014). RNA-seq differential expression studies: more sequence or more replication? *Bioinformatics* 30, 301–304. doi: 10.1093/bioinformatics/btt688
- Lopez-Maestre, H., Carnelossi, E. A. G., Lacroix, V., Burlet, N., Mugat, B., Chambeyron, S., et al. (2017). Identification of misexpressed genetic elements in hybrids between *Drosophila*-related species. *Sci. Rep.* 7:40618. doi: 10.1038/srep40618
- Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* 15:550.
- Marcet-Houben, M., and Gabaldón, T. (2015). Beyond the whole-genome duplication: phylogenetic evidence for an ancient interspecies hybridization in the Baker's yeast lineage. *PLoS Biol.* 13:e1002220. doi: 10.1371/journal.pbio.1002220
- McClintock, B. (1984). The significance of responses of the genome to challenge. *Science* 226, 792–801. doi: 10.1126/science.15739260
- McManus, C. J., Coolon, J. D., Duff, M. O., Eipper-Mains, J., Graveley, B. R., and Wittkopp, P. J. (2010). Regulatory divergence in *Drosophila* revealed by mRNA-seq. *Genome Res.* 20, 816–825. doi: 10.1101/gr.102491.109
- Metzger, B. P. H., Wittkopp, P. J., and Coolon, J. D. (2017). Evolutionary dynamics of regulatory changes underlying gene expression divergence among *saccharomyces* species. *Genome Biol. Evol.* 9, 843–854. doi: 10.1093/gbe/evx035
- Mixão, V., and Gabaldón, T. (2018). Hybridization and emergence of virulence in opportunistic human yeast pathogens. *Yeast* 35, 5–20. doi: 10.1002/yea.3242
- Mixão, V., Hansen, A. P., Saus, E., Boekhout, T., Lass-Flörl, C., and Gabaldón, T. (2019). Whole-Genome sequencing of the opportunistic yeast pathogen *Candida inconspicua* uncovers its hybrid origin. *Front. Genet.* 10:383. doi: 10.3389/fgene.2019.00383
- Möller, M., Habig, M., Freitag, M., and Stukenbrock, E. H. (2018). Extraordinary genome instability and widespread chromosome rearrangements during vegetative growth. *Genetics* 210, 517–529. doi: 10.1534/genetics.118.301050

- Möller, M., and Stukenbrock, E. H. (2017). Evolution and genome architecture in fungal plant pathogens. *Nat. Rev. Microbiol.* 15, 756–771. doi: 10.1038/nrmicro.2017.76
- Morales, L., and Dujon, B. (2012). Evolutionary role of interspecies hybridization and genetic exchanges in yeasts. *Microbiol. Mol. Biol. Rev.* 76, 721–739. doi: 10.1128/MMBR.00022-12
- Nomaguchi, T., Maeda, Y., Yoshino, T., Asahi, T., Tirichine, L., Bowler, C., et al. (2018). Homoeolog expression bias in allopolyploid oleaginous marine diatom *Fistulifera solaris*. *BMC Genomics* 19:330. doi: 10.1186/s12864-018-4691-0
- Pérez-Torrado, R., González, S. S., Combina, M., Barrio, E., and Querol, A. (2015). Molecular and enological characterization of a natural *Saccharomyces uvarum* and *Saccharomyces cerevisiae* hybrid. *Int. J. Food Microbiol.* 204, 101–110. doi: 10.1016/j.ijfoodmicro.2015.03.012
- Plissonneau, C., Stürchler, A., and Croll, D. (2016). The evolution of orphan regions in genomes of a fungal pathogen of wheat. *mBio* 7:e1231-16. doi: 10.1128/mBio.01231-1216
- Pryszcz, L. P., Németh, T., Gácsér, A., and Gabaldón, T. (2014). Genome comparison of *Candida orthopsilosis* clinical strains reveals the existence of hybrids between two distinct subspecies. *Genome Biol. Evol.* 6, 1069–1078. doi: 10.1093/gbe/evu082
- Pryszcz, L. P., Németh, T., Saus, E., Ksiezopolska, E., Hegedúsová, E., Nosek, J., et al. (2015). The genomic aftermath of hybridization in the opportunistic pathogen *Candida metapsilosis*. *PLoS Genet.* 11:e1005626. doi: 10.1371/journal.pgen.1005626
- Quinlan, A. R. (2014). BEDTools: the swiss-army tool for genome feature analysis. *Curr. Protoc. Bioinformatics* 47, 11.12.1–11.12.34. doi: 10.1002/0471250953.bi1112s47
- Rando, O. J., and Chang, H. Y. (2009). Genome-wide views of chromatin structure. *Annu. Rev. Biochem.* 78, 245–271. doi: 10.1146/annurev.biochem.78.071107.134639
- Rieseberg, L. H. (1997). Hybrid origins of plant species. *Ann. Rev. Ecol. Syst.* 28, 359–389. doi: 10.1146/annurev.ecolsys.28.1.359
- Robinson, J. T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E. S., Getz, G., et al. (2011). Integrative genomics viewer. *Nat. Biotechnol.* 29, 24–26.
- Ross-Innes, C. S., Stark, R., Teschendorff, A. E., Holmes, K. A., Ali, H. R., Dunning, M. J., et al. (2012). Differential oestrogen receptor binding is associated with clinical outcome in breast cancer. *Nature* 481, 389–393. doi: 10.1038/nature10730
- Salvadó, Z., Arroyo-López, F. N., Guilmón, J. M., Salazar, G., Querol, A., and Barrio, E. (2011). Temperature adaptation markedly determines evolution within the genus *Saccharomyces*. *Appl. Environ. Microbiol.* 77, 2292–2302. doi: 10.1128/AEM.01861-10
- Scannell, D. R., Zill, O. A., Rokas, A., Payen, C., Dunham, M. J., Eisen, M. B., et al. (2011). The awesome power of yeast evolutionary genetics: new genome sequences and strain resources for the *Saccharomyces sensu stricto* Genus. G3 1, 11–25. doi: 10.1534/g3.111.000273
- Schaeffe, B., Emerson, J. J., Wang, T.-Y., Lu, M.-Y. J., Hsieh, L.-C., and Li, W.-H. (2013). Inheritance of gene expression level and selective constraints on trans- and cis-regulatory changes in yeast. *Mol. Biol. Evol.* 30, 2121–2133. doi: 10.1093/molbev/mst114
- Schep, A. N., Buenrostro, J. D., Denny, S. K., Schwartz, K., Sherlock, G., and Greenleaf, W. J. (2015). Structured nucleosome fingerprints enable high-resolution mapping of chromatin architecture within regulatory regions. *Genome Res.* 25, 1757–1770. doi: 10.1101/gr.192294.115
- Schraiber, J. G., Mostovoy, Y., Hsu, T. Y., and Brem, R. B. (2013). Inferring evolutionary histories of pathway regulation from transcriptional profiling data. *PLoS Comput. Biol.* 9:e1003255. doi: 10.1371/journal.pcbi.1003255
- Schröder, M. S., Martínez de San Vicente, K., Prandini, T. H. R., Hammel, S., Higgins, D. G., Bagagli, E., et al. (2016). Multiple origins of the pathogenic yeast *Candida orthopsilosis* by separate hybridizations between two parental species. *PLoS Genet.* 12:e1006404. doi: 10.1371/journal.pgen.1006404
- Schwenk, K., Brede, N., and Streit, B. (2008). Introduction. Extent, processes and evolutionary impact of interspecific hybridization in animals. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 363, 2805–2811. doi: 10.1098/rstb.2008.0055
- Selmecki, A., Bergmann, S., and Berman, J. (2005). Comparative genome hybridization reveals widespread aneuploidy in *Candida albicans* laboratory strains. *Mol. Microbiol.* 55, 1553–1565.
- Selmecki, A. M., Dulmage, K., Cowen, L. E., Anderson, J. B., and Berman, J. (2009). Acquisition of aneuploidy provides increased fitness during the evolution of antifungal drug resistance. *PLoS Genet.* 5:e1000705. doi: 10.1371/journal.pgen.1000705
- Session, A. M., Uno, Y., Kwon, T., Chapman, J. A., Toyoda, A., Takahashi, S., et al. (2016). Genome evolution in the allotetraploid frog *Xenopus laevis*. *Nature* 538, 336–343. doi: 10.1038/nature19840
- Smukowski Heil, C. S., DeSevo, C. G., Pai, D. A., Tucker, C. M., Hoang, M. L., and Dunham, M. J. (2017). Loss of heterozygosity drives adaptation in hybrid yeast. *Mol. Biol. Evol.* 34, 1596–1612. doi: 10.1093/molbev/msx098
- Steenwyk, J. L., and Rokas, A. (2018). Copy number variation in fungi and its implications for wine yeast genetic diversity and adaptation. *Front. Microbiol.* 9:288. doi: 10.3389/fmicb.2018.00288
- Teixeira, M. C., Monteiro, P. T., Palma, M., Costa, C., Godinho, C. P., Pais, P., et al. (2018). YEASTRACT: an upgraded database for the analysis of transcription regulatory networks in *Saccharomyces cerevisiae*. *Nucleic Acids Res.* 46, D348–D353. doi: 10.1093/nar/gkx842
- Thompson, D. A., and Regev, A. (2009). Fungal regulatory evolution: cis and trans in the balance. *FEBS Lett.* 583, 3959–3965. doi: 10.1016/j.febslet.2009.11.032
- Tirosh, I., Reikavav, S., Levy, A. A., and Barkai, N. (2009). A yeast hybrid provides insight into the evolution of gene expression regulation. *Science* 324, 659–662. doi: 10.1126/science.1169766
- Tirosh, I., Sigal, N., and Barkai, N. (2010). Divergence of nucleosome positioning between two closely related yeast species: genetic basis and functional consequences. *Mol. Syst. Biol.* 6:365. doi: 10.1038/msb.2010.20
- Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D. R., et al. (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protoc.* 7, 562–578. doi: 10.1038/nprot.2012.016
- Waters, A. J., Makarevitch, I., Noshay, J., Burghardt, L. T., Hirsch, C. N., Hirsch, C. D., et al. (2017). Natural variation for gene expression responses to abiotic stress in maize. *Plant J.* 89, 706–717.
- Wickham, H. (2016). *Programming With ggplot2. In: ggplot2. Use R!*. Cham: Springer.
- Wimalasena, T. T., Greetham, D., Marvin, M. E., Liti, G., Chandelia, Y., Hart, A., et al. (2014). Phenotypic characterisation of *Saccharomyces* spp. yeast for tolerance to stresses encountered during fermentation of lignocellulosic residues to produce bioethanol. *Microb. Cell Fact.* 13:47. doi: 10.1186/1475-2859-13-47
- Wittkopp, P. J., Haerum, B. K., and Clark, A. G. (2004). Evolutionary changes in cis and trans gene regulation. *Nature* 430, 85–88.
- Wu, J., Lin, L., Xu, M., Chen, P., Liu, D., Sun, Q., et al. (2018). Homoeolog expression bias and expression level dominance in resynthesized allopolyploid *Brassica napus*. *BMC Genomics* 19:586. doi: 10.1186/s12864-018-4966-5
- Yoo, M.-J., Szadkowski, E., and Wendel, J. F. (2013). Homoeolog expression bias and expression level dominance in allopolyploid cotton. *Heredity* 110, 171–180. doi: 10.1038/hdy.2012.94
- Zerbino, D. R., Achuthan, P., Akanni, W., Amodé, M. R., Barrell, D., Bhai, J., et al. (2018). Ensembl 2018. *Nucleic Acids Res.* 46, D754–D761. doi: 10.1093/nar/gkx1098
- Zhang, M., Liu, X.-K., Fan, W., Yan, D.-F., Zhong, N.-S., Gao, J.-Y., et al. (2018). Transcriptome analysis reveals hybridization-induced genome shock in an interspecific F1 hybrid from *Camellia*. *Genome* 61, 477–485. doi: 10.1139/gen-2017-2105
- Zhang, Y., Liu, T., Meyer, C. A., Eickhout, J., Johnson, D. S., Bernstein, B. E., et al. (2008). Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* 9:R137. doi: 10.1186/gb-2008-9-9-r137

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2020 Hovhannisyan, Saus, Ksiezopolska, Hinks Roberts, Louis and Gabaldón. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



OPEN ACCESS

Edited by:

Ekaterina Shelest,
German Centre for Integrative
Biodiversity Research (iDiv), Germany

Reviewed by:

Suriyan Cha-um,
National Science and Technology
Development Agency (NSTDA),
Thailand
Padhmanand Sudhakar,
Earlham Institute (EI), United Kingdom

***Correspondence:**

Benildo G. de los Reyes
benildo.reyes@ttu.edu

[†] These authors have contributed
equally to this work

***Present address:**

Glenn B. Gregorio,
Institute of Crop Science, University
of the Philippines Los Baños,
Los Baños, Philippines
Rakesh Kumar Singh,
International Center for Biosaline
Agriculture, Dubai,
United Arab Emirates

Specialty section:

This article was submitted to
Systems Biology,
a section of the journal
Frontiers in Genetics

Received: 13 August 2020

Accepted: 05 October 2020

Published: 23 October 2020

Citation:

Pabuayon ICM, Kitazumi A,
Gregorio GB, Singh RK and
de los Reyes BG (2020) Contributions
of Adaptive Plant Architecture
to Transgressive Salinity Tolerance
in Recombinant Inbred Lines of Rice:
Molecular Mechanisms Based on
Transcriptional Networks.
Front. Genet. 11:594569.
doi: 10.3389/fgene.2020.594569

Contributions of Adaptive Plant Architecture to Transgressive Salinity Tolerance in Recombinant Inbred Lines of Rice: Molecular Mechanisms Based on Transcriptional Networks

Isaiah Catalino M. Pabuayon^{1†}, Ai Kitazumi^{1†}, Glenn B. Gregorio^{2†},
Rakesh Kumar Singh^{2*} and Benildo G. de los Reyes^{1*}

¹ Department of Plant and Soil Science, Texas Tech University, Lubbock, TX, United States, ² International Rice Research Institute, Los Baños, Philippines

Genetic novelties are important nucleators of adaptive speciation. Transgressive segregation is a major mechanism that creates genetic novelties with morphological and developmental attributes that confer adaptive advantages in certain environments. This study examined the morpho-developmental and physiological profiles of recombinant inbred lines (RILs) from the salt-sensitive IR29 and salt-tolerant Pokkali rice, representing the total range of salt tolerance including the outliers at both ends of the spectrum. Morpho-developmental and physiological profiles were integrated with a hypothesis-driven interrogation of mRNA and miRNA transcriptomes to uncover the critical genetic networks that have been rewired for novel adaptive architecture. The transgressive super-tolerant FL510 had a characteristic small tiller angle and wider, more erect, sturdier, and darker green leaves. This unique morphology resulted in lower transpiration rate, which also conferred a special ability to retain water more efficiently for osmotic avoidance. The unique ability for water retention conferred by such adaptive morphology appeared to enhance the efficacy of defenses mediated by Na⁺ exclusion mechanism (*SalTol*-effects) inherited from Pokkali. The super-tolerant FL510 and super-sensitive FL499 had the smallest proportions of differentially expressed genes with little overlaps. Genes that were steadily upregulated in FL510 comprised a putative cytokinin-regulated genetic network that appeared to maintain robust growth under salt stress through well-orchestrated cell wall biogenesis and cell expansion, likely through major regulatory (*OsRR23*, *OsHK5*) and biosynthetic (*OsIPT9*) genes in the cytokinin signaling pathway. Meanwhile, a constitutively expressed cluster in FL510 prominently featured two transcription factors (*OsIBH1*, *TAC3*) that control tiller angle and growth habit through the brassinosteroid signaling pathway. Both the putative cytokinin-mediated and brassinosteroid-mediated clusters appeared to function as highly coordinated network synergies in FL510. In contrast, both networks appeared to be sub-optimal and inferior in the other RILs and parents as they were disjointed and highly fragmented. Transgressively expressed miRNAs (*miR169*, *miR397*, *miR827*) were also identified as

prominent signatures of FL510, with functional implications to mechanisms that support robust growth, homeostasis, and osmotic stress avoidance. Results of this study demonstrate how genetic recombination creates novel morphology that complements inducible defenses hence transgressive adaptive phenotypes.

Keywords: genetic novelty, transgressive segregation, salinity, plant architecture, growth habit, network rewiring

INTRODUCTION

Adaptation to the dynamic nature of abiotic and biotic factors in the environment is crucial for plant survival and reproduction (Anderson et al., 2011). Success in an ecological niche is optimized by environmental pressures that interface with genetic and epigenetic potentials across a founder population by virtue of genotype-by-environment interaction (GxE). The GxE dynamics translate into adaptive growth, developmental and reproductive habits through intricately programmed networks of gene expression (Van Wallendael et al., 2019). Recent studies supported the role of natural hybridization between genetically distant parents as important mechanism for the creation of rare developmental, morphological, and/or adaptive traits that are outliers from the majority of the population, i.e., genetic novelties. These novelties have been proposed to set foundations for new phylogenetic lineages that adapt to new ecological niches (Wagner and Lynch, 2010).

Engineering the new generation of agricultural crops with adaptive capacity to emerging ecological dynamics amidst the marginalization created by climate change, natural resource deterioration, and environmental degradation has become the core of plant breeding in the 21st century. Similar to ecological and evolutionary dynamics, key to this goal are novel phenotypes that are above and beyond what has already been achieved by earlier breeding paradigms (de Los Reyes, 2019). For instance, the Green Revolution in the 1960's was hallmarked by the creation of novel plant architecture that was optimal for managed environments with ideal water and nutrient inputs. In rice, pivotal to the success of the Green Revolution was the novel semi-dwarf architecture conferred by the *sd1* mutation, with lower than normal levels of bioactive gibberellic acid (GA) that prevented extensive stem elongation (De Datta et al., 1968; Chandler, 1992; Monna et al., 2002; Sasaki et al., 2002; Hedden, 2003; Peng et al., 2010). Semi-dwarf cultivars exhibited a special adaptive trait that avoided lodging, which is a negative offshoot of enhanced yield created by efficient mobilization of resources from vegetative source to reproductive sinks. Increased seed yield would otherwise have negative pre-harvest impacts in the native landraces with tall and lanky stature, because of inability to support the heavy weight of the panicles created by efficient source-sink partitioning. Therefore, ideotype breeding created an adaptive plant architecture for water-rich and nutrient-rich environments (Vergara, 1988; Khush, 1995a).

While the Green Revolution enabled dramatic improvements in yield, it also led to dramatic increase in water and nutrient requirements to support the full genetic potentials of modern cultivars. The semi-dwarf, high-yielding cultivars had a narrow

ecological niche beyond well-managed agricultural ecosystems. For instance, *sd1* is linked to an inferior allele of the drought tolerance QTL *qDTY1.1* that is important for maintaining yield in upland landraces, which normally thrive with limited water (Vikram et al., 2016). While the successful modification of rice ideotype through *sd1* created a large net gain in grain yield under well-managed environments, the technology also led to unanticipated trade-offs because of the loss of other attributes that would allow a wider ecological adaptive capacity. There is a renewed vision to recover the lost attributes to create the 21st century breeds of crops with minimal yield penalty under marginalized environments (Khush, 2001; Khush, 2005).

In response to the paradigm switch triggered by the burgeoning ecological threats to agriculture, in the 1990's, the International Rice Research Institute (IRRI) proposed to remodel the developmental and morphological architecture of the rice plant, to create the '*new plant-type*' (NPT) for enhanced performance to more limiting ecological settings. The aim was to create a new ideotype with fewer but fully productive tillers, sturdy stems, dark green, thick, and erect leaves, and vigorous roots, toward even greater increases in harvest index (Peng et al., 1994; Khush, 1995b). The Green Revolution varieties typically produce 30–40 tillers, of which only half produce panicles even under water-rich and nutrient-rich environments. Such ideotype creates issues with nutrient allocation, with vegetative tillers creating a '*sink*' for nutrients instead of '*source*' for seed development. Thus, the NPT combined the features that prevented pre-harvest losses through lodging, fine-tuned biomass production for channeling photosynthate toward grain development, and suppressed unproductive vegetative growth, giving as much as 1.5 tons ha⁻¹ improvement in yield (Khush, 2013). The NPT represents another example of the importance of adaptive domestication by breeding to address the burgeoning environmental threats to modern agriculture.

The phenomenon of transgressive segregation, which is observed in both natural and artificial populations created by plant breeding, is characterized by the occurrence of minority phenotypic outliers relative to parental range across a segregating or recombinant population derived from genetically divergent parents. In addition to the classic explanations attributing complementation and epistatic interactions as major mechanisms behind transgressive traits, the possible roles of coupling and uncoupling effects and genetic network rewiring have also been recently proposed (Vega and Frey, 1980; Rieseberg et al., 1999; Dittrich-Reed and Fitzpatrick, 2013; de Los Reyes, 2019). Combined with the paradigms of genomic biology, the potential of transgressive individuals

for enhanced yield of crops have been established, but its true potential for adaptive traits is yet to be determined (DeVicente and Tanksley, 1993).

A modern view on phenotypic variance for complex traits was recently presented by the *Omnigenic Theory* through the synergistic interaction between the relatively fewer major-effect ‘core genes’ and the more numerous but minute-effect ‘peripheral genes’ scattered throughout the genome (Boyle et al., 2017). A further extension of the *Omnigenic Theory* may also accommodate the potential roles of not just the protein-coding but also the non-protein-coding regions (ncRNAs) of the genome as both core and peripheral components (Carrington and Ambros, 2003; Mallory and Vaucheret, 2006; Khraiwesh et al., 2012). This study represents a holistic re-examination of the phenomenon of transgressive segregation for an adaptive trait (salt tolerance) in rice by integrating the potential contributions of both transcriptional and post-transcriptional regulatory mechanisms. Guided by the visions of the *Omnigenic Theory*, the aim of this study was to illustrate the molecular synergies in context of the genetic network paradigm behind the outlier stress-adaptive phenotypes across a recombinant inbred (RIL) population derived from the improved high-yielding cultivar IR29 (salt-sensitive) and an Indian landrace Pokkali (salt-tolerant; Pabuayon et al., 2020).

Results showed that salt stress defense capacity encoded by the *SalTol* QTL was not the main driver behind transgressive super-tolerance or super-sensitivity (Pabuayon et al., 2020). Rather, transgressive salt-tolerance can be explained to larger extents by rewired genetic networks that involved regulatory transcription factors and miRNAs that determine plant architecture, growth habit, and stress avoidance. This study provides yet another example that the optimization of adaptive potential requires the modification not only of the capacity for defense and repair but more importantly, the reconfiguration of developmental attributes to allow growth adjustment and stress avoidance, similar to the new ideotypes created by rice breeding during the Green Revolution and post-Green Revolution era.

MATERIALS AND METHODS

Plant Growth Conditions

The comparative panels of genotypes used in this study were selected from earlier studies that identified with transgressive segregants at $EC = 12 \text{ dS m}^{-1}$ during V4–V12 stage of vegetative growth (Pabuayon et al., 2020). Four recombinant inbred lines (RILs) derived from IR29 (*Xian/indica*; salt-sensitive) \times Pokkali (*Aus*; salt-tolerant) representing different levels of salinity tolerance were used for morpho-developmental profiling and RNA-Seq experiments. These included FL510 (super-tolerant), FL499 (super-sensitive), FL478 (tolerant), and FL454 (sensitive). Plants were grown in Lubbock, Texas under standard greenhouse conditions at 30–35°C day, 24–26°C night; 20 to 30% RH; 12-hour photoperiod with an average light intensity of $500 \mu\text{mol m}^{-2}\text{s}^{-1}$ (photosynthetic photon flux density).

Seeds were germinated at 30°C in petri dishes lined with wet filter paper for 2 days and grown in seedling trays with standard peat moss potting mix for 14 days. Individual plants were transplanted to 2.27-liter (0.6-gallon) hydroponic buckets with 1 g L^{-1} Peter’s Professional 20-20-20 General Fertilizer at pH 5.8 supplemented with 0.4 g L^{-1} $\text{FeSO}_4 \cdot 7\text{H}_2\text{O}$. The plants were grown until maximum tillering stage (V10–V12; Counce et al., 2000), and salinity stress was introduced by increasing the electrical conductivity (EC) of the hydroponic solution to $EC = 12 \text{ dS m}^{-1}$ ($\sim 120 \text{ mM}$) with NaCl. Tissue samples (shoots) for RNA extraction were collected prior to stress (control; 0 h) and 24, 48, 72, and 144 h after stress.

Separate sets of plants were used for morpho-developmental and physiological profiling. For these measurements, salinity stress was not imposed. Plants were grown to their maximum potential to enable precise description of morphology and growth habit without stress limitations, thus allowing the examination of the differences in inherent morpho-developmental potentials in relation to their salinity tolerance capacities. Individual seedlings were transplanted into buckets with Turface MVP® and submerged in the hydroponics solution. Plants were grown to the same maximum tillering stage as in the transcriptome experiment described above (V10–V12).

Morpho-Developmental and Physiological Profiling and Statistical Analysis

The overall vegetative morphology and growth habit of each genotype in the comparative panel were characterized. Width of leaf blades was measured at the widest point (middle area) using a digital caliper. Three leaves per plant were measured at random, and six plants were used for this analysis ($n = 18$). Tiller angle and plant height were measured via image analysis using Fiji software (Schindelin et al., 2012). The furthest tiller was used as a reference point and its angle from the base of the plant from the vertical axis perpendicular to the ground was measured. Plant height was measured as the length from the visible base to its highest point ($n = 6$). Shoot and root biomass was measured from three individual plants. Samples were oven dried at 70°C for at least 5 days and weighed. Tiller number was counted as the number of individual tillers stemming from the base of the plant shoot ($n = 6$). Stomatal conductance was measured in terms of flux ($\text{mmol/m}^2\text{s}$, $n = 5$) using the SC-1 Leaf Porometer (Meter Group, Inc., Pullman, WA, United States).

Statistical analysis was conducted on R 4.0.1 (R Core Team, 2020). Analysis of Variance (ANOVA) and Tukey’s HSD *post hoc* tests were conducted using the ‘agricolae’ package with significance tests set at $P < 0.05$ (De Mendiburu and De Mendiburu, 2019). Data was visualized using the ‘ggplot2’ package (Wickham and Chang, 2008). The clustering dendrogram and heatmap was made with the package ‘gplots’ with default parameters (Warnes et al., 2016).

RNA-Seq and Transcriptional Network Modeling

The temporal transcriptome experiments were designed to capture both the immediate (24, 48, 72 h) and long-term (144 h) responses to salt stress relative to control (0 h) to enable dissection of different patterns of co-expression through a temporal and inter-genotypic transcriptome matrix. Parallel libraries were constructed to capture both the mRNA and sRNA/miRNA reads by RNA-Seq. Total RNA was extracted from frozen leaf tissues using miRVanaTM miRNA Isolation Kit (Invitrogen, Carlsbad, CA, United States) and used to construct time-course (0, 24, 48, 72, 144 h) RNA-Seq libraries. The sRNA/miRNA libraries were also made from the same RNA samples using Bioo Scientific NextFlexTM Small RNA-seq Kit V3 (Bioo Scientific Corporation, Austin, TX, United States). Strand-specific 150-bp paired-end (mRNA) and single-read (sRNA/miRNA) RNA-Seq libraries were sequenced on Illumina HiSeq3000 with two replicates (Oklahoma Medical Research Foundation, OK, United States). Sequence output from mRNA-Seq and sRNA-Seq libraries (PRJNA378253) were preprocessed with Cutadapt (v2.10) and mapped against the Nipponbare genome (IRGSP-1.0) and miRBase (*Oryza sativa* mature miRNA, v22.1) using HISAT2 (v2.1.0) and blat (v36), respectively (Kent, 2002; Fahlgren et al., 2010; Martin, 2011; Sakai et al., 2013; Kim et al., 2015; Stark et al., 2019). Transcript read counts were normalized by trimmed mean M-values (TMM) by Subread (v1.5.2) and differential expression was examined by edgeR (v3.24.3) using a false detection rate of 0.05 (McCarthy et al., 2012; Liao et al., 2013). Expression values were analyzed and shown as the log₂ fold-change (log₂-FC) relative to control for each genotype, allowing the comparative analysis of expression patterns among the different genotypes.

Assignment of genes to relevant metabolic pathways and gene ontologies were done through the Kyoto Encyclopedia of Genes and Genomes (Kanehisa and Goto, 2000; Moriya et al., 2007). Enrichment of GO terms was calculated with R package 'goseq' and a cutoff of $P < 0.05$ was set for identifying significant ontologies (Young et al., 2010). Visualization of data and computation of relevant Pearson's correlation coefficient (PCC) matrices were performed with R 4.0.1 (R Core Team, 2020). PCC was used to determine the strength and direction of the relationship between genes. This information was then used to create gene networks. As there was no specific central gene selected, correlation coefficients were used as a threshold for filtering connections (edges) between genes (nodes). This allowed the discernment between co-expression and reverse co-expression of genes and to demonstrate the cohesion or fragmentation of a gene network in a genotype. Networks were constructed and visualized using Cytoscape (Smoot et al., 2010).

Upset graphs were created using the package 'UpsetR' (Conway et al., 2017) and were used to show the distribution of genes in different classes. Heatmaps were created using the package 'gplots' (Warnes et al., 2016). Alluvial graphs were created using the 'alluvial' package (Bojanowski and Edwards, 2018) to show the similarity of gene expression patterns between genotypes.

RESULTS

Morphological and Developmental Novelties of Transgressive Segregants From IR29 × Pokkali

Transgressive segregation for salt tolerance was uncovered in a recombinant inbred population (F₈-RIL) of rice derived from the improved high-yielding *Xian/Indica* cultivar IR29 (salt-sensitive) and the Indian *Aus* landrace Pokkali (salt-tolerant), based on integrative molecular, biochemical, macro-physiological, and real-time growth profiling approach to phenotyping (Pabuayon et al., 2020). Pokkali is a known donor of seedling-stage (V2–V4) salt-tolerance by the Na⁺ exclusion mechanism encoded by the *SalTol* QTL (Thomson et al., 2010). Systems-level characterization of each individual across the population identified two representative F₈-RILs as clear outliers relative to parental range of growth potential under salt stress. FL510 represents a positive transgressive (super-tolerant) progeny and FL499 represents a negative transgressive (super-sensitive) progeny. Two other sibling RILs, FL478, and FL454, represented the salt-tolerance similar to the donor parent Pokkali and salt-sensitivity similar to the recipient parent IR29, respectively. Along with the parents, these four sibling RILs (salt-tolerant FL478, salt-sensitive FL454, super-tolerant FL510, and super-sensitive FL499) comprised the minimal comparative panel for in-depth characterization.

A key observation was the unique developmental and morphological attributes that highlight the transgressive nature of FL510. In terms of overall morphology and growth habit, the representative RILs in the panel were either similar to one of the parents (FL478 is similar to IR29) or hybrid for parental attributes (FL454, FL499). The clear exception was FL510, which had a modified architecture that resembled neither of the parents (**Figure 1A**). FL478 and IR29 had semi-dwarf stature (average height of 82.72 and 78.94 cm, respectively) with profuse tillering habit, characteristics of the Green Revolution plant architecture (**Figure 1B**). FL454 and FL499 had tall statures (average height of 121.17 and 133.20 cm, respectively) with slender and lanky leaves similar to Pokkali, and intermediate tillering habit. Distinct developmental and morphological features of FL510 were highlighted by intermediate height (average of 99.88 cm), low tiller angle, and thick, erect, dark green leaves, similar to the NPT.

The leaf morphology of FL454 based on width of leaf blades (average of 14.54 mm) appeared to have been inherited from Pokkali, while the FL478 leaf morphology resembled that of IR29. FL510 and FL499 were clear transgressives for leaf morphology having significantly wider blades (average width of 16.69 mm) than Pokkali, and narrower (average width of 10.46 mm) than IR29, respectively (**Figure 1C**). Width of leaf blades is known to contribute to higher transpiration rates in maple (Bauerle and Bowden, 2011). In rice, it has been associated with better yield, suggesting its positive contributions to photosynthetic capacity (Fujita et al., 2009). Wider leaf blade is also inversely correlated with leaf rolling under drought,

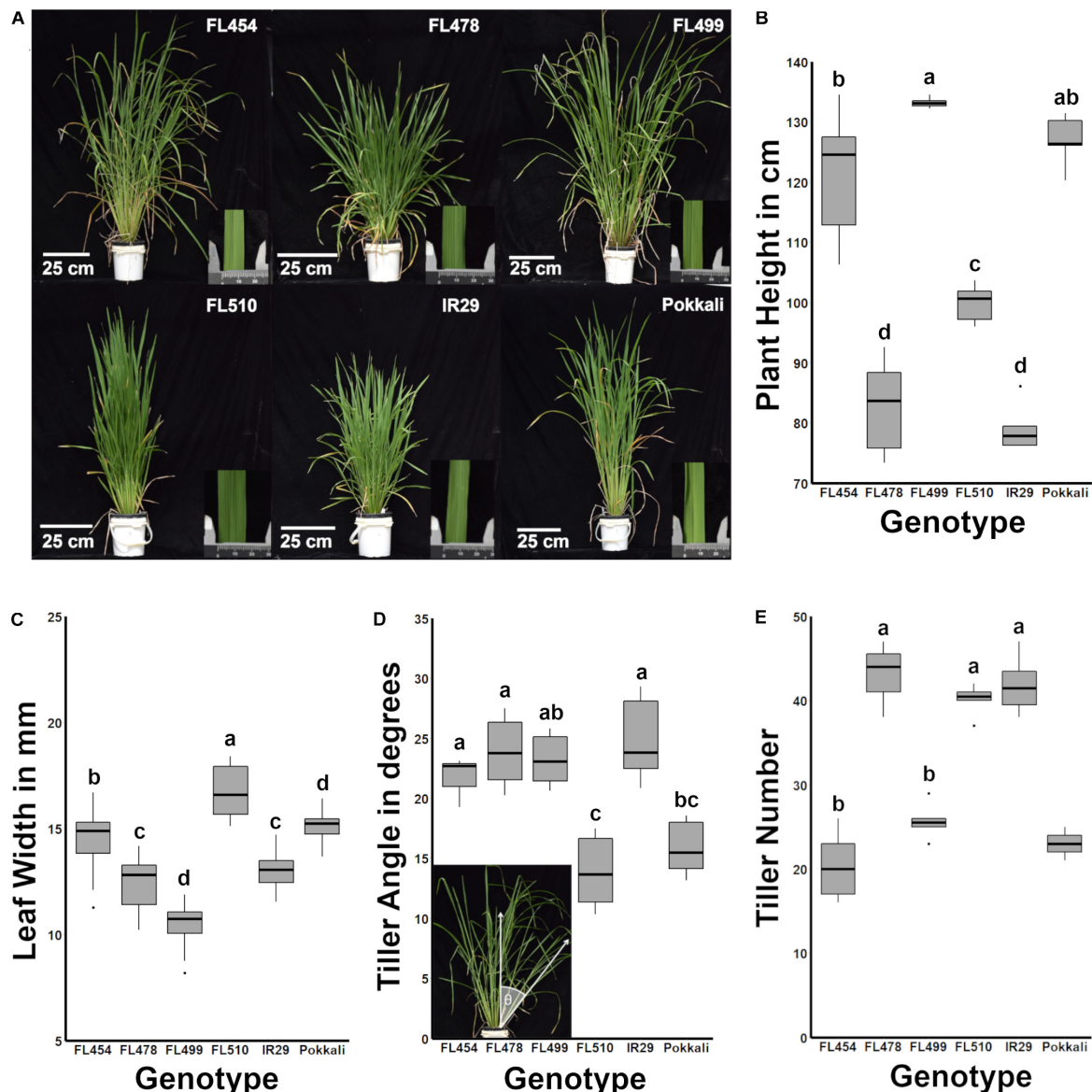


FIGURE 1 | Morphological and growth habit profiles across the IR29 \times Pokkali minimal comparative panel. The genotypic panel was grown under optimal conditions and compared morphologically and physiologically. **(A)** Gross-morphological differences highlighted the unique architecture of FL510. The inset photograph beside each plant shows a representative leaf under a caliper (units in mm). **(B)** Similarities across genotypes based on plant height ($P < 0.05$). FL454 and FL499, which are both sensitive to salt, are more similar to the tolerant donor parent Pokkali in terms of plant height, while FL478 is more similar to IR29. In contrast, FL510 is uniquely intermediate between the parents. **(C)** Box plots of leaf blade widths ($n = 18$) highlighting the uniqueness of FL510. In comparison, FL454 and FL478 are similar to the parents Pokkali and IR29, respectively. The super-sensitive FL499 has narrower leaves. **(D)** The tiller angle of each genotype ($n = 6$) was measured as the angle of the farthest tiller to its base as demonstrated in the inset photo and are shown as box plots. FL454, FL478, and FL499 are not significantly different from IR29 and have wider tiller angles compared to Pokkali and FL510 ($P < 0.05$). Tiller angle in FL499 was similar to Pokkali although it has a much higher mean value. **(E)** Number of tillers per plant showed that the genotypes with smaller statures (IR29, FL478, FL510) are more similar to each other with significantly higher values than the large-stature genotypes (Pokkali, FL454, FL499; $P < 0.05$). Taken together, morphological profiles indicate a compact architecture of FL510.

suggesting that wider leaves contribute to osmotic stress tolerance (Cal et al., 2019).

Aside from thick, dark green, erect leaves, the tiller angle measured as the angle from the main axis of the plant perpendicular to the ground profoundly differentiated the transgressive super-tolerant FL510 from its parents and

siblings. While FL454, FL478, and FL499 had similar angles as IR29, by virtue of its erect growth habit, FL510 had the smallest tiller angle even in comparison to Pokkali, which had the smaller angle of the two parents (**Figure 1D**). The uniquely compact stature of FL510 was also supported by tiller size and number, having the smallest mean tiller number

among the smaller-stature genotypes. The other RILs with larger stature had lower tiller number, yet more spreading growth habit. IR29 and FL478 had smaller tillers with wider growth angle, which created a more open architecture and larger surface area (**Figure 1E**). Based on morphology and vegetative growth habit (tillering), the transgressive super-tolerant FL510 is unique by virtue of its compact, erect stature and intermediate height, all of which contribute to smaller surface area.

Potential Effects of Plant Morphology on Water Retention

Leaf angle and how it shapes the overall surface area of the plant is an important aspect of transpiration. Upright (erect) leaf growth or tiller growth in the case of rice has been correlated with lower rates of transpiration than leaves or tillers growing at wider angles (Van Zanten et al., 2010). To further investigate the potential significance of the novel morphology and growth habit of FL510 to its unique ability to maintain robust growth under salt stress, the stomatal conductance was compared across the panel as a measure of transpiration rate. Results showed that FL510 indeed had the lowest stomatal conductance, significantly different from its sibling RILs (**Figure 2A**). The highest stomatal conductance was observed in the salt sensitive FL454, which also had among the widest tiller angles and narrowest leaf blades across the panel. The tolerant FL478 and super-sensitive FL499, both of which had tiller angle comparable to FL454, had the next highest stomatal conductance. These results are consistent with the inverse trends between morphology and stomatal conductance as observed in the other RILs.

In terms of biomass, FL499 had the highest shoot dry weight while IR29 had the smallest (**Figure 2B**). FL478 had slightly lower biomass than FL499 but was not statistically different from FL510 and Pokkali. FL454 was slightly smaller than the FL510 and Pokkali and was not statistically different from IR29. Differences in root biomass were not statistically significant (**Figure 2C**). The overall trends in shoot biomass are consistent with the uniquely compact architecture of FL510. Leaves growing at wider angles from the main axis of the plant have increased incidence of solar radiation, causing rapid elevation of internal leaf temperature and consequently, higher transpiration rates (Murchie et al., 1999; Van Zanten et al., 2010). The unique morphology and growth habit of the super-tolerant FL510, especially its narrow tiller angle appears to contribute to the reduction of the impact of osmotic stress under salinity as water is more efficiently retained by virtue of modulated transpiration. At the same time, the wider but erect, dark green leaves of FL510 appear to compensate for potential impediments to photosynthesis, despite reduction in gas intake. These features are likely allowing the plant to meet its energetic requirements while also modulating metabolic rate.

Traits that are most variable across genotypes (plant height, shoot dry weight, tiller angle, leaf width, and tiller number) were used to assess the overall developmental similarities across the panel. As shown in the hierarchical clustering dendrogram, the larger-stature (FL499, FL454, Pokkali) and smaller-stature (IR29,

FL478, FL510) genotypes formed two distinct clades (**Figure 2D**). IR29 and FL478 were much closer to each other than FL510, due to the unique properties of FL510 for leaf width, tiller angle, plant height, and shoot dry weight. On the other hand, the main difference between Pokkali and the two larger-stature RILs was tiller angle, but with high similarity to FL454 with respect to other parameters. These results also highlight the uniqueness of the combination of traits that created FL510. While FL478 mainly inherited its morphology from IR29, FL510 combined the smaller stature of IR29 with a compact tiller angle of Pokkali. However, it also had transgressively wider leaves, which is likely a result of network rewiring.

General Trends in the mRNA Transcriptomes of Recombinant Inbred Lines

Based on multiple layers of information from all available data (**Figures 1, 2** in this paper and from Pabuayon et al., 2020), the overall morphology and growth habit of FL510 appeared to provide physiological advantage and flexibility under salt stress. This is likely through an inherent capacity to reduce the impact of osmotic stress while also maintaining adequate level of photosynthesis, hence avoiding severe penalties to growth. In combination with other defense mechanisms conferred by *SalTol* (Na^+ exclusion), the transgressive super-tolerant FL510 appeared to have both the necessary repair/avoidance and physiological maintenance attributes to sustain better growth under severe salt stress, more so than its tolerant donor parent Pokkali and tolerant sibling FL478. To begin unraveling the genetic mechanisms behind such potentially unique synergy of growth, morphology, and physiological defenses in FL510, we conducted an in-depth profiling of the salt stress response mRNA transcriptomes across the comparative panel. The aim was to identify some of the critical regulators and targets (core and peripheral) involved in the expression of the inherent phenotypic potential of FL510.

The mapping statistics and coverage of the temporal mRNA-Seq datasets across the comparative panel before (0 h; control) and during (24, 48, 72, 144 h) salt stress ($\text{EC} = 12 \text{ dS m}^{-1}$) are summarized in **Supplementary Table 1** (PRJNA378253: SRR11528266 to SRR11528295). Based on the distribution of mRNA abundances, the super-tolerant FL510 and super-sensitive FL499 appeared to exhibit the least changes in the transcriptome as indicated by the proportions of differentially expressed protein-coding genes (**Figure 3A**). The number of upregulated genes was highest in the tolerant FL478 and sensitive parent IR29 (**Figure 3B**), while the number of downregulated genes was highest in the sensitive FL454 (**Figure 3C**).

Common responses to $\text{EC} = 12 \text{ dS m}^{-1}$ were also evident across the comparative panel based on the overlapping patterns in gene expression. For example, the tolerant FL478 and its sensitive parent IR29 shared a total of 1,160 upregulated genes, which is more than double the number of upregulated genes shared by FL478 with its tolerant donor parent Pokkali (**Figure 3B**). Pokkali and IR29 had the third highest numbers of shared upregulated genes across all pair-wise comparisons, indicating major overlaps between their response mechanisms at least at

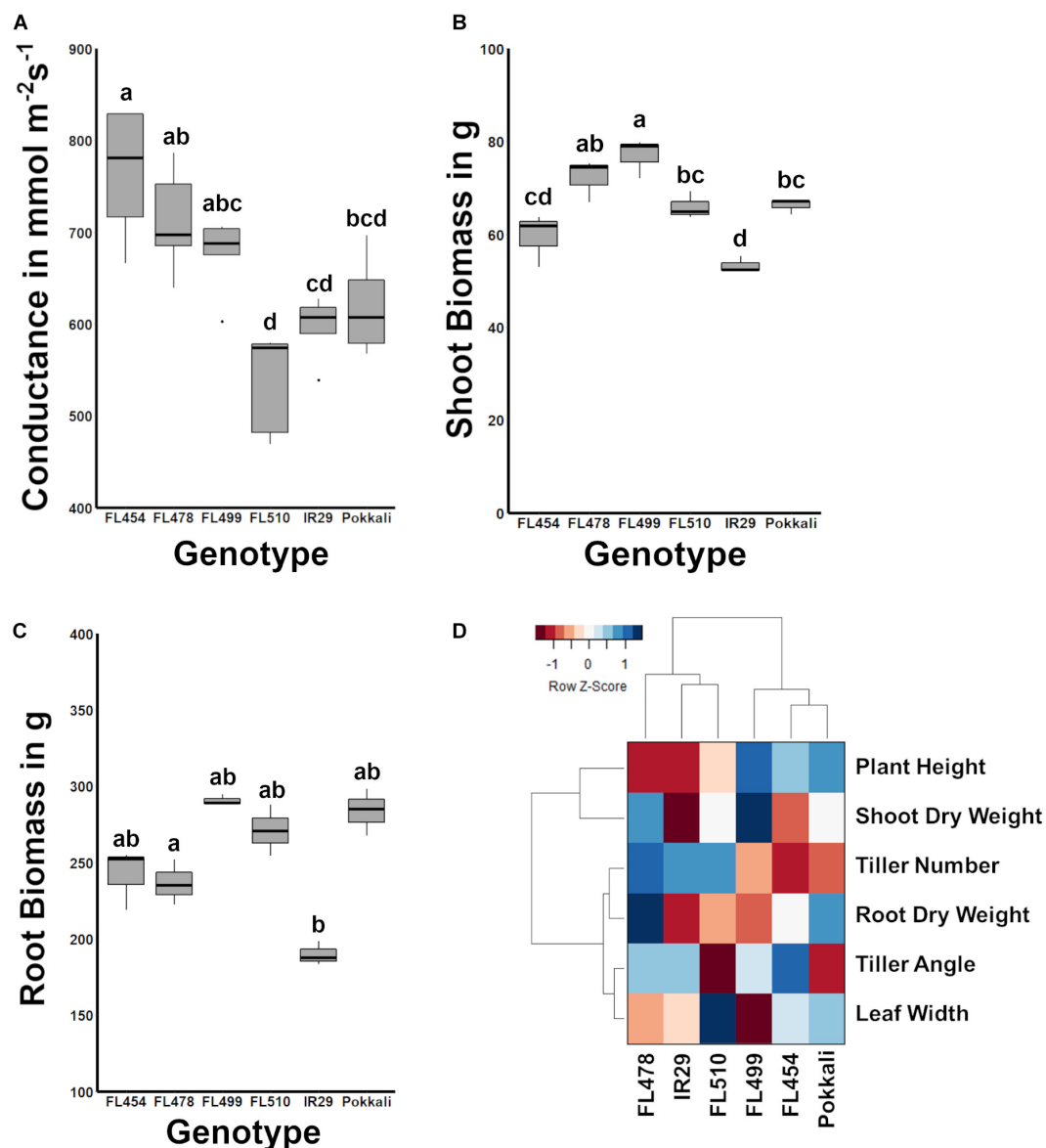


FIGURE 2 | Comparison of stomatal conductance and biomass potential across the IR29 × Pokkali comparative panel, and measure of overall relatedness across the panel based on morpho-developmental and physiological profiles. **(A)** The width of leaf blades and tiller angle consequently affect transpiration. FL510, which had the smallest tiller angle and broadest leaf blades also had the smallest mean stomatal conductance, significantly lower than the other recombinant inbred lines (RILs; $P < 0.05$). While the stomatal conductance of FL510 was not statistically different from IR29 and Pokkali, it was lowest mean among the genotypes. **(B)** Shoot biomass indicated that FL510 had smaller exposed surface area compared to the other genotypes, while its shoot dry weight was comparable to the other genotypes. **(C)** Root biomass was not significantly different among the different genotypes, indicating that the difference in architecture is mainly due to shoot growth. **(D)** Hierarchical clustering dendrogram of morpho-physiological traits highlighting the similarity of FL510 and FL478 to IR29, and the similarity of FL454 and FL499 to Pokkali. The heatmap summarizes the widest deviation of FL510, which lie on its wider leaves and lesser tiller angle, to the common architecture of FL478 and IR29.

the transcriptome level, despite their contrasting sensitivity to salt stress. In contrast, the transgressive super-tolerant FL510 was unique based on the small proportion of gene expression changes shared with any of the other genotypes. Consistent with its superior growth under $EC = 12 \text{ dS m}^{-1}$, it has the least overlap in gene expression with the sensitive FL454 and super-sensitive FL499. The two inferior RILs, i.e., sensitive FL454 and super-sensitive FL499, also shared the largest number

of upregulated genes. Meanwhile, the tolerant donor parent Pokkali and its sensitive RIL FL454 had the largest number of shared downregulated genes, followed by IR29 and FL478 (Figure 3C). Pokkali and its sensitive RIL FL454 also exhibited similar proportions of downregulated genes, suggesting that downregulation of gene expression could be a critical component of the response mechanisms in Pokkali and FL454, more so than the other genotypes.

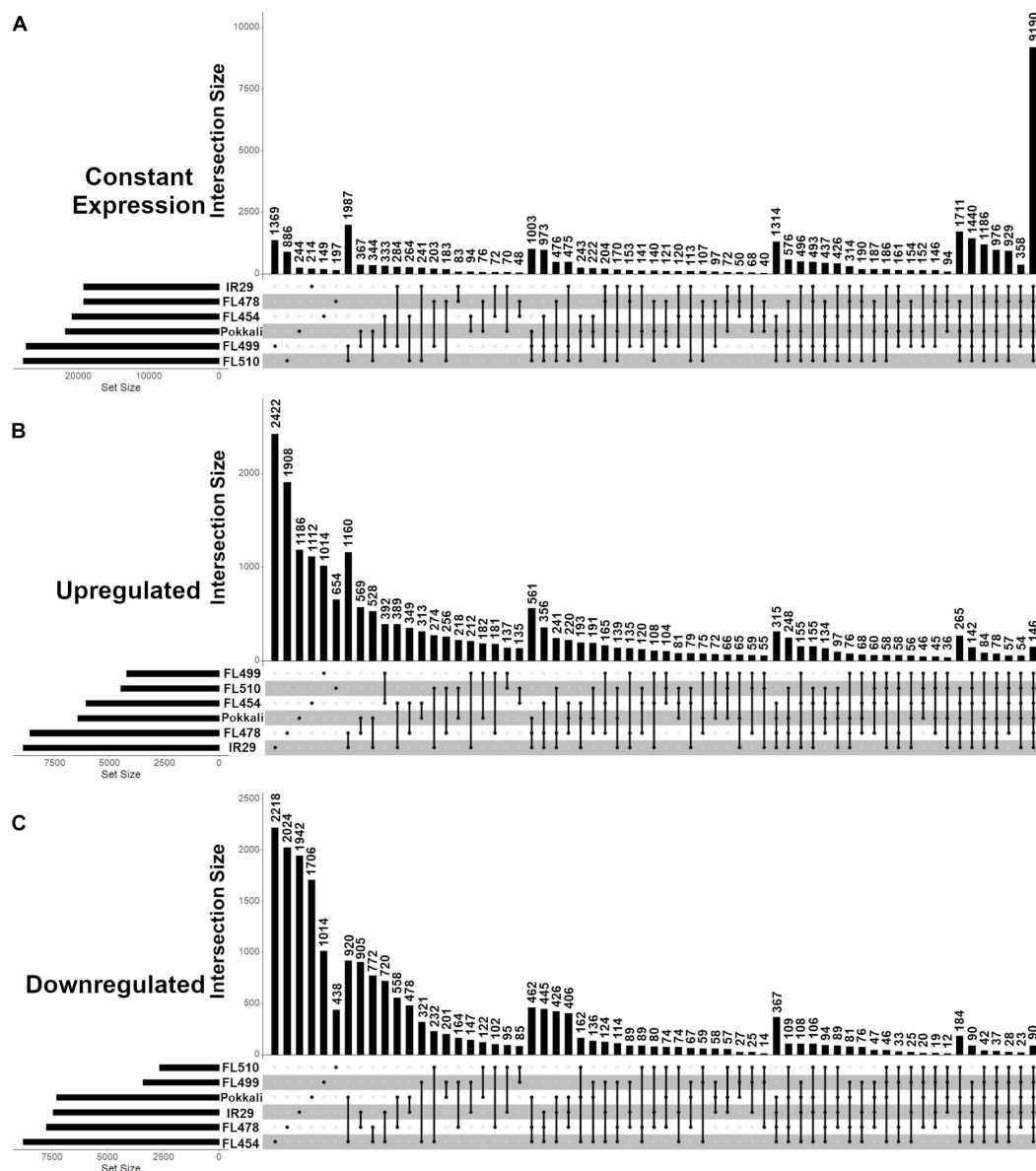


FIGURE 3 | Distribution of protein-coding genes with unchanged (constant), upregulated or downregulated expression across the IR29 × Pokkali comparative panel as revealed by temporal RNA-Seq analysis before (0 h) and during (24, 48, 72, 144 h) salt stress at EC = 12 ds m⁻¹. Analysis was performed from V5 to V12 stage of vegetative growth. **(A)** Genes with unchanged (constant) expression had no significant changes in transcript abundance (< 2 and > -2 log₂-fold) across all time-points. **(B)** Upregulated genes had significant increase in transcript abundance (> 2 log₂-fold) relative to 0 h in at least one time point at EC = 12 ds m⁻¹. **(C)** Downregulated genes had significant decrease in transcript abundance (< -2 log₂-fold) relative to 0 h in at least one time point at EC = 12 ds m⁻¹. Upset graphs accommodate all possible pair-wise comparisons and corresponding set intersections across pair-wise comparisons. The horizontal bar graph at the bottom left represents the total set size in each genotype, while the vertical bar graph represents the abundance of genes in each set, as dictated by the matrix at the bottom. The matrix indicates which genotype(s) combine for the vertical bar graph. The super-tolerant FL510 and super-sensitive FL499 had the least number of differentially expressed genes as shown in **(A–C)**. The tolerant FL478 and sensitive parent IR29 had the highest number of upregulated genes, followed by the tolerant donor parent Pokkali and sensitive FL454. Meanwhile, FL454 has the highest number of downregulated genes, followed by FL478, IR29, and Pokkali. The overlap between the subsets of stress responsive genes in FL510 and FL499 are minimal, indicating their responses are unique from each other.

The transgressive super-tolerant FL510 and super-sensitive FL499 had the largest overlaps in their mRNA transcriptomes when the subset of genes whose expression did not changed at EC = 12 ds m⁻¹ were considered (**Figure 3A**). The number of such genes is disproportionately

large compared to any pair-wise comparison, consistent with the fact that the total number of differentially expressed genes were smallest in the two transgressive segregants. The implications of this common signature may be totally unrelated in FL510 and FL499, given their widely contrasting

phenotypes at $EC = 12 \text{ dS m}^{-1}$. The small proportion of differentially expressed genes in FL510 may be an indication of less systemic perturbations, while the regulatory machinery for stress response may be inherently defective or inadequate in FL499.

Global transcriptome patterns ($n = 34,935$ unique protein-coding transcripts) were examined across the panel to trace the parental origins of the expression signatures in each RIL. FL510 and FL499 had the smallest numbers of genes that were not expressed in both control and stress conditions, compared to the other genotypes. General trends also highlighted the uniqueness of the two transgressive segregants by virtue of non-parental expression signatures in a large number of genes (Figure 4). For instance, the super-tolerant FL510 and super-sensitive FL499 had the smallest proportions of genes with the expected parental profiles, hence they also had the highest numbers of genes that deviated significantly from the expected parental profiles (transgressive). However, distinct subsets of

genes were transgressively expressed in the super-tolerant FL510 and super-sensitive FL499.

The tolerant FL478 had the largest number of genes whose expression profiles were similar from either parent, of which a much greater number were from the sensitive parent IR29. This trend was not surprising given that FL478 is a specific selection in IRRI's breeding program for the growth and developmental attributes of IR29, and limited introgression from Pokkali including *SalTol* (Walia et al., 2005; Cotsaftis et al., 2011; Chowdhury et al., 2016). Consistent with these trends, FL478 also had the smallest number of genes with Pokkali-type expression profiles (Figure 4). Meanwhile, the sensitive FL454 had the smallest number of genes with non-parental expression profile, and the highest number of genes with Pokkali-type expression profiles (Figure 4).

The patterns revealed by tracing the parental origins of expression signature for all annotated protein-coding genes in the transcriptome matrix are consistent with the true

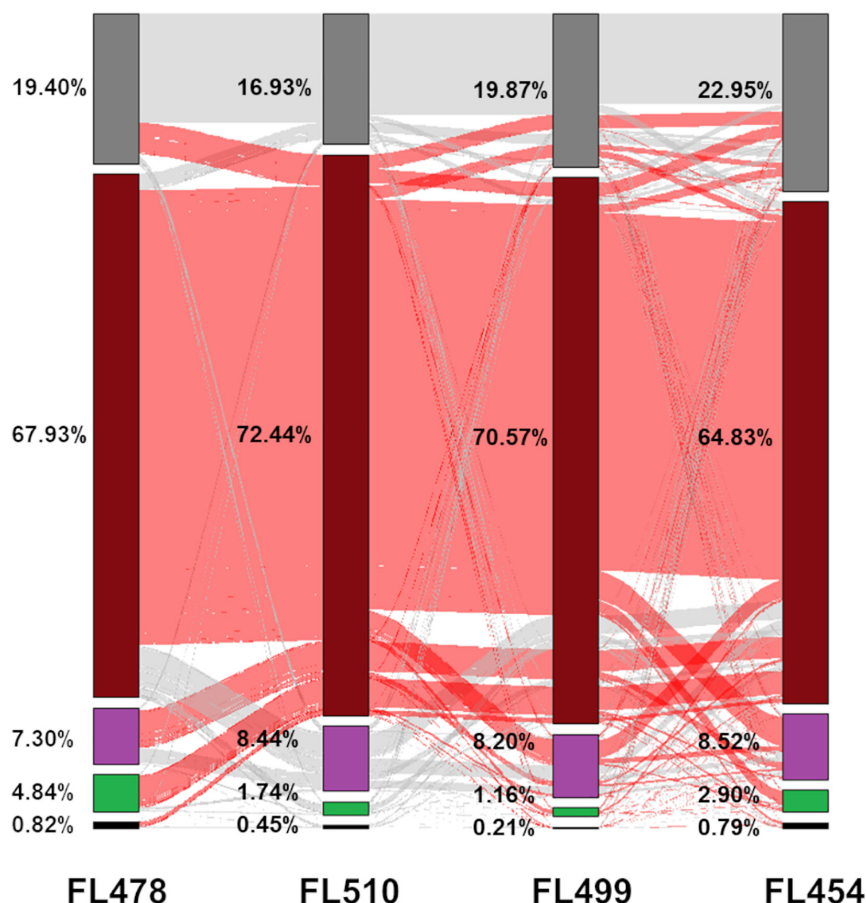


FIGURE 4 | Analysis of the transmission of gene expression signatures from parents to offspring. Each protein-coding gene in the comparative transcriptome matrix was examined to trace the parental origin of respective expression patterns and classified into five categories: (1) complete inheritance (black), meaning that expression profiles were the same in both parents and in the RIL; (2) inherited from IR29 (green), meaning that the RIL had the same expression as IR29; (3) inherited from Pokkali (violet), meaning that the RIL had the same expression as Pokkali; (4) non-parental (brick red), meaning the expression pattern of the gene followed neither parent (deviant or transgressive); or (5) genes was excluded in the analysis due to very low transcript abundance (near zero) or extremely outlying values (gray). The proportion of genes shared with another category is denoted by the red and gray lines in between the bars.

recombinant nature and genetic uniqueness of each RIL from the IR29 × Pokkali F₈ population. In addition to expression signatures that were clearly inherited from either parent, each RIL showed significant deviations from parental gene expression profiles and the number of such deviant genes was highest among the two transgressive segregants at the opposite ends of the phenotypic spectrum. Trends uncovered from the global mRNA transcriptome data suggest that after eight rounds of recombination, network rewiring appeared to be most pronounced among the two phenotypic outliers in the population, i.e., FL510 and FL499.

Cytokinin-Mediated Network May Contribute to Enhanced Growth Potential Under Salt Stress

The unique configuration of the regulatory network mediated by the Myb-type Multi-pass transcription factor *OsMPS* (Os02g0618400) in the transgressive super-tolerant FL510 has been previously reported (Schmidt et al., 2013; Pabuayon et al., 2020). It was proposed that the unique configuration of the *OsMPS* network in FL510 and its activation during the critical initial 24 h of salt stress facilitate an efficient integration of normal growth responses with stress responses, thus contributing to a large positive net gain in the overall potential of FL510. The unique molecular signature of FL510 as revealed by the global trends in the mRNA transcriptome (Figure 3) further supported the initial hypotheses of network rewiring (Pabuayon et al., 2020). To address this hypothesis further, we performed an integrative analysis of unique transcriptome signatures as a means to reveal the components of rewired networks associated with the novel growth and morphological attributes of FL510 (Figures 1, 2).

Our approach was to mine the global transcriptome matrix across the comparative panel for co-expression modules that are unique to FL510 by PCC. The first level of selection for network components included genes that were upregulated by at least twofold ($>2 \log_2\text{-FC}$) across all time-points in FL510 in order to establish the total inducible transcriptome across all genotypes (Figure 5A). To establish the unique signatures of FL510, genes that made it to the initial shortlist was further filtered for those that were upregulated during the first 24 h of stress only in FL510 but downregulated in the inferior RILs. This excluded all other upregulated genes with potentially negative effects in the inferior RILs.

The first co-expressed cluster with a unique and transgressive signature in FL510 is comprise of 244 genes enriched with molecular and biological functions associated with the regulation of growth and development, and plant architecture (Figure 5B). The transgressive nature of this network in the super-tolerant FL510 was reiterated by the fact that FL478 was a clear combination of network signatures from both parents. This network of 244 genes appeared to be regulated primarily through the cytokinin signaling pathway as indicated by signature signaling genes such as *OsRR23* (Os02t0796500-01; B-type response regulator) and *OsHK5* (Os10t0362300-02; cytokinin signal receptor protein), biosynthetic genes such as *OsIPT9*

(Os01t0968700-02; Isopentenyl transferase), and deactivation genes such as N-glucosyltransferase (Os03t0824600-00; cytokinin conjugation enzyme; **Supplementary Table 2**). The *OsIPT9*, *OsRR23*, and *OsHK5* have been reported to modulate adaptive morphology and growth under saline conditions, chemical toxicity, nutrient deficiency, and many biotic factors (Sharan et al., 2017; Ghosh et al., 2018; Nongpiur et al., 2019). The timely and robust expression of this cytokinin-mediated network in the transgressive super-tolerant FL510 is consistent with its unique adaptive morphology that may confer osmotic stress avoidance through regulated transpiration and robust photosynthesis (Figure 2).

Previous studies presented evidence that disjointed or fragmented organization of regulatory networks for growth adjustment and defense-related cellular process contributes to stress sensitivity in certain genotypes of rice (Kitazumi et al., 2018). To understand if the organization of the cytokinin-mediated network is a critical factor in the superior phenotype of FL510 and inferior phenotype of FL499 and other siblings, homologous network models were constructed across the comparative panel by PCC. Co-expression of orthologous genes was defined by network edges with high correlation coefficients of $r \leq -0.95$ (inverse correlation) or $r \geq 0.95$ (direct correlation).

The network model in the super-tolerant FL510 appeared to be well-connected based on consistently high r values, indicative of tight co-expression of component genes amongst each other and with their putative network hubs, *OsIPT9*, *OsRR23*, and *OsHK5* (Figure 6A). In contrast, the orthologous networks in other RILs and in the parents appeared fragmented and disorganized hence inferior to FL510 network (Figures 6B–F). Component genes in these networks form several smaller sub-clusters with different patterns of co-expression that deviate from the general patterns evident in FL510. In FL478, there are clusters missing from the main network relative to the network in FL510 (Figures 6B,D). Furthermore, some of network hubs appeared to be missing in the inferior genotypes (Figures 6C,E,F). Apparently, the deconstruction of the cytokinin-mediated network is worst in the inferior RILs. For instance, in the sensitive FL454, the component genes are not as tightly bound together compared with FL510 and FL478. There are many ‘outlier or straggler’ genes without significant linkages to the network because of weak r values. In the super-sensitive FL499, the network appeared to be lacking any order as most component genes had contrasting expression. This model suggests that the network is practically absent in this genotype. These results further illuminate why FL510 had minimal growth penalty, due to the optimal organization of growth-related genetic networks. Much of the growth enhancement of FL510 could also be attributed to its optimal architecture that provides better adaptive potential. Meanwhile, other genotypes, especially FL454 and the most inferior FL499, have high growth penalty due to the loss of synergy for an organized network.

Detailed examination of all functionally annotated genes in the FL510 network suggests an interplay of growth and development with mechanisms of stress mitigation by cross-talks with cytokinin signaling (**Supplementary Table 2**). For example, the regulatory transcription factor *OsHAP5J/OsNF-YC8*

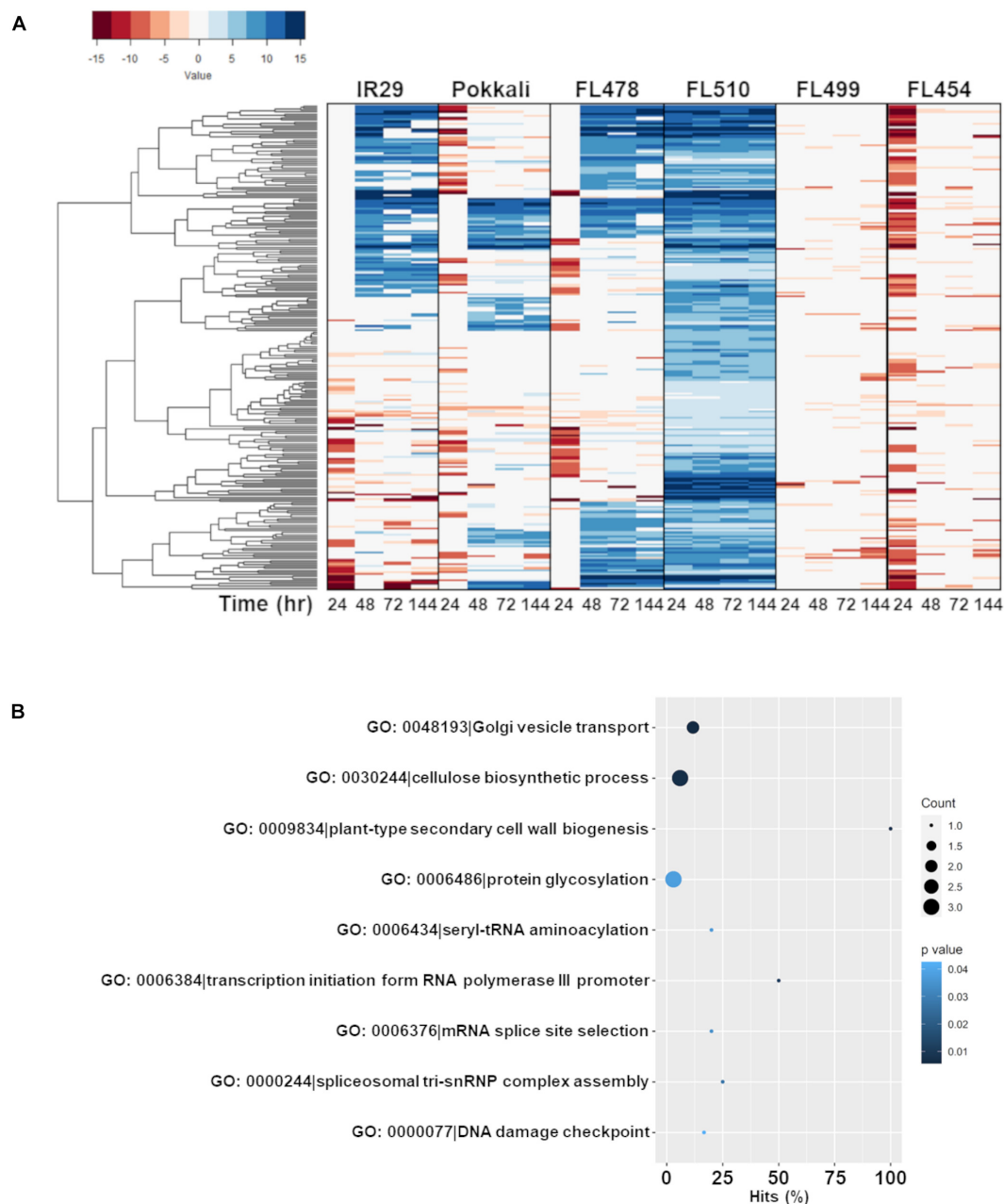


FIGURE 5 | Comparison of the expression profiles of the subset of FL510-upregulated genes across all genotypes and their gene ontology enrichment. Genes that were upregulated by at least twofold ($> 2 \log_2\text{-FC}$) in FL510 across all time-points are included in this subset ($n = 244$). **(A)** Expression patterns of the subset of 244 genes across all six genotypes in IR29 \times Pokkali RIL the comparative panel. While these genes had steady upregulation in FL510, only a fraction were upregulated in the other genotypes. Most of these genes were downregulated in super-sensitive FL499 and sensitive FL454. Upregulation of these genes were significantly delayed in sensitive parent IR29, tolerant donor parent Pokkali, and tolerant FL478. Note: expression under control (0 h) is not displayed as expression data is shown as \log_2 -fold expression relative to control. **(B)** This cluster of FL510-upregulated genes is enriched with functions related to growth and cellulose deposition, indicative of continuous capacity for growth even under salt stress ($P < 0.05$).

(Os01t0580400-01) plays an important role in both development and salt tolerance through ABA and other hormones (Alam et al., 2015; Li et al., 2016). The *OsWRKY13* (Os01t0750100-03) and *OsTGAP1/OsbZIP37* (Os04t0637000-02) transcription factors and the biosynthetic gene *OsOPR8* (Os02t0559400-01)

are components of jasmonic acid (JA) response mechanisms, further reinforcing the importance of JA in the integration of growth and stress-related responses (Qiu et al., 2007; Yoshida et al., 2017). Network compositions also seem to suggest a potential interplay with brassinosteroid signaling, by virtue

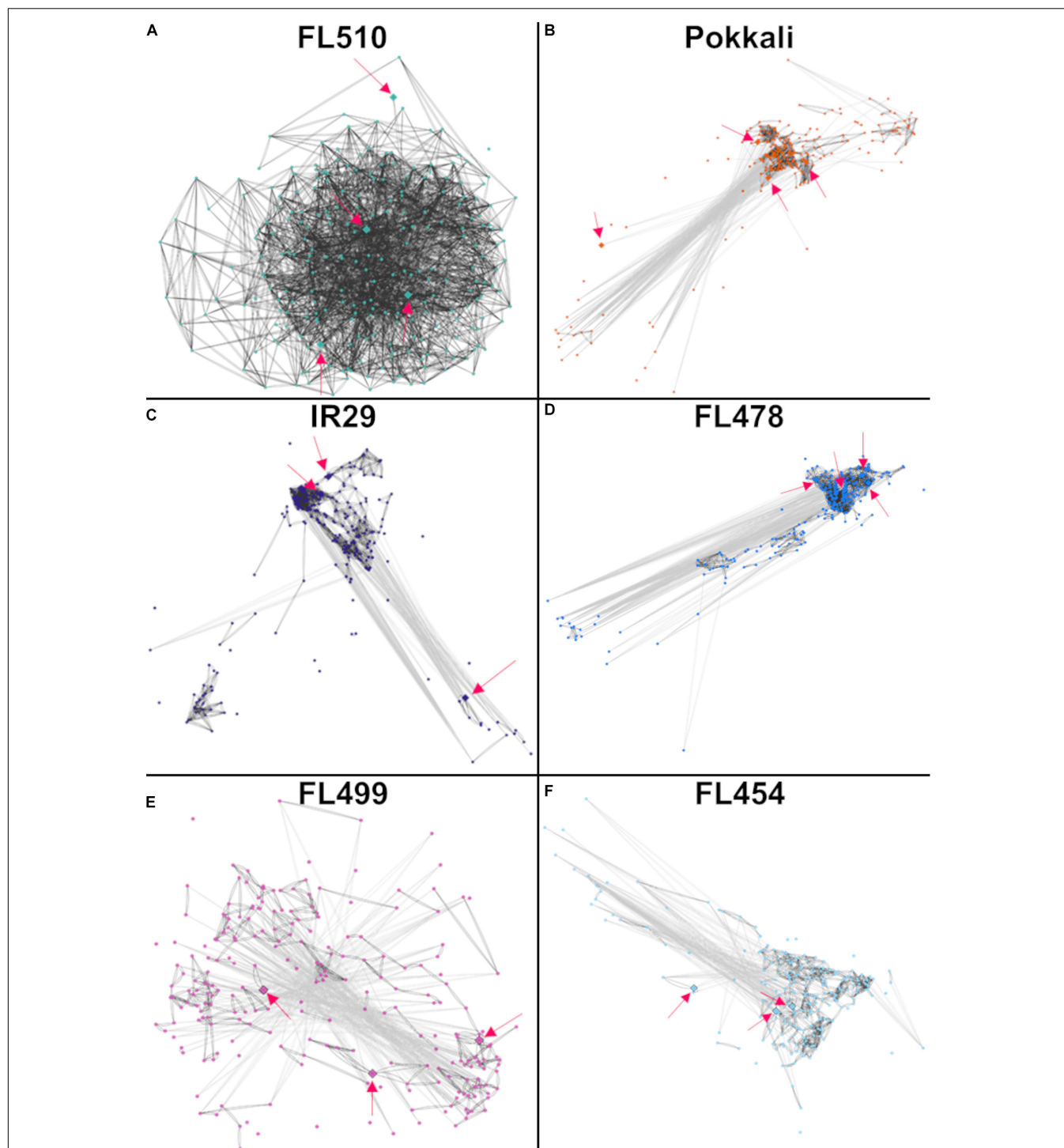


FIGURE 6 | Models of the putative cytokinin-mediated transcriptional network comprised of 244 genes that were upregulated in FL510 as shown in **Figure 5**. Co-expression networks were constructed by Pearson correlation coefficients (PCC). Edges represent r values of at least -0.95 or 0.95 . Dark gray lines indicate direct co-expression between two genes (nodes) and light gray lines indicate reverse co-expression. The distances between nodes are also reflective of the relationship between genes, with lower co-expression values having larger distances. Pink arrows indicate the position of the identified cytokinin-related genes, which are also denoted as diamond nodes. **(A)** The network in FL510 does not show negative co-expression between nodes. In contrast, the network appeared disjointed in **(B)** sensitive parent IR29, **(C)** tolerant donor parent Pokkali, and **(D)** tolerant FL478 with only partial conservation of the network connectivity observed in FL510. **(E)** The network in the super-sensitive FL499 is extensively fragmented, indicating a lack of functional coordination. **(F)** The network in the sensitive FL454 shows a different coordinated cluster that formed from their coordinated downregulation at 24 h. The inferior genotypes excluded one cytokinin-related genes (different in each genotype) as potential hub, further illustrating the fragmented nature of such networks relative to FL510.

of the upregulation of *OsBZR4* (Os02t0233200-00; Bai et al., 2007).

Genetic Network for Novel Adaptive Plant Architecture

The second cluster that formed a unique co-expression network in the transgressive super-tolerant FL510 is comprised of 118 genes with constitutively high expression (i.e., consistent expression not affected by salt stress) only in FL510 at a threshold of $\log_2\text{-FC} > 2$ (Figure 7A). We hypothesized that constitutive networks for the maintenance of plant form and growth habit might be critical to the unique adaptive morphology of FL510 (Figures 1, 2). Similar observations have been reported in *Solanum* and other angiosperms (De Almeida et al., 2014; Ichihashi et al., 2014). This network of 118 constitutively expressed genes is enriched with molecular functions such as response to photo-oxidative stress, proteasome assembly, spindle assembly, photosystem I assembly, and stomatal complex development (Figure 7B). Genes involved in floral organ development such as *FRIGIDA*-like protein (Os03t0794900-03), *HD3A* (Os06t0157700-01; heading date), and *OsFD5* (Os06t0724000-01; homolog of *OsFD1*) were also quite prominent in this network (Supplementary Table 3; Tamaki et al., 2007; Choi et al., 2011; Brambilla et al., 2017). These flowering regulator genes had stable expression in the transgressive super-tolerant FL510, but various patterns of upregulation were evident in the other genotypes, suggestive of perturbed developmental programming due to salt stress.

Most interestingly, enrichment of genes involved in the regulation of plant vegetative morphogenesis was quite evident in this network but only in FL510 (Figure 7A; Supplementary Table 3). This group includes prominent regulators of vegetative growth in grasses such as the transcription factor *OsIBH1* (Os04t0660100-02). The *OsIBH1* is a negative regulator of cell elongation through the brassinosteroid signaling pathway, leading to reduced tiller angle and erect growth habit. The effect of this transcription factor is modulated by negative regulators such as *OsBZR1* (Os07g0580500; *BRASSINAZOLE RESISTANT-1*) and *OsILH1/OsHLLH154* (Os04g0641700; *INCREASED LEAF INCLINATION 1*; Zhang et al., 2009). Also included in this cluster is another transcription factor *TAC3* (Os03t0726700-01; *TILLER ANGLE CONTROL-3*), which determines tiller angle in rice (Dong et al., 2016). The transgressive expression of these transcription factors in FL510, acting as potential hubs of vegetative morphogenesis network, are likely contributing to the unique morphology of FL510, which resembles the NPT architecture by virtue of its erect and broad leaves (Peng et al., 1994).

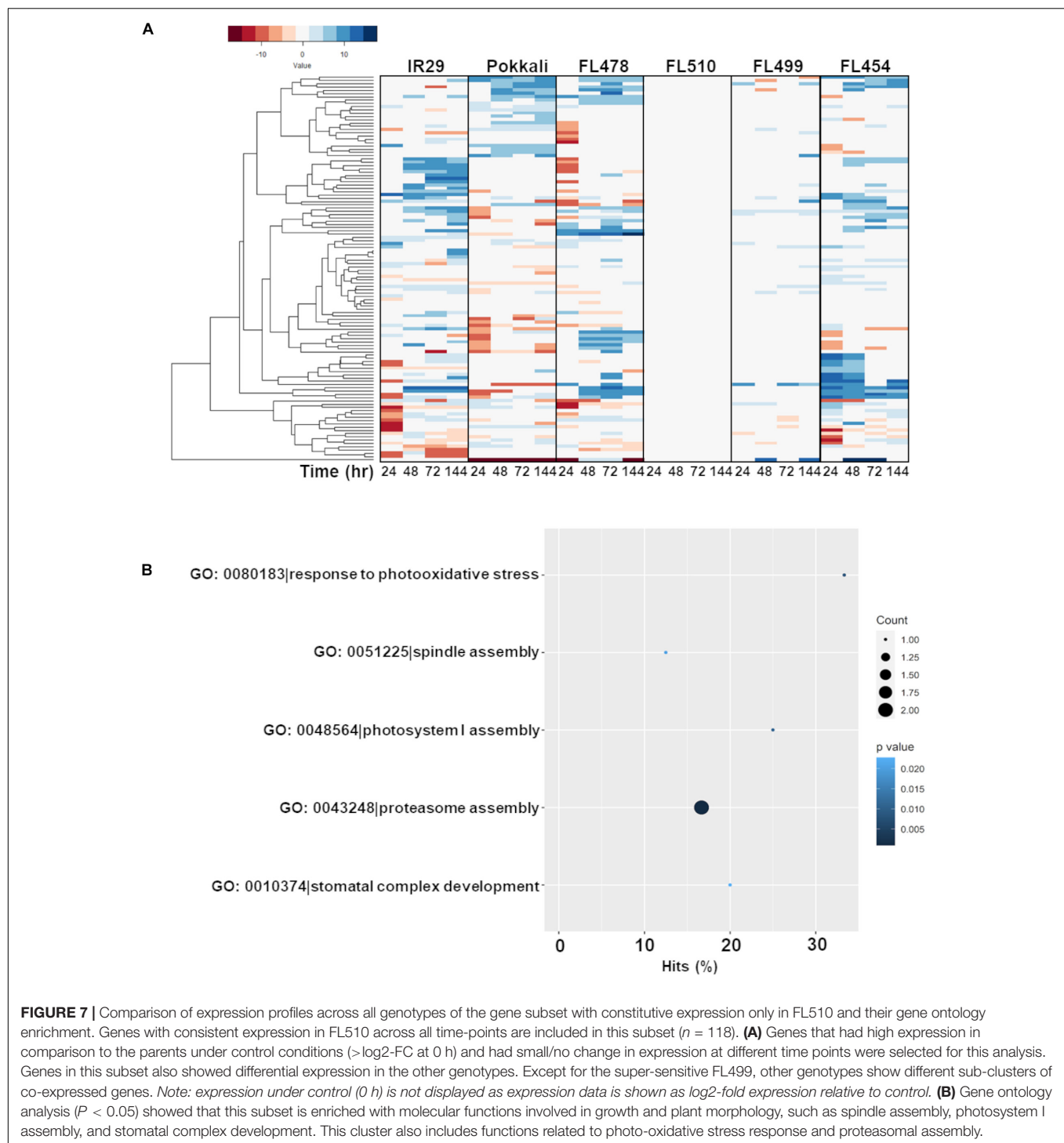
Models of the constitutive vegetative morphogenesis network potentially mediated by *OsIBH1* and *TAC3* transcription factors through the brassinosteroid signaling pathway were constructed across the comparative panel in order to understand its potential significance to the unique attributes of FL510 (Figure 8). A modified threshold for constitutively expressed genes of $r \leq -0.8$ (inverse correlation) or $r \geq 0.8$ (direct correlation) was used for this analysis. In FL510, all the nodes are connected

by edges indicating strong co-expression (dark gray lines; Figure 8A). Except for the tolerant donor parent Pokkali, the other genotypes displayed a lack of inter-connectivity among the smaller sub-clusters, due to the strong co-expression within each sub-cluster or reverse co-expression between the small sub-clusters (Figures 8B–F). This trend reflects the similar trends observed in the cytokinin-mediated network for growth regulation (Figure 6), depicting the lack of coordination among the network components in the inferior genotypes. These trends also suggest the unperturbed nature of vegetative growth and morphogenesis in FL510 compared to the other genotypes.

The small tiller angle and erect leaf growth in the super-tolerant FL510 could have a major contribution to its ability to mitigate the effects of Na^+ toxicity, as an offshoot of decreased water content from osmotic stress (Munns and Tester, 2008). The ability of a plant to exclude Na^+ from functional plant tissues may be supplemented by the ability to retain water to further reduce cellular Na^+ concentration. This may occur through a reduced transpiration rate, which also limits the uptake of solutes including Na^+ (Figure 2A). Thus, osmotic stress may be reduced by the inherent advantages from the unique architecture of FL510. This may be aided by sustained expression of other genes involved in cuticle formation such as *OsHSD1/LGF1* (Os11t0499600-01; hydroxysteroid dehydrogenase; Supplementary Table 1; Zhang et al., 2016). Genes involved in cutin, suberin, and wax biosynthesis showed sustained expression in the super-tolerant FL510 and its tolerant sibling FL478 (Supplementary Figure 1). Taken together, the transgressive nature of the constitutive brassinosteroid vegetative morphogenesis network suggests that the novel architecture of FL510 may complement the defenses acquired from Pokkali to minimize metabolic requirements under marginal conditions.

miRNA Signatures Associated With Transgressive Salt Tolerance

A series of miRNA-Seq libraries was also constructed in parallel to the mRNA-Seq libraries in order to reveal potential post-transcriptional regulatory signatures associated with transgressive phenotypes. The mapping statistics and coverage of the miRNA-Seq libraries are summarized in Supplementary Table 1 (PRJNA378253: SRR12213131 to SRR12213160). The list of non-redundant miRNAs expressed across the comparative matrix is summarized in Figure 9A according to the classification of agronomically important miRNA families (Zhang et al., 2017). While the great majority of salt stress-responsive miRNAs identified across the data matrix had strikingly similar expression profiles across the comparative panel, FL510 appeared to be the most unique by virtue of the narrow ranges of miRNA abundances compared (Figure 9B). This unique profile mirrors the more modulated changes in mRNA abundances in FL510 as shown in Figure 3, indicative of fine-tuned transcriptome and minimal system perturbation. In contrast, the other genotypes had wide ranges of miRNA abundances with extreme outliers, indicating sporadic changes in expression under salt stress (Figure 9B).



Among the shortlist of candidate salt stress-responsive miRNA families with unique signatures are *miR397*, *miR169*, and *miR827* (Zhang et al., 2017). *miR397* has a primarily developmental function and known as a positive regulator of grain size and panicle branching in grasses by modulating the expression of laccase genes (Zhang et al., 2013). Additionally, *miR397* has been reported to be involved in the inhibition of

lignin production that supports secondary cell wall biogenesis and plant growth through cell expansion (Lu et al., 2013; Huang et al., 2016). Laccase genes are also involved in other types of defenses by acting on flavonoids (Turlapati et al., 2011). Upregulation of *miR397* at different temporal patterns was evident across the comparative panel except in the tolerant FL478 and super-tolerant FL510, where it was hardly detectable

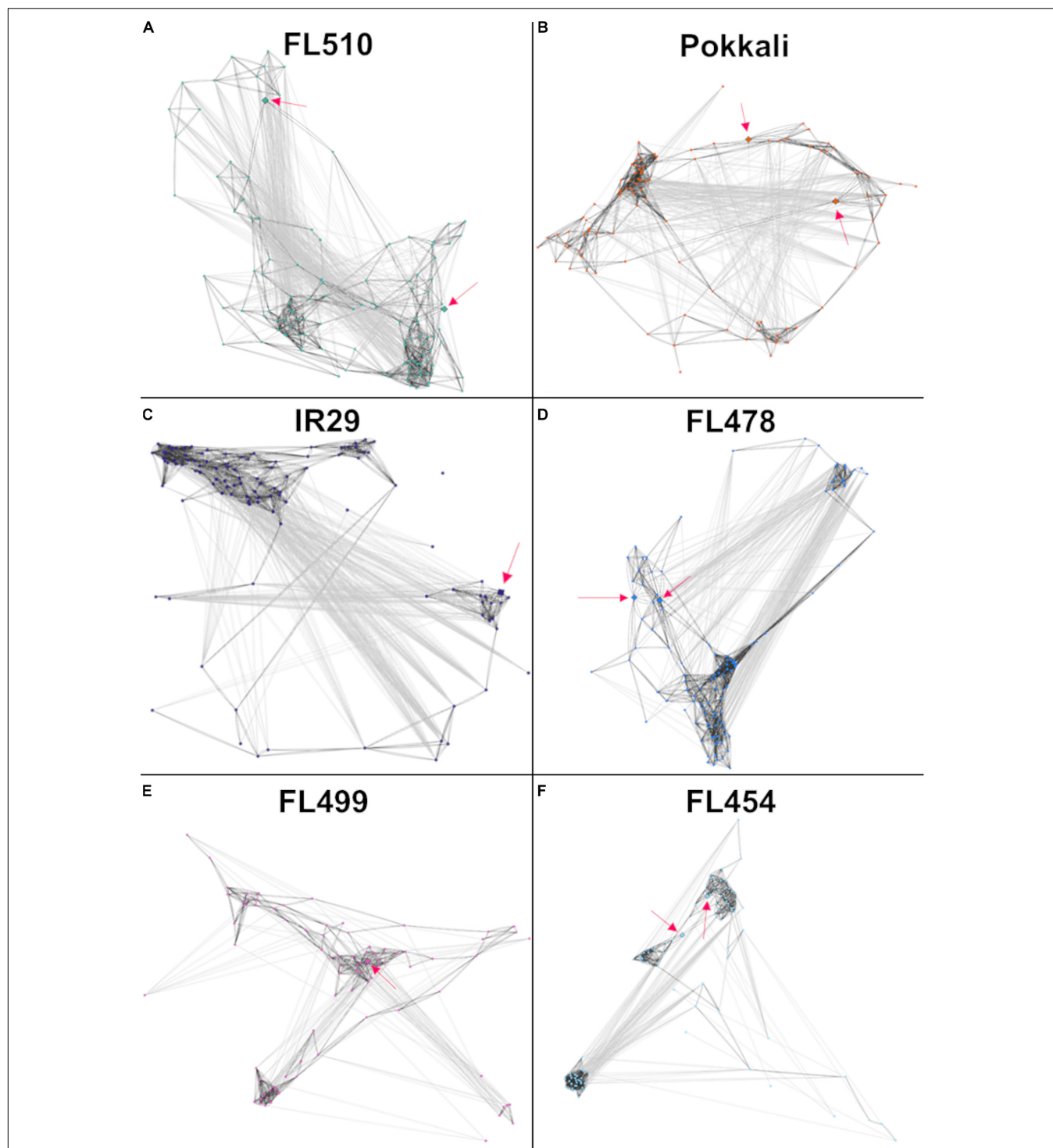


FIGURE 8 | Models of the putative brassinosteroid-mediated transcriptional network comprised of 118 genes with constitutive expression only in FL510 as shown in **Figure 7**. Thresholds for PCC values were adjusted to -0.8 and 0.8, as this gene set had lower correlations due to constitutive expression in FL510. Dark gray lines indicate direct correlation, while light gray lines indicate inverse correlation. Pink arrows indicate the position of the two key genes in this group, namely *OsIBH1* (Os04t0660100-02) and *TAC3* (Os03t0726700-01), which are connected to brassinosteroid signaling and likely network hubs. **(A)** The network of FL510 is shown as a reference well-organized network in comparisons with the other genotypes. Various magnitudes of network fragmentation are shown for **(B)** sensitive parent Pokkali, **(C)** tolerant donor parent IR29, **(D)** tolerant FL478, **(E)** super-sensitive FL499, and **(F)** sensitive FL454. In all the genotypes, strongly co-expressed clusters can be observed. While still showing negative correlations, all the genes are still connected through positive co-expression in FL510 and Pokkali. In the other genotypes, the connectivity is significantly reduced or absent, creating clusters that have opposite expression patterns. These networks may indicate that these gene clusters are not unique among the different genotypes, but rather it is their constant expression in FL510 that makes it distinct.

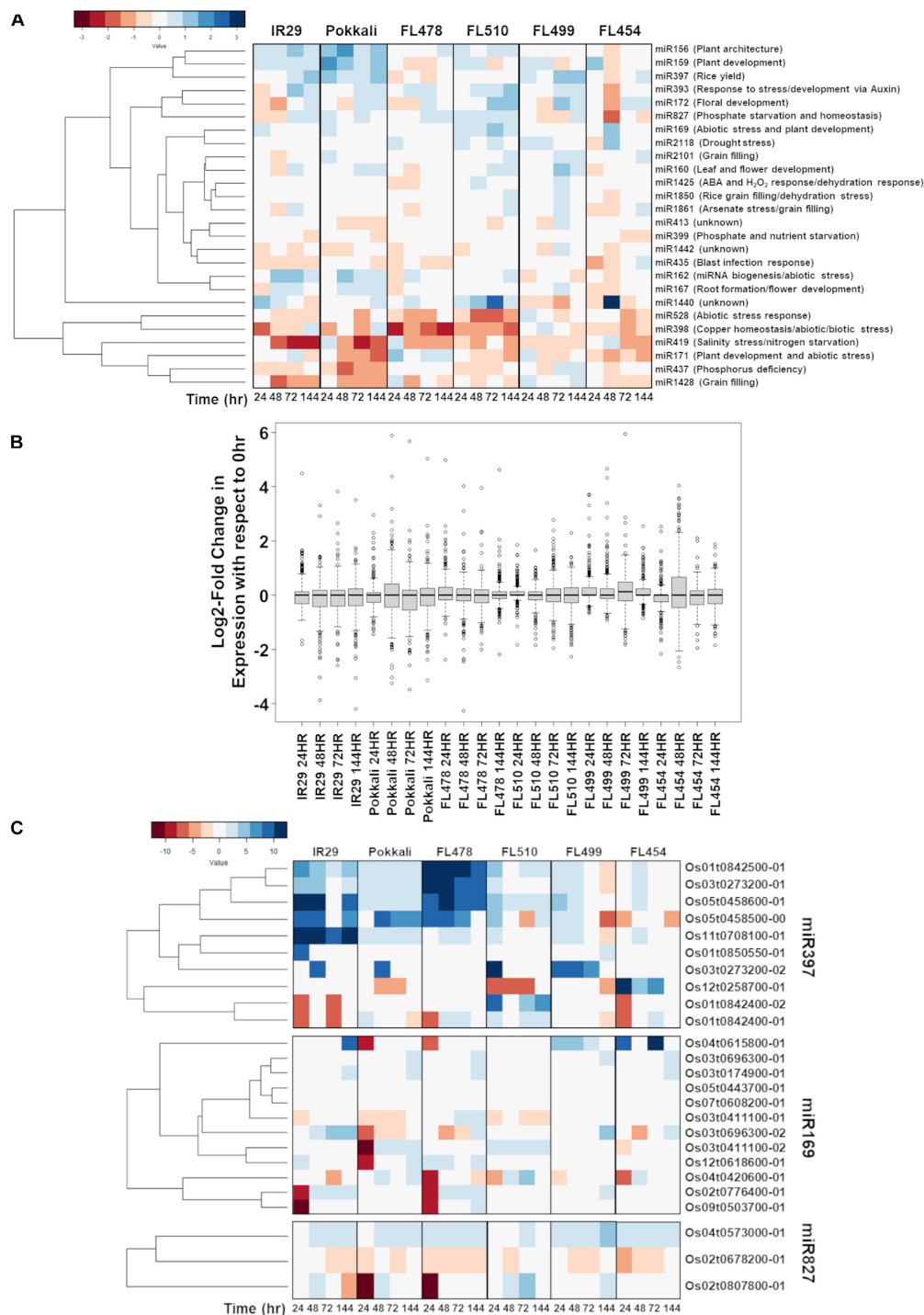


FIGURE 9 | Comparison of the temporal co-expression of major miRNA families and their target genes across the comparative panel of IR29 × Pokkali recombinant inbred lines (RILs). **(A)** Mean expression of all miRNAs detected by RNA-Seq sequencing according to the classification of agronomically important miRNAs (Zhang et al., 2017) to minimize overlap between highly similar but distinctly annotated miRNA. **(B)** The miRNA expression across the different genotypes is similar, but expression in FL510 had smaller magnitude of changes with respect to control (0 h; i.e., narrow ranges). *miR169*, *miR827*, and *miR397* were uniquely expressed in the super-tolerant FL510. In comparison, the other genotypes have much wider ranges of values. **(C)** Expression of the downstream targets of miRNAs across the comparative panel. The *miR397*-target laccase genes were highly upregulated in FL478 and IR29, which suggest flavonoid oxidation as a response to stress. In FL510, the transgressive downregulation of *OsLAC29* (Os12t0258700-01) is consistent with its reported upregulation in salt-sensitive rice cultivars. *OsLAC4* (Os01t0842400-01, Os01t0842400-02) was transgressively upregulated in FL510, consistent with its reported function in maintaining structural integrity of leaves in rice. Differential expression of *OsHAP2E* (Os03t0411100-01, Os03t0411100-02) Pokkali and FL510 suggests targeted regulation of specific splice variant. *OsSPX-MFS1,2* (Os04t0573000-01, Os02t0678200-01) genes showed differential regulation except in FL510.

(Figure 9A). Given its possible role as a negative regulator of lignin production and secondary cell wall formation, the upregulation of *miR397* among the inferior RILs (FL454, FL499) with poor growth under salt stress, and its downregulation in the superior RILs (FL478, FL510) that grew better under salt stress, appeared to be correlated with the differential growth capacity of inferior and superior RILs.

A survey of the *miR397*-target laccase genes across the mRNA-Seq dataset showed strong upregulation particularly in IR29 and FL478 (Figure 9C). As the expression of laccase genes is strongly induced by salt stress in these genotypes, the cause may not be connected to lignin metabolism, but rather toward flavonoid oxidation processes as stress defense component. Meanwhile, Pokkali, FL499, and FL454 had strong upregulation of a laccase gene, i.e., *OsLAC14* (Os05t0458500-00) in Pokkali, *OsLAC11* (Os03t0273200-02) in FL499, and *OsLAC29* (Os12t0258700-01) in FL454. The *OsLAC29* was consistently downregulated in FL510, suggesting that this may be the main laccase gene affected by *miR397* induction. Upregulation of this laccase gene has also been reported in salt-sensitive rice cultivars, including IR29 (Liu et al., 2017). While the data presented does not show the same pattern observed in the previous study, upregulation in the sensitive FL454 and downregulation in the tolerant Pokkali and FL510 indicate that its expression may be detrimental toward tolerance. In contrast, *OsLAC4/OsLacc/OsLAC17* (Os01t0842400-01, Os01t0842400-02) showed transgressive upregulation in FL510. The *Arabidopsis* homolog of this gene is important for lignin deposition in stems to maintain normal growth (Berthet et al., 2011). In rice, this laccase gene was found to be downregulated in mutants with rolled leaf phenotype (Fang et al., 2012). This suggests that the unique expression of *OsLAC4* in FL510 may be relevant to its rigid architecture.

The *miR169* has been reported to improve drought tolerance when overexpressed in tomato by virtue of its roles in the regulation of osmotic imbalance caused by dehydration (Zhang et al., 2011). Sustained upregulation of *miR169* under salt stress was evident only in the super-tolerant FL510, reflecting a transgressive profile (Figure 9A). The unique upregulation of *miR169* in the super-tolerant FL510 may be suggestive of its possible contributions to osmotic adjustment mechanisms. Genes potentially targeted by *miR169* belong to the *Nuclear Factor Y (NF-Y)/Heme Activator Protein (HAP)* family. We found that the expressions of *miR169*-target genes were sparse especially in FL510, FL454, and FL499 (Figure 9C). However, the expression of *OsHAP2E* (Os03t0411100-01, Os03t0411100-02) in FL510 and Pokkali were unique. The longer transcript variant (Os03t0411100-01) was downregulated, while the shorter (Os03t0411100-02) was upregulated, indicating a possible difference in their sensitivity to miRNA-directed degradation. *OsHAP2E* overexpression has been implicated with salinity and drought tolerance (Alam et al., 2015).

Similar to *miR169*, *miR827* was detected at high abundance across the entire duration of salt stress (sustained; transgressive) only in FL510. The *miR827* plays a role in the regulation of responses to phosphate starvation especially under salinity and drought (Lin et al., 2010; Kant et al., 2011; Ferdous et al., 2017;

Shukla et al., 2018). The targets of *miR827* showed a unique expression in FL510 (Figure 9C). There was minimal change in *OsSPX-MFS1* and 2 (Os04t0573000-01 and Os02t0678200-01, respectively), while *OsWAK20* (Os02t0807800-01) was only upregulated in FL510. *OsSPX-MFS* genes are involved in phosphate starvation during salinity (Wang et al., 2015; Wang et al., 2017). It has also been reported that *OsSPX-MFS* genes are expressed in opposite patterns during phosphate starvation, with *OsSPX-MFS1* being downregulated and *OsSPX-MFS2* being upregulated (Lin et al., 2010). Expression of these genes in the inferior genotypes indicates that there is a surplus of phosphate requiring transport to the tonoplast to preserve homeostasis. In contrast, minimal changes in FL510 are indicative of less systemic perturbation. The specific co-upregulation only in the super-tolerant FL510 of two miRNA families (*miR169*, *miR827*) with shared roles in osmotic stress mechanisms is consistent with the growth and morphological attributes of FL510 (Figures 1, 2), implying potential significance to osmotic stress tolerance. Overall, the unique miRNA profiles of FL510 revealed another layer of information that points to the importance of osmotic stress tolerance to its total phenotypic potential, by virtue of its novel morphology and growth habit.

DISCUSSION

Transgressive segregation is observed both in nature and in artificial populations created by plant breeding. However, such individuals exceeding the parental phenotypic range arise only in very small proportions of the population thus also referred to as genetic novelties (Rieseberg et al., 1999; Wagner and Lynch, 2010). Transgressive segregation in natural populations have been proposed as an important driver of adaptive speciation through the nucleation of new phylogenetic lineages with novel ecological niche. This theory has important implications to modern plant breeding paradigms as crop domestication and improvement by directed mating aim for similar outcomes as adaptive speciation, i.e., crops that maintain productivity under marginal environments (de Los Reyes, 2019).

In light of the current paradigms for developing novel adaptive phenotypes for use in marginal agriculture, modern plant breeding and biotechnological approaches must be guided with how evolutionary processes lead to optimization and fine-tuning to create the most adapted individuals. The modern reductionist approach to genetic manipulation considered mostly the inducible defense components of such a complex trait as adaptation, while neglecting the importance of inherent constitutive morpho-developmental attributes as important aspects of adaptive traits. Looking back at the success of the Green Revolution in the 1960's, classical ideotype breeding created the most optimal plant architecture to maximize plant productivity potential under water- and nutrient-rich environments (Hedden, 2003). During the more recent times, a NPT was created in rice that further limits wastage of resources on extensive vegetative growth by only producing few but productive tillers. It was successful in out-yielding *indica* rice cultivars and even spurred the creation of "super" rice hybrids in China (Laza et al.,

2003; Peng et al., 2008). However, for salinity, morphological modifications are often overlooked, as the primary concern in mitigating its effects is excluding Na^+ . Therefore, salinity tolerance is thought to hinge upon the differences in the capacity of Na^+ transporter alleles to translocate toxic ions. The classic paradigm of ideotype breeding must be reintegrated in modern biotechnology and genomics-based crop improvement to create the next generation of crops with high adaptive capacities to marginal conditions, in a manner that is guided by evolutionary principles.

The importance of novel adaptive morphology and development in re-envisioning the current approaches to crop genetic manipulation can best be illustrated by evolutionary examples. The evolution of halophytic plants is a good illustration of how specialized tissues and organs work in synergy with inducible genetic defense mechanisms to provide the plant the capacity to thrive under saline environments. Different species rely on different mechanisms such as succulence, salt secretion, or ion compartmentalization at much higher capacities than glycophytes (Flowers and Colmer, 2008, 2015). There are many commonalities in the mechanisms used by plants for survival under marginal environments (de los Reyes et al., 2018). However, halophytes expand on these common mechanisms through unique morphological modifications (Breckle, 2002). For example, halophytes can expand their capacity to absorb salt by developing succulent leaves or stems. Others like *Atriplex centralasiatica* develop salt glands to secrete NaCl from leaves (Yuan et al., 2016). *Porteresia coarctata* (syn: *Oryza coarctata*) has the capacity to excrete salt from its leaves through microscopic hairs (Flowers et al., 1990; Brar et al., 1997).

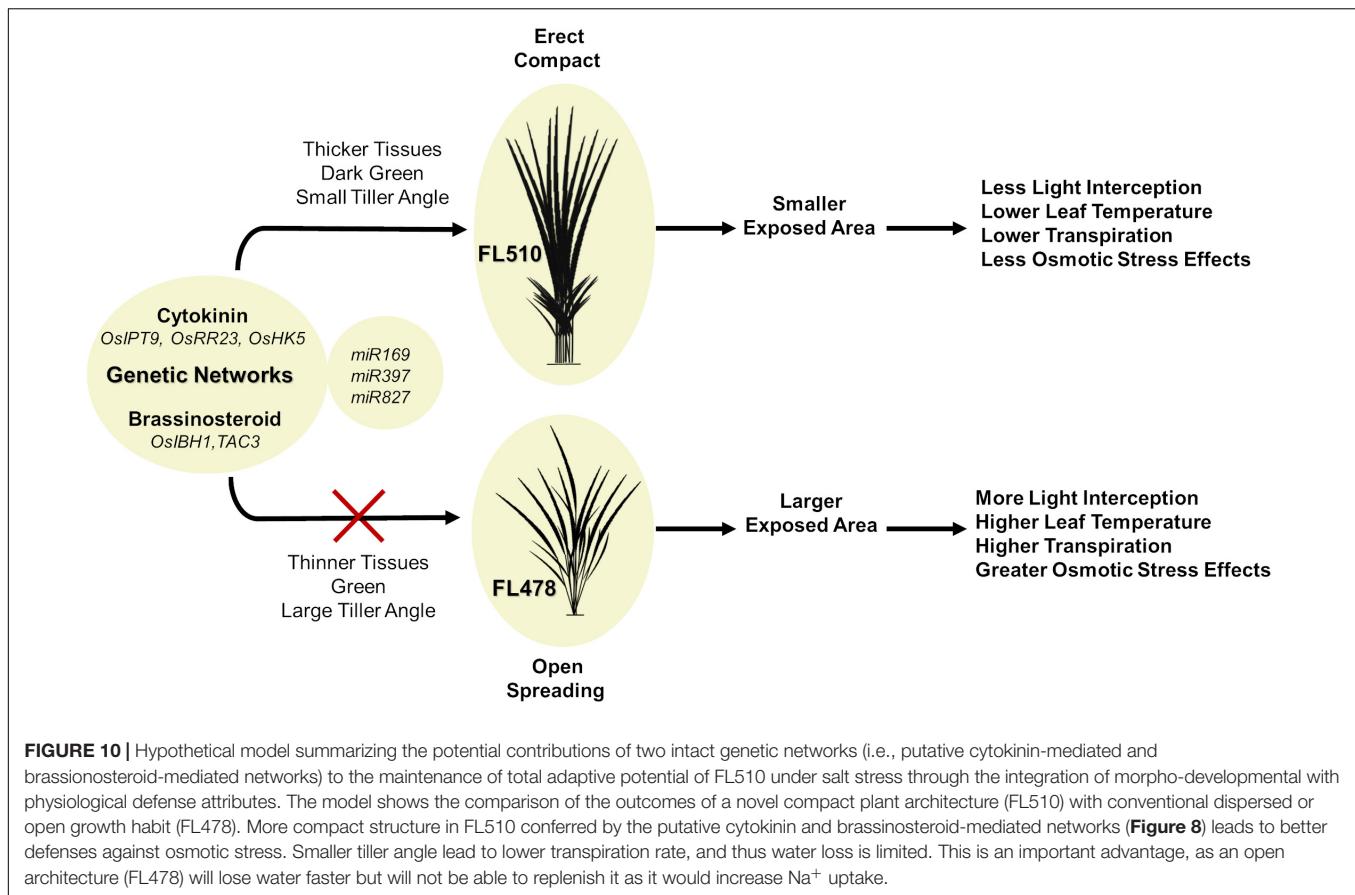
Improving plant performance through morphological modifications has been done for a variety of conditions. For example, *OsSPL14* has been used to improve panicle branching and increase yields in rice (Miura et al., 2010). Morphological traits have been used in conjunction with molecular markers to improve drought tolerance (Manickavelu et al., 2006). More prominently, root system architecture has always been a popular target for improving yield performance under drought (Henry et al., 2011; Kulkarni et al., 2017; Ye et al., 2018). In some cases, characteristics that may impress as very minor can have profound effects in agronomic performance, as in the case of the *DRO1* gene for root architecture (Uga et al., 2013).

Results of this study point to the importance of plant architectural modification created through transgressive segregation, an important process in natural adaptive evolution, in complementing the effects of inducible defenses against salt stress. In the previous study conducted on the same genotypic comparative panel, it was hypothesized that FL510 was more metabolically efficient under salinity stress compared to its siblings and parents (Pabuayon et al., 2020). Both FL510 and FL478 inherited a superior Na^+ exclusion mechanism from the donor parent Pokkali; however, the mechanism of Na^+ exclusion will eventually become overloaded, unsustainable, and ineffective under long-term salinity stress. It will then become necessary to prolong the efficacy of Na^+ exclusion to extend survival through other means that may be directly or indirectly connected to inducible defenses. By identifying such mechanisms, many of

the contributing components behind the transgressive nature of FL510 have been illuminated.

A hypothetical model that integrates the different components behind the novelty of transgressive segregants is shown in **Figure 10**. The most prominent feature that makes FL510 unique is its plant architecture. Its compact structure is the result of combining the smaller stature of its salt-sensitive parent IR29 and the low tiller angle property of its salt-tolerant parent Pokkali. Additionally, its height is intermediate, and its leaf width is transgressively larger compared to its parents and other siblings as a consequence of network rewiring. Studies on leaf width and its association with drought stress suggest critical roles in osmotic stress tolerance and lesser degree of leaf rolling (Cal et al., 2019). Compact structure confers an advantage through the lowering of transpiration rate compared to individuals with more open growth habit. While the open structure allows for more light interception, this also has a trade-off of being more exposed for gas exchange, which allows more water to be released to the air. Under saline conditions, this can be a disadvantage, as water loss translates to increased cellular Na^+ concentration. Uptake for water (solute) will also inevitably bring in more Na^+ , thereby further necessitating Na^+ exclusion, which also brings to a risk of exhausting such capacity faster. If water uptake is limited, the plant experiences osmotic stress, with immediate effects on growth capacity (Munns and Tester, 2008). In contrast, a compact plant architecture retains water more efficiently because of smaller exposed surface area that lowers leaf internal temperature and transpiration rate. This mitigates the effects osmotic stress, allowing a longer period by which water uptake and consequently Na^+ intake could be limited. The net result is the prolongation of the effectiveness of Na^+ exclusion mechanism. While these factors do not directly affect Na^+ exclusion mechanism *per se*, they are necessary to preserve efficacy and help extend the time duration until perturbation and injuries could no longer be controlled.

We reported earlier that except for FL510 and FL499, most RILs derived from IR29 \times Pokkali had intermittent changes in gene expression especially at the short-term (Pabuayon et al., 2020). In contrast, there was minimal induction or repression of genes under salinity in FL510. Genes that were transgressively upregulated in FL510 within the critical first 24 h of salt stress had mainly growth-related functions. This indicates that FL510 either did not experience the same magnitude of stress like the other genotypes, or it continued growth to enhance its capacity to sustain itself despite the sub-optimal conditions. New growth also allows accumulation of more Na^+ in older leaves, assuring the continued function of younger and more active leaves (Yeo and Flowers, 1982). It is also important to have adequate supply of photosynthate in source organs to fuel reproductive growth and grain filling. Immediate shift toward creating new tissues and organs instead of entering stasis, as the case was in the inferior genotypes, may have prevented additional accumulation of Na^+ that would hamper other cellular functions. While the parents and tolerant sibling FL478 showed upregulation of portions of the network in FL510 at much later time-points, such limited responses apparently did not elicit the same positive effects as in FL510. This study also points to the importance of maintaining



constitutive networks that sustain morphogenesis and growth. Typically, genes which are responsive to stress are given more stock as it is often perceived that response equates to adaptation. However, genes that confer adaptation even before the onset of stress could be expressed steadily. Two genes that are potential culprits of the unique architecture of FL510, *OsIBH1* and *TAC3*, were in fact constitutively and transgressively expressed.

It is also important to underscore the additive nature of the genetic networks involved in plant morphology and growth under stress. Co-expression networks allow the visualization of the interconnectivity between genes that may or may not have direct effect with each other, consistent with the Omnigenic theory. It also forms a baseline in determining which genes potentially interact with one another that can be further supplemented with other gene interaction data. In this study, the key genes involved in the cytokinin-mediated network were detected as a complete set in the superior genotypes, but not in the inferior genotypes (**Figure 6**). Additionally, in the tolerant but non-transgressive FL478, the core genes had strong co-expression, but the network was incomplete due to gene sub-clusters that were inversely expressed relative to the major hubs. The same pattern was evident in the putative brassinosteroid-mediated network (**Figure 8**). In the inferior genotypes IR29 and FL499, one of either *TAC3* or *OsIBH1* transcription factors that may be functioning as hubs of the network was missing. Therefore, maximizing the genetic potential of a plant entails

having intricately coordinated expression of many genes that form a network, as observed in FL510. These results point to the cumulative effects of different components the network, consistent with the *Omnigenic Theory*. It may also be possible to use the results from the network analysis to model and quantify the phenotypic effects of synchronized expression in a network, or as training model for genomic selection.

The miRNA genes that were transgressively expressed in the super-tolerant FL510 provide additional support to the ability to mitigate the initial impact of salinity through osmotic effects. Early and consistent upregulation of *miR169* and *miR827* in FL510 but not in the other genotypes suggest that the initial response was active rather than the passive (Pabuayon et al., 2020). This may also explain the low number of differentially expressed genes in FL510, as its response is finely regulated and targeted, requiring only a much smaller group of genes. Specifically, regulation of *OsLAC4* is consistent with the integration of growth and stress responses. The regulation of *OsSPX-MFS* genes also provides an insight into the robust nature of FL510. Phosphate homeostasis is important under salinity stress, as excess phosphate can inhibit growth, compounding the effects of increased salt concentration (Aslam et al., 1996). In comparison, there is minimal regulation of these genes in FL510 suggesting a relatively unperturbed phosphate homeostasis. Tolerance is often associated with survival rate under unfavorable conditions. However, in the case of FL510, tolerance translates

more toward minimal system perturbations. Externally, this is manifested by low growth penalty, and internally by more refined changes in the transcriptome. System perturbations exacerbate the injuries as reactions can be energetically wasteful.

Fine-scale characterization of a transgressive segregant such as FL510 opens new paradigms in breeding for adaptive traits since novel phenotypes can arise through optimal complementation of different traits that may otherwise seem trivial. Repeated recombination events provide more genomic permutations that could lead to genetic novelties by network rewiring. Thus, outlier individuals arising from a population, while seemingly unimportant, can provide insights to different types of possible synergism that are optimal under different environments. This study also points to the importance of traits that are often overlooked. For example, during selection in plant breeding, the capacity to sustain growth under salinity stress is often viewed of secondary importance to Na^+ exclusion capacity. However, the two traits should be complementary, as one trait feeds and expands the capacity of the other. This concept goes back to the *Omnigenic Theory*, which points to seemingly unimportant peripheral genes working in the background of the core genes to create a fine-tuned synergy (Boyle et al., 2017).

In this study, the fragmentation of gene networks in inferior genotypes is suggestive of a loss of contributory peripheral genes, with cumulative effects toward lowering the overall adaptive potential of the plant. Complete networks in FL510 is necessary for maximized phenotypic performance. Therefore, manipulation or addition of singular alleles or genes will not maximize a plant's potential. It is through the synergy of the entire system that the overall phenotypic potential can be fully realized, and this is possible only through the process of genetic recombination. This study also highlights the power of genetic recombination to produce phenotypes which appear to be combinations of minute and indistinguishable traits from the parents yet creating to novel morphologies. The next step to ascertain the effects of these genes would be to investigate loss of function or gain of function mutants with an aim of reconstructing similar morpho-developmental features as FL510. While this is exciting at a theoretical level, it may be challenging to achieve given that FL510 has a unique stacking of beneficial traits. A more feasible approach for further validation would be the analysis of interactome networks, which would entail cis-element and protein-protein interaction analysis. Additionally, the unique trait configuration of FL510 means that the best way to validate the results from this study is to be able to select new genotypes using the genomic model of FL510.

The study also presents a new viewpoint in addressing the problem of what traits should be targeted for crop improvement. Since abiotic stress conditions are multifaceted, approaches for mitigation should also attempt to cover each effect as much as possible. Addressing only one part of the issue, such as using single major QTL is insufficient. Additionally, it is necessary to not just examine phenotypes under stress but also under normal states to discover new mechanisms that can be utilized to improve phenotypic performance. It is necessary to determine the traits that are complementary to the baseline mechanisms (de los Reyes et al., 2018). The study contributes to a comprehensive pipeline

for creating models for larger populations in the future, for genomic selection of new genotypes with maximized phenotypic performance. Traits do not operate in a vacuum with respect to each other. Rather, they act in synergy with one another, or can act antagonistically.

DATA AVAILABILITY STATEMENT

The RNA-Seq data used in this article are publicly available in the NCBI Short Read Archive (PRJNA378253: SRR11528266–SRR11528295 and SRR12213131–SRR12213160).

AUTHOR CONTRIBUTIONS

IP, AK, and BR designed the experiments and wrote the manuscript. IP and AK performed the experiments and data analysis. RS and GG created the recombinant inbred populations and generated initial plant phenotyping data. BR conceptualized the project and designed the experiments. All authors contributed to the article and approved the submitted version.

ACKNOWLEDGMENTS

This work was supported by NSF-IOS Plant Genome Research Program Grant-1602494 and Bayer CropScience Endowed Professorship. Genomic computations were performed using the supercomputing facilities at the ROIS National Institute of Genetics, Mishima, Japan, and Texas Tech University High-Performance Computing Cluster. Next-Gen sequencing was performed at the Oklahoma Medical Research Foundation, Norman, OK.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fgene.2020.594569/full#supplementary-material>

Supplementary Figure 1 | Expression of genes related cutin, suberin, and wax biosynthesis across the different genotypes used in the study. Genes that belong to this biosynthetic pathway in KEGG (ko00073) were surveyed for their expression across the genotypes in the panel and were clustered together hierarchically. Among the genotypes, the two tolerant recombinant inbred lines (RILs) FL510 and FL478 mainly showed upregulation, while the others had both up- and downregulation.

Supplementary Table 1 | Summary and mapping statistics of the RNA-seq libraries constructed, processed, and analyzed for the parental genotypes and selected recombinant inbred lines (RILs).

Supplementary Table 2 | Genes uniquely upregulated in FL510 at all timepoints immediately after imposition of stress and downregulated in sensitive genotypes FL499 and FL454.

Supplementary Table 3 | Genes that show high basal expression in FL510 compared to the parental lines ($\log_2\text{-FC} > 2$) and showed no differential expression under salinity stress.

REFERENCES

- Alam, M. M., Tanaka, T., Nakamura, H., Ichikawa, H., Kobayashi, K., Yaeno, T., et al. (2015). Overexpression of a rice heme activator protein gene (OsHAP2E) confers resistance to pathogens, salinity and drought, and increases photosynthesis and tiller number. *Plant Biotechnol. J.* 13, 85–96. doi: 10.1111/pbi.12239
- Anderson, J. T., Willis, J. H., and Mitchell-Olds, T. (2011). Evolutionary genetics of plant adaptation. *Trends Genet.* 27, 258–266. doi: 10.1016/j.tig.2011.04.001
- Aslam, M., Flowers, T. J., Qureshi, R. H., and Yeo, A. R. (1996). Interaction of phosphate and salinity on the growth and yield of rice (*Oryza sativa* L.). *J. Agron. Crop Sci.* 176, 249–258. doi: 10.1111/j.1439-037X.1996.tb00469.x
- Bai, M.-Y., Zhang, L.-Y., Gampala, S. S., Zhu, S.-W., Song, W.-Y., Chong, K., et al. (2007). Functions of OsBZR1 and 14-3-3 proteins in brassinosteroid signaling in rice. *Proc. Natl. Acad. Sci. U.S.A.* 104, 13839–13844.
- Bauerle, W. L., and Bowden, J. D. (2011). Predicting transpiration response to climate change: insights on physiological and morphological interactions that modulate water exchange from leaves to canopies. *Hortscience* 46, 163–166. doi: 10.21273/HORTSCI.46.2.163
- Berthet, S., Demont-Caulet, N., Pollet, B., Bidzinski, P., Cézard, L., Le Bris, P., et al. (2011). Disruption of *LACCASE4* and 17 results in tissue-specific alterations to lignification of *Arabidopsis thaliana* stems. *Plant Cell* 23, 1124–1137. doi: 10.1105/tpc.110.082792
- Bojanowski, M., and Edwards, R. (2018). *Alluvial: R package for Creating Alluvial Diagrams*. 2016. R package version: 0.1-2.
- Boyle, E. A., Li, Y. I., and Pritchard, J. K. (2017). An expanded view of complex traits: from polygenic to omnigenic. *Cell* 169, 1177–1186. doi: 10.1016/j.cell.2017.05.038
- Brambilla, V., Martignago, D., Goretti, D., Cerise, M., Somssich, M., De Rosa, M., et al. (2017). Antagonistic transcription factor complexes modulate the floral transition in rice. *Plant Cell* 29, 2801–2816.
- Brar, D., Elloran, R., Talag, J., Abbasi, F., and Khush, G. (1997). Cytogenetic and molecular characterization of an intergeneric hybrid between *Oryza sativa* L. and *Porteresia coarctata* (Roxb.) Tateoka. *Rice Genet. Newsl.* 14, 43–44.
- Breckle, S.-W. (2002). "Salinity, halophytes and salt affected natural ecosystems," in *Salinity: Environment - Plants - Molecules*, eds A. Läuchli and U. Lüttge (Dordrecht: Springer Netherlands), 53–77. doi: 10.1007/0-306-48155-3_3
- Cal, A. J., Sanciango, M., Rebollo, M. C., Luquet, D., Torres, R. O., McNally, K. L., et al. (2019). Leaf morphology, rather than plant water status, underlies genetic variation of rice leaf rolling under drought. *Plant Cell Environ.* 42, 1532–1544. doi: 10.1111/pce.13514
- Carrington, J. C., and Ambros, V. (2003). Role of microRNAs in plant and animal development. *Science* 301, 336–338. doi: 10.1126/science.1085242
- Chandler, R. F. (1992). *An Adventure in Applied Science: A History of the International Rice Research Institute*. Los Baños: International Rice Research Institute.
- Choi, K., Kim, J., Hwang, H.-J., Kim, S., Park, C., Kim, S. Y., et al. (2011). The *FRIGIDA* complex activates transcription of *FLC*, a strong flowering repressor in *Arabidopsis*, by recruiting chromatin modification factors. *Plant Cell* 23, 289–303. doi: 10.1105/tpc.110.075911
- Chowdhury, A. D., Haritha, G., Sunitha, T., Krishnamurthy, S. L., Divya, B., Padmavathi, G., et al. (2016). Haplotyping of rice genotypes using simple sequence repeat markers associated with salt tolerance. *Rice Sci.* 23, 317–325.
- Conway, J. R., Lex, A., and Gehlenborg, N. (2017). UpSetR: an R package for the visualization of intersecting sets and their properties. *Bioinformatics* 33, 2938–2940. doi: 10.1093/bioinformatics/btx364
- Cotsaftis, O., Plett, D., Johnson, A. A. T., Walia, H., Wilson, C., Ismail, A. M., et al. (2011). Root-specific transcript profiling of contrasting rice genotypes in response to salinity stress. *Mol. Plant* 4, 25–41.
- Counce, P. A., Keisling, T. C., and Mitchell, A. J. (2000). A uniform, objective, and adaptive system for expressing rice development. *Crop Sci.* 40, 436–443. doi: 10.2135/cropsci2000.402436x
- De Almeida, A. M. R., Yockteng, R., Schnable, J., Alvarez-Buylla, E. R., Freeling, M., and Specht, C. D. (2014). Co-option of the polarity gene network shapes filament morphology in angiosperms. *Sci. Rep.* 4:6194. doi: 10.1038/srep06194
- De Datta, S. K., Tauro, A. C., and Balaoing, S. N. (1968). Effect of plant type and nitrogen level on the growth characteristics and grain yield of indica rice in the tropics 1. *Agron. J.* 60, 643–647. doi: 10.2134/agronj1968.00021962006000060017x
- de Los Reyes, B. G. (2019). Genomic and epigenomic bases of transgressive segregation – new breeding paradigm for novel plant phenotypes. *Plant Sci.* 288:110213. doi: 10.1016/j.plantsci.2019.110213
- de los Reyes, B. G., Kim, Y. S., Mohanty, B., Kumar, A., Kitazumi, A., Pabuayon, I. C. M., et al. (2018). "Cold and water deficit regulatory mechanisms in rice: optimizing stress tolerance potential by pathway integration and network engineering," in *Rice Genomics, Genetics and Breeding*, eds T. Sasaki and M. Ashikari (Singapore: Springer Singapore), 317–359.
- De Mendiburu, F., and De Mendiburu, M. F. (2019). *Package 'Agricolae'. R package version, 1.2–8*.
- DeVicente, M. C., and Tanksley, S. D. (1993). QTL analysis of transgressive segregation in an interspecific tomato cross. *Genetics* 134, 585–596.
- Dittrich-Reed, D. R., and Fitzpatrick, B. M. (2013). Transgressive hybrids as hopeful monsters. *Evol. Biol.* 40, 310–315. doi: 10.1007/s11692-012-9209-0
- Dong, H., Zhao, H., Xie, W., Han, Z., Li, G., Yao, W., et al. (2016). A novel tiller angle gene, *TAC3*, together with *TAC1* and *D2* largely determine the natural variation of tiller angle in rice cultivars. *PLoS Genet.* 12:e1006412. doi: 10.1371/journal.pgen.1006412
- Fahlgren, N., Jogdeo, S., Kasschau, K. D., Sullivan, C. M., Chapman, E. J., Laubinger, S., et al. (2010). MicroRNA gene evolution in *Arabidopsis lyrata* and *Arabidopsis thaliana*. *Plant Cell* 22, 1074–1089.
- Fang, L., Zhao, F., Cong, Y., Sang, X., Du, Q., Wang, D., et al. (2012). Rolling-leaf14 is a 2OG-Fe (II) oxygenase family protein that modulates rice leaf rolling by affecting secondary cell wall formation in leaves. *Plant Biotechnol. J.* 10, 524–532. doi: 10.1111/j.1467-7652.2012.00679.x
- Ferdous, J., Whitford, R., Nguyen, M., Brien, C., Langridge, P., and Tricker, P. J. (2017). Drought-inducible expression of Hv-miR827 enhances drought tolerance in transgenic barley. *Funct. Integr. Genomics* 17, 279–292. doi: 10.1007/s10142-016-0526-8
- Flowers, T., Flowers, S., Hajibagheri, M., and Yeo, A. (1990). Salt tolerance in the halophytic wild rice, *Porteresia coarctata* Tateoka. *New Phytol.* 114, 675–684. doi: 10.1111/j.1469-8137.1990.tb00439.x
- Flowers, T. J., and Colmer, T. D. (2008). Salinity tolerance in halophytes. *New Phytol.* 179, 945–963. doi: 10.1111/j.1469-8137.2008.02531.x
- Flowers, T. J., and Colmer, T. D. (2015). Plant salt tolerance: adaptations in halophytes. *Ann. Bot.* 115, 327–331. doi: 10.1093/aob/mcu267
- Fujita, D., Santos, R. E., Ebron, L. A., Teleanco-Yanoria, M. J., Kato, H., Kobayashi, S., et al. (2009). Development of introgression lines of an Indica-type rice variety, IR64, for unique agronomic traits and detection of the responsible chromosomal regions. *Field Crops Res.* 114, 244–254. doi: 10.1016/j.fcr.2009.08.004
- Ghosh, A., Shah, M. N. A., Jui, Z. S., Saha, S., Fariha, K. A., and Islam, T. (2018). Evolutionary variation and expression profiling of Isopentenyl transferase gene family in *Arabidopsis thaliana* L. and *Oryza sativa* L. *Plant Gene* 15, 15–27. doi: 10.1016/j.plgene.2018.06.002
- Hedden, P. (2003). The genes of the green revolution. *Trends Genet.* 19, 5–9. doi: 10.1016/S0168-9525(02)00009-4
- Henry, A., Gowda, V. R. P., Torres, R. O., McNally, K. L., and Serraj, R. (2011). Variation in root system architecture and drought response in rice (*Oryza sativa*): phenotyping of the OryzaSNP panel in rainfed lowland fields. *Field Crops Res.* 120, 205–214. doi: 10.1016/j.fcr.2010.10.003
- Huang, J.-H., Qi, Y.-P., Wen, S.-X., Guo, P., Chen, X.-M., and Chen, L.-S. (2016). Illumina microRNA profiles reveal the involvement of miR397a in *Citrus* adaptation to long-term boron toxicity via modulating secondary cell-wall biosynthesis. *Sci. Rep.* 6:22900. doi: 10.1038/srep22900
- Ichihashi, Y., Aguilar-Martínez, J. A., Farhi, M., Chitwood, D. H., Kumar, R., Millon, L. V., et al. (2014). Evolutionary developmental transcriptomics reveals a gene network module regulating interspecific diversity in plant leaf shape. *Proc. Natl. Acad. Sci. U.S.A.* 111, E2616–E2621. doi: 10.1073/pnas.1402835111
- Kanehisa, M., and Goto, S. (2000). KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* 28, 27–30. doi: 10.1093/nar/28.1.27
- Kant, S., Peng, M., and Rothstein, S. J. (2011). Genetic regulation by NLA and microRNA827 for maintaining nitrate-dependent phosphate homeostasis in *Arabidopsis*. *PLoS Genet.* 7:e1002021. doi: 10.1371/journal.pgen.1002021
- Kent, W. J. (2002). BLAT—the BLAST-like alignment tool. *Genome Res.* 12, 656–664. doi: 10.1101/gr.229202

- Khraiwesh, B., Zhu, J. K., and Zhu, J. (2012). Role of miRNAs and siRNAs in biotic and abiotic stress responses of plants. *Biochim. Biophys. Acta* 1819, 137–148. doi: 10.1016/j.bbagr.2011.05.001
- Khush, G. S. (1995a). Breaking the yield frontier of rice. *Geojournal* 35, 329–332. doi: 10.1007/BF00989140
- Khush, G. S. (1995b). Modern varieties—their real contribution to food supply and equity. *Geojournal* 35, 275–284. doi: 10.1007/BF00989135
- Khush, G. S. (2001). Green revolution: the way forward. *Nat. Rev. Genet.* 2, 815–822. doi: 10.1038/35093585
- Khush, G. S. (2005). What it will take to feed 5.0 billion rice consumers in 2030. *Plant Mol. Biol.* 59, 1–6. doi: 10.1007/s11103-005-2159-5
- Khush, G. S. (2013). Strategies for increasing the yield potential of cereals: case of rice as an example. *Plant Breed.* 132, 433–436. doi: 10.1111/pbr.1991
- Kim, D., Langmead, B., and Salzberg, S. L. (2015). HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods* 12, 357–360. doi: 10.1038/nmeth.3317
- Kitazumi, A., Pabuayon, I. C. M., Ohyanagi, H., Fujita, M., Osti, B., Shenton, M. R., et al. (2018). Potential of *Oryza officinalis* to augment the cold tolerance genetic mechanisms of *Oryza sativa* by network complementation. *Sci. Rep.* 8:16346. doi: 10.1038/s41598-018-34608-z
- Kulkarni, M., Soolanayakanahally, R., Ogawa, S., Uga, Y., Selvaraj, M. G., and Kagale, S. (2017). Drought response in wheat: key genes and regulatory mechanisms controlling root system architecture and transpiration efficiency. *Front. Chem.* 5:106. doi: 10.3389/fchem.2017.00106
- Laza, M. R., Peng, S., Akita, S., and Saka, H. (2003). Contribution of biomass partitioning and translocation to grain yield under sub-optimum growing conditions in irrigated rice. *Plant Prod. Sci.* 6, 28–35. doi: 10.1626/pp.6.28
- Li, Q., Yan, W., Chen, H., Tan, C., Han, Z., Yao, W., et al. (2016). Duplication of OsHAP family genes and their association with heading date in rice. *J. Exp. Bot.* 67, 1759–1768.
- Liao, Y., Smyth, G. K., and Shi, W. (2013). The Subread aligner: fast, accurate and scalable read mapping by seed-and-vote. *Nucleic Acids Res.* 41:e108. doi: 10.1093/nar/gkt214
- Lin, S.-I., Santi, C., Jobet, E., Lacut, E., El Kholi, N., Karlowski, W. M., et al. (2010). Complex regulation of two target genes encoding SPX-MFS proteins by rice miR827 in response to phosphate starvation. *Plant Cell Physiol.* 51, 2119–2131. doi: 10.1093/pcp/pcq170
- Liu, Q., Luo, L., Wang, X., Shen, Z., and Zheng, L. (2017). Comprehensive analysis of rice laccase gene (*OsLAC*) family and ectopic expression of *OsLAC10* enhances tolerance to copper stress in *Arabidopsis*. *Int. J. Mol. Sci.* 18:209. doi: 10.3390/ijms18020209
- Lu, S., Li, Q., Wei, H., Chang, M.-J., Tunlaya-Anukit, S., Kim, H., et al. (2013). Ptr-miR397a is a negative regulator of laccase genes affecting lignin content in *Populus trichocarpa*. *Proc. Natl. Acad. Sci. U.S.A.* 110, 10848–10853.
- Mallory, A. C., and Vaucheret, H. (2006). Functions of microRNAs and related small RNAs in plants. *Nat. Genet.* 38, S31–S36. doi: 10.1038/ng1791
- Manickavelu, A., Nadarajan, N., Ganesh, S. K., Gnanamalar, R. P., and Chandra Babu, R. (2006). Drought tolerance in rice: morphological and molecular genetic consideration. *Plant Growth Regul.* 50, 121–138. doi: 10.1007/s10725-006-9109-3
- Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.j.* 17:10. doi: 10.14806/ej.17.1.200
- McCarthy, D. J., Chen, Y., and Smyth, G. K. (2012). Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Res.* 40, 4288–4297. doi: 10.1093/nar/gks042
- Miura, K., Ikeda, M., Matsubara, A., Song, X.-J., Ito, M., Asano, K., et al. (2010). OsSPL14 promotes panicle branching and higher grain productivity in rice. *Nat. Genet.* 42, 545–549.
- Monna, L., Kitazawa, N., Yoshino, R., Suzuki, J., Masuda, H., Maehara, Y., et al. (2002). Positional cloning of rice semidwarfing gene, sd-1: rice “green revolution gene” encodes a mutant enzyme involved in gibberellin synthesis. *DNA Res.* 9, 11–17. doi: 10.1093/dnares/9.1.11
- Moriya, Y., Itoh, M., Okuda, S., Yoshizawa, A. C., and Kanehisa, M. (2007). KAAAS: an automatic genome annotation and pathway reconstruction server. *Nucleic Acids Res.* 35, W182–W185. doi: 10.1093/nar/gkm321
- Munns, R., and Tester, M. (2008). Mechanisms of salinity tolerance. *Annu. Rev. Plant Biol.* 59, 651–681. doi: 10.1146/annurev.arplant.59.032607.092911
- Murchie, E. H., Chen, Y.-Z., Hubbart, S., Peng, S., and Horton, P. (1999). Interactions between senescence and leaf orientation determine *in Situ* patterns of photosynthesis and photoinhibition in field-grown rice. *Plant Physiol.* 119, 553–564. doi: 10.1104/pp.119.2.553
- Nongpiur, R. C., Gupta, P., Sharan, A., Singh, D., Singla-Pareek, S. L., and Pareek, A. (2019). “The two-component system: transducing environmental and hormonal signals,” in *Sensory Biology of Plants*, ed. S. Sopory (Singapore: Springer), 247–278.
- Pabuayon, I. C. M., Kitazumi, A., Cushman, K. R., Singh, R. K., Gregorio, G. B., Dhath, B., et al. (2020). Transgressive segregation for salt tolerance in rice due to physiological coupling and uncoupling and genetic network rewiring. *bioRxiv [Preprint]* doi: 10.1101/2020.06.25.171603
- Peng, S., Huang, J., Cassman, K. G., Laza, R. C., Visperas, R. M., and Khush, G. S. (2010). The importance of maintenance breeding: a case study of the first miracle rice variety-IR8. *Field Crops Res.* 119, 342–347.
- Peng, S., Khush, G., and Cassman, K. (1994). “Evolution of the new plant ideotype for increased yield potential,” in *Breaking the Yield Barrier: Proceedings of a Workshop on Rice Yield Potential in Favorable Environments*, ed. K. G. Cassman (Los Banos: International Rice Research Institute), 5–20.
- Peng, S., Khush, G. S., Virk, P., Tang, Q., and Zou, Y. (2008). Progress in ideotype breeding to increase rice yield potential. *Field Crops Res.* 108, 32–38. doi: 10.1016/j.fcr.2008.04.001
- Qiu, D., Xiao, J., Ding, X., Xiong, M., Cai, M., Cao, Y., et al. (2007). OsWRKY13 mediates rice disease resistance by regulating defense-related genes in salicylate- and jasmonate-dependent signaling. *Mol. Plant Microbe Interact.* 20, 492–499.
- R Core Team (2020). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.
- Rieseberg, L. H., Archer, M. A., and Wayne, R. K. (1999). Transgressive segregation, adaptation and speciation. *Heredity* 83, 363–372. doi: 10.1038/sj.hdy.6886170
- Sakai, H., Lee, S. S., Tanaka, T., Numa, H., Kim, J., Kawahara, Y., et al. (2013). Rice annotation project database (RAP-DB): an integrative and interactive database for rice genomics. *Plant Cell Physiol.* 54:e6.
- Sasaki, A., Ashikari, M., Ueguchi-Tanaka, M., Itoh, H., Nishimura, A., Swapan, D., et al. (2002). A mutant gibberellin-synthesis gene in rice. *Nature* 416, 701–702.
- Schindelin, J., Arganda-Carreras, I., Frise, E., Kaynig, V., Longair, M., Pietzsch, T., et al. (2012). Fiji: an open-source platform for biological-image analysis. *Nat. Methods* 9, 676–682.
- Schmidt, R., Schippers, J. H., Mieulet, D., Obata, T., Fernie, A. R., Guiderdoni, E., et al. (2013). MULTIPASS, a rice R2R3-type MYB transcription factor, regulates adaptive growth by integrating multiple hormonal pathways. *Plant J.* 76, 258–273.
- Sharan, A., Soni, P., Nongpiur, R. C., Singla-Pareek, S. L., and Pareek, A. (2017). Mapping the ‘two-component system’ network in rice. *Sci. Rep.* 7:9287.
- Shukla, P. S., Borza, T., Critchley, A. T., Hiltz, D., Norrie, J., and Prithiviraj, B. (2018). *Ascophyllum nodosum* extract mitigates salinity stress in *Arabidopsis thaliana* by modulating the expression of miRNA involved in stress tolerance and nutrient acquisition. *PLoS One* 13:e0206221. doi: 10.1371/journal.pone.0206221
- Smoot, M. E., Ono, K., Ruscheinski, J., Wang, P.-L., and Ideker, T. (2010). Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics* 27, 431–432. doi: 10.1093/bioinformatics/btq675
- Stark, R., Grzelak, M., and Hadfield, J. (2019). RNA sequencing: the teenage years. *Nat. Rev. Genet.* 20, 631–656. doi: 10.1038/s41576-019-0150-2
- Tamaki, S., Matsuo, S., Wong, H. L., Yokoi, S., and Shimamoto, K. (2007). Hd3a protein is a mobile flowering signal in rice. *Science* 316, 1033–1036. doi: 10.1126/science.1141753
- Thomson, M. J., De Ocampo, M., Egdane, J., Rahman, M. A., Sajise, A. G., Adorada, D. L., et al. (2010). Characterizing the *Saltol* quantitative trait locus for salinity tolerance in rice. *Rice* 3, 148–160.
- Turlapati, P. V., Kim, K.-W., Davin, L. B., and Lewis, N. G. (2011). The laccase multigene family in *Arabidopsis thaliana*: towards addressing the mystery of their gene function (s). *Planta* 233, 439–470. doi: 10.1007/s00425-010-1298-3
- Uga, Y., Sugimoto, K., Ogawa, S., Rane, J., Ishitani, M., Hara, N., et al. (2013). Control of root system architecture by DEEPER ROOTING 1 increases rice yield under drought conditions. *Nat. Genet.* 45, 1097–1102.
- Van Wallendael, A., Soltani, A., Emery, N. C., Peixoto, M. M., Olsen, J., and Lowry, D. B. (2019). A molecular view of plant local adaptation: incorporating stress-response networks. *Annu. Rev. Plant Biol.* 70, 559–583.

- Van Zanten, M., Pons, T. L., Janssen, J. A. M., Voesenek, L. A. C. J., and Peeters, A. J. M. (2010). On the relevance and control of leaf angle. *Crit. Rev. Plant Sci.* 29, 300–316.
- Vega, U., and Frey, K. J. (1980). Transgressive segregation in inter and intraspecific crosses of barley. *Euphytica* 29, 585–594. doi: 10.1007/BF00023206
- Vergara, B. S. (1988). Raising the yield potential of rice. *Philipp. Technol. J.* 13, 3–19.
- Vikram, P., Kadam, S., Singh, B. P., Lee, Y. J., Pal, J. K., Singh, S., et al. (2016). Genetic diversity analysis reveals importance of green revolution gene (*Sd1* locus) for drought tolerance in rice. *Agric. Res.* 5, 1–12.
- Wagner, G. P., and Lynch, V. J. (2010). Evolutionary novelties. *Curr. Biol.* 20, R48–R52. doi: 10.1016/j.cub.2009.11.010
- Walia, H., Wilson, C., Condamine, P., Liu, X., Ismail, A. M., Zeng, L., et al. (2005). Comparative transcriptional profiling of two contrasting rice genotypes under salinity stress during the vegetative growth stage. *Plant Physiol.* 139, 822–835.
- Wang, C., Yue, W., Ying, Y., Wang, S., Secco, D., Liu, Y., et al. (2015). Rice SPX-major family superfamily3, a vacuolar phosphate efflux transporter, is involved in maintaining phosphate homeostasis in rice. *Plant Physiol.* 169, 2822–2831.
- Wang, D., Lv, S., Jiang, P., and Li, Y. (2017). Roles, regulation, and agricultural application of plant phosphate transporters. *Front. Plant Sci.* 8:817. doi: 10.3389/fpls.2017.00817
- Warnes, G. R., Bolker, B., Bonebakker, L., Gentleman, R., and Huber, W. (2016). 'gplots'. *Various R Programming Tools for Plotting Data*. Available online at: <https://CRAN.R-project.org/package=gplots> (accessed June 8, 2020).
- Wickham, H., and Chang, W. (2008). *ggplot2: An Implementation of the Grammar of Graphics*. R package version 0.7.
- Ye, H., Roorkiwal, M., Valliyodan, B., Zhou, L., Chen, P., Varshney, R. K., et al. (2018). Genetic diversity of root system architecture in response to drought stress in grain legumes. *J. Exp. Bot.* 69, 3267–3277.
- Yeo, A. R., and Flowers, T. J. (1982). Accumulation and localisation of sodium ions within the shoots of rice (*Oryza sativa*) varieties differing in salinity resistance. *Physiol. Plant.* 56, 343–348. doi: 10.1111/j.1399-3054.1982.tb00350.x
- Yoshida, Y., Miyamoto, K., Yamane, H., Nishizawa, Y., Minami, E., Nojiri, H., et al. (2017). OsTGAP1 is responsible for JA-inducible diterpenoid phytoalexin biosynthesis in rice roots with biological impacts on allelopathic interaction. *Physiol. Plant.* 161, 532–544. doi: 10.1111/ppl.12638
- Young, M. D., Wakefield, M. J., Smyth, G. K., and Oshlack, A. (2010). Gene ontology analysis for RNA-seq: accounting for selection bias. *Genome Biol.* 11:R14. doi: 10.1186/gb-2010-11-2-r14
- Yuan, F., Leng, B., and Wang, B. (2016). Progress in studying salt secretion from the salt glands in recretohalophytes: how do plants secrete salt? *Front. Plant Sci.* 7:977. doi: 10.3389/fpls.2016.00977
- Zhang, H., Zhang, J., Yan, J., Gou, F., Mao, Y., Tang, G., et al. (2017). Short tandem target mimic rice lines uncover functions of miRNAs in regulating important agronomic traits. *Proc. Natl. Acad. Sci. U.S.A.* 114, 5277–5282. doi: 10.1073/pnas.1703752114
- Zhang, L. Y., Bai, M. Y., Wu, J., Zhu, J. Y., Wang, H., Zhang, Z., et al. (2009). Antagonistic HLH/bHLH transcription factors mediate brassinosteroid regulation of cell elongation and plant development in rice and *Arabidopsis*. *Plant Cell* 21, 3767–3780.
- Zhang, X., Zou, Z., Gong, P., Zhang, J., Ziaf, K., Li, H., et al. (2011). Over-expression of microRNA169 confers enhanced drought tolerance to tomato. *Biotechnol. Lett.* 33, 403–409.
- Zhang, Y.-C., Yu, Y., Wang, C.-Y., Li, Z.-Y., Liu, Q., Xu, J., et al. (2013). Overexpression of microRNA OsmiR397 improves rice yield by increasing grain size and promoting panicle branching. *Nat. Biotechnol.* 31, 848–852.
- Zhang, Z., Cheng, Z. J., Gan, L., Zhang, H., Wu, F. Q., Lin, Q. B., et al. (2016). OsHSD1, a hydroxysteroid dehydrogenase, is involved in cuticle formation and lipid homeostasis in rice. *Plant Sci.* 249, 35–45.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2020 Pabuayon, Kitazumi, Gregorio, Singh and de los Reyes. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



The Effects of Gene Duplication Modes on the Evolution of Regulatory Divergence in Wild and Cultivated Soybean

OPEN ACCESS

Edited by:

Deborah A. Triant,
University of Virginia, United States

Reviewed by:

Stefan Laurent,
Max Planck Institute for Plant
Breeding Research, Germany
Thomas Kono,
University of Minnesota Twin Cities,
United States
Kevin Silverstein,
University of Minnesota Twin Cities,
United States

*Correspondence:

Scott A. Jackson
scott.jackson@bayer.com;
sjackson@uga.edu
Chunming Xu
xucm848@nenu.edu.cn

[†]These authors have contributed
equally to this work

*Present address:

Scott A. Jackson,
Bayer Crop Science, Chesterfield,
MO, United States

Specialty section:

This article was submitted to
Systems Biology,
a section of the journal
Frontiers in Genetics

Received: 31 August 2020

Accepted: 04 November 2020

Published: 08 December 2020

Citation:

Zhao N, Ding X, Lian T, Wang M,
Tong Y, Liang D, An Q, Sun S,
Jackson SA, Liu B and Xu C (2020)
The Effects of Gene Duplication
Modes on the Evolution of Regulatory
Divergence in Wild and Cultivated
Soybean. *Front. Genet.* 11:601003.
doi: 10.3389/fgene.2020.601003

Na Zhao^{1,2†}, Xiaoyang Ding^{3†}, Taotao Lian², Meng Wang², Yan Tong², Di Liang², Qi An²,
Siwen Sun¹, Scott A. Jackson^{4*†}, Bao Liu² and Chunming Xu^{2*}

¹ Department of Agronomy, Jilin Agricultural University, Changchun, China, ² Key Laboratory of Molecular Epigenetics of Ministry of Education (MOE), Northeast Normal University, Changchun, China, ³ Soybean Research Institute, Jilin Academy of Agricultural Sciences, Changchun, China, ⁴ Center for Applied Genetic Technologies, University of Georgia, Athens, GA, United States

Regulatory changes include divergence in both *cis*-elements and *trans*-factors, which play roles in organismal evolution. Whole genome duplications (WGD) followed by diploidization are a recurrent feature in the evolutionary history of angiosperms. Prior studies have shown that duplicated genes have different evolutionary fates due to variable selection constraints and results in genomic compositions with hallmarks of paleopolyploidy. The recent sequential WGDs and post-WGD evolution in the common ancestor of cultivated soybean (*Glycine max*) and wild soybean (*Glycine soja*), together with other models of gene duplication, have resulted in a highly duplicated genome. In this study, we investigated the transcriptional changes in *G. soja* and *G. max*. We identified a sizable proportion of interspecific differentially expressed genes (DEGs) and found parental expression level dominance of *G. max* in their F1 hybrids. By classifying genes into different regulatory divergence types, we found the *trans*-regulatory changes played a predominant role in transcriptional divergence between wild and cultivated soybean. The same gene ontology (GO) and protein family (Pfam) terms were found to be over-represented in DEGs and genes of *cis*-only between JY47 and GS, suggesting the substantial contribution of *cis*-regulatory divergences to the evolution of wild and cultivated soybeans. By further dissecting genes into five different duplication modes, we found genes in different duplication modes tend to accumulate different types of regulatory differences. A relatively higher proportion of *cis*-only regulatory divergences was detected in singleton, dispersed, proximal, and tandem duplicates than WGD duplicates and genome-wide level, which is in line with the prediction of gene balance hypothesis for the differential fates of duplicated genes post-WGD. The numbers of *cis*-only and *trans*-only regulated genes were similar for singletons, whereas there were more genes of *trans*-only than *cis*-only in the rest duplication types, especially in WGD in which there were two times more *trans*-only genes than that in *cis*-only type. Tandem duplicates showed the highest proportion of *trans*-only genes probably due to

some special features of this class. In summary, our results demonstrate that genes in different duplication modes have different fates in transcriptional evolution underpinned by *cis*- or *trans*-regulatory divergences in soybean and likely in other paleopolyploid higher organisms.

Keywords: soybean, hybrid, regulatory divergence, duplicate gene, *Glycine max*, *Glycine soja*

INTRODUCTION

Cultivated soybean (*Glycine max* L. Merr.) is believed to be domesticated from wild soybean (*Glycine soja* Sieb. and Zucc.) in East Asia 6,000–9,000 years ago (Kim et al., 2012). However, recent genomic studies suggested that soybean domestication was a complex process involving introgressions between wild and domesticated soybeans (Kim et al., 2010; Li et al., 2014; Wang et al., 2019). Although the origin and domestication of soybean are still under debate, the two species have accumulated enormous genetic and phenotypic changes since their divergence (Gong, 2020). Nevertheless, *G. max* and *G. soja* can be hybridized to form fertile offspring with mostly normal meiotic chromosome pairing. Phenotypic differences between *G. max* and *G. soja* can arise from functional divergence of gene products as well as regulatory divergence of their expression. The evolution in gene products has historically received more attention because they can be easily detected. With the development of new technologies, methods for identifying the genetic changes that underlie expression changes have been developed (Wittkopp et al., 2004; McManus et al., 2010). Transcriptional regulation includes two major components: *cis*-acting elements (i.e., promoters, enhancers, and silencers) and *trans*-acting factors (i.e., transcription factors and non-coding regulatory RNAs). Gene expression is controlled by biochemical interactions between *cis*-acting elements and *trans*-acting factors. Regulatory divergence, including both *cis*- and *trans*-acting changes, can be inferred through comparing differences in gene expression between two genotypes to differences in allelic expression in their F1 hybrids (Wittkopp et al., 2008). Previous studies showed that *trans*-regulatory divergence often make larger contributions to gene expression differences than *cis*-regulatory divergence within species, whereas *cis*-regulatory divergence makes either similar or greater contributions to gene expression divergence between species (Zhuang and Adams, 2007; Wittkopp et al., 2008; Emerson et al., 2010; Goncalves et al., 2012; Lemmon et al., 2014; Xu et al., 2014; Guerrero et al., 2016; Wu et al., 2016). *Cis*-regulatory changes preferentially accumulate over time which fits the theory that *trans*-regulatory changes are selected against by purifying selection and many *cis*-regulatory changes are selected for by positive selection (Prud'homme et al., 2007; Emerson et al., 2010; Coolon et al., 2014). Domesticated plants have experienced unique evolutionary bottlenecks which may lead to differences in the relative contributions of *cis*- and *trans*-regulatory divergence relative to undomesticated taxa (Lemmon et al., 2014).

Whole genome duplications (WGD) or polyploidization are prevalent and recurring throughout the evolutionary histories of all flowering plants (Jiao et al., 2011; Wei and Ge, 2011). Two ancestral WGD events occurred in the common ancestor

of seed plants and the common ancestor of angiosperms, respectively (Jiao et al., 2011). The majority of genes duplicated by WGD will return to a single copy over evolutionary time, whereas some duplicated genes will be retained. The fates of duplicated genes following WGD have attracted much interest. Several models have been proposed to explain the loss or retention of duplicated genes. The neofunctionalization and sub-functionalization hypotheses predict that duplicated copies evolve neutrally (Innan and Kondrashov, 2010) and are retained by acquiring new function or reciprocal loss-of-function mutations (He and Zhang, 2005). Another widely accepted hypothesis is the gene balance hypothesis that states the stoichiometry of members of multisubunit complexes affects the function of the whole due to the kinetics and mode of assembly (Birchler and Veitia, 2010). The gene balance hypothesis predicts that all gene duplicates are not retained equally and that loss of dosage-sensitive WGD genes in an interacting balance relationship with others will be selected against in post-WGD evolutionary processes (Birchler and Veitia, 2007, 2010). This has been supported with evidence that WGD-derived duplicated genes are enriched in signal transduction components and transcription factors in multiple plant species. Meanwhile, these functional categories were found to be under-represented in genes duplicated by small-scale duplications e.g., tandem duplication (Blanc and Wolfe, 2004; Maere et al., 2005; Chapman et al., 2006; Xu et al., 2018). Since genes in different duplication modes are the result of and/or under different selection pressures, it is interesting to investigate the relationship of gene duplication mode and types of regulatory divergence.

Besides the two ancient WGDs, *G. max* and *G. soja* experienced two additional sequential WGD events; one occurred about 59 MYA in the common ancestor of legumes and the other about 8–13 MYA in the *Glycine* lineage (Schmutz et al., 2010; Chen et al., 2020). More than 75% of genes in the paleopolyploid soybean genome are multiple copies, and most of these resulted from the WGD events (Schmutz et al., 2010). Recent studies of duplicated genes in soybean showed that genes in different duplication modes have different expressions and gene body DNA methylation profiles (Xu et al., 2018). The functional classification and expression divergence of WGD genes supported different hypotheses of duplicate gene evolution (Xu et al., 2018). The WGD genes in soybean were found to be enriched in *Glycine* transcription factors and transcription regulation functions, which fits the gene balance hypothesis (Xu et al., 2018) and indicates variable constraints on the evolution of genes derived from different duplication modes. In this study, we investigated the transcriptional changes and regulatory divergences as well as their functional preference and relationship with duplication modes in *G. max* and *G. soja*. We reveal the

effects of gene duplication modes on the evolution of gene expression and regulation in wild and cultivated soybean.

MATERIALS AND METHODS

Growth Condition, RNA Extraction, and Sequencing

Jiyu47 (JY47) is a soybean elite cultivar which is mainly planted in northeast China. The wild soybean GS was collected from middle China. The hybrid between wild and cultivated soybean was created using JY47 and GS as paternal and maternal parents respectively. The seeds of the three genotypes were planted into soil and grown in a growth chamber under 18-h light and 6-h dark cycles. The temperatures were 25 and 22°C in day and night, respectively. The plants were grown until the first trifoliolate was fully developed; then, the second trifoliolate leaf was harvested and frozen in liquid nitrogen. For each genotype, three individuals were harvested and stored separately. RNA was extracted for each individual plant using the Trizol method according to the manufacturer's instruction. The total RNA samples were sent to a sequencing company for library construction and sequencing. The sequencing platform was Novaseq 6000. The raw reads were cleaned to remove adapter contamination, low quality reads, and reads with more than 5% N bases. At least 5 Gb of clean bases were produced for each sample.

RNA-seq Data Processing, Mapping, and Identifying Differentially Expressed Genes

Equal-amount reads from both parental samples were mixed and served as the *in silico* hybrid. Three *in silico* hybrid replicates were created using different parental samples. Then, RNA-seq data were mapped to cultivated soybean reference genome (Williams 82, version: a2v1) using STAR (version 2.7.0d) with settings to report the alignments of uniquely mapped reads (Dobin et al., 2013). Gene expression data were filtered, and genes whose average read counts were bigger than 10 and less than 1,000 were kept. Gene expression levels between genotypes were normalized and compared using DESeq2 with default setting (Wald test) (Love et al., 2014). The differentially expressed genes (DEGs) were identified using a cutoff of FDR adjusted p -value < 0.05. The same processes were conducted using wild soybean genome (GCF_004193785.1) as the reference to examine the impacts of mapping preference on the DEG analysis. A detailed description of command and parameters can be found in the supplementary notes.

DNA Sequencing Data Processing and SNP Calling

The raw DNA sequencing data were filtered to remove adapter contamination, low-quality reads, and reads with more than 5% N bases, then trimmed using "Trimmomatic-0.39" with the parameter "LEADING:5 TRAILING:5 MINLEN:75" (Bolger et al., 2014). Clean reads of the two parental genotypes were mapped against the cultivated reference genome using BWA

with default settings (Li and Durbin, 2009). Variants were called using the HaplotypeCaller tool, then both parental genotypes were jointly genotyped using the GenotypeGVCFs tool in GATK (version 4.1.3.0). The raw variants were filtered using VariantFiltration with a setting of "QD < 2.0, QUAL < 30.0, SOR > 3.0, FS > 60.0, MQ < 40.0, MQRankSum < -12.5, and ReadPosRankSum < -8.0." Then, only bi-allelic SNPs with genotype quality >20 and sample depth >5 were kept. Equal amounts of DNA sequencing reads were mixed and mapped to the reference genome. A detailed description of commands and parameters can be found in the Supplementary Materials.

Calculating Allelic Expression

The BAM files generated from mapping F1 and *in silico* hybrid RNA-seq data and mixed DNA data were used for allelic analysis. Allelic read counts were calculated at each SNP site using ASEReadCounter tool in GATK. The mapped RNA or DNA reads covering these sites were assigned to JY47 or GS based on the SNPs. SNPs were filtered to remove sites with biased parental DNA read counts (binomial test p -value < 0.05 for 1:1 ratio) in the mixed DNA sample. Genes with less than two SNPs between parental genotypes were excluded in further analysis. For each gene, the allele specific expression was calculated by summing the number of JY47 reads or GS reads in the body region. A detailed description of commands and parameters can be found in the Supplementary Materials.

Assignment of Regulatory Divergence Types and Duplication Modes

The regulatory divergence types were assigned using the method described in McManus et al. (2010). Briefly, the relative allelic expression of every gene was tested in F1 hybrid (named H comparison) and *in silico* hybrid (named P comparison) using binomial test against the null hypothesis of 1:1 respectively, and compared between F1 and *in silico* hybrid (named T comparison) using Fisher's exact test. The difference was classified as significant in any comparison with the FDR adjusted p -value < 0.05. For the relative allelic expression of a gene, the significance in P comparison was considered evidence of parental expression divergence. The expression difference in F1 hybrid (significant in H comparison) was considered evidence of *cis*-regulatory divergence. The parental expression divergence was considered due to *trans*-regulatory changes if the allelic expression was not different in the F1 hybrid (no significance in H comparison) and the ratios of allelic expression were different between the parental mix (*in silico* hybrid) and F1 hybrid (significant in T comparison). The regulatory divergence types were further classified into seven types using the following criteria: *cis*-only: significant in comparison P and H but not significant in T. *trans*-only: significant in comparison P and T, but not H. *cis* + *trans*: significant in comparison P, H, and T, moreover, the log2-transformed allelic expression ratio has the same sign in F1 and *in silico* hybrid. In the *cis* + *trans* type, the *cis*- and *trans*-regulatory divergences favor expression of the same allele. *cis***trans*: significant in comparison P, H, and T, besides, the log2-transformed allelic expression ratio has the opposite

sign in F1 and *in silico* hybrid. In the *cis*trans* type, the *cis*- and *trans*-regulatory divergences favor expression of the opposite alleles. Compensatory: significant in comparison H and T, but not in P. In the compensatory type, the *cis*- and *trans*-regulatory divergences compensate each other. Conserved: no significance in any of the three comparisons. Ambiguous: all other patterns. Genes were classified into five duplication modes using the method described in Xu et al., 2018. The protein sequences of all genes were aligned to each other using blastp program, then the gene duplication modes were assigned using MCScanX (Wang et al., 2012). Genes in singleton mode had no hits in the all-to-all blastp search. Genes in dispersed mode are dispersed paralogs interrupted by many genes on the same chromosome or non-collinear on different chromosome. Genes in proximal mode are paralogs interrupted by fewer than 20 genes. Genes in tandem mode are clusters of consecutive tandem duplicates. Genes in WGD mode are paralogs in collinear chromosome regions. The WGD genes were further classified into old and young ages based on the distribution of K_s values (Xu et al., 2018). Briefly, the K_s values between WGD duplicates were calculated using “add_ka_and_ks_to_collinearity.pl” in MCScanX, and the average K_s value for each collinear block was calculated. The collinear blocks were then clustered into three groups using a *k*-means method ($k = 3$) in R; then genes were classified as young duplicates if present only in the cluster with least mean K_s value. The other genes were classified as old duplicates because they were found in at least one old cluster.

GO and Pfam Enrichment Analysis

DEGs between genotypes and genes assigned into different regulatory divergence types were used for functional enrichment analysis. gene ontology (GO) or protein family (Pfam) terms containing less than five expressed genes were removed from further analysis. A one-tail hypergeometric test was used to test whether a GO or Pfam term was over-represented in DEGs or genes of different regulatory types. The raw *p*-values were adjusted using the FDR method, and only terms whose adjusted *p*-value less than 0.05 were classified as significantly over-represented.

RESULTS

Gene Expression Changes in Cultivated and Wild Soybean, and Their Hybrid F1

The cultivated soybean JY47 (*G. max*) and the wild soybean GS (*G. soja*) are dramatically different in morphology, while F1 hybrids between them show intermediate phenotype for many traits, such as plant height and leaf size (Supplementary Figure 1). RNA-seq reads were mapped to the reference genome of cultivated soybean cv. Williams 82 (version a2v1) and the gene expression values were calculated and compared between genotypes. Consistent with morphological differences, 12,677 genes were identified as DEGs between JY47 and GS, which accounted for 43.40% of all expressed genes (29,235) in the leaf tissue. There were nearly equal amounts of up-regulated genes in JY47 (6,321 genes) and GS (6,356 genes) compared with

each other. Three mixtures using equal amounts of maternal and paternal data from three pairs of parental individuals were constructed and served as *in silico* “hybrids.” The gene expression values detected in the *in silico* “hybrids” represent additive mid-parental expression levels. In the comparison between F1 hybrids and *in silico* “hybrids,” 493 genes were found to be differentially expressed (non-additive). Interestingly, the down-regulated genes (353 genes) in F1 hybrid were twofold more than the up-regulated genes (140 genes) as compared to *in silico* “hybrids” indicating complicated regulatory interactions in the F1 hybrids. When compared to the two parental genotypes, the F1 hybrids showed more DEGs with GS (5,210) than with JY47 (1,008) (Table 1), indicating the dominant role of regulatory alleles from cultivated soybean. To examine whether the observed parental expression level dominance is due to mapping preference of reads from JY47 to the cultivated reference genome, we performed the same DEGs analysis using a wild soybean reference genome and found the same trend (Supplementary Table 1).

Regulatory Divergence Between the Wild and Cultivated Soybean Genotypes

To further address the evolution of expression divergence between the wild and cultivated soybean genotypes, we classified the genes into seven regulatory divergence types based on their allelic expression patterns in the *in silico* “hybrids” and F1 hybrids. In total, 7,132 genes were interrogated, the majority of which were found to be conserved (3,333 genes) or ambiguous (1,432 genes) (Figure 1); 533 genes were diverged in a *cis*-only pattern, while 1,265 were in *trans*-only pattern suggesting *trans*-regulatory changes play a predominant role in the expression divergence between the wild and cultivated soybean genotypes (Figure 1). A relatively lower fraction of genes were found in the other three more complex types (233 in *cis* + *trans*, 145 in *cis*trans*, and 191 in compensatory patterns) (Figure 1).

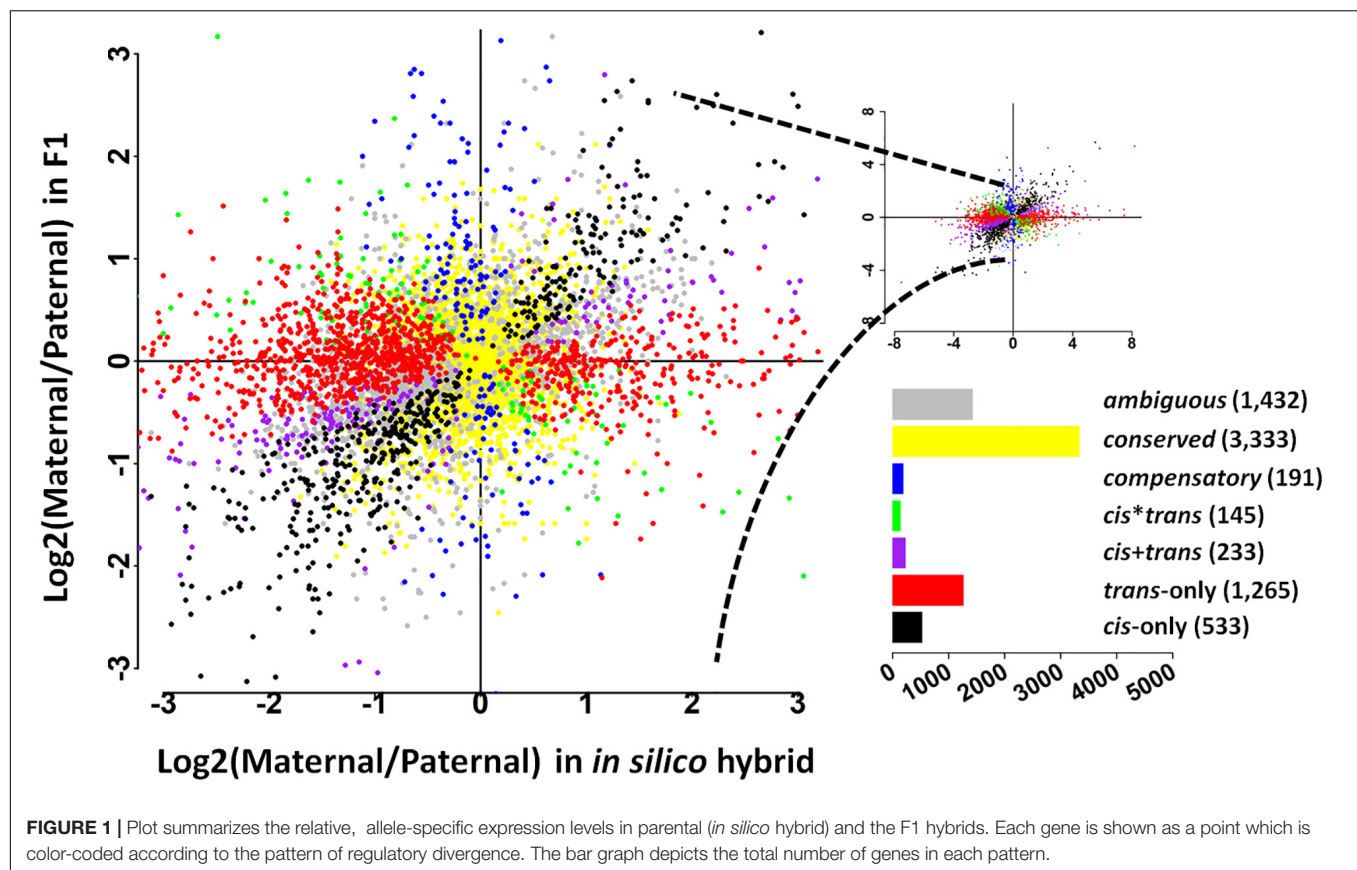
The Relationship Between Duplication Mode and Regulatory Divergence

To address the relationship between gene duplication modes and type of regulatory divergence, we classified all chromosomal genes into five different categories based on their duplicate states in the reference genome as singleton, dispersed, proximal,

TABLE 1 | Summary of differentially expressed genes in each comparison between genotypes.

Comparisons	DEGs	Up-regulated ^a	Down-regulated ^b
GS vs. JY47	12,677 (43.4%)	6,356 (21.8%)	6,321 (21.6%)
GS vs. F1	10,048 (34.4%)	4,838 (16.6%)	5,210 (17.8%)
JY47 vs. F1	1,753 (6.0%)	745 (2.5%)	1,008 (3.5%)
F1 vs. Mix*	493 (1.7%)	140 (0.5%)	353 (1.2%)

^aNumber and fraction of DEGs up-regulated in the former genotype. ^bNumber and fraction of DEGs down-regulated in the former genotype. *Mix was constructed by equal amounts of maternal and paternal data and served *in silico* “hybrids.”



tandem, and WGD/large segmental duplication (WGD for short). We calculated the distribution of genes in different regulatory divergence types for each duplication mode. In singletons, we found the same number of genes diverged in *cis*-only and *trans*-only patterns (11% *cis*-only/*trans*-only) but significantly more genes in *trans*-only than *cis*-only divergence type in the other three duplication modes (chi-squared test p -value < 0.01) (Table 2). The difference in the proportion of genes subject to *cis*-only and *trans*-only patterns was the highest in WGD genes where there were two times

more genes in *trans*-only pattern (1,110 genes) than *cis*-only pattern (428 genes) (Table 2). All duplication modes except for WGD mode showed higher proportions *cis*-only genes as compared to genome-wide levels, and differences were statistically significant for singleton and proximal modes (chi-squared test p -value < 0.05). Furthermore, the conserved regulatory type accounted for 47.23% of WGD genes, which was the highest, while similar proportions of conserved genes were found in singleton (46.96%) and dispersed (46.56%) genes, whereas proportions in proximal (34.31%) and tandem

TABLE 2 | Number and proportion of genes in different regulatory patterns for each duplicate mode.

	<i>Cis</i> -only	<i>Trans</i> -only	<i>Cis</i> + <i>trans</i>	<i>Cis</i> * <i>trans</i>	Compensatory	Conserved	Ambiguous
Singleton	22 (12.15%)	22 (12.15%)	11 (6.08%)	2 (1.10%)	9 (4.97%)	85 (46.96%)	30 (16.57%)
Dispersed	45 (9.11%)	76 (15.38%)	14 (2.83%)	7 (1.42%)	17 (3.44%)	230 (46.56%)	105 (21.26%)
Proximal	15 (14.71%)	18 (17.65%)	11 (10.78%)	2 (1.96%)	4 (3.92%)	35 (34.31%)	17 (16.67%)
Tandem	22 (11.06%)	44 (22.11%)	11 (5.53%)	6 (3.02%)	6 (3.02%)	74 (37.19%)	36 (18.09%)
WGD	428 (6.98%)	1100 (17.93%)	185 (3.01%)	128 (2.09%)	155 (2.53%)	2,898 (47.23%)	1,242 (20.24%)
Total	532 (7.48%)	1,260 (17.72%)	232 (3.26%)	145 (2.04%)	191 (2.69%)	3,322 (46.71%)	1,430 (20.11%)

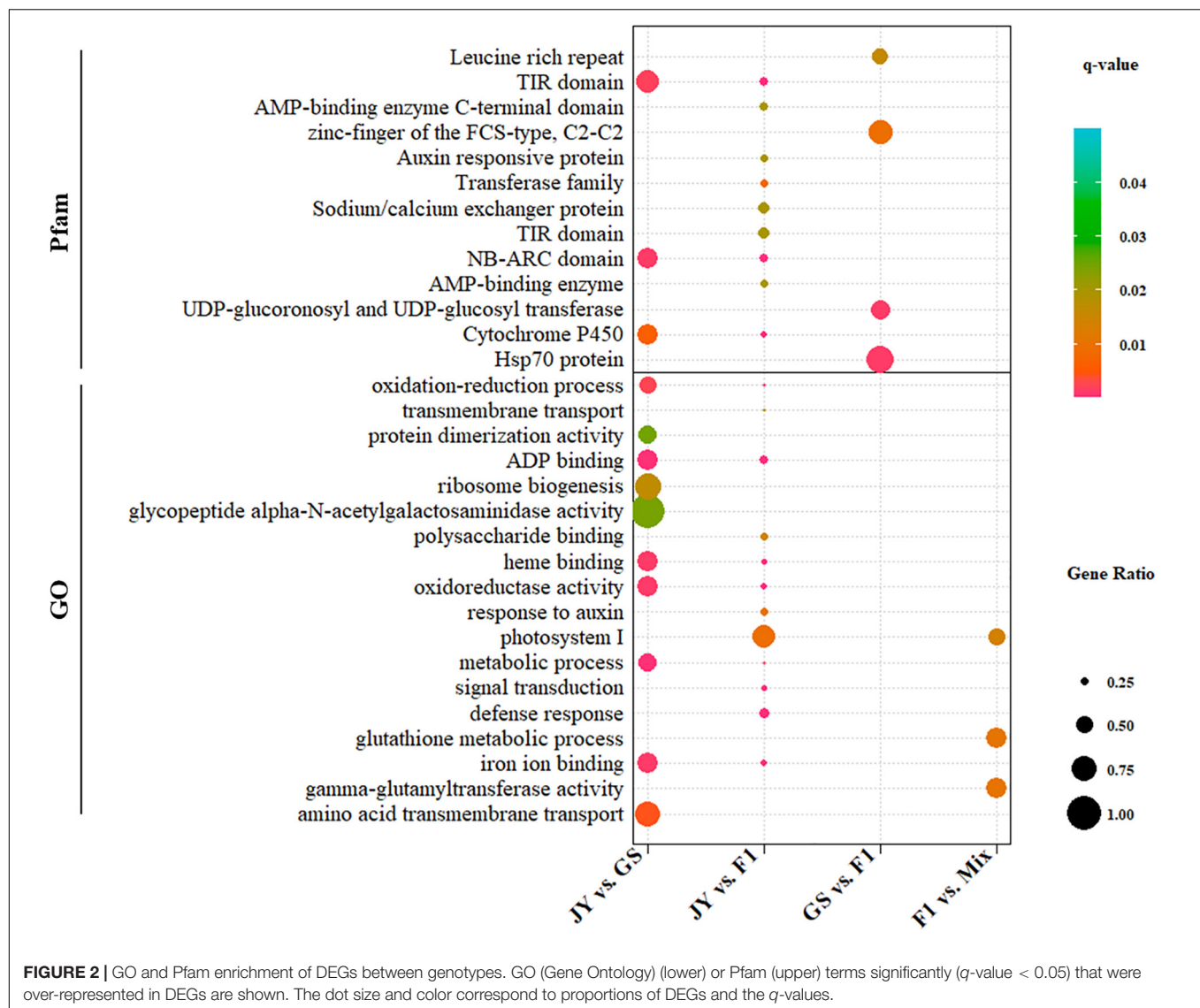


FIGURE 2 | GO and Pfam enrichment of DEGs between genotypes. GO (Gene Ontology) (lower) or Pfam (upper) terms significantly (q -value < 0.05) that were over-represented in DEGs are shown. The dot size and color correspond to proportions of DEGs and the q -values.

(37.19%) mode were significantly lower than the genome-wide level (Table 2).

Functional Enrichment of DEGs and Genes in Different Regulatory Divergence Types

We performed GO and Pfam enrichment analysis for DEGs and genes in different regulatory divergence types. DEGs between cultivated and wild soybean showed significant enrichment in ADP-binding (GO:0043531), oxidation-reduction related terms (GO:0016705, GO:0055114) and other GO terms, totaling to five GO terms (q -value < 0.05) (Figure 2 and Supplementary Material). Three GO terms, gamma-glutamyltransferase activity, glutathione metabolic process, and photosystem I, were over-represented in non-additive genes (Figure 2 and Supplementary Material). Three protein domains, NB-ARC domain, TIR domain, and cytochrome P450 were over-represented in DEGs

between the two parental genotypes; however, no protein domain was over-represented in the non-additive genes (Figure 2 and Supplementary Material).

For genes in different regulatory divergence types, ADP binding GO term was over-represented in genes of *cis*-only type, whereas no GO terms were over-represented in other regulatory types except protein binding (GO:0005515) in conserved pattern (Supplementary Material). Different protein families (Pfam domain) were over-represented in genes of *cis*-only and *trans*-only patterns. NB-ARC domain and TIR domain were over-represented in genes of *cis*-only pattern, while the response regulator receiver domain and Hsp70 protein domain were over-represented in *trans*-only pattern genes (Supplementary Material). No over-represented domain was found in the rest of the regulatory divergence patterns. The top over-represented GO term (ADP binding) and Pfam terms (NB-ARC domain and TIR domain) were the same in DEGs between GS and JY and genes of *cis*-only divergence.

DISCUSSION

In line with the differences in plant morphology, a large number of genes were found to be differentially expressed between JY47 and GS, indicating that domestication and subsequent evolution/improvement have dramatically shaped the transcriptomes of *G. max* and *G. soja*. Commonality of the top over-represented GO and Pfam terms in DEGs and genes subject to *cis*-only regulatory divergence between JY47 and GS (Supplementary Material) suggests the substantial contribution of *cis*-regulatory divergence in the evolution and diversification of wild and cultivated soybeans. A few studies have shown that gene expression changes played a role in the domestication and improvement of soybean (Dai et al., 2018; Zhang et al., 2019; Miao et al., 2020). One example is the *GmCYP78A* gene family of which there are three members: two, *GmCYP78A70* (Glyma.01G061100) and *GmCYP78A57* (Glyma.02G119600), were derived from a single ancestor during the latest WGD ~13 Mya, and the third copy *GmCYP78A72* (Glyma.19G240800) was duplicated from *GmCYP78A57* (Dai et al., 2018). These genes show expression divergence among tissues and positive correlation with leaf size and seed weight in different cultivars; furthermore, population genetic results indicate two underwent intense selection during soybean domestication and/or improvement (Dai et al., 2018). In our study, *GmCYP78A70* and *GmCYP78A57* showed detectable expression in leaf (Supplementary Figure 2) and the expression *GmCYP78A70* in cultivated soybean was statistically higher than in wild soybean consistent with the previous study. Genome-wide expression levels in the hybrid were biased toward the cultivated soybean JY47, indicating parental expression level dominance (Table 1 and Supplementary Table 1). A similar phenomenon has been found in cotton (Flagel et al., 2008) which was shown to result from the up- or down-regulation of gene copy (allele/homeolog) from the non-dominant parent (Yoo et al., 2013). Expression level dominance can be caused by *trans*-regulatory interactions, which accords with our findings of a large proportion of genes subject to *trans*-regulatory divergence.

The relative contribution of *cis*- and *trans*-regulatory divergence in evolution has been broadly studied. *Trans*-regulatory divergence has been found to play a dominant role in the regulatory divergence within species, while *cis*-regulatory divergence makes either similar or greater contribution to gene expression divergence between species (Zhuang and Adams, 2007; Wittkopp et al., 2008; Emerson et al., 2010; Goncalves et al., 2012; Lemmon et al., 2014; Xu et al., 2014; Guerrero et al., 2016; Wu et al., 2016), which fits the prediction of different types of selection acting on the two types of regulatory divergences (Prud'homme et al., 2007; Emerson et al., 2010; Coolon et al., 2014). Besides the complicated divergence and domestication history, *G. max* and *G. soja* have a highly duplicated genome due to the two recent WGDs in their common ancestor occurred about 13 MYA and 59 MYA (Schmutz et al., 2010). The gene balance hypothesis predicts that all gene duplicates are not equally retained following a WGD (Edger and Pires, 2009; Birchler, 2019); therefore, genes resulting from different duplication modes in the soybean genome

have experienced different selection constraints. Prior studies have shown there are abundant genes in different duplication modes, but >60% of genes remain collinear in the soybean genome (Xu et al., 2018). The WGD genes in soybean were found to be enriched in transcription factors and transcription regulation functions, which is in line with the gene balance hypothesis. In soybean, different duplication modes are distinct from each other in DNA methylation and expression profiles as well as enriched functional categories (Xu et al., 2018), suggesting varied constraints on the evolution of genes in different modes.

In this study, we revealed the effects of duplication mode on the evolution of regulatory divergence between wild and cultivated soybean. We found that genes from different duplication modes tended to accumulate different types of regulatory divergence. Relative higher proportions of *cis*-only regulatory divergence were detected in singleton, dispersed, proximal, and tandem modes than in genes from a WGD and genome-wide levels (Table 2), consistent with the prediction of gene balance hypothesis that genes in these duplication modes are less involved in regulatory networks (Edger and Pires, 2009). However, at genome-wide scale, *trans*-regulatory changes were found to play a predominant role in the expression divergence between *G. soja* and *G. max* (Figure 1). We found that as majority constituents to the soybean genome, WGD genes are more likely to be affected by *trans*-regulatory changes than by *cis*-regulatory changes, leading to the observed more *trans*-regulatory changes in genome-wide scale (Table 2). Some WGD duplicates may have conserved regulatory regions following whole genome duplications. These paralogs can be regulated by the same *trans*-acting factors which can lead to amplified effects of *trans*-regulatory changes in these genes. Furthermore, the retained WGD genes are more likely involved in regulatory network according to gene balance hypothesis. Transcription factors are usually dosage-sensitive and preferentially retained following WGDs due to dosage constraint, which has also been supported in a previous study in soybean (Xu et al., 2018). In this study, we observed a high proportion of WGD genes in *trans*-only regulatory type. Genes affected by *trans*-regulatory divergence were more likely to be the targets of transcription factors. Here, our results suggest that not only the transcription factors but also many of their targets have been retained in the collinear blocks in the soybean genome which have experienced transcriptional divergence. However, it is still not clear how the diverged *trans*-acting factors are released from purifying selections and gene balance constraints. The proportion of conserved genes was highest in WGD mode suggests they are under stronger purifying selection than genes in other duplication modes. The expression coordinates of retained WGD paralogs were decreased and transcriptional divergence increased over time in soybean (Xu et al., 2018). Expression divergence indicating subfunctionalization and/or neofunctionalization contributes to the maintenance of most duplicated regulatory genes in *Arabidopsis* after each round of duplication (Duarte et al., 2006). A recent study in *Paramecium* and yeast revealed that WGD genes were retained due to dosage

constraint followed by divergence in expression level and eventual deterministic gene loss through dosage subfunctionalization (Gout and Lynch, 2015). Our results revealed the divergence of regulatory network during post-WGD evolution, which is consistent with findings in yeast demonstrating rapid divergence and increase in complexity of networks after polyploidization (Teichmann and Babu, 2004; Gu et al., 2005). Thus, gene/genome duplication plays a key role in network evolution. Together, it is clear that genes in different duplication modes which are under and/or resulting from selection pressures have differential effects on transcriptional evolution due to *cis*- and *trans*-regulatory divergence and that retained WGD genes are prone to *trans*-regulatory divergence. We further classified WGD genes into young and old WGD duplicates based on their Ks values. Most WGD genes (32,993) were young duplicates. A higher proportion of *cis*-only genes (7.11%) but lower proportion of *trans*-only genes (17.59%) were found in young WGD duplicates than in old duplicates (*cis*-only: 5.72%, *trans*-only: 21.14%) (Supplementary Table 2). The proportions of *cis*-only genes in both young and old WGD duplicates were lower than in the other duplicate modes. This is probably due to the large amount of young WGD genes, some of which were less likely to be subject to gene balance constraints and more susceptible to *cis*-regulatory changes than old WGD genes.

Tandem duplicates have the lowest proportion of genes in conserved patterns, suggesting higher divergence rates in these genes. We have shown previously that tandem duplicate genes in the soybean genome are enriched for stress related functions (Xu et al., 2018). Also, there is no evidence implicating that this type of duplicated genes are subject to gene balance constraint (Edger and Pires, 2009). Interestingly, we found the highest proportion of genes due to *trans*-only divergences in the tandem duplicate mode. A recent study also showed that subfunctionalization of expression evolves slowly in tandem duplicates possibly because they are coregulated by shared genomic elements (Lan and Pritchard, 2016). We suggest that coregulation, together with preference of some *trans*-acting factors for tandem duplicates, may have given rise to the observed high *trans*-regulatory divergence in this type of duplicates.

REFERENCES

- Birchler, J. A. (2019). Genomic balance plays out in evolution. *Plant Cell* 31, 1186–1187. doi: 10.1105/tpc.19.00329
- Birchler, J. A., and Veitia, R. A. (2007). The gene balance hypothesis: from classical genetics to modern genomics. *Plant Cell* 19, 395–402. doi: 10.1105/tpc.106.049338
- Birchler, J. A., and Veitia, R. A. (2010). The gene balance hypothesis: implications for gene regulation, quantitative traits and evolution. *New Phytol.* 186, 54–62. doi: 10.1111/j.1469-8137.2009.03087.x
- Blanc, G., and Wolfe, K. H. (2004). Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes. *Plant Cell* 16, 1667–1678. doi: 10.1105/tpc.021345
- Bolger, A. M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30, 2114–2120. doi: 10.1093/bioinformatics/btu170

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: <https://www.ncbi.nlm.nih.gov/>, PRJNA660310 and <https://www.ncbi.nlm.nih.gov/>, PRJNA660313.

AUTHOR CONTRIBUTIONS

CX designed the research. NZ, XD, TL, MW, YT, DL, QA, and SS performed the experiments and analyzed the data. NZ and CX wrote the manuscript. SJ and BL revised the manuscript. All authors read and approved the final manuscript.

FUNDING

This work was supported by the China National Novel Transgenic Organisms Breeding Project (2016ZX08004-004); the National Natural Science Foundation of Jilin, China, 20200201032JC; the United States National Science Foundation (1539838); and the Fundamental Research Funds for the Central Universities.

ACKNOWLEDGMENTS

The authors appreciate the support from the China National Novel Transgenic Organisms Breeding Project (2016ZX08004-004); the National Natural Science Foundation of Jilin, China, 20200201032JC, the United States National Science Foundation (1539838), and the Fundamental Research Funds for the Central Universities.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fgene.2020.601003/full#supplementary-material>

- Chapman, B. A., Bowers, J. E., Feltus, F. A., and Paterson, A. H. (2006). Buffering of crucial functions by paleologous duplicated genes may contribute cyclicity to angiosperm genome duplication. *Proc. Natl. Acad. Sci. U.S.A.* 103, 2730–2735. doi: 10.1073/pnas.0507782103
- Chen, H. T., Zeng, Y., Yang, Y. Z., Huang, L. L., Tang, B. L., Zhang, H., et al. (2020). Allele-aware chromosome-level genome assembly and efficient transgene-free genome editing for the autotetraploid cultivated alfalfa. *Nat. Commun.* 11:2494.
- Coolon, J. D., Mcmanus, C. J., Stevenson, K. R., Graveley, B. R., and Wittkopp, P. J. (2014). Tempo and mode of regulatory evolution in *Drosophila*. *Genome Res.* 24, 797–808. doi: 10.1101/gr.163014.113
- Dai, A. H., Yang, S. X., Zhou, H. K., Tang, K. Q., Li, G., Leng, J. T., et al. (2018). Evolution and expression divergence of the CYP78A subfamily genes in soybean. *Genes* 9:611. doi: 10.3390/genes9120611
- Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., et al. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 29, 15–21. doi: 10.1093/bioinformatics/bts635

- Duarte, J. M., Cui, L., Wall, P. K., Zhang, Q., Zhang, X., Leebens-Mack, J., et al. (2006). Expression pattern shifts following duplication indicative of subfunctionalization and neofunctionalization in regulatory genes of *Arabidopsis*. *Mol. Biol. Evol.* 23, 469–478. doi: 10.1093/molbev/msj051
- Edger, P. P., and Pires, J. C. (2009). Gene and genome duplications: the impact of dosage-sensitivity on the fate of nuclear genes. *Chromosome Res.* 17, 699–717. doi: 10.1007/s10577-009-9055-9
- Emerson, J. J., Hsieh, L. C., Sung, H. M., Wang, T. Y., Huang, C. J., Lu, H. H., et al. (2010). Natural selection on cis and trans regulation in yeasts. *Genome Res.* 20, 826–836. doi: 10.1101/gr.101576.109
- Flagel, L., Udall, J., Nettleton, D., and Wendel, J. (2008). Duplicate gene expression in allopolyploid *Gossypium* reveals two temporally distinct phases of expression evolution. *BMC Biol.* 6:16. doi: 10.1186/1741-7007-6-16
- Goncalves, A., Leigh-Brown, S., Thybert, D., Stefflova, K., Turro, E., Flicek, P., et al. (2012). Extensive compensatory cis-trans regulation in the evolution of mouse gene expression. *Genome Res.* 22, 2376–2384. doi: 10.1101/gr.142281.112
- Gong, Z. Z. (2020). Flowering phenology as a core domestication trait in soybean. *J. Integr. Plant Biol.* 62, 546–549. doi: 10.1111/jipb.12934
- Gout, J. F., and Lynch, M. (2015). Maintenance and loss of duplicated genes by dosage subfunctionalization. *Mol. Biol. Evol.* 32, 2141–2148. doi: 10.1093/molbev/msv095
- Gu, X., Zhang, Z., and Huang, W. (2005). Rapid evolution of expression and regulatory divergences after yeast gene duplication. *Proc. Natl. Acad. Sci. U.S.A.* 102, 707–712. doi: 10.1073/pnas.0409186102
- Guerrero, R. F., Posto, A. L., Moyle, L. C., and Hahn, M. W. (2016). Genome-wide patterns of regulatory divergence revealed by introgression lines. *Evolution* 70, 696–706. doi: 10.1111/evo.12875
- He, X., and Zhang, J. (2005). Rapid subfunctionalization accompanied by prolonged and substantial neofunctionalization in duplicate gene evolution. *Genetics* 169, 1157–1164. doi: 10.1534/genetics.104.037051
- Innan, H., and Kondrashov, F. (2010). The evolution of gene duplications: classifying and distinguishing between models. *Nat. Rev. Genet.* 11, 97–108. doi: 10.1038/nrg2689
- Jiao, Y., Wickett, N. J., Ayyampalayam, S., Chanderbali, A. S., Landherr, L., Ralph, P. E., et al. (2011). Ancestral polyploidy in seed plants and angiosperms. *Nature* 473, 97–100. doi: 10.1038/nature09916
- Kim, M. Y., Lee, S., Van, K., Kim, T. H., Jeong, S. C., Choi, I. Y., et al. (2010). Whole-genome sequencing and intensive analysis of the undomesticated soybean (*Glycine soja* Sieb. and Zucc.) genome. *Proc. Natl. Acad. Sci. U.S.A.* 107, 22032–22037.
- Kim, M. Y., Van, K., Kang, Y. J., Kim, K. H., and Lee, S. H. (2012). Tracing soybean domestication history: from nucleotide to genome. *Breed Sci.* 61, 445–452. doi: 10.1270/jsbbs.61.445
- Lan, X., and Pritchard, J. K. (2016). Coregulation of tandem duplicate genes slows evolution of subfunctionalization in mammals. *Science* 352, 1009–1013. doi: 10.1126/science.1248411
- Lemmon, Z. H., Bukowski, R., Sun, Q., and Doebley, J. F. (2014). The role of cis regulatory evolution in maize domestication. *PLoS Genet.* 10:e1004745. doi: 10.1371/journal.pgen.1004745
- Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25, 1754–1760. doi: 10.1093/bioinformatics/btp324
- Li, Y. H., Zhou, G., Ma, J., Jiang, W., Jin, L. G., Zhang, Z., et al. (2014). De novo assembly of soybean wild relatives for pan-genome analysis of diversity and agronomic traits. *Nat. Biotechnol.* 32, 1045–1052.
- Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* 15:550.
- Maere, S., De Bodt, S., Raes, J., Casneuf, T., Van Montagu, M., Kuiper, M., et al. (2005). Modeling gene and genome duplications in eukaryotes. *Proc. Natl. Acad. Sci. U.S.A.* 102, 5454–5459. doi: 10.1073/pnas.0501102102
- McManus, C. J., Coolon, J. D., Duff, M. O., Eipper-Mains, J., Graveley, B. R., and Wittkopp, P. J. (2010). Regulatory divergence in *Drosophila* revealed by mRNA-seq. *Genome Res.* 20, 816–825. doi: 10.1101/gr.102491.109
- Miao, L., Yang, S. N., Zhang, K., He, J. B., Wu, C. H., Ren, Y. H., et al. (2020). Natural variation and selection in GmSWEET39 affect soybean seed oil content. *New Phytol.* 225, 1651–1666. doi: 10.1111/nph.16250
- Prud'homme, B., Gompel, N., and Carroll, S. B. (2007). Emerging principles of regulatory evolution. *Proc. Natl. Acad. Sci. U.S.A.* 104(Suppl. 1), 8605–8612. doi: 10.1073/pnas.0700488104
- Schmutz, J., Cannon, S. B., Schlueter, J., Ma, J., Mitros, T., Nelson, W., et al. (2010). Genome sequence of the palaeopolyploid soybean. *Nature* 463, 178–183.
- Teichmann, S. A., and Babu, M. M. (2004). Gene regulatory network growth by duplication. *Nat. Genet.* 36, 492–496. doi: 10.1038/ng1340
- Wang, X. T., Chen, L. Y., and Ma, J. X. (2019). Genomic introgression through interspecific hybridization counteracts genetic bottleneck during soybean domestication. *Genome Biol.* 20:22.
- Wang, Y., Tang, H., DeBarry, J. D., Tan, X., Li, J., Wang, X., et al. (2012). MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res.* 40:e49. doi: 10.1093/nar/gkr1293
- Wei, R. X., and Ge, S. (2011). Evolutionary history and complementary selective relaxation of the duplicated PI genes in grasses. *J. Integr. Plant Biol.* 53, 682–693. doi: 10.1111/j.1744-7909.2011.01058.x
- Wittkopp, P. J., Haerum, B. K., and Clark, A. G. (2004). Evolutionary changes in cis and trans gene regulation. *Nature* 430, 85–88. doi: 10.1038/nature02698
- Wittkopp, P. J., Haerum, B. K., and Clark, A. G. (2008). Regulatory changes underlying expression differences within and between *Drosophila* species. *Nat. Genet.* 40, 346–350. doi: 10.1038/ng.77
- Wu, Y., Sun, Y., Wang, X., Lin, X., Sun, S., Shen, K., et al. (2016). Transcriptome shock in an interspecific F1 triploid hybrid of *Oryza* revealed by RNA sequencing. *J. Integr. Plant Biol.* 58, 150–164. doi: 10.1111/jipb.12357
- Xu, C., Bai, Y., Lin, X., Zhao, N., Hu, L., Gong, Z., et al. (2014). Genome-wide disruption of gene expression in allopolyploids but not hybrids of rice subspecies. *Mol. Biol. Evol.* 31, 1066–1076. doi: 10.1093/molbev/msu085
- Xu, C., Nadon, B. D., Kim, K. D., and Jackson, S. A. (2018). Genetic and epigenetic divergence of duplicate genes in two legume species. *Plant Cell Environ.* 41, 2033–2044.
- Yoo, M. J., Szadkowski, E., and Wendel, J. F. (2013). Homoeolog expression bias and expression level dominance in allopolyploid cotton. *Heredity* 110, 171–180. doi: 10.1038/hdy.2012.94
- Zhang, D., Zhang, H. Y., Hu, Z. B., Chu, S. S., Yu, K. Y., Lv, L. L., et al. (2019). Artificial selection on GmOLEO1 contributes to the increase in seed oil during soybean domestication. *Plos Genet.* 15:e1008267. doi: 10.1371/journal.pgen.1008267
- Zhuang, Y., and Adams, K. L. (2007). Extensive allelic variation in gene expression in populus F1 hybrids. *Genetics* 177, 1987–1996. doi: 10.1534/genetics.107.080325

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2020 Zhao, Ding, Lian, Wang, Tong, Liang, An, Sun, Jackson, Liu and Xu. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



Epigenetics as Driver of Adaptation and Diversification in Microbial Eukaryotes

Agnes K. M. Weiner^{1*} and Laura A. Katz^{1,2*}

¹ Department of Biological Sciences, Smith College, Northampton, MA, United States, ² Program in Organismic and Evolutionary Biology, University of Massachusetts Amherst, Amherst, MA, United States

Keywords: epigenetics, adaptation, speciation, chromatin modification, non-protein-coding RNA, protists

OPEN ACCESS

Edited by:

Ekaterina Shelest,
German Centre for Integrative
Biodiversity Research (iDiv), Germany

Reviewed by:

Rosa María Bermudez-Cruz,
Instituto Politécnico Nacional de
México (CINVESTAV), Mexico
Douglas Chalker,
Washington University in St. Louis,
United States

*Correspondence:

Agnes K. M. Weiner
aweiner@smith.edu
orcid.org/0000-0002-9917-5235
Laura A. Katz
lkatz@smith.edu
orcid.org/0000-0002-9138-4702

Specialty section:

This article was submitted to
Evolutionary and Population Genetics,
a section of the journal
Frontiers in Genetics

Received: 15 December 2020

Accepted: 15 February 2021

Published: 16 March 2021

Citation:

Weiner AKM and Katz LA (2021)
Epigenetics as Driver of Adaptation
and Diversification in Microbial
Eukaryotes. *Front. Genet.* 12:642220.
doi: 10.3389/fgene.2021.642220

INTRODUCTION

Microbial eukaryotes, i.e., protists, represent the bulk of eukaryotic diversity in terms of species diversity and biomass. Protists are globally distributed in all ecosystems and play important roles in food webs and nutrient cycles. To date it remains enigmatic how protist diversity is generated, especially in lineages with large populations in ecosystems without apparent dispersal barriers (i.e., many marine species, species that encyst). We argue that epigenetic processes, such as chromatin modification and/or regulation by small non-protein-coding RNAs (npc-RNAs) that rapidly modify genomes and gene expression states, play important roles in driving phenotypic plasticity, differential adaptation and ultimately diversification of protists. Our argument is based on two recent developments in epigenetic research: (1) it is now clear that epigenetic processes were present in the last eukaryotic common ancestor (LECA) and are widespread across eukaryotes, and (2) numerous studies have demonstrated that at least some epigenetic marks can be inherited across generations. Given this, we suggest to combine morphometrics, genomics, and epigenomics for research on adaptability and diversification in microbial eukaryotes.

DIVERSITY OF MICROBIAL EUKARYOTES

Many lineages of protists have a tremendous species diversity, which is reflected in a wide variety of morphologies and ecological functions (e.g., Adl et al., 2019). In addition, research of the last two decades has unearthed a large amount of cryptic diversity, suggesting a decoupling of morphological and molecular evolution (e.g., Katz et al., 2005; Šlapeta et al., 2005; Darling and Wade, 2008; Oliverio et al., 2014). Protists occur globally in all ecosystems, and while some species are endemic to certain areas, others have a cosmopolitan distribution and vast population sizes (e.g., Ryšánek et al., 2015; Faure et al., 2019). Large-scale barcoding studies revealed that some closely related cryptic species are able to co-occur in close biogeographical proximity (e.g., Amato et al., 2007; Weiner et al., 2014; Badger et al., 2017; Tucker et al., 2017). In addition, protists show a variety of complex life cycles, sometimes alternating sexual and asexual generations (e.g., Grell, 1973; Parfrey et al., 2008). Given these characteristics, the enormous species diversity is perhaps not surprising. However, which (molecular/epigenetic) mechanisms allow for speciation in microbes, especially in habitats with seemingly unlimited dispersal potential, remains unresolved. Several groups have hypothesized that differential adaptation to environmental factors may be the underlying driver for diversification in sympatry (e.g., Ryšánek et al., 2016; Irwin et al., 2017; Škaloud et al., 2019). However, for gene flow between populations to be overcome, mechanisms leading to the establishment of reproductive isolation would have to be fast and efficient. We argue that in order for rapid diversification to be achieved, epigenetic processes

that regulate gene activity and that may be influenced by the environment play important roles in establishing phenotypic plasticity; if the epigenetic marks are inherited across generations—what we refer to as “epigenetic assimilation”—they can provide a fitness advantage to members of the population and ultimately lead to differential adaptation that drives speciation (Figure 1).

EPIGENETICS IN MICROBIAL EUKARYOTES

A prerequisite for our model of epigenetics as driver of ecological speciation in protists (Figure 1) to be valid is the widespread existence of epigenetic phenomena in microbial eukaryotes. Eukaryotic epigenetics comprises processes such as chromatin and DNA modifications and regulation by npc-RNAs (e.g., Razin and Riggs, 1980; Ng and Bird, 1999; Shabalina and Koonin, 2008), that are thought to have evolved originally for mediating genome conflict between mobile genetic elements and host genomes (Lisch, 2009; Fedoroff, 2012). The effects of epigenetics include, among others, gene activation or silencing, and altering genome structures through DNA elimination or polyploidization (e.g., Liu and Wendel, 2003; Bernstein and Allis, 2005). Most epigenetic research has focused on animals and plants, yet it was recently confirmed that the basic epigenetic gene toolkit was present in LECA and is now widespread throughout the eukaryotic tree of life (Aravind et al., 2014; Weiner et al., 2020). This highlights the importance of epigenetics for the functioning of eukaryotic genomes. For the majority of protists, however, knowledge on their epigenetics remains limited, mostly because many are uncultivable and annotated reference genomes are lacking. What is known so far mostly stems from research on model organisms, such as ciliates (Alveolata) and human pathogens [e.g., *Plasmodium* (Alveolata) and *Trypanosoma* (Excavata)].

In ciliates, which contain both a germline and somatic nucleus within one cell, epigenetics plays key roles in distinctions between the two genomes and in elimination of DNA during the development of a new somatic nucleus during reproduction (e.g., Chicoine and Allis, 1986; Jahn and Klobutcher, 2002; Chalker et al., 2013; Pilling et al., 2017). Small npc-RNAs, such as “scan RNAs” and “macronuclear RNAs,” bind to homologous regions in the genome or direct histone modifications (e.g., H3K9 methylation) in those regions to mark them for either retention or elimination (Chen et al., 2014; Swart et al., 2014). Similarly, npc-RNAs, so-called “template RNAs,” were found to be involved in the reordering of scrambled genes in some ciliates (e.g., Garnier et al., 2004; Nowacki et al., 2011). Another phenomenon of genome dynamics that is likely driven by epigenetics is the determination of ploidy levels throughout the life cycle. Many protist lineages, such as some Foraminifera (Rhizaria), ciliates and Amoebozoa have been observed to undergo significant changes in ploidy, sometimes containing thousands of copies of the genome that later are eliminated again (Parfrey et al., 2008; Bellec and Katz, 2012; Goodkov et al., 2020). In the case of ciliates, research suggested that RNA interference, which is part

of the “epigenetic toolkit,” is driving these changes (Heyse et al., 2010).

In addition to these large-scale modifications to the genome architecture, epigenetic processes are involved in changes to the morphology or physiology of protists. This is especially prevalent in parasites, in which epigenetics controls virulence and cell differentiation through regulation of gene expression and thus plays an important role in host-pathogen interaction (e.g., Croken et al., 2012; Gomez-Diaz et al., 2012). For example, the formation of cysts (an important life cycle stage for host infection) in *Toxoplasma* (Apicomplexa), *Acanthamoeba* (Amoebozoa), and *Giardia* (Excavata) is driven at least partly by epigenetic mechanisms such as histone acetylation and methylation (e.g., H3K18 acetylation and H3R17 methylation in *Toxoplasma*; Saksouk et al., 2005; Dixon et al., 2010; Sonda et al., 2010; Moon et al., 2017; Lagunas-Rangel and Bermudez-Cruz, 2019). Antigenic variation, a strategy used by many pathogens (e.g., *Trypanosoma brucei*, *Giardia lamblia*, *Giardia duodenalis*, and *Plasmodium falciparum*) to avoid the host immune system, also is achieved through epigenetic regulation of gene expression (Kulakova et al., 2006; Elias and Faria, 2009; Juarez-Reyes and Castano, 2019; Lagunas-Rangel and Bermudez-Cruz, 2019). Their genomes contain many genes for surface proteins and the timing of gene expression is at least partly epigenetically regulated, e.g., through histone methylation (H3K4) or acetylation (H3K9) of the *var* genes in *Plasmodium falciparum* (e.g., Freitas-Junior et al., 2005; Guizetti and Scherf, 2013; Duffy et al., 2014). In this way, pathogenic protists are able to rapidly react and adapt to a changing host environment.

A further aspect of cell physiology that seems to involve epigenetics is mating type determination in ciliates (e.g., Pilling et al., 2017). While most ciliate species have different mating types, their number varies greatly (up to a 100; Phadke and Zufall, 2009) and so do the molecular mechanisms for mating type determination (e.g., Orias et al., 2017). For *Paramecium tetraurelia* it could be shown that the difference between its two mating types lies in the presence/absence of a transmembrane protein, whose expression is regulated by “scan RNAs” (Singh et al., 2014).

Despite the fact that details on the exact molecular processes and the genes/enzymes involved often remain scarce, the above-mentioned examples highlight the ubiquity and importance of epigenetics in the life histories of microbial eukaryotes.

THE POTENTIAL ROLE OF EPIGENETICS IN DRIVING ADAPTATION AND DIVERSIFICATION

Ecological speciation through differential adaptation to environmental factors may be a plausible explanation for diversification in protists considering their often-large population sizes and wide biogeographic distribution. Here, we focus on the role of epigenetics in these events, yet we acknowledge that bottlenecks and drift likely are also important drivers of diversity in protists, especially in lineages with small

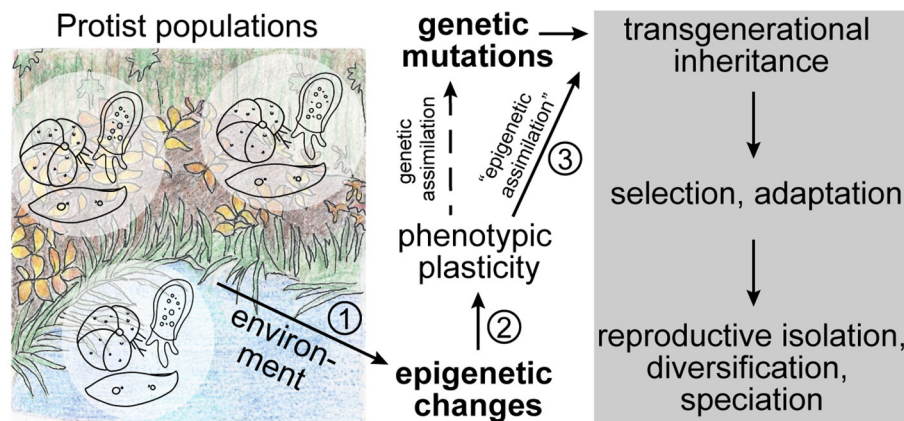


FIGURE 1 | Theoretical sequence of events in ecological speciation driven by epigenetics. In addition to genetic mutations, naturally occurring protist populations experience epigenetic modifications that may be stochastic or be triggered by the environment. These modifications can lead to phenotypic plasticity in the population through changes in genome structure or gene expression states. If the epigenetic modification is followed by a genetic mutation, it may be fixed in the genome through genetic assimilation. However, if the epigenetic mark itself is stably inherited (i.e., “epigenetically assimilation”) across generations, it may represent a selectable advantage that can lead to an increase in fitness of the population and ultimately to adaptation, diversification, and speciation without changes to the genome. The numbers indicate the most critical steps in this sequence of events that we discuss throughout the text.

populations and restricted distribution. However, the effects of these population genetic phenomena on epigenetics remain largely unknown.

In order to elucidate the interactions between ecology, epigenetics, and evolution that are the basis of our suggested model, special consideration has to be placed on the following questions (**Figure 1**): (1) does the environment trigger epigenetic variations, (2) can epigenetic modifications lead to phenotypic plasticity, and (3) are environmentally acquired epigenetic marks stably inherited to establish reproductive isolation and speciation? Over the last few years research efforts investigating these interactions have rapidly increased, yet so far mostly focusing on multicellular model species (e.g., Smith and Ritchie, 2013; Vogt, 2017; Boskovic and Rando, 2018; Perez and Lehner, 2019).

The notion that the environment influences epigenetic modifications is by now well-established. Many studies have focused on the effects of stress, toxin exposure or nutrition on epigenetic marks (e.g., Yaish et al., 2011; Collotta et al., 2013; Tiffon, 2018; Weyrich et al., 2019), and research on natural non-model systems showed epigenetic variability in populations across ecological gradients (e.g., Foust et al., 2016; Mcnew et al., 2017; Johnson and Kelly, 2020; Wogan et al., 2020). These studies usually focus on patterns of DNA methylation as this epigenetic modification is better understood and easy to analyze through bisulfite sequencing methods (e.g., Meissner et al., 2005; Smallwood et al., 2014). To our knowledge, few data exist on similar studies of protists, yet we argue that due to their ubiquitous occurrence across a wide range of environments, protist populations hold great promise for investigating environmental effects on epigenetic variation.

Elucidating the influence of epigenetics on phenotypic plasticity is more challenging as it can be difficult to rule

out underlying genetic influences. However, recently progress has been made, mostly through experimental modifications to epigenetic marks on DNA or histones and the investigation of subsequent effects on the phenotype (e.g., Kronholm et al., 2016; Verhoeven et al., 2016). Research on a natural system was able to show that epigenetic modifications were more likely than genetic variability to have shaped the behavioral reproductive isolation in fish species (Smith et al., 2016). Similarly, epigenetic mechanisms were found to be responsible for phenotypic plasticity in asexual lineages allowing them to respond to environmental fluctuations (Castonguay and Angers, 2012). A further striking example of rapid phenotypic plasticity induced by epigenetics can be found in protist lineages that use antigenic variation through epigenetically regulated changes in gene expression to adjust to changing environments (see above).

For epigenetics to act as driver of speciation, it is important that epigenetic marks are stably inherited across generations, at least until reproductive isolation is established and/or genetic assimilation has occurred (**Figure 1**; discussed in Rey et al., 2016). The stable inheritance of epigenetic marks has long been debated as they were assumed to be eliminated during reproduction and only affect the current generation (discussed in: Richards, 2006), a view that is in line with the concepts of the modern synthesis and the rejection of the idea that acquired traits can be passed down to future generations (discussed in: Jablonka and Lamb, 2008; Bonduriansky, 2012). However, in recent years, examples of soft inheritance through transgenerational inheritance of epigenetic marks became more numerous (e.g., Richards, 2006; Bond and Finnegan, 2007; Perez and Lehner, 2019). Again, important examples can be found among protists, such as ciliates, in which acquired changes to the morphology or physiology, such as doublet morphology and mating types, are inherited to progeny without changes in the

underlying nucleotide sequence (e.g., Pilling et al., 2017; Neeb and Nowacki, 2018). In addition, experimental evolution on the unicellular algae *Chlamydomonas* (Archaeplastida) showed that epigenetic variation is stably inherited across generations and thus influences adaptability of the organism (Kronholm et al., 2017).

CONCLUSION

In recent years, a large amount of research has been published that focuses on the role of epigenetics in ecological speciation. It has been shown that environmentally induced epigenetic modifications can lead to differential gene expression and phenotypic plasticity. If these epigenetic marks are stably inherited across generations (“epigenetic assimilation”) and increase the fitness of the population, they could be substrate for selection and thus represent a first step toward ecological speciation (Figure 1).

While detailed information on the molecular processes of epigenetics in microbial eukaryotes remains scarce, its prominent role in shaping genome dynamics and driving phenotypic plasticity even across generations makes it likely that epigenetics is involved in generating their tremendous diversity. This, as well as their short generation times, make protists interesting model systems for studying the influence of epigenetics on adaptation and speciation. The model of ecological speciation driven by epigenetics presented here is consistent with the idea of rapid diversification in lineages with large population

sizes and therefore weak genetic drift. Recent improvements in the sensitivity of high-throughput sequencing techniques to sequence the genomes, transcriptomes, and epigenomes of non-model microbes make this an exciting time to combine molecular, morphological, and epigenetic approaches for elucidating the origin of species diversity and a species’ response to changing environmental conditions.

AUTHOR CONTRIBUTIONS

AW and LK developed the ideas and wrote the paper. All authors contributed to the article and approved the submitted version.

FUNDING

LK was supported by grants from the National Institute of Health (grant number R15HG010409) and the National Science Foundation (grant numbers OCE-1924570, DEB-1651908, and DEB-1541511).

ACKNOWLEDGMENTS

We thank the editors for organizing the Frontiers Research Topic Gene Regulation as a Driver of Adaptation and Speciation and two reviewers for helpful comments on our manuscript. We thank Eleanor Goetz and Caitlin Timmons for the drawings in the figure.

REFERENCES

- Adl, S. M., Bass, D., Lane, C. E., Lukes, J., Schoch, C. L., Smirnov, A., et al. (2019). Revisions to the classification, nomenclature, and diversity of eukaryotes. *J. Eukaryot. Microbiol.* 66, 4–119. doi: 10.1111/jeu.12691
- Amato, A., Kooistra, W. H., Ghiron, J. H. L., Mann, D. G., Pröschold, T., and Montresor, M. (2007). Reproductive isolation among sympatric cryptic species in marine diatoms. *Protist* 158, 193–207. doi: 10.1016/j.protis.2006.10.001
- Aravind, L., Burroughs, A. M., Zhang, D. P., and Iyer, L. M. (2014). Protein and DNA modifications: evolutionary imprints of bacterial biochemical diversification and geochemistry on the provenance of eukaryotic epigenetics. *Cold Spring Harb. Perspect. Biol.* 6:a016063. doi: 10.1101/cshperspect.a016063
- Badger, M., Tucker, S. J., Grattepanche, J. D., and Katz, L. A. (2017). Rapid turnover of ciliate community members in New England tide pools. *Aquat. Microb. Ecol.* 80, 43–54. doi: 10.3354/ame01839
- Bellec, L., and Katz, L. A. (2012). Analyses of chromosome copy number and expression level of four genes in the ciliate *Chilodonella uncinata* reveal a complex pattern that suggests epigenetic regulation. *Gene* 504, 303–308. doi: 10.1016/j.gene.2012.04.067
- Bernstein, E., and Allis, C. D. (2005). RNA meets chromatin. *Genes Dev.* 19, 1635–1655. doi: 10.1101/gad.1324305
- Bond, D. M., and Finnegan, E. J. (2007). Passing the message on: inheritance of epigenetic traits. *Trends Plant Sci.* 12, 211–216. doi: 10.1016/j.tplants.2007.03.010
- Bonduriansky, R. (2012). Rethinking heredity, again. *Trends Ecol. Evol.* 27, 330–336. doi: 10.1016/j.tree.2012.02.003
- Boskovic, A., and Rando, O. J. (2018). Transgenerational epigenetic inheritance. *Ann. Rev. Genet.* 52, 21–41. doi: 10.1146/annurev-genet-120417-031404
- Castonguay, E., and Angers, B. (2012). The key role of epigenetics in the persistence of asexual lineages. *Genet. Res. Int.* 2012:534289. doi: 10.1155/2012/534289
- Chalker, D. L., Meyer, E., and Mochizuki, K. (2013). Epigenetics of ciliates. *Cold Spring Harb. Perspect. Biol.* 5:a017764. doi: 10.1101/cshperspect.a017764
- Chen, X., Bracht, J. R., Goldman, A. D., Dolzhenko, E., Clay, D. M., Swart, E. C., et al. (2014). The architecture of a scrambled genome reveals massive levels of genomic rearrangement during development. *Cell* 158, 1187–1198. doi: 10.1016/j.cell.2014.07.034
- Chicoine, L. G., and Allis, C. D. (1986). Regulation of histone acetylation during macronuclear differentiation in *Tetrahymena*: evidence for control at the level of acetylation and deacetylation. *Dev. Biol.* 116, 477–485. doi: 10.1016/0012-1606(86)90148-X
- Collotta, M., Bertazzi, P. A., and Bollati, V. (2013). Epigenetics and pesticides. *Toxicology* 307, 35–41. doi: 10.1016/j.tox.2013.01.017
- Croken, M. M., Nardelli, S. C., and Kim, K. (2012). Chromatin modifications, epigenetics, and how protozoan parasites regulate their lives. *Trends Parasitol.* 28, 202–213. doi: 10.1016/j.pt.2012.02.009
- Darling, K. F., and Wade, C. M. (2008). The genetic diversity of planktic foraminifera and the global distribution of ribosomal RNA genotypes. *Mar. Micropaleontol.* 67, 216–238. doi: 10.1016/j.marmicro.2008.01.009
- Dixon, S. E., Stilger, K. L., Elias, E. V., Naguleswaran, A., and Sullivan, W. J. (2010). A decade of epigenetic research in *Toxoplasma gondii*. *Mol. Biochem. Parasitol.* 173, 1–9. doi: 10.1016/j.molbiopara.2010.05.001
- Duffy, M. F., Selvarajah, S. A., Josling, G. A., and Petter, M. (2014). Epigenetic regulation of the *Plasmodium falciparum* genome. *Brief. Funct. Genom.* 13, 203–216. doi: 10.1093/bfpg/elt047
- Elias, M. C., and Faria, M. (2009). Are there epigenetic controls in *Trypanosoma cruzi*? *Ann. N. Y. Acad. Sci.* 1178, 285–290. doi: 10.1111/j.1749-6632.2009.05008.x

- Faure, E., Not, F., Benoiston, A.-S., Labadie, K., Bittner, L., and Ayata, S.-D. (2019). Mixotrophic protists display contrasted biogeographies in the global ocean. *ISME J.* 13, 1072–1083. doi: 10.1038/s41396-018-0340-5
- Fedoroff, N. V. (2012). Presidential address. Transposable elements, epigenetics, and genome evolution. *Science* 338, 758–767. doi: 10.1126/science.338.6108.758
- Foust, C. M., Preite, V., Schrey, A. W., Alvarez, M., Robertson, M. H., Verhoeven, K. J., et al. (2016). Genetic and epigenetic differences associated with environmental gradients in replicate populations of two salt marsh perennials. *Mol. Ecol.* 25, 1639–1652. doi: 10.1111/mec.13522
- Freitas-Junior, L. H., Hernandez-Rivas, R., Ralph, S. A., Montiel-Condado, D., Ruvalcaba-Salazar, O. K., Rojas-Meza, A. P., et al. (2005). Telomeric heterochromatin propagation and histone acetylation control mutually exclusive expression of antigenic variation genes in malaria parasites. *Cell* 121, 25–36. doi: 10.1016/j.cell.2005.01.037
- Garnier, O., Serrano, V., Duharcourt, S., and Meyer, E. (2004). RNA-mediated programming of developmental genome rearrangements in *Paramecium tetraurelia*. *Mol. Cell. Biol.* 24, 7370–7379. doi: 10.1128/MCB.24.17.7370-7379.2004
- Gomez-Diaz, E., Jorda, M., Angel Peinado, M., and Rivero, A. (2012). Epigenetics of host-pathogen interactions: the road ahead and the road behind. *PLoS Pathog.* 8:e1003007. doi: 10.1371/journal.ppat.1003007
- Goodkov, A. V., Berdieva, M. A., Podlipaeva, Y. I., and Demin, S. Y. (2020). The chromatin extrusion phenomenon in *Amoeba proteus* cell cycle. *J. Eukaryot. Microbiol.* 67, 203–208. doi: 10.1111/jeu.12771
- Grell, K. G. (1973). *Protozoology*. Berlin: Springer-Verlag.
- Guizetti, J., and Scherf, A. (2013). Silence, activate, poise and switch! mechanisms of antigenic variation in *Plasmodium falciparum*. *Cell. Microbiol.* 15, 718–726. doi: 10.1111/cmi.12115
- Heyse, G., Jonsson, F., Chang, W. J., and Lipps, H. J. (2010). RNA-dependent control of gene amplification. *Proc. Natl. Acad. Sci. U.S.A.* 107, 22134–22139. doi: 10.1073/pnas.1009284107
- Irwin, N. A., Sabetrasekh, M., and Lynn, D. H. (2017). Diversification and phylogenetics of mobilid peritrichs (ciliophora) with description of *Urceolaria parakorschelti* sp. nov. *Protist* 168, 481–493. doi: 10.1016/j.protis.2017.07.003
- Jablonka, E., and Lamb, M. J. (2008). The epigenome in evolution: beyond the modern synthesis. *Becmihuk BOTuC*, 12, 242–254.
- Jahn, C. L., and Klobutcher, L. A. (2002). Genome remodeling in ciliated protozoa. *Annu. Rev. Microbiol.* 56, 489–520. doi: 10.1146/annurev.micro.56.012302.160916
- Johnson, K. M., and Kelly, M. W. (2020). Population epigenetic divergence exceeds genetic divergence in the Eastern oyster *Crassostrea virginica* in the Northern Gulf of Mexico. *Evol. Appl.* 13, 945–959. doi: 10.1111/eva.12912
- Juarez-Reyes, A., and Castano, I. (2019). Chromatin architecture and virulence-related gene expression in eukaryotic microbial pathogens. *Curr. Genet.* 65, 435–443. doi: 10.1007/s00294-018-0903-z
- Katz, L. A., Mcmanus, G. B., Snoeyenbos-West, O. L. O., Griffin, A., Pirog, K., Costas, B., et al. (2005). Reframing the “Everything is everywhere” debate: evidence for high gene flow and diversity in ciliate morphospecies. *Aquat. Microb. Ecol.* 41, 55–65. doi: 10.3354/ame041055
- Kronholm, I., Bassett, A., Baulcombe, D., and Collins, S. (2017). Epigenetic and genetic contributions to adaptation in *Chlamydomonas*. *Mol. Biol. Evol.* 34, 2285–2306. doi: 10.1093/molbev/msx166
- Kronholm, I., Johannesson, H., and Ketola, T. (2016). Epigenetic control of phenotypic plasticity in the filamentous fungus *Neurospora crassa*. *Genes Genom. Genet.* 6, 4009–4022. doi: 10.1534/g3.116.033860
- Kulakova, L., Singer, S. M., Conrad, J., and Nash, T. E. (2006). Epigenetic mechanisms are involved in the control of *Giardia lamblia* antigenic variation. *Mol. Microbiol.* 61, 1533–1542. doi: 10.1111/j.1365-2958.2006.05345.x
- Lagunas-Rangel, F. A., and Bermudez-Cruz, R. M. (2019). Epigenetics in the early divergent eukaryotic *Giardia duodenalis*: an update. *Biochimie* 156, 123–128. doi: 10.1016/j.biochi.2018.10.008
- Lisch, D. (2009). Epigenetic regulation of transposable elements in plants. *Annu. Rev. Plant Biol.* 60, 43–66. doi: 10.1146/annurev.arplant.59.032607.092744
- Liu, B., and Wendel, J. F. (2003). Epigenetic phenomena and the evolution of plant allopolyploids. *Mol. Phylogenet. Evol.* 29, 365–379. doi: 10.1016/S1055-7903(03)00213-6
- Mcnew, S. M., Beck, D., Sadler-Riggleman, I., Knutie, S. A., Koop, J. A., Clayton, D. H., et al. (2017). Epigenetic variation between urban and rural populations of Darwin’s finches. *BMC Evol. Biol.* 17:183. doi: 10.1186/s12862-017-1025-9
- Meissner, A., Gnirke, A., Bell, G. W., Ramsahoye, B., Lander, E. S., and Jaenisch, R. (2005). Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis. *Nucleic Acids Res.* 33, 5868–5877. doi: 10.1093/nar/gki901
- Moon, E. K., Hong, Y., Lee, H. A., Quan, F. S., and Kong, H. H. (2017). DNA Methylation of gene expression in *Acanthamoeba castellanii* encystation. *Korean J. Parasitol.* 55, 115–120. doi: 10.3347/kjp.2017.55.2.115
- Neeb, Z. T., and Nowacki, M. (2018). RNA-mediated transgenerational inheritance in ciliates and plants. *Chromosoma* 127, 19–27. doi: 10.1007/s00412-017-0655-4
- Ng, H. H., and Bird, A. (1999). DNA methylation and chromatin modification. *Curr. Opin. Genet. Dev.* 9, 158–163. doi: 10.1016/S0959-437X(99)80024-0
- Nowacki, M., Shetty, K., and Landweber, L. F. (2011). RNA-Mediated epigenetic programming of genome rearrangements. *Annu. Rev. Genom. Hum. Genet.* 12, 367–389. doi: 10.1146/annurev-genom-082410-101420
- Oliverio, A., Lahr, D. J. G., Nguyen, T., and Katz, L. A. (2014). Cryptic diversity within morphospecies of testate (shelled) amoebae in New England bogs and fens. *Protist* 165, 196–207. doi: 10.1016/j.protis.2014.02.001
- Orias, E., Singh, D. P., and Meyer, E. (2017). Genetics and epigenetics of mating type determination in *Paramecium* and *Tetrahymena*. *Annu. Rev. Microbiol.* 71, 133–156. doi: 10.1146/annurev-micro-090816-093342
- Parfrey, L. W., Lahr, D. J. G., and Katz, L. A. (2008). The dynamic nature of eukaryotic genomes. *Mol. Biol. Evol.* 25, 787–794. doi: 10.1093/molbev/msn032
- Perez, M. F., and Lehner, B. (2019). Intergenerational and transgenerational epigenetic inheritance in animals. *Nat. Cell Biol.* 21, 143–151. doi: 10.1038/s41556-018-0242-9
- Phadke, S. S., and Zufall, R. A. (2009). Rapid diversification of mating systems in ciliates. *Biol. J. Linn. Soc.* 98, 187–197. doi: 10.1111/j.1095-8312.2009.01250.x
- Pilling, O. A., Rogers, A. J., Gulla-Devaney, B., and Katz, L. A. (2017). Insights into transgenerational epigenetics from studies of ciliates. *Eur. J. Protistol.* 61, 366–375. doi: 10.1016/j.ejop.2017.05.004
- Razin, A., and Riggs, A. D. (1980). DNA Methylation and gene-function. *Science* 210, 604–610. doi: 10.1126/science.6254144
- Rey, O., Danchin, E., Mirouze, M., Loot, C., and Blanchet, S. (2016). Adaptation to global change: a transposable element-epigenetics perspective. *Trends Ecol. Evol.* 31, 514–526. doi: 10.1016/j.tree.2016.03.013
- Richards, E. J. (2006). Inherited epigenetic variation—revisiting soft inheritance. *Nat. Rev. Genet.* 7, 395–401. doi: 10.1038/nrg1834
- Ryšánek, D., Holzing, A., and Škaloud, P. (2016). Influence of substrate and pH on the diversity of the aeroterrestrial alga *Klebsormidium* (Klebsormidiales, Streptophyta): a potentially important factor for sympatric speciation. *Phycologia* 55, 347–358. doi: 10.2216/15-110.1
- Ryšánek, D., Hrkčková, K., and Škaloud, P. (2015). Global ubiquity and local endemism of free-living terrestrial protists: phylogeographic assessment of the streptophyte alga *Klebsormidium*. *Environ. Microbiol.* 17, 689–698. doi: 10.1111/1462-2920.12501
- Saksouk, N., Bhatti, M. M., Kieffer, S., Smith, A. T., Musset, K., Garin, J., et al. (2005). Histone-modifying complexes regulate gene expression pertinent to the differentiation of the protozoan parasite *Toxoplasma gondii*. *Mol. Cell. Biol.* 25, 10301–10314. doi: 10.1128/MCB.25.23.10301-10314.2005
- Shabalina, S. A., and Koonin, E. V. (2008). Origins and evolution of eukaryotic RNA interference. *Trends Ecol. Evol.* 23, 578–587. doi: 10.1016/j.tree.2008.06.005
- Singh, D. P., Saudemont, B., Guglielmi, G., Arnaiz, O., Gout, J. F., Prajer, M., et al. (2014). Genome-defence small RNAs adapted for epigenetic mating-type inheritance. *Nature* 509, 447–452. doi: 10.1038/nature13318
- Škaloud, P., Škaloudová, M., Doskočilová, P., Kim, J. I., Shin, W., and Dvůrák, P. (2019). Speciation in protists: spatial and ecological divergence processes cause rapid species diversification in a freshwater chrysophyte. *Mol. Ecol.* 28, 1084–1095. doi: 10.1111/mec.15011
- Šlapeta, J., Moreira, D., and López-García, P. (2005). The extent of protist diversity: insights from molecular ecology of freshwater eukaryotes. *Proc. R. Soc. Lond. B Biol. Sci.* 272, 2073–2081. doi: 10.1098/rspb.2005.3195
- Smallwood, S. A., Lee, H. J., Angermueller, C., Krueger, F., Saadeh, H., Peat, J., et al. (2014). Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity. *Nat. Methods* 11, 817–820. doi: 10.1038/nmeth.3035

- Smith, G., and Ritchie, M. G. (2013). How might epigenetics contribute to ecological speciation? *Curr. Zool.* 59, 686–696. doi: 10.1093/czoolo/59.5.686
- Smith, T. A., Martin, M. D., Nguyen, M., and Mendelson, T. C. (2016). Epigenetic divergence as a potential first step in darter speciation. *Mol. Ecol.* 25, 1883–1894. doi: 10.1111/mec.13561
- Sonda, S., Morf, L., Bottova, I., Baetschmann, H., Rehrauer, H., Cafilisch, A., et al. (2010). Epigenetic mechanisms regulate stage differentiation in the minimized protozoan *Giardia lamblia*. *Mol. Microbiol.* 76, 48–67. doi: 10.1111/j.1365-2958.2010.07062.x
- Swart, E. C., Wilkes, C. D., Sandoval, P. Y., Arambasic, M., Sperling, L., and Nowacki, M. (2014). Genome-wide analysis of genetic and epigenetic control of programmed DNA deletion. *Nucleic Acids Res.* 42, 8970–8983. doi: 10.1093/nar/gku619
- Tiffon, C. (2018). The impact of nutrition and environmental epigenetics on human health and disease. *Int. J. Mol. Sci.* 19:3425. doi: 10.3390/ijms19113425
- Tucker, S. J., Mcmanus, G. B., Katz, L. A., and Grattepanche, J. D. (2017). Distribution of abundant and active planktonic ciliates in coastal and slope waters off New England. *Front. Microbiol.* 8:2178. doi: 10.3389/fmicb.2017.02178
- Verhoeven, K. J. F., Vonholdt, B. M., and Sork, V. L. (2016). Epigenetics in ecology and evolution: what we know and what we need to know. *Mol. Ecol.* 25, 1631–1638. doi: 10.1111/mec.13617
- Vogt, G. (2017). Facilitation of environmental adaptation and evolution by epigenetic phenotype variation: insights from clonal, invasive, polyploid, and domesticated animals. *Environ. Epigenet.* 3:dvx002. doi: 10.1093/eep/dvx002
- Weiner, A. K., Cerón-Romero, M. A., Yan, Y., and Katz, L. A. (2020). Phylogenomics of the epigenetic toolkit reveals punctate retention of genes across eukaryotes. *Genome Biol. Evol.* 12, 2196–2210. doi: 10.1093/gbe/evaa198
- Weiner, A. K., Weinkauff, M. F., Kurasawa, A., Darling, K. F., Kucera, M., and Grimm, G. W. (2014). Phylogeography of the tropical planktonic foraminifera lineage *Globigerinella* reveals isolation inconsistent with passive dispersal by ocean currents. *PLoS ONE* 9:e92148. doi: 10.1371/journal.pone.0092148
- Weyrich, A., Lenz, D., and Fickel, J. (2019). Environmental change-dependent inherited epigenetic response. *Genes* 10:4. doi: 10.3390/genes10010004
- Wogan, G. O. U., Yuan, M. L., Mahler, D. L., and Wang, I. J. (2020). Genome-wide epigenetic isolation by environment in a widespread *Anolis* lizard. *Mol. Ecol.* 29, 40–55. doi: 10.1111/mec.15301
- Yaish, M. W., Colasanti, J., and Rothstein, S. J. (2011). The role of epigenetic processes in controlling flowering time in plants exposed to stress. *J. Exp. Bot.* 62, 3727–3735. doi: 10.1093/jxb/err177

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2021 Weiner and Katz. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



Promoter Architecture and Transcriptional Regulation of Genes Upregulated in Germination and Coleoptile Elongation of Diverse Rice Genotypes Tolerant to Submergence

Bijayalaxmi Mohanty*

NUS Environmental Research Institute, National University of Singapore, Singapore, Singapore

OPEN ACCESS

Edited by:

Deborah A. Triant,
University of Virginia, United States

Reviewed by:

Nisha Singh,
Cornell University, United States
Ishaan Gupta,
Indian Institute of Science Education
and Research, India

*Correspondence:

Bijayalaxmi Mohanty
eriv97@nus.edu.sg;
bijayalaxmi.mohanty@gmail.com

Specialty section:

This article was submitted to
Systems Biology,
a section of the journal
Frontiers in Genetics

Received: 09 December 2020

Accepted: 08 February 2021

Published: 16 March 2021

Citation:

Mohanty B (2021) Promoter
Architecture and Transcriptional
Regulation of Genes Upregulated
in Germination and Coleoptile
Elongation of Diverse Rice Genotypes
Tolerant to Submergence.
Front. Genet. 12:639654.
doi: 10.3389/fgene.2021.639654

Rice has the natural morphological adaptation to germinate and elongate its coleoptile under submerged flooding conditions. The phenotypic deviation associated with the tolerance to submergence at the germination stage could be due to natural variation. However, the molecular basis of this variation is still largely unknown. A comprehensive understanding of gene regulation of different genotypes that have diverse rates of coleoptile elongation can provide significant insights into improved rice varieties. To do so, publicly available transcriptome data of five rice genotypes, which have different lengths of coleoptile elongation under submergence tolerance, were analyzed. The aim was to identify the correlation between promoter architecture, associated with transcriptional and hormonal regulation, in diverse genotype groups of rice that have different rates of coleoptile elongation. This was achieved by identifying the putative *cis*-elements present in the promoter sequences of genes upregulated in each group of genotypes (tolerant, highly tolerant, and extremely tolerant genotypes). Promoter analysis identified transcription factors (TFs) that are common and unique to each group of genotypes. The candidate TFs that are common in all genotypes are MYB, bZIP, AP2/ERF, ARF, WRKY, ZnF, MADS-box, NAC, AS2, DOF, E2F, ARR-B, and HSF. However, the highly tolerant genotypes interestingly possess binding sites associated with HY5 (bZIP), GBF3, GBF4 and GBF5 (bZIP), DPBF-3 (bZIP), ABF2, ABI5, bHLH, and BES/BZR, in addition to the common TFs. Besides, the extremely tolerant genotypes possess binding sites associated with bHLH TFs such as BEE2, BIM1, BIM3, BM8 and BAM8, and ABF1, in addition to the TFs identified in the tolerant and highly tolerant genotypes. The transcriptional regulation of these TFs could be linked to phenotypic variation in coleoptile elongation in response to submergence tolerance. Moreover, the results indicate a cross-talk between the key TFs and phytohormones such as gibberellic acid, abscisic acid, ethylene, auxin, jasmonic acid, and brassinosteroids, for an altered transcriptional regulation leading to differences in germination and coleoptile elongation under submergence. The information derived from the current *in silico* analysis can potentially assist in developing new rice breeding targets for direct seeding.

Keywords: rice, germination and coleoptile elongation, submergence tolerance, promoter *cis*-element, transcription factor, phytohormones, plant breeding

INTRODUCTION

Rice is one of the major cereal crops and staple food in Southeast Asian countries. In most of these countries, people sow seeds in the flooded rice fields through direct seeding that helps in reducing cost and manpower. During this process, rice seeds encounter flooding/submergence stress as they are exposed to hypoxia and even anoxia. Rice has the natural morphological adaptation to germinate and elongate its coleoptile under submerged flooding condition and also even in anoxic conditions. This coleoptile reaches the surface of water to get O₂ for aerobic respiration to get sufficient energy for the development of roots and shoots (Narsai et al., 2015). However, there is phenotypic variation in coleoptile elongation and the tolerance to submergence at the germination stage among different genotypes and which could be due to natural variation. Tolerant genotypes germinate and elongate their coleoptile and hydrolyze starch to sugars for glycolysis and enhanced ethanolic fermentation to generate ATP (Lee et al., 2009). Besides, the plant hormone ethylene plays a major role during this adaptation (Ma et al., 2010; Steffens et al., 2012). Although a number of genetic studies provide information that the extent of tolerance to flooding in rice at the germination stage is due to natural variation (Septiningsih et al., 2013; Baltazar et al., 2014; Hsu and Tung, 2015), the molecular mechanism behind this phenotypic variation is largely unknown. A comprehensive understanding of the gene regulation of different genotypes tolerant to submergence/flooding can provide a significant

insight into breeding of improved tolerant rice varieties for direct seeding cultivation.

A number of microarray and RNA-sequencing studies performed on rice germination under anoxia/hypoxia have shown the up-/downregulation of a number of stress-responsive genes (Lasanthi-Kudahettige et al., 2007; Narsai et al., 2009; Shingaki-Wells et al., 2011; Narsai et al., 2015; Hsu and Tung, 2017; Wu and Yang, 2020). During such stress conditions, transcription factors (TFs) play a key role in regulating the expression of genes that support them to survive through tolerance mechanism. The molecular mechanism of submergence has been studied on the rice genotype FR13A, which is able to survive 2 weeks as it has *Submergence 1* (*SUB1*) locus that encodes genes such as *SUB1A*, *SUB1B*, and *SUB1C* that belong to the ethylene-response factor (ERF) subgroup VII (Xu et al., 2006). Among those three genes, *SUB1A* helps rice plants to tolerate submergence (Fukao and Bailey-Serres, 2008a,b). In addition, it has also been shown that *SNORKELs* genes such as *SNORKEL 1* (*SK1*) and *SNORKEL 2* (*SK2*) can also induce tolerance in rice plants (deep-water floating rice) by inducing GA for rapid internode elongation to rescue the plants from drowning (Hattori et al., 2009). These *SNORKELs* also belong to ERF subgroup VII. In rice coleoptiles, a number of ERFs such as *ERF60*, *ERF67*, and *ERF68* genes have been shown to be upregulated under anoxic conditions and also belong to the ERF subgroup VII (Lin et al., 2019). These data favor the understanding that rice coleoptile elongation is promoted by ethylene during submerged conditions. Besides, it has been shown that auxin also plays a key role in coleoptile elongation during submergence (Ishizawa and Esashi, 1983; Wu and Yang, 2020). Auxin-dependent differential growth in rice coleoptile is shown to be due to the effect of the TF auxin response factor 1 (*OsARF1*) (Waller et al., 2002). Likewise, TF WRKY is also involved in the regulation of tolerance to rice under submergence (Viana et al., 2018). High accumulation of WRKY in both shoots and roots was observed in response to submergence. The expression of *OsWRKY11* and *OsWRKY56* was more than 100-fold in comparison to controls in the root tissues. These results support the importance of TFs in the regulation of tolerance of rice plants under submerged/flooding conditions. Recent studies of the roles of auxin in rice coleoptile elongation suggest that it plays a key role in the cell division and tropism of rice coleoptiles under submergence (Wu and Yang, 2020).

For such transcriptional regulation, the expression of a specific TF can regulate the expression of a number of specific sets of stress-responsive genes by binding to their cognate *cis*-acting elements in the promoters of specific genes (Mizoi et al., 2012). However, there are very limited studies performed on the *cis*-element enrichment analysis in the promoter of such stress-responsive genes. The studies so far revealed the identification of a number of putative *cis*-acting elements that can be potentially associated with TF families such as ARF, ERF, MYB, WRKY, bZIP, E2F, and ZnF (Mohanty et al., 2012; Lakshmanan et al., 2014; Sharma et al., 2018). This association of TFs also provides useful potential links with different hormonal signaling as it regulates the downstream genes through interactions with other regulatory molecules of signaling pathways (Song et al., 2005).

Abbreviations: ABA, Absciscic Acid; ABF1/2, ABA-responsive element binding factor 1/2; ABI3/4/5, Absciscic Acid-Insensitive 3/4/5; ABR1, AP2-like ABA repressor 1; ABRE, Absciscic Acid Response Element; AGL, Agamous-like; AP2/B3, APETALA2/B3; AP2/ERF, APETALA2/Ethylene-Responsive Element; ARR-B-Type-B Arabidopsis Response Regulator; ARF, Auxin Response Factor; ATHB4 Arabidopsis thaliana Homeobox 2; AUX1, Auxin influx carrier Auxin Resistant 1; BAM 8, BZR1-BAM 8; BAM1, β -amylase (BAM)-like 2; BEE, BR Enhanced Expression; BEH4, BES1/BZR1 Homolog 4; BES1, BRASSINOSTEROID INSENSITIVE1-ETHYL METHANESULFONATE-SUPPRESSOR 1; bHLH, basic Helix-Loop-Helix; BIM, BES1-Interacting MYC-Like; BIN2, Brassinosteroid Insensitive 2; BPC, Basic Penta Cysteine; BR, Brassinosteroid; bZIP, Basic Leucine Zipper; BZR1/2, Brassinazole-Resistant 1/2; CARG-box, 'C-A rich-G-box; CAMTA, Calmodulin binding transcription factor; CBF, C-repeat Binding Factor; CCA1, circadian clock associated 1; C2H2: CCHH Zinc Finger; C3H, CCCH Zinc Finger; CCCH CPRF 5/6/7, Common Plant Regulatory Factor 5/6/7; cpm1, coleoptile photomorphogenesis 1; DOF, DNA-Binding with One Finger; DPBF-1/2, Dc3 Promoter Binding Factor 1/2; DREB, Dehydration-Responsive Element-Binding; E2F, E2 factor; EIL, Ethylene-Insensitive 3 (EIN3)-like protein3 (EIL3); EIN, Ethylene-Insensitive 3; ERF, Ethylene Response Factor; GA, Gibberellic Acid; GARE, Gibberellic Acid Response Element; GBF, G-Box Binding Factor; HD, ZIP-Homeodomain Leucine Zipper; HSF, Heat Shock Factor; HY5, Elongated Hypocotyl 5; IAA, Indole-3-Acetic Acid; JA, Jasmonate; JA, Jasmonic Acid; JARE, Jasmonic acid Response Element; JAZ, Jasmonate ZIM domain; LBD, Asymmetric Leaves 2 (ASL2)/Lateral organ boundaries domain; MADS-box, Minichromosome Maintenance Factor 1 (MCM1 from *Saccharomyces cerevisiae*), Agamous (AG; from *Arabidopsis thaliana*), Deficiens (DEF; from *Antirrhinum majus*), and Serum Response Factor (SRF; from *Homo sapiens*); MYB, Myeloblastosis; NAC, NAM, ATAF/2, and CUC2; NAM, No apical meristem; PIF1/3/4/7, Phytochrome Interacting Factor 1/3/4/7; RAP, Related to AP2; RAV, Related to ABI3/VP1; RNA, seq-RNA sequencing; ROS, Reactive oxygen species; SUB1, Submergence 1; SK1, SNORKEL1; SK2, SNORKEL2; TCP, TEOSINTE BRANCHED 1; TFs, Transcription Factors; TSS, Transcription Start Site; ZCT 1, 2, 3, Zinc Finger Catharanthus Transcription Factor 1, 2, and 3; ZnF, Zinc Finger.

The genome-wide gene expression profile generated by microarray/transcriptome analyses on rice germination and coleoptile elongation under hypoxia/anoxia have revealed the involvement of a common molecular mechanism that is associated with carbohydrate metabolism, fermentation, hormone induction, cell division, and expansion (Lasanthi-Kudahettige et al., 2007; Shingaki-Wells et al., 2011; Narsai et al., 2015; Hsu and Tung, 2017). However, the level of tolerance to submergence/flooding during germination and coleoptile elongation could be due to the natural variation. The molecular basis of this variation due to transcriptional regulation is unknown. Hence, the aim was to analyze and identify the specific promoter architecture of gene expression data of different groups of diverse rice genotypes that display variation in the rate of coleoptile elongation under submerged conditions (Hsu and Tung, 2017). These analysis can provide a hypothesis on the key TFs that are associated with genetic variation of different genotypes of rice germination and coleoptile elongation. Further, it can be linked to common and/or unique transcriptional modules associated with different genotypes.

MATERIALS AND METHODS

Extraction of Promoter Sequences of Common Genes Upregulated in Different Groups of Diverse Genotypes of Rice Tolerant to Submergence

Common upregulated genes in the coleoptiles of different genotype groups of rice under submerged conditions were identified from published RNA-sequencing data (Hsu and Tung, 2017). The five diverse rice genotypes-the *japonica* variety Nipponbare, two recombinant inbred lines F291 and F274-2a generated from a cross between Nipponbare and IR64, and two natural genotypes originated from Southeast Asia [8391 from Laos (IRGC 94599) and 8753 from Indonesia (IRGC 54313)]-were used for the promoter architecture analysis. To identify the promoter *cis*-element content of the common upregulated genes of different rice genotypes based on their extent of tolerance in terms of elongation of coleoptiles, common genes upregulated in different groups of genotypes (grouping done by Hsu and Tung, 2017) were extracted. The analysis was performed for all five tolerant genotypes (Nipponbare, F291, F274-2a, 8391, and 8753), which have 23 upregulated genes (**Supplementary Table 1**); four highly tolerant genotypes (F291, F274-2a, 8391, and 8753), which have 16 upregulated genes (**Supplementary Table 2**); and two extremely tolerant natural genotypes (8391 and 8753), which have 27 upregulated genes (**Supplementary Table 3**). The promoter sequences [-1000, +200 nucleotide] relative to the experimentally verified Transcription Start Site (TSS) were extracted for these common upregulated genes from our in-house rice promoter sequence database (Mohanty et al., 2012).

To validate the identification of promoter architecture analysis of the three groups of genotypes that have varying degrees of tolerance to submergence, two additional analyses were performed. In one of the analysis that belongs to the intermediate

TABLE 1 | Potential putative *cis*-elements identified in the promoters of upregulated genes in tolerant genotypes (Nipponbare, F291, F274-2a, 8391, and 8753).

Cis-elements	Motifs	Associated TFs	% (TIC), E-value*
AT-hook/PE1-like	TTTTTTT	MYB (PF1)	61 (16.56), 1e-004
GT-element-like	CTGAAAAAG	MYB (GT-1/GT-3b)	65 (12.79), 3e-004
GARE-like	CAACAACA	MYB (R1, R2R3)	74 (11.70), 0e+000
MYB-box-like	GGTGGGCG	MYB (R2R3)	78 (10.89), 3e-004
	CAACAACA	MYB (R2R3)	74 (11.70), 0e+000
	GACAAATT	MYB (R2R3)	65 (12.58), 7e-004
	CAAAACCA	MYB (R2R3)	61 (13.34), 4e-004
As-1/ocs-like	CGTTGATC	bZIP (Gr. D, I, S)	70 (11.38), 1e-004
	CTGAAAAAA	bZIP (Gr. D, S)	65 (12.79), 3e-004
GCN4 motif	CAACAACA	bZIP (RISBZ1, Gr. G)	74 (11.70), 0e+000
	GACAAATT	bZIP (RISBZ1, Gr. G)	65 (12.58), 6e-004
CAMTA5 binding site-like	CCACACAA	bZIP (CAMTA5)	65 (12.57), 6e-004
GCC-box-like	AGTGGGCG	ERF (I, IV, VII, X)	78 (10.89), 3e-004
	CGCCGCCG	ERF (I, IV, VII, X)	70 (14.42), 9e-005
	CGCCGCCTC	ERF (I, IV, VII, X)	57 (14.27), 2e-004
ERE-like	GCCGAGAG	ERF/RAP2.3, RAP2.10 (Gr. III)	65 (11.42), 3e-004
	GCGGCCATT	ERF/RAP2.2, RAP2.3 (Gr. III)	52 (12.47), 4e-004
	TGCTCGCCG	ERF/RAP2.6, RAP2.10 (Gr. VII)	52 (13.56), 2e-004
		ERF/RAP2.6 (Gr. VII)	
RAV1a-like	CAACAACA	ERF/RAV1 (Gr. II, III and RAV1)	74 (11.70), 0e+000
CRT/DRE-like	CGCCGCCTC	ERF (Gr. III, IV)	57 (14.27), 2e-004
JA response element-like	CTTTGATC	ERF (ERF Gr. VI, VIII, IX)	70 (11.38), 1e-004
ABRE-like	ATTTAGCG	ABI3 (B3 domain of AP2/)	61 (12.23), 2e-004
	TGCTTGCCG	ABI3 (B3 domain)	52 (13.66), 2e-004
ABR-binding site-like	GCCGAGAG	AP2-like ABA repressor 1 (ABR1)	65 (11.42), 3e-004
	TGCTCGCCG	AP2-like ABA repressor 1 (ABR1)	52 (13.66), 2e-004
Ethylene-insensitive 3 binding site-like	AAATGCAAA	EIL3 (EIN3)	61 (13.85), 9e-005
	GAATGCAAA	EIL3 (EIN3)	61 (13.85), 9e-005
Aux-RE-like	CAACAACA	ARF1	74 (11.70), 0e+000
	ACGAGACC	ARF1, ARF5, ARF7	52 (11.65), 3e-004
W-box-like	GTCAAATT	WRKY (Gr. I, IIa, IIc, III)	56 (15.03), 9e-005
Zing finger binding site-like	CAACAACA	ZnF (C2H2-type)	74 (11.70), 0e+000
	GCCGAGAG	ZnF (C2H2-type)	65 (11.42), 3e-004
IDD binding site-like	CAACAACA	IDD 2, 4, 5, 7 (ZnF)	74 (11.70), 0e+000
	CCACACAA	IDD (ZnF)	65 (12.57), 6e-004
Zinc-finger-binding site-like	AGTGGGCG	STZ (Salt tolerant zinc finger)	78 (10.89), 3e-004
DRE-like	CAACAACA	NAC (38)	74 (11.70), 0e+000
	GCCGAGAG	NAC (65)	65 (11.42), 3e-004
	AAATGCAAA	NAC (11, 45)	61 (13.85), 9e-005
	TGCTTGCCG	NAC (58)	52 (13.66), 2e-004
	TGCTCGCCG	NAC (62)	52 (13.66), 2e-004
	ACGAGACC	NAC (NAM)	52 (11.65), 3e-004
	ACGAGACC	NAP(NAC-like, activated by AP3/P1)	52 (11.65), 3e-004

(Continued)

TABLE 1 | Continued

Cis-elements	Motifs	Associated TFs	% (TIC), E-value*
AAAAG/CTTTT-element-like	CTGAAAAAA	DOF (DOF1/4/11/22)	65(12.79), 3e-004
	CTGAAAAAG	DOF (DOF1/4/11/22)	65(12.79), 3e-004
Heat shock binding factor element-like	GCCGAGAG	HSF(HSF3)	65 (11.42), 3e-004
	ATTTAGCG	HSF(HSFA4A, HSF7)	61 (12.23), 2e-004
	TTCTTCTG	HSF (HSFA6B, HSF4, HSF7)	52 (13.96), 3e-004
S2-binding site-like	CCACACAA	AS2 (LBD16, 19)	65 (12.57), 6e-004
	GACAAATT	AS2 (LBD16)	65 (12.58), 7e-004
	CTGAAAAAG	AS2 (LBD16, 18)	65 (12.79), 3e-004
	ATTTAGCG	AS2 (LBD2)	61 (12.23), 2e-004
	GCGGCCATT	AS2 (LBD23)	52 (12.47), 4e-004
	TTCTTCTG	AS2 (LBD30)	52 (13.96), 2e-004
E2F binding site-like	ATTTAGCG	E2F-like protein, E2Fa,	61 (12.23), 2e-004
	ATGGGACT	E2Fc (DEL1)	52 (12.87), 3e-004
	GCGGCCATT	E2F (DEL1)	52 (12.47), 4e-004
		E2F (DEL2)	
TCP binding site-like	ACGAGACC	TCP (TCP 3, 24)	52 (11.65), 3e-004
HD-ZIP binding site-like	AAATGCAAA	HD-ZIP (ATHB4)	61 (13.85), 9e-005
ARR-14 binding element-like	ACGAGACC	ARR-B (ARR14)	52 (11.65), 3e-004
CARG box-binding site-like	GACAAATT	MADS-box (AGL 6, 15)	65 (12.58), 6e-004
	GCCGAGAG	MADS-box (AGL 95)	65 (11.42), 3e-004
	AAATGCAAA	MADS-box (AGL16)	61 (13.85), 9e-005
	ATTTAGCG	MADS-box (AGL16)	61 (12.23), 2e-004
GATA binding site-like	ACGAGACC	GATA1	52 (11.65), 3e-004
	GCGGCCATT	GATA1	52 (12.47), 4e-004
DBP-binding site-like	TAAATATA	DBP1	56 (13.75), 1e-005
TATA-box-like	TAAATATA	TBP	56 (13.75), 1e-005

% = percent occurrence among all upregulated genes, TIC = total information content of homology, E-value* = E-value of homology with promoter database entry.

tolerant group, two highly tolerant genotypes (F291 and F274-2a) that have 24 upregulated genes (**Supplementary Table 4**) were analyzed to compare with the group that has four highly tolerant genotypes (F291, F274-2a, X8391, and X8753). In the other analysis, two genotypes (Nipponbare and F274-2a) that have 84 upregulated genes (**Supplementary Table 5**) were analyzed for comparison with the group that has five tolerant genotypes (Nipponbare, F291, F274-2a, X8391, and X8753) from different backgrounds. The comparisons of the two groups of genotypes with respective similar groups of genotypes will support the analysis.

Identification of Putative *cis*-Elements and Associated TFs

Putative *cis*-elements were identified in the promoter regions of each set of common upregulated genes by the “The Dragon Motif Builder program” having EM2 option with a threshold value of 0.875 (Huang et al., 2005), which is similar to our earlier detection (Mohanty et al., 2012, 2016; Lakshmanan et al., 2014). The program identified a total of 30 motifs having a length of 8–10 nucleotides for each set of common genes. Motifs were selected having a threshold value of 10^{-3} and more than 50% occurrence.

The biological significance of these motifs was verified by Transcription Factor Binding databases such as TRANSFAC (Matys et al., 2003¹), PLACE database (Higo et al., 1999²), AGRIS (Davuluri et al., 2003; Yilmaz et al., 2011³), and PlantPAN 3.0 (Chow et al., 2019⁴). Putative *cis*-elements in the promoters of the upregulated genes were identified based on their sites for different plant TFs present in plant genes with a minimum sequence length of five nucleotides. The cutoffs for core and matrix similarities were more than 75%. TF genes, significantly upregulated, were identified from the RNA-sequencing data (Hsu and Tung, 2017) and annotated based on the RAP genome annotations (Itoh et al., 2007; Tanaka et al., 2008⁵). The methods used to analyze the data have been summarized and presented in **Figure 1**.

RESULTS

Identification of Putative *cis*-Elements in the Upregulated Genes of Tolerant, Highly Tolerant, and Extremely Tolerant Diverse Rice Genotypes in Response to Submergence

Promoter regions of common genes upregulated in diverse genotypes of rice germination and coleoptile elongation in response to submergence tolerance contain *cis*-elements/TF binding sites that are responsible for the regulation of genes associated with various hormone signaling as well as metabolic pathways. Our analysis identified a number of common putative *cis*-elements in the promoter regions of the upregulated genes in tolerant (Nipponbare, F291, F274-2a, 8391, and 8753), highly tolerant (F291, F274-2a, 8391, and 8753), and extremely tolerant (8391 and 8753) genotypes. Putative *cis*-elements that are present in all genotypes are found to be associated with many known TFs such as MYB, bZIP, AP2/ERF, ARF, WRKY, ZnF, MADS-box, NAC, AS2, DOF, E2F, ARR-B, and HSF (**Tables 1–3**). Interestingly, both highly tolerant and extremely tolerant genotypes, in addition to the above binding sites for the associated TFs, also possess unique binding sites that are associated with TFs such as HY5 (bZIP), GBF3, 4, and 5 (bZIP), DPBF-3 (bZIP), ABF2, ABI5, bHLH (basic helix-loop-helix), and BES/BZR (BR-responsive TFs) (**Table 2**). Moreover, the extremely tolerant genotypes that have highly elongated coleoptiles comprise more specific binding sites that are associated with bHLH TFs such as BEE2, BIM1, BIM3, BM8, and BAM8 and bZIP TF ABF1 (**Table 3**). Hence, the following TF regulatory modules are identified as transcriptional activators/repressors that could be involved in controlling the regulation of germination and variation in coleoptile elongation of diverse rice genotypes in response to submergence tolerance.

¹www.gene-regulation.com

²http://www.dna.affrc.go.jp/htdocs/PLACE/

³http://arabidopsis.med.ohio-state.edu/

⁴http://PlantPAN.itps.ncku.edu.tw

⁵http://rapdb.dna.affrc.go.jp/

TABLE 2 | Potential putative *cis*-elements identified in the promoters of upregulated genes in highly tolerant genotypes (F291, F274-2a, 8391, and 8753).

<i>Cis</i> -elements	Motifs	Associated TFs	%(TIC), <i>E</i> -value*
AT-hook/PE1-like	TTAAAAAC	MYB (PF1)	63 (13.49), 2e-004
	TTAAAAATA	MYB (PF1)	63 (16.52), 2e-005
	TATTAAAAA	MYB (PF1)	63 (15.86), 5e-005
	TTAATTTTT	MYB (PF1)	56 (15.46), 1e-004
GT-element-like	CAACCACA	MYB (GT-1)	81 (11.63), 3e-004
	ATTTGATTT	MYB (GT-1)	81 (13.98), 5e-005
	CCAACCAA	MYB (GT-1)	69 (11.71), 1e-004
	TTTGTATTTA	MYB (GT-3b)	63 (13.49), 2e-004
	TGCATGTA	MYB (GT-3a)	50 (13.95), 3e-005
	CCAACCAA	MYB (R1, R2R3)	69 (11.71), 1e-004
GARE-like	TAAACAAA	MYB (R2R3)	94 (12.79), 1e-004
MYB-box-like	TAAACAGA	MYB (R2R3)	94 (12.79), 1e-004
	AAACCACA	MYB (R2R3)	81 (11.63), 3e-004
	CAACCACA	MYB (R2R3)	81 (11.63), 3e-004
	GCTAGCTAGA	MYB (R2R3)	81 (12.81), 3e-004
	CAAGCTGC	MYB (R2R3)	69 (12.41), 1e-004
	CCAAACAA	MYB (R2R3)	69 (11.71), 1e-004
	CCAACCAA	MYB (R2R3)	69 (11.71), 1e-004
	AAGCTGAG	MYB-like (CDC5)	69 (11.15), 9e-005
	TTAATTTTT	MYB-like (CCA1, CDA-1)	56 (15.46), 1e-004
As-1/ocs-like	ATTTGATTT	bZIP (Gr. D, I, S)	81 (13.98), 5e-005
	TGAAGCTT	bZIP (Gr. D, I, S)	69 (12.23), 6e-005
	GATCGTGA	bZIP (Gr. D, I, S)	56 (13.61), 5e-005
ABRE-like	CAAGCTGC	bZIP (Gr. A)	69 (12.41), 1e-004
	GATCGTGA	bZIP (Gr. A)	56 (13.61), 5e-005
	TGCATGTA	bZIP (Gr. A)	50 (13.95), 3e-005
GCN4 motif-like	TAAACAAA	bZIP (RISBZ1, Gr. G)	94 (12.79), 1e-004
	CCAAACAA	bZIP (RISBZ1, Gr. G)	69 (11.71), 1e-004
	AGAAAGTG	bZIP (RISBZ1, Gr. G)	50 (12.42), 5e-004
HY-5 binding site-like	ATTTGATTT	bZIP (Gr. H)(HY5)	81 (13.98), 5e-005
G-box-like	GATCGTGA	bZIP (Gr. G) (GBF3, 5, 6)	56 (13.61), 5e-005
	TGCATGTA	bZIP (Gr. G) (GBF3, 5)	50 (13.95), 3e-005
CAMTA5 binding site-like	GATCGTGA	bZIP (CAMTA5)	56 (13.61), 5e-005
ABRE-like (DPBF binding site-like)	GATCGTGA	bZIP (DPBF-3)	56 (13.61), 5e-005
	TGCATGTA	bZIP (DPBF-3) (Opaque-2)	50 (13.95), 3e-005
ABRE-like	GATCGTGA	ABI5 (bZIP)	56 (13.61), 5e-005
ABRE-like	GATCGTGA	ABF2 (bZIP)	56 (13.61), 5e-005
ERE-like	GCTCCATC	ERF/RAP 2.4 (Gr. III)	69 (13.21), 8e-005
JA response element-like	ATTTGATTT	ERF (Gr. VI, VIII, IX)	81 (13.98), 5e-005
ABRE-like	AAGTCAAA	ERF (Gr. VI, VIII, IX)	69 (13.75), 4e-004
	CAAGCTGC	ABI3/V1P1 (B3 domain of AP2/ERF)	69 (12.41), 1e-004
	GCATGGGC		69 (11.61), 5e-005
	TGCATGTA	FUS3, (Similar to VP1/ABI3-like proteins)	50 (13.95), 3e-005
B3-binding site-like	AGAAAGTG	FUS3, (Similar to VP1/ABI3-like proteins)	
		AP2/B3-like	50 (12.42), 5e-004
Ethylene-insensitive 3 binding site-like	TCTTCCAT	EIL3 (EIN3)	50 (13.56), 6e-005
Aux-RE-like	TAAACAAA	ARF1	94 (12.79), 1e-004
	CCAAACAA	ARF1	69 (11.71), 1e-004
	TTTGTATTTA	ARF1	63 (13.49), 2e-004
W-box-like	ATTTGATTT	WRKY (Gr. I, IIa, IIc, III)	81 (13.98), 5e-005
	AAGTCAAA	WRKY (Gr. I, IIa, IIc, III)	69 (13.75), 4e-004

(Continued)

TABLE 2 | Continued

<i>Cis</i> -elements	Motifs	Associated TFs	%(TIC), <i>E</i> -value*
Zinc-finger-binding site-like	ATTTGATTT	ZnF (ZCT1, ZCT2, ZCT3)	81 (13.98), 5e-005
	GATCGTGA	ZnF (GAL3)	56 (13.61), 5e-005
Zing finger binding site-like	TAAACAAA	ZnF (C2H2-type)	94 (12.79), 1e-004
	CCAAACAA	ZnF (C2H2-type)	69 (11.71), 1e-004
IDD binding site-like	TAAACAAA	IDD2, 4, 5, 7 (ZnF)	94 (12.79), 1e-004
	CCAAACAA	IDD2, 4, 5, 7 (ZnF)	69 (11.71), 1e-004
DRE-like	CAAGCTGC	NAC (5, 45, 62, 71, 96)	69 (12.41), 1e-004
	TGAAGCTT	VND (2, 3, 4, 6) SND3	69 (12.23), 6e-005
	TCATGGGC	NAC (45, 58)VND 6	69 (11.61), 5e-005
	AAGCTGAG	NAC (57, 58, 87, 92)	69 (11.15), 9e-005
	GATCGTGA	CUC3	56 (13.51), 5e-005
	AGAAAGTG	NAC (45)	50 (12.42), 5e-004
	TGCATGTA	NAC (46, 47, 55, 58, 92)	50 (13.95), 3e-005
		NAP(NAC-like, activated by AP ₃ /P ₁) NST1 (NAC) ATAF1NAC (5, 45, 62, 7, 96) VND (2, 3, 4, 6) NAC (46, 57, 78) CUC1, CUC2	
AAAAG/CTTTT-element-like	TATTAAGA	DOF (DOF1/4/11/22)	63 (15.86), 5e-005
	AGAAAGTG	OBP3	50 (12.42), 5e-004
Heat shock binding factor element-like		DOF-type zinc finger	
	AGAAAGTG	HSF (HSFB2A)	50 (12.42), 5e-004
S2-binding site-like	AAACCACA	AS2 (LBD16, 19)	81 (11.63), 3e-004
	GCTCCATC	AS2 (LBD13)	69 (13.21), 8e-005
	GATCGTGA	AS2 (LBD2, 19)	56 (13.61), 5e-005
TCP binding site-like	CAACCACA	TCP (TCP 15, 16)	81 (11.63), 3e-004
	GCTCCATC	TCP (TCP3)	69 (13.21), 8e-005
HD-ZIP binding site-like	CAACCACA	HD-ZIP (ATHB7)	81 (11.63), 3e-004
	ATTTGATTT	HD-ZIP (ATHB6, 7, 12 and 13)	81 (13.98), 5e-005
	TGCATGTA	HD-ZIP (ATHB4)	50 (13.95), 3e-005
ARR10-binding element-like	ATTTGATTT	ARR-B (ARR10)	81 (13.98), 5e-005
CArG box-binding site-like	TAAACAAA	MADS (AGL 6, 15)	94 (12.79), 1e-004
	ATTAGCG	MADS (AGL 6, 15)	65 (12.58), 6e-004
	TCTCCAT	MADS (AGL 16)	50(13.56), 6e-005
	AGAAAGTG	MADS box (AGL 6, 15, and 16)	50 (12.42), 5e-004
GATA binding site-like	CAACCACA	GATA1	81 (11.63), 3e-004
	CCAACCAA	GATA1	69 (11.71), 1e-004
E-box-like	GATCGTGA	BES/BZR (BES1/BZR1)	56 (13.61), 5e-005
	TGCATGTA	BES/BZR (BES/BZR1 homologue 2 and 3)	50 (13.95), 3e-005
E-box-like/G-box-like	GATCGTGA	bHLH (Gr. III, VII)	56 (13.61), 5e-005
	TGCATGTA	bHLH (Gr. III, VII)	50 (13.95), 3e-005
MBF1 binding element	TCCTCCTC	MBF1 (MBF1c)	56 (13.42), 2e-004
BPC-binding site-like	TTTCTCTC	BPC (BPC1, 6)	81 (12.26), 3e-004
	GTTCTCTC	BPC (BPC1)	81 (12.26), 3e-004
	AGAAAGTG	BPC (BPC1)	50 (12.42), 5e-004
DBP-binding site-like	TATTAAGAAA	DBP	63 (15.86), 5e-005
	TTAATTTTT	DBP	56 (15.46), 1e-004
TATA-box-like	GTAATTATA	TBP	56 (15.47), 2e-005

% = percent occurrence among all upregulated genes, TIC = total information content of homology, *E*-value* = *E*-value of homology with promoter database entry.

TABLE 3 | Potential putative *cis*-elements identified in the promoters of upregulated genes in extremely tolerant genotypes (8391 and 8753).

<i>Cis</i> -elements	Motifs	Associated TFs	% (TIC), <i>E</i> -value*
AT-hook/PE1-like	AAAAATAT ACAAAAA	MYB (PF1) MYB (PF1)	67 (13.84), 3e-004 52 (16.17), 0e+000
GT-element-like	CATTTGTT AGACGTGG	MYB (GT-3) MYB (GT-3a)	63 (11.57), 6e-005 63 (10.82), 9e-005
GARE-like	CATTTGTT	MYB (R1, R2R3)	63 (11.57), 6e-005
MYB-box-like	TGCTACTG AGAACATAG CATTTGTT ACAAAAA	MYB (R2R3) MYB (R2R3) MYB (R2R3) MYB (R2R3)	81 (11.45), 2e-005 70 (12.40), 2e-004 63 (11.57), 6e-005 52 (16.17), 0e+000
As-1/ocs-like	AAATTTGA CATTGTGA TTGAAAAAT AGACGTGG	bZIP (Gr. D, I, S) bZIP (Gr. D, S) bZIP (Gr. D, S) bZIP (Gr. D, I, S)	67 (13.30), 4e-004 67 (12.12), 9e-005 63 (14.06), 2e-004 63 (10.82), 9e-005
C-box-like/G-box-like	AGACGTGG AGACGTGG AGTCGTGG	bZIP (Gr. A, B, H) bZIP (Gr. G) (GBF 1, 3, 5, 6) bZIP (Gr. G) (GBF 6) bZIP (Gr. G), (CPRF5, CPRF6, CPRF7)	63 (10.82), 9e-005 63 (10.82), 9e-005 63 (10.82), 9e-005 63 (10.82), 9e-005
HY-5 binding site-like	AGACGTGG	bZIP (Gr. H) (HY5)	63 (10.82), 9e-005
GCN4 motif	CATTTGTT ACAAAAA	bZIP (RISBZ1, Gr. G) bZIP (RISBZ1, Gr. G)	63 (11.57), 6e-005 52 (16.17), 0e+000
CAMTA5 binding site-like	AGTCGTGG	bZIP (CAMTA5)	63 (10.82), 9e-005
ABRE-like (DPBF binding site-like)	AGACGTGG	bZIP (DPBF-3)	63 (10.82), 9e-005
ABRE-like	AGACGTGG AGTCGTGG	ABI5 (bZIP) ABI5 (bZIP)	63 (10.82), 9e-005 63 (10.82), 9e-005
ABRE-like	AGACGTGG AGTCGTGG	ABF1, ABF2 (bZIP) ABF2 (bZIP)	63 (10.82), 9e-005 63 (10.82), 9e-005
ERE-like	AGACGTGG AGTCGTGG	ERF/RAP2.3, (Gr. III) ERF/RAP2.3, RAP2.6 RAP2.10 (Gr. III), ERF (Gr. VII)	63 (10.82), 9e-005 63 (10.82), 9e-005
DRE/CRT-like	AGTCGTGG	ERF/DDF1, CBF3, CBF4, (Gr. III, IV)	63 (10.82), 9e-005
ABRE-like	AGAAAGTA	AP2/B3 (Gr. II) Related to RAV2	63 (12.66), 8e-005
JA response element-like	AAATTTGA	ERF (Gr. VI, VIII, IX)	67 (13.30), 4e-004
Ethylene-insensitive 3 binding site-like	AGAACATAG	EIL3 (EIN3)	70 (12.40), 2e-004
Aux-RE-like	CATTTGTT TTGAAAAAT	ARF1 ARF16	63 (11.57), 6e-005 63 (14.06), 2e-004
W-box-like	TTGAAAAAT	WRKY (Gr. I, IIa, IIc, III)	63 (14.06), 2e-004
Zinc finger binding element-like	TCAAATTAA	ZnF (ZCT1, ZCT2, ZCT3)	59 (13.81), 1e-005
Zinc finger binding site-like	TTGAAAAAT AGAAAGTA GAAATCCT TAGTAGTA	ZnF (C2H2-type) ZnF (C2H2-type) ZnF (C2H2-type) ZnF (C2H2-type)	63 (14.06), 2e-004 63 (12.66), 8e-005 63 (11.69), 2e-004 52 (13.63), 2e-005
IDD binding site-like	TTGAAAAAT	IDD 2, 5 (ZnF)	63 (14.06), 2e-004

(Continued)

TABLE 3 | Continued

<i>Cis</i> -elements	Motifs	Associated TFs	% (TIC), <i>E</i> -value*
DRE-like	TGCTACTG CATTGTGA AGACGTGG AGTCGTGG AGACGTGG AGTCGTGG	NAC (62, 96) NAC (46) NAC (34, 42, 46, 47, 55, 70, 94) NAC (45, 46, 47) NAC (34, 42, 45, 55, 58) NAC (NAM) NAP (NAC-like, activated by AP ₃ /P ₁) NAP (NAC-like, activated by AP ₃ /P ₁)	81 (11.45), 2e-005 67 (12.12), 9e-005 63 (10.82), 9e-005 63 (10.82), 9e-005 63 (10.82), 9e-005 63 (10.82), 9e-005
ATAF1-binding site-like	AGTCGTGG	ATAF1(NAC)	63 (10.82), 9e-005
AAAAG/CTTTT-element-like	AGAAAGTA	DOF (DOF 4.5)	63 (12.66), 8e-005
Heat shock binding factor element-like	AGAACATAG CATTGTGA GAAATCCT	HSF (HSFB2A HSF) HSF (HSFB2A) (HSFA8)	70 (12.40), 2e-004 67 (12.12), 9e-005 63 (11.69), 2e-004
S2-binding site-like	CATTGTGA AGACGTGG AGTCGTGG ACAAAAA	AS2 (LBD16, 18, 19) AS2 (LBD2) AS2 (LBD23) AS2 (LBD16)	67 (12.12), 9e-005 63 (10.82), 9e-005 63 (10.82), 9e-005 52 (16.17), 0e+000
HD-ZIP binding site-like	AAAATTAG	ATHB3	55 (12.95), 5e-005
ARR14-binding element-like	AGAACATAG	ARR-B (ARR14)	70 (12.40), 2e-004
CArG box-binding site-like	ACAAAAA AGAAAGTA AAGATTGCA	MADS (AGL 15, 16) MADS (AGL 6) MADS (AGL 6)	52 (16.17), 0e+000 63 (12.66), 8e-005 74 (12.50), 2e-004
E-box-like	AGACGTGG	BES/BZR (BES1/BZR1)	63 (10.82), 9e-005
E-box-like/G-box-like	AGACGTGG AGTCGTGG	bHLH (Gr. III, VII) bHLH (Gr. III, VII)	63 (10.82), 9e-005 63 (10.82), 9e-005
E-box-like	AGACGTGG	BEE2 (bHLH) BIM1, BIM3 (bHLH)	63 (10.82), 9e-005
BBRE-element-like/G-box-like	AGACGTGG	BAM8 (bHLH): BAMs	63 (10.82), 9e-005
BRRE-element/G-box-like	AGACGTGG	BEH4 (bHLH)	63 (10.82), 9e-005
DBP-binding site-like	TCAAATTAA	DBP	59 (13.81), 1e-005
TATA-box-like	TAAATATA	TBP	56 (13.75), 1e-005

% = percent occurrence among all upregulated genes, TIC = total information content of homology, *E*-value* = *E*-value of homology with promoter database entry.

MYB Regulatory Module

The promoter analysis of common upregulated genes in the tolerant, highly tolerant, and extremely tolerant genotypes identified high enrichment of a number of MYB-associated putative *cis*-elements such as AT-hook/PE1-like, GT-element-like, gibberellic acid response element (GARE)-like, and MYB-box-like. The MYB-box-like elements were highly enriched in the highly tolerant genotypes and, besides the promoters, also possess MYB-box related-like elements associated with MYB-like (CDC5) and MYB-like (CCA1, CDA-1) TFs (Tables 1-3). These elements are most likely

TABLE 4 | List of upregulated transcription factors with potential significance to the identified putative *cis*-elements among upregulated genes in five genotypes from diverse background in response to submergence tolerance.

TF Family	Locus_ID (Annotation)*	Fold increase**
MYB	Os12g0125000 (MYB-like DNA-binding domain-containing protein)	5.90
	Os11g0128500 (MYB-like DNA-binding domain-containing protein)	5.57
	Os05g0553400 (Putative MYB-related transcription factor)	5.14
	Os01g0298400 (Putative typical P-type R2R3 MYB protein)	4.55
	Os12g0567300 (MYB transcription factor domain-containing protein: R2R3 MYB)	3.24
	Os01g0874300 (Putative MYB-related protein; Putative MYB2)	3.20
	Os05g0140100 (R2R3 MYB transcription factor)	2.44
bZIP	Os11g0152700 (bZIP transcription factor: transcription factor HBP-1)	5.48
	Os12g0547600 (Calmodulin-binding protein, putative, expressed)	3.10
	Os05g0129300 (bZIP transcription factor)	2.51
ERF	Os01g0968800 (DREB transcription factor like)	Infinite
	Os06g0127100 (Dehydration-responsive element-binding protein 1C)	Infinite
	Os01g0868000 (AP2/ERF transcription factor like)	6.86
	Os09g0522100 (Similar to Dehydration-responsive element-binding protein 1H)	6.44
	Os01g0140700 (Similar to RAV family protein: AP2/ERF and B3 domain-containing protein)	6.16
	Os02g0677300 (Similar to CRT/DRE binding factor 1)	4.71
	Os02g0654700 (AP2/ERF family protein)	4.65
	Os02g0656600 (Similar to DRE binding factor 2B)	4.52
	Os05g0549800 (Similar to DNA-binding protein RAV1)	4.17
	Os03g0184500 (B3 domain-containing protein: ABVP1 transcription factor)	3.58
	Os01g0693400 (RAV family protein)	2.89
ARF	Os12g0601400 (Auxin-responsive protein IAA31)	2.51
WRKY	Os05g0583000 (Similar to WRKY transcription factor 8)	Infinite

(Continued)

TABLE 4 | Continued

TF Family	Locus_ID (Annotation)*	Fold increase**
ZnF	Os05g0537100 (WRKY transcription factor 10, WRKY transcription factor 7)	4.27
	Os01g0750100 (Similar to WRKY transcription factor 13)	2.71
	Os03g0437200 (C2H2-type zinc finger protein, abscisic acid-induced antioxidant defense, Water stress and oxidative stress tolerance)	6.82
	Os03g0820300 (C2H2 transcription factor protein)	6.16
	Os12g0113700 (Zinc finger, C3HC4 type family protein)	6.04
	Os02g0646200 (Zinc finger, B-box domain-containing protein)	4.43
	Os10g0456800 (CHY zinc finger family protein)	2.81
	Os01g0303600 (Zinc finger, RING/FYVE/PHD-type domain-containing protein)	3.82
	Os06g0340200 (Zinc finger, RING-CH-type domain-containing protein)	3.33
	Os03g0329200 (Zinc finger CCCH domain-containing protein 23)	3.27
NAC	Os03g0764100 (Zinc finger transcription factor ZF1)	2.60
	Os09g0486500 (Zinc finger A20 and AN1 domain-containing stress-associated protein 1)	2.49
	Os05g0128200 (Zinc finger CCCH domain-containing protein 33)	2.21
	Os11g0154500 (No apical meristem (NAM) protein domain-containing protein; NAC-domain-containing protein 90)	5.83
	Os03g0815100 (Similar to OsNAC6 protein)	4.94
	Os01g0884300 (NAC domain-containing protein 6)	3.50
	Os07g0684800 (Similar to NAM/CUC2-like protein)	3.33
	Os07g0225300 (OsNAC3 protein; NAC domain-containing protein 67)	3.27
MADS-box	Os04g0580700MADS-box transcription factor 17	6.07
HSF	Os08g0471000 (Heat stress transcription factor B-4a, HSF20)	Infinite
	Os09g0526600 (Heat stress transcription factor B-2c, HSF 3)	4.51
	Os09g0456800 (Heat stress transcription factor B-1)	4.05
	Os02g0232000 (Similar to Heat shock transcription factor 29, HSF 5)	3.67

(Continued)

TABLE 4 | Continued

TF Family	Locus_ID (Annotation)*	Fold increase**
PhD-finger	Os03g0302200 (PHD-finger family protein)	2.65
JAZ	Os10g0391400 (Jasmonate ZIM-domain (JAZ) protein, Negative regulation of JA signal transduction pathway)	7.81
	Os03g0180800 (Jasmonate ZIM-domain protein 3)	6.04
	Os08g0428400 protein (ZIM transcription factor; Jasmonate ZIM-domain protein 9)	2.49
HD ZIP	Os05g0129700 (Homeobox protein knotted-1-like 10)	4.82
	Os06g0140400 (Homeobox-leucine zipper protein HOX28)	4.50
	Os03g0198600 (Homeodomain-leucine zipper transcription factor)	4.46
	Os06g0140700 (Homeobox-leucine zipper protein HOX2)	3.49
	Os03g0188900 (Homeobox-leucine zipper protein HOX13)	3.21
	Os09g0528200 (Similar to Homeobox-leucine zipper protein HOX6Homeobox-leucine zipper protein HOX6)	2.83
bHLH	Os01g0773800 (Basic helix-loop-helix protein 185)	Infinite
	Os03g0188400 (Basic helix-loop-helix protein)	3.84
	Os07g0628500 (Basic helix-loop-helix dimerization region bHLH domain-containing protein)	3.31
	Os03g0135700 Basic helix-loop-helix transcription factor	3.16
	Os07g0143200 (Phytochrome-interacting bHLH factor)	2.44
G-box binding protein	G-box binding protein; G-box binding protein-like (B12D-like protein); Os06g0246000 protein	2.70

*Information based on RAP-DB (<http://rapdb.dna.affrc.go.jp/>); **based on RNA-seq data of Hsu and Tung (2017).

associated with the upregulation of a number of MYB genes (Table 4) such as *Os12g0125000* (MYB-like DNA-binding domain-containing protein), *Os11g0128500* (MYB-like DNA-binding domain-containing protein), *Os05g0553400* (Similar to MYB-related TF), *Os01g0298400* (Putative typical P-type R2R3 MYB protein), *Os12g0567300* (MYB TF domain-containing protein, R2R3 MYB), *Os01g0874300* (Putative MYB-related protein, MYB2), and *Os05g0140100* (R2R3 MYB TF) (Table 4).

It has been highlighted that MYB TFs represent a major protein in rice and are involved in the transcriptional regulation

of many developmental processes as well as abiotic and biotic stress conditions (Katiyar et al., 2012). Additionally, it has already been shown both experimentally and *in silico* analysis that MYB TFs that bind to MYB-box/GT-element-like elements play a key role in the transcriptional regulation of rice during germination under submergence and anoxia (Dolferus et al., 2003; Mohanty et al., 2012; Lakshmanan et al., 2014). In this analysis, MYB-box-like elements are more highly enriched compared to GARE-like elements. Under flooding/submergence/anoxic conditions, the plant hormone gibberellic acid (GA) plays a major role in rice and other plants such as barley (Gubler et al., 2002; Kaneko et al., 2002). It activates the endosperm reserve to aleurone layers for the induction of enzyme α -amylases for the hydrolysis of starch, protein, and cell wall reserve (Woodger et al., 2003; Lee et al., 2014). Under submerged/flooding conditions, α -amylases play a major role in providing sugar substrates by hydrolyzing starch for glycolysis and alcohol fermentation to generate ATP. Recently, Abdulmajid et al. (2020) have screened favorable rice genotypes for coleoptile elongation length sensitivity to exogenous gibberellin under submerged conditions. However, in this analysis, pyrimidine-box-like elements and GARE-like elements are less enriched compared to MYB-box-like elements, which are associated with MYB (R1, R2R3). A similar pattern was also identified in our previous analysis for the transcriptional regulation of coleoptile germination and elongation of the *japonica* rice under anoxia (Mohanty et al., 2012). It appears that submergence tolerance for rice germination may not be completely mediated by GA. Recently, it has been shown that two MYB TFs that bind to the same *cis*-element regulate the on/off switch of α -Amy expression (Chen et al., 2019). MYBS1 activates the expression of α -Amy during sugar starvation and promotes nuclear import of MYBS1, whereas MYBS2 behaves in the opposite manner during sugar provision. They have also shown that there is no enhancement of submergence tolerance in rice seedlings when the expression of MYBS2 was reduced. However, α -Amy is necessary for growth of rice seedlings under submergence as reduced seedling growth was observed in MYBS2-Ox lines under submergence. High enrichment of MYB-box-like element associated with R2R3 MYB and upregulation of R2R3 MYB genes in the tolerant genotypes suggest that abscisic acid (ABA) could also be playing a role together with other hormones in the germination and coleoptile elongation of rice under submergence.

bZIP Regulatory Module

As-1/ocs-like, GCN4 motif, and CAMTA5 binding site-like were highly enriched in the upregulated genes of tolerant, highly tolerant, and extremely tolerant genotypes (Tables 1-3). However, upregulated genes of highly tolerant and extremely tolerant genotypes possess specific ABRE-like elements associated with bZIP (Gr. A), bZIP (Gr. G) (GBF 3, 5, and 6), bZIP (DPBF-3), ABI5 (bZIP), and ABF2 (bZIP) and C-box-like/G-box-like and HY-5 binding site-like elements associated with GBF and HY5, respectively (Tables 2, 3). These binding sites could be associated with the upregulation of genes such as *Os12g0547600* (Calmodulin-binding protein, putative, expressed), *Os11g0152700* (bZIP transcription factor

79; transcription factor HBP-1), and *Os05g0129300* (bZIP transcription factor) (Table 4). bZIP TFs play a key role in plant growth, development, and abiotic and biotic stress conditions in *Arabidopsis* and rice (Zhang et al., 2015; Yang et al., 2019). Although the role of bZIP in response to abiotic stresses has been well studied in *Arabidopsis*, a few cases have been characterized in rice (Tang et al., 2012, 2016; Zhang et al., 2017; Yang et al., 2019). Moreover, substantial enrichment of bZIP-associated elements suggest the significance of ABRE-like/CAMTA5/C-box/G-box-like elements in response to ABA signaling in response to submergence. The role of bZIP in response to submergence/flooding stress in rice is not well studied yet. In rice, *OsZIP45*, an ortholog of maize GBF1, was shown to be induced by hypoxia (de Vetten and Ferl, 1995). Later, *OsABF1*, an ABA-responsive element (ABRE) binding bZIP TF, has been shown to be induced during different abiotic stresses, such as anoxia, drought, cold, salinity, and ABA in rice seedlings (Hossain et al., 2010). ABA plays a key role in plant physiology, development, seed maturation, dormancy, and responses to a number of abiotic stress conditions such as drought, cold, and salt (Dar et al., 2017). Promoter regions of many genes possess ABRE-responsive elements that are responsible for ABA-dependent regulation. These elements interact with various ABA-responsive TFs that regulate ABA response particularly (Kim et al., 2004). Recently, induction of bZIP in response to short- and long-term hypoxia in tomato root has been observed (Safavi-Rizi et al., 2020). In the previous analysis, we identified both high enrichment of As1/ocs-like element, ABRE-like, G-box-like, GCN4-like, and CAMTA5-like in the promoters of genes upregulated in anoxia (Mohanty et al., 2012), glycolysis and fermentation (Lakshmanan et al., 2014), and in the wild-type rice cultivar (Kinmaze) during germination and coleoptile elongation (Mohanty et al., 2016). In rice, all genotypes tolerant to submergence also showed upregulation of bZIP TF and a calmodulin-binding protein. Identification of CAMTA5 binding site-like elements associated with CAMTA5 TF proposes that it could be involved in regulating auxin transport and homeostasis (Galon et al., 2010). High enrichment of ABA-regulated bZIP TFs suggests a cross-talk between sugar and ABA-signaling during germination and coleoptile elongation under submergence. Besides, there could be some mutual enhancement between ABA and ethylene and ethylene could repress ABA activity depending on the requirement.

Interestingly, binding sites associated with bZIPs such as HY5, DPBF-3, and GBF 3, 5, and 6 were identified in both highly tolerant and extremely tolerant rice genotypes. HY5 is a key integrating factor for light and ABA pathways, and it stimulates ABA signaling pathway by binding to the promoter of ABI5 (Chen et al., 2008). It also regulates cell elongation and proliferation besides its other roles in plant growth and development (Jing et al., 2013). The presence of high enrichment of ABRE-like elements associated with different bZIP TFs especially in both highly tolerant and extremely tolerant genotypes suggests a possible temporal role of ABA signaling in response to longer coleoptile elongation.

ERF Regulatory Module

A number of putative *cis*-elements associated with different groups of ERF TFs were identified in the promoter regions of the upregulated genes in tolerant, highly tolerant, and extremely tolerant genotypes in response to submergence (Tables 1–3). Genes in tolerant genotypes were highly enriched with GCC-box-like elements associated with Groups VI, VIII, and IX ERF TFs; ERE-like elements associated with Groups I, IV, and VII TFs; jasmonic acid response element (JARE)-like elements associated with Gr. VI, VIII, and IX ERF TFs; and RAV-1-like element associated with ERF/RAV1 (Gr. II) TFs (Table 1). There is also enrichment of ABRE-like elements associated with Group II ABI3 AP2/B3 related to RAV TF, CRT/DRE-like elements associated with Group IV ERF TF and ABRE-binding site-like associated with ABR1 TF (Table 1). The enrichment of *cis*-elements related to the ERF group is less compared to their presence in tolerant genotypes (Tables 2, 3). GCC-box-like elements are also absent in these groups (Tables 2, 3). However, these genes possess binding sites associated with ERF (Gr. VI, VIII, and IX), ABI3 AP2/B3 related to RAV and ERF/RAP 2.4, and ERF/RAP 2.3, 2.6, and 2.10 (Tables 2, 3). The enrichment pattern of these elements could be correlated with the upregulation of a number of ERF genes such as *Os01g0968800* (DREB transcription factor-like), *Os06g0127100* (Dehydration-responsive element-binding protein 1C), *Os01g0868000* (AP2/ERF transcription factor-like), *Os09g0522100* (Dehydration-responsive element-binding protein 1H), *Os01g0140700* (Similar to RAV2: AP2/ERF and B3 domain-containing protein), *Os02g0677300* (Similar to CRT/DRE binding factor 1), *Os02g0654700* (AP2/ERF family protein, abiotic stress response), *Os02g0656600* (DRE binding factor 2B), *Os03g0184500* (B3 domain-containing protein, ABIVP1 transcription factor), and *Os01g0693400* (RAV family protein) (Table 4).

In rice, ERF VII plays a key role in flooding, hypoxia, and submergence conditions (Bailey-Serres et al., 2012; Bui et al., 2015; Gibbs et al., 2015). *SUBMERGENCE 1A* (*Sub1A*), an ERF-type TF, regulates submergence/flooding tolerance in rice. Although the *japonica* cultivar “Nipponbare” lacks this gene (Fukao et al., 2006; Xu et al., 2006), it germinates and elongates its coleoptile under anoxia and submergence conditions. Hence, this shows that there could be some other mechanism that is independent of *SUB1A* in these genotypes to tolerate anoxia and submergence (Lee et al., 2009). An earlier evidence by Ishizawa and Esashi (1988) showed that ethylene is important for the transport of sucrose from the scutellum to the coleoptile during germination and coleoptile elongation of rice seeds. Recently, it has been reported that ethylene plays an important role in the signal transduction pathway dependent on ethylene and oxygen under hypoxia/submergence conditions. In contrast, two ERFs such as *SK1* and *SK2* are involved in submergence adaptation in deep-water rice by rapidly elongating the internodes through the action of GA response (Hattori et al., 2009).

In *Arabidopsis*, members of ERF-VII such as ERF71/HRE2, ERF72/RAP2.3, ERF73/HRE1, ERF74/RAP2.12, and ERF75/RAP2.2 are induced under limited oxygen conditions to regulate many hypoxia-induced genes involved in fermentation,

TABLE 5 | Potential putative *cis*-elements identified in the promoters of upregulated genes in two intermediately tolerant genotypes (F291 and F274-2a).

<i>Cis</i>-elements	Motifs	Associated TFs	% (TIC), <i>E</i>-value*
AT-hook/PE1-like	AGAAAAATG	MYB (PF1)	58 (14.34), 2e-004
GT-element-like	TTTGTTC	MYB (GT-3)	71 (13.29), 2e-004
	CATGTGTG	MYB (GT-3a)	58 (12.27), 2e-004
	TTTACTCT	MYB (GT1/GT2)	50 (12.92), 1e-004
	CTCTTAAA	MYB (GT1/GT2)	50 (12.19), 0e+000
GARE-like	TTTGTTC	MYB (R1, R2R3)	71 (13.29), 2e-004
MYB-box-like	CCACCATG	MYB (R2R3)	79 (11.60), 3e-005
	TTTGTTC	MYB (R2R3)	71 (13.29), 2e-004
	GCAAGGTG	MYB (R2R3)	67 (12.65), 3e-005
	GGTTCGTC	MYB (R2R3)	58 (11.20), 6e-005
	GATGGTATT	MYB (R2R3)	54 (13.12), 2e-004
MYB-box related-like	GCCACCAT	MYB-like	63 (12.88), 3e-004
As-1/ocs-like	GCAAGGTG	bZIP (Gr. D, I, S)	67 (12.65), 3e-005
	AAGTTTGA	bZIP (Gr. D, I, S)	63 (12.83), 1e-004
	CATGTGTG	bZIP (Gr. D, I, S)	58 (12.27), 2e-004
GCN4 motif	TTTGTTC	bZIP (RISBZ1, Gr. G)	71 (13.29), 2e-004
	AGAAAGTG	bZIP (RISBZ1, Gr. G)	54 (13.04), 6e-005
G-box-like	GCAAGGTG	bZIP (Gr. G) (GBF 3, 5)	67 (12.65), 3e-005
	CATGTGTG	bZIP (Gr. G) (GBF 3, 5, 6)	58 (12.27), 2e-004
ABRE-like	GCAAGGTG	bZIP (DPBF-3)	67 (12.65), 3e-005
(DPBF binding site-like)	CATGTGTG	bZIP (DPBF-3)	58 (12.27), 2e-004
		(Opaque-2)	
ABRE-like	GCAAGGTG	ABI5 (bZIP)	67 (12.65), 3e-005
	CATGTGTG	ABI5 (bZIP)	58 (12.27), 2e-004
ABRE-like	CATGTGTG	ABF2 (bZIP)	58 (12.27), 2e-004
GCC-box-like	GCCACCAT	ERF (I, IV, VII, X)	63 (12.88), 3e-004
	GCCGAAA	ERF (I, IV, VII, X)	50 (12.53), 3e-004
ERE-like	GCAAGGTG	ERF/RAP2.1, RAP2.4,	67 (12.65), 3e-005
	GCCACCAT	RAP2.9, RAP2.10	63 (12.88), 3e-004
	GGTTCGTC	ERF/RAP2.3, RAP2.6 (Gr. III)	58 (11.20), 6e-005
	GCCGAAA	ERF/RAP2.3, RAP2.6 (Gr. III)	50 (12.53), 3e-004
		RAP2.11(Gr. III)	
CRT/DRE-like	GCCACCAT	ERF (Gr., III, IV)	63 (12.88), 3e-004
	GCCGAAA	ERF (Gr., III, IV)	50 (12.53), 3e-004
JA response element-like	AAGTTTGA	ERF (Gr. VI, VIII, IX)	63 (12.83), 1e-004
		ERF (Gr. VI, VIII, IX)	69 (13.75), 4e-004
ABRE-like	CATGTGTG	FUS3, (Similar to VP1/ABI3-like proteins)	58 (12.27), 2e-004
		FUS3, (Similar to VP1/ABI3-like proteins)	69 (11.61), 5e-005
		FUS3, (Similar to VP1/ABI3-like proteins)	50 (13.95), 3e-005
ABRE-like	AGAAAGTG	AP2/B3 (Gr. II)	54 (13.04), 6e-005
	AGAAAAATG	Related to RAV2	58 (14.34), 2e-004
		AP2/B3 (Gr. II)	
		Related to RAV2	
ABR-binding site-like	GCGGGAGA	AP2-like ABA repressor 1 (ABR1)	54 (11.84), 9e-005
Ethylene-insensitive 3 binding site-like	TTTGTTC	EIL3 (EIN3)	71 (13.29), 2e-004
	GCCGAAA	EIN2	50 (12.53), 3e-004
Aux-RE-like	TTTGTTC	ARF1	71 (13.29), 2e-004
	AGAAAAATG	ARF16	58 (14.34), 2e-004
W-box-like	TTTGTTC	WRKY (Gr. I, IIa, IIc, III)	71 (13.29), 2e-004
Zinc-finger-binding site-like	GCAAGGTG GGTTCGTC	ZnF (C2H2-type)	71 (13.29), 2e-004
	AGAAAGTG	ZnF (C2H2)	58 (11.20), 6e-005
	GATGCGATT	ZnF (C2H2)	54 (13.04), 6e-005
	GCCGAAA	ZnF (CCCH-type)	54 (13.12), 2e-004
		ZnF (PhD)	50 (12.53), 3e-004
		ZnF (PhD)	
IDD binding site-like	TTTGTTC	IDD 2, 4, 5, 7 (ZnF)	71 (13.29), 2e-004
	AGAAAAATG	IDD 4, 5 (ZnF)	58 (14.34), 2e-004

(Continued)

TABLE 5 | Continued

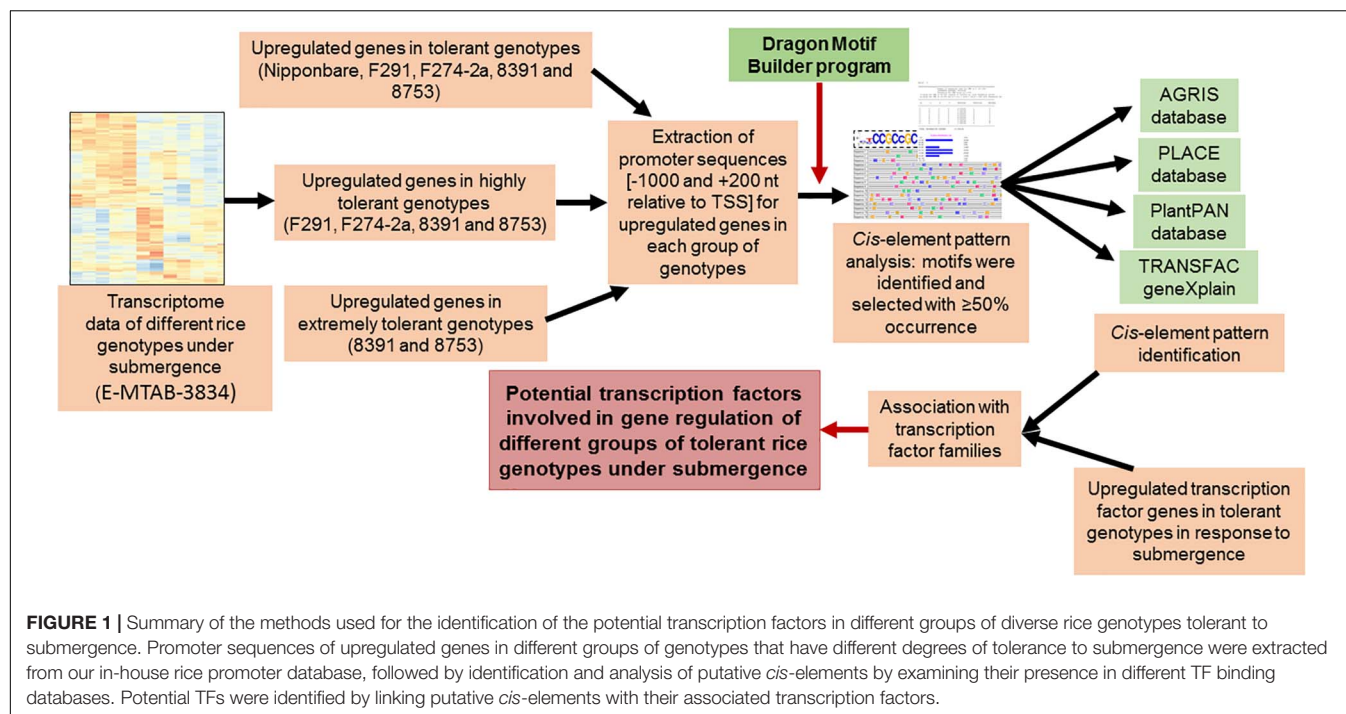
Cis-elements	Motifs	Associated TFs	% (TIC), E-value*
DRE-like	CCACCATG	NAC (16, 92)	79 (11.60), 3e-005
	GGTTCGTC	NAC (46, 55, 58) NAP (NAC-like, activated by AP ₃ /P ₁)	58 (11.20), 6e-005
	CATGTGTG		58 (12.27), 2e-004
	AGAAAAATG	NAM	58 (14.34), 2e-004
	AGAAAGTG	NAC (57, 71, 83, 103)	54 (13.04), 6e-005
	CTCTTAA	CUC1, CUC2, CUC3	50 (12.19), 0e+000
	AGCCGTAG	NAC (5, 11)	50 (12.65), 8e-005
		NAC (5, 11, 20, 28, 45, 50, 58, 62, 71, 75, 96)	
		VND (2, 3, 4, 6) CUC1, CUC2, CUC3	
		NAC (5, 62)	
AAAAG/CTTTT-element-like	ATCCCTTT	DOF (PBF)	67 (11.52), 1e-004
	AGAAAAATG	DOF-type zinc finger	58 (14.34), 2e-004
	AGAAAGTG	DOF (DOF 4)	54 (13.04), 6e-005
		DOF (DOF 5.7)	
Heat shock binding factor element-like	CTCCCCCT	HSF (HSFB2A)	67 (12.51), 2e-004
	AGAAAAATG	HSF (HSFB2A)	58 (14.34), 2e-004
	AGAAAGTG	HSF (HSFB2A)	54 (13.04), 6e-005
S2-binding site-like	CTCCCCCT	AS2 (LBD13)	67 (12.51), 2e-004
	GCCACCAT	AS2 (LBD23)	63 (12.88), 3e-004
	GCCGGA	AS2 (LBD2, 13, 16)	50 (12.53), 3e-004
TCP binding site-like	GGTTCCTC	TCP (3, 20)	58 (11.20), 6e-005
ARR14-binding element-like	GATGCGATT	ARR-B (ARR14)	54 (13.12), 2e-004
CArG box-binding site-like	CCACCATG	MADS (AGL 42, 55)	79 (11.60), 3e-005
	GCCACCAT	MADS (AGL 55)	63 (12.88), 3e-004
	AGAAAAATG	MADS box (AGL 6, 15, 16)	58 (14.34), 2e-004
	AGAAAGTG	MADS box (AGL 6, 15, 16)	54 (13.04), 6e-005
	TTTACTCT	MADS (AGL 16)	50 (12.92), 1e-004
GATA binding site-like	TCTCCAT	GATA1	67 (12.72), 3e-005
	GCAAGGTG	GATA1	67 (12.65), 3e-005
E-box-like	CATGTGTG	BES/BZR (BES1/BZR1 homologue 2, 3, 4)	58 (12.27), 2e-004
E-box-like/G-box-like	GAAGTAAC	bHLH (Gr. III, VII)	63 (11.53), 1e-004
	CATGTGTG AGCCGTAG	bHLH (Gr. III, VII)	58 (12.27), 2e-004
		bHLH (Gr. III, VII)	50 (12.65), 8e-005
BPC-binding site-like	AGAAAAATG	BPC (BPC1)	58 (14.34), 2e-004
	AGAAAGTG	BPC (BPC1)	54 (13.04), 6e-005

% = percent occurrence among all upregulated genes, TIC = total information content of homology, E-value* = E-value of homology with promoter database entry.

sugar metabolism, and ethylene biosynthesis. Besides, it has been shown that RAP2.12 is present in the plasma membrane in an inactive state in the presence of oxygen and then moves to the nucleus under hypoxia/submergence conditions (Kosmacz et al., 2015). This mechanism shows the fast response of the TF to oxygen shortage in order to protect plant cells. In *Arabidopsis*, a higher survival rate under low oxygen condition was observed by overexpressing RAPs (RAP2.2, RAP2.3, and RAP2.12) (Papdi et al., 2015; Yao et al., 2017). In this analysis, GCC-box-like elements associated with ERF (I, IV, VII, and X) and ERE-like elements associated with ERF (Gr. III) such as RAP2.2, RAP2.3, RAP2.6, and RAP2.10 were identified. However, identification of moderate enrichment of RAV1a-like elements can be associated with the expression of *Os05g0549800* (AP2/ERF and B3 domain-containing protein: similar to DNA-binding protein RAV1). In *Arabidopsis*, during seed germination and early seedling

development, RAV1 plays a key role in ABA signaling by repressing the expression of ABI3, ABI4, and ABI5 by binding to the 5'-CAACA-3' motif within the promoter of ABI3, ABI4, and ABI5. These results suggest that RAV1 could also be acting as a repressor of ABA signaling. In addition to ABA inhibition, ethylene also inhibits the biosynthesis of jasmonate (JA) during coleoptile elongation of etiolated rice seedlings (Xiong et al., 2017). Although the role of ERF-VIIs has been well studied in response to submergence, it is still a complex process and the regulatory mechanism is still not clearly understood yet.

The upregulated genes in the tolerant genotypes are highly enriched with ABR1 (AP2-like ABA repressor 1) binding-site-like motifs associated with ABR1 TF belonging to AP2-domain-containing protein group X in rice. TF ABR1 functions as a negative regulator of ABA response during seed germination. The expression of this TF is induced during various abiotic



stresses, such as drought, cold, and salt in *Arabidopsis* (Pandey et al., 2005). In rice, the group X *OsERFs* are also closely related to the ortholog of *ABR1* in *Arabidopsis* (Mishra et al., 2013) and show significant expression during different developmental stages and response to various abiotic stresses. Interestingly, the expression of *AtERF#111/ABR1* in *Arabidopsis* was shoot specific and induced under submergence and hypoxia (van Veen et al., 2016; Yeung et al., 2018). Later, the regulation of *AtERF#111* expression was suggested to be related to mechanical stress during submergence as it was regulated by WRKY18, 33, and 40 (Birkenbihl et al., 2017; Bäuml et al., 2019). Significant induction of TFs WRKY18, WRKY33, and WRKY40 is reported to be induced under both submergence and anoxia as well as wounding stress (Hsu et al., 2013; Tsai et al., 2014; Wang et al., 2015). This shows that *ABR1* related to wounding/pathogen response was also induced during submergence to stimulate the immune response against the threat of wounding or pathogen infection after flooding (Bäuml et al., 2019). This TF could be involved in the immune response mechanism during rice germination and coleoptile elongation in tolerant genotypes in response to submergence.

EIN3/EIL1 Regulatory Module

The promoter analysis of the upregulated genes in all tolerant, highly tolerant, and extremely tolerant genotypes identified the presence of putative EIN3 binding-like elements in 50–70% of the genes (Tables 1–3). This binding site can be associated with EIN3/EIL1 TF. However, there was no expression of this TF in tolerant genotypes. This TF is the master transcriptional regulator of ethylene signaling as it is required for the activation of the ethylene pathway. It regulates the transcription of ethylene-responsive genes under different environmental and

spatiotemporal conditions in *Arabidopsis* (An et al., 2010; Dolgikh et al., 2019). Besides, EIN3/EIL1 controls multiple transcriptional cascades. It also targets genes such as *ERF1*, *PIF3*, and *CBF1/2/3*, which are important regulators during different abiotic stress conditions and developmental processes (Zhang et al., 2011; Shi et al., 2012; Zhong et al., 2012). It has been reported that EIN3 is involved in the regulation of a secondary transcriptional ethylene response that includes TFs such as AP2/ERFs: *ERF1*, *ERF5*, *WRKY14/47*, *PIF3*, *NAC6*, and *RAP2.2* (Chang et al., 2013). EIN3 binding also modulates feedback regulation of the ethylene signaling pathway and integrates between different hormone-mediated pathways. These results suggest that it could be playing a key role in rice germination and coleoptile elongation mainly to activate ethylene signaling and ERF TFs and could also act as a mediator in regulating other hormone signaling and TFs.

ARF Regulatory Module

Aux-RE-like elements associated with ARF TFs are identified in all tolerant, highly tolerant, and extremely tolerant genotypes, and this could be associated with the expression of ARF gene such as *Os12g0601400* (Auxin-responsive protein IAA31) (Tables 1–4). The phytohormone auxin plays a key role in plant growth and development. It has been suggested that the elongation of rice coleoptile under submerged conditions could be cooperatively regulated by the endogenous concentration of ethylene and auxin (Ishizawa and Esashi, 1983). Moreover, it was shown that external IAA addition had a positive effect on the initial elongation of coleoptile (Breviaro et al., 1992), whereas ethylene enhanced the later stage of elongation (Hoson et al., 1992). Auxin mainly promotes cell division. It has been shown that coleoptile in rice elongates rapidly in response to auxin treatment

(Perrot-Rechenmann, 2010). *OsARF*, a rice homolog of the ARF, was found to be positively correlated with auxin-dependent differential growth in rice coleoptiles (Waller et al., 2002). Recently, Wu and Yang (2020) have revealed the involvement of auxin signaling in regulating rice coleoptile elongation as well as regulation of carbohydrate metabolism and secondary metabolism under submergence. Besides, auxin plays a key role in cell division during coleoptile elongation in the *japonica* rice under submergence (Nghie et al., 2020). Differences in auxin transport determine the length of coleoptile while availability of higher auxin level determines the final length of coleoptile under submerged conditions. They have also claimed that the long coleoptile in rice under submergence is due to an increase in auxin transport by the influx carrier AUX1. Besides, it also regulates the expression of other TFs. The presence of Aux-RE-like elements in the promoters of upregulated genes present in all groups and significant expression of ARF gene suggest a role of ARF and auxin signaling in coleoptile elongation under submergence.

WRKY Regulatory Module

Promoter analysis of the upregulated genes identified enrichment of W-box-like elements associated with WRKY TF in all tolerant, highly tolerant, and extremely tolerant genotypes, and those could be associated with the upregulation of WRKY genes such as *Os05g0583000* (Similar to WRKY transcription factor 8), *Os05g0537100* (WRKY transcription factor 10, WRKY transcription factor 7), and *Os01g0750100* (WRKY transcription factor 13) (Tables 1–4). Identification of W-box-like elements also coincides with the previous identification in the genes upregulated during germination and coleoptile elongation in rice (Mohanty et al., 2012; Lakshmanan et al., 2014). WRKY plays a key role in plant growth and development, secondary metabolite biosynthesis, and response to a number of abiotic and biotic stresses (Phukan et al., 2016). Increase in the expression of WRKY has also been shown by microarray data in *Arabidopsis* (Loreti et al., 2005) and rice (Lasanthi-Kudahettige et al., 2007) in response to oxygen deficiency. The role of WRKY in the regulation of submergence tolerance in rice has also been reported by Viana et al. (2018). Also, in transgenic *Arabidopsis* expressing a sunflower WRKY, *HaWRKY76* was found to be tolerant to flooding (complete submergence) by preserving carbohydrates, mainly sucrose and starch, through the repression of fermentation pathways (Raineri et al., 2015). Upregulation of WRKYs at different time points of submergence stress was also observed in maize (Campbell et al., 2016). Similar observations were also identified in *alcohol dehydrogenase 1* (*ADH1*)-deficient mutant of rice where WRKY could be playing a key role in cell survival rather than elongation (Mohanty et al., 2016). Interestingly, W-box elements have been identified in the promoter of upregulated WRKY genes such as *OsWRKY11*, *OsWRKY56*, and *OsWRKY62* during submergence (Viana et al., 2018). This shows the self-regulation of WRKY in response to submergence stress. Overall, it shows that WRKY could be playing a role by self-regulating through an on–off switch depending on the condition in response to submergence.

A number of WRKY TFs are known to be ABA responsive and are involved in ABA signaling pathways. However, in *Arabidopsis*, three WRKY genes such as *AtWRKY18*, *AtWRKY40*, and *AtWRKY60* negatively regulate ABA signaling. Among them, *AtWRKY 40* binds to the promoter of *ABI4* and *ABI5* and represses the expression of ABA-responsive genes (Shang et al., 2010). *ABI4* itself and *WRKY 9* positively regulate *ABI4* expression in *Arabidopsis* (Chen et al., 2013), whereas *ABI3/VP1* (Feng et al., 2014) and Basic Pentacysteine (BPC) negatively regulate *ABI4* (Mu et al., 2017). However, the TF *WRKY 6* induces the expression of *RAV1* by binding to the W-box motif present in its promoter and this represses the expression of *ABI3*, *ABI4*, and *ABI5* (Huang et al., 2016).

In addition to the high enrichment of putative *cis*-elements associated with WRKY TF, high enrichment of putative *cis*-elements associated with *AB3/VP1* and *ABI5* was also identified. Recently, the role of WRKYs in hypoxia tolerance has been demonstrated in *Arabidopsis* (Tang et al., 2021). Through genetic and molecular experiments, it has been shown that WRKYs synergistically (*WRKY33* interacts with *WRKY12*) increase the activation of TF *RAP2.2* to increase the hypoxia tolerance in *Arabidopsis* and support the role of WRKY in hypoxia tolerance. The analysis of promoter architecture suggests its self-regulation depending on its requirement to activate other TFs to enhance submergence tolerance.

ZnF Regulatory Module

The promoter analysis results identified a high enrichment of zinc-finger binding element-like associated with ZnF TF in all tolerant, highly tolerant, and extremely tolerant genotypes (Tables 1–3). These elements are probably associated with the upregulation of different ZnF genes such as *Os03g0437200* (C2H2-type zinc finger protein, ABA-induced antioxidant defense, water stress, and oxidative stress tolerance), *Os03g0820300* (C2H2 transcription factor), *Os12g0113700* (Zinc finger, C3HC4-type family protein), *Os02g0646200* (Zinc finger, B-box domain-containing protein), *Os10g0456800* (CHY zinc finger family protein), *Os01g0303600* (Zinc finger, RING/FYVE/PHD-type domain-containing protein), *Os06g0340200* (Zinc finger, RING-CH-type domain-containing protein), *Os03g0329200* (Zinc finger CCCH domain-containing protein 23), *Os03g0764100* (Zinc finger transcription factor ZF1), *Os09g0486500* (Zinc finger A20 and AN1 domain-containing stress-associated protein 1), and *Os05g0128200* (Zinc finger CCCH domain-containing protein 33) (Table 4). The enrichment of zinc finger binding-like elements are higher in extremely tolerant genotypes compared to tolerant genotypes. This suggests a key role of ZnF in submergence tolerance although the molecular mechanism is not known yet. ZnF family TFs are known to be involved in important transcriptional regulation of plant responses to different abiotic stresses, such as drought, temperature, light, and salt (Wang et al., 2018). In rice, ZFP and C2H2 ZnF proteins are induced by a number of abiotic stresses including cold, drought, and salt (Jin et al., 2018). In our previous analysis, we identified high enrichment of zinc finger binding element-like elements associated with ZnF TFs as well as upregulation of a number of C2H2 and other ZnF proteins

in the upregulated genes related to anaerobic metabolism in rice under anoxia (Lakshmanan et al., 2014). We also found upregulation of the expression of ZnF TF in wild-type rice under complete submergence (Mohanty et al., 2016). Besides, it has been reported that ZnF TF was upregulated in both rice and *Arabidopsis* in response to hypoxia/anoxia (Loreti et al., 2005; Lasanthi-Kudahettige et al., 2007; Pandey and Kim, 2012). In rice, a CCCH-type zinc finger protein was significantly induced by hypoxia/submergence stress indicating a key role during hypoxia/submergence stress in rice (Pandey and Kim, 2012). A B-box type zinc finger protein displayed significantly higher expression in soybean tolerant genotypes in response to flooding treatment compared to sensitive genotypes (Yu et al., 2019).

NAC Regulatory Module

High enrichment of DRE-like elements associated with NAC TFs is identified in the promoters of upregulated genes in all tolerant, highly tolerant, and extremely tolerant genotypes (Tables 1–3). This high enrichment can be associated with the activities of several NAC genes such as *Os11g0154500* [No apical meristem (NAM) protein domain-containing protein; NAC-domain-containing protein 90], *Os03g0815100* (Similar to OsNAC6 protein), *Os01g0884300* (NAC domain-containing protein 6), *Os07g0684800* (Similar to NAM/CUC2-like protein), and *Os07g0225300* (OsNAC3 protein) (Table 4). Members of NAC family TFs play a major role in regulating abiotic and biotic stresses in *Arabidopsis* and many other crops (Nakashima et al., 2012; Shao et al., 2015; Yuan et al., 2019a,b). In many cases, NAC TF is regulated through the ABA-dependent signal transduction pathway (Chen et al., 2014). In addition, there is evidence regarding the interaction of NAC and JA signaling pathway. Regulation of NAC through both ABA and JA to abiotic stress tolerance has been elucidated in *Arabidopsis* and other plants (Bu et al., 2008; Yoshii et al., 2010). It also maintains membrane integrity during abiotic stresses in *Arabidopsis* (Yong et al., 2019). It could be regulating germination and coleoptile elongation to maintain membrane integrity in cross-talk with other hormones such as ABA and JA. Moreover, reactive oxygen and nitrogen species (ROS and RNS) that accumulate during submergence/hypoxia activate a number of TFs including Heat Shock Factor (HSFs) and NAC families to control the homeostasis of the harmful molecules (Gonzali et al., 2015).

DOF Regulatory Module

Sequence motif “AAAAG/CTTTT”-element-like was identified in the promoters of all tolerant, highly tolerant, and extremely tolerant upregulated genes, which can be associated with DOF (DNA binding with one finger) TFs. These TFs are mainly plant-specific (Noguero et al., 2013) and are involved in the regulation of various processes in plant metabolism, seed germination, phytochrome response, and various developmental processes (Noguero et al., 2013; Wu et al., 2015). However, the biological function of this TF is not well studied in rice yet. Experiments on *OsDof3* suggested that it interacts with GAMYB to induce the expression of *RAmy1A* to facilitate GA signaling during the germination of rice seeds (Washio, 2003). In this analysis, DOF could be interacting with GA and other plant hormones

at the initial stage of germination to activate the expression of stress-responsive genes to tolerate submergence during rice germination and coleoptile elongation.

HSF Regulatory Module

Enrichment of heat shock binding factor element-like in the promoter regions of upregulated genes of all tolerant, highly tolerant, and extremely tolerant genotypes could potentially be correlated with the upregulation of HSF genes such as *Os08g0471000* (Heat stress transcription factor B-4a, HSF20), *Os09g0526600* (Heat stress transcription factor B-2c, HSF 3), *Os09g0456800* (Heat stress transcription factor B-1), and *Os02g0232000* (Similar to Heat shock transcription factor 29, HSF 5) (Tables 1–4). HSFs are shown to be involved in heat stress and also other abiotic stresses (Nishizawa et al., 2008). Induction of a number of HSPs and HSTs in response to anoxia/hypoxia has been reported in *Arabidopsis* and rice (Loreti et al., 2005; Lasanthi-Kudahettige et al., 2007). HSF proteins play a major role in protecting cellular responses under stress conditions by preventing the misfolding and denaturation of proteins (Liberek et al., 2008) and also in the activation of HSP pathway during either anoxia or heat stress to cope with the production of ROS.

AS2/LBD Regulatory Module

Identification of a high enrichment of S2-binding-like elements in all tolerant, highly tolerant, and extremely tolerant genotypes proposes their possible association with Lateral organ boundaries domain (LBD) family TF AS2 (Tables 1–4). However, we did not see any expression of this gene in any of the tolerant genotypes. This TF plays a key role in different abiotic stress conditions. It is one of the hypoxia-induced TFs in *Arabidopsis* during flooding stress (Licausi et al., 2010). ARF7 and ARF19 TFs are known to regulate the expression of *LBD16* and *LBD29* for lateral organ development in *Arabidopsis* (Okushima et al., 2007). Recently, it has been shown that these TF genes could be regulated by auxin signaling under submergence (Wu and Yang, 2020), although the function is not known yet. This element could be present in the upregulated genes of all groups for lateral organ development later once the coleoptile reaches the surface of water for O₂ availability.

E2F Regulatory Module

A moderate enrichment of E2F-binding site-like elements associated with E2F TFs was identified in the upregulated genes of the tolerant genotypes (Table 1). E2F TF plays a role in the regulation of cell division in rice coleoptile elongation under anoxia (Kosugi and Ohashi, 2002). The presence of E2F-binding site-like elements associated with E2F TFs was also detected in the promoters of upregulated genes under anoxia in rice coleoptile elongation and in wild-type rice demonstrating coleoptile elongation under anoxia (Mohanty et al., 2012, 2016).

TCP Regulatory Module

TCP binding site-like elements associated with TCP TF are enriched in the promoters of upregulated genes in tolerant and highly tolerant genotypes (Tables 1, 2). This TF plays a key role in cell growth, different hormone response pathways, and abiotic stress responses (Li et al., 2005; Danisman, 2016). The role

of TCP in submergence tolerance is not known yet. However, in *Arabidopsis*, the activity of *AtTCP14* is necessary for seed germination and there is a functional relationship between this TF and GA (Tatematsu et al., 2008). Subsequently, both TCP14 and TCP15 were reported to regulate cell proliferation (Kieffer et al., 2011) and required for GA-dependent regulation of seed germination in *Arabidopsis* (Resentini et al., 2015). Identification of TCP-binding sites in tolerant and highly tolerant genotypes suggests their roles in initial cell proliferation of coleoptile during germination under submergence in rice.

HD-ZIP (ATHB4) Regulatory Module

Promoter analysis identified the presence of HD-ZIP binding site-like elements associated with HD-ZIP TFs in the promoters of all upregulated genes of all tolerant, highly tolerant, and extremely tolerant genotypes (Tables 1–3). The enrichment of this element is most probably associated with the activities of a number of HD-ZIP genes such as *Os06g014040* (Homeobox protein knotted-1-like 10), *Os05g0129700* (Homeobox-leucine zipper protein HOX28), *Os03g0198600* (Homeodomain-leucine zipper transcription factor), *Os06g0140700* (Homeobox-leucine zipper protein HOX2), *Os03g0188900* (Homeobox-leucine zipper protein HOX13), and *Os09g0528200* (Similar to Homeobox-leucine zipper protein HOX6) (Table 4). HD-ZIP is involved in the regulation of many developmental processes and response to different abiotic stresses in many plants (Annapurna et al., 2016; Tang et al., 2019; Zhao et al., 2019). However, the role of HD-ZIP in response to submergence stress is not known yet. In *Arabidopsis* seedlings, *AtHB4* gene was shown to regulate both shade avoidance and hormone-mediated development, particularly BR (Sorin et al., 2009). *AtHB4* acts downstream of P1F1 and was involved in the activation of hypocotyl cell wall composition and elongation in response to short day plants (Capella et al., 2015). The presence of this binding site in highly tolerant and extremely tolerant genotypes could be playing a role in coleoptile elongation by mediating the BR signaling pathway.

ARR-B Regulatory Module

ARR-B binding element-like associated with type-B ARRs was identified among upregulated genes in all tolerant, highly tolerant, and extremely tolerant genotypes (Tables 1–3). Type-B ARRs such as ARR1-ARR2, ARR10-ARR14, and ARR18-ARR21 play a key role as positive regulators of cytokinin signaling (Argyros et al., 2008). Besides, they also play an important role in the downregulation of ABA activity in response to cytokinin. Moreover, the binding sites of ARR-B TFs are known to be associated with metabolism of BR, BR signaling, and transcriptional regulation by TFs such as BZR2; BEH1, 2, 3, and 4; BIM1; and MYB30 (Zubo and Schaller, 2020). Genes that belong to ARR-B are also involved in the induction of GA biosynthesis and the reduction of GA perception (Marín-de la Rosa et al., 2015). These binding sites were identified in the genes associated with ethylene biosynthesis and signal transduction. Since genes in all genotypes possess these binding sites, it suggests that they could be playing a key role together with other hormones to regulate the elongation and germination of coleoptile in response to submergence tolerance.

MADS-box (AGL) Regulatory Module

High percentage of CARG box-binding site-like elements corresponding to MADS box (AGL) TF were identified in the upregulated genes of all tolerant, highly tolerant, and extremely tolerant genotypes (Tables 1–3), and this could be due to the induction of *Os04g0580700* (MADS-box transcription factor 17) (Table 4). MADS box (AGL) TFs are key regulators of many developmental processes in plants (Smaczniak et al., 2012). However, their role in seed germination is unknown except for a few MADS-box genes such as *AGL25*, *AGL67*, and *AGL 21* (Chiang et al., 2009; Bassel et al., 2011; Yu et al., 2017). *AGL25* is reported to be involved in temperature-dependent seed germination by inducing the GA biosynthetic pathway and the ABA catabolic pathway, whereas *AGL67* and *AGL21* act as negative regulators of seed germination. *AGL21* incorporates both hormone signal and environmental signal to ABA signaling by balancing ABI5 to prevent germination under adverse conditions (Yu et al., 2017).

BES/BZR Regulatory Module

Interestingly, E-box-like elements associated with BES/BZR TFs involved in Brassinosteroid (BR) signaling pathway were identified only in the upregulated genes of highly tolerant and extremely tolerant genotypes that had longer coleoptile elongation in response to submergence (Tables 2, 3). BR is one of the most important plant steroidal hormones that regulate a wide range of plant growth, development including cell elongation and seed germination, and responses to biotic and abiotic stresses (Guo et al., 2013; Tong et al., 2014; Zhang et al., 2014; Li et al., 2016; Ahammed et al., 2020). It has also been shown earlier that BRs can increase coleoptile length in rice (Yamamuro et al., 2000). However, the molecular mechanism of BR that controls the germination and coleoptile elongation of rice seeds is not well known yet. High level of BR inactivates BIN2 and activates dephosphorylation of TFs BZR1 and BZR2/BES1 by protein phosphatase for their accumulation in the nucleus to increase their DNA-binding activity for the expression of BR-responsive genes (Wang et al., 2012). BZR1 and BZR2/BES1 regulate a large number of structural and metabolic genes as well as cell wall biogenesis and regulatory genes (Yu et al., 2011). Gene expression profiles have also revealed the induction of the expression of many cell wall extension and loosening enzymes and expansins (Guo et al., 2009). Several studies have reported that rice has a well-preserved BR signaling pathway like *Arabidopsis* (Li et al., 2009; Tong and Chu, 2012; Tong et al., 2014). It also regulates developmental processes through a cross-talk interaction with other signaling pathways. In lowland rice plants, both BRs and GA act antagonistically in response to submergence tolerance (Schmitz et al., 2013). They have indicated that BR induces GA catabolic genes and also DELLA proteins to limit GA level during submergence, and the cross-talk between BR and GA depends on tissues and hormone levels in rice (Tong et al., 2014). Although ABA and BR signaling are known to be antagonist of each other, recently it has been reported that both BR and ABA co-regulate to enhance seed germination in *Arabidopsis*. BES1, an important TF of the BR signaling pathway, assists in seed germination by

weakening ABA signaling pathway induction by interfering the transcriptional activity of ABI5 (Zhao et al., 2019).

BR also activates different stress adaptive signaling pathways by directly or indirectly regulating different stress-responsive TFs such as bZIP, MYB, WRKY, NAC, DREB, etc. through BIN2 and key BZR1/BES1 TFs (Sharma et al., 2017). Genetic analysis shows that BZR1 and PIF, a bHLH TF, directly interact and promote cell elongation and etiolation (Oh et al., 2012). Moreover, BZR1 and PIF4 along with ARF6 promote genes that are involved in cell expansion (Oh et al., 2014). Both E-box motifs and G-box (CACGTG) motifs are associated with TF BZR1/2 and PIFs, and these are highly enriched in ARF6 binding regions. BEE2 is a bHLH TF, which is shown to be involved in the regulation of cell elongation by BR (Friedrichsen et al., 2002). BEE2 also interacts with ARF6 like PIF6 to promote cell elongation (Oh et al., 2014). These findings suggest a key role of BR-responsive TFs in cell elongation.

bHLH Regulatory Module

E-box-like/G-box-like elements associated with bHLH (Gr. III and VII) including PIFs were identified in both highly tolerant and extremely tolerant genotypes (Tables 2, 3). Interestingly, a set of E-box-like and BBRE-element-like/G-box-like elements associated with TFs such as BEE2, BIM1, BIM3, BAM8, and BEH4 that belong to bHLH were identified only in the upregulated genes of the two extremely tolerant genotypes. These elements could be associated with the induction of different bHLH TF genes such as *Os01g0773800* (bHLH protein 185), *Os03g0188400* (bHLH protein), *Os07g0628500* (bHLH dimerization region bHLH domain-containing protein), *Os03g0135700* (bHLH transcription factor), and *Os07g0143200* (Phytochrome-interacting bHLH factor) (Table 4). bHLH TFs are involved in the regulation of many cellular processes (Zhao et al., 2020) such as seed germination (Penfield et al., 2005), light signaling (Leivar et al., 2008), hormone signaling (Fernández-Calvo et al., 2011), responses to wounding, drought, salt, and low temperature (Sun et al., 2018). They also play a key role in BR-responsive gene expression to support coleoptile elongation, which has already been discussed earlier.

Phytochrome-interacting factors (PIFs) are bHLH TFs. Besides the bHLH domain, they have active phytochrome A/B binding domains. They are involved in various physiological processes such as seed germination, photomorphogenesis, shade responses, flowering time, and leaf senescence (Leivar and Quail, 2011; Casal, 2013; Sakuraba et al., 2014). Besides, PIFs promote cell elongation under low Red:far Red light conditions by inducing the transcription of genes related to growth (Paik et al., 2017). Involvement of PIFs in signaling responses (GH3, IAA, and ARF), cell wall modification, and elongation has been studied (Zhang et al., 2013; Pfeiffer et al., 2014). In addition, they are involved in a variety of hormone-response pathways such as GA, BR, JA, ethylene, and nitric oxide (Mazzella et al., 2014; Paik et al., 2017). However, the study of phototropism of rice coleoptile under submergence is not known yet. Interestingly, it was suggested that submerged coleoptiles exhibited only a slight R-induced growth inhibition (Pjon and Furuya, 1974). Studies show that they are less phototropic compared to other

gramineae coleoptiles such as maize and oats (Neumann and Iino, 1997). Experiments on the *coleoptile photomorphogenesis 1* (*cpm1*) mutant suggest that this gene has a role in phytochrome-mediated inhibition of coleoptile growth. Later studies on phototropism rice coleoptile demonstrates that auxin is involved in this process (Haga et al., 2005).

Identification of Putative *cis*-Elements in the Upregulated Genes of Intermediately Tolerant Diverse Rice Genotypes in Response to Submergence Tolerance

To compare and validate the identification of *cis*-elements in tolerant, highly tolerant, and extremely tolerant genotype groups, upregulated genes in the intermediate group, which has two highly tolerant genotypes, F291 and F274-2a, were also analyzed (Table 5). The analysis results identified highly enriched significant *cis*-elements that are associated with MYB, bZIP, ABI5, different groups of AP2/ERF, AP2/B3, EIL, ARF, WRKY, ZnF, NAC, DOF, HSF, AS2, TCP, ARR-B, MADS box, GATA, BES/BZR, bHLH, and BPC2 (Table 5). Interestingly, the analysis results for the intermediate group are exactly similar to the identification of *cis*-elements and their associated TFs in highly tolerant genotypes (Table 2). In addition, it is slightly different in terms of promoter architecture content from the analysis that was performed for extremely tolerant genotypes (Table 3). Hence, the analysis results for both intermediately tolerant genotypes and highly tolerant genotypes suggest that they have a common gene regulatory mechanism that allows longer coleoptile elongation of rice seeds during germination, in response to submergence tolerance.

Similarly, to validate the identification results, analysis was performed for a set of upregulated genes that have the moderate genotype Nipponbare and highly tolerant genotype F274-2a. The analysis results identified *cis*-elements that are associated with TFs such as MYB, bZIP, ERF, ABI3, AP2/B3, ARF, ZnF, NAC, HSF, AS2, TCP, ARR-B, MADS box, GATA, DBP1, and TBP (Table 6). Remarkably, the potential TFs that were identified also show similar promoter architecture with results obtained for tolerant genotypes (Table 1). The analysis results also did not identify specific *cis*-elements that are associated with BES/BZR and bHLH TFs present in highly tolerant and extremely tolerant genotypes (Tables 2, 3).

DISCUSSION

The global climate variation causes extreme weather changes and often causes severe rain, which is harmful for rice seed germination and seedling growth. Although rice has the unique ability to tolerate such conditions, different genotypes of rice show different degrees of tolerance to submergence by elongating their coleoptiles. Hence, it is essential to understand the regulatory mechanism that helps in tolerating such adverse conditions. It is a complex process and involves interactive mechanisms of both metabolic and transcriptional regulation and hormonal signaling. A number of studies show that different

TABLE 6 | Potential putative *cis*-elements identified in the promoters of upregulated genes in a group that has one tolerant genotype and one highly tolerant genotype (Nipponbare and F274-2a).

Cis-elements	Motifs	Associated TFs	% (TIC), E-value*
AT-hook/PE1-like	TTTTTCA	MYB (PF1)	55 (13.88), 3e-004
	AAAAAATA	MYB (PF1)	50 (15.40), 1e-004
	GTTTTTTT	MYB (PF1)	50 (15.39), 4e-004
GT-element-like	TGGTTTGT GGGGAAAA	MYB (GT-3)	69 (11.62), 1e-004
	AAAATATCT	MYB (GT-1/GT-3)	64 (12.14), 4e-004
		MYB (GT-1)	58 (13.64), 2e-004
Pyrimidine box-like	TTTTTCA	MYB (R1, R2R3)	55 (13.88), 3e-004
GARE-like	TGGTTTGT	MYB (R1, R2R3)	69 (11.62), 1e-004
MYB-box-like	TGGTTTGT	MYB (R2R3)	69 (11.62), 1e-004
	AAAACCAA	MYB (R2R3)	64 (12.42), 2e-004
	AACCATGC	MYB (R2R3)	57(11.65), 4e-004
As-1/ocs-like	TTTTTCA	bZIP (Gr. D, I, S)	55 (13.88), 3e-004
	CTGCAGGC	bZIP (Gr. D, I, S)	57 (11.69), 5e-004
GCN4 motif	TGGTTTGT	bZIP (RISBZ1, Gr. G)	69 (11.62), 1e-004
GCC-box-like	CGCCGCCGC	ERF (I, IV, VII, X)	52 (15.43), 7e-004
		ERF (I, IV, VII, X)	
ERE-like	CTGCCGCGC	ERF/RAP2.1, RAP2.2, RAP2.3, RAP2.6, RAP2.10,	57 (11.69), 5e-004
	CTGCAGGC	RAP2.11, RAP2.12 (Gr. III)	57 (11.69), 5e-004
		ERF/RAP2.2, RAP2.12 (Gr. VII)	
CRT/DRE-like	CTGCCGCGC	ERF (Gr., III, IV)	57 (11.69), 5e-004
ABRE-like	AACCATGC	ABI3/V1P1 (B3 domain)	57(11.65), 4e-004
ABRE-like		AP2/B3 (Gr. II)	56 (12.79), 2e-004
		Related to RAV2	
ABR-binding site-like	CTGCCGCGC	AP2-like ABA repressor 1 (ABR1)	57 (11.69), 5e-004
Aux-RE-like	TGGTTTGT	ARF1	69 (11.62), 1e-004
Zing finger binding site-like	GGGGAAAA TATATGTA	ZnF (C2H2-type)	64 (12.14), 4e-004
		ZnF (C2H2-type)	62 (12.72), 1e-004
		ZnF (C2H2-type)	58 (13.64), 2e-004
DRE-like	ACTCTTCC AACCATGC	NAC 20	69 (11.29), 3e-004
		NAC (4, 20, 38, 50, 57, 58, 70, 83, 103) SND3, CUC1, CUC2, CUC3	57(11.65), 4e-004
Heat shock binding factor element-like	ACTCCCCC	HSF (HSFB2A)	69 (11.29), 3e-004
	GGGGAAAA	HSF(HSFB2A)	64 (12.14), 4e-004
S2-binding site-like	ACTCCCCC	AS2 (LBD13)	69 (11.29), 3e-004
	CTCCTCCT	AS2 (LBD13)	57 (13.60), 3e-005
CARG box-binding site-like	ACTCTTCC GGGGAAAA	MADS-box (AG 16)	69 (11.29), 3e-004
		MADS-box (AG 15, 16)	64 (12.14), 4e-004
GATA binding site-like	TGGTTTGT	GATA 1	69 (11.62), 1e-004
DBP-binding site-like	AAAATTAT	DBP1	65 (12.40), 3e-004
	GAAATATT	DBP1	51 (12.80), 3e-004
TATA-box-like	GAAATATT	TBP	51 (12.80), 3e-004

% = percent occurrence among all upregulated genes, TIC = total information content of homology, E-value * = E-value of homology with promoter database entry.

types of hormones mainly activate transcriptional regulation that enables different plant metabolic processes to tolerate any adverse conditions due to climate change (Nemhauser et al., 2006). During rice germination and coleoptile elongation under submergence, although ethylene and ERF factors are known to be involved, the role of other TFs and hormonal signaling is not well understood yet. Therefore, the promoter architecture of upregulated genes of tolerant genotypes with different rates of coleoptile elongation could give some indication regarding the *cis*-elements content and their association with specific TFs. This information would support in providing a regulatory role of potential TFs and different hormonal signaling pathways

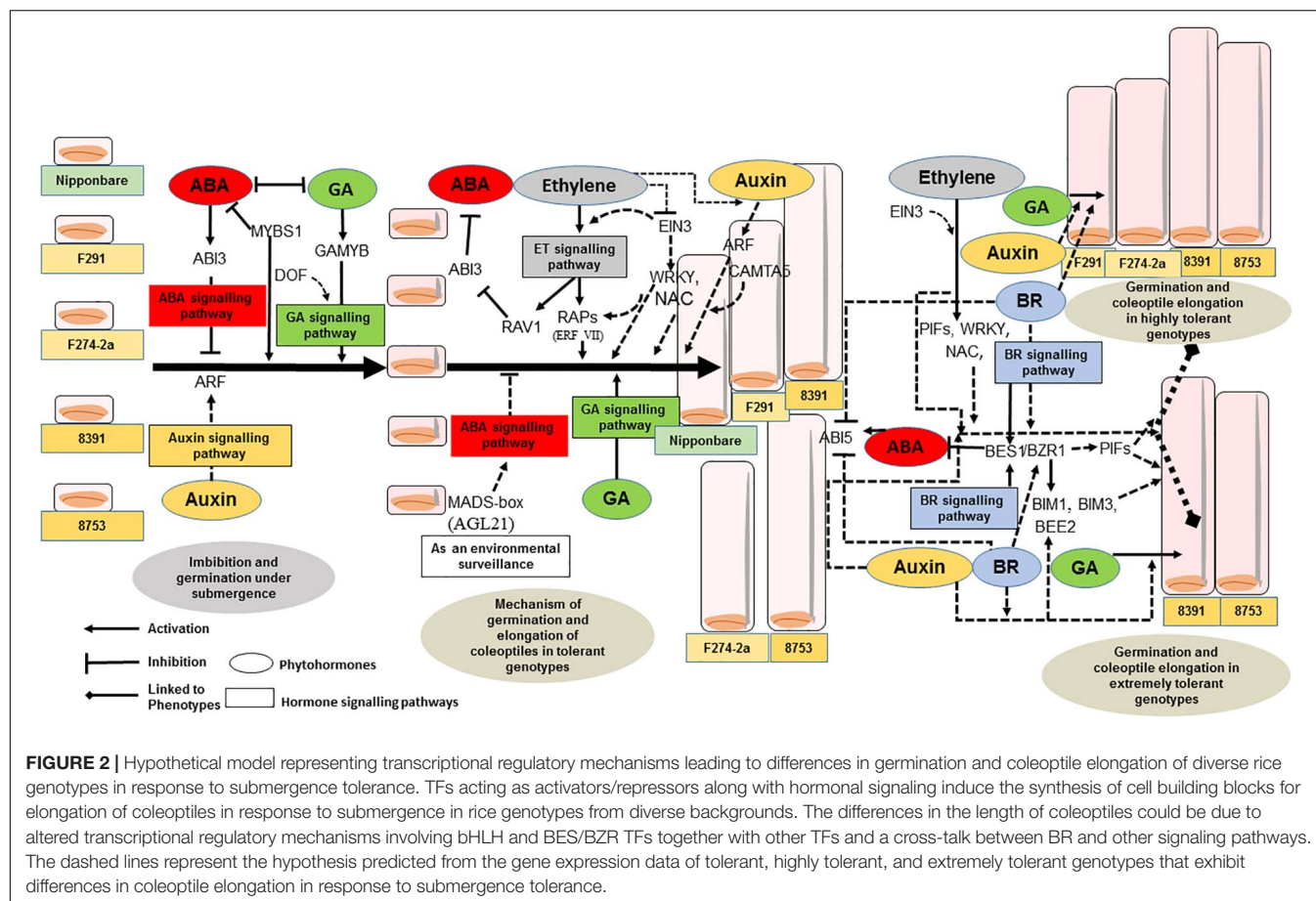
involved in the transcriptional regulation of submergence tolerance mechanism. In this study, promoter architecture was analyzed for different sets of common genes associated with different degrees of submergence tolerance such as tolerant genotypes (Nipponbare, two recombinant inbred lines F291 and F274-2a, and two natural genotypes 8391 and 8753), highly tolerant genotypes with longer coleoptile elongation (F291, F274-2a, 8391, and 8753), and extremely tolerant natural genotypes with the longest coleoptile elongation (8391 and 8753). The promoter architecture analysis of upregulated genes for tolerant genotypes including *O. sativa japonica* identified the presence of putative *cis*-elements that are associated with TFs such as

MYB, bZIP, AP2/ERF, ARF, EIN3, ABI3, ABR1, WRKY, ZnF, MADS-box, NAC, AS2, DOF, HD-ZIP, E2F, ARR-B, and HSF (Table 1). For the verification of these findings, promoters of the upregulated genes of a group that has tolerant genotypes, Nipponbare and F274-2a, were analyzed. The analysis identified most of the *cis*-elements that are associated with TFs identified in tolerant genotypes (Table 6). Interestingly, in addition to these TFs, there was identification of more specific binding sites associated with different specific TFs such as ABRE-binding site associated with bZIP, ABI5, ABF2, and bZIP (DPBF3); HY-5 binding site-like associated with HY5; G-box-like associated with GBF3, 5, and 6; E-box-like associated with BES/BZR TF; and E-box/G-box-like elements associated with bHLH (PIF7) in highly tolerant genotypes (Table 2). Moreover, the promoter architecture of the extremely tolerant genotypes (longest coleoptile elongation) is quite fascinating. They contain binding sites that are present in both tolerant and highly tolerant genotypes as well as higher enrichment of more binding sites such as ABRE-like associated with both ABF1 and ABF2 (bZIP), E-box-like/G-box-like and BBRE-element-like/G-box-like associated with bHLH, BEE2(bHLH), and BAM8 (bHLH) involved in BR-mediated signaling (Table 3). These binding sites are completely absent in the tolerant genotypes and less enriched or absent in highly tolerant genotypes. To support these findings, an additional promoter architecture analysis was performed by taking an intermediate genotype group that has two highly tolerant genotypes, F291 and F274-2a. The putative *cis*-element analysis of this group also identified less enrichment or complete absence of those specific binding sites that are present in the extremely tolerant genotypes (Table 5). These findings clearly show that the promoter architecture varies from genotype to genotype and the difference in tolerance mechanism and coleoptile elongation could be dependent on the presence of specific binding elements that are associated with specific transcriptional regulation by TFs in the promoter of genes upregulated during submergence.

Based on the promoter architecture of genes upregulated in all the three groups of different genotypes, the results suggest that there is involvement of both MYBs such as MYB (R1, R2R3) and MYB (R2R3) during the initial stage, i.e., imbibition and germination where both hormones ABA and GA play a key role. Upon imbibition of rice seed, the GA content gradually increases and breaks dormancy by inducing the secretion of hydrolytic enzymes and the endogenous ABA level decreases rapidly due to the induction of ABA catabolic genes (Figure 2). GAMYB is a GA-responsive R1, R2R3 TF that induces the expression of α -amylase gene (*RAmy3D*) in the aleurone layer for starch degradation to provide substrates for the germination and coleoptile elongation of rice seeds (Perata et al., 1997) under submergence. During this process, DOF TF interacts with GAMYB for induction of the expression of α -amylase to enable GA signaling (Zou et al., 2008; Figure 2). The high enrichment of MYB-box-like elements associated with MYB R2R3 TF and ABA signaling suggest that these TFs that are involved in seed maturation might be stored in rice seeds to initiate initial germination process during the early stage of germination (Sreenivasulu et al., 2008). High enrichment of ERF-VII TFs

including RAP2.2, RAP2.3, RAP2.6, RAP2.10, and RAV1 shows their involvement and the importance of the hormone ethylene in germination and coleoptile elongation of rice seeds in response to submergence. The gaseous hormone ethylene is required for the transport of sucrose to the coleoptile (Ishizawa and Esashi, 1988). This is again supported by the higher expression of two ERF genes such as *LOC_Os01g21120* in all genotypes [higher expression in tolerant genotypes compared to sensitive genotype (IR64) and higher expression of *LOC_Os07g47790* in the tolerant genotypes] (Hsu and Tung, 2017). The role of ethylene in coleoptile elongation is also supported by the identification of EIN3-binding site elements in all tolerant, highly tolerant, and extremely tolerant genotypes (Tables 1–3). EIN3 becomes active in the dark and regulates the ethylene signaling pathway to transcribe ethylene-responsive genes. Since this TF has a unique role in regulating multiple transcriptional regulation, it could also be involved in feedback regulation by ethylene and regulation of other hormones and TFs (Dolgikh et al., 2019; Figure 2). Besides, high enrichment of ERE-like elements associated with ERF TF (Gr. III) such as RAP2.2, RAP2.3, RAP2.6, and RAP2.10 suggests their roles in fermentation and sugar metabolism in response to submergence (Tables 1–3). Identification of RAV-like elements associated with RAV1 implies their role in repressing the ABA signaling pathway by inhibiting the expression of ABI3, ABI4, and ABI5 by binding to the 5′–CAACA–3′ *cis*-element in their promoters (Feng et al., 2014; Figure 2). The enrichment of binding sites associated with ERF TFs including GCC-box-like elements associated with ERF (I, IV, VII, and X) was highest in tolerant genotypes and less enriched in the extremely tolerant genotypes. High enrichment of RAP2.3, RAP2.6, RAP2.10, and ERF (Gr. VII) could be associated with a common role of ethylene in all genotypes. The enrichment of ARF binding sites associated with TF ARF and the expression of ARF gene suggest the role of auxin in an auxin-dependent elongation of rice coleoptile through the interaction with ethylene (Ishizawa and Esashi, 1984; Breviaro et al., 1992). Interaction of both hormones also inhibited root elongation in rice seedlings (Qin et al., 2017) and plays an important role in cell division and regulation of carbohydrate metabolism (Wu and Yang, 2020). Auxin upregulates GA biosynthesis genes and regulates the expression of GA metabolism genes through the action of Aux/IAA and ARF (Frigerio et al., 2006). Moreover, the co-regulation of both ethylene and auxin happens at the level of transcription as well as transport response. Besides AuxRE-like elements, there was identification of CAMTA5 binding site-like elements associated with bZIP (CAMTA5) TF. This TF could be involved in regulating auxin transport and homeostasis (Galon et al., 2010).

It is known that bZIP 11 (Gr. S1) is induced by sucrose and can regulate sugar metabolism. It is most probably involved in adaptations to carbon starvation in *Arabidopsis* (Ma et al., 2011). bZIP also activates genes involved in trehalose metabolism and amino acid metabolism in response to stress (Ma et al., 2011; Weiste and Dröge-Laser, 2014). In rice, *OsTPP7* encodes trehalose-6-phosphate phosphatase and is involved in the induction of starch mobilization during germination and coleoptile elongation (Kretschmar et al., 2015). However, this



gene was found to be upregulated in Nipponbare and the two RIL genotypes, F291 and F274-2a, but downregulated in two natural genotypes, and that could be due to structural variation (Hsu and Tung, 2017). Identification of as-1/ocs-like elements associated with bZIP (Gr. D, I, S) in tolerant, highly tolerant, and extremely tolerant genotypes propose their possible role in carbon and amino acid metabolism during coleoptile elongation in response to submergence. The majority of the bZIP TFs play a key role in ABA signaling pathways. G-box-related *cis*-elements associated with bZIPs are found in auxin-induced promoters and function as modulators of auxin-induced transcription in *Arabidopsis* (Weiste and Dröge-Laser, 2014). High enrichment of ABRE-like and G-box-like elements associated with bZIP TF could probably be involved in auxin-induced transcription for the longer elongation of coleoptiles in response to submergence in highly tolerant and extremely tolerant genotypes (Figure 2).

High enrichment of MYB-box-like elements associated with MYB R2R3 TF in the promoter regions of upregulated genes in all groups of genotypes indicates the involvement of pre-existing ABA not only at the beginning of germination but also at a later stage of germination through activation-inhibition machinery (Tables 1-3). It has been reported that ethylene-ABA interaction inhibits root growth in rice seedlings (Ma et al., 2014). In the dark, ethylene plays a double role in rice: stimulates coleoptile elongation and inhibits root growth (Ma et al., 2010,

2013; Yang et al., 2015), which is different from the response of ethylene in *Arabidopsis* (Bleecker and Kende, 2000). Moreover, it is motivating to find the presence of E-box-like/G-box-like elements associated with bHLH TFs such as PIF3 and PIF7 and BES/BZR TF in the promoters of upregulated genes in the highly tolerant and extremely tolerant genotypes (Tables 2, 3). Among them, the enrichment of the binding sites are much higher in the extremely tolerant natural genotypes. bHLH TFs are major players in phytochrome signal transduction and expression of BR-responsive genes (Duek and Fankhauser, 2005; Serna, 2007). PIFs and the hormone BR could be contributing in the regulation of longer cell elongation of rice coleoptiles under submergence (Paik et al., 2017) compared to moderate elongation in the *O. sativa japonica* genotype (Figure 2). Among the highly tolerant and extremely tolerant genotypes, E-box-like/G-box-like elements and BBRE-element-like/G-box-like element are more enriched in the extremely tolerant natural genotypes compared to the two recombinant inbred lines (RILs; F291 and F274-2a) derived from a cross between Nipponbare and IR64 (Tables 2, 3). BZR1 and PIF4 together with ARF4 induce genes that are involved in cell expansion (Oh et al., 2014). Moreover, BEE2 also regulates cell elongation by BR and interacts with ARF6 (Oh et al., 2014). These data indicate a potential involvement of BEE2, PIFs (bHLH), and BR-mediated signaling pathways contributing to the phenotypes of the two natural genotypes in

regulating the highly elongated coleoptile growth compared to the RILs (**Figure 2**). Additionally, BR is reported to stimulate seed germination and inactivate the negative effect of ABA on seed germination by a negative feedback mechanism that modulates ABA signaling (Steber and McCourt, 2001; Xi et al., 2010).

During submergence/flooding stress, rice coleoptiles also suffer from low osmotic stress in addition to hypoxia/anoxia stress. A number of TFs such as AREB/ABFs are known to regulate ABA signaling during such stress conditions (Yoshida et al., 2010). In this promoter architecture analysis, identification of ABRE-like elements associated with ABF1 and ABF2 TFs could be involved in osmotic stress tolerance in the highly tolerant and extremely tolerant genotypes where the coleoptile elongation is higher compared to the moderate genotype *O. sativa japonica*. Among these two groups, there is high enrichment of ABF1 and ABF2 associated motifs in the extremely tolerant genotypes compared to highly tolerant genotypes (**Tables 2, 3**). As these TFs are associated with ABA signaling, the endogenous ABA level could be acting as an on-off switch having a spatiotemporal regulation to activate/repress other TFs and hormonal signaling pathways during the germination and coleoptile elongation under submergence (**Figure 2**). The enrichment of ABRE-like elements associated with both ABI3 and ABI5 TFs in all five tolerant genotypes is complicated as both are positive regulators of ABA signaling and involved in the arrest of seed germination (Lopez-Molina et al., 2002). They have shown that ABI5 is a bZIP TF, and it acts downstream of ABI3 to inhibit germination. A MADS-box TF AGL21 positively regulates the expression of ABI5 and responds to a number of environmental stresses and plant hormones during seed germination (Yu et al., 2017). It is reported that it could be acting as a surveillance integrator to incorporate external environmental signals and endogenous hormonal signals to ABA signaling for the regulation of seed germination and early post-germination growth. In this analysis, identification of putative *cis*-elements associated with ABI3 and ABI5 and high enrichment of MADS-box (AGL) TFs indicate a probable role of AGL as an environmental surveillance integrator to safeguard seed germination process by inducing ABA signaling through ABI5 TF (**Figure 2**). Although the putative *cis*-element associated with ABI3 is present in all groups, the elements associated with ABI5 are more enriched in highly tolerant and extremely tolerant genotypes. Also, during seed germination in *Arabidopsis*, JA enhances ABA activation to inhibit seed germination through JAZ repressors of the JA signaling pathway by regulating ABI3 and ABI5 TFs (Pan et al., 2020). Upregulation of the expression of JA ZIM-domain protein indicates their possible role in the activation of ABI3 and ABI5 for the spatiotemporal safeguard seed germination process.

The role of WRKY TFs in response to coleoptile elongation in rice under submergence is not well studied yet. However, studies on WRKY TFs reported that it plays a role in hypoxic stress response in persimmon (Zhu et al., 2019), rice (Shiono et al., 2014; Mohanty et al., 2016), and sunflower (Raineri et al., 2015), and submergence tolerance in rice (Viana et al., 2018). Recent studies on WRKY stated that a regulatory module composed of WRKY33 and WRKY12 together with RAP2.2 plays a key role in hypoxia tolerance in *Arabidopsis* (Tang

et al., 2021). During submergence stress, interaction of WRKY 33 with WRKY12 upregulates the expression of RAP2.2, which acts downstream of both WRKY33 and WRKY12. RAP2.2 is an ethylene response TF that normally regulates genes associated with ethylene production, metabolism, and induction of genes encoding sugar metabolism and fermentation pathway enzymes (Hinz et al., 2010).

The identification of high enrichment of DRE-like elements associated with NAC TF in all groups of genotypes and expression of NAC genes suggest their role in regulating germination and coleoptile elongation together with other hormones to maintain membrane integrity. Similarly, the HD-ZIP (AtHB) gene acts downstream of P1F1 and regulates activation of hypocotyl cell wall composition and elongation through hormone-mediated development, particularly BR (Sorin et al., 2009; Capella et al., 2015). However, tolerant genotypes lack binding sites for BR-regulated TFs such as BES/BZR as well as bHLH (PIFs), whereas these are highly enriched particularly in highly tolerant and extremely tolerant genotypes (**Tables 2, 3**). In darkness, PIFs are normally active and involved in the regulation of gene expression to stimulate the skotomorphogenic response (Leivar et al., 2008). BZR1 and PIF4 interact with each other and regulate BR-induced gene expression (Martínez et al., 2018). There is also identification of ARR-B type binding site-like associated with ARR-B TFs. Type-B ARR binding sites are linked to the transcriptional regulation by TFs such as BZR2; BEH1, 2, 3, 4; BIM1; and MYB30 (Zubo and Schaller, 2020). MYB30 cooperates with BES1 and regulates the expression of BR-induced genes (Lee et al., 2009). Moreover, BR and GA are known to regulate cell elongation in rice, and BR regulates cell elongation by controlling GA metabolism (Tong et al., 2014). It seems that there is a cross-talk between BR, GA, and other hormones for the contribution of coleoptile elongation (**Figure 2**).

Interestingly, BBRE-element-like/G-box-like/E-box-like elements associated with TF BIM1, BIM3, and BEH1 are highly enriched in the extremely tolerant natural genotypes that have maximum coleoptile elongation (**Figure 2** and **Table 3**). These genes may play a role in longer elongation of their coleoptiles compared to the highly tolerant genotypes. Besides the contribution of the BES1/BZR1 and PIF4 interaction for the longer elongation of coleoptiles in the extremely tolerant genotypes, BEE2 could also be involved in the BR signaling pathway for the maximum elongation of coleoptile in those two genotypes. Additionally, BAM8 (β -amylase-like proteins), a bHLH TF, could be acting as a metabolic sensor by interacting with the BR signaling pathway (Soyk et al., 2014).

Besides these binding sites, there were identifications of binding sites for TFs such as ABR1, HD-ZIP (ATHB4), TCP, ZNF, E2F, AS2 (LBD), NAC, and HSF (**Tables 1-3, 5, 6**). The roles of these TFs have not been studied in response to submergence stress. However, they could be involved in different protective roles during germination and coleoptile elongation in response to submergence such as HSF in protecting and preventing cellular responses (Liberek et al., 2008), ABR1TF in relation to mechanical stress, and NAC TF in controlling homeostasis of the cells. C2H2 zinc finger proteins could be involved in targeting the antioxidant genes associated with ROS scavenging

(Han et al., 2020). TFs such as E2F and TCP could be involved in cell elongation and proliferation in a GA dependent pathway (Kieffer et al., 2011).

CONCLUSION

Germination and coleoptile elongation of diverse rice genotypes vary in response to submergence. Promoter architecture of upregulated genes associated with different tolerant genotypes suggests a fine-tuning at the transcriptional level that affects the phenotype of different genotypes. The germination and elongation of rice coleoptiles in response to submergence tolerance in all three genotype groups could be due to a combination of mechanisms involving different TFs such as MYB, bZIP, AP2/ERF, ARF, WRKY, ZnF, MADS-box, NAC, AS2, DOF, E2F, ARR-B, and HSF, and hormonal regulation by GA, ABA, ethylene, JA, and auxin. There could be re-balance between GA and ABA to activate other TFs and stress-responsive genes for the elongation of coleoptile to escape the submergence stress. Interestingly, the longer coleoptile elongation in the highly tolerant and extremely tolerant genotypes could be due to the involvement of additional TFs such as bHLH and BES/BZR along with BR signaling and a cross-talk with other signaling pathways. Moreover, the maximum elongation in the two extremely tolerant natural genotypes might be due to the additional transcriptional regulatory mechanism governed by TFs such as BEE2, BIM1, BIM3, and BAM8, which are not present in the other two genotype groups. The variation in coleoptile elongation in different groups of genotypes is certainly due to the difference in transcriptional regulatory mechanism controlled by specific TFs along with a synergistic cross-talk interaction

between different hormones, which needs further experimental validation. This analysis provides a potential mechanism of transcriptional regulation across rice genotypes from diverse backgrounds, which may be helpful for rice breeding targets for direct seeding to improve rice production.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author/s.

AUTHOR CONTRIBUTIONS

BM designed and analyzed the data, and wrote the manuscript.

ACKNOWLEDGMENTS

BM would like to thank NUS Environmental Research Institute, National University of Singapore, Singapore, for providing the facility to do the work and Dr. Pallavi Panda and Miss Puravi Panda for proofreading the manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fgene.2021.639654/full#supplementary-material>

REFERENCES

- Abdulmajid, D., Ali, N., Eltahawy, M. S., Liu, E., Dang, X., and Mining, D. H. (2020). Mining favorable alleles for rice coleoptile elongation length sensitivity to exogenous gibberellin under submergence condition. *J. Plant Growth Regul.* doi: 10.1007/s00344-020-10196-z
- Ahammed, G. J., Li, X., Liu, A., and Chen, S. (2020). Brassinosteroids in plant tolerance to abiotic stress. *J. Plant Growth Regul.* 39, 1451–1464. doi: 10.1007/s00344-020-10098-0
- An, F., Zhao, Q., Ji, Y., Li, W., Jiang, Z., Yu, X., et al. (2010). Ethylene-induced stabilization of ETHYLENE INSENSITIVE3 and EIN3-LIKE1 is mediated by proteasomal degradation of EIN3 binding F-box 1 and 2 that requires EIN2 in *Arabidopsis*. *Plant Cell* 22, 2384–2401. doi: 10.1105/tpc.110.07.6588
- Annapurna, B., Khurana, J. P., and Mukesh, J. (2016). Characterization of rice homeobox genes, OsHOX22 and OsHOX24, and over-expression of OsHOX24 in transgenic *Arabidopsis* suggest their role in abiotic stress response. *Front. Plant Sci.* 7:627. doi: 10.3389/fpls.2016.00627
- Argyros, R. D., Mathews, D. E., Chiang, Y. H., Palmer, C. M., Thibault, D. M., Etheridge, N., et al. (2008). Type B response regulators of *Arabidopsis* play key roles in cytokinin signaling and plant development. *Plant Cell* 20, 2102–2116. doi: 10.1105/tpc.108.059584
- Bailey-Serres, J., Fukao, T., Gibbs, D. J., Holdsworth, M. J., Lee, S. C., Licausi, F., et al. (2012). Making sense of low oxygen sensing. *Trends Plant Sci.* 17, 129–138. doi: 10.1016/j.tplants.2011.12.004
- Baltazar, M. D., Ignacio, J. C. I., Thomson, M. J., Ismail, A. M., Mendioro, M. S., and Septiningsih, E. M. (2014). QTL mapping for tolerance of anaerobic germination from IR64 and the aus landrace Nanhi using SNP genotyping. *Euphytica* 197, 251–260. doi: 10.1007/s10681-014-1064-x
- Bassel, G. W., Lan, H., Glaab, E., Gibbs, D. J., Gerjets, T., Krasnogor, N., et al. (2011). Genome-wide network model capturing seed germination reveals coordinated regulation of plant cellular phase transitions. *Proc. Natl. Acad. Sci. U.S.A.* 108, 9709–9714. doi: 10.1073/pnas.1100958108
- Bäumler, J., Riber, W., Klecker, M., Müller, L., Dissmeyer, N., Weig, A. R., et al. (2019). AtERF111/ABR1 is a transcriptional activator involved in the wounding response. *Plant J.* 100, 969–990.
- Birkenbihl, R. P., Kracher, B., and Somssich, I. E. (2017). Induced genome-wide binding of three *Arabidopsis* WRKY transcription factors during early MAMP-triggered immunity. *Plant Cell* 29, 20–38. doi: 10.1105/tpc.16.00681
- Bleecker, A. B., and Kende, H. (2000). Ethylene: a gaseous signal molecule in plants. *Annu. Rev. Cell Dev. Biol.* 16, 1–18. doi: 10.1146/annurev.cellbio.16.1.1
- Breviario, D., Giani, S., Di Vietri, P., and Coraggio, I. (1992). Auxin and growth regulation of rice coleoptile segments: molecular analysis. *Plant Physiol.* 98, 488–495. doi: 10.1104/pp.98.2.488
- Bu, Q., Jiang, H., Li, C.-B., Zhai, Q., Zhang, J., Wu, X., et al. (2008). Role of the *Arabidopsis thaliana* NAC transcription factors ANAC019 and ANAC055 in regulating jasmonic acid-signaled defense responses. *Cell Res.* 18, 756–767. doi: 10.1038/cr.2008.53
- Bui, L. T., Giuntoli, B., Kosmacz, M., Parlanti, S., and Licausi, F. (2015). Constitutively expressed ERF-VII transcription factors redundantly activate the core anaerobic response in *Arabidopsis thaliana*. *Plant Sci.* 236, 37–43. doi: 10.1016/j.plantsci.2015.03.008
- Campbell, M. T., Proctor, C. A., Dou, Y., Schmitz, A. J., Phansak, P., Kruger, G. R., et al. (2016). Genetic and molecular characterization of submergence response

- identifies Sub1o6 as a major submergence tolerance locus in maize. *PLoS One* 10:e0120385. doi: 10.1371/journal.pone.0120385
- Capella, M., Ribone, P. A., Arce, A. L., and Chan, R. L. (2015). *Arabidopsis thaliana* HomeoBox 1 (AtHB1), a homeodomain-leucine zipper I (HD-Zip I) transcription factor, is regulated by phytochrome-interacting factor 1 to promote hypocotyl elongation. *New Phytol.* 207, 669–682. doi: 10.1111/nph.13401
- Casal, J. J. (2013). Photoreceptor signaling networks in plant responses to shade. *Annu. Rev. Plant Biol.* 64, 403–427. doi: 10.1146/annurev-arplant-050312-120221
- Chang, K. N., Zhong, S., Weirauch, M. T., Hon, G., Pelizzola, M., Li, H., et al. (2013). Temporal transcriptional response to ethylene gas drives growth hormone cross-regulation in *Arabidopsis*. *eLife* 2:e00675.
- Chen, H., Zhang, J., Neff, M. M., Hong, S. W., Zhang, H., et al. (2008). Integration of light and abscisic acid signaling during seed germination and early seedling development. *Proc. Natl. Acad. Sci. U.S.A.* 105, 4495–4500. doi: 10.1073/pnas.0710778105
- Chen, L., Zhang, L., Li, D., Wang, F., and Yu, D. (2013). WRKY8 transcription factor functions in the TMV-cg defense response by mediating both abscisic acid and ethylene signaling in *Arabidopsis*. *Proc. Natl. Acad. Sci. U.S.A.* 110, E1963–E1971.
- Chen, X., Wang, Y., Lv, B., Li, J., Luo, L., Lu, S., et al. (2014). The NAC family transcription factor OsNAP confers abiotic stress response through the ABA pathway. *Plant Cell Physiol.* 55, 604–619. doi: 10.1093/pcp/pct204
- Chen, Y. S., Ho, T. H. D., Liu, L., Lee, D. H., Lee, C. H., Chen, Y. R., et al. (2019). Sugar starvation-regulated MYBS2 and 14-3-3 protein interactions enhance plant growth, stress tolerance, and grain weight in rice. *Proc. Natl. Acad. Sci. U.S.A.* 116, 21925–21935. doi: 10.1073/pnas.1904818116
- Chiang, G. C., Barua, D., Kramer, E. M., Amasino, R. M., and Donohue, K. (2009). Major flowering time gene, flowering locus C, regulates seed germination in *Arabidopsis thaliana*. *Proc. Natl. Acad. Sci. U.S.A.* 106, 11661–11666. doi: 10.1073/pnas.0901367106
- Chow, C. N., Lee, T. Y., Hung, Y. C., Li, G. Z., Tseng, K. C., Liu, Y. H., et al. (2019). PlantPAN3.0: a new and updated resource for reconstructing transcriptional regulatory networks from ChIP-seq experiments in plants. *Nucleic Acids Res.* 8, D1155–D1163.
- Danisman, S. (2016). TCP transcription factors at the interface between environmental challenges and the plant's growth responses. *Front. Plant Sci.* 7:1930. doi: 10.3389/fpls.2016.01930
- Dar, N. A., Amin, I., Wani, W., Wani, S. A., Shikari, A. B., Wani, S. H., et al. (2017). Abscisic acid: a key regulator of abiotic stress tolerance in plants. *Plant Gene* 11, 106–111. doi: 10.1016/j.plgene.2017.07.003
- Davuluri, R. V., Sun, H., Palaniswamy, S. K., Matthews, N., Molina, C., Kurtz, M., et al. (2003). AGRIS: *Arabidopsis* gene regulatory information server, an information resource of *Arabidopsis* cis-regulatory elements and transcription factors. *BMC Bioinform.* 4:25. doi: 10.1186/1471-2105-4-25
- de Vetten, N. C., and Ferl, R. J. (1995). Characterization of a maize G-box binding factor that is induced by hypoxia. *Plant J.* 7, 589–601. doi: 10.1046/j.1365-3113.1995.7040589.x
- Dolferus, R., Klok, E. J., Delessert, C., Wilson, S., Ismond, K. P., Good, A. G., et al. (2003). Enhancing the anaerobic response. *Ann. Bot.* 91, 111–117. doi: 10.1093/aob/mcf048
- Dolgikh, V. A., Pukhovaya, E. M., and Zemlyanskaya, E. V. (2019). Shaping ethylene response: the role of EIN3/EIL1 transcription factors. *Front. Plant Sci.* 10:1030. doi: 10.3389/fpls.2019.01030
- Duek, P. D., and Fankhauser, C. (2005). bHLH class transcription factors take centre stage in phytochrome signalling. *Trends Plant Sci.* 10, 51–54. doi: 10.1016/j.tplants.2004.12.005
- Feng, C. Z., Chen, Y., Wang, C., Kong, Y. H., Wu, W. H., and Chen, Y. F. (2014). *Arabidopsis* RAV1 transcription factor, phosphorylated by SnRK2 kinases, regulates the expression of ABI3, ABI4, and ABI5 during seed germination and early seedling development. *Plant J.* 80, 654–668.
- Fernández-Calvo, P., Chini, A., Fernández-Barbero, G., Chico, J. M., Gimenez-Ibanez, S., Geerinck, J., et al. (2011). The *Arabidopsis* bHLH transcription factors MYC3 and MYC4 are targets of JAZ repressors and act additively with MYC2 in the activation of jasmonate responses. *Plant Cell* 23, 701–715. doi: 10.1105/tpc.110.080788
- Friedrichsen, D. M., Nemhauser, J., Muramitsu, T., Maloof, J. N., Alonso, J., Ecker, J. R., et al. (2002). Three redundant brassinosteroid early response genes encode putative bHLH transcription factors required for normal growth. *Genetics* 162, 1445–1456.
- Frigerio, M., Alabadi, D., Perez-Gomez, J., Garcia-Carcel, L., Phillips, A. L., Hedden, P., et al. (2006). Transcriptional regulation of gibberellin metabolism genes by auxin signaling in *Arabidopsis*. *Plant Physiol.* 142, 553–563. doi: 10.1104/pp.106.084871
- Fukao, T., and Bailey-Serres, J. (2008a). Ethylene—a key regulator of submergence responses in rice. *Plant Sci.* 175, 43–51. doi: 10.1016/j.plantsci.2007.12.002
- Fukao, T., and Bailey-Serres, J. (2008b). Submergence tolerance conferred by Sub1A is mediated by SLR1 and SLR1 restriction of gibberellin responses in rice. *Proc. Natl. Acad. Sci. U.S.A.* 105, 16814–16819. doi: 10.1073/pnas.0807821105
- Fukao, T., Xu, K., Ronald, P. C., and Bailey-Serres, J. (2006). A variable cluster of ethylene response factor-like genes regulates metabolic and developmental acclimation responses to submergence in rice. *Plant Cell Online* 18, 2021–2034. doi: 10.1105/tpc.106.043000
- Galon, Y., Aloni, R., Nachmias, D., Snir, O., Feldmesser, E., Scrase-Field, S., et al. (2010). Calmodulin-binding transcription activator 1 mediates auxin signaling and responds to stresses in *Arabidopsis*. *Planta* 232, 165–178. doi: 10.1007/s00425-010-1153-6
- Gibbs, D. J., Conde, J. V., Berckhan, S., Prasad, G., Mendiondo, G. M., and Holdsworth, M. J. (2015). Group VII ethylene response factors coordinate oxygen and nitric oxide signal transduction and stress responses in plants. *Plant Physiol.* 169, 23–31. doi: 10.1104/pp.15.00338
- Gonzali, S., Loreti, E., Cardarelli, F., Novi, G., Parlanti, S., and Pucciariello, C. (2015). Universal stress protein HRU1 mediates ROS homeostasis under anoxia. *Nat. Plants* 1:15151.
- Gubler, F., Chandler, P. M., White, R. G., Llewellyn, D. J., and Jacobsen, J. V. (2002). Gibberellin signaling in barley aleurone cells. Control of SLN1 and GAMYB expression. *Plant Physiol.* 129, 191–200. doi: 10.1104/pp.010918
- Guo, H., Li, L., Aluru, M., Aluru, S., and Yin, Y. (2013). Mechanisms and networks for brassinosteroid regulated gene expression. *Curr. Opin. Plant Biol.* 16, 545–553. doi: 10.1016/j.pbi.2013.08.002
- Guo, H. Q., Li, L., Ye, H. X., Yu, X. F., Algreen, A., and Yin, Y. H. (2009). Three related receptor-like kinases are required for optimal cell elongation in *Arabidopsis thaliana*. *Proc. Natl. Acad. Sci. U.S.A.* 106, 7648–7653. doi: 10.1073/pnas.0812346106
- Haga, K., Takano, M., Neumann, R., and Iino, M. (2005). The rice COLEOPTILE PHOTOTROPISM 1 gene encoding an ortholog of *Arabidopsis* NPH3 is required for phototropism of coleoptiles and lateral translocation of auxin. *Plant Cell* 17, 103–115. doi: 10.1105/tpc.104.028357
- Han, G., Lu, C., Guo, J., Qiao, Z., Sui, N., Qiu, N., et al. (2020). C2H2 zinc finger proteins: master regulators of abiotic stress responses in plants. *Front. Plant Sci.* 11:115. doi: 10.3389/fpls.2020.00115
- Hattori, Y., Nagai, K., Furukawa, S., Song, X.-J., Kawano, R., Sakakibara, H., et al. (2009). The ethylene response factors SNORKEL1 and SNORKEL2 allow rice to adapt to deep water. *Nature* 460, 1026–1030. doi: 10.1038/nature08258
- Higo, K., Ugawa, Y., Iwamoto, M., and Korenaga, T. (1999). Plant cis-acting regulatory DNA elements (PLACE) database. *Nucleic Acids Res.* 27, 297–300. doi: 10.1093/nar/27.1.297
- Hinz, M., Wilson, I. W., Yang, J., Buerstenbinder, K., Llewellyn, D., Dennis, E. S., et al. (2010). *Arabidopsis* RAP2.2: an ethylene response transcription factor that is important for hypoxia survival. *Plant Physiol.* 153, 757–772. doi: 10.1104/pp.110.155077
- Hoson, T., Masuda, Y., and Pilet, P. E. (1992). Auxin content in air and water grown rice coleoptiles. *J. Plant Physiol.* 139, 685–689. doi: 10.1016/s0176-1617(11)81711-6
- Hossain, M. A., Cho, J.-I. I., Han, M., Ahn, C. H., Jeon, J. S., An, G., et al. (2010). The ABRE-binding bZIP transcription factor OsABF2 is a positive regulator of abiotic stress and ABA signaling in rice. *J. Plant Physiol.* 167, 1512–1520. doi: 10.1016/j.jplph.2010.05.008
- Hsu, F. C., Chou, M. Y., Chou, S. J., Li, Y. R., Peng, H. P., and Shih, M. C. (2013). Submergence confers immunity mediated by the WRKY22 transcription factor in *Arabidopsis*. *Plant Cell* 25, 2699–2713. doi: 10.1105/tpc.113.114447
- Hsu, S.-K., and Tung, C.-W. (2015). Genetic mapping of anaerobic germination associated QTLs controlling coleoptile elongation in rice. *Rice* 8:38.

- Hsu, S.-K., and Tung, C.-W. (2017). RNA-Seq analysis of diverse rice genotypes to identify the genes controlling coleoptile growth during submerged germination. *Front. Plant Sci.* 8:762. doi: 10.3389/fpls.2017.00762
- Huang, E., Yang, L., Chowdhary, R., Kassim, A., and Bajic, V. B. (2005). "An algorithm for *ab-initio* DNA motif detection," in *Information Processing and Living System*. World Scientific, eds V. B. Bajic, and T. W. Tan, (London: Imperial College Press), 611–614. doi: 10.1142/9781860946882_0004
- Huang, Y., Feng, C. Z., Ye, Q., Wu, W. H., and Chen, Y. F. (2016). Arabidopsis WRKY6 transcription factor acts as a positive regulator of abscisic acid signaling during seed germination and early seedling development. *PLoS Genet.* 12:e1005833. doi: 10.1371/journal.pgen.1005833
- Ishizawa, K., and Esashi, Y. (1983). Cooperation of ethylene and auxin in the growth-regulation of rice coleoptile segments. *J. Exp. Bot.* 34, 74–82. doi: 10.1093/jxb/34.1.74
- Ishizawa, K., and Esashi, Y. (1984). Osmoregulation in rice coleoptile elongation as promoted by cooperation between IAA and ethylene. *Plant Cell Physiol.* 25, 495–504.
- Ishizawa, K., and Esashi, Y. (1988). Action mechanism of ethylene in the control of sugar translocation in relation to rice coleoptile growth. I. Sucrose metabolism. *Plant Cell Physiol.* 29, 1311–1341.
- Itoh, T., Tanaka, T., Barrero, R. A., Yamasaki, C., Fujii, Y., Hilton, P. B., et al. (2007). Curated genome annotation of *Oryza sativa* ssp. japonica and comparative genome analysis with *Arabidopsis thaliana*. *Genome Res.* 17, 175–183.
- Jin, Y. M., Piao, R., Yan, Y. F., Chen, M., Wang, L., He, H., et al. (2018). Overexpression of a new zinc finger protein transcription factors *OsCTZFP8* improves cold tolerance in rice. *Int. J. Genomics.* 2018, 1–13. doi: 10.1155/2018/5480617
- Jing, Y., Zhang, D., Wang, X., Tang, W., Wang, W., Huai, J., et al. (2013). *Arabidopsis* chromatin remodeling factor PICKLE interacts with transcription factor HY5 to regulate hypocotyl cell elongation. *Plant Cell* 25, 242–256. doi: 10.1105/tpc.112.105742
- Kaneko, M., Itoh, H., Ueguchi, M., Ashikari, M., and Matsuoka, M. (2002). The α -Amylase induction in endosperm during rice seed germination is caused by gibberellin synthesized in epithelium. *Plant Physiol.* 128, 1264–1270. doi: 10.1104/pp.010785
- Katiyar, A., Smita, S., Lenka, S. K., Rajwanshi, R., Chinnusamy, V., and Bansa, K. C. (2012). Genome-wide classification and expression analysis of MYB transcription factor families in rice and *Arabidopsis*. *BMC Genomics* 13:544. doi: 10.1186/1471-2164-13-544
- Kieffer, M., Master, V., Waites, R., and Davies, B. (2011). TCP14 and TCP15 affect internode length and leaf shape in *Arabidopsis*. *Plant J.* 68, 147–158. doi: 10.1111/j.1365-3113x.2011.04674.x
- Kim, S., Kang, J. Y., Cho, D. I., Park, J. H., and Kim, S. Y. (2004). ABF2, an ABRE-binding bZIP factor, is an essential component of glucose signaling and its overexpression affects multiple stress tolerance. *Plant J.* 40, 75–87. doi: 10.1111/j.1365-3113x.2004.02192.x
- Kosmacz, M., Parlanti, S., Schwarzländer, M., Kragler, F., Licausi, F., and Van Dongen, J. T. (2015). The stability and nuclear localization of the transcription factor RAP2.12 are dynamically regulated by oxygen concentration. *Plant Cell Environ.* 38, 1094–1103. doi: 10.1111/pce.12493
- Kosugi, S., and Ohashi, Y. (2002). E2F sites that can interact with E2F proteins cloned from rice are required for meristematic tissue-specific expression of rice and tobacco proliferating cell nuclear antigen promoters. *Plant J.* 29, 45–59. doi: 10.1046/j.1365-3113x.2002.01196.x
- Kretschmar, T., Pelayo, M. A., Trijatmiko, K. R., Gabunada, L. F., Alam, R., Jimenez, R., et al. (2015). A trehalose-6-phosphate phosphatase enhances anaerobic germination tolerance in rice. *Nat. Plants* 24:15124.
- Lakshmanan, M., Mohanty, B., Lim, S.-H., Ha, S.-H., and Lee, D.-Y. (2014). Metabolic and transcriptional regulatory mechanisms underlying the anoxic adaptation of rice coleoptile. *Arabidopsis* 6:Plu026.
- Lasanthi-Kudahettige, R., Magneschi, L., Loreti, E., Gonzali, S., Licausi, F., Novi, G., et al. (2007). Transcript profiling of the anoxic rice coleoptile. *Plant Physiol.* 144, 218–231. doi: 10.1104/pp.106.093997
- Lee, K. W., Chen, P. W., Lu, C. A., Chen, S., Ho, T. H., and Yu, S. M. (2009). Coordinated responses to oxygen and sugar deficiency allow rice seedlings to tolerate flooding. *Sci. Signal.* 2:ra61. doi: 10.1126/scisignal.2000333
- Lee, K. W., Chen, P. W., and Yu, S. M. (2014). Metabolic adaptation to sugar/O₂ deficiency for anaerobic germination and seedling growth in rice. *Plant Cell Environ.* 37, 2234–2244.
- Leivar, P., Monte, E., Oka, Y., Liu, T., Carle, C., Castillon, A., et al. (2008). Multiple phytochrome-interacting bHLH transcription factors repress premature seedling photomorphogenesis in darkness. *Curr. Biol.* 18, 1815–1823. doi: 10.1016/j.cub.2008.10.058
- Leivar, P., and Quail, P. H. (2011). PIFs: pivotal components in a cellular signaling hub. *Trends Plant Sci.* 16, 19–28. doi: 10.1016/j.tplants.2010.08.003
- Li, C., Potuschak, T., Colón-Carmona, A., Gutiérrez, R. A., and Doerner, P. (2005). *Arabidopsis* TCP20 links regulation of growth and cell division control pathways. *Proc. Natl Acad. Sci. U.S.A.* 102, 12978–12983. doi: 10.1073/pnas.0504039102
- Li, L., Yu, X., Thompson, A., Guo, M., Yoshida, S., Asami, T., et al. (2009). Arabidopsis MYB30 is a direct target of BES1 and cooperates with BES1 to regulate brassinosteroid-induced gene expression. *Plant J.* 58, 275–286. doi: 10.1111/j.1365-3113x.2008.03778.x
- Li, Q. F., Xiong, M., Xu, P., Huang, L. C., Zhang, C. Q., and Liu, Q. Q. (2016). Dissection of brassinosteroid-regulated proteins in rice embryos during germination by quantitative proteomics. *Sci. Rep.* 6:34583. doi: 10.1038/srep34583
- Liberek, K., Lewandowska, A., and Zietkiewicz, S. (2008). Chaperones in control of protein disaggregation. *Embo J.* 27, 328–335. doi: 10.1038/sj.emboj.7601970
- Licausi, F., vanDongen, J. T., Giuntoli, B., Novi, G., Santaniello, A., Geigenberger, P., et al. (2010). HRE1 and HRE2, two hypoxia-inducible ethylene response factors, affect anaerobic responses in *Arabidopsis thaliana*. *Plant J.* 62, 302–315. doi: 10.1111/j.1365-3113x.2010.04149.x
- Lin, C. C., Chao, Y. T., Chen, W. C., Ho, H. Y., Chou, M. Y., Li, Y. R., et al. (2019). Regulatory cascade involving transcriptional and N-end rule pathways in rice under submergence. *Proc. Natl Acad. Sci. U.S.A.* 116, 3300–3309. doi: 10.1073/pnas.1818507116
- Lopez-Molina, L., Mongrand, S., Mclachlin, D. T., Chait, B. T., and Chua, N. H. (2002). ABI5 acts downstream of ABI3 to execute an ABA-dependent growth arrest during germination. *Plant J.* 32, 317–328. doi: 10.1046/j.1365-3113x.2002.01430.x
- Loreti, E., Poggi, A., Novi, G., Alpi, A., and Perata, P. (2005). A genome-wide analysis of the effects of sucrose on gene expression in *Arabidopsis* seedlings under anoxia. *Plant Physiol.* 137, 1130–1138. doi: 10.1104/pp.104.057299
- Ma, B., Chen, S. Y., and Zhang, J. S. (2010). Ethylene signaling in rice. *Chin. Sci. Bull.* 55, 2204–2210. doi: 10.1007/s11434-010-3192-2
- Ma, B., He, S. J., Duan, K. X., Yin, C. C., Chen, H., Yang, C., et al. (2013). Identification of rice ethylene-response mutants and characterization of MHZ7/OsEIN2 in distinct ethylene response and yield trait regulation. *Mol. Plant* 6, 1830–1848. doi: 10.1093/mp/sst087
- Ma, B., Yin, C. C., He, S. J., Lu, X., Zhang, W. K., Lu, T. G., et al. (2014). Ethylene-induced inhibition of root growth requires abscisic acid function in rice (*Oryza sativa* L.) seedlings. *PLoS Genet.* 10:e1004701. doi: 10.1371/journal.pgen.1004701
- Ma, J., Hanssen, M., Lundgren, K., Hernández, L., Delatte, T., Ehlert, A., et al. (2011). The sucrose-regulated *Arabidopsis* transcription factor bZIP11 reprograms metabolism and regulates trehalose metabolism. *New Phytol.* 191, 733–745. doi: 10.1111/j.1469-8137.2011.03735.x
- Marín-de la Rosa, N., Pfeiffer, A., Hill, K., Locascio, A., Bhalerao, R. P., Miskolczi, P., et al. (2015). Genome wide binding site analysis reveals transcriptional coactivation of cytokinin-responsive genes by DELLA proteins. *PLoS Genet.* 11:e1005337. doi: 10.1371/journal.pgen.1005337
- Martinez, C., Espinosa-Ruiz, A., de Lucas, M., Bernardo-García, S., Franco-Zorrilla, J. M., and Prat, S. (2018). PIF4-induced BR synthesis is critical to diurnal and thermomorphogenic growth. *Embo J.* 37:e99552.
- Matys, V., Fricke, E., Geffers, R., Gossling, E., Haubrock, M., Hehl, R., et al. (2003). TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res.* 31, 74–378.
- Mazzella, M. A., Casal, J. J., Muschietti, J. P., and Fox, A. R. (2014). Hormonal networks involved in apical hook development in darkness and their response to light. *Front. Plant Sci.* 5:52. doi: 10.3389/fpls.2014.00052
- Mishra, M., Kanwar, P., Singh, A., Pandey, A., Kapoor, S., and Pandey, G. K. (2013). Plant omics: genome-wide analysis of ABA Repressor1 (ABR1) related genes in rice during abiotic stress and development. *Omics J. Int. Biol.* 17, 439–450. doi: 10.1089/omi.2012.0074
- Mizoi, J., Shinozaki, K., and Yamaguchi-Shinozaki, K. (2012). AP2/ERF family transcription factors in plant abiotic stress responses. *Biochim. Biophys. Acta Gene Regul. Mech.* 1819, 86–96. doi: 10.1016/j.bbargm.2011.08.004

- Mohanty, B., Hearth, V., Wijaya, E., Reyes, B. D., and Lee, D. Y. (2012). Patterns of *cis*-element enrichment reveal potential regulatory modules in the transcriptional regulation of anoxia response of japonica rice. *Gene* 511, 235–242. doi: 10.1016/j.gene.2012.09.048
- Mohanty, B., Takahashi, H., de los Reyes, B. G., Wijaya, E., Nakazono, E., and Lee, D.-Y. (2016). Transcriptional regulatory mechanism of alcohol dehydrogenase 1-deficient mutant of rice for cell survival under complete submergence. *Rice* 9:51.
- Mu, Y., Zou, M., Sun, X., He, B., Xu, X., Liu, Y., et al. (2017). BASIC PENTACYSTEINE proteins repress ABSCISIC ACID INSENSITIVE4 expression via direct recruitment of the Polycomb-Repressive Complex 2 in Arabidopsis root development. *Plant Cell Physiol.* 58, 607–621.
- Nakashima, K., Takasaki, H., Mizoi, J., Shinozaki, K., and Yamaguchi-Shinozaki, K. (2012). NAC transcription factors in plant abiotic stress responses. *Biochim. Biophys. Acta* 1819, 97–103.
- Narsai, R., Edwards, J. M., Roberts, T. H., Whelan, J., Joss, G. H., and Atwell, B. J. (2015). Mechanisms of growth and patterns of gene expression in oxygen deprived rice coleoptiles. *Plant J.* 82, 25–40. doi: 10.1111/tpj.12786
- Narsai, R., Howell, K. A., Carroll, A., Ivanova, A., Millar, A. H., and Whelan, J. (2009). Defining core metabolic and transcriptomic responses to oxygen availability in rice embryos and young seedlings. *Plant Physiol.* 151, 306–322. doi: 10.1104/pp.109.142026
- Nemhauser, J. L., Hong, F., and Chory, J. (2006). Different plant hormones regulate similar processes through largely nonoverlapping transcriptional responses. *Cell* 126, 467–475. doi: 10.1016/j.cell.2006.05.050
- Neumann, R., and Iino, M. (1997). Phototropism of rice (*Oryza sativa* L.) coleoptiles: fluence-response relationships, kinetics and photogravitropic equilibrium. *Planta* 201, 288–292. doi: 10.1007/s004250050068
- Nghi, K. N., Tagliani, A., Mariotti, L., Weits, D. A., Perata, P., and Pucciariello, C. (2020). Auxin is required for the long coleoptile trait in *japonica* rice under submergence. *New Phytol.* 229, 85–93. doi: 10.1111/nph.16781
- Nishizawa, A., Yabuta, Y., and Shigeoka, S. (2008). Galactinol and raffinose constitute a novel function to protect plants from oxidative damage. *Plant Physiol.* 147, 1251–1263. doi: 10.1104/pp.108.122465
- Noguero, M., Atif, R. M., Ochatt, S., and Thompson, R. D. (2013). The role of the DNA-binding One Zinc Finger (DOF) transcription factor family in plants. *Plant Sci.* 209, 32–45. doi: 10.1016/j.plantsci.2013.03.016
- Oh, E., Zhu, J. Y., Bai, M. Y., Arenhart, R. A., Sun, Y., and Wang, Z. Y. (2014). Cell elongation is regulated through a central circuit of interacting transcription factors in the *Arabidopsis* hypocotyl. *eLife* 3:e03031.
- Oh, E., Zhu, J. Y., and Wang, Z. Y. (2012). Interaction between BZR1 and PIF4 integrates brassinosteroid and environmental responses. *Nat. Cell Biol.* 14, 802–809. doi: 10.1038/ncb2545
- Okushima, Y., Fukaki, H., Onoda, M., Theologis, A., and Tasaka, M. (2007). ARF7 and ARF19 regulate lateral root formation via direct activation of LBD/ASL genes in *Arabidopsis*. *Plant Cell* 19, 118–130. doi: 10.1105/tpc.106.047761
- Paik, I., Kathare, P. K., Kim, J. I., and Huq, E. (2017). Expanding roles of PIFs in signal integration from multiple processes. *Mol. Plant* 10, 1035–1046. doi: 10.1016/j.molp.2017.07.002
- Pan, J., Hu, Y., Wang, H., Guo, Q., Chen, Y., Howe, G. A., et al. (2020). Molecular mechanism underlying the synergetic effect of jasmonate on abscisic acid signaling during seed germination in *Arabidopsis*. *Plant Cell* 32, 3846–3865. doi: 10.1105/tpc.19.00838
- Pandey, D. M., and Kim, S. R. (2012). Identification and expression analysis of hypoxia stress inducible CCCH-type zinc finger protein genes in rice. *J. Plant Biol.* 55, 489–497. doi: 10.1007/s12374-012-0384-4
- Pandey, G. K., Grant, J. J., Cheong, Y. H., Kim, B. G., Li, L., and Luan, S. (2005). ABR1, an APETALA2-domain transcription factor that functions as a repressor of ABA response in *Arabidopsis*. *Plant Physiol.* 139, 1185–1193. doi: 10.1104/pp.105.066324
- Papdi, C., Perez-Salamo, I., Joseph, M. P., Giuntoli, B., Bogre, L., Koncz, C., et al. (2015). The low oxygen, oxidative and osmotic stress responses synergistically act through the ethylene response factor VII genes RAP2.12, RAP2.2 and RAP2.3. *Plant J.* 82, 772–784. doi: 10.1111/tpj.12848
- Penfield, S., Josse, E. M., Kannangara, R., Gilday, A. D., Halliday, K. J., and Graham, I. A. (2005). Cold and light control seed germination through the bHLH transcription factor SPATULA. *Curr. Biol.* 15, 1998–2006. doi: 10.1016/j.cub.2005.11.010
- Perata, P., Guglielminetti, L., and Alpi, A. (1997). Mobilization of endosperm reserves in cereal seeds under anoxia. *Ann. Bot.* 79, 49–56. doi: 10.1093/oxfordjournals.aob.a010306
- Perrot-Rechenmann, C. (2010). Cellular responses to auxin: division versus expansion. *Cold Spring Harb. Perspect. Biol.* 2:a001446. doi: 10.1101/cshperspect.a001446
- Pfeiffer, A., Shi, H., Tepperman, J. M., Zhang, Y., and Quail, P. H. (2014). Combinatorial complexity in a transcriptionally centered signaling hub in *Arabidopsis*. *Mol. Plant* 7, 1598–1618. doi: 10.1093/mp/ssu087
- Phukan, U. J., Jeena, G. S., and Shukla, R. K. W. R. K. Y. (2016). Transcription factors: molecular regulation and stress responses in plants. *Front. Plant Sci.* 7:760. doi: 10.3389/fpls.2016.00760
- Pjón, C. J., and Furuya, M. (1974). Phytochrome action in *Oryza sativa* L. VII. Effects of light and aeration on coleoptile growth under submerged conditions. *Plant Cell Physiol.* 15, 663–668. doi: 10.1093/oxfordjournals.pcp.a075051
- Qin, H., Zhang, Z., Wang, J., Chen, X., Wei, P., and Huang, R. (2017). The activation of OsEIL1 on YUC8 transcription and auxin biosynthesis is required for ethylene-inhibited root elongation in rice early seedling development. *PLoS Genet.* 13:e1006955. doi: 10.1371/journal.pgen.1006955
- Raineri, J., Ribichich, K. F., and Chan, R. L. (2015). The sunflower transcription factor HaWRKY76 confers drought and flood tolerance to *Arabidopsis thaliana* plants without yield penalty. *Plant Cell Rep.* 34, 2065–2080. doi: 10.1007/s00299-015-1852-3
- Resentini, F., Felipe-Benavent, A., Colombo, L., Blázquez, M. A., Alabadi, D., and Masiero, S. (2015). TCP14 and TCP15 mediate the promotion of seed germination by gibberellins in *Arabidopsis thaliana*. *Mol. Plant* 8, 482–485. doi: 10.1016/j.molp.2014.11.018
- Safavi-Rizi, V., Herde, M., and Stöhr, C. (2020). RNA-Seq reveals novel genes and pathways associated with hypoxia duration and tolerance in tomato root. *Sci. Rep.* 10:1692.
- Sakuraba, Y., Jeong, J., Kang, M. Y., Kim, J., Paek, N. C., and Choi, G. (2014). Phytochrome-interacting transcription factors PIF4 and PIF5 induce leaf senescence in *Arabidopsis*. *Nat. Commun.* 5:4636.
- Schmitz, A. J., Folsom, J. J., Jikamaru, Y., Ronald, P., and Walia, H. (2013). SUB1A-mediated submergence tolerance response in rice involves differential regulation of the brassinosteroid pathway. *New Phytol.* 198, 1060–1070.
- Septiningsih, E. M., Ignacio, J. C., Sendon, P. M., Sanchez, D. L., Ismail, A. M., and Mackill, D. J. (2013). QTL mapping and confirmation for tolerance of anaerobic conditions during germination derived from the rice landrace Ma-Zhan Red. *Theor. Appl. Genet.* 126, 1357–1366.
- Serna, L. (2007). bHLH proteins know when to make a stoma. *Trends Plant Sci.* 12, 483–485. doi: 10.1016/j.tplants.2007.08.016
- Shang, Y., Yan, L., Liu, Z. Q., Cao, Z., Mei, C., Xin, Q., et al. (2010). The Mg-chelatase H subunit of *Arabidopsis* antagonizes a group of WRKY transcription repressors to relieve ABA-responsive genes of inhibition. *Plant Cell*, 22, 1909–1935. doi: 10.1105/tpc.110.073874
- Shao, H. B., Wang, H. Y., and Tang, X. L. (2015). NAC transcription factors in plant multiple abiotic stress responses: progress and prospects. *Front. Plant Sci.* 6:902. doi: 10.3389/fpls.2015.00902
- Sharma, I., Kaur, N., and Pati, P. K. (2017). Brassinosteroids: a promising option in deciphering remedial strategies for abiotic stress tolerance in rice. *Front. Plant Sci.* 8:2151. doi: 10.3389/fpls.2017.02151
- Sharma, N., Dang, T. M., Singh, N., Ruzicic, S., Mueller-Roeber, B., Baumann, U., et al. (2018). Allelic variants of OsSUB1A cause differential expression of transcription factor genes in response to submergence in rice. *Rice* 11:2.
- Shi, Y., Tian, S., Hou, L., Huang, X., Zhang, X., Guo, H., et al. (2012). Ethylene signaling negatively regulates freezing tolerance by repressing expression of CBF and type-A ARR genes in *Arabidopsis*. *Plant Cell* 24, 2578–2595. doi: 10.1105/tpc.112.098640
- Shingaki-Wells, R. N., Huang, S., Taylor, N. L., Carroll, A. J., Zhou, W., and Millar, A. H. (2011). Differential molecular responses of rice and wheat coleoptiles to anoxia reveal novel metabolic adaptations in amino acid metabolism for tissue tolerance. *Plant Physiol.* 156, 1706–1724. doi: 10.1104/pp.111.175570
- Shiono, K., Yamauchi, T., Yamazaki, S., Mohanty, B., Malik, A. I., Nagamura, Y., et al. (2014). Microarray analysis of laser-microdissected tissues indicates the biosynthesis of suberin in the outer part of roots during formation of a

- barrier to radial oxygen loss in rice (*Oryza sativa*). *J. Exp. Bot.* 65, 4795–4806. doi: 10.1093/jxb/eru235
- Smaczniak, C., Immink, R. G. H., Angenent, G. C., and Kaufmann, K. (2012). Developmental and evolutionary diversity of plant MADS-domain factors: insights from recent studies. *Development* 139, 3081–3098. doi: 10.1242/dev.074674
- Song, C. P., Agarwal, M., Ohta, M., Guo, Y., Halfter, U., Wang, P., et al. (2005). Role of an *Arabidopsis* AP2/EREBP-type transcriptional repressor in abscisic acid and drought stress responses. *Plant Cell* 17, 2384–2396. doi: 10.1105/tpc.105.033043
- Sorin, C., Salla-Martret, M., Bou-Torrent, J., Roig-Villanova, I., and Martínez-García, J. F. (2009). ATHB4, a regulator of shade avoidance, modulates hormone response in *Arabidopsis* seedlings. *Plant J.* 59, 266–277. doi: 10.1111/j.1365-3113X.2009.03866.x
- Soyk, S., Simková, K., Zürcher, E., Luginbühl, L., Brand, L. H., Vaughan, K. C., et al. (2014). The enzyme-like domain of *Arabidopsis* nuclear α -amylases is critical for DNA sequence recognition and transcriptional activation. *Plant Cell* 26, 1746–1763. doi: 10.1105/tpc.114.123703
- Sreenivasulu, N., Usadel, B., Winter, A., Radchuk, V., Scholz, U., Stein, N., et al. (2008). Barley grain maturation and germination: metabolic pathway and regulatory network commonalities and differences highlighted by new MapMan/PageMan profiling tools. *Plant Physiol.* 146, 1738–1758. doi: 10.1104/pp.107.111781
- Steber, C. M., and McCourt, P. (2001). A role for brassinosteroids in germination in *Arabidopsis*. *Plant Physiol.* 125, 763–769. doi: 10.1104/pp.125.2.763
- Steffens, B., Kovalev, A., Gorb, S. N., and Sauter, M. (2012). Emerging roots alter epidermal cell fate through mechanical and reactive oxygen species signaling. *Plant Cell* 24, 3296–3306. doi: 10.1105/tpc.112.101790
- Sun, X., Wang, Y., and Sui, N. (2018). Transcriptional regulation of bHLH during plant response to stress. *Biochem. Bioph. Res. Co.* 503, 397–401. doi: 10.1016/j.bbrc.2018.07.123
- Tanaka, T., Antonio, B. A., Kikuchi, S., Matsumoto, T., Nagamura, Y., Numa, H., et al. (2008). The rice annotation project database (RAP-DB). *Nucleic Acids Res.* 36, D1028–D1033. doi: 10.1093/nar/gkm978
- Tang, H., Bi, H., Liu, B., Lou, S., Song, Y., Tong, S., et al. (2021). WRKY33 interacts with WRKY12 protein to up-regulate RAP2.2 during submergence induced hypoxia response in *Arabidopsis thaliana*. *New Phytol.* 229, 106–125. doi: 10.1111/nph.17020
- Tang, N., Ma, S., Zong, W., Yang, N., Lv, Y., Yan, C., et al. (2016). MODD mediates deactivation and degradation of OsbZIP46 to negatively regulate ABA signaling and drought resistance in rice. *Plant Cell* 28, 2161–2177. doi: 10.1105/tpc.16.00171
- Tang, N., Zhang, H., Li, X., Xiao, J., and Xiong, L. (2012). Constitutive activation of transcription factor OsbZIP46 improves drought tolerance in rice. *Plant Physiol.* 158, 1755–1768. doi: 10.1104/pp.111.190389
- Tang, Y., Wang, J., Bao, X., Liang, M., Lou, H., Zhao, J., et al. (2019). Genome-wide identification and expression profile of HD-ZIP genes in physic nut and functional analysis of the *JcHDZ16* gene in transgenic rice. *BMC Plant Biol.* 19:298. doi: 10.1186/s12870-019-1920-x
- Tatematsu, K., Nakabayashi, K., Kamiya, Y., and Nambara, E. (2008). Transcription factor AtTCP14 regulates embryonic growth potential during seed germination in *Arabidopsis thaliana*. *Plant J.* 53, 42–52. doi: 10.1111/j.1365-3113X.2007.03308.x
- Tong, H., Xiao, Y., Liu, D., Gao, S., Liu, L., Yin, Y., et al. (2014). Brassinosteroid regulates cell elongation by modulating gibberellin metabolism in rice. *Plant Cell* 26, 4376–4393. doi: 10.1105/tpc.114.132092
- Tong, H. N., and Chu, C. C. (2012). Brassinosteroid signaling and application in rice. *J. Genet. Genomics* 39, 3–9. doi: 10.1016/j.jgg.2011.12.001
- Tsai, K. J., Lin, C. Y., Ting, C. Y., and Shih, M. C. (2014). Ethylene plays an essential role in the recovery of *Arabidopsis* during post-anaerobiosis reoxygenation. *Plant Cell Environ.* 37, 2391–2405. doi: 10.1111/pce.12292
- van Veen, H., Vashisht, D., Akman, M., Girke, T., Mustroph, A., Reinen, E., et al. (2016). Transcriptomes of eight *Arabidopsis thaliana* accessions reveal core conserved, genotype- and organ-specific responses to flooding stress. *Plant Physiol.* 172, 668–689. doi: 10.1104/pp.16.00472
- Viana, V. E., Marini, N., Busanello, C., Pegoraro, C., Fernando, J. A., Da Maia, L. C., et al. (2018). Regulation of rice responses to submergence by WRKY transcription factors. *Biol. Plant* 62, 551–560. doi: 10.1007/s10535-018-0806-3
- Waller, F., Furuya, M., and Nick, P. (2002). OsARF1, an auxin response factor from rice, is auxin-regulated and classifies as a primary auxin responsive gene. *Plant Mol. Biol.* 50, 415–425. doi: 10.1023/A:1019818110761
- Wang, C., Ding, Y., Yao, J., Zhang, Y., Sun, Y., Colee, J., et al. (2015). *Arabidopsis* Elongator subunit 2 positively contributes to resistance to the necrotrophic fungal pathogens *Botrytis cinerea* and *Alternaria brassicicola*. *Plant J.* 83, 1019–1033. doi: 10.1111/tpj.12946
- Wang, K., Ding, Y., Cai, C., Chen, Z., and Zhu, C. (2018). The role of C2H2 zinc finger proteins in plant responses to abiotic stresses. *Physiol. Plant.* 165, 690–700. doi: 10.1111/pp.12728
- Wang, Z. Y., Bai, M. Y., Oh, E., and Zhu, J. Y. (2012). Brassinosteroid signaling network and regulation of photomorphogenesis. *Annu. Rev. Genet.* 46, 701–724. doi: 10.1146/annurev-genet-102209-163450
- Washio, K. (2003). Functional dissections between GAMYB and DOF transcription factors suggest a role for protein-protein associations in the gibberellin-mediated expression of the *RAmy1A* gene in the rice aleurone. *Plant Physiol.* 133, 850–863. doi: 10.1104/pp.103.027334
- Weiste, C., and Dröge-Laser, W. (2014). The *Arabidopsis* transcription factor bZIP11 activates auxin-mediated transcription by recruiting the histone acetylation machinery. *Nat. Commun.* 5:3883. doi: 10.1038/ncomms4883
- Woodger, F. J., Gubler, F., Pogson, B., and Jacobsen, J. V. (2003). The role of GAMYB transcription factors in GA-regulated gene expression. *J. Plant Growth Regul.* 22, 176–184. doi: 10.1007/s00344-003-0025-8
- Wu, Q., Li, D., Li, D., Liu, X., Zhao, X., Li, X., et al. (2015). Overexpression of *OsDof12* affects plant architecture in rice (*Oryza sativa* L.). *Front. Plant Sci.* 6:833. doi: 10.3389/fpls.2015.00833
- Wu, Y. S., and Yang, C. Y. (2020). Comprehensive transcriptomic analysis of auxin responses in submerged rice coleoptile growth. *Int. J. Mol. Sci.* 21:1292. doi: 10.3390/ijms21041292
- Xi, W., Liu, C., Hou, X., and Yu, H. (2010). MOTHER OF FT AND TFL1 regulates seed germination through a negative feedback loop modulating ABA signaling in *Arabidopsis*. *Plant Cell* 22, 1733–1748. doi: 10.1105/tpc.109.073072
- Xi, W., and Yu, H. (2010). MOTHEROF TFL1 regulates seed germination and fertility relevant to the brassinosteroid signaling pathway. *Plant Signal. Behav.* 5, 1315–1317. doi: 10.4161/psb.5.10.13161
- Xiong, Q., Ma, B., Lu, X., Huang, Y.-H., He, S.-J., Yang, C., et al. (2017). Ethylene-inhibited jasmonic acid biosynthesis promotes mesocotyl/coleoptile elongation of etiolated rice seedlings. *Plant Cell* 29, 1053–1072. doi: 10.1105/tpc.16.00981
- Xu, K., Xu, X., Fukao, T., Canlas, P., Maghirang-Rodriguez, R., Heuer, S., et al. (2006). Sub1A is an ethylene-response-factor-like gene that confers submergence tolerance to rice. *Nature* 442, 705–708. doi: 10.1038/nature04920
- Yamamuro, C., Ihara, Y., Wu, X., Noguchi, T., Fujioka, S., Takatsuto, S., et al. (2000). Loss of function of a rice brassinosteroid insensitive1 homolog prevents internode elongation and bending of the lamina joint. *Plant Cell* 12, 1591–1606. doi: 10.1105/tpc.12.9.1591
- Yang, C., Lu, X., Ma, B., Chen, S. Y., and Zhang, J. S. (2015). Ethylene signaling in rice and *Arabidopsis*: conserved and diverged aspects. *Mol. Plant* 4, 495–505. doi: 10.1016/j.molp.2015.01.003
- Yang, Y., Li, J., Li, H., Yang, Y., Guang, Y., and Zhou, Y. (2019). The bZIP gene family in watermelon: Genome-wide identification and expression analysis under cold stress and root-knot nematode infection. *Peer J.* 7:e7878. doi: 10.7717/peerj.7878
- Yao, Y., He, R. J., Xie, Q. L., Zhao, X. H., Deng, X. M., He, J. B., et al. (2017). ETHYLENE RESPONSE FACTOR 74 (ERF74) plays an essential role in controlling a respiratory burst oxidase homolog D (RbohD)-dependent mechanism in response to different stresses in *Arabidopsis*. *New Phytol.* 213, 1667–1681. doi: 10.1111/nph.14278
- Yeung, E., van Veen, H., Vashisht, D., Paiva, A. L. S., Hummel, M., Rankenb, T., et al. (2018). A stress recovery signaling network for enhanced flooding tolerance in *Arabidopsis thaliana*. *Proc. Natl Acad. Sci. U.S.A.* 115, E6085–E6094. doi: 10.1101/276519
- Yilmaz, A., Mejia-Guerra, M. K., Kurz, K., Liang, X., Welch, L., and Grotewold, E. (2011). AGRIS: the *Arabidopsis* gene regulatory information server, an update. *Nucleic Acids Res.* 39, D1118–D1122. doi: 10.1093/nar/gkq1120
- Yong, Y., Zhang, Y., and Lyu, Y. (2019). A stress-responsive NAC transcription factor from tiger lily (*LINAC2*) interacts with LIDREB1 and LIZHFD4 and enhances various abiotic stress tolerance in *Arabidopsis*. *Int. J. Mol. Sci.* 20:3225. doi: 10.3390/ijms20133225

- Yoshida, T., Fujita, Y., Sayama, H., Kidokoro, S., Maruyama, K., Mizoi, J., et al. (2010). AREB1, AREB2, and ABF3 are master transcription factors that cooperatively regulate ABRE dependent ABA signalling involved in drought stress tolerance and require ABA for full activation. *Plant J.* 61, 672–685. doi: 10.1111/j.1365-313X.2009.04092.x
- Yoshii, M., Yamazaki, M., Rakwal, R., Kishi-Kaboshi, M., Miyao, A., and Hirochika, H. (2010). The NAC transcription factor RIM1 of rice is a new regulator of jasmonate signaling. *Plant J.* 61, 804–815. doi: 10.1111/j.1365-313X.2009.04107.x
- Yu, L. H., Wu, J., Zhang, Z. S., Miao, Z. Q., Zhao, P. X., Wang, Z., et al. (2017). *Arabidopsis* MADS-Box transcription factor AGL21 acts as environmental surveillance of seed germination by regulating ABI5 expression. *Mol. Plant.* 10, 834–845. doi: 10.1016/j.molp.2017.04.004
- Yu, X., Li, L., Zola, J., Aluru, M., Ye, H., Foudree, A., et al. (2011). A brassinosteroid transcriptional network revealed by genome-wide identification of BES1 target genes in *Arabidopsis thaliana*. *Plant J.* 65, 634–646. doi: 10.1111/j.1365-313X.2010.04449.x
- Yu, Z., Chang, F., Lv, W., Sharmin, R. A., Wang, Z., Kong, J., et al. (2019). Identification of QTN and candidate gene for seed-flooding tolerance in soybean [*Glycine max* (L.) Merr.] using Genome-Wide Association Study (GWAS). *Genes* 10:957. doi: 10.3390/genes10120957
- Yuan, X., Wang, H., Cai, J., Bi, Y., Li, D., and Song, F. (2019a). Rice NAC transcription factor ONAC066 functions as a positive regulator of drought and oxidative stress response. *BMC Plant Biol.* 19:278. doi: 10.1186/s12870-019-1883-y
- Yuan, X., Wang, H., Cai, J., Li, D., and Song, F. (2019b). NAC transcription factors in plant immunity. *Phytopathol. Res.* 1:3. doi: 10.1186/s42483-018-0008-0
- Zhang, C., Bai, M. Y., and Chang, K. (2014). Brassinosteroid-mediated regulation of agronomic traits in rice. *Plant Cell Rep.* 33, 683–696. doi: 10.1007/s00299-014-1578-7
- Zhang, C., Li, C., Liu, J., Lv, Y., Yu, C., Li, H., et al. (2017). The *OsABF1* transcription factor improves drought tolerance by activating the transcription of COR413-TM1 in rice. *J. Exp. Bot.* 68, 4695–4707. doi: 10.1093/jxb/erx260
- Zhang, L., Li, Z., Quan, R., Li, G., Wang, R., and Huang, R. (2011). An AP2 domain-containing gene, ESE1, targeted by the ethylene signaling component EIN3 is important for the salt response in *Arabidopsis*. *Plant Physiol.* 157, 854–865. doi: 10.1104/pp.111.179028
- Zhang, L., Xia, C., Zhao, G., Liu, J., Jia, J., and Kong, X. (2015). A novel wheat bZIP transcription factor, TabZIP60, confers multiple abiotic stress tolerances in transgenic *Arabidopsis*. *Physiol. Plant* 153, 538–554. doi: 10.1111/pp.12261
- Zhang, Y., Mayba, O., Pfeiffer, A., Shi, H., Tepperman, J. M., Speed, T. P., et al. (2013). A quartet of PIF bHLH factors provides a transcriptionally centered signaling hub that regulates seedling morphogenesis through differential expression-patterning of shared target genes in *Arabidopsis*. *PLoS Genet.* 9:e1003244. doi: 10.1371/journal.pgen.1003244
- Zhao, Q., Fan, Z., Qiu, L., Che, Q., Wang, T., Li, Y., et al. (2020). *MdbHLH130*, an Apple bHLH transcription factor, confers water stress resistance by regulating stomatal closure and ROS homeostasis in transgenic tobacco. *Front. Plant Sci.* 11:543696. doi: 10.3389/fpls.2020.543696
- Zhao, X., Dou, L., Gong, Z., Wang, X., and Mao, T., (2019). BES1 hinders ABSCISIC ACID INSENSITIVE5 and promotes seed germination in *Arabidopsis*. *New Phytol.* 221, 908–918. doi: 10.1111/nph.15437
- Zhong, S., Shi, H., Xue, C., Wang, L., Xi, Y., Li, J., et al. (2012). A molecular framework of light-controlled phytohormone action in *Arabidopsis*. *Curr. Biol.* 22, 1530–1535. doi: 10.1016/j.cub.2012.06.039
- Zhu, Q. G., Gong, Z. Y., Huang, J., Grierson, D., Chen, K. S., and Yin, X. R. (2019). High-CO₂/hypoxia-responsive transcription factors *DkERF24* and *DkWRKY1* interact and activate *DkPDC2* promoter. *Plant Physiol.* 180, 621–633. doi: 10.1104/pp.18.01552
- Zou, X., Neuman, D., and Shen, Q. J. (2008). Interactions of two transcriptional repressors and two transcriptional activators in modulating gibberellin signaling in aleurone cells. *Plant Physiol.* 148, 176–186. doi: 10.1104/pp.108.123653
- Zubo, Y. O., and Schaller, G. E. (2020). Role of the cytokinin-activated Type-B response regulators in hormone crosstalk. *Plants* 9:166. doi: 10.3390/plants9020166

Conflict of Interest: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2021 Mohanty. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



***KLF4*, a Key Regulator of a Transitive Triplet, Acts on the TGF- β Signaling Pathway and Contributes to High-Altitude Adaptation of Tibetan Pigs**

OPEN ACCESS

Edited by:

Deborah A. Triant,
University of Virginia, United States

Reviewed by:

Lorna Grindlay Moore,
University of Colorado, United States
Anna Shestakova,
University of Michigan, United States

*Correspondence:

Xibiao Wang
wangxibiao@neau.edu.cn
Zhipeng Wang
wangzhipeng@neau.edu.cn

[†]These authors have contributed
equally to this work and share first
authorship

Specialty section:

This article was submitted to
Evolutionary and Population Genetics,
a section of the journal
Frontiers in Genetics

Received: 11 November 2020

Accepted: 10 March 2021

Published: 15 April 2021

Citation:

Wang T, Guo Y, Liu S, Zhang C,
Cui T, Ding K, Wang P, Wang X and
Wang Z (2021) *KLF4*, a Key Regulator
of a Transitive Triplet, Acts on
the TGF- β Signaling Pathway
and Contributes to High-Altitude
Adaptation of Tibetan Pigs.
Front. Genet. 12:628192.
doi: 10.3389/fgene.2021.628192

**Tao Wang^{1,2†}, Yuanyuan Guo^{1,2†}, Shengwei Liu^{1,2}, Chaoxin Zhang^{1,2}, Tongyan Cui^{1,2},
Kun Ding³, Peng Wang⁴, Xibiao Wang^{1*} and Zhipeng Wang^{1,2*}**

¹ College of Animal Science and Technology, Northeast Agricultural University, Harbin, China, ² Bioinformatics Center, Northeast Agricultural University, Harbin, China, ³ College of Computer Science and Technology, Inner Mongolia Normal University, Hohhot, China, ⁴ HeiLongJiang Provincial Husbandry Department, Harbin, China

Tibetan pigs are native mammalian species on the Tibetan Plateau that have evolved distinct physiological traits that allow them to tolerate high-altitude hypoxic environments. However, the genetic mechanism underlying this adaptation remains elusive. Here, based on multitissue transcriptional data from high-altitude Tibetan pigs and low-altitude Rongchang pigs, we performed a weighted correlation network analysis (WGCNA) and identified key modules related to these tissues. Complex network analysis and bioinformatics analysis were integrated to identify key genes and three-node network motifs. We found that among the six tissues (muscle, liver, heart, spleen, kidneys, and lungs), lung tissue may be the key organs for Tibetan pigs to adapt to hypoxic environment. In the lung tissue of Tibetan pigs, we identified *KLF4*, *BCL6B*, *EGR1*, *EPAS1*, *SMAD6*, *SMAD7*, *KDR*, *ATOH8*, and *CCN1* genes as potential regulators of hypoxia adaption. We found that *KLF4* and *EGR1* genes might simultaneously regulate the *BCL6B* gene, forming a *KLF4*–*EGR1*–*BCL6B* complex. This complex, dominated by *KLF4*, may enhance the hypoxia tolerance of Tibetan pigs by mediating the TGF- β signaling pathway. The complex may also affect the PI3K-Akt signaling pathway, which plays an important role in angiogenesis caused by hypoxia. Therefore, we postulate that the *KLF4*–*EGR1*–*BCL6B* complex may be beneficial for Tibetan pigs to survive better in the hypoxia environments. Although further molecular experiments and independent large-scale studies are needed to verify our findings, these findings may provide new details of the regulatory architecture of hypoxia-adaptive genes and are valuable for understanding the genetic mechanism of hypoxic adaptation in mammals.

Keywords: Tibetan pig, multitissue, transcriptome, hypoxia adaptation, gene network

INTRODUCTION

Hypoxia is a significant environmental characteristic of high altitude, which exerts a marked impact on biological organisms and imposes extreme physiological challenges in mammals. The Tibetan pig was originally distributed at altitudes of 2,900–4,300 m in the Tibetan Plateau (Ai et al., 2014). Physiological studies showed that Tibetan pigs have evolved physiological adaptations to high-altitude hypoxia, such as a thicker alveolar septum with more highly developed capillaries (Ma et al., 2019) and a larger and strong heart (Li et al., 2013). Therefore, they represent a suitable animal model for exploring the molecular mechanism of hypoxia adaptation in high-altitude organisms.

With the development of sequencing technology, the majority of studies have explored the genetic basis of hypoxic adaptation in Tibetan pigs from the perspective of selection signals (Li et al., 2013, 2016; Ai et al., 2014; Huang et al., 2019; Ma et al., 2019; Shang et al., 2020) or by using differential expression analysis between differential conditional gene expression in one tissue based on the transcriptome (Jia et al., 2016; Zhang B. et al., 2017) up to the present. Although previous studies have identified the *EPAS1*, *HIF1A*, *EGLN1*, *RGCC*, *KLF6*, *TGFB2*, *EGLN3*, and *ACE* genes related to hypoxia, these genes may only explain a minority of genetic variance due to the case of the missing heritability. Therefore, the most detailed solution to the missing heritability problem would involve identifying all causal genetic variants (Young, 2019) and exploring related gene networks that have facilitated high-altitude adaptation of Tibetan pigs.

The adaptation of Tibetan pigs to hypoxia is a very complex biological process that may involve multiple genes and transcriptional regulation among genes. The gene network provides a systemic view of gene regulation by the coordinated activity of multiple genes and regulatory factors and serves as a medium for understanding the mechanism of gene regulation (Narang et al., 2015). Based on the gene expression profile, a gene network was constructed by quantitative modeling, which can be used for rational design of molecular approaches to target specific biological processes (Nishio et al., 2008) and infer new biological functions (Cheng and Gerstein, 2012; McLeay et al., 2012). Although gene expression status cannot completely determine gene function, constructing gene network based on gene expression profile may be a feasible method to explore the mechanism of hypoxia adaptation. Moreover, the gene network cannot only intuitively elucidate the regulatory relationship between genes but also identify important hub genes. These hub genes represent candidates for further experimental investigation and potential biomarkers for complex traits (Döhr et al., 2005; Buckingham and Rigby, 2014; Chen et al., 2016).

Transcription factors (TFs) and microRNAs (miRNAs) regulate gene expression at the transcriptional and posttranscriptional levels, respectively. They coordinately control the dynamics and output of gene transcription and tightly control spatial and temporal patterns of gene expression. Therefore, constructing a gene regulatory network involving TFs and miRNAs is helpful in understanding the regulatory mechanism of genes in adaptation to hypoxia.

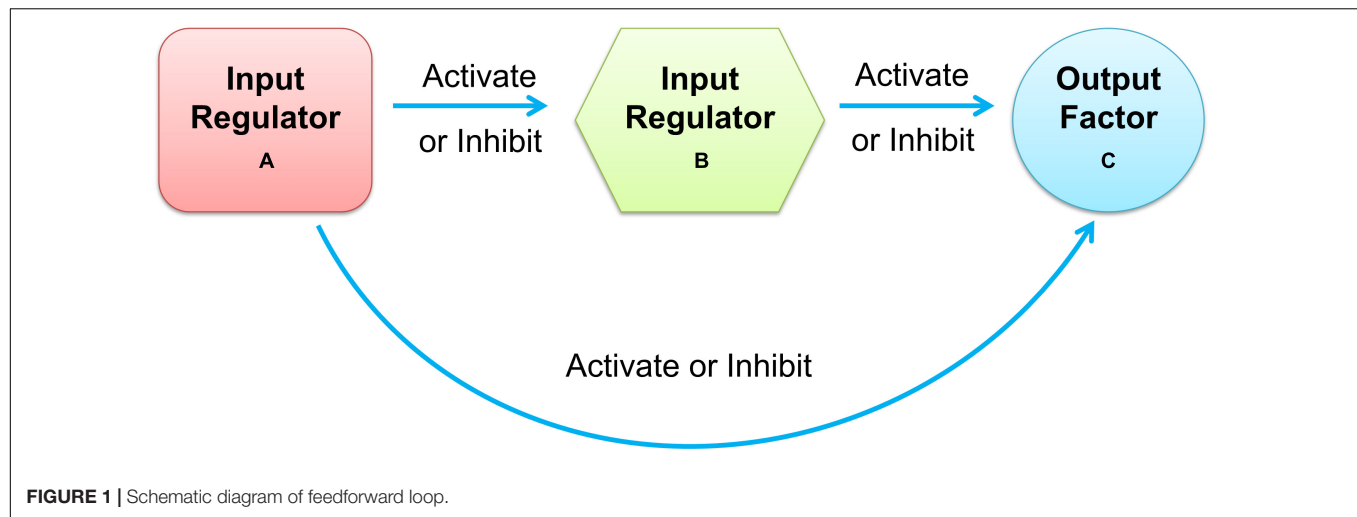
Moreover, most cellular tasks are not performed by individual genes but by groups of functionally associated genes, generally referred to as modules. In a gene regulatory network, modules appear as groups of densely interconnected nodes, also called communities or clusters (Adamcsek et al., 2006). Among these clusters of gene regulatory networks, size-3 network motifs were suggested to be recurring circuit elements that carry out key information processing tasks. The three-node motif included 13 types of connected subgraphs, such as V-out, 3-Chain, feed forward loop (FFL), 3-Loop, and Clique. Among them, the FFL motif consists of two input regulators, A and B, and one output factor C, where regulators A and B regulate target factor C together, and A also regulates B, as shown in **Figure 1**. According to the regulation functions (activate or inhibit) between the three elements in FFL, it can be divided into two categories: coherent feed-forward loop and incoherent feed-forward loop (Mangan et al., 2003; Alon, 2007). In the coherent feed-forward loop, the regulators strengthen each other's functions, and have the effect of controlling stability and resisting noise in biological networks (Le and Kwon, 2013). In the incoherent feed-forward loop, the regulator performs the opposite function to speed up the response and suppress the delay (Kim et al., 2008; Lan and Tu, 2013). FFL plays an important role in biological processing (Milo et al., 2002; Mangan and Alon, 2003), which appears in hundreds of gene systems in *Escherichia coli*, yeast, fruit fly, and humans. The FFL motif governs many aspects of normal cell functions, such as creating bistable switches of gene expression in developing tissues for spatial avoidance, controlling the time sequence of gene expression to create temporal avoidance, and minimizing expression fluctuation against noise (Shalgi et al., 2009).

The Tibetan pig and Rongchang pig are two indigenous pig breeds in China. Rongchang pigs are cross-fertile relatives of Tibetan pigs, living in geographically neighboring low-altitude regions (Tang et al., 2017). Ai et al. (2014) found that Tibetan pigs had a close genetic distance with Rongchang pigs through neighbor-joining (NJ) tree analysis. In this study, based on transcriptional data from six tissues in Tibetan pigs and Rongchang pigs, the key module of lung tissues was identified by constructing a gene network. By integrating complex network analysis and bioinformatics analysis, we identified key genes and size-3 network motifs and found that *KLF4*, a key regulator of the complex, may enhance the survival ability of Tibetan pigs by mediating the TGF- β signaling pathway in the hypoxia environments. This study provides a valuable clue to further understand the molecular mechanism of adaptability in high-altitude hypoxia.

MATERIALS AND METHODS

Gene Expression Data Collection

The protein-encoding genes and miRNA expression profile data of six tissues (muscle, liver, heart, spleen, kidneys, and lungs) from three Tibetan pigs and three Rongchang pigs were obtained from the Gene Expression Omnibus (GEO) database at the National Center for Biotechnology Information (NCBI) under



accession numbers GSE93855 (provided by Tang et al., 2017) and GSE124418 (uploaded by Long et al., 2019), respectively. For details about the experimental animals, please refer to the **Supplementary Material**. Gene list, expressed in each tissue, was updated based on the Sus scrofa 11.1 genome assembly. Taking into account that genes are with very low expression and are less reliable and indistinguishable from the sampling noise, we selected the top 50% of protein-encoding genes of the median absolute deviation (MAD) of expression level.

Co-expression Network Analysis

Network analysis was performed according to the protocol of the WGCNA R package (Langfelder and Horvath, 2008). We used the following criteria to identify the key module of each tissue: (1) the p -value of the correlation between the module and the tissue was less than 3.97×10^{-4} (0.05/126) using the Bonferroni correction method, and (2) the median of the gene significance (GS) value was greater than 0.8. In addition, we calculated the fundamental topology concepts of each key module, including density, mean cluster coefficient, centralization, and heterogeneity.

Analysis of Gene Expression Patterns in Multiple Tissues

In this study, we used the Mfuzz package in R (Kumar and Futschik, 2007) to identify multitissue expression patterns of each gene in each key module. Based on the fuzzy c -means algorithm, this software implements soft clustering methods for microarray data analysis, which makes the clustering process less sensitive to noise and effectively reflects the strength of a gene's association with a given cluster.

Gene Tissue-Specific Analysis

We used the tissue-specificity index (TSI, τ) (Yanai et al., 2005) to grade the scalar measure of the specificity of an expression profile, which ranged from 0 for housekeeping genes to 1 for strictly TS genes. According to Yanai et al. (2005), genes with TSI > 0.9 were considered TS genes.

Functional Enrichment Analysis of Genes in Key Modules

We used the online software DAVID (v6.8) (Huang et al., 2009) to perform functional enrichment analysis of genes in each key module with all protein-encoding genes on pig genome as the background genes set, including gene ontology (GO) and KEGG pathway analysis.

Identification of Hub Genes in Key Modules

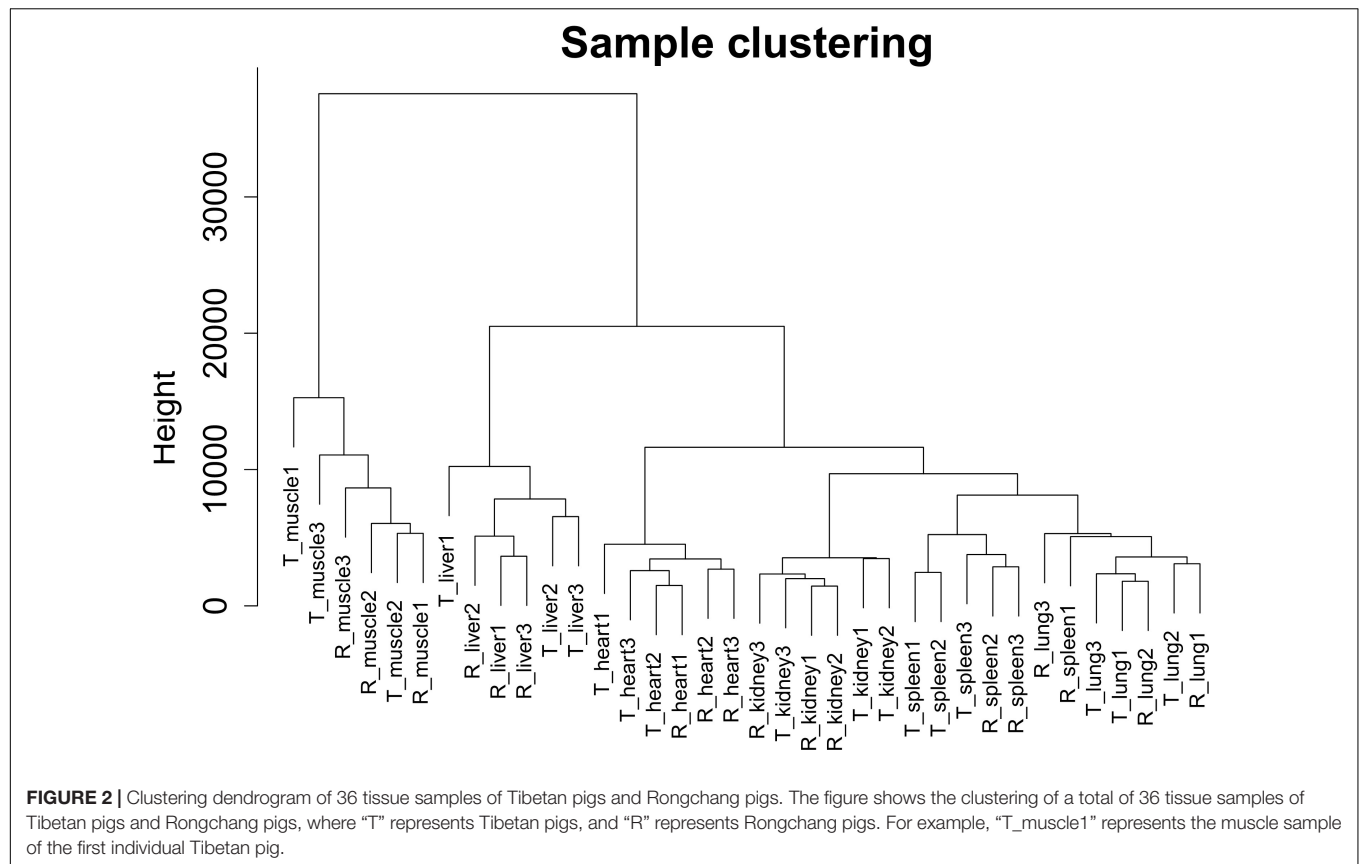
We identified the hub genes in each key module according to the following criteria: (1) GS value of the gene ≥ 0.8 , (2) module membership (MM) value of the gene ≥ 0.95 , and (3) in each module, Kwithin ranked in the top 20% of the genes. The module membership (MM) is defined as the correlation of the module eigengenes and the gene expression profile. The Kwithin value represents the degree of connectivity of edges located under the same module as the gene.

Gene Regulatory Network Construction

We used the TFBSTools package in R (Tan and Lenhard, 2016) to predict the target genes of TFs in each key module of Tibetan pigs. The relScore value was set to 0.85, and other parameters were defaulted. Based on the miRanda tool (Enright et al., 2003), we predicted target genes of the miRNAs, and the Tot Score and Tot Energy were set to 140 and -20 , respectively. The gene regulatory network in each Tibetan pig tissue was constructed by combining TFs, miRNAs, target genes, co-expressed genes, hub genes, and their interactions.

Motif Analysis of the Gene Regulatory Network

The three-node motifs in the gene regulatory network of each tissue were obtained using mfinder1.2 (Kashtan et al., 2004). The number of random networks was set to 10,000. The Z score describes the significance of the difference between the frequency of motifs in the real network and that in the corresponding



randomized network. The significance profile (SP) is the vector of Z scores normalized to length 1, describing the statistical significance of each motif in the network (Milo et al., 2004). We constructed the triad significance profile (TSP) of the six tissues from Tibetan pigs, which display certain relations between subgraph types.

Identification of Important Genes and Size-3 Subgraphs in the Lung-Specific Gene Regulation Network

We calculated the importance scores of each node, edge, and size-3 sub-graph in the lung-specific gene regulatory network. Each node was scored according to the connectivity, differential expression between different conditions, tissue-specific expression, and TF characteristics. The score of each edge was the weight value of the edge from WGCNA. The score of each candidate size-3 sub-graph was calculated by combining the node score and the edge score. The calculation formula and specific details are referred to the **Supplementary Material**.

Verification of Important Genes in Lung Tissue

The lung tissue expression profiles of three Tibetan sheep and three yaks were obtained from the GEO database (accession: GSE93855) (Tang et al., 2017), and the expression profiles in the lung tissue of four Diqing Tibetan pigs were collected from

another dataset (accession: GSE84409) (Jia et al., 2016). We used WGCNA to perform module analysis. The Hmisc package in R was used to statistically test the correlation between genes.

RESULTS

WGCNA and Identification of Key Modules in Tissues

In this study, we selected 5,723 protein-coding genes (top 50% of the MAD value) for subsequent analysis. Cluster analysis revealed that different samples from the same tissue of Tibetan pigs and Rongchang pigs clustered together, and no outlier samples were observed, as shown in **Figure 2**.

We constructed a co-expression network for Tibetan pigs and Rongchang pigs. To fulfill the criteria of approximate scale-free topology, the soft threshold power β was set to 20 [the scale-free topological index $R^2 = 0.85$ for Tibetan pigs (**Figure 3A**) and $R^2 = 0.80$ for Rongchang pigs (**Supplementary Figure 1A**)]. Through hierarchic clustering and dynamic branch cutting procedures, 21 and 20 merged modules were identified in the co-expression network of Tibetan pigs and Rongchang pigs, respectively. Clustering of the modules is shown in **Figure 3B**.

Next, the GS values of genes contained in these modules and the correlation between each module and different tissues in Tibetan pigs (**Figure 3C**) were calculated. According to the screening criteria, key modules from six tissues (muscle,

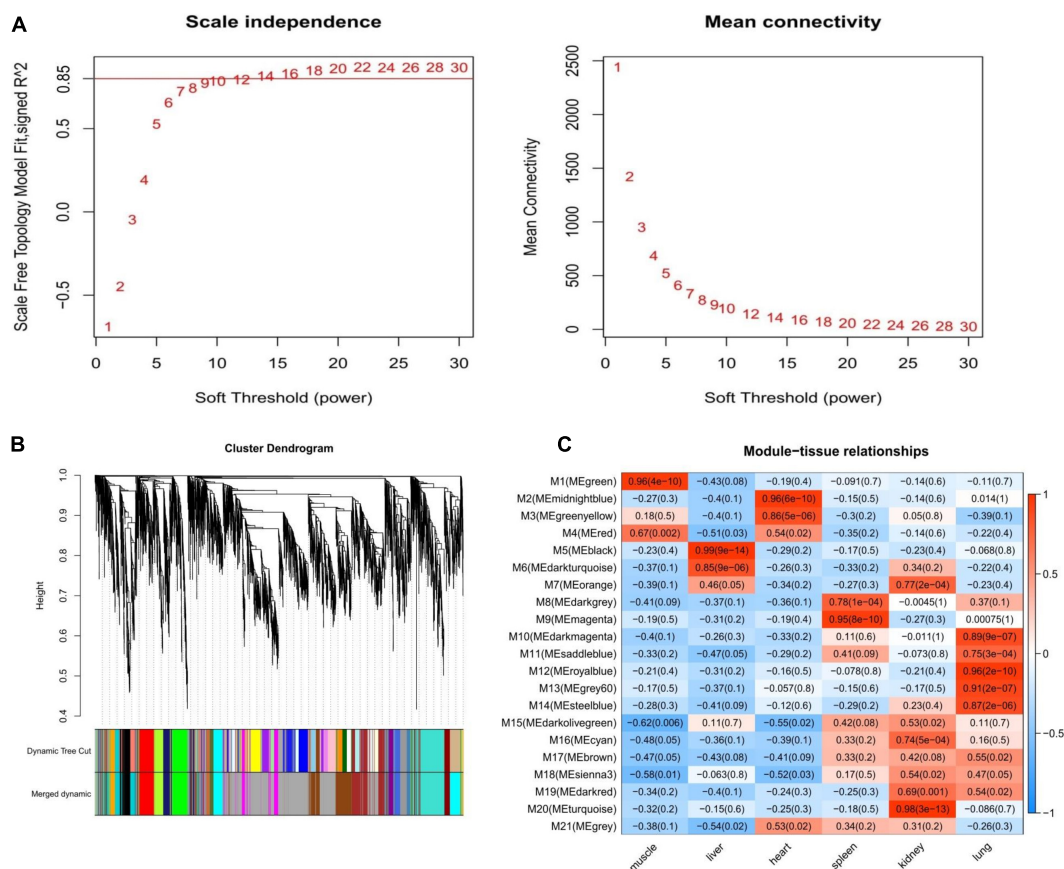


FIGURE 3 | Weighted gene co-expression network analysis of Tibetan pigs. **(A)** Network topology of different soft-thresholding power of Tibetan pig co-expression network. The left panel displays the influence of soft-thresholding power (x -axis) on scale-free fit index (y -axis). The right panel shows the influence of soft-thresholding power (x -axis) on the mean connectivity (degree, y -axis). **(B)** Gene clustering module of Tibetan pig co-expression network. The dissimilarity was based on topological overlap. The “Merged dynamic” is the result of merging modules with a correlation higher than 0.9. The y -axis is the distance determined by the extent of topological overlap. **(C)** Heatmap of the correlation between module eigengenes and the six tissues of Tibetan pigs. The x -axis is the six tissues of Tibetan pigs, and the y -axis is the module eigengene (ME). In the heatmap, red represents high adjacency (positive correlation), and blue represents low adjacency (negative correlation). In brackets is the p -value of the correlation test.

liver, heart, spleen, kidneys, and lungs) in Tibetan pigs were determined. These modules contained 267, 215, 157, 201, 420, and 350 genes, respectively. The list of genes and the co-expression network for each tissue are shown in **Supplementary Table 1** and **Supplementary Figure 2**, respectively. The gene list of each key module of Rongchang pigs are shown in **Supplementary Table 1**.

Network Topology Analysis

We calculated the network topology of the key module of each tissue from the Tibetan pigs and Rongchang pigs, including density, mean cluster coefficient, centralization, and heterogeneity. The results are shown in **Table 1**. We observed that the network density and the clustering coefficient of Tibetan pig lung and heart tissues were the lowest, while those of the spleen were the highest. These network concepts indicated that the key modules of the lungs and heart were a sparse network. The network topology of Rongchang pigs was similar to those of Tibetan pigs.

Multitissue Gene Expression Patterns

According to the analysis of gene expression patterns, we found that compared with other tissues, Tibetan pigs and Rongchang pigs had the largest differences in gene expression patterns of the key modules of lung tissue. In the key module of lung tissue, gene expression patterns in multitissues could be divided into eight clusters. Compared with other tissues, the level of gene expression in the lung tissues of Tibetan pigs was the highest in all clusters (**Figure 4A**). However, the genes in the lung tissues of Rongchang pigs are expressed as the highest only in cluster 1, cluster 5, and cluster 6 (**Figure 4B**). The different gene expression patterns may be caused by physiological changes in the hypoxia environment of Tibetan pigs or interspecies differences between Tibetan pigs and Rongchang pigs.

Tissue-Specific Gene Analysis

A total of 210 and 206 genes were identified as tissue specific ($\tau > 0.9$) in the key modules of Tibetan pigs and Rongchang pigs, respectively. For the gene details, see **Supplementary Table 2**. For

TABLE 1 | The fundamental network topology concepts of key modules in Tibetan pig and Rongchang pig tissues.

Pig breed	Tissue	Key module	DS	MCC	CL	HG
Tibetan pig	Muscle	M1	0.03	0.26	0.10	1.12
	Liver	M5	0.05	0.21	0.13	1.10
	Heart	M2	0.03	0.14	0.09	0.91
	Spleen	M9	0.12	0.28	0.17	0.80
	Kidney	M20	0.05	0.17	0.11	0.84
	Lung	M22	0.03	0.13	0.08	0.82
Rongchang	Muscle	M14	0.05	0.18	0.12	1.01
	Liver	M8	0.05	0.20	0.13	1.08
	Heart	M21	0.05	0.18	0.11	0.97
	Spleen	M3	0.08	0.22	0.14	0.81
	Kidney	M13	0.07	0.20	0.14	0.83
	Lung	M1	0.05	0.17	0.10	0.76

DS, network density; MCC, mean cluster coefficient; CL, network centralization; HG, network heterogeneity.

Tibetan pigs, there were 32, 50, 23, 36, 47, and 22 TS genes in the key modules of the muscle, liver, heart, spleen, kidneys and lungs, respectively. There are more TS genes in the lung tissues of Tibetan pigs than in Rongchang pigs. Compared with the other five tissues, the number of TS genes in the lung has the largest difference between the two pig breeds.

Functional Enrichment Analysis of Genes in Key Modules

To further understand the biological functions of genes in each key module in Tibetan pigs and Rongchang pigs, we conducted gene function enrichment analysis. After the Benjamini correction, we identified significant pathway enrichment in Tibetan pigs and Rongchang pigs, as shown in **Supplementary Table 3**. Compared with Rongchang pigs, there were 10, 4, 1, and 13 pathways in the muscle, lungs, heart, and spleen that were only significantly enriched in Tibetan pigs, as shown in **Table 2**. Pathways enriched only in Tibetan pig lungs, including regulated cell growth, proliferation, migration, and apoptosis include focal adhesion (ssc04510), ECM-receptor

interaction (ssc04512), PI3K–Akt signaling pathway (ssc04151), and TGF- β signaling pathway (ssc04350). In addition, 30 pathways were significantly enriched only in Rongchang pigs (see **Supplementary Table 4**).

Identification of Hub Genes in Key Modules

According to the screening criteria of hub genes, we have determined the hub genes of each key module of Tibetan pigs and Rongchang pigs. Compared with Rongchang pigs, Tibetan pigs had more hub genes in the liver, kidneys, and lung tissues. There was no hub gene overlap between the lung tissues of Tibetan pigs and Rongchang pigs. In addition, eight hub genes were TS gene in the lung tissue of Tibetan pigs, while Rongchang pigs only have one. **Table 3** summarizes the hub gene information in Tibetan pigs and Rongchang pigs.

Gene Regulatory Network Construction

The gene regulatory network of Tibetan pig tissues was constructed by combining TFs, miRNAs, target genes, co-expression genes, and hub genes. There were 115, 80, 35, 117, 160, and 157 nodes (genes) and 986, 1,786, 298, 1,976, 5,315, and 1,075 edges (regulatory relationship) in the gene regulatory network of the muscle, liver, heart, spleen, kidneys, and lungs, respectively, as shown in **Figure 5**. There were 9, 3, 1, 3, 3, and 16 TFs, respectively, in the gene regulatory network of each tissue. In total, 35 TFs belonged to 10 TF families, among which 10 TFs were also hub genes. According to the PWM provided by the CIS-BP database, 20 TFs target genes were predicted. We found that these 20 TFs regulate 237 genes (94 genes are hub genes) in each tissue key modules, predicting a total of 408 regulatory relationships. Through the prediction of miRNA target genes, we found that genes in the key modules of the muscle, liver, heart, spleen, kidneys, and lungs were regulated by 8, 3, 3, 2, 4, and 6 miRNAs, respectively. **Table 4** summarizes the information of TFs, miRNAs, target genes, and hub genes in the gene regulatory network of six tissues in Tibetan pigs.

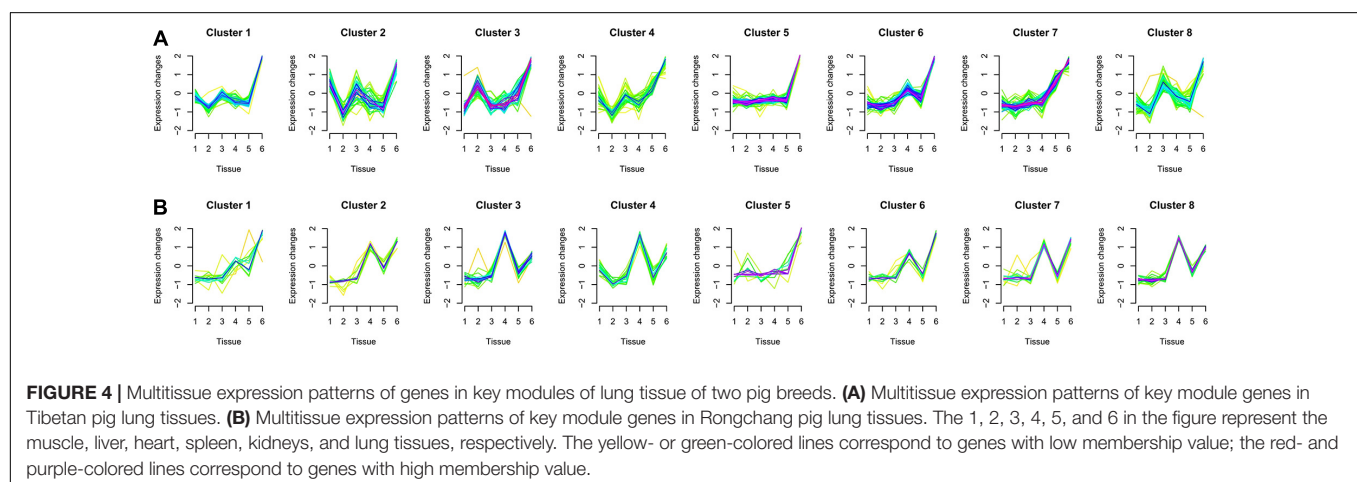


TABLE 2 | Pathways that are only significantly enriched in Tibetan pig tissue modules.

Tissue	Category	ID	Term	Benjamini
Muscle	Cellular components	GO:0031595	Nuclear proteasome complex	3.97E-03
		GO:0008540	Proteasome regulatory particle, base subcomplex	1.02E-02
	KEGG_Pathway	ssc03050	Proteasome	3.15E-04
		ssc01200	Carbon metabolism	6.64E-03
		ssc04152	AMPK signaling pathway	1.38E-02
		ssc04261	Adrenergic signaling in cardiomyocytes	2.18E-02
		ssc04931	Insulin resistance	3.64E-02
		ssc05169	Epstein-Barr virus infection	3.69E-02
		ssc04722	Neurotrophin signaling pathway	3.83E-02
		ssc04921	Oxytocin signaling pathway	4.11E-02
Heart	KEGG_Pathway	ssc05412	Arrhythmogenic right ventricular cardiomyopathy (ARVC)	1.85E-02
Spleen	Biological progresses	GO:0006412	Translation	6.86E-15
		GO:0001731	Formation of translation preinitiation complex	1.32E-02
		GO:0006446	Regulation of translational initiation	1.81E-02
	Cellular components	GO:0022627	Cytosolic small ribosomal subunit	6.49E-17
		GO:0022625	Cytosolic large ribosomal subunit	8.99E-08
		GO:0016282	Eukaryotic 43S preinitiation complex	1.29E-04
		GO:0033290	Eukaryotic 48S preinitiation complex	1.76E-04
		GO:0005852	Eukaryotic translation initiation factor 3 complex	2.17E-03
		GO:0005683	U7 snRNP	2.35E-02
		GO:0042105	Alpha-beta T-cell receptor complex	2.35E-02
	Molecular function	GO:0003735	Structural constituent of ribosome	1.59E-18
		GO:0003743	Translation initiation factor activity	9.59E-03
	KEGG_Pathway	ssc03010	Ribosome	4.10E-19
Lung	KEGG_Pathway	ssc04510	Focal adhesion	3.69E-05
		ssc04512	ECM-receptor interaction	2.98E-04
		ssc04151	PI3K-Akt signaling pathway	9.26E-03
		ssc04350	TGF-beta signaling pathway	1.80E-02

TABLE 3 | Hub gene information of key modules in Tibetan pigs and Rongchang pigs.

Pig breed	Tissue	Num of hubs	Num of overlapping hubs*	Num of hub (TSI > 0.9)	Num of TFs in hub
Tibetan pig	Muscle	23	22	11	1
	Liver	41	20	30	0
	Heart	20	2	8	1
	Spleen	40	6	20	1
	Kidney	81	45	26	1
	Lung	32	0	8	6
Rongchang	Muscle	61	22	25	7
	Liver	39	20	30	0
	Heart	26	2	3	0
	Spleen	123	6	0	13
	Kidney	68	45	26	0
	Lung	14	0	1	0

*"Num of overlapping hubs" is the number of overlapping hub genes detected by Tibetan pigs and Rongchang pigs in the same tissue module.

Identification of Gene Regulatory Network Motifs

In gene networks, some motifs displayed much higher frequencies than expected in randomized networks (Ravasz

et al., 2002; Shen-Orr et al., 2002), and these motifs were suggested to be recurring circuit elements that perform key information-processing tasks (Rosenfeld et al., 2002; Shen-Orr et al., 2002; Mangan and Alon, 2003). Among them, the motif composed of three nodes contains 13 kinds, including V-out, 3-Chain, FFL, 3-Loop, Clique, and so on. Using the mfinder1.2 software, we identified 8,894, 13,067, 993, 19,899, 78,959, and 14,692 motifs in the gene regulatory networks of the muscle, liver, heart, spleen, kidneys, and lung tissues in Tibetan pigs, respectively. There were significant differences in the distribution of motifs among the different gene regulatory networks ($p < 2.2E-16$). The motif information in the gene regulatory networks of the six tissues is shown in **Table 5**.

To analyze the statistical significance of each motif type, we generated 10,000 random networks representing a conservation rule. The distribution of TSP in the gene regulatory network of the lung tissues is shown in **Figure 6**. We found that the frequency of FFL, Regulated mutual, Regulating mutual, and Clique motifs in the lung tissue gene regulatory network was significantly different from that of random networks ($p < 1E-04$). In the muscle and heart tissue gene regulatory network, Regulated mutual and Clique motif were significant motif types, while V-out, Semi clique, and Clique motif were significant in the gene regulatory network of the kidneys.

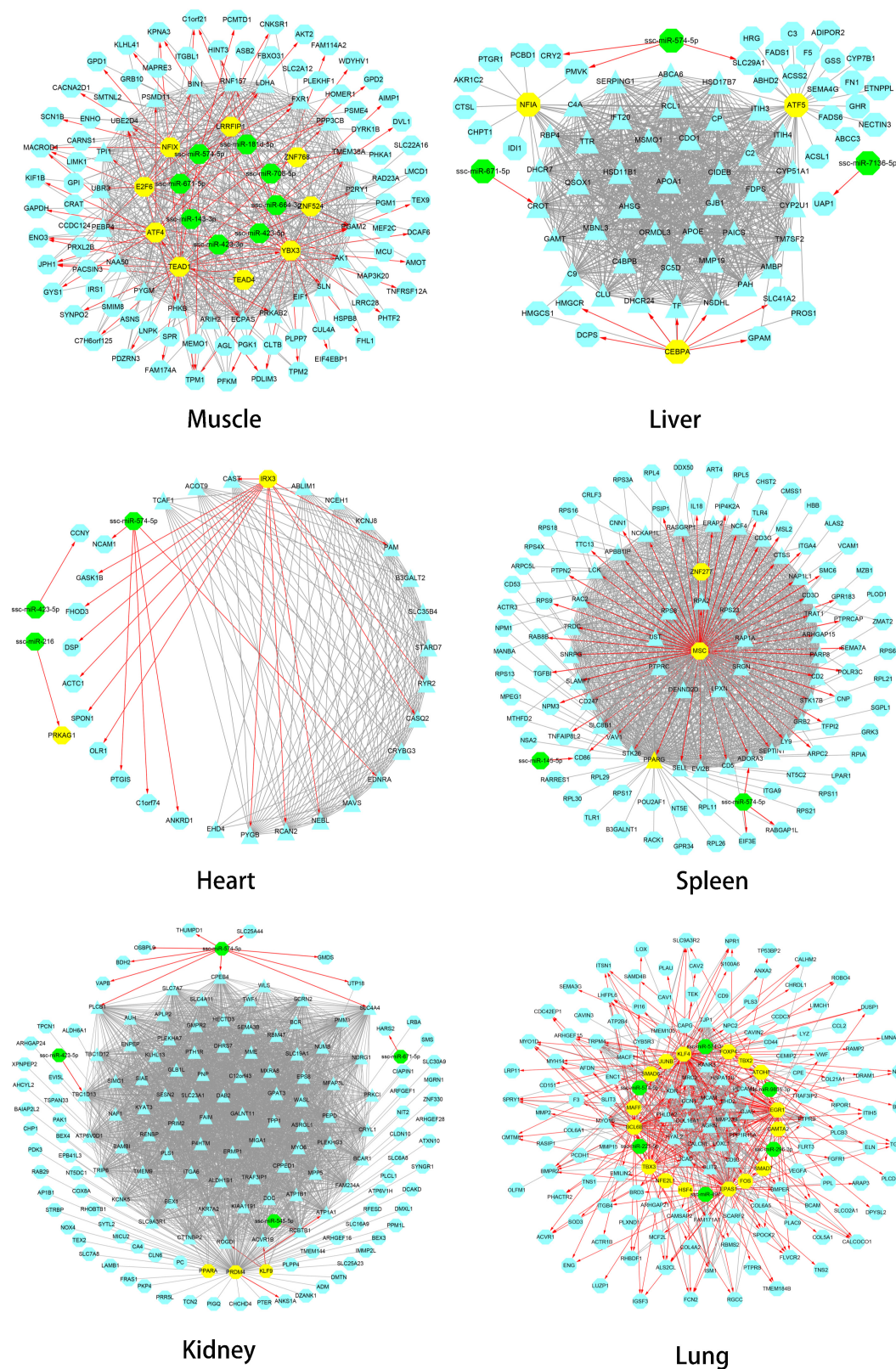


FIGURE 5 | Gene regulatory network of six tissues of Tibetan pigs. In each network in the figure, the yellow dots represent TFs, the green dots represent miRNAs, and the hub genes are represented by triangles. The red edges with arrows represent the regulatory relationship between TFs and miRNAs and target genes. The gray edge indicates that there is only a co-expression relationship between the two genes.

TABLE 4 | Detailed information of gene regulatory networks in six tissues of Tibetan pigs.

Tissue	Num of TFs	Num of TF target genes	Num of hub regulated by TFs	Num of miRNAs	Num of miRNAs target genes	Num of hub regulated by miRNAs
Muscle	9	49	17	8	12	1
Liver	3	7	3	3	5	1
Heart	1	13	7	3	8	1
Spleen	3	55	33	2	4	1
Kidney	3	4	3	4	13	5
Lung	16	118	31	6	12	3

Motif Analysis of the Gene Regulatory Network in the Lung Tissues

We further analyzed FFL, Regulated mutual, and Regulated mutual motifs in the gene regulatory network of lung tissues. All FFL motifs in the gene regulatory network of lung tissues were $TF_1 \rightarrow TF_2$, including $KLF4 \rightarrow EPAS1$, $KLF4 \rightarrow BCL6B$, $KLF4 \rightarrow FOS$, $EGR1 \rightarrow BCL6B$, $EGR1 \rightarrow EPAS1$, $BCL6B \rightarrow EPAS1$, $TBX3 \rightarrow EPAS1$, and $TBX3 \rightarrow BCL6B$. Then, the two TFs shared a target gene. As a result, 51 target genes were regulated, including four TFs, forming 13 FFLs, and 21 hub genes, forming 71 FFLs. In addition, three of these target genes were both TF and hub gene, forming eight FFLs.

There were two main types of Regulating mutual motifs. One includes two TFs regulating each other, including $EGR1-KLF4$, $EGR1-TBX3$, and $KLF4-TBX3$, and jointly regulating the same target gene. A total of 47 target genes were regulated, including six TFs, forming 12 complexes, and 27 hub genes, forming 49 complexes. Among these target genes, four target genes were both TF and hub genes, forming 10 complexes. The other type of Regulating mutual motifs includes two TFs that were co-expressed and shared a target gene. We found that *FOS* and *JUNB* co-expressed and co-regulated the *DUSP1* gene.

In the Regulated mutual motif, one TF regulated two genes, and there was a co-expression relationship between the two target genes. It was composed of TFs, including *EGR1*, *KLF4*, *EPAS1*, *BCL6B*, and *TBX3*, and their regulated target genes, forming a total of 810 complexes. Of these complexes, there were eight in which both target genes are TFs and 593 in which both are hub genes. In the Clique motif, only the $EGR1-KLF4-TBX3$ motif was the mutual regulation of these three TFs, and the remaining motifs were co-expressed relationships among genes.

Identification of Important Genes and Regulatory Relationships Related to Hypoxia in the Lung Gene Regulatory Network

Formulas (8) and (10) were used to evaluate the importance of each gene and the three-node motif, including FFL, Regulating mutual, and Regulated mutual type motif, in the gene regulatory network of lung tissues. We found that the several important top genes were *KLF4*, *BCL6B*, *EGR1*, *SMAD6*, and *EPAS1* transcription factor genes, which were also hub genes. The top 25% of the node importance scores in the Tibetan pig lung gene regulatory network are shown in **Table 6**.














The Regulating mutual motif formed by $KLF4-EGR1-BCL6B$ was the most important motif based on the motif score. We call it the “ $KLF4-EGR1-BCL6B$ ” complex. In this motif, the *KLF4* and *EGR1* genes regulate the same target gene, *BCL6B*. This complex preferred to synergistically regulate the *EPAS1*, *KDR*, *SMAD6*, *SMAD7*, *CCN1*, and *ATOH8* genes (**Figure 7**), which comprised 18 motifs (**Table 7**). The “ $KLF4-EGR1-BCL6B$ ” complex may coordinately regulate the *SMAD6* and *SMAD7* genes, which play an important role in the TGF- β signaling pathway. *EPAS1* is an important hypoxia-inducible factor. This complex may also indirectly regulate *SMAD6* and *SMAD7* genes by regulating the *EPAS1* gene. This complex may also regulate the *KDR* gene, which is involved in the PI3K-Akt signaling pathway. Both TGF- β and PI3K-Akt signaling pathways play an important role in hypoxia response (Chen et al., 2006; Ambalavanan et al., 2008; Jia et al., 2016; Qi et al., 2019).

Validation of Important Genes in Lung Tissue

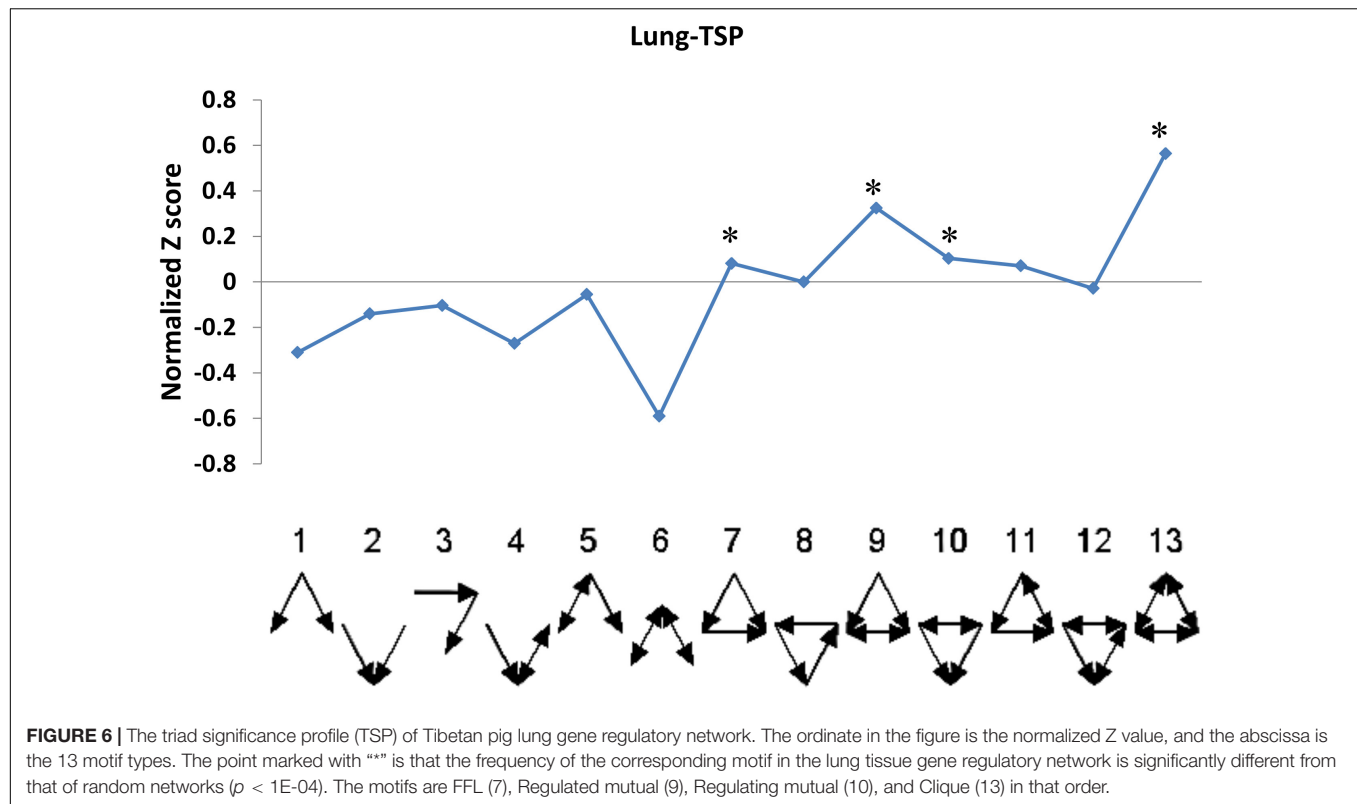
To confirm the relationship between *KLF4*, *EGR1*, *BCL6B*, *SMAD6*, *EPAS1*, *KDR*, *SMAD7*, *CCN1*, *ATOH8*, and *MMP23B* genes, we used lung tissue transcriptomic data from the Tibetan sheep, yak, and Diqing Tibetan pig population for validation via the co-expression network analysis. We identified the key module of the lung tissues for the three validation groups using WGCNA. The network topology of these key modules, including density, mean cluster coefficient, centralization, and heterogeneity, were similar to those of Songpan Tibetan pigs. There were 151, 150, and 73 overlap genes between gene sets of the lung key module for the three validation groups and Tibetan pigs, respectively.

Due to using commercially available Agilent Whole Porcine Genome Oligo (4 × 44 K) Microarrays for Diqing Tibetan pig, there was no probe annotation information for the *BCL6B*, *CCN1*, *ATOH8*, and *MMP23B* genes. In Diqing Tibetan pig lung tissues, six genes, including *KLF4*, *EGR1*, *EPAS1*, *SMAD6*, *SMAD7*, and *KDR*, were highly expressed and were significantly positively correlated between genes, except for between *KDR* gene and others (as shown in the **Supplementary Table 5**). Overall, there were 11 and nine edges (the co-expression relationship) among the above genes found in the Songpan Tibetan pig and Diqing Tibetan pig (**Figure 8A**), respectively. So, 81.82% (9/11) of the co-expression relationships in the above genes were confirmed. Based on the WGCNA for the Diqing Tibetan pig lung tissue, we found the *KLF4*, *EGR1*, and *SMAD7* genes

TABLE 5 | Motif information in regulatory networks of six tissues in Tibetan pigs.

Motif	Name	Muscle			Liver			Heart			Spleen			Kidney			Lung		
		Counts	Z*	P*	Counts	Z	P	Counts	Z	P	Counts	Z	P	Counts	Z	P	Counts	Z	P
	V-out	378	-5.29	1.00	21	-0.79	0.92	67	-2.18	0.99	0	–	1.00	43	1.58	0.05	5,160	-8.94	1.00
	V-in	24	0.46	0.38	0	–	1.00	0	–	1.00	0	–	1.00	0	–	1.00	18	-3.69	1.00
	3-Chain	21	-0.05	0.57	0	–	1.00	0	–	1.00	0	–	1.00	209	-1.97	0.96	133	-2.76	1.00
	Mutual in	408	-2.12	0.98	145	-0.08	0.61	84	-2.18	0.99	0	–	1.00	0	–	1.00	380	-7.70	1.00
	Mutual out	902	1.64	0.05	26	0.41	0.39	0	–	1.00	0	–	1.00	0	–	1.00	3,152	-0.89	0.82
	Mutual V	4,363	-12.93	1.00	2,094	-45.36	1.00	232	-3.14	0.99	8,623	-23.86	1.00	0	–	1.00	3,098	-15.01	1.00
	FFL	3	-0.14	0.59	0	–	1.00	0	–	1.00	0	–	1.00	469	-1.50	0.81	135	2.51	<1E-4
	3-Loop	0	–	1.00	0	–	1.00	0	–	1.00	0	–	1.00	29,452	-28.23	1.00	0	-0.01	1.00
	Regulated mutual	140	5.47	<1E-4	3	0.79	0.32	21	2.18	1.00	0	–	1.00	0	–	1.00	810	9.34	<1E-4
	Regulating mutual	9	-0.36	0.71	0	–	1.00	0	–	1.00	0	–	1.00	0	–	1.00	88	2.67	<1E-4
	Mutual and 3-Chain	4	0.16	0.40	0	–	1.00	0	–	1.00	0	–	1.00	4	-1.58	0.98	28	1.88	0.03
	Semi clique	309	-1.68	0.95	9	-0.41	0.76	0	–	1.00	0	–	1.00	65	1.97	0.04	382	-2.11	0.98
	Clique	2,333	16.14	<1E-4	10,769	45.80	<1E-4	589	3.14	0.01	11,276	23.86	<1E-4	48,717	28.52	<1E-4	1,308	15.15	<1E-4

*"Z" and "P" in the table are the Z score and p-value calculated by the mfinder1.2 software for each motif.



were clustered into one module, while *EPAS1* and *SMAD6* were clustered into the other module.

The *KLF4*, *BCL6B*, *EPAS1*, *EGR1*, *SMAD6*, *KDR*, *CCN1*, and *ATOH8* genes had the highest expression in lung tissues compared with the other five tissues of Tibetan sheep (muscle, liver, heart, spleen, and kidneys). There were 14 and 24 significantly co-expression relationships among the above genes, which were identified in Tibetan sheep (see **Figure 8B** and the **Supplementary Table 5**) and Songpan Tibetan pig, respectively. In total, 58.33% (14/24) of the relationships between genes were validated. Except for *EGR1* and *ATOH8*, the other genes were all clustered into the same key module related to the lung using WGCNA.

With the exception of *ATOH8* and *MMP23B*, the other genes were most highly expressed in the lung tissues of yak compared with the other five tissues. We detected 12 and 24 significantly positive correlations in the above genes of Tibetan yak lung tissue (see **Figure 8C** and the **Supplementary Table 5**) and Songpan Tibetan pig. We successfully verified 50% (12/24) of the relationships among genes. After performing WGCNA, we found that *KLF4*, *BCL6B*, *EPAS1*, *SMAD6*, and *SMAD7* were clustered into key modules related to the yak lungs.

DISCUSSION

Many previous studies primarily focused on identifying differentially expressed genes through gene expression profile analysis, but interactions between genes in different cell states

may not have been fully considered (Kostka and Spang, 2004). Moreover, differences in gene expression are not equal to differences in gene action. Compared with expression level analysis, network-based analysis not only captures local patterns but also identifies global patterns in a biological context, revealing molecular regulation details of hub genes at the network level. At the same time, the hub genes related to biological processes identified through gene network analysis also provide clues for subsequent molecular studies.

In this study, we detected the gene regulatory network related to Tibetan pig lung tissues. An appropriate sample size is critical to the planning and interpretation of genetic studies, whether they are descriptive or analytical. The small sample sizes will result in imprecise estimates in a descriptive study and failure to achieve statistical significance in an analytic or comparative study (Weber and Hoo, 2018). Due to the sample size limitation of this study, this may have limited the generalizability of the results of the research, and further independent tests may be required to verify our findings. However, our research methods and results can provide some valuable clues for the study of the hypoxia adaptation mechanism of Tibetan pigs. Combining topological characteristics, differential expression, and tissue-specific expression, we identified a list of genes that may be related to hypoxia in Tibetan pig lung tissue, such as *EPAS1*, *LOXL1*, *KLF4*, *EGR1*, *BCL6B*, *SMAD6*, *SMAD7*, *KDR*, *MMP23B*, and miR-296.

The *EPAS1* gene found in this study may be related to the adaptation to the hypoxic environment. The *EPAS1* gene encodes one subunit of hypoxia-inducible factor (HIF),

TABLE 6 | The top 25% of S_{node} genes in the Tibetan pig lung gene regulatory network.

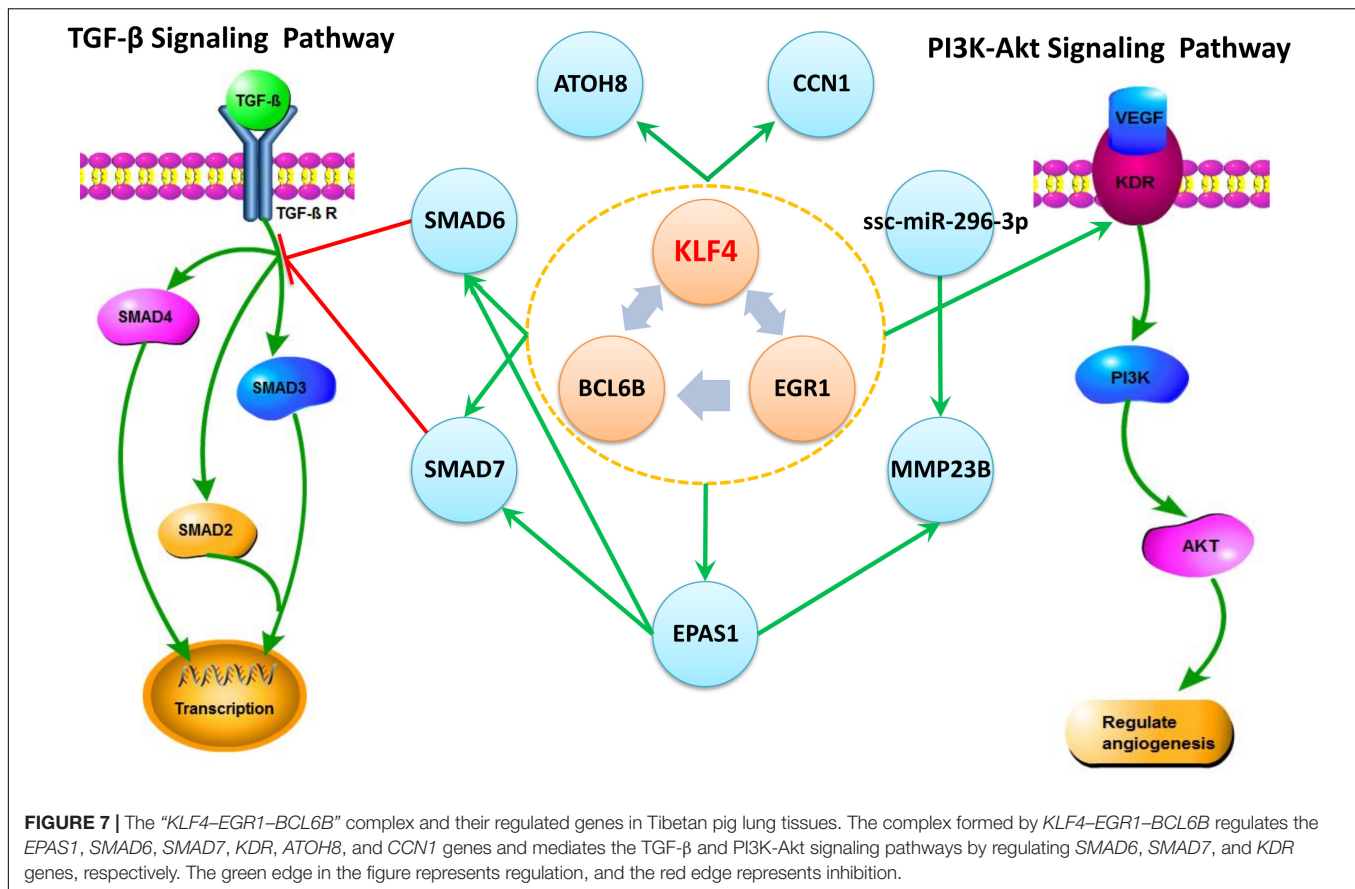
Ranking	Gene	Full name	Type (Hub/TF)*	S_{node}
1	<i>KLF4</i>	Kruppel-like factor 4	Hub & TF	0.4485
2	<i>BCL6B</i>	BCL6B transcription repressor	Hub & TF	0.3506
3	<i>EGR1</i>	Early growth response 1	Hub & TF	0.3184
4	<i>TBX3</i>	T-box 3	TF	0.1885
5	<i>EPAS1</i>	Endothelial PAS domain protein 1	Hub & TF	0.1789
6	<i>SMAD6</i>	SMAD family member 6	Hub & TF	0.1516
7	<i>MAFF</i>	MAF bZIP transcription factor F	TF	0.1189
8	<i>CCN1</i>	Cellular communication network factor 1	Hub	0.0814
9	<i>GJA5</i>	Gap junction protein, alpha 5	Hub	0.0784
10	<i>JUNB</i>	JunB proto-oncogene, AP-1 transcription factor subunit	TF	0.0744
11	<i>CALCRL</i>	Calcitonin receptor like receptor	Hub	0.0712
12	<i>FOS</i>	Fos proto-oncogene, AP-1 transcription factor subunit	TF	0.0707
13	<i>JCAD</i>	Junctional cadherin 5 associated	Hub	0.0705
14	<i>MRC2</i>	Mannose receptor C type 2	Hub	0.0683
15	<i>PECAM1</i>	Platelet and endothelial cell adhesion molecule 1	Hub	0.0646
16	<i>TJP1</i>	Tight junction protein 1	Hub	0.0645
17	<i>CD93</i>	CD93 molecule	Hub	0.0636
18	<i>LHFPL6</i>	LHFPL tetraspan subfamily member 6	TG	0.0611
19	<i>COL16A1</i>	Collagen, type XVI, alpha 1	Hub	0.0588
20	<i>PTPRB</i>	Protein tyrosine phosphatase receptor type B	Hub	0.0570
21	<i>SMAD7</i>	SMAD family member 7	Hub	0.0556
22	<i>MCAM</i>	Melanoma cell adhesion molecule	Hub	0.0540
23	<i>HYAL2</i>	Hyaluronidase 2	Hub	0.0525
24	<i>SLIT2</i>	Slit guidance ligand 2	Hub	0.0515
25	<i>HSPA12B</i>	Heat shock protein family A (Hsp70) member 12B	Hub	0.0495
26	<i>PPP1R15A</i>	Protein phosphatase 1, regulatory subunit 15A	Hub	0.0464
27	<i>EHD2</i>	EH domain containing 2	Hub	0.0461
28	<i>PHLDA2</i>	Pleckstrin homology like domain family A member 2	Hub	0.0451
29	<i>KANK3</i>	KN motif and ankyrin repeat domains 3	Hub	0.0442
30	<i>MMP23B</i>	Matrix metalloproteinase 23B	Hub	0.0394
31	<i>LOXL1</i>	Lysyl oxidase like 1	Hub	0.0333
32	<i>TBX2</i>	T-box 2	TF	0.0251
33	<i>FAM171A1</i>	Family with sequence similarity 171 member A1	Hub	0.0250
34	<i>KDR</i>	Kinase insert domain receptor	Hub	0.0212
35	<i>MYO1C</i>	Myosin IC	Hub	0.0211
36	<i>ATOH8</i>	Atonal bHLH transcription factor 8	TF	0.0193
37	<i>AGRN</i>	Agrin	Hub	0.0178
38	<i>NPC2</i>	NPC intracellular cholesterol transporter 2	TG	0.0169
39	<i>PLAC9</i>	Placenta associated 9	TG	0.0163
40	<i>SLC9A3R2</i>	SLC9A3 regulator 2	TG	0.0157

*There are four types of genes, including "Hub&TF," "Hub," "TF," and "TG." Among them, type "Hub & TF" represents that the gene is both TF and hub genes. Type "Hub" represents that the gene is a hub gene, and type "TF" represents that the gene is TF. Type "TG" means that the gene is neither hub nor TF, but only a target gene of TF.

which show multifarious effects involved in complex oxygen sensing (Henderson et al., 2005) and regulation of angiogenesis, hemoglobin concentration, and erythrocytosis (Beall et al., 2010). In the Tibetan human population, the *EPAS1* gene is involved in the chronic hypoxia response, and it has been shown to have a strong signature of selection (Bigham et al., 2010; Simonson et al., 2010; Peng et al., 2011; Xu et al., 2011). Moreover, Li et al. (2019) show that the mutant genotype frequencies of the rs13419896, rs1868092, and rs4953354 loci in the *EPAS1* gene are significantly higher in the Tibetan population than in the plains population. Under plateau hypoxic conditions, the plains population was able

to acclimate rapidly to hypoxia through increasing *EPAS1* mRNA expression and changing the hemoglobin conformation. The *EPAS1* gene also has obvious selection signature in other plateau animals, such as Tibetan horses (Liu et al., 2019) and Tibetan pigs (Ma et al., 2019) and has been identified as a key evolutionary molecule adapted to the plateau hypoxic environment.

Angiogenesis was an adaptive response to tissue hypoxia (Fong, 2008). A majority of the identified hub genes participated in the angiogenesis process, such as *LOXL1*, *KLF4*, and *EGR1*. The *LOXL1* gene is essential for the stability and strength of elastic vessels and tissues (Li et al., 2020) and may play



important roles in the enhanced angiogenesis promoted by hypoxia (Xie et al., 2017). The *KLF4* gene tended to be pleiotropic. Not only does it promote pulmonary angiogenesis and blood transport (Ghaleb and Yang, 2017) and accelerate the acquisition and transport of oxygen but also it protects the lungs from hypoxia (Shatat et al., 2014). The *EGR1* gene stimulates and induces the process of angiogenesis (Adamson and Mercola, 2002; Sheng et al., 2018). Angiogenesis is conducive to the increase in oxygen transport. Therefore, we infer that these genes might contribute to obtaining and transporting oxygen better in hypoxic environments, by involving in the angiogenesis process.

Based on the Tibetan pig lung tissue-specific gene network, we found that *KLF4* and *EGR1* simultaneously regulated the *BCL6B* gene, forming the *KLF4-EGR1-BCL6B* complex, which might be dominated by the *KLF4* gene and affect the expression of *EPAS1*, *SMAD6*, *SMAD7*, *CCN1*, *KDR*, and *ATOH8*. These key genes and regulatory relationships were validated in the lung tissues of Tibetan pigs from Jia et al. (2016) and Tibetan sheep and yak from Tang et al. (2017). After a large literature review and verification of gene function annotation, we postulate that the *KLF4-EGR1-BCL6B* complex might be beneficial for Tibetan pigs to survive better in hypoxic environments.

The *KLF4*, *EGR1*, and *BCL6B* genes jointly regulate the *SMAD6* and *SMAD7* genes, which are important regulators

of the TGF-β signaling pathway. In the TGF-β signaling pathway, the SMAD family genes are very important signal transduction and regulatory molecules. *SMAD6* and *SMAD7* are antagonists of the TGF-β gene family. High expression of *SMAD7* inhibited the transcription of *SMAD2* and *SMAD3* gene induced by the *TGF-β* gene and antagonizes tissue fibrosis (Yan et al., 2009). Therefore, the *KLF4-EGR1-BCL6B* complex in Tibetan pig lungs may mediate the TGF-β signaling pathway by regulating the expression of *SMAD6* and *SMAD7*, thereby enhancing the antifibrotic effect of the lungs.

Moreover, the *KLF4-EGR1-BCL6B* complex might regulate the *KDR* gene, which was primarily expressed in pulmonary vascular endothelial cells and has important proangiogenic activity (Melincovici et al., 2018). The *KDR* gene is an important regulator of the PI3K-Akt signaling pathway. Jia et al. (2016) and Qi et al. (2019) found that the PI3K-Akt signaling pathway was involved in hypoxia adaptation in both Tibetan pigs and yaks. Under hypoxic conditions, the combination of *KDR* and *VEGF* activates the downstream *PI3K* gene, thereby regulating proliferation and differentiation of neovascular endothelial cells and playing an important role in the development of angiogenesis (Graupera and Potente, 2013). Therefore, the *KLF4-EGR1-BCL6B* complex may act on the PI3K-Akt pathway by mediating the *KDR* gene and accelerating the acquisition and transportation of oxygen under hypoxic conditions.

TABLE 7 | The motifs formed between the “*KLF4–EGR1–BCL6B*” complex and its regulatory genes in the lung.

Motif type	Name	Motif	S_{motif}	Ranking
	Regulating mutual	<i>KLF4–EGR1–BCL6B</i>	1.1695	1
		<i>KLF4–EGR1–EPAS1</i>	1.0766	2
		<i>KLF4–EGR1–CCN1</i>	1.0129	13
		<i>KLF4–EGR1–SMAD6</i>	1.0073	16
		<i>KLF4–EGR1–MMP23B</i>	0.9544	43
		<i>KLF4–EGR1–ATOH8</i>	0.9293	65
		<i>KLF4–EGR1–SMAD7</i>	0.9031	97
		<i>KLF4–EGR1–KDR</i>	0.9025	98
	Regulated mutual	<i>KLF4–SMAD6–BCL6B</i>	1.0108	15
	FFL	<i>KLF4–BCL6B–EPAS1</i>	1.0717	3
		<i>KLF4–BCL6B–CCN1</i>	1.0212	9
		<i>EGR1–BCL6B–EPAS1</i>	1.0006	19
		<i>KLF4–BCL6B–ATOH8</i>	0.9463	50
		<i>KLF4–EPAS1–SMAD6</i>	0.9423	55
		<i>KLF4–BCL6B–KDR</i>	0.9199	77
		<i>EGR1–BCL6B–CCN1</i>	0.9125	87
		<i>KLF4–EPAS1–MMP23B</i>	0.8839	117
		<i>EGR1–EPAS1–SMAD6</i>	0.8806	121

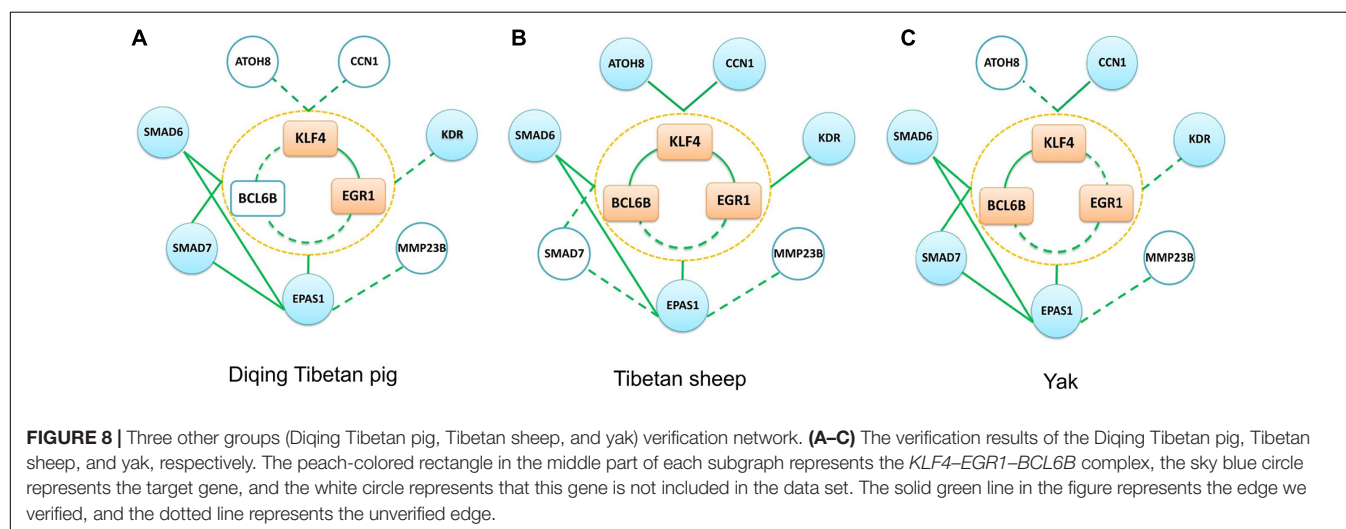
In addition, the *KLF4–EGR1–BCL6B* complex also regulated the *ATOH8*, *CCN1*, and *EPAS1* genes. High expression of *CCN1* suppressed pulmonary vascular smooth muscle contraction in response to hypoxia (Lee et al., 2015). The *ATOH8* gene participates in the *ALK-1/SMAD/ATOH8* axis, which attenuated the hypoxic response in endothelial cells in the pulmonary circulation and might help prevent the development of pulmonary arterial hypertension (Morikawa et al., 2019). The *MMP23B* gene was a member of the MMP gene family, and MMP matrix metalloproteinases played an important role in tissue remodeling and angiogenesis (Białkowska et al., 2020). Moreover, *MMP23B* is regulated by *EPAS1* and ssc-miR-296-3p. Studies

had shown that miR-296 can regulate angiogenesis (Anand and Cheresch, 2011; Li et al., 2018).

The co-expression and network analysis were performed in three validation groups (Tibetan sheep, yak, and Diqing Tibetan pig). Comparing pigs, sheep, and cow living in normal oxygen content environments, some genes, such as *KLF4*, *EGR1*, *EPAS1*, *SMAD6*, and *KDR* genes, are overlapped in the key module of the Tibetan pig, Tibetan sheep, and yak lung tissues. As stated above, these genes improve the tolerance of Tibetan pigs to hypoxic environment through involving in angiogenesis and antagonizing lung tissue fibrosis.

Due to using porcine oligo microarrays for the Diqing Tibetan pig, some genes, such as *BCL6B*, do not have probe annotation information. We did not observe the *KLF4–EGR1–BCL6B* complex in the Diqing Tibetan pig lungs, but the *KLF4* and *EGR1* genes might jointly correlate with *SMAD6*, *SMAD7*, and *EPAS1* genes. In the Tibetan sheep lung, the *BCL6B* gene did not significantly correlate with the *EGR1* gene (p value = 0.0839), due to the limitation of sample size. So, further experiments in a large validation population, such as ChIP-seq, would help the demonstration of the regulation function of the complex *KLF4–EGR1–BCL6B*.

Although we identified the *KLF4* gene as a key gene in the lung tissue of different species, and related to the *SMAD6* and *EPAS1* genes, there are some different co-expression relationships in the gene regulatory network of Tibetan pig, Tibetan sheep, and yak. We deem that the following reasons could have contributed to the observed differences. First, the study populations of different species have various genetic backgrounds. Many previous studies have also shown that there are different anatomical structures of tissues, physiological and biochemical indexes, and molecular mechanisms in the environment adaptation of plateau animals. Second, there are differences in the sampling methods and genetic drift events among studies on the same species. Third, gene regulatory programs display a wide range of characteristics, depending on where they are in the body and what stage in its life cycle. To control a cell's behavior in different space and time,



different gene expression profiles and regulation relationship will be observed.

In addition to the lung tissues, the heart also plays an important role in high-altitude hypoxia adaptation. Qi et al. (2019) showed that the heart and lung tissues were identified as the two key organs of yak hypoxia adaptation. In this study, we identified some specific expression genes related to hypoxia in the heart tissue of Songpan Tibetan pigs, such as *EGLN3*, *RYR2*, *EDNRA*, and *EGLN3* (egl-9 family hypoxia inducible factor 3), that likely plays an important role in cellular adaptation to hypoxic conditions by participating in the HIL-1 signaling pathway. Zhang B. et al. (2017) used the transcriptomic and proteomic data of Tibetan pig heart tissues to identify the *EGLN3* gene as important candidate genes for high-altitude adaptation. Moreover, the *EGLN1* gene, a member of the same family of *EGLN3*, has been determined to be involved in the hypoxia adaptation of Tibetan humans (Bigham et al., 2010; Simonson et al., 2010; Xiang et al., 2013).

RYR2 and *EDNRA* may play crucial roles in heart development, heart rhythm stabilization, and signal transduction to cope with hypoxia. Zhang et al. (2014) found that the *RYR2* gene was related to the hypoxia adaptability of Tibetan gray wolves. Low oxygen environment can increase *EDNRA* gene expression (Minchenko et al., 2019; Zhang Y. et al., 2017). However, Zhang B. et al. (2017) did not identify the *RYR2* and *EDNRA* genes as hypoxia-related candidate genes by comparative mRNA and protein expression profiles in heart tissues of Tibetan and Yorkshire pigs, respectively. These inconsistencies may be a result of differences in the study population in the genetic backgrounds, population structure, developmental stages, and environmental factors. This also suggests that many key genes, in reducing hypoxia injury that exists within the Tibetan pig genome, have yet to be discovered.

CONCLUSION

In summary, through gene network analysis, we found that lung tissues may play an important role in hypoxia adaptation in Tibetan pigs. We comprehensively profiled the gene regulatory network of Tibetan pig lung tissues, identifying a series of candidate genes related to hypoxia and discovering that *KLF4* is likely the core regulator of the *KLF4-EGR1-BCL6B* complex, which may mediate the TGF- β signaling pathway and improve the anti-hypoxic ability of Tibetan pigs. Although gene function is not entirely dependent on gene expression regulation, these findings may provide valuable clues and better understanding in exploring the underlying molecular mechanisms of hypoxia adaptation.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: The data underlying this article are

available in Gene Expression Omnibus (GEO) database at <https://www.ncbi.nlm.nih.gov/geo/>, and can be accessed with GSE93855, GSE124418, and GSE84409.

AUTHOR CONTRIBUTIONS

ZW, XW, and TW conceived the project. TW, YG, SL, and CZ performed the bioinformatics and data analysis. TW and ZW wrote the manuscript. TC, KD, and PW collected the samples and data. All authors read and approved the final manuscript.

FUNDING

This work was financially supported by the Natural Science Foundation of China (No. 32070571), the Academic Backbone Project of Northeast Agricultural University (No. 15XG14), and the NEAU Research Founding for Excellent Young Teachers (2010RCB29).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fgene.2021.628192/full#supplementary-material>

Supplementary Figure 1 | Weighted gene co-expression network analysis of Rongchang pigs. **(A)** Analysis of network topology of Rongchang pig showed that it met the scale-free topology threshold of 0.8 when $\beta = 20$. The left panel shows the scale-free fit index as a function of the soft-threshold power. The right panel displays the mean connectivity as a function of the soft-threshold power. **(B)** The dissimilarity was based on topological overlap. The "Merged dynamic" is the result of merging modules with a correlation higher than 0.9. The y-axis is the distance determined by the extent of topological overlap. **(C)** Heatmap displaying the correlations and significant differences between gene modules and six tissues of Rongchang pigs. Red represents high adjacency (positive correlation) and blue represents low adjacency (negative correlation). In brackets is the p -value of the correlation test.

Supplementary Figure 2 | The co-expression network of six tissues key modules of Tibetan pig. The co-expression network of muscle, liver, heart, spleen, kidney and lung in the figure shows the co-expression relationship of weight above 0.35, 0.35, 0.25, 0.35, 0.35, and 0.25, respectively. The dark blue circles in the figure represent the hub genes of each network.

Supplementary Table 1 | List of genes in the key modules of the six tissues of Tibetan pigs and Rongchang pigs.

Supplementary Table 2 | List of tissue-specific genes for Tibetan pigs and Rongchang pigs. The first two columns in the table are the gene name and Ensembl ID. Columns 3–5 are the position coordinates of the gene. Columns 6–7 are information about the tissue-specific genes of Tibetan pigs, and the seventh column represents the TSI value of the Tibetan pig TS genes. Columns 8–9 are the information of Rongchang pig TS genes, and the ninth column represents the TSI value of the Rongchang pig TS genes.

Supplementary Table 3 | Significantly enriched signaling pathways in key modules of various tissues in Tibetan pigs and Rongchang pigs.

Supplementary Table 4 | Pathways that are only significantly enriched in Rongchang pig tissue modules.

Supplementary Table 5 | Correlation coefficients and p -values between candidate genes in the three validation groups (Diqing Tibetan pig, Tibetan sheep and yak).

REFERENCES

- Adamcsek, B., Palla, G., Farkas, I. J., Derényi, I., and Vicsek, T. (2006). CFinder: locating cliques and overlapping modules in biological networks. *Bioinformatics* 22, 1021–1023. doi: 10.1093/bioinformatics/btl039
- Adamson, E. D., and Mercola, D. (2002). Egr1 transcription factor: multiple roles in prostate tumor cell growth and survival. *Tumour Biol.* 23, 93–102. doi: 10.1159/000059711
- Ai, H., Yang, B., Li, J., Xie, X., Chen, H., and Ren, J. (2014). Population history and genomic signatures for high-altitude adaptation in Tibetan pigs. *BMC Genomics* 15:834. doi: 10.1186/1471-2164-15-834
- Alon, U. (2007). Network motifs: theory and experimental approaches. *Nat. Rev. Genet.* 8, 450–461. doi: 10.1038/nrg2102
- Ambalavanan, N., Nicola, T., Hagood, J., Bulger, A., Serra, R., Murphy-Ullrich, J., et al. (2008). Transforming growth factor-beta signaling mediates hypoxia-induced pulmonary arterial remodeling and inhibition of alveolar development in newborn mouse lung. *Am. J. Physiol. Lung. Cell. Mol. Physiol.* 295, L86–L95. doi: 10.1152/ajplung.00534.2007
- Anand, S., and Cheresch, D. A. (2011). MicroRNA-mediated regulation of the angiogenic switch. *Curr. Opin. Hematol.* 18, 171–176. doi: 10.1097/MOH.0b013e328345a180
- Beall, C. M., Cavalleri, G. L., Deng, L., Elston, R. C., Gao, Y., Knight, J., et al. (2010). Natural selection on EPAS1 (HIF2alpha) associated with low hemoglobin concentration in Tibetan highlanders. *Proc. Natl. Acad. Sci. U.S.A.* 107, 11459–11464. doi: 10.1073/pnas.1002443107
- Bialkowska, K., Marciniak, W., Muszyńska, M., Baszuk, P., Gupta, S., Jaworska-Bieniek, K., et al. (2020). Polymorphisms in MMP-1, MMP-2, MMP-7, MMP-13 and MT2A do not contribute to breast, lung and colon cancer risk in polish population. *Hered. Cancer Clin. Pract.* 18:16. doi: 10.1186/s13053-020-00147-w
- Bigham, A., Bauchet, M., Pinto, D., Mao, X., Akey, J. M., Mei, R., et al. (2010). Identifying signatures of natural selection in Tibetan and Andean populations using dense genome scan data. *PLoS Genet.* 6:e1001116. doi: 10.1371/journal.pgen.1001116
- Buckingham, M., and Rigby, P. W. (2014). Gene regulatory networks and transcriptional mechanisms that control myogenesis. *Dev. Cell* 28, 225–238. doi: 10.1016/j.devcel.2013.12.020
- Chen, J., Alvarez, M. J., Talos, F., Dhruv, H., Rieckhof, G. E., and Iyer, A. (2016). Identification of causal genetic drivers of human disease through systems-level analysis of regulatory networks. *Cell* 166:1055. doi: 10.1016/j.cell.2016.07.036
- Chen, Y., Feng, J., Li, P., Xing, D., Zhang, Y., Serra, R., et al. (2006). Dominant negative mutation of the TGF-beta receptor blocks hypoxia-induced pulmonary vascular remodeling. *J. Appl. Physiol.* (1985) 100, 564–571. doi: 10.1152/japplphysiol.00595.2005
- Cheng, C., and Gerstein, M. (2012). Modeling the relative relationship of transcription factor binding and histone modifications to gene expression levels in mouse embryonic stem cells. *Nucleic Acids Res.* 40, 553–568. doi: 10.1093/nar/gkr752
- Döhr, S., Klingenhoff, A., Maier, H., Hrabé de Angelis, M., Werner, T., and Schneider, R. (2005). Linking disease-associated genes to regulatory networks via promoter organization. *Nucleic Acids Res.* 33, 864–872. doi: 10.1093/nar/gki230
- Enright, A. J., John, B., Gaul, U., Tuschl, T., Sander, C., and Marks, D. S. (2003). MicroRNA targets in *Drosophila*. *Genome Biol.* 5:R1. doi: 10.1186/gb-2003-5-1-r1
- Fong, G. H. (2008). Mechanisms of adaptive angiogenesis to tissue hypoxia. *Angiogenesis* 11, 121–140. doi: 10.1007/s10456-008-9107-3
- Ghaleb, A. M., and Yang, V. (2017). Krüppel-like factor 4 (KLF4): What we currently know. *Gene* 611, 27–37. doi: 10.1016/j.gene.2017.02.025
- Graupera, M., and Potente, M. (2013). Regulation of angiogenesis by PI3K signaling networks. *Exp. Cell Res.* 319, 1348–1355. doi: 10.1016/j.yexcr.2013.02.021
- Henderson, J., Withford-Cave, J. M., Duffy, D. L., Cole, S. J., Sawyer, N. A., Gulbin, J. P., et al. (2005). The EPAS1 gene influences the aerobic-anaerobic contribution in elite endurance athletes. *Hum. Genet.* 118, 416–423. doi: 10.1007/s00439-005-0066-0
- Huang, D. W., Sherman, B. T., and Lempicki, R. A. (2009). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* 4, 44–57. doi: 10.1038/nprot.2008.211
- Huang, M., Yang, B., Chen, H., Zhang, H., Wu, Z., Ai, H., et al. (2019). The fine-scale genetic structure and selection signals of Chinese indigenous pigs. *Evol. Appl.* 13, 458–475. doi: 10.1111/eva.12887
- Jia, C., Kong, X., Koltjes, J. E., Gou, X., Yang, S., Yan, D., et al. (2016). Gene co-expression network analysis unraveling transcriptional regulation of high-altitude adaptation of Tibetan Pig. *PLoS One* 11:e0168161. doi: 10.1371/journal.pone.0168161
- Kashtan, N., Itzkovitz, S., Milo, R., and Alon, U. (2004). Efficient sampling algorithm for estimating subgraph concentrations and detecting network motifs. *Bioinformatics* 20, 1746–1758. doi: 10.1093/bioinformatics/bth163
- Kim, D., Kwon, Y. K., and Cho, K. H. (2008). The biphasic behavior of incoherent feed-forward loops in biomolecular regulatory networks. *Bioessays* 30, 1204–1211. doi: 10.1002/bies.20839
- Kostka, D., and Spang, R. (2004). Finding disease specific alterations in the co-expression of genes. *Bioinformatics* 20, i194–i199. doi: 10.1093/bioinformatics/bth909
- Kumar, L., and Futschik, M. E. (2007). Mfuzz: a software package for soft clustering of microarray data. *Bioinformatics* 2, 5–7. doi: 10.6026/97320630002005
- Lan, G., and Tu, Y. (2013). The cost of sensitive response and accurate adaptation in networks with an incoherent type-1 feed-forward loop. *J. R. Soc. Interface* 10:20130489. doi: 10.1098/rsif.2013.0489
- Langfelder, P., Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* 9:559. doi: 10.1186/1471-2105-9-559
- Le, D. H., and Kwon, Y. K. (2013). A coherent feedforward loop design principle to sustain robustness of biological networks. *Bioinformatics* 29, 630–637. doi: 10.1093/bioinformatics/btt026
- Lee, S. J., Zhang, M., Hu, K., Lin, L., Zhang, D., and Jin, Y. (2015). CCN1 suppresses pulmonary vascular smooth muscle contraction in response to hypoxia. *Pulm. Circ.* 5, 716–722. doi: 10.1086/683812
- Li, C., Li, X., Xiao, J., Liu, J., Fan, X., Fan, F., et al. (2019). Genetic changes in the EPAS1 gene between tibetan and han ethnic groups and adaptation to the plateau hypoxic environment. *PeerJ*. 7:e7943. doi: 10.7717/peerj.7943
- Li, G., Schmitt, H., Johnson, W. M., Lee, C., Navarro, I., Cui, J., et al. (2020). Integral role for lysyl oxidase-like-1 in conventional outflow tissue function and behavior. *FASEB J.* 34, 10762–10777. doi: 10.1096/fj.202000702RR
- Li, H., Ouyang, X. P., Jiang, T., Zheng, X., He, P., and Zhao, G. (2018). MicroRNA-296: a promising target in the pathogenesis of atherosclerosis? *Mol. Med.* 24:12. doi: 10.1186/s10020-018-0012-y
- Li, M., Jin, L., Ma, J., Tian, S., Li, R., and Li, X. (2016). Detecting mitochondrial signatures of selection in wild Tibetan pigs and domesticated pigs. *Mitochondrial DNA A DNA Mapp. Seq. Anal.* 27, 747–752. doi: 10.3109/19401736.2014.913169
- Li, M., Tian, S., Jin, L., Zhou, G., Li, Y., Zhang, Y., et al. (2013). Genomic analyses identify distinct patterns of selection in domesticated pigs and Tibetan wild boars. *Nat. Genet.* 45, 1431–1438. doi: 10.1038/ng.2811
- Liu, X., Zhang, Y., Li, Y., Pan, J., Wang, D., Chen, W., et al. (2019). EPAS1 gain-of-function mutation contributes to high-altitude adaptation in Tibetan horses. *Mol. Biol. Evol.* 36, 2591–2603. doi: 10.1093/molbev/msz158
- Long, K., Feng, S., Ma, J., Zhang, J., Jin, L., Tang, Q., et al. (2019). Small non-coding RNA transcriptome of four high-altitude vertebrates and their low-altitude relatives. *Sci. Data* 6:192. doi: 10.1038/s41597-019-0204-5
- Ma, Y., Han, X., Huang, C., Zhong, L., Adeola, A. C., Irwin, D. M., et al. (2019). Population genomics analysis revealed origin and high-altitude adaptation of Tibetan pigs. *Sci. Rep.* 9:11463. doi: 10.1038/s41598-019-47711-6
- Mangan, S., and Alon, U. (2003). Structure and function of the feed-forward loop network motif. *Proc. Natl. Acad. Sci. U.S.A.* 100, 11980–11985. doi: 10.1073/pnas.2133841100
- Mangan, S., Zaslaver, A., and Alon, U. (2003). The coherent feedforward loop serves as a sign-sensitive delay element in transcription networks. *J. Mol. Biol.* 334, 197–204. doi: 10.1016/j.jmb.2003.09.049
- McLeay, R. C., Lesluyes, T., Cuellar Partida, G., and Bailey, T. L. (2012). Genome-wide in silico prediction of gene expression. *Bioinformatics* 28, 2789–2796. doi: 10.1093/bioinformatics/bts529

- Melincovici, C. S., Boşca, A. B., Şuşman, S., Mărginean, M., Mişu, C., Istrate, M., et al. (2018). Vascular endothelial growth factor (VEGF) – key factor in normal and pathological angiogenesis. *Rom. J. Morphol. Embryol.* 59, 455–467.
- Milo, R., Itzkovitz, S., Kashtan, N., Levitt, R., Shen-Orr, S., Ayzenshtat, I., et al. (2004). Superfamilies of evolved and designed networks. *Science* 303, 1538–1542. doi: 10.1126/science.1089167
- Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D., and Alon, U. (2002). Network motifs: simple building blocks of complex networks. *Science* 298, 824–827. doi: 10.1126/science.298.5594.824
- Minchenko, D. O., Tsybmal, D. O., Riabovol, O. O., Viletska, Y. M., Lahanovska, Y. O., Sliusar, M. Y., et al. (2019). Hypoxic regulation of EDN1, EDNRA, EDNRB, and ECE1 gene expressions in ERN1 knockdown U87 glioma cells. *Endocr. Regul.* 53, 250–262. doi: 10.2478/enr-2019-0025
- Morikawa, M., Mitani, Y., Holmborn, K., Kato, T., Koinuma, D., Maruyama, J., et al. (2019). The ALK-1/SMAD/ATOH8 axis attenuates hypoxic responses and protects against the development of pulmonary arterial hypertension. *Sci. Signal.* 12:eay4430. doi: 10.1126/scisignal.aay4430
- Narang, V., Ramli, M. A., Singhal, A., Kumar, P., de Libero, G., Poidinger, M., et al. (2015). Automated identification of core regulatory genes in human gene regulatory networks. *PLoS Comput. Biol.* 11:e1004504. doi: 10.1371/journal.pcbi.1004504
- Nishio, Y., Usuda, Y., Matsui, K., and Kurata, H. (2008). Computer-aided rational design of the phosphotransferase system for enhanced glucose uptake in *Escherichia coli*. *Mol. Syst. Biol.* 4:160. doi: 10.1038/msb4100201
- Peng, Y., Yang, Z., Zhang, H., Cui, C., Qi, X., Luo, X., et al. (2011). Genetic variations in Tibetan populations and high-altitude adaptation at the Himalayas. *Mol. Biol. Evol.* 28, 1075–1081. doi: 10.1093/molbev/msq290
- Qi, X., Zhang, Q., He, Y., Yang, L., Zhang, X., Shi, P., et al. (2019). The transcriptomic landscape of yaks reveals molecular pathways for high altitude adaptation. *Genome Biol. Evol.* 11, 72–85. doi: 10.1093/gbe/evy264
- Ravasz, E., Somera, A. L., Mongru, D. A., Oltvai, Z. N., and Barabási, A. L. (2002). Hierarchical organization of modularity in metabolic networks. *Science* 297, 1551–1555. doi: 10.1126/science.1073374
- Rosenfeld, N., Elowitz, M. B., and Alon, U. (2002). Negative autoregulation speeds the response times of transcription networks. *J. Mol. Biol.* 323, 785–793. doi: 10.1016/S0022-2836(02)00994-4
- Shalgi, R., Brosh, R., Oren, M., Pilpel, Y., and Rotter, V. (2009). Coupling transcriptional and post-transcriptional miRNA regulation in the control of cell fate. *Aging* 1, 762–770. doi: 10.18632/aging.100085
- Shang, P., Li, W., Tan, Z., Zhang, J., Dong, S., Wang, K., et al. (2020). Population genetic analysis of ten geographically isolated tibetan pig populations. *Animals (Basel)* 10:1297. doi: 10.3390/ani10081297
- Shatat, M. A., Tian, H., Zhang, R., Tandon, G., Hale, A., Fritz, J. S., et al. (2014). Endothelial Krüppel-like factor 4 modulates pulmonary arterial hypertension. *Am. J. Respir. Cell Mol. Biol.* 50, 647–653. doi: 10.1165/rcmb.2013-0135OC
- Sheng, J., Liu, D., Kang, X., Chen, Y., Jiang, K., and Zheng, W. (2018). Egr-1 increases angiogenesis in cartilage via binding Netrin-1 receptor DCC promoter. *J. Orthop. Surg. Res.* 13:125. doi: 10.1186/s13018-018-0826-x
- Shen-Orr, S. S., Milo, R., Mangan, S., and Alon, U. (2002). Network motifs in the transcriptional regulation network of *Escherichia coli*. *Nat. Genet.* 31, 64–68. doi: 10.1038/ng881
- Simonson, T. S., Yang, Y., Huff, C. D., Yun, H., Qin, G., Witherspoon, D. J., et al. (2010). Genetic evidence for high-altitude adaptation in Tibet. *Science* 329, 72–75. doi: 10.1126/science.1189406
- Tan, G., and Lenhard, B. (2016). TFBSTools: an R/bioconductor package for transcription factor binding site analysis. *Bioinformatics* 32, 1555–1556. doi: 10.1093/bioinformatics/btw024
- Tang, Q., Gu, Y., Zhou, X., Jin, L., Guan, J., Liu, R., et al. (2017). Comparative transcriptomics of 5 high-altitude vertebrates and their low-altitude relatives. *Gigascience* 6, 1–9. doi: 10.1093/gigascience/gix105
- Weber, E. J., and Hoo, Z. H. (2018). Why sample size estimates? *Emerg. Med. J.* 35, 755–756. doi: 10.1136/emered-2018-207763
- Xiang, K., Ouzhuluobu, Peng, Y., Yang, Z., Zhang, X., Cui, C., et al. (2013). Identification of a Tibetan-specific mutation in the hypoxic gene EGLN1 and its contribution to high-altitude adaptation. *Mol. Biol. Evol.* 30, 1889–1898. doi: 10.1093/molbev/mst090
- Xie, Q., Xie, J., Tian, T., Ma, Q., Zhang, Q., Zhu, B., et al. (2017). Hypoxia triggers angiogenesis by increasing expression of LOX genes in 3-D culture of ASCs and ECs. *Exp. Cell Res.* 352, 157–163. doi: 10.1016/j.yexcr.2017.02.011
- Xu, S., Li, S., Yang, Y., Tan, J., Lou, H., Jin, W., et al. (2011). A genome-wide search for signals of high-altitude adaptation in Tibetans. *Mol. Biol. Evol.* 28, 1003–1011. doi: 10.1093/molbev/msq277
- Yan, X., Liu, Z., and Chen, Y. (2009). Regulation of TGF-beta signaling by Smad7. *Acta Biochim. Biophys. Sin. (Shanghai)* 41, 263–272. doi: 10.1093/abbs/gmp018
- Yanai, I., Benjamin, H., Shmoish, M., Chalifa-Caspi, V., Shklar, M., Ophir, R., et al. (2005). Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. *Bioinformatics* 21, 650–659. doi: 10.1093/bioinformatics/bti042
- Young, A. I. (2019). Solving the missing heritability problem. *PLoS Genet.* 15:e1008222. doi: 10.1371/journal.pgen.1008222
- Zhang, B., Chamba, Y., Shang, P., Wang, Z., Ma, J., Wang, L., et al. (2017). Comparative transcriptomic and proteomic analyses provide insights into the key genes involved in high-altitude adaptation in the Tibetan pig. *Sci. Rep.* 7:3654. doi: 10.1038/s41598-017-03976-3
- Zhang, W., Fan, Z., Han, E., Hou, R., Zhang, L., Galaverni, M., et al. (2014). Hypoxia adaptations in the grey wolf (*Canis lupus chanco*) from Qinghai-Tibet Plateau. *PLoS Genet.* 10:e1004466. doi: 10.1371/journal.pgen.1004466
- Zhang, Y., Gou, W., Ma, J., Zhang, H., Zhang, Y., and Zhang, H. (2017). Genome methylation and regulatory functions for hypoxic adaptation in Tibetan chicken embryos. *PeerJ* 5:e3891. doi: 10.7717/peerj.3891

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2021 Wang, Guo, Liu, Zhang, Cui, Ding, Wang, Wang and Wang. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



Positive Selection in Gene Regulatory Factors Suggests Adaptive Pleiotropic Changes During Human Evolution

Vladimir M. Jovanovic^{1,2}, Melanie Sarfert¹, Carlos S. Reyna-Blanco^{3,4},
Henrike Indrischek^{5,6,7}, Dulce I. Valdivia⁸, Ekaterina Shelest⁹ and Katja Nowick^{1*}

¹ Human Biology and Primate Evolution, Freie Universität Berlin, Berlin, Germany, ² Bioinformatics Solution Center, Freie Universität Berlin, Berlin, Germany, ³ Department of Biology, University of Fribourg, Fribourg, Switzerland, ⁴ Swiss Institute of Bioinformatics, Fribourg, Switzerland, ⁵ Max Planck Institute of Molecular Cell Biology and Genetics, Dresden, Germany, ⁶ Max Planck Institute for the Physics of Complex Systems, Dresden, Germany, ⁷ Center for Systems Biology Dresden, Dresden, Germany, ⁸ Evolutionary Genomics Laboratory and Genome Topology and Regulation Laboratory, Genetic Engineering Department, Center for Research and Advanced Studies of the National Polytechnic Institute (CINVESTAV-Irapuato), Irapuato, Mexico, ⁹ Centre for Enzyme Innovation, University of Portsmouth, Portsmouth, United Kingdom

OPEN ACCESS

Edited by:

Susana Seixas,
University of Porto, Portugal

Reviewed by:

Magdalena Gayà-Vidal,
Research Center in Biodiversity
and Genetic Resources
(CIBIO-InBIO), Portugal
Martin Kuhlwillm,
Pompeu Fabra University, Spain

*Correspondence:

Katja Nowick
katja.nowick@fu-berlin.de

Specialty section:

This article was submitted to
Evolutionary and Population Genetics,
a section of the journal
Frontiers in Genetics

Received: 31 January 2021

Accepted: 19 April 2021

Published: 17 May 2021

Citation:

Jovanovic VM, Sarfert M,
Reyna-Blanco CS, Indrischek H,
Valdivia DI, Shelest E and Nowick K
(2021) Positive Selection in Gene
Regulatory Factors Suggests
Adaptive Pleiotropic Changes During
Human Evolution.
Front. Genet. 12:662239.
doi: 10.3389/fgene.2021.662239

Gene regulatory factors (GRFs), such as transcription factors, co-factors and histone-modifying enzymes, play many important roles in modifying gene expression in biological processes. They have also been proposed to underlie speciation and adaptation. To investigate potential contributions of GRFs to primate evolution, we analyzed GRF genes in 27 publicly available primate genomes. Genes coding for zinc finger (ZNF) proteins, especially ZNFs with a Krüppel-associated box (KRAB) domain were the most abundant TFs in all genomes. Gene numbers per TF family differed between all species. To detect signs of positive selection in GRF genes we investigated more than 3,000 human GRFs with their more than 70,000 orthologs in 26 non-human primates. We implemented two independent tests for positive selection, the branch-site-model of the PAML suite and aBSREL of the HyPhy suite, focusing on the human and great ape branch. Our workflow included rigorous procedures to reduce the number of false positives: excluding distantly similar orthologs, manual corrections of alignments, and considering only genes and sites detected by both tests for positive selection. Furthermore, we verified the candidate sites for selection by investigating their variation within human and non-human great ape population data. In order to approximately assign a date to positively selected sites in the human lineage, we analyzed archaic human genomes. Our work revealed with high confidence five GRFs that have been positively selected on the human lineage and one GRF that has been positively selected on the great ape lineage. These GRFs are scattered on different chromosomes and have been previously linked to diverse functions. For some of them a role in speciation and/or adaptation can be proposed

based on the expression pattern or association with human diseases, but it seems that they all contributed independently to human evolution. Four of the positively selected GRFs are KRAB-ZNF proteins, that induce changes in target genes co-expression and/or through arms race with transposable elements. Since each positively selected GRF contains several sites with evidence for positive selection, we suggest that these GRFs participated pleiotropically to phenotypic adaptations in humans.

Keywords: primate, transcription factor, speciation, great apes, archaic humans, gene regulatory evolution, phenotypic evolution, KRAB-ZNF

INTRODUCTION

Phenotypic differences between individuals and species could be partly explained by the sequence differences in coding parts of genes, and partly by the variation in gene regulatory mechanisms (Lewontin, 1974; Wray, 2007; Wittkopp and Kalay, 2012; Lappalainen et al., 2013; Orgogozo et al., 2015; Perdomo-Sabogal et al., 2016; Anderson et al., 2020). The latter can be caused by changes in the DNA sequence of a regulatory region of a gene that could affect its expression (Siepel and Arbiza, 2014), as well as by changes in the sequence of so-called gene regulatory factors (GRFs) that could affect their target genes (Nowick et al., 2011; Perdomo-Sabogal et al., 2014). GRFs are involved in gene regulation in various ways, such as defining timing and tissue-specificity of a gene's expression. They include proteins such as transcription factors that bind directly to DNA, cofactors that bind to the transcription factors, histone modifying enzymes, and (long) non-coding RNAs (Latchman, 1997; Zhu et al., 2013; Perdomo-Sabogal et al., 2014; Li et al., 2015; Wingender et al., 2015). GRFs usually display pleiotropic characteristics and regulate more than one gene, hence it has been assumed that their sequence, especially of functional domains, should be subject to long-term constraints and conserved even between species (Wagner and Lynch, 2008; Perdomo-Sabogal et al., 2014; Anderson et al., 2020). However, it has also been suggested that the gene regulatory mechanisms evolve under less selective constraints, compared to their target genes (e.g., Anderson et al., 2020). This led to the description of GRFs as having domain-islands of conservation "in a sea of divergence" (Wagner and Lynch, 2008). Non-deleterious evolutionary changes in GRFs regularly occur both within and outside functionally important regions in homeodomain- and zinc-finger (ZNF) proteins, among other GRF families, exemplifying their role for driving intra- and interspecific morphologic innovations and phenotypic diversity (Wagner and Lynch, 2008; Nowick et al., 2013; Perdomo-Sabogal et al., 2014).

Among the most intriguing questions of phenotypic diversity between species are the striking differences between humans and great apes (Nickel et al., 2008; Varki et al., 2008), but also between great apes and other primates. This particular phenotypic diversity cannot be attributed to the sequence differences alone, but must involve expression changes as well (King and Wilson, 1975; Khaitovich et al., 2005). The genetic dissimilarity between humans and their closest relatives, chimpanzees, has been estimated to be 1.2% in average, with slightly higher dissimilarity in non-coding compared to coding regions (Elango et al., 2006;

Kronenberg et al., 2018). Taking non-alignable parts of the genome into account, i.e., insertions, deletion, rearrangements, the difference amounts to 3–4%. In contrast, the dissimilarity between humans and the rhesus macaque, a more distant primate species, was estimated to be substantially higher with 6.46% in average, or up to 9.24% when considering small insertions and deletions (Gibbs et al., 2007; Su et al., 2016). Some of the sequence changes could be the outcome of neutral evolution, whereas others could also be the result of ongoing adaptive interactions among genomes and the environment and hence positive selection (Varki et al., 2008). A paramount example for non-neutral selection in a human GRF has been demonstrated for the FOXP2 gene, where two codons seem to be positively selected in humans in comparison to chimpanzees (Enard et al., 2002). Given the genes' phenotype association, these selected changes were linked to language skills, one of the most distinctive human capabilities (Fisher, 2019).

Depending on the method used for dating, the human lineage diverged approximately 5.5–11.5 million years ago (Ma) from its closest lineage of chimpanzee and bonobo (Patterson et al., 2006; Langergraber et al., 2012; Amster and Sella, 2016; Besenbacher et al., 2019). The resulting human phenotype has been traditionally seen as driven by ongoing adaptations to local environments and niches (Varki et al., 2008; Lachance and Tishkoff, 2013; Jeong and Di Rienzo, 2014). The identification of genes evolving by positive selection can reveal the route in which organisms adapt to their environment (Casola and Hahn, 2009), and answer some substantially important biological questions (Su et al., 2016), for instance, how a specific phenotype arose. It has been long hypothesized that identifying the genes that have been positively selected along the human lineage, in contrast to neutral and purifying selection in their closest relatives (great apes, primates), could offer insight into the biologically significant genetic changes that distinctly characterize humans (Clark et al., 2003; Mundy and Cook, 2003; Nickel et al., 2008; Daub et al., 2017; Goodwin and de Guzman Strong, 2017). Consequently, many studies to date have tested the human genome for signatures of positive selection using several approaches. Several studies aimed at detecting adaptive changes in protein-coding genes in genome-wide scans of a set of primate species (Nielsen et al., 2005; Voight et al., 2006; Sabeti et al., 2007; Kosiol et al., 2008; Goodwin and de Guzman Strong, 2017), usually without a major overlap of positively selected genes among these analyses. Meanwhile, other studies chose an approach to detect the selection on a polygenic level, within groups of genes unified by the function of encoded

proteins (e.g., Nowick et al., 2011; Daub et al., 2013, 2017; Afanasyeva et al., 2018).

From some of the first genome-wide scans of primate genomes for positive selection (Nielsen et al., 2005; Gibbs et al., 2007), it was clear that even in the closest primate lineages the adaptive selection pressures could have undergone different paths and left footprints in different genes. In their comparison of macaque, chimpanzee and human genomes, Gibbs et al. (2007) found that only one human gene and as many as 12 chimpanzee genes were uniquely under positive selection, suggesting a lineage-specific selection. Nevertheless, common selective pressures may create uniform selection patterns across a whole set of species (Schultz and Sackton, 2019), necessitating for broader studies of selection on the branch-level (e.g., Daub et al., 2017).

The aim of our study was to identify genes with signatures of positive selection among the primate GRFs. We specifically focused on two branches in the primate tree, the great apes (Hominidae) and human (Hominina) lineages. Since the power to detect positive selection depends on the number of available sequences (Anisimova et al., 2001; Gayà-Vidal and Albà, 2014), we included all 27 currently available primate genomes to add power to our analysis. There are several lists or databases that compile regulatory factors (e.g., Ravasi et al., 2010; Tripathi et al., 2013; Lambert et al., 2018). For this study we chose 3,344 genes from a published human GRF catalog (Perdomo-Sabogal and Nowick, 2019), which we consider to be the most comprehensive GRF catalog to date. Interestingly, positive selection of some of the GRFs from that catalog has been previously proposed among primate species (for instance, 3 of 36 genes in Nielsen et al., 2005; 35 of 187 genes in Su et al., 2016), albeit with fewer species included in the analyses, and at population level within humans (Perdomo-Sabogal and Nowick, 2019). Positively selected mutations are rarely observed as polymorphic sites (Gayà-Vidal and Albà, 2014). Rather, they should have been rapidly fixed by adaptive selection (Gayà-Vidal and Albà, 2014; Slodkowitz and Goldman, 2020). Interestingly, FOXP2 (mentioned above) was shown not to be recently positively selected on the human lineage after thorough investigation of its variation within modern humans (Atkinson et al., 2018; Fisher, 2019). Therefore, investigating the polymorphism of positively selected codons at population level enables the exclusion of potential false positive candidates identified on the species level.

Here, we compile a high quality set of primate GRFs under positive selection by (1) taking advantage of the completeness of our input data (2) by extensive filtering and curation of the input data to reduce false positives and (3) by verification of potential sites under positive selection by inclusion of chimpanzee and human population variation data.

MATERIALS AND METHODS

Compilation of a Primate GRF Data Set

Starting with the list of Ensembl IDs of human GRFs (**Supplementary Data 1** from Perdomo-Sabogal and Nowick, 2019), the orthologous coding sequences from 27 primate genomes, including human, available at Ensembl/Compara

(Vilella et al., 2009) and NCBI GenBank were downloaded using *biomaRt* (Durinck et al., 2005) and *rentrez* (Winter, 2017) R packages. Thus all gene sequences were from the Ensembl release 100¹ (Yates et al., 2020) and NCBI GenBank Release 237², both from April 2020.

The age of the GRF relative to the species tree was taken from GenTree³ (Shao et al., 2019). In parallel, genome-wide prediction of transcription factor sequences was made. The protein sequences were downloaded from Ensembl. InterPro (Blum et al., 2021) domain annotations were run for each proteome using Blast2GO (Götz et al., 2008). The resulting genome-wide domain prediction tables were confronted with a manually curated collection of TF-type DNA-binding domains (DBD) using an R script. To be considered as a TF, a protein had to possess at least one TF-type DBD. The TF-type DBD list was collected as described in Shelest (2017). In brief, InterPro database was scanned for DNA binding domains excluding non-TF DBD types (such as, e.g., helicases, nucleases, DNA repair enzymes, etc.). The obtained set was confronted with the DBD list from the DBD database (Wilson et al., 2008), which helped to clean the set from non-TF domains, and then was additionally cleaned from redundancies. Plant-specific DBDs were not included in the final list. The proteins were further arranged in TF family groups as described in Shelest (2017).

Alternative splicing could produce false positive results in the positive selection analysis, if non-orthologous exons were aligned. Therefore, for each gene the human MANE (Matched Annotation between NCBI and EBI) transcript isoform was selected as the representative human sequence, and a temporary dataset was created, which contained that isoform and the sequences of all isoforms of non-human orthologs. These orthologous sequences were then clustered using the *MMseqs2* program (Steinegger and Söding, 2017), with at least 80% identity of sequences within clusters, and a cluster containing the human sequence was selected for further analyses.

The selected sequence cluster was stored in a *DNAStrSet* and converted into an *AAStrSet* containing the protein sequences using the *translate* function from the R package *biostrings* (Pagès et al., 2019). A multiple sequence alignment was then created from the amino acid fasta file by MAFFT (Katoh and Standley, 2013). With the output alignment and the original DNA sequence, the codon alignment was created using the program *PAL2NAL* (Suyama et al., 2006). The phylogenetic species tree of primates, needed for the analyses, was downloaded from the 10kTrees Project v.3⁴ (Accessed March 1st 2020, Arnold et al., 2010). If the ortholog was not found in all genomes, this species tree was adjusted with the function *drop.tip* from the R package *ape* (Paradis and Schliep, 2019). Using one universal topology of the species tree made the analyses robust to differences in substitution rates among genes, but also to the fact that gene regulatory factors frequently produces distorted gene trees (Anderson et al., 2020).

¹www.ensembl.org/

²www.ncbi.nlm.nih.gov/search/

³gentree.ioz.ac.cn

⁴<https://10ktrees.nunn-lab.org/Primates/>

Branch-Site Analysis of Positive Selection

Selective pressure acting on protein-coding genes is regularly quantified by estimating the ratio of non-synonymous to synonymous substitutions (ω) between the coding parts of homologs. We detected branches under positive selection by employing two different maximum likelihood methods: the branch-site model (Yang and Nielsen, 2002) using CODEML of the PAML v4.9 suite (Yang, 2007), and aBSREL (Smith et al., 2015) of the HyPhy v2.5 suite⁵. As applied here, both methods calculate the probability of positive selection ($\omega > 1$) of a fraction of sites in a predefined foreground branch of the species tree, namely both the human lineage (taxonomically, subtribe Hominina), and the great apes branch (taxonomically, family Hominidae). The age of divergence of these branches from their sister branches is 5.5–11.5 Ma, 16–26 Ma, respectively (e.g., Dos Reis et al., 2018; Besenbacher et al., 2019).

ABSREL additionally allows for different selection pressures (ω) acting on different branches, while the CODEML branch-site model assumes constant ω -values for the respective site classes in all background branches (Yang and Nielsen, 2002; Smith et al., 2015). The human lineage is especially interesting as it could shed more light onto the phenotypic evolution of human species, while the great ape lineage could contribute to our knowledge about the divergence of great apes from other Old World monkeys and gibbons. In both, CODEML and aBSREL methods, the empirical p -values were obtained assuming a χ^2 distribution of the log-ratio tests (LRT). Multiple testing of a large number of GRFs was accounted for by the Benjamini-Hochberg method.

Given that alignments of non-homologous positions are known to frequently cause false positives in such analyses (Fletcher and Yang, 2010), we visually inspected those GRF alignments that showed signs of positive selection. If necessary, the alignments were manually corrected in MEGA X software (Kumar et al., 2018) and the analyses repeated. When the LRT suggested positive selection in the CODEML framework, the Bayes empirical Bayes (BEB, included in PAML v4.9) method was used to calculate posterior probabilities and identify codons that might represent positively selected sites (PSS) in the foreground branch (Yang, 2007). In the same manner, we ran MEME from HyPhy v2.5 (Murrell et al., 2012) in order to detect PSS for genes that were positively selected according to aBSREL. Our further analyses focused on candidate genes and codons, for which positive selection was supported by both methods.

Further Analysis of Positively Selected GRFs and Sites

In order to comprehensively understand the adaptiveness of positively selected sites, we need to define the impact of the particular codon change on the phenotype. The positively selected gene candidates, retrieved as explained above, were manually investigated by mining several genetic and protein databases. We searched for publications covering functions and associations of positively selected GRFs with human phenotypes

and diseases in UniProt⁶ (UniProt Consortium, 2019), Ensembl (see text footnote 1; Yates et al., 2020), NCBI databases (see text footnote 2), OMIM⁷ (Online Mendelian Inheritance in Man, 2020) and FANTOM5/FANTOM CAT⁸ (Hon et al., 2017; Kawaji et al., 2017). In order to uncover common pathways and themes in our set of positively selected gene candidates, we looked at the gene ontology (GO) and KEGG pathway classifications, excluding the gene expression related terms. We used the Expression Atlas⁹ (Papatheodorou et al., 2020), ProteomicsDB¹⁰ (Samaras et al., 2020), and Bgee¹¹ (Bastian et al., 2021) to investigate the candidates' expression pattern among human and other available primates' tissues and potential differential expression. Possible interactions and co-expression between the positively selected GRFs were investigated by performing string-based protein-protein interaction network analysis with STRING v11¹² (Szklarczyk et al., 2019), calculating the proteome co-regulatory network with ProteomeHD (Kustatscher et al., 2019), and by mining the EdgeExpressDB (FANTOM4-EEDB¹³, Kawaji et al., 2009). The position of positively selected codons in relation to protein domains and functional sites was checked at UniProt and manually, following the specific protein-related literature. A short summary on used databases is given in **Table 1**.

The variation of the PSS in modern humans was investigated in the Ensembl genome browser. The population data therein include, among others, results from 1,000 Genomes project, NCBI ALFA, Gambian Genome Variation Project, GnomAD, TOPMed, ExAC, and Korea1K project. If the ancestral (pre-selection) codon variant was recorded in these projects, the frequency of polymorphisms among populations, their phenotype correlates and calculation of mutual linkage disequilibrium was further investigated in NCBI dbSNP database and Ensembl. In order to be conservative in detecting PSS, we discarded the PSS that showed polymorphisms in modern human populations. However, that does not fully discard the possibility that these PSS could have been positively selected in some time period, either in recent times (but not reaching fixation yet), or in ancient times (so that nowadays new variants relaxed from the selective pressure appear). The sequence and variation of PSS in archaic humans (Vindija and Altai Neanderthal, Denisovans) was read from the Ancient Genome Browser of the Max Planck Institute for Evolutionary Anthropology, Leipzig¹⁴. Taking archaic variation into consideration, we could approximately date the positive selection process to before or after separation of Neanderthals and Denisovans from Anatomically Modern Humans. The polymorphism of human PSS in non-human great apes was investigated in data from a number of publicly available and published datasets

⁶www.uniprot.org/

⁷<https://www.omim.org>

⁸<https://fantom.gsc.riken.jp/cat/v1/>

⁹www.ebi.ac.uk/gxa/home

¹⁰www.proteomicsdb.org

¹¹<https://bgee.org/>

¹²www.string-db.org/

¹³<https://fantom.gsc.riken.jp/4/edgeexpress/subnet/>

¹⁴bioinf.eva.mpg.de/jbrowse/

⁵www.hyphy.org

TABLE 1 | List of mined databases with the description of stored information and covered topic.

Database	Description
Ensembl	Provides access to genomes, their annotation information, domains, structures, external links and some analysis tools. In addition, it contains information on variation for human and chimpanzee genomes, and population-based distribution of the variation
EMBL-EBI Expression Atlas	Provides the freely available information about gene and protein expression, from microarray, bulk and single cell RNA-Seq studies
NCBI (National Center for Biotechnology Information)	Provides access to gene, genome and protein sequences, structure and annotation information, publications, as well as information on genome variation (for instance, SNPs)
UniProt	Contains various general information on proteins, their sequence and structure, function, domains and ontology
OMIM (Online Mendelian Inheritance in Man)	Contains information on known mendelian disorders and focuses on the relationship between phenotype and genotype
GO (Gene Ontology)	Contains information on the functions of genes, together with their hierarchical classification into functional categories
Kyoto Encyclopedia of Genes and Genomes (KEGG)	Provides information on a large array of high-level functions of genes and proteins, collecting their orthologs, metabolic pathways, disease-related network variation etc.
ProteomicsDB	Provides information on human proteome, isoforms of proteins, expression per tissue, and other analytics
Bgee	Retrieve and compare gene expression patterns between animal species
STRING	Contains information on protein-protein interactions
ProteomeHD	Contains information on co-regulation between the proteins, with additional analytics and GO terms
EdgeExpressDB (FANTOM4-EEDB)	Provides information on co-expression networks between expressed components of mammalian genomes
FANTOM CAT (FANTOM5)	Provides atlases of functional parts of mammalian genomes such as promoters, enhancers, lncRNAs and miRNAs, together with metadata

(Prüfer et al., 2012; Prado-Martinez et al., 2013; Scally et al., 2013; De Manuel et al., 2016; Fonsere et al., 2021), providing information from 111 chimpanzees, 17 bonobos, 42 gorillas and 10 orangutans (**Supplementary Table 4**). For each gene that includes PSS, we aligned all matching reads from these individuals, which allowed us to infer the state and variation of PSS among great apes.

RESULTS

Not every human GRF had orthologs in all 26 other primate species' genomes. One of the reasons is the incompleteness of many genomes under investigation, that limits the possibility of multispecies sequence alignment and comparison (Kronenberg et al., 2018). Furthermore, we excluded read-through transcripts and GRFs originating from recent duplication events in the human lineage, i.e., after Homo-Pan divergence, as no thorough orthology relationship could be established for those cases. Even though it is recommended to include duplicated loci into genome-wide scans for selection (Han et al., 2009), recently duplicated genes often experience gene conversion, which has

been shown to elevate false detection of positive selection in paralogs in both site and branch-site models (Casola and Hahn, 2009). Our conservative approach resulted in a set of 3,221 protein-coding genes, with 72,086 non-human orthologs (**Supplementary Table 1**).

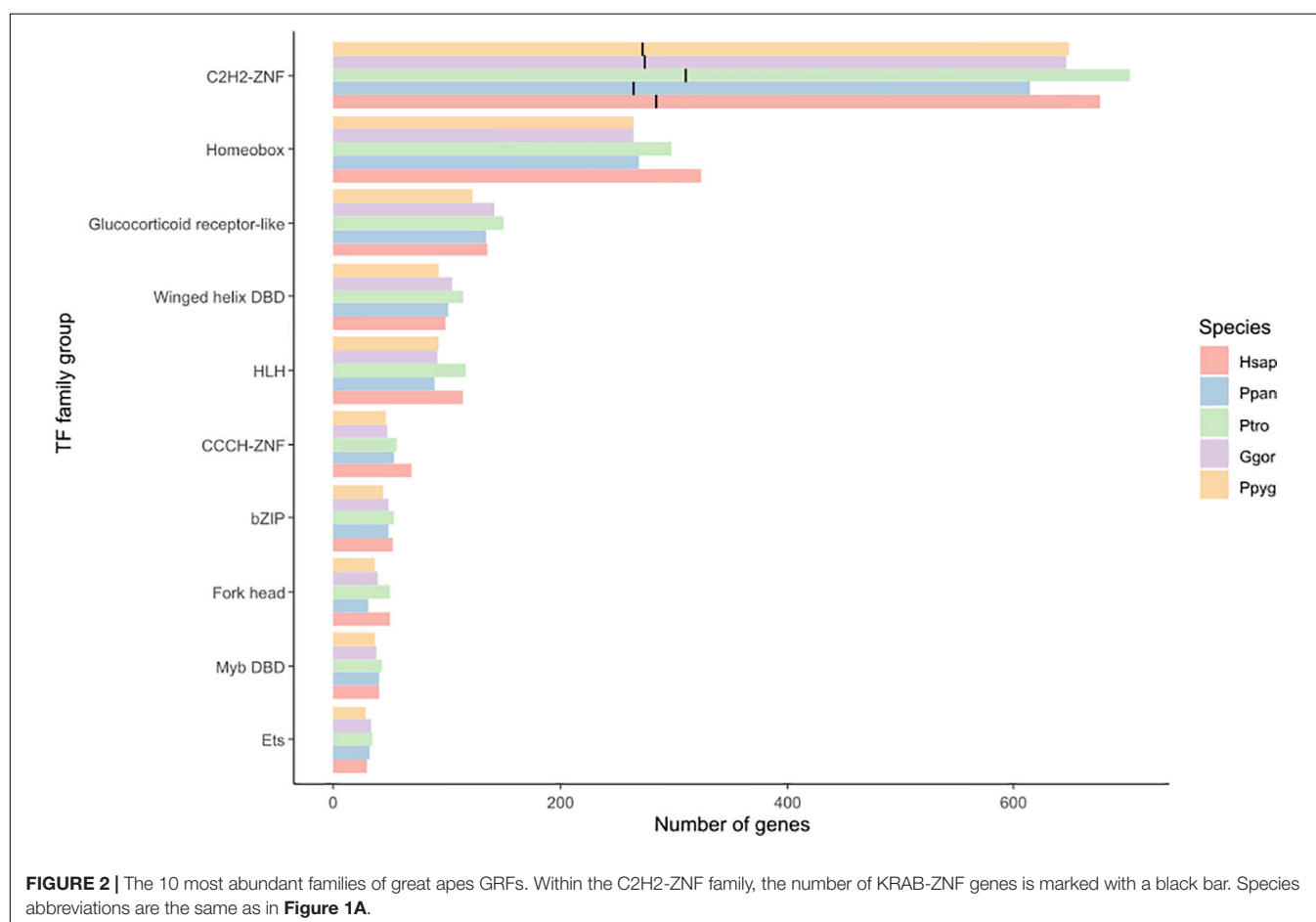
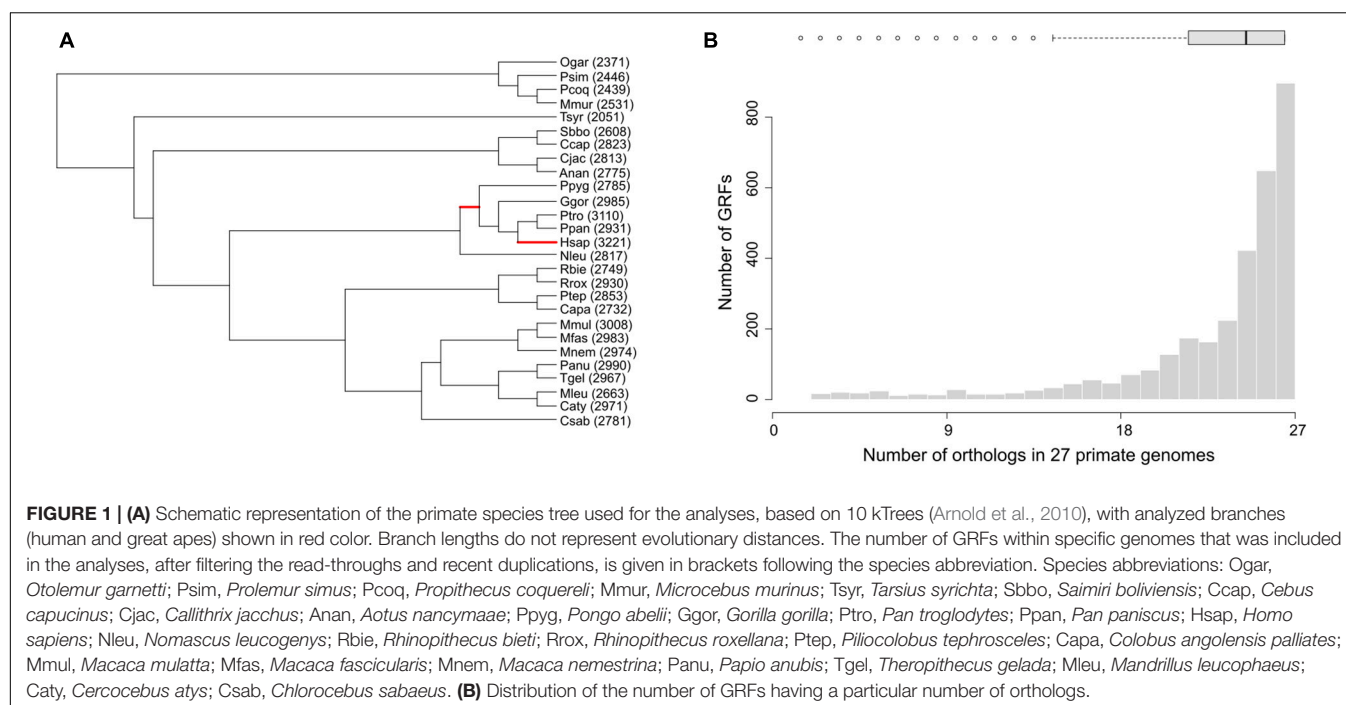
The distribution of the total number of detected orthologs in each genome, after applying the filters, as well as the number of GRFs that arose in the specific clades, is shown in **Figure 1A**. Only 3,044 human GRFs from our dataset were dated by GenTree, and of those, 78 arose within the primate clade. However, 177 studied GRFs were not dated by GenTree, half of them belonging to the zinc finger family previously seen as harboring many primate-specific genes (Nowick and Stubbs, 2010). The median number of orthologs per GRF was 25, meaning that we could identify orthologs in almost all investigated species. Nevertheless, some GRFs have clearly fewer orthologs, either due to their recent origin within primates or due to missing data (**Figure 1B**).

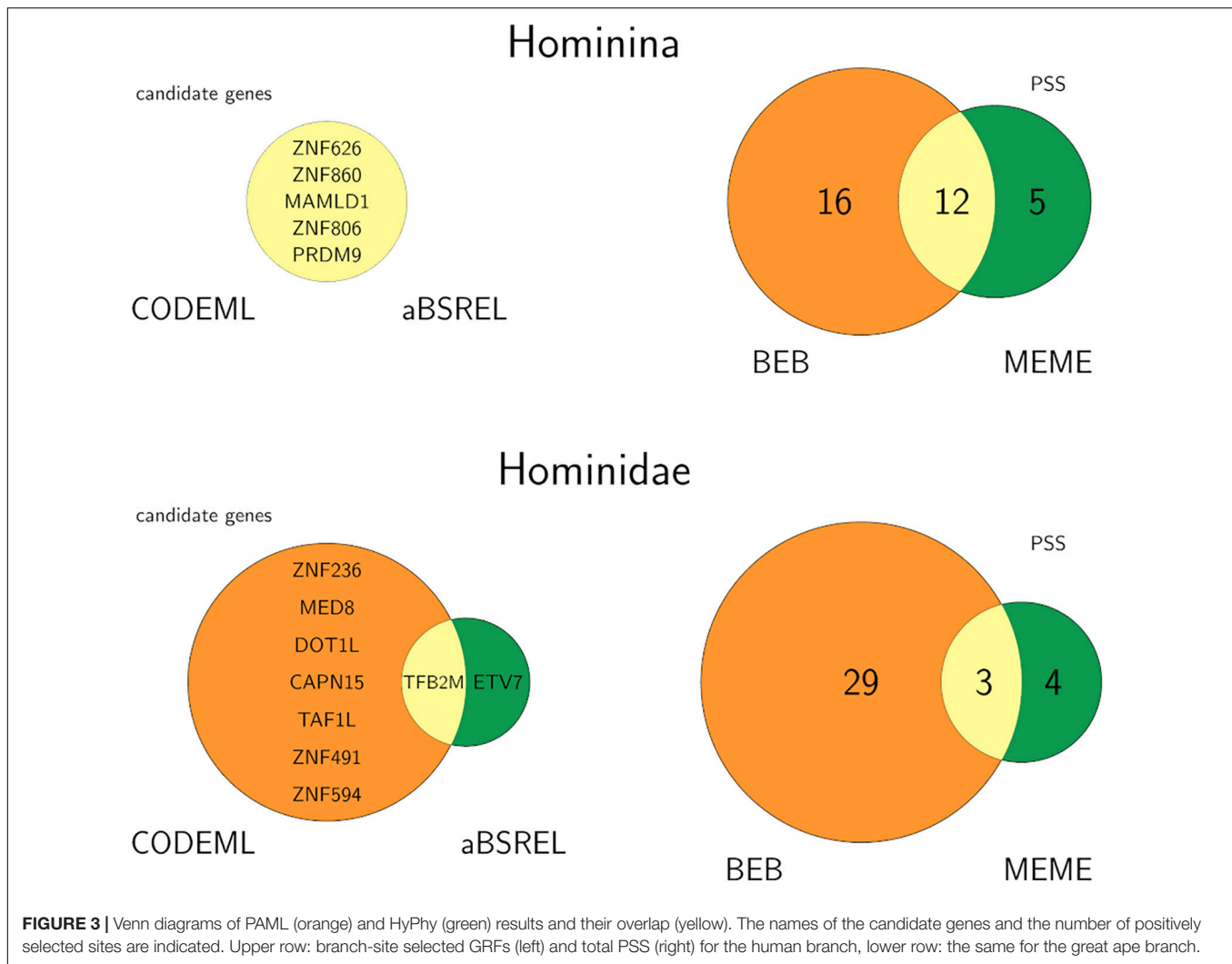
Transcription factors are usually classified into families based on their DNA-binding domain (Wingender et al., 2013; Wingender et al., 2015; Shelest, 2017). The most common GRF family in the analyzed great ape genomes were zinc fingers (especially C2H2-ZNF) and homeobox, followed by glucocorticoid receptors (**Figure 2**). Although differences in the number of genes per GRF family between species exist, they were not significant (Chi-squared test, $p = 1$; **Supplementary Table 2**). Within the C2H2-ZNF family, KRAB-ZNF proteins were the most numerous, with 40–42% of all GRFs in great ape species (**Figure 2**).

Positive Selection in the Human Lineage

Since every method is known to produce false positives we decided to perform our analyses with two commonly used packages, PAML and HyPhy, and to keep only the candidates detected by both (**Figure 3**). CODEML detected 52, and HyPhy 61 candidate genes, before correcting for multiple testing. For the branch-site analysis both procedures indicated the same five genes for positive selection in the human (Hominina) lineage: MAMLD1, and four KRAB-zinc-finger containing proteins (PRDM9, ZNF626, ZNF806, and ZNF860) (**Supplementary Table 3**). To learn more about these five candidates, we next investigated their evolutionary age and expression patterns. All candidates seem to have arisen at very different time points. According to GenTree, MAMLD1 arose within the land vertebrate clade (Tetrapoda), PRDM9 is seen as common for placental mammals (even though several studies identify it as the earliest in its protein family, being present already in the ancestors of chordates; Birtle and Ponting, 2006; Imbeault et al., 2017; Helleboid et al., 2019), while three of them appeared in primate clade: ZNF626 within Simiiformes, ZNF860 within Catarrhini, and ZNF806 originated within great apes. This indicates that changes in relatively old and relatively young genes show signs of positive selection on the human lineage.

Further, it seems that different tissues could be affected by these changes. MAMLD1, ZNF626, and ZNF860 are ubiquitously expressed in human tissues, while PRDM9 and ZNF806 are predominantly expressed in testes, developing ovaries, and parts of the central nervous system. Interestingly, MAMLD1 is





additionally seen as important for the male gonad development (GO:BP term 0008584), and has the highest expression in gonads in comparison to other organs in almost all primates included in the Bgee database. In the FantomCAT both, PRDM9 and MAMLD1, were associated with testes as well, whereas ZNF860 was associated with B-cells and ZNF626 with middle temporal gyrus. A SNP nearby ZNF626 was associated with bipolar disorder. Finally, no co-expression and protein-protein interaction between all the positively selected candidates was found in FANTOM4-EEDB database nor by STRING analysis, and there was not enough expression data for building a proteome co-regulatory network of these genes at ProteomeHD. Taken together, we did not find a common expression pattern nor sufficient data indicating a functional link between the five candidate genes, but it is worth mentioning that at least three of them might have important roles in gonads or the nervous system.

Within the five positively selected genes in the human branch, a total of 33 codons was detected as positively selected sites (PSS) by at least one of the BEB or MEME procedures

(**Supplementary Table 3**). Of those, 12 within three genes were detected by both procedures (**Figure 3** and **Table 2**).

In MAMLD1 there were two candidate PSS detected, that involved an exchange of amino acids with different physico-chemical properties. These codon sequences are fixed without variation in human populations (**Table 2**). At the same time, they are not variable among bonobo, chimpanzee and gorilla populations (**Supplementary Table 4**). These features make them ideal candidates for positively selected sites.

PRDM9 exhibits a strong signature of positive selection, which empowers the identification of seven PSS, all of which but one are distributed among six of 14 zinc-finger domains present in the protein. Most changes cause alterations of amino acid properties. The codons 573, 629, and 657 are located between the histidine residues of the zinc finger domain that coordinate the zinc ion, while the codons 591 and 737 are at the α -helix positions – 1 and 6, respectively, that specify DNA-binding (Brayer et al., 2008; Oliver et al., 2009). These positions were also among the three positions found to be under positive selection by Oliver et al. (2009). PSS in PRDM9 show variation within humans, with

TABLE 2 | The 11 positively selected codons (PSS) within three genes with positive selection in the human branch that were detected by BEB and MEME, along with the respective nucleotide and amino acid (in brackets) changes.

Gene/PSS	Nucl(AA) change	SNP in modern human	Decision
MAMLD1			
726	AGT (S) > AGA (R)	/	True PSS
728	GGC (G) > GAC (D)	/	True PSS
PRDM9			
155	CCT (P) > TCT (S)	/	True PSS
573	ACA (T) > ATA (I)	rs199686868	True PSS
591	CGG CAG GTT (R Q V) > TGG (W)	rs200381384	True PSS
629	ACA (T) > AGA (R)	rs112192848	True PSS
657	ACA (T) > AGA (R)	rs112679149	True PSS
681	AG A T (R S) > ACT (T)	rs6875787	Minor allele
737	TGT ATT (C I) > AGA (R)	/	True PSS
ZNF860			
219	CAA (Q) > CTA (L)	/	True PSS
348	GAC (D) > GAA (E)	rs13064905	False positive
464	A C GT (S R) > CAT (H)	rs1808125	False positive

The existence of non-synonymous single nucleotide variation (SNP) in modern humans is also included, as well as the decision on PSS after applying a rigorous quality check.

occasional appearance of the ancestral state, i.e., the sequence seen in great ape genomes (Table 2). These mutations are rare in most human population samples, reaching for the closely positioned SNPs rs199686868 and rs200381384 (codons 573 and 591) frequencies of 0.02 and 0.06 in Gambian and Korean populations, respectively. Even though positioned in codons that are close to one another in the genome, these SNPs do not exhibit linkage disequilibrium in the Gambian population (Ensembl) and can therefore be seen as independent and furthermore, not a result of a recent selective sweep. Codons 629 and 657 represent the same substitution, at the same position in relation to functional histidines of two neighboring ZNF domains. This is a result of already recognized concerted evolution within PRDM9 (Oliver et al., 2009; Schwartz et al., 2014). The population-wide alignment of Pan PRDM9 sequences showed high diversity in the number of ZNF domains, but also in their sequence at the DNA-binding sites (Groeneveld et al., 2012). The homologous DNA-binding position (−1) in the fourth ZNF domain of any Pan sequence does not harbor the human-specific codon 591 (TGG). The homologous position (+6) of codon 737 is seen in two recognized PRDM9 zinc finger alleles in Pan (alleles D and Z in Groeneveld et al., 2012), with one other DNA-binding position (+3) of these alleles being the same as in humans. Population stratification of both Homo and Pan PRDM9 sequences was previously shown by Schwartz et al. (2014).

Codon 681 of PRDM9 constitutes an exception of this set of PSS as the ancestral codon state (AGT) is described as rs6875787. Not reported in the 1,000 Human Genomes Project, this polymorphism was nonetheless found in other population genomics projects (for instance, in GnomAD genomes dataset), with frequencies of the “ancestral” nucleotide G of around 0.75 in the overall modern human population (Figure 4A),

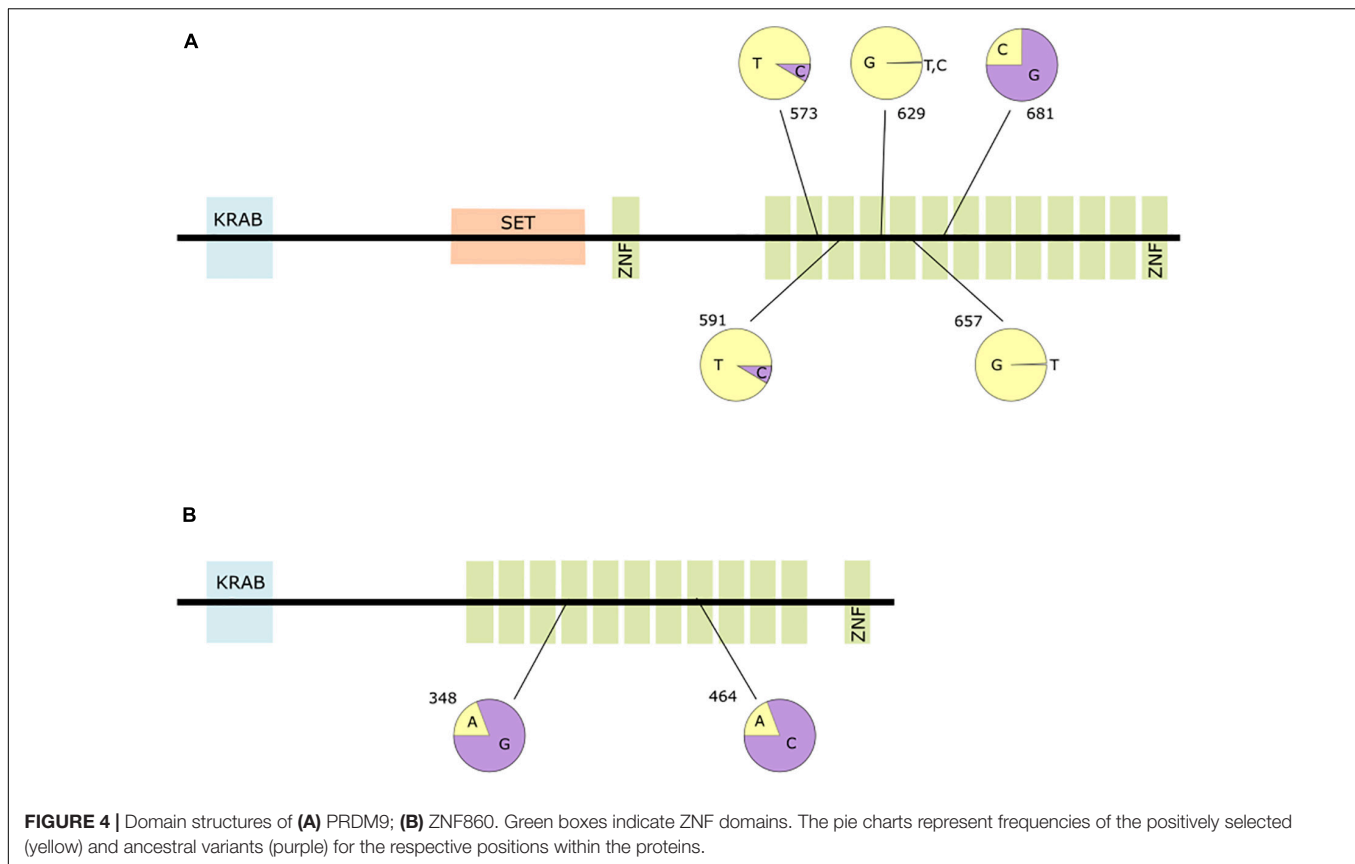
and “positively selected” C of around 0.25. Moreover, we cannot conclude about the presence of this polymorphism in Neanderthal populations, due to limited data. Most of the chimpanzee and bonobo ZNF alleles of PRDM9 have AGT at the DNA-binding position 6 (Groeneveld et al., 2012). This PSS candidate constitutes rather a case of human genetic variation where one of two major alleles was chosen to be the reference. It may also be that this region is located within a region biased for the sequencing technology used in some genome projects leading to an omission of that SNP. In any case, we rule it out as a true positively selected site in the human lineage.

Two of three PSS candidates in ZNF860 (codons 348 and 464) lay within the zinc finger domains. Their ancestral codon states, as well as the ancestral codon states of the candidate obtained by MEME only (codon 609), are reaching the frequencies up to 0.80 and 0.90 in African populations (Figure 4B). They more likely represent the ancient variations within the human lineage, than reverse mutations to the ancestral states. The archaic humans have either ancestral, modern or heterozygous states of these codons. These three codons also stand in linkage disequilibrium, with correlation (r^2) between alleles 0.5–1. This implies that the alleles we got as positively selected are part of one haplo-block. We regard these codons as false positives, induced by variation that is not present in the human reference genome, as well as a potential result of bottleneck in non-African human populations. This leaves ZNF860 with only one PSS, codon 219. This codon shows no variation within the analyzed bonobo, chimpanzee, gorilla and orangutan samples (Supplementary Table 4).

In total, our strict analysis recovered nine codons that show signs of positive selection. All nine PSS had the same sequence in the archaic humans (Neanderthals and Denisovans). Even with the caveat of limited data for these genomes, there was no polymorphism detected at the PSS in the high coverage sequenced archaic humans. It can be concluded that the adaptive selection happened along the human branch, before the divergence between Neanderthals, Denisovans and anatomically modern humans.

In ZNF626, seven sites were detected to be positively selected by the BEB method only. Interestingly, three of them are located within the KRAB domain of the protein. Additionally, a frameshift mutation occurred in the human gene within codon position 503, degenerating its last, 12th zinc finger domain. Fourteen codons after the frameshift were detected by both MEME and BEB, and were thereafter excluded as false positives.

We found that ZNF806 is part of a series of duplications within great apes. First, ZNF285 was duplicated in apes branch (Hominoidea) giving rise to ZNF285B, that is positioned nearby, but on the other strand of chromosome 19. The high similarity of those recently formed paralogs can lead to genome assembly mistakes as seen in our analysis for the fragmented gibbon genome. For this reason, gibbon ZNF285 and ZNF285B had to be excluded from our respective analyses. The next duplication yielded ZNF806 in all great ape genomes, positioned at human chromosome 2 or its homolog 2B, or on an unplaced scaffold in the orang-utan genome. Yet another duplication happened in the Homo/Pan branch yielding a paralog, present in all three available genomes at chromosome 20. Although all four paralogous genes



have similar sequences, their orthology relationship was not resolved in Ensembl/Compara. The gene tree built in MEGA X (Kumar et al., 2018) supports a duplication of ZNF806.

Positive Selection in Great Apes

In the great ape lineage, CODEML and aBSREL gave different results, overlapping only in one gene (Figure 3). Positive selection was detected by CODEML in eight GRFs (CAPN15, DOT1L, MED8, TAF1L, TFB2M, ZNF236, ZNF491, ZNF594), and by aBSREL in two (ETV7, TFB2M). The single overlap was gene TFB2M, which is expressed in all human tissues. TFB2M is a nuclear gene that is a part of the mitochondrial transcription initiation complex, and as such it is required for basal transcription of mitochondrial DNA (Falkenberg et al., 2002) but also for replication and packaging of mtDNA and ribosome biogenesis (Bonawitz et al., 2006). This gene is seen as having a critical role in mitochondrial DNA gene expression, and mutations in the gene or deviation in its expression have been associated with mitochondrial DNA depletion syndromes, Parkinson disease (Grünewald et al., 2016), and autism spectrum disorder (Park et al., 2018). Within TFB2M, both BEB and MEME methods detected three PSS that are not located within any domain. None of these PSS were variable within human and non-human population data we analyzed.

As the overlap between the methods was too small to investigate the potential coexpression and interaction between the positively selected genes, we included all the positively

selected candidates in the great apes branch in these analyses. However, similar to the positively selected GRFs in the human lineage, there was no co-expression, protein-protein interaction, and co-regulatory network of these genes found by STRING and ProteomeHD.

DISCUSSION

In our study, starting with a list of over 3,000 human GRFs, we identified five and one orthologs with significant signs of positive selection in the human and great apes lineages, respectively. To our knowledge, three of the identified GRFs, namely ZNF626, ZNF806, and ZNF860, have not been reported in previous analyses of positive selection among primate species (for instance in Nielsen et al., 2005; Su et al., 2016; Van Der Lee et al., 2017). The number of our candidate genes is small in contrast to the genome-wide studies (e.g., Su et al., 2016; Van Der Lee et al., 2017), but is in line with studies that estimated the proportion of adaptive amino acid substitutions as low in humans (Fay et al., 2001; Zhang and Li, 2005; Boyko et al., 2008). Furthermore, we focused only on GRFs, which constitute about one sixth of all protein-coding sequences (20,448 in human reference genome, assembly GRCh38.p13, Ensembl). We included only 78 of 254 primate specific GRFs (Shao et al., 2019), as the others were found in <3 genomes. We took effort to reduce the number of false positives that likely falsely increased the number of positively selected

genes in previous studies. Firstly, in comparison to previous analyses of positive selection within primates, we included more genomes—all the 27 available from Ensembl—thus substantially increasing the power to detect signatures of positive selection. Secondly, we excluded very divergent sequences from our sets of orthologs by performing an ortholog clustering step. High divergence might lead to the saturation of non-synonymous changes and thus affect the power of our branch-site tests (Gharib and Robinson-Rechavi, 2013). Our approach is, however, limited by the incompleteness of non-human primate genomes and the GRF sequences therein. Third, we manually investigated alignments that resulted in detection of positive selection to exclude those cases, where the respective signal was caused by alignment of non-homologous codons. In most cases, improving the alignments resulted in losing the statistical significance. Fourth, we consider only those candidates as reliable, which were detected by both, CODEML and aBSREL, methods. This certainly helped to deflate the false positive rate for detection of selected GRFs in the human lineage, in scope of the recent finding that around 35% of the human genome is subject to incomplete lineage sorting among the African apes (Kronenberg et al., 2018).

Our findings highlight the role of population data leveraged to the detection of signs of positive selection. Given only reference genomes, we detected 12 codons within five positively selected genes in the human lineage. However, looking into modern human population variance, we have found the PSS sequence variants that are the same as seen in the ancestral lineages (present in at least some of the great ape genomes). Some of them were common polymorphisms, some had very low frequencies. The first are either balanced polymorphisms or the result of a selective sweep within distinct human populations. They are, however, not fixed in the human genome, and were discarded as false positives. The low frequency of the latter polymorphisms allowed us to keep them as good candidates. Together with the PSS that showed no polymorphism, they comprise a set of nine positively selected codons in three genes (MAMLD1, PRDM9, and ZNF860). All of them were present in Neanderthal and Denisovan genomes, thus dating the adaptive selection episode(s) to after the divergence of the Pan/Homo branches (~6.5–7.5 Ma; Amster and Sella, 2016), and before the divergence among lineages that led to Neanderthal, Denisovan and anatomically modern humans 765,000–550,000 years ago.

In order to clearly identify sequence changes that could distinguish human and primate adaptive phenotypes, genetic variation and population data also need to be analyzed from non-human primate species. As seen in one of our positively selected genes, PRDM9, the variation among chimpanzees and bonobos is high (Groeneveld et al., 2012), and there are alleles/haplotypes that are the same as in the human genome. Based on our results, balanced polymorphisms in the human genome, like the ones seen in ZNF860 and PRDM9, as well as polymorphism resulting from incomplete lineage sorting and present in some populations of great apes, are prone to be wrongly indicated as PSS. Those cases remain a methodological challenge for the detection of adaptive selection from (only) reference genomes. We thus investigated variation of the detected PSS in 180 non-human great ape individuals. Three PSS detected for the

human lineage showed no variation in either human, or any great ape populations (Table 2 and Supplementary Table 4), strongly indicating that they do not represent incomplete lineage sorting among great apes, but are rather cases of true positive selection. Similarly, the candidate PSS for the great ape lineage within TFB2B are monomorphic, and therefore most likely the true positively selected sites. All in all, our strict approach and inclusion of variation data from great ape individuals led us to obtaining a short but high confidence list of GRFs with signs of positive selection on the human or great ape lineage, and PSS within them.

The adaptive importance of candidate genes should also be seen in their function and interconnections. However, our analysis yielded a small number of positively selected genes. Only one gene, TFB2M, was recovered as positively selected in the great apes branch and having sites under positive selection. This gene is crucial for proper functions of mitochondria as part of the mitochondrial transcription initiation complex that is necessary for expression of all genes encoded in the mitochondrial genome. Mitochondria and mitochondrially-encoded genes are essential for providing energy for all cellular functions. Impaired mitochondria have been associated with several diseases, including mitochondrial DNA depletion syndromes (Correia et al., 2011), diabetes, Parkinson's, deafness and cancer (Wallace, 2005). The effect of a particular SNP within this gene, c.790C > T, has been seen as delaying the unloading of DNA from TFB2M, thus increasing the mitochondrial DNA expression (Park et al., 2018). If the change in the ease of DNA-TFB2M detachment was also influenced by the PSS we detected, we could speculate at this point, that it might be related to the expression level of respiratory chain complex genes, and may have allowed better energy production efficiency for tissues including the brain.

In the human lineage we detected five positively selected GRFs. Four of them belong to the GRFs that possess a KRAB domain together with zinc finger domains, so called KRAB-ZNF proteins. KRAB-ZNFs themselves constitute the largest class of GRFs within the human genome (Mark et al., 1999; Yang et al., 2017). It is worth mentioning, that even though KRAB-ZNFs represent ~40% of the C2H2-ZNF protein family in great apes, they constitute 80% of positively selected GRFs in the human lineage in this study. Our results are thus in agreement with earlier findings that KRAB-ZNF proteins evolve rapidly (Nowick and Stubbs, 2010; Nowick et al., 2011; Zhao and Kishino, 2020). Some additional and previously known modes of their evolution, such as changes of zinc-finger copy numbers and loss of KRAB domains (Nowick and Stubbs, 2010; Nowick et al., 2011; Shao et al., 2019) were not within the scope of our analysis, but could also have happened by natural selection.

KRAB-ZNF proteins have been implicated in many important functions, such as genomic imprinting, cell differentiation, metabolic control, brain development, but also phenomena like sexual dimorphism and speciation (Nowick et al., 2013; Jacobs et al., 2014; Yang et al., 2017). Recently, it was discovered that at least some KRAB-ZNFs, such as ZNF91/93, are important for recognition and transcriptional silencing of transposable elements (Jacobs et al., 2014; Helleboid et al., 2019). The

KRAB-ZNF family is also enriched among genes with differential expression between human and chimpanzee prefrontal cortex (Nowick et al., 2009).

Three of the KRAB-ZNFs identified in this study (ZNF626, ZNF806, and ZNF860) are largely unexplored to date. ZNF626 was pointed out as a candidate gene involved in posttraumatic stress disorder in European American individuals of the United States Army (Stein et al., 2016) and seems to be associated with bipolar disorder (Hon et al., 2017). It is highly expressed in the middle temporal gyrus, which is involved in language processing, for instance while reading (Acheson and Hagoort, 2013). Hippocampus-specific somatic mutations within ZNF806 have been identified in 9 out of 17 patients with sporadic Alzheimer's disease (Parcerisas et al., 2014). Previous association of the ZNF806 SNP rs4953961 with tardive dystonia, one of the serious types of extrapyramidal symptoms that antipsychotics can cause, was shown to be erroneous and probably relatable to similar genomic regions (Kanahara et al., 2021). In our study, we have revealed that ZNF806 is in a group with three paralogous ZNF sequences, one of which can be a potential candidate for this symptom. ZNF860 has been associated with early-onset type 2 diabetes mellitus and prostate cancer, and its higher expression is seen as an indicator for gastric cancer (Dmitriev et al., 2015; Yamada et al., 2018; Pan et al., 2019). These findings indicate that these positively selected ZNF genes are playing a role in complex phenotypes. Interestingly, two of those genes, ZNF626 and ZNF806, may be associated with the brain.

PRDM9, on the other hand, is a well-studied gene. It specifies the sites of meiotic DNA double-strand breaks that initiate meiotic recombination in mice and humans. PRDM9 is known to bind to specific DNA sequences with its DNA binding domain, to induce methylation to adjacent nucleosomes, and to recruit or activate the meiotic machinery (Baudat et al., 2010; Billings et al., 2013). Although its function can be seen as essential, a human adult knock-out was reported, pointing to differences in humans vs. non-primate mammals, and supporting the possibility of alternative mechanisms of localizing human meiotic crossover (Narasimhan et al., 2016). PRDM9 was previously reported as a candidate for positive selection in a number of studies (e.g., Oliver et al., 2009; Schwartz et al., 2014; Daub et al., 2015). At least in mice, this gene has been considered as a speciation gene causing infertility in hybrids (Mihola et al., 2009), and a similar role has been proposed for the primate clade (Daub et al., 2015). In addition to that, Schwartz et al. (2014) have speculated that positive selection at positions dedicated to DNA binding and specificity can lead to differential usage of binding motifs, which may result in abovementioned hybrid sterility and contribute to speciation in the primate lineage. Our work further supports a major role for PRDM9 in speciation of humans.

Our fifth candidate, MAMLD1, seems to be important for sex determination and the development of male genitalia. Mutations in MAMLD1 have been found to cause hypospadias type 2, a disorder of sex development in which the male urethral opening is moved ventrally, and genitalia of XY individuals can appear female-like (Fukami et al., 2006). Changing the position of urethral opening will also functionally

block successful mating, in terms of delivering sperm into the genital tract of females. Indeed, strong differences in size and morphology of testicles and penis exist between humans and chimpanzees and seem to be related to their mating strategies (Harcourt et al., 1981; Brindle and Opie, 2016). Genes involved in reproduction are considered prime candidates for driving speciation. Several other studies of positive selection in the human genome have also disclosed genes involved in spermatogenesis and transcriptional regulation (e.g., Clark et al., 2003; Gayà-Vidal and Albà, 2014). While PRDM9 might create a species barrier at the postzygotic level, MAMLD1 might have been involved in establishing a barrier prior to fertilization.

However, no co-expression, interaction or co-regulation among our candidate genes was previously reported. It may be speculated that the sets of genes they regulate are independent, or acting in different pathways, such that the epistasis among them could not be detected with the currently available data. The seeming independence and the possibility of participation in complex phenotypes can be accounted for if we include the potential pleiotropic effect. Namely, since GRFs usually regulate the expression of several to many genes, they can induce various physiological and morphological consequences within cells, tissues or at the level of whole organisms (Stern, 2000; Wagner and Lynch, 2008; Wagner and Zhang, 2011). These consequences could be independently adaptive. It has been shown that small mutations, even within a single gene, may provide a rapid path to phenotypic adaptation (Linnen et al., 2013). The different PSS that we identified within one gene, can of course add to the pleiotropic effect and be associated with different traits. This has been reported before for some genes, where different polymorphic sites had different trait associations (Flint and Mackay, 2009; Mackay et al., 2009). Yet another plausible explanation for the lack of interaction among our candidate genes is that the selective pressures on them were acting at different timepoints after the split of the human lineage. There might have been millions of years between the selective events, so that they can likely be considered to have occurred independently from each other.

Taken together, our study points out six candidate GRFs that experienced positive selection in great apes and human branches. These GRFs did not show common patterns of co-expression or co-regulation. Hence, we concluded that the effect of several PSS within some of these genes, could have had pleiotropic effects on different phenotypic traits, and that the effect of all candidate GRFs may have been epistatic toward the same goal—adaptation. Detection of mainly KRAB-ZNF genes as positively selected GRFs in the human lineage, along with the recent duplication events for at least one of them (ZNF806), lead us to propose that these proteins are driving human-specific phenotypes by shifting target genes co-expression (as proposed by Nowick et al., 2009), and through arms race with transposable elements (Imbeault et al., 2017; Yang et al., 2017; Warren et al., 2020). The association with the brain for at least some of them further supports the notion that phenotypic and cognitive differences in the primate brain might have been caused by adaptive changes in regulatory factors.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author/s.

AUTHOR CONTRIBUTIONS

KN, HI, CR-B, and VJ conceptualized and developed the research idea and designed the study. CR-B, ES, DV, MS, and VJ built code pipelines and analyzed the data. VJ, MS, and KN interpreted the obtained results and prepared the manuscript. All authors have read, discussed and approved the final submitted manuscript.

REFERENCES

- Acheson, D. J., and Hagoort, P. (2013). Stimulating the brain's language network: syntactic ambiguity resolution after TMS to the inferior frontal gyrus and middle temporal gyrus. *J. Cogn. Neurosci.* 25, 1664–1677. doi: 10.1162/jocn_a_00430
- Afanasyeva, A., Bockwoldt, M., Cooney, C. R., Heiland, I., and Gossmann, T. I. (2018). Human long intrinsically disordered protein regions are frequent targets of positive selection. *Genome Res.* 28, 975–982. doi: 10.1101/gr.232645.117
- Amster, G., and Sella, G. (2016). Life history effects on the molecular clock of autosomes and sex chromosomes. *Proc. Natl. Acad. Sci. U S A.* 113, 1588–1593. doi: 10.1073/pnas.1515798113
- Anderson, J. A., Vilgalys, T. P., and Tung, J. (2020). Broadening primate genomics: new insights into the ecology and evolution of primate gene regulation. *Curr. Opin. Genet. Dev.* 62, 16–22. doi: 10.1016/j.gde.2020.05.009
- Anisimova, M., Bielawski, J. P., and Yang, Z. (2001). Accuracy and power of the likelihood ratio test in detecting adaptive molecular evolution. *Mol. Biol. Evol.* 18, 1585–1592. doi: 10.1093/oxfordjournals.molbev.a003945
- Arnold, C., Matthews, L. J., and Nunn, C. L. (2010). The 10kTrees Website: A New Online Resource for Primate Phylogeny. *Evol. Anthropol.* 19, 114–118. doi: 10.1002/evan.20251
- Atkinson, E. G., Audesse, A. J., Palacios, J. A., Bobo, D. M., Webb, A. E., Ramachandran, S., et al. (2018). No evidence for recent selection at FOXP2 among diverse human populations. *Cell* 174, 1424–1435. doi: 10.1016/j.cell.2018.06.048
- Bastian, F. B., Roux, J., Niknejad, A., Comte, A., Fonseca Costa, S. S., De Farias, T. M., et al. (2021). The Bgee suite: integrated curated expression atlas and comparative transcriptomics in animals. *Nucleic Acids Res.* 49, D831–D847. doi: 10.1093/nar/gkaa793
- Baudat, F., Buard, J., Grey, C., Fledel-Alon, A., Ober, C., Przeworski, M., et al. (2010). PRDM9 is a major determinant of meiotic recombination hotspots in humans and mice. *Science* 327, 836–840. doi: 10.1126/science.1183439
- Besenbacher, S., Hvilsom, C., Marques-Bonet, T., Mailund, T., and Schierup, M. H. (2019). Direct estimation of mutations in great apes reconciles phylogenetic dating. *Ecol. Evol.* 3, 286–292. doi: 10.1038/s41559-018-0778-x
- Billings, T., Parvanov, E. D., Baker, C. L., Walker, M., Paigen, K., and Petkov, P. M. (2013). DNA binding specificities of the long zinc-finger recombination protein PRDM9. *Genome Biol.* 14:R35. doi: 10.1186/gb-2013-14-4-r35
- Birtle, Z., and Ponting, C. P. (2006). Meisetz and the birth of the KRAB motif. *Bioinformatics* 22, 2841–2845. doi: 10.1093/bioinformatics/btl498
- Blum, M., Chang, H. Y., Chuguransky, S., Grego, T., Kandasamy, S., Mitchell, A., et al. (2021). The InterPro protein families and domains database: 20 years on. *Nucleic Acids Res.* 49, D344–D354. doi: 10.1093/nar/gkaa977
- Bonawitz, N. D., Clayton, D. A., and Shadel, G. S. (2006). Initiation and beyond: multiple functions of the human mitochondrial transcription machinery. *Mol. Cell* 24, 813–825. doi: 10.1016/j.molcel.2006.11.024
- Boyko, A. R., Williamson, S. H., Indap, A. R., Degenhardt, J. D., Hernandez, R. D., Lohmueller, K. E., et al. (2008). Assessing the evolutionary impact of amino acid

FUNDING

This work was supported by the Volkswagen Foundation within the initiative “Evolutionary Biology” (KN) and the Deutsche Forschungsgemeinschaft as part of the SPP 2205 (KN and MS). We further acknowledge funding from the Freie Universität Berlin for Open Access Publishing.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fgene.2021.662239/full#supplementary-material>

- mutations in the human genome. *PLoS Genet.* 4:e1000083. doi: 10.1371/journal.pgen.1000083
- Brayer, K. J., Kulshreshtha, S., and Segal, D. J. (2008). The protein-binding potential of C2H2 zinc finger domains. *Cell Biochem. Biophys.* 51, 9–19. doi: 10.1007/s12013-008-9007-6
- Brindle, M., and Opie, C. (2016). Postcopulatory sexual selection influences baculum evolution in primates and carnivores. *Proc. Biol. Sci.* 283:20161736. doi: 10.1098/rspb.2016.1736
- Casola, C., and Hahn, M. W. (2009). Gene conversion among paralogs results in moderate false detection of positive selection using likelihood methods. *J. Mol. Evol.* 68, 679–687. doi: 10.1007/s00239-009-9241-6
- Clark, A. G., Glanowski, S., Nielsen, R., Thomas, P. D., Kejariwal, A., Todd, M. A., et al. (2003). Inferring nonneutral evolution from human-chimp-mouse orthologous gene trios. *Science* 302, 1960–1963. doi: 10.1126/science.1088821
- Correia, R. L., Oba-Shinjo, S., Uno, M., Huang, N., and Marie, S. K. (2011). Mitochondrial DNA depletion and its correlation with TFAM, TFB1M, TFB2M and POLG in human diffusely infiltrating astrocytomas. *Mitochondrion* 11, 48–53. doi: 10.1016/j.mito.2010.07.001
- Daub, J. T., Dupanloup, I., Robinson-Rechavi, M., and Excoffier, L. (2015). Inference of evolutionary forces acting on human biological pathways. *Genome Biol. Evol.* 7, 1546–1558. doi: 10.1093/gbe/evv083
- Daub, J. T., Hofer, T., Cutivet, E., Dupanloup, I., Quintana-Murci, L., Robinson-Rechavi, M., et al. (2013). Evidence for polygenic adaptation to pathogens in the human genome. *Mol. Biol. Evol.* 30, 1544–1558. doi: 10.1093/molbev/mst080
- Daub, J. T., Moretti, S., Davydov, I. I., Excoffier, L., and Robinson-Rechavi, M. (2017). Detection of pathways affected by positive selection in primate lineages ancestral to humans. *Mol. Biol. Evol.* 34, 1391–1402. doi: 10.1093/molbev/msx083
- De Manuel, M., Kuhlilm, M., Frandsen, P., Sousa, V. C., Desai, T., Prado-Martinez, J., et al. (2016). Chimpanzee genomic diversity reveals ancient admixture with bonobos. *Science* 354, 477–481. doi: 10.1126/science.aag2602
- Dmitriev, A. A., Rosenberg, E. E., Krasnov, G. S., Geraschenko, G. V., Gordiyuk, V. V., Pavlova, T. V., et al. (2015). Identification of novel epigenetic markers of prostate cancer by NotI-microarray analysis. *Dis. Markers* 2015:241301. doi: 10.1155/2015/241301
- Dos Reis, M., Gunnell, G. F., Barba-Montoya, J., Wilkins, A., Yang, Z., and Yoder, A. D. (2018). Using phylogenomic data to explore the effects of relaxed clocks and calibration strategies on divergence time estimation: primates as a test case. *Syst. Biol.* 67, 594–615. doi: 10.1093/sysbio/syy001
- Durinck, S., Moreau, Y., Kasprzyk, A., Davis, S., De Moor, B., Brazma, A., et al. (2005). BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics* 21, 3439–3440. doi: 10.1093/bioinformatics/bti525
- Elango, N., Thomas, J. W., NISC Comparative Sequencing Program, and Yi, S. Y. (2006). Variable molecular clocks in hominoids. *Proc. Natl. Acad. Sci. U S A.* 103, 1370–1375. doi: 10.1073/pnas.0510716103
- Enard, W., Przeworski, M., Fisher, S. E., Lai, C. S., Wiebe, V., Kitano, T., et al. (2002). Molecular evolution of FOXP2, a gene involved in speech and language. *Nature* 418, 869–872. doi: 10.1038/nature01025

- Falkenberg, M., Gaspari, M., Rantanen, A., Trifunovic, A., Larsson, N. G., and Gustafsson, C. M. (2002). Mitochondrial transcription factors B1 and B2 activate transcription of human mtDNA. *Nat. Genet.* 31, 289–294. doi: 10.1038/ng909
- Fay, J. C., Wyckoff, G. J., and Wu, C. I. (2001). Positive and negative selection on the human genome. *Genetics* 158, 1227–1234. doi: 10.1093/genetics/158.3.1227
- Fisher, S. E. (2019). Human genetics: the evolving story of FOXP2. *Curr. Biol.* 29, R65–R67. doi: 10.1016/j.cub.2018.11.047
- Fletcher, W., and Yang, Z. (2010). The effect of insertions, deletions, and alignment errors on the branch-site test of positive selection. *Mol. Biol. Evol.* 27, 2257–2267. doi: 10.1093/molbev/msq115
- Flint, J., and Mackay, T. F. C. (2009). Genetic architecture of quantitative traits in flies, mice and humans. *Genome Res.* 19, 723–733. doi: 10.1101/gr.086660.108
- Fontser, C., Alvarez-Estape, M., Lester, J., Arandjelovic, M., Kuhlwlum, M., Dieguez, P., et al. (2021). Maximizing the acquisition of unique reads in noninvasive capture sequencing experiments. *Mol. Ecol. Resour.* 21, 745–761. doi: 10.1111/1755-0998.13300
- Fukami, M., Wada, Y., Miyabayashi, K., Nishino, I., Hasegawa, T., Nordenskjöld, A., et al. (2006). CXorf6 is a causative gene for hypospadias. *Nat. Genet.* 38, 1369–1371. doi: 10.1038/ng1900
- Gaya-Vidal, M., and Albà, M. M. (2014). Uncovering adaptive evolution in the human lineage. *BMC Genomics* 15:599. doi: 10.1186/1471-2164-15-599
- Gharib, W. H., and Robinson-Rechavi, M. (2013). The branch-site test of positive selection is surprisingly robust but lacks power under synonymous substitution saturation and variation in GC. *Mol. Biol. Evol.* 30, 1675–1686. doi: 10.1093/molbev/mst062
- Gibbs, R. A., Rogers, J., Katze, M. G., Bumgarner, R., Weinstock, G. M., Mardis, E. R., et al. (2007). Evolutionary and biomedical insights from the rhesus macaque genome. *Science* 316, 222–234. doi: 10.1126/science.1139247
- Goodwin, Z. A., and de Guzman Strong, C. (2017). Recent positive selection in genes of the mammalian epidermal differentiation complex locus. *Front. Genet.* 7:227. doi: 10.3389/fgene.2016.00227
- Götz, S., Garcia-Gomez, J. M., Terol, J., Williams, T. D., Nagaraj, S. H., Nueda, M. J., et al. (2008). High-throughput functional annotation and data mining with the Blast2GO suite. *Nucleic Acids Res.* 36, 3420–3435. doi: 10.1093/nar/gkn176
- Groeneveld, L. F., Atencia, R., Garriga, R. M., and Vigilant, L. (2012). High diversity at PRDM9 in chimpanzees and bonobos. *PLoS One* 7:e39064. doi: 10.1371/journal.pone.0039064
- Grünewald, A., Rygiel, K. A., Hepplewhite, P. D., Morris, C. M., Picard, M., and Turnbull, D. M. (2016). Mitochondrial DNA Depletion in Respiratory Chain-Deficient Parkinson Disease Neurons. *Ann. Neurol.* 79, 366–378. doi: 10.1002/ana.24571
- Han, M. V., Demuth, J. P., McGrath, C. L., Casola, C., and Hahn, M. W. (2009). Adaptive evolution of young gene duplicates in mammals. *Genome Res.* 19, 859–867. doi: 10.1101/gr.085951.108
- Harcourt, A. H., Harvey, P. H., Larson, S. G., and Short, R. V. (1981). Testis weight, body weight and breeding system in primates. *Nature* 293, 55–57. doi: 10.1038/293055a0
- Helleboid, P. Y., Heusel, M., Duc, J., Piot, C., Thorball, C. W., Coluccio, A., et al. (2019). The interactome of KRAB zinc finger proteins reveals the evolutionary history of their functional diversification. *EMBO J.* 38:e101220. doi: 10.15252/embj.2018101220
- Hon, C. C., Ramiłowski, J. A., Harshbarger, J., Bertin, N., Rackham, O. J., Gough, J., et al. (2017). An atlas of human long non-coding RNAs with accurate 5' ends. *Nature* 543, 199–204. doi: 10.1038/nature21374
- Imbeault, M., Helleboid, P. Y., and Trono, D. (2017). KRAB zinc-finger proteins contribute to the evolution of gene regulatory networks. *Nature* 543, 550–554. doi: 10.1038/nature21683
- Jacobs, F. M., Greenberg, D., Nguyen, N., Haeussler, M., Ewing, A. D., Katzman, S., et al. (2014). An evolutionary arms race between KRAB zinc-finger genes ZNF91/93 and SVA/L1 retrotransposons. *Nature* 516, 242–245. doi: 10.1038/nature13760
- Jeong, C., and Di Rienzo, A. (2014). Adaptations to local environments in modern human populations. *Curr. Opin. Genet. Dev.* 29, 1–8. doi: 10.1016/j.gde.2014.06.011
- Kanahara, N., Nakata, Y., and Iyo, M. (2021). Genetic association study detected misalignment in previous whole exome sequence: association study of ZNF806 and SART3 in tardive dystonia. *Psychiatr. Genet.* 31, 29–31. doi: 10.1097/YPG.0000000000000263
- Katoh, K., and Standley, D. M. (2013). MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Mol. Biol. Evol.* 30, 772–780. doi: 10.1093/molbev/mst010
- Kawaji, H., Kasukawa, T., Forrest, A., Carninci, P., and Hayashizaki, Y. (2017). The FANTOM5 collection, a data series underpinning mammalian transcriptome atlases in diverse cell types. *Sci. Data* 4:170113. doi: 10.1038/sdata.2017.113
- Kawaji, H., Severin, J., Lizio, M., Waterhouse, A., Katayama, S., Irvine, K. M., et al. (2009). The FANTOM web resource: from mammalian transcriptional landscape to its dynamic regulation. *Genome Biol.* 10:R40. doi: 10.1186/gb-2009-10-4-r40
- Khaitovich, P., Hellmann, I., Enard, W., Nowick, K., Leinweber, M., Franz, H., et al. (2005). Parallel patterns of evolution in the genomes and transcriptomes of humans and chimpanzees. *Science* 309, 1850–1854. doi: 10.1126/science.1108296
- King, M. C., and Wilson, A. C. (1975). Evolution at two levels in humans and chimpanzees. *Science* 188, 107–116. doi: 10.1126/science.1090005
- Kosiol, C., Vinař, T., da Fonseca, R. R., Hubisz, M. J., Bustamante, C. D., Nielsen, R., et al. (2008). Patterns of positive selection in six mammalian genomes. *PLoS Genet.* 4:e1000144. doi: 10.1371/journal.pgen.1000144
- Kronenberg, Z. N., Fiddes, I. T., Gordon, D., Murali, S., Cantsilieris, S., Meyerson, O. S., et al. (2018). High-resolution comparative analysis of great ape genomes. *Science* 360:eaar6343. doi: 10.1126/science.aar6343
- Kumar, S., Stecher, G., Li, M., Knyaz, C., and Tamura, K. (2018). MEGA X: molecular evolutionary genetics analysis across computing platforms. *Mol. Biol. Evol.* 35, 1547–1549. doi: 10.1093/molbev/msy096
- Kustatscher, G., Grabowski, P., Schrader, T. A., Passmore, J. B., Schrader, M., and Rappsilber, J. (2019). Co-regulation map of the human proteome enables identification of protein functions. *Nat. Biotechnol.* 37, 1361–1371. doi: 10.1038/s41587-019-0298-5
- Lachance, J., and Tishkoff, S. A. (2013). Population genomics of human adaptation. *Annu. Rev. Ecol. Evol. S* 44, 123–143. doi: 10.1146/annurev-ecolsys-110512-135833
- Lambert, S. A., Jolma, A., Campitelli, L. F., Das, P. K., Yin, Y., Albu, M., et al. (2018). The human transcription factors. *Cell* 172, 650–665. doi: 10.1016/j.cell.2018.01.029
- Langergraber, K. E., Prüfer, K., Rowney, C., Boesch, C., Crockford, C., Fawcett, K., et al. (2012). Generation times in wild chimpanzees and gorillas suggest earlier divergence times in great ape and human evolution. *Proc. Natl. Acad. Sci. U S A.* 109, 15716–15721. doi: 10.1073/pnas.1211740109
- Lappalainen, T., Sammeth, M., Friedländer, M. R., 't Hoen, P. A. C., Monlong, J., Rivas, M. A., et al. (2013). Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* 501, 506–511. doi: 10.1038/nature12531
- Latchman, D. S. (1997). Transcription factors: an overview. *Int. J. Biochem. Cell B* 29, 1305–1312. doi: 10.1016/S1357-2725(97)00085-X
- Lewontin, R. C. (1974). *The genetic basis of evolutionary change*. New York: Columbia University Press.
- Li, M., Hada, A., Sen, P., Olufemi, L., Hall, M. A., Smith, B. Y., et al. (2015). Dynamic regulation of transcription factors by nucleosome remodeling. *eLife* 4:e06249. doi: 10.7554/eLife.06249
- Linnen, C. R., Poh, Y. P., Peterson, B. K., Barrett, R. D., Larson, J. G., Jensen, J. D., et al. (2013). Adaptive evolution of multiple traits through multiple mutations at a single gene. *Science* 339, 1312–1316. doi: 10.1126/science.1233213
- Mackay, T. F., Stone, E. A., and Ayroles, J. F. (2009). The genetics of quantitative traits: challenges and prospects. *Nature Rev. Genet.* 10, 565–577. doi: 10.1038/nrg2612
- Mark, C., Abrink, M., and Hellman, L. (1999). Comparative analysis of KRAB zinc finger proteins in rodents and man: evidence for several evolutionarily distinct subfamilies of KRAB zinc finger genes. *DNA Cell Biol.* 18, 381–396. doi: 10.1089/104454999315277
- Mihola, O., Trachtulec, Z., Vlcek, C., Schimenti, J. C., and Forejt, J. (2009). A mouse speciation gene encodes a meiotic histone H3 methyltransferase. *Science* 323, 373–375. doi: 10.1126/science.1163601
- Mundy, N. I., and Cook, S. (2003). Positive selection during the diversification of class I vomeronasal receptor-like (V1RL) genes, putative pheromone receptor

- genes, in human and primate evolution. *Mol. Biol. Evol.* 20, 1805–1810. doi: 10.1093/molbev/msg192
- Murrell, B., Wertheim, J. O., Moola, S., Weighill, T., Scheffler, K., and Pond, S. L. K. (2012). Detecting individual sites subject to episodic diversifying selection. *PLoS Genet.* 8:e1002764. doi: 10.1371/journal.pgen.1002764
- Narasimhan, V. M., Hunt, K. A., Mason, D., Baker, C. L., Karczewski, K. J., Barnes, M. R., et al. (2016). Health and population effects of rare gene knockouts in adult humans with related parents. *Science* 352, 474–477. doi: 10.1126/science.aac8624
- Nickel, G. C., Tefft, D. L., Goglin, K., and Adams, M. D. (2008). An empirical test for branch-specific positive selection. *Genetics* 179, 2183–2193. doi: 10.1534/genetics.108.090548
- Nielsen, R., Bustamante, C., Clark, A. G., Glanowski, S., Sackton, T. B., Hubisz, M. J., et al. (2005). A scan for positively selected genes in the genomes of humans and chimpanzees. *PLoS Biol.* 3:e170. doi: 10.1371/journal.pbio.0030170
- Nowick, K., and Stubbs, L. (2010). Lineage-specific transcription factors and the evolution of gene regulatory networks. *Brief. Funct. Genomics* 9, 65–78. doi: 10.1093/bfpg/elp056
- Nowick, K., Carneiro, M., and Faria, R. (2013). A prominent role of KRAB-ZNF transcription factors in mammalian speciation? *Trends Genet.* 29, 130–139. doi: 10.1016/j.tig.2012.11.007
- Nowick, K., Fields, C., Gernat, T., Caetano-Anolles, D., Kholina, N., and Stubbs, L. (2011). Gain, loss and divergence in primate zinc-finger genes: a rich resource for evolution of gene regulatory differences between species. *PLoS One* 6:e21553. doi: 10.1371/journal.pone.0021553
- Nowick, K., Gernat, T., Almaas, E., and Stubbs, L. (2009). Differences in human and chimpanzee gene expression patterns define an evolving network of transcription factors in brain. *Proc. Natl. Acad. Sci. U S A.* 106, 22358–22363. doi: 10.1073/pnas.0911376106
- Oliver, P. L., Goodstadt, L., Bayes, J. J., Birtle, Z., Roach, K. C., Phadnis, N., et al. (2009). Accelerated evolution of the Prdm9 speciation gene across diverse metazoan taxa. *PLoS Genet.* 5:e1000753. doi: 10.1371/journal.pgen.1000753
- Online Mendelian Inheritance in Man (2020). *McKusick-Nathans Institute of Genetic Medicine*. Baltimore, MD: Johns Hopkins University.
- Orgogozo, V., Morizot, B., and Martin, A. (2015). The differential view of genotype-phenotype relationships. *Front. Genet.* 6:179. doi: 10.3389/fgene.2015.00179
- Pagès, H., Abouyoun, P., Gentleman, R., and DebRoy, S. (2019). *Biostrings: Efficient manipulation of biological strings. R package version 2.54.0*. vienna: R Core Team.
- Pan, H. X., Bai, H. S., Guo, Y., and Cheng, Z. Y. (2019). Bioinformatic analysis of the prognostic value of ZNF860 in recurrence-free survival and its potential regulative network in gastric cancer. *Eur. Rev. Med. Pharmacol.* 23, 162–170. doi: 10.26355/eurrev_201901_16760
- Papathodorou, I., Moreno, P., Manning, J., Fuentes, A. M. P., George, N., Fexova, S., et al. (2020). Expression Atlas update: from tissues to single cells. *Nucleic Acids Res.* 48, D77–D83. doi: 10.1093/nar/gkz947
- Paradis, E., and Schliep, K. (2019). ape 5.3: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics* 35, 526–528. doi: 10.1093/bioinformatics/bty633
- Parcerisas, A., Rubio, S. E., Muhaisen, A., Gomez-Ramos, A., Pujadas, L., Puiggros, M., et al. (2014). Somatic signature of brain-specific single nucleotide variations in sporadic Alzheimer's disease. *J. Alzheimers Dis.* 42, 1357–1382. doi: 10.3233/JAD-140891
- Park, C. B., Choi, V. N., Jun, J. B., Kim, J. H., Lee, Y., Lee, J., et al. (2018). Identification of a rare homozygous c. 790C>T variation in the TFB2M gene in Korean patients with autism spectrum disorder. *Biochem. Biophys. Res. Commun.* 507, 148–154. doi: 10.1016/j.bbrc.2018.10.194
- Patterson, N., Richter, D. J., Gnerre, S., Lander, E. S., and Reich, D. (2006). Genetic evidence for complex speciation of humans and chimpanzees. *Nature* 441, 1103–1108. doi: 10.1038/nature04789
- Perdomo-Sabogal, Á., and Nowick, K. (2019). Genetic variation in human gene regulatory factors uncovers regulatory roles in local adaptation and disease. *Genome Biol. Evol.* 11, 2178–2193. doi: 10.1093/gbe/evz131
- Perdomo-Sabogal, Á., Kanton, S., Walter, M. B. C., and Nowick, K. (2014). The role of gene regulatory factors in the evolutionary history of humans. *Curr. Opin. Genet. Dev.* 29, 60–67. doi: 10.1016/j.gde.2014.08.007
- Perdomo-Sabogal, Á., Nowick, K., Piccini, I., Sudbrak, R., Lehrach, H., Yaspo, M. L., et al. (2016). Human lineage-specific transcriptional regulation through GA-binding protein transcription factor alpha (GABPA). *Mol. Biol. Evol.* 33, 1231–1244. doi: 10.1093/molbev/msw007
- Prado-Martinez, J., Sudmant, P. H., Kidd, J. M., Li, H., Kelley, J. L., Lorente-Galdos, B., et al. (2013). Great ape genetic diversity and population history. *Nature* 499, 471–475. doi: 10.1038/nature12228
- Prüfer, K., Munch, K., Hellmann, I., Akagi, K., Miller, J. R., Walenz, B., et al. (2012). The bonobo genome compared with the chimpanzee and human genomes. *Nature* 486, 527–531. doi: 10.1038/nature11128
- Ravasi, T., Suzuki, H., Cannistraci, C. V., Katayama, S., Bajic, V. B., Tan, K., et al. (2010). An atlas of combinatorial transcriptional regulation in mouse and man. *Cell* 140, 744–752. doi: 10.1016/j.cell.2010.01.044
- Sabeti, P. C., Varilly, P., Fry, B., Lohmueller, J., Hostetter, E., Cotsapas, C., et al. (2007). Genome-wide detection and characterization of positive selection in human populations. *Nature* 449, 913–918. doi: 10.1038/nature06250
- Samaras, P., Schmidt, T., Frejno, M., Gessulat, S., Reinecke, M., Jarzab, A., et al. (2020). ProteomicsDB: a multi-omics and multi-organism resource for life science research. *Nucleic Acids Res.* 48, D1153–D1163. doi: 10.1093/nar/gkz974
- Scally, A., Yngvadottir, B., Xue, Y., Ayub, Q., Durbin, R., and Tyler-Smith, C. (2013). A genome-wide survey of genetic variation in gorillas using reduced representation sequencing. *PLoS One* 8:e65066. doi: 10.1371/journal.pone.0065066
- Schultz, A. J., and Sackton, T. B. (2019). Immune genes are hotspots of shared positive selection across birds and mammals. *eLife* 8, e41815. doi: 10.7554/eLife.41815
- Schwartz, J. J., Roach, D. J., Thomas, J. H., and Shendure, J. (2014). Primate evolution of the recombination regulator PRDM9. *Nat. Commun.* 5:4370. doi: 10.1038/ncomms5370
- Shao, Y., Chen, C., Shen, H., He, B. Z., Yu, D., Jiang, S., et al. (2019). GenTree, an integrated resource for analyzing the evolution and function of primate-specific coding genes. *Genome Res.* 29, 682–696. doi: 10.1101/gr.238733.118
- Shelest, E. (2017). Transcription factors in fungi: TFome dynamics, three major families, and dual-specificity TFs. *Front. Genet.* 8:53. doi: 10.3389/fgene.2017.00053
- Siepel, A., and Arbiza, L. (2014). Cis-regulatory elements and human evolution. *Curr. Opin. Genet. Dev.* 29, 81–89. doi: 10.1016/j.gde.2014.08.011
- Slodkowitz, G., and Goldman, N. (2020). Integrated structural and evolutionary analysis reveals common mechanisms underlying adaptive evolution in mammals. *Proc. Natl. Acad. Sci. U S A.* 117, 5977–5986. doi: 10.1073/pnas.1916786117
- Smith, M. D., Wertheim, J. O., Weaver, S., Murrell, B., Scheffler, K., and Kosakovsky Pond, S. L. (2015). Less is more: an adaptive branch-site random effects model for efficient detection of episodic diversifying selection. *Mol. Biol. Evol.* 32, 1342–1353. doi: 10.1093/molbev/msv022
- Stein, M. B., Chen, C. Y., Ursano, R. J., Cai, T., Gelernter, J., Heeringa, S. G., et al. (2016). Genome-wide association studies of posttraumatic stress disorder in 2 cohorts of US Army soldiers. *JAMA Psychiat.* 73, 695–704. doi: 10.1001/jamapsychiatry.2016.0350
- Steinberger, M., and Söding, J. (2017). MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat. Biotechnol.* 35, 1026–1028. doi: 10.1038/nbt.3988
- Stern, D. L. (2000). Evolutionary developmental biology and the problem of variation. *Evolution* 54, 1079–1109. doi: 10.1111/j.0014-3820.2000.tb00544.x
- Su, Z., Zhang, J., Kumar, C., Molony, C., Lu, H., Chen, R., et al. (2016). Species specific exome probes reveal new insights in positively selected genes in nonhuman primates. *Sci. Rep.* 6:33876. doi: 10.1038/srep33876
- Suyama, M., Torrents, D., and Bork, P. (2006). PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Res.* 34, W609–W612. doi: 10.1093/nar/gkl315
- Szklarczyk, D., Gable, A. L., Lyon, D., Junge, A., Wyder, S., Huerta-Cepas, J., et al. (2019). STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.* 47, D607–D613. doi: 10.1093/nar/gky1131
- Tripathi, S., Christie, K. R., Balakrishnan, R., Huntley, R., Hill, D. P., Thommesen, L., et al. (2013). Gene Ontology annotation of sequence-specific DNA binding transcription factors: setting the stage for a large-scale curation effort. *Database* 2013:bat062. doi: 10.1093/database/bat062

- UniProt Consortium (2019). UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.* 47, D506–D515. doi: 10.1093/nar/gky1049
- Van Der Lee, R., Wiel, L., Van Dam, T. J., and Huynen, M. A. (2017). Genome-scale detection of positive selection in nine primates predicts human-virus evolutionary conflicts. *Nucleic Acids Res.* 45, 10634–10648. doi: 10.1093/nar/gkx704
- Varki, A., Geschwind, D. H., and Eichler, E. E. (2008). Human uniqueness: genome interactions with environment, behaviour and culture. *Nat. Rev. Genet.* 9, 749–763. doi: 10.1038/nrg2428
- Vilella, A. J., Severin, J., Ureta-Vidal, A., Heng, L., Durbin, R., and Birney, E. (2009). EnsemblCompara GeneTrees: complete, duplication-aware phylogenetic trees in vertebrates. *Genome Res.* 19, 327–335. doi: 10.1101/gr.073585.107
- Voight, B. F., Kudaravalli, S., Wen, X., and Pritchard, J. K. (2006). A map of recent positive selection in the human genome. *PLoS Biol.* 4:e72. doi: 10.1371/journal.pbio.0040072
- Wagner, G. P., and Lynch, V. J. (2008). The gene regulatory logic of transcription factor evolution. *Trends Ecol. Evol.* 23, 377–385. doi: 10.1016/j.tree.2008.03.006
- Wagner, G. P., and Zhang, J. (2011). The pleiotropic structure of the genotype–phenotype map: the evolvability of complex organisms. *Nat. Rev. Genet.* 12, 204–213. doi: 10.1038/nrg2949
- Wallace, D. C. (2005). A mitochondrial paradigm of metabolic and degenerative diseases, aging, and cancer: a dawn for evolutionary medicine. *Annu. Rev. Genet.* 39, 359–407. doi: 10.1146/annurev.genet.39.110304.095751
- Warren, W. C., Harris, R. A., Haukness, M., Fiddes, I. T., Murali, S. C., Fernandes, J., et al. (2020). Sequence diversity analyses of an improved rhesus macaque genome enhance its biomedical utility. *Science*, 370:eabc6617. doi: 10.1126/science.abc6617
- Wilson, D., Charoensawan, V., Kummerfeld, S. K., and Teichmann, S. A. (2008). DBD - taxonomically broad transcription factor predictions: new content and functionality. *Nucleic Acids Res.* 36, D88–D92. doi: 10.1093/nar/gkm964
- Wingender, E., Schoeps, T., and Dönitz, J. (2013). TFClass: an expandable hierarchical classification of human transcription factors. *Nucleic Acids Res.* 41, D165–D170. doi: 10.1093/nar/gks1123
- Wingender, E., Schoeps, T., Haubrock, M., and Dönitz, J. (2015). TFClass: a classification of human transcription factors and their rodent orthologs. *Nucleic Acids Res.* 43, D97–D102. doi: 10.1093/nar/gku1064
- Winter, D. J. (2017). rentrez: an R package for the NCBI eUtils API. *R J.* 9, 520–526. doi: 10.32614/RJ-2017-058
- Wittkopp, P. J., and Kalay, G. (2012). Cis-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence. *Nat. Rev. Genet.* 13, 59–69. doi: 10.1038/nrg3095
- Wray, G. A. (2007). The evolutionary significance of cis-regulatory mutations. *Nat. Rev. Genet.* 8, 206–216. doi: 10.1038/nrg2063
- Yamada, Y., Kato, K., Oguri, M., Horibe, H., Fujimaki, T., Yasukochi, Y., et al. (2018). Identification of four genes as novel susceptibility loci for early-onset type 2 diabetes mellitus, metabolic syndrome, or hyperuricemia. *Biomed. Rep.* 9, 21–36. doi: 10.3892/br.2018.1105
- Yang, P., Wang, Y., and Macfarlan, T. S. (2017). The role of KRAB-ZFPs in transposable element repression and mammalian evolution. *Trends Genet.* 33, 871–881. doi: 10.1016/j.tig.2017.08.006
- Yang, Z. (2007). PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* 24, 1586–1591. doi: 10.1093/molbev/msm088
- Yang, Z., and Nielsen, R. (2002). Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. *Mol. Biol. Evol.* 19, 908–917. doi: 10.1093/oxfordjournals.molbev.a004148
- Yates, A. D., Achuthan, P., Akanni, W., Allen, J., Allen, J., Alvarez-Jarreta, J., et al. (2020). Ensembl 2020. *Nucleic Acids Res.* 48, D682–D688. doi: 10.1093/nar/gkz966
- Zhang, L., and Li, W.-H. (2005). Human SNPs reveal no evidence of frequent positive selection. *Mol. Biol. Evol.* 22, 2504–2507. doi: 10.1093/molbev/msi240
- Zhao, X. W., and Kishino, H. (2020). Multiple Isolated Transcription Factors Act as Switches and Contribute to Species Uniqueness. *Genes* 11:1148. doi: 10.3390/genes11101148
- Zhu, J., Fu, H., Wu, Y., and Zheng, X. (2013). Function of lncRNAs and approaches to lncRNA-protein interactions. *Sci. China Life Sci.* 56, 876–885.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2021 Jovanovic, Sarfert, Reyna-Blanco, Indrischek, Valdivia, Shelest and Nowick. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



Gene Expression Modification by an Autosomal Inversion Associated With Three Male Mating Morphs

Jasmine L. Loveland^{1*}, David B. Lank² and Clemens Küpper¹

¹ Research Group for Behavioural Genetics and Evolutionary Ecology, Max Planck Institute for Ornithology, Seewiesen, Germany, ² Department of Biological Sciences, Simon Fraser University, Burnaby, BC, Canada

OPEN ACCESS

Edited by:

Katja Nowick,
Freie Universität Berlin, Germany

Reviewed by:

Sara Lipshutz,
Indiana University Bloomington,
United States
Rui Faria,
Centro de Investigacao em
Biodiversidade e Recursos Geneticos
(CIBIO-InBIO), Portugal

*Correspondence:

Jasmine L. Loveland
jloveland@orn.mpg.de

Specialty section:

This article was submitted to
Systems Biology,
a section of the journal
Frontiers in Genetics

Received: 14 December 2020

Accepted: 22 April 2021

Published: 04 June 2021

Citation:

Loveland JL, Lank DB and
Küpper C (2021) Gene Expression
Modification by an Autosomal
Inversion Associated With Three Male
Mating Morphs.
Front. Genet. 12:641620.
doi: 10.3389/fgene.2021.641620

Chromosomal inversions are structural rearrangements that frequently provide genomic substrate for phenotypic diversity. In the ruff *Philomachus pugnax*, three distinct male reproductive morphs (Independents, Satellites and Faeders) are genetically determined by a 4.5 Mb autosomal inversion. Here we test how this stable inversion polymorphism affects gene expression in males during the lekking season. Gene expression may be altered through disruptions at the breakpoints and the accumulation of mutations due to suppressed recombination. We used quantitative PCR to measure expression of 11 candidate inversion genes across three different tissues (liver, adrenal glands and gonads) and tested for allelic imbalance in four inversion genes across 12 males of all three morphs (8 Independents, 2 Satellites, 2 Faeders). We quantified transcripts of *CENPN*, an essential gene disrupted by the inversion at the proximal breakpoint, at different exons distributed near and across the breakpoint region. Consistent with dosage dependent gene expression for the breakpoint gene *CENPN*, we found that expression in Independents was broadly similar for transcripts segments from inside and outside the inversion regions, whereas for Satellites and Faeders, transcript segments outside of the inversion showed at least twofold higher expression than those spanning over the breakpoint. Within the inversion, observed expression differences for inversion males across all four genes with allele-specific primers were consistent with allelic imbalance. We further analyzed gonadal expression of two inversion genes, *HSD17B2* and *SDR42E1*, along with 12 non-inversion genes related to steroid metabolism and signaling in 25 males (13 Independents, 7 Satellites, 5 Faeders). Although we did not find clear morph differentiation for many individual genes, all three morphs could be separated based on gene expression differences when using linear discriminant analysis (LDA), regardless of genomic location (i.e., inside or outside of the inversion). This was robust to the removal of genes with the highest loadings. Pairwise correlations in the expression of genes showed significant correlations for 9–18 pairs of genes within morphs. However, between morphs, we only found a single association between genes *SDR42E1* and *AROM* for Independents and Satellites. Our results suggest complex and wide-ranging changes in gene expression caused by structural variants.

Keywords: chromosomal inversion, alternative reproduction strategies, steroidogenic pathway, *Philomachus pugnax*, *SDR42E1*, aromatase, *HSD17B2*, *CENPN*

INTRODUCTION

Chromosomal inversions are genomic rearrangements that occur in animals and plants and are associated with local adaptation and speciation (Kirkpatrick, 2010; Dagilis and Kirkpatrick, 2016; Wellenreuther and Bernatchez, 2018; Faria et al., 2019). Inversion polymorphisms, which can be maintained by balancing selection (Kirkpatrick, 2010; Wellenreuther and Bernatchez, 2018), frequently provide the genetic basis for morphological and behavioral diversity (Fuller et al., 2016; Küpper et al., 2016; Lamichhaney et al., 2016; Lindtke et al., 2017; Huang et al., 2018). Changes in gene expression provide one possible mechanism for chromosomal inversions to increase phenotypic variation (Naseeb et al., 2016; Said et al., 2018). Quantifying gene expression differences between inversion alleles is therefore an important step toward characterizing the molecular mechanisms that lead from genotypic to phenotypic variation.

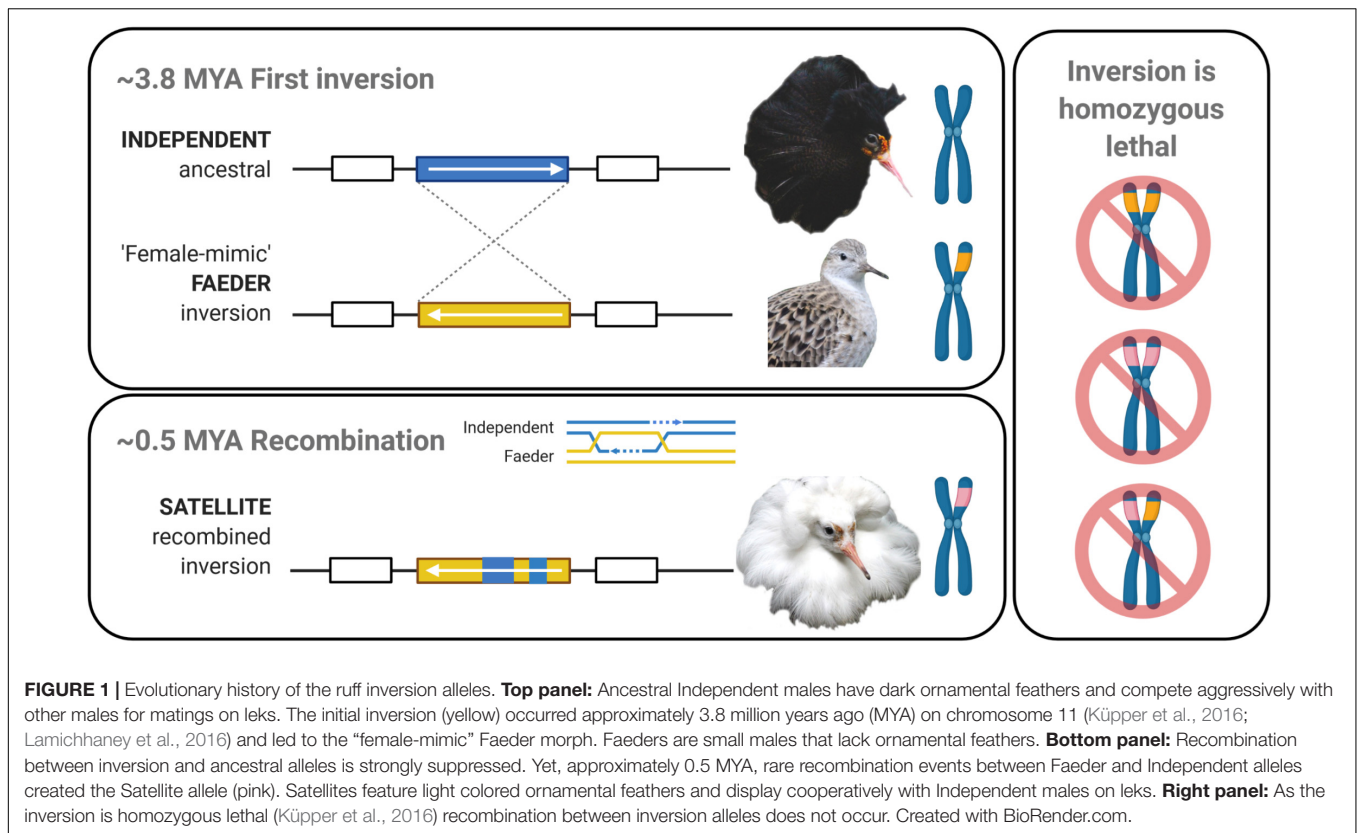
A signature feature of inversions is that they are subject to suppressed recombination, which leads to high genetic differentiation between inverted and non-inverted alleles (Kirkpatrick, 2010). While sequence divergence in regulatory and coding regions alike, requires time for mutations to appear, the new chromosomal rearrangement has the potential to immediately alter gene expression. Consequences of an inversion thus include the loss of a full transcript if a gene is located at a breakpoint, and the disruption of cis-regulatory elements that control transcription of genes near a breakpoint (Huang et al., 2018; Faria et al., 2019; Yan et al., 2020). For example, approximately half of haemophilia A cases are explained by the impact of an inversion breakpoint that disrupts a coagulation factor gene (Lakich et al., 1993; Antonarakis et al., 1995; Brinke et al., 1996; Cumming, 2004). Allelic imbalance, a change in expression levels between non-inverted and inverted haplotypes, can also be observed for genes in inversion regions (Sun et al., 2018). This imbalance is caused by mutations in regulatory regions. Inversion alleles may then either become over- or underexpressed in heterozygotes. In white-throated sparrows (*Zonotrichia albicollis*), a systematic study on neural tissue revealed that reduced expression of the inversion allele can lead to dosage compensation with the non-inversion allele becoming more expressed (Sun et al., 2018). As a result, in morphs that are heterozygous for the inversion, many genes may show allelic imbalance although total gene expression of these loci is not necessarily altered in comparison to the ancestral morph.

An inherent difficulty in pin-pointing causal associations between an inversion and its associated traits is that variation in gene expression can be restricted, in any number of combinations, to a specific time of the year, sex, tissue type and life stage, to name a few (Imsland et al., 2012; Fuller et al., 2016; Wang et al., 2017; Newhouse et al., 2019). For example, in the Rose-comb chicken (*Gallus gallus*), the comb phenotype is due to ectopic expression of an inversion gene (*MNR2*) in comb tissue during a narrow window of embryonic development (Imsland et al., 2012; Wang et al., 2017). In addition, low fertility and poor sperm motility Rose-comb traits are associated with truncated transcripts and increased testicular expression of an inversion gene (*CCDC108*), most pronounced in birds homozygous for

the R1 inversion allele (Imsland et al., 2012). Interestingly, both *MNR2* and *CCDC108* are located near the distal breakpoint of the inversion (Wang et al., 2017) and their phenotypic consequences, in both R1 and R2 inversion alleles can be traced to exon shuffling near breakpoints (Imsland et al., 2012).

Through suppressed recombination, inversions often form supergenes (Schwander et al., 2014), where certain allele combinations are fixed and evolve together. In some prominent examples of adaptive behavioral diversity, these supergenes have captured multiple loci involved in hormone signaling and metabolism (Merritt et al., 2018, 2020; Horton et al., 2020). Steroid hormones often have widespread pleiotropic effects and modulate gene expression genome-wide across many different tissues (Beato, 1993). In addition, steroids such as testosterone and estrogen are extremely important for determining the sexual differentiation of the brain (Steimer and Hutchison, 1980; Balthazart and Ball, 1995; Court et al., 2020). These two hormones can have long-lasting organizational effects on brain development, which is an effective way to exert an entire morph-specific neural circuit by virtue of the delicate relationship between genetic sex, hormonal environment and genetic background.

Here we investigate the impact of an autosomal inversion on gene expression in the ruff (*Philomachus pugnax*). This lekking shorebird is characterized by a stable polymorphism that includes three phenotypically and genetically distinct male mating morphs: Independents, Satellites, and Faeders (Lank et al., 2013; **Figure 1**). Male morph type is determined by a 4.5 Mb inversion located on chromosome 11 that contains no more than 125 genes, with Satellites and Faeders carrying distinct dominant inversion alleles (Küpper et al., 2016; Lamichhaney et al., 2016). Inversion homozygotes are non-viable (**Figure 1**), thus, all Satellites and Faeders are inversion heterozygotes (i.e., carry one ancestral and one inversion haplotype). During the breeding season, these three morphs showcase alternative reproductive tactics (ARTs) with discrete differences in aggression and courtship displays (Jukema and Piersma, 2006; Küpper et al., 2016). Independent males with ornamental plumage, most often predominantly dark in color, establish courts on leks and are highly aggressive. Independent males form temporary alliances with Satellite males, a morph with lightly colored ornamented feathers, and together they engage in semi-cooperative courtship displays. In contrast, the third morph, female-mimicking Faeder males, do not display any male-typical ornamentation or courtship behaviors (Jukema and Piersma, 2006). Hormonally, inversion morphs consistently have the lowest levels of circulating testosterone, but also higher levels of androstenedione (a testosterone precursor and metabolite), compared to Independent males (Küpper et al., 2016; Loveland et al., 2021). Whether and how higher levels of androstenedione in inversion morphs could produce physiological, morphological or behavioral differences compared to Independents is not known, in part because androstenedione is only a weak activator of the androgen receptor. Given the important role of sex hormones in priming and maintaining seasonal social behavior, the behavioral differences among morphs are presumed to be connected to their respective androgenic differences (Küpper et al., 2016;



Loveland et al., 2021). In a recent study, we found that the ability to synthesize testosterone in inversion morphs is severely impaired: stimulation of the hypothalamic–pituitary–gonadal axis with gonadotropin-releasing hormone (GnRH) induced a robust increase in androstenedione, but only a subdued increase in testosterone levels, compared to Independent males (Loveland et al., 2021).

The morphological and behavioral differences between Satellites and Faeders, both inversion carriers, must derive from their particular inversion haplotypes. The Faeder haplotype evolved first 3.8 million years ago (MYA) and the Satellite haplotype appeared just 0.5 MYA, following a rare recombination between the Faeder inversion and one or several Independent alleles (Lamichhaney et al., 2016; **Figure 1**). Consistent with their respective evolutionary history and current sequence similarities, Satellites are in some ways behaviorally and phenotypically intermediate to Independents and Faeders. It follows that divergent regions of the inversion should contain genes or regulatory elements that help explain why certain traits are shared by only two morphs (i.e., Independents and Satellites or Faeders and Satellites). For example, only Satellites and Independents engage in courtship in the lekking competition, but both Satellites and Faeders have a similar androgenic profile and response range, relatively larger testes sizes, and show an adaptive decrease in aggression. Yet, in other aspects, such as body size, pituitary progesterone receptor and gonadal *STAR* gene expression, however, Satellites occupy an intermediate position between Independents and Faeders

(Jukema and Piersma, 2006; Küpper et al., 2016; Loveland et al., 2021). Consequently, Satellites provide a unique opportunity to test how recent recombination has affected gene expression variation between morphs. Furthermore, the genetic basis for androgenic differences between inversion morphs has been suggested to derive from expression differences in the *HSD17B2* (hydroxysteroid 17-beta dehydrogenase 2) gene (Küpper et al., 2016) precisely because the enzymatic function of this protein is to convert testosterone to androstenedione (Küpper et al., 2016; Loveland et al., 2021).

In this study, we compare gene expression among ruff male morphs. We sampled ruffs during the breeding season, as this is the time when morph differences in behavior, hormone profiles and appearance are most pronounced. We focused on genes associated with hormone synthesis and/or signaling, located both inside and outside of the inversion, and assayed three tissue types: gonads, adrenal glands and liver. We chose gonads and adrenals because they are major sources for the synthesis of steroid hormones and their precursors, and are involved in the regulation of aggression during breeding and non-breeding contexts (Heimovics et al., 2018). We selected the liver because it is responsible for steroid metabolic processing and can provide information on peripheral mechanisms to offset high circulating testosterone (Bentz et al., 2019). In addition, the liver has proven useful to detect the emergence of gene expression differences in sexual dimorphism during ontogeny (Cox et al., 2017; Cox, 2020).

First, we asked whether genes from the inversion that were previously suggested as candidates for morph-specific traits and

fitness (Küpper et al., 2016) differ in expression between inversion morphs and Independent males across tissues, and we tested for allelic imbalance in a subset of these genes in the heterozygous inversion morphs. We hypothesized that the *CENPN* gene, which is interrupted at the proximate breakpoint in inversion carriers, would show reduced expression from the inversion allele relative to the ancestral allele. The *CENPN* gene is the major candidate gene for homozygous lethality of the ruff inversion because it forms part of the constitutive centromere-associated network (CCAN), a protein complex crucial for mitotic centromere assembly (Cheeseman and Desai, 2008; Carroll et al., 2009; Küpper et al., 2016; Yan et al., 2019).

We further predicted that the *HSD17B2* gene, the major inversion-based candidate gene to explain observed morph differences in circulating testosterone and androstenedione, would have higher gonadal expression in inversion morphs because it is the sole enzyme responsible for this enzymatic step. Second, we investigated whether gonadal expression of two key inversion genes, *HSD17B2* and *SDR42E1* (short chain dehydrogenase/reductase family 42E, member 1), in the context of a dozen non-inversion genes involved in steroid synthesis and signaling, could point to specific steps or genes in testosterone synthesis, or steroid hormones at large, that may be atypical in inversion morphs. We selected *SDR42E1* for several reasons. First, inversion morphs share three major deletions surrounding *HSD17B2* and *SDR42E1* genes, which makes a case for gene expression changes due to the elimination or disruption of cis-regulatory elements more likely (Küpper et al., 2016; Lamichhaney et al., 2016). One of these deletions is 5.2 kb in size and located in between *SDR42E1* and *HSD17B2*, genes that are transcribed in opposing directions.

Second, *SDR42E1* belongs to the so-called “extended” family of short-chain dehydrogenases/reductases (SDRs) superfamily of enzymes (Persson et al., 2009) and its proposed functions are binding, oxidoreductase activity and 3-beta-hydroxy-delta5-steroid dehydrogenase activity (NCBI AceView database; Thierry-Mieg and Thierry-Mieg, 2006). This last proposed function is ascribed because it contains typical HSD3 β domains that could confer enzymatic activity that is partially or fully redundant with other HSD3 β enzymes, such as HSD3B2. HSD3B2 is a major enzyme involved in the biosynthesis of all steroid hormones, including metabolizing pregnenolone to progesterone (Albalat et al., 2011), as well as DHEA (dehydroepiandrosterone) to androstenedione. Thus, previously *SDR42E1* was speculated to regulate progesterone synthesis (Lamichhaney et al., 2016) and by extension, possibly also androstenedione synthesis. Intriguingly, Faeders have reduced expression of the progesterone receptor gene in the pituitary compared to Independents (Loveland et al., 2021) meaning that differential *SDR42E1* expression could alert toward a possible role for progesterone in explaining morph differences.

Third, with regards to allelic expression, we predicted that due to greater sequence similarity between Independents and Satellites in recently recombined regions, genes in these areas (*ZFPM1*, *ZDHHC7*, *ZNF469*) would be similarly expressed in Independents and Satellites, but differently in Faeders. For example, Faeders have 12 and 49 non-synonymous mutations in

the *ZFPM1* and *ZNF469* genes, respectively, that are unique to the Faeder haplotype (Küpper et al., 2016). In contrast, for the inversion gene *SPATA2L*, which is located in an area that has high genetic differentiation between Independents and both inversion morphs, we predicted similar expression in Satellites and Faeders versus Independents.

The analysis of inversion genes' expression offers opportunities to begin to disentangle the mechanistic relationship between function in hormone producing tissues and their phenotypic consequences at the organismal level. Genes that contain androgen response elements, which are short palindromic sequences where androgen receptors bind to regulate transcription, are regulated by androgens including testosterone and its metabolites (i.e., dihydrotestosterone). Given the order of magnitude difference in circulating testosterone levels between Independents and inversion morphs (Küpper et al., 2016; Loveland et al., 2021), we expected that testosterone regulated genes should have overall lower expression in inversion morphs. In addition, because co-expression patterns between genes often indicates they are also functionally associated, we expected that correlations between the expression levels of pairs of genes in Independents, whether positive or negative, might be weakened or absent in inversion morphs.

MATERIALS AND METHODS

Birds and Housing

We sampled 35 adult ruff males (17 Independents, 9 Satellites, 9 Faeders) from a captive breeding flock at Simon Fraser University. The mean age \pm standard error (SE) was 5.11 ± 0.4 yrs. This captive population was originally established from eggs collected near Oulu, Finland, in 1985, 1989, and 1990 (Lank et al., 2013). For at least 3 weeks prior to sample collection, males were group-housed in same-sex pens with visual access to females in an outdoor aviary with unrestricted access to food and water and an area to bathe. All housing and procedures (permit #1232B-17) were approved by the Animal Care Committee of Simon Fraser University operating under guidelines from the Canadian Council on Animal Care.

Tissue Collection and Hormone Analysis

We monitored males to select ones that exhibited their respective morph-specific behaviors according to previously established ethograms (van Rhijn, 1973; Lank et al., 1999) and their behavior was video recorded immediately before killing. No pharmacological manipulations of any kind were performed prior to tissue collection. We collected tissue samples during the breeding seasons of 3 years: 2017 (June 11–17), 2018 (June 5–16), 2019 (June 8–13). We sampled birds in the morning between 8:00–12:15, except for two birds in 2019 that were sampled at 13:00 and 15:00, respectively. Sampling details have been described by Loveland et al. (2021). We dissected samples of the right liver lobe, left and right adrenal glands in their entirety, and left and right gonads. All tissues were rinsed briefly in phosphate-buffered saline (PBS) and dried with a paper tissue before preserving in RNAlater (Ambion) according

to manufacturer's instructions. We weighed gonads to calculate gonadosomatic index ($GSI = \text{gonad mass} \times 100/\text{body mass}$) before storing one gonad in RNAlater, and the other was frozen in Neg50 medium (Thermo Fisher Scientific) or in aluminum foil on dry ice; the gonad preserved in RNAlater was used for RNA extractions. These samples were kept at 4°C for 1–3 days and then stored frozen at –20°C until RNA extraction at the Max Planck Institute for Ornithology in Seewiesen, Germany. We measured testosterone levels from plasma, described by Loveland et al. (2021).

RNA Extraction and cDNA Synthesis

We extracted RNA from adrenal gland and liver samples and gonad using the RNeasy Minikit with modifications as described by Loveland et al. (2021). In addition, for adrenal glands, we mixed 200 µl of the lysed homogenate with 400 µl lysis buffer before proceeding with the steps in the standard protocol. This modification to the protocol was performed to prevent possible clogging of the column and to allow preservation of lysed homogenate for future extractions. We measured RNA concentration with a NanoDrop and assessed RNA quality with the Bioanalyzer RNA nanochip (Agilent). Only samples with a 260/280 ratio of ≥ 1.8 and RINs ≥ 5 were used. We excluded four gonad RNA samples from further analysis due to low yield. The average RINs for RNA from gonads, adrenals and liver were 7.4, 8.8, and 7, respectively. For all tissues, we synthesized 1 µg of RNA into cDNA using the iScript cDNA synthesis kit (Bio-Rad) in 20 µl reactions according to manufacturer's instructions. For comparisons of expression of genes located within the inversion in adrenals, gonads and liver, all cDNA synthesis was performed on the same number of samples (8 Independents, 2 Satellites and 2 Faeders) in 2018. In a second experiment performed in 2019, cDNA was freshly synthesized from gonadal RNA (14 Independents, 8 Satellites, 9 Faeders) to analyze the expression of two inversion genes (*HSD17B2* and *SDR42E1*) along with 12 genes associated with sex hormone synthesis and signaling. We always diluted cDNA 10-fold before use as template in qPCR assays. Sample sizes for tissue RNA extractions for data presented in **Figures 2, 3** were 8 Independents 2 Satellites and 2 Faeders; and in **Figures 5, 6** were: 17 Independents, 9 Satellites, and 9 Faeders; see notes above for exclusions prior to cDNA synthesis based on RNA yield.

Primer Design

To examine whether male morphs differ in gene expression across tissues, we selected 11 genes with morph-specific SNPs within the inversion to analyze allele-specific expression across three tissue types (gonads, liver and adrenal glands). Details on *CENPN* gene structure are shown in **Figure 2** whereas the inversion genes we selected are depicted in **Figure 3**. We selected an additional 12 non-inversion genes associated with sex hormone synthesis and signaling to analyze expression in gonads (**Figure 4**). All primers were designed in PrimerBlast using Genbank ruff coding sequences (CDS) as queries; for accession numbers and primer details see **Supplementary Table 1** and Loveland et al. (2021).

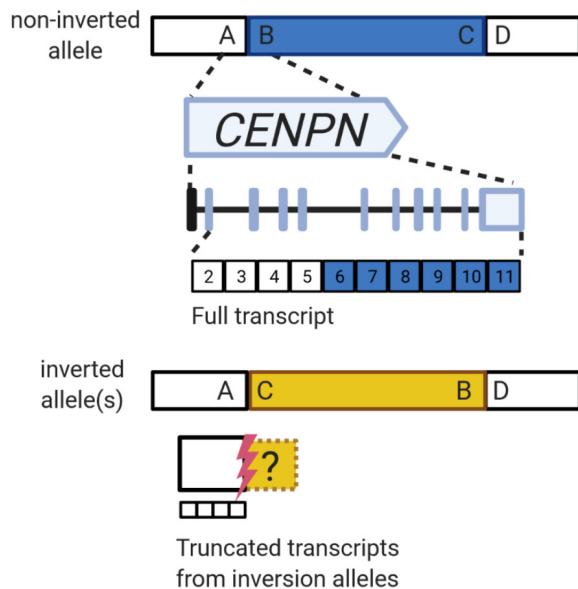
Primer Design for Genes Within the Inversion

We designed allele-specific primers such that the target amplicons spanned at least one exon-exon boundary when possible. We manually annotated exon boundaries by a combination of the following: CDS as queries against the chicken genome (v5.0) with the BLAT tool (Kent, 2002) in the UCSC genome browser (Kent et al., 2002); assessment of Augustus v3.1 gene predictions for chicken orthologs and for ruff (NCBI *Calidris pugnax* Annotation Release 100, software version 6.5). For BLAT results, exon-exon boundaries were called based on their agreement between ruff and chicken. In case of discrepancies between the predicted site of a boundary, we applied the chicken annotation and designed the primer pair to not anneal directly across the boundary, since the exact location was uncertain. For each gene that had morph-specific SNPs, we designed one primer ("common primer") to anneal to ancestral and inversion haplotypes and a second primer to be allele-specific (**Supplementary Table 2**). We extracted sequence data from inversion contigs generated from genome assemblies and mapping data reported in Küpper et al. (2016), which were based on high coverage sequencing ($>80\times$) of two individuals of each morph. We confirmed specificity of allele-specific primer pairs by testing them on liver-derived cDNA from one individual of each morph, by assessing melt curves of amplicons and cycle thresholds. Standard curves were performed on serially diluted cDNA in the range of 1:5–1:100. All allele-specific primers showed amplification efficiency for the target allele in the standard curve above 1.9, but no amplification or low efficiency (<1.3) with unambiguously late amplification in the exponential phase of the common allele when the target allele was absent; an example is shown in **Supplementary Figure 1**. The details for the design for *CENPN* primers that span across the inversion breakpoint are given in **Figure 2A**. For five inversion genes (*CENPN*, *BCO1*, *SLC7A5*, *PLCG2*, *TERF2*) that did not involve allele-specific measurements across tissues we assumed an amplification efficiency of two for all morphs; for all other genes we used gene efficiencies from standard curves.

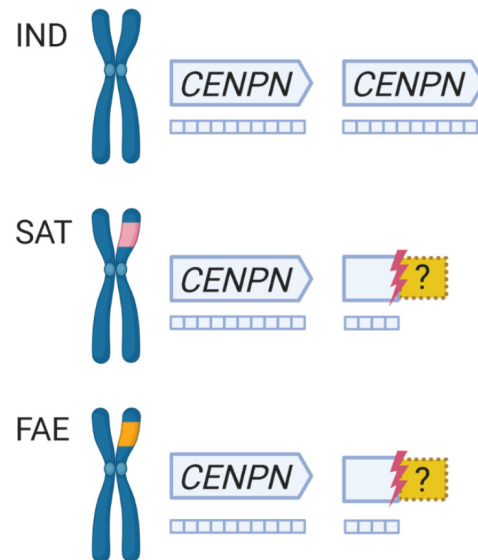
qPCR Conditions

We performed qPCR experiments with SsoAdvanced Universal SYBR Green Super mix (Bio-Rad) as described in Loveland et al. (2021) with the modification that for inversion genes and allele-specific assays, the qPCR was run for 40 cycles. A report of efficiencies for all primer pairs corresponding to non-inversion genes not published in Loveland et al. (2021) is given in **Supplementary Table 3**. For all experiments, we ran each sample in duplicate and used the following conditions: 10 µl reaction with 1X SYBR mix, 400 nM of each primer and 7 ng of cDNA template. To control for interplate effects, each plate contained samples of Independent, Satellite and Faeder males and both reference and target genes were assayed on the same plate. To test for any possible genomic carryover from RNA extraction to cDNA synthesis, we ran the qPCR with cDNA synthesis negative controls (i.e., no reverse transcriptase) for a subset of samples (24 of 36 samples in 2018); none showed any amplification.

A *CENPN* transcripts by allele



B Expected transcripts by morph



C *CENPN* expression

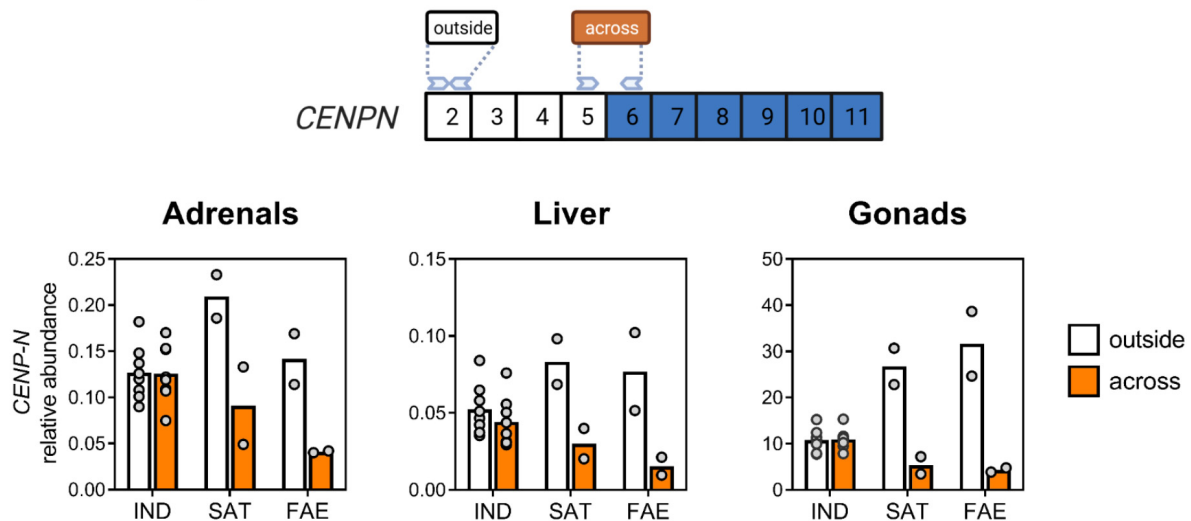


FIGURE 2 | The inversion breakpoint gene *CENPN* has reduced expression in inversion morphs, across tissues. **(A)** Schematic of the order of DNA segments (A-D) on the ancestral allele. The B-C segment (blue) corresponds to the 4.5 Mb inversion region. Segment A flanks the proximal breakpoint, and D flanks the distal breakpoint. The *CENPN* gene spans across the proximal breakpoint and the full transcript contains 11 exons, with the first exon being a non-coding UTR (black). Exons are colored based on their position relative to the corresponding proximal breakpoint: white for outside and blue for inside the corresponding region that becomes inverted. In the inversion allele(s) the order of DNA segments is changed to A, C, B, D with the inversion segment C-B (yellow) reversed from its ancestral orientation. The proximal breakpoint (red bolt) interrupts the *CENPN* gene between exons 5 and 6, thus a truncated four-exon *CENPN* transcript is expected from inversion alleles. Uncertainty about where transcription stops continuing into the inversion is denoted by the dashed yellow box. Exon structure redrawn from NCBI *Calidris pugnax* Genome Viewer. **(B)** Independents produce full *CENPN* transcripts from two ancestral alleles, whereas inversion morphs, which are always heterozygotes for the inversion, are predicted to produce full transcripts from one ancestral allele and only truncated transcripts from their respective inversion allele. **(C)** *CENPN* expression in adrenal glands, liver and gonads for all three morphs plotted as abundance relative to two reference genes (see section “Materials and Methods” for details). In Independent males, expression of exon 2 (white), which is outside of the inverted region, and exons 5 and 6 (orange), which span across the breakpoint, are near identical, whereas in Satellites and Faeders, expression of exons 5 and 6 is much lower than that of exon 2. The amplicon containing exons 5 and 6 spans across the breakpoint and is thus only present on the non-inverted allele in inversion morphs. Sample sizes for adrenals and liver were (8 Independents, 2 Faeders and 2 Satellites) and for gonads were (7 Independents, 2 Faeders, and 2 Satellites). Created with BioRender.com.

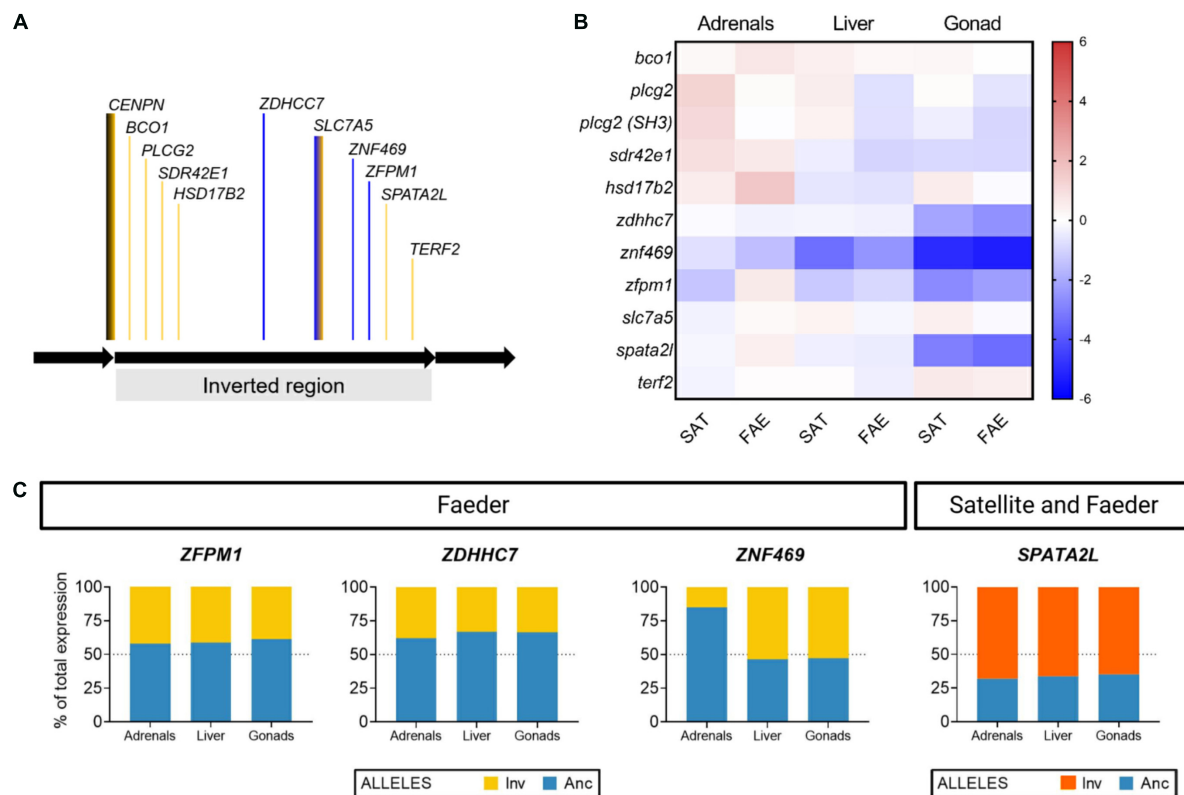


FIGURE 3 | Tissue-specific expression and allelic imbalance of inversion genes. **(A)** Relative location of genes (not drawn to scale) within the inversion that were examined in this study, shown in the orientation of the ancestral non-inverted allele. The line color indicates whether a given gene is in a region with high genetic differentiation between Independents and Satellites (yellow) or high genetic differentiation between Satellites and Faeders (blue); therefore, blue regions correspond to recently recombined regions within the Satellite inversion. *CENPN* spans across the proximal breakpoint (black and yellow line) and *SLC7A5* spans across regions from the most ancient inversion as well as recently recombined areas (blue and yellow line). **(B)** Heatmap visualization of the expression of inversion genes in inversion morphs (Satellites $N = 2$, Faeders $N = 2$), color shading scale indicates increased (red) or reduced expression (blue) relative to Independents ($N = 8$). **(C)** Allelic imbalance in *ZFPM1*, *ZDHCC7*, *ZNF469*, and *SPATA2L* genes varies in magnitude, direction and tissue type. *ZFPM1*, *ZDHCC7*, *ZNF469* are located in regions that underwent recombination and created the Satellite allele. For each bird, we calculated the contribution of expression from each allele (inversion and ancestral) to the total expression of each gene, and plot the means for each inversion morph (Satellites, $N = 2$; Faeders, $N = 2$) as percent of total expression. In Faeders, both *ZFPM1* and *ZDHCC7* show reduced expression originating from the inversion allele of similar magnitude in all three tissues. In contrast, *ZNF469* had reduced expression of the inversion allele only in the adrenal glands. *SPATA2L* showed increased expression of the inversion allele in all three tissues of both Satellites and Faeders.

Relative Abundance

We calculated relative abundance as the expression of target genes relative to two reference genes as described in Loveland et al. (2021). For the qPCR experiments involving only inversion genes on cDNA from gonads, liver and adrenals, we used *GAPDH* and *RPL30* as reference genes, whereas for qPCR experiments involving *HSD17B2* and *SDR42E1* with the 12 non-inversion gene set on cDNA from gonads, we used the reference genes *HPRT1* and *RPL32*.

Statistics

The *CENPN* gene is interrupted by the inversion and has been proposed as the main reason why the inversion is homozygous lethal (Küpper et al., 2016). To test whether morphs differed in the relative abundance of *CENPN* transcripts from “outside” (present in all alleles) vs. “across” the proximal inversion breakpoint (present in ancestral allele only), we divided expression of the “across” fragment by the expression of the

“outside” fragment. As both amplicons are present in the full ancestral transcript we expected this ratio to equal “1” in Independents and conversely, “0.5” in inversion morphs due to the “across” amplicon being absent in inversion alleles (Figure 2B). We tested the mean ratio of each group (i.e., Independent and inversion morphs) with a one-sample *t*-test against a hypothesized mean of one, which assumes that only full transcripts from the ancestral allele are being expressed. In addition, we used the “across” amplicon as an estimate of the abundance of non-truncated *CENPN* transcripts, and compared Independent vs. inversion morph means with a Mann-Whitney U-test, given limited sample sizes. To test allelic imbalance for genes *ZFPM1*, *ZDHCC7*, *ZNF469*, *SPATA2L* we asked whether the contribution of expression from the inversion allele to the total levels (i.e., ancestral allele amplicon + inversion-specific amplicon) differed from 50% with a one-sample *t*-test. To correct for multiple testing (4 genes, 3 tissues) we applied a Bonferroni-Dunn correction.

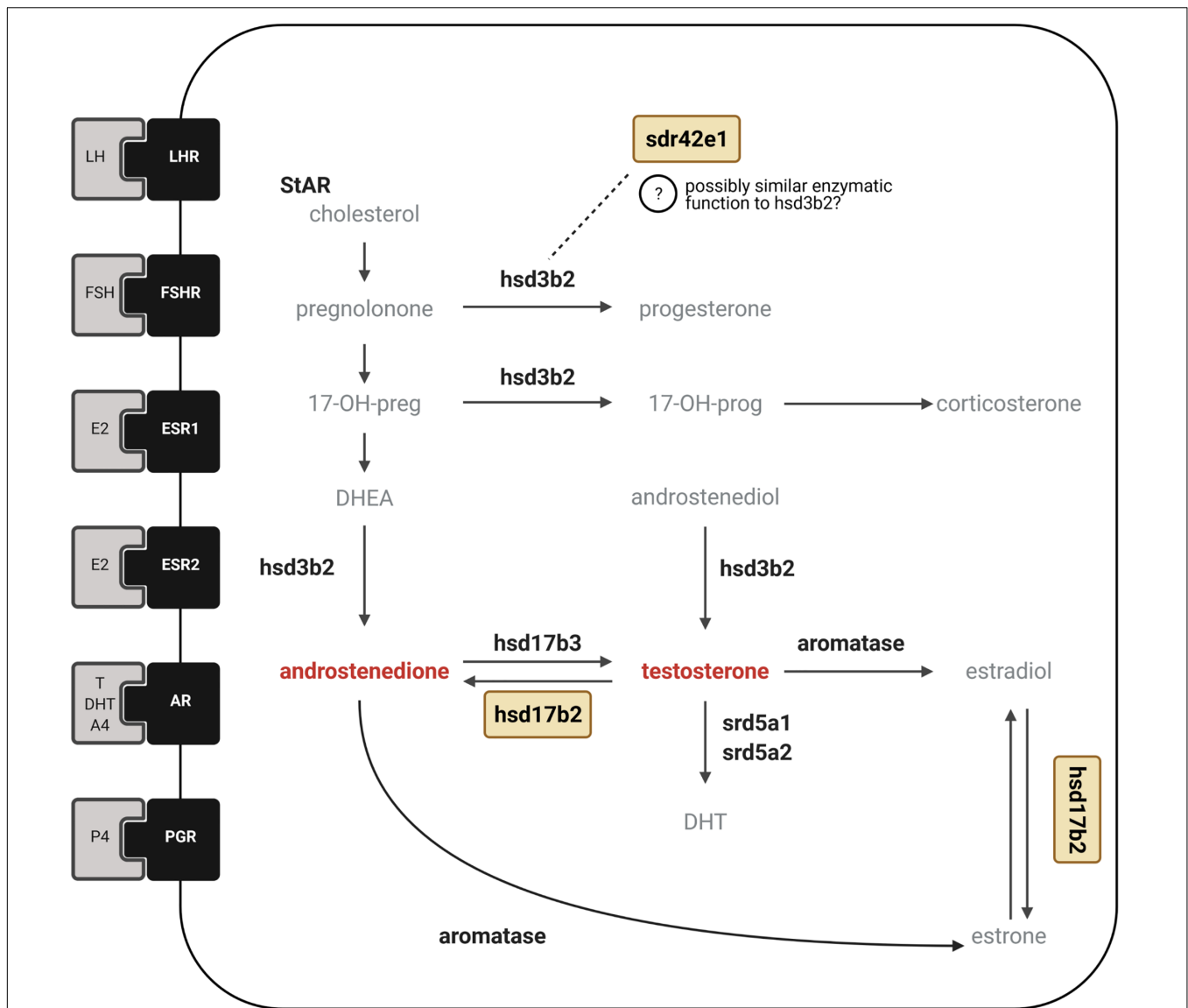


FIGURE 4 | Candidate genes involved in sex hormone synthesis and signaling. Simplified diagram showing the major enzymes involved in pathways for synthesis of steroid hormones with receptors for sex hormones and gonadotropins depicted as membrane-bound. The two inversion genes that encode HSD17B2 and SDR42E1 proteins are shown in yellow boxes. In the ruff, circulating levels of testosterone are drastically reduced in inversion males compared to Independent males. Conversely, inversion morphs have greater levels of androstenedione than Independent males. The *HSD17B2* gene is located in the inversion region and is responsible for the back conversion of testosterone to androstenedione, as well as estradiol to estrone, and has been suggested to be an important candidate that could explain androgenic hormonal differences among ruff morphs (Küpper et al., 2016; Lamichhaney et al., 2016). The *SDR42E1* gene encodes an oxidoreductase enzyme with unknown substrate specificity (indicated with question mark) but has been suggested to have potentially similar enzymatic activity as the HSD3B2 enzyme in progesterone synthesis (Lamichhaney et al., 2016) due to conserved domains characteristic of HSD3 β short-chain dehydrogenase enzymes. We measured the expression of 14 genes associated with sex hormone synthesis and signaling encoding six hormone receptors LHR, FSHR, ESR1, ESR2, AR, PGR; six steroidogenic enzymes shown in bold, HSD3B2, HSD17B2, HSD17B3, aromatase, SRD5A1, SRD5A2. Created with BioRender.com.

To generate the heatmap visualization of the expression of genes from the inversion we calculated for each gene, the mean expression in Independents divided by the expression in an individual inversion morph bird. The means for Satellites and Faeders were then plotted separately on a logarithmic (log2) scale. As, except for *CENPN*, there were no statistically clear expression differences across tissues, we used a heatmap to visualize the results on how expression differed across tissues and morphs.

To collectively analyze the expression of genes associated with sex hormone synthesis and signaling, we performed linear discriminant analysis (LDA) in the R package (MASS) with morph as a preset class, and included the full 14 gene dataset (Figure 5A and Supplementary Figures 3A, 4A). With this analysis we identified axes loadings that explain most of the variation between morphs, such that the separation between morphs is maximized while the variation within morphs is kept

minimal. First, we analyzed two 13 gene datasets, each with one of the two top genes with heaviest loadings removed (*ESR2* or *HSD17B2*) (**Figures 5B,C**) and one 12 gene dataset with both *ESR2* and *HSD17B2* removed (**Figure 5D**). We then proceeded to remove in step-wise fashion the top two genes with highest positive and negative loadings on LD1 and ended with three genes as the last dataset (**Supplementary Figures 3B–L**). This approach showed how robust morph clustering was to the removal of strongly influential genes. In the second approach, we instead removed in a step-wise fashion genes that had the lightest positive and negative loadings to arrive at the minimal dataset that would preserve morph clustering (**Supplementary Figures 4B–H**).

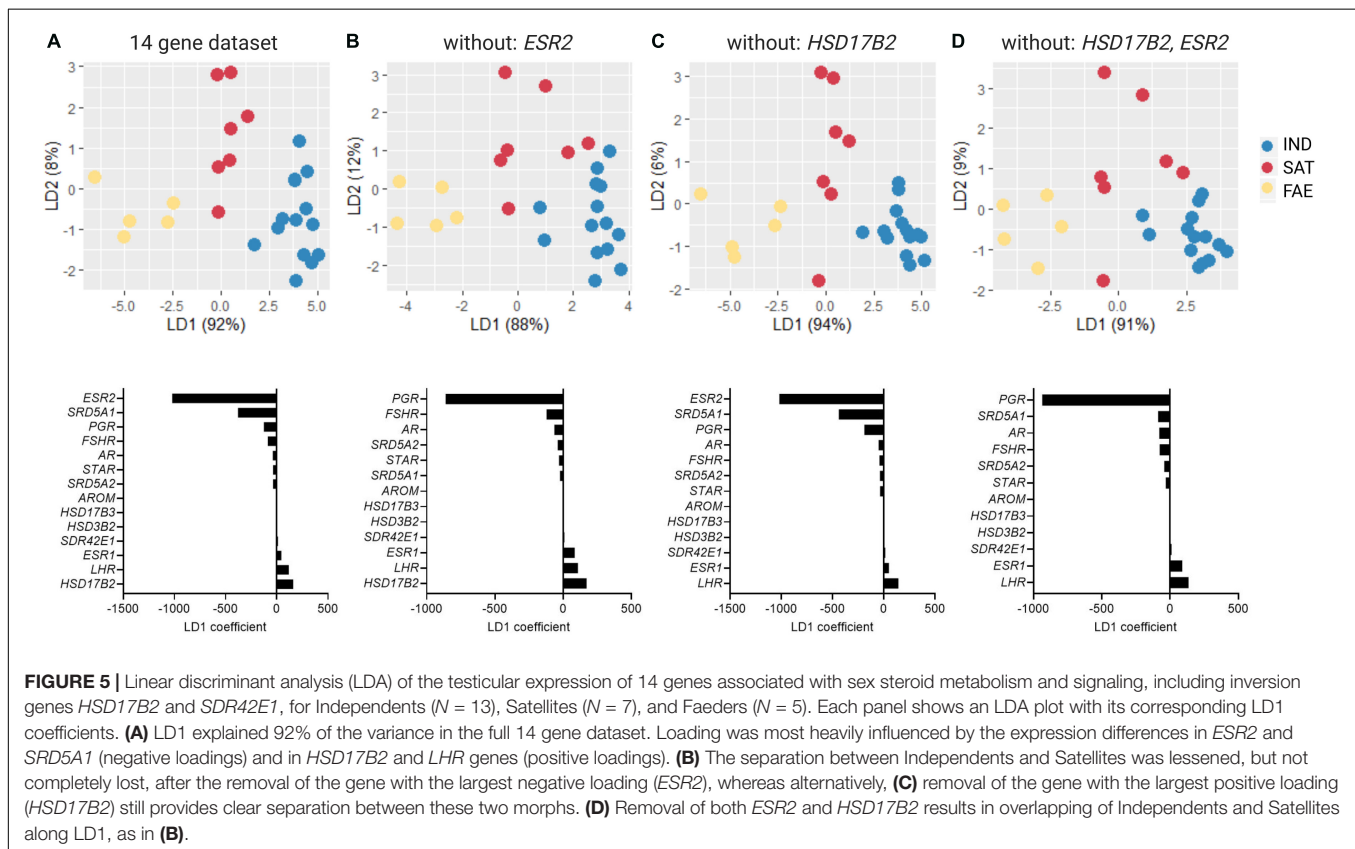
To test for morph-specific correlations of gene expression, we calculated correlation matrices for each morph on the 14 gene dataset (i.e., sex hormone synthesis and signaling genes including two inversion genes *HSD17B2* and *SDR42E1*), plus circulating testosterone and GSI values. We expected this analysis would identify genes with expression associated with testosterone, as activated androgen receptor translocated into the nucleus regulates the expression of genes that contain androgen response elements in their respective regulatory regions (Tan et al., 2015). In addition, we tested for correlations of gene expression with GSI, to explore potential links with spermatogenesis. We predicted that correlations between gene expression and testosterone in Independents would be absent in inversion morphs as these have lower testosterone levels. Similarly, we predicted differences in co-expression between our candidate

genes between Independents and inversion morphs. This analysis rendered a total of 360 correlations. The corresponding *p*-values were adjusted with a False Discovery Rate (FDR) of 10% with the Benjamini and Hochberg method (Benjamini and Hochberg, 1995). We used GraphPad Prism (version 8) and R (version 3.6.2) for statistical analyses.

RESULTS

CENPN Expression

For Independent males the relative abundance of the “across” amplicon of the *CENPN* gene was not different from the relative abundance of the “outside” amplicon in gonads ($t_6 = 0.5$, $p = 0.63$) and adrenal glands ($t_7 = 0.16$, $p = 0.88$), but showed a statistically clear difference in the liver ($t_7 = 3.06$, $p = 0.02$) (**Figure 2C**). In contrast, in inversion morphs, the relative abundance of the “across” amplicon was always lower in all tissues, with gonads ($t_3 = 34.92$, $p < 0.001$), liver ($t_3 = 14.47$, $p < 0.001$), and adrenals ($t_3 = 8.63$, $p = 0.003$) often not even reaching 50% of the “outside” amplicon expression (**Figure 2C**). Using the “across” amplicon as an estimate of levels of non-truncated *CENPN* transcripts, inversion morphs had reduced expression compared to Independents, that was statistically clear in gonads (Mann-Whitney $U = 0$, $p = 0.006$) and liver (Mann-Whitney $U = 4$, $p = 0.048$), but not quite in adrenal glands (Mann-Whitney $U = 5$, $p = 0.07$).



Expression of Inversion Genes and Allelic Imbalance Across Tissues

Overall, there were no clear morph differences in the expression of inversion genes other than *CENPN*, but we note that because of small sample sizes for these particular assays (7–8 Independents, 2 Satellites, 2 Faeders) the lack of clear differences should be viewed with caution. Nonetheless, across the three tissues, the gonads showed the most pronounced decreases in expression of inversion morphs compared to Independents (**Figure 3B**). We found evidence of allelic imbalance (AI) in expression from the inversion allele(s) for *ZFPM1*, *ZDHHC7*, *ZNF469*, and *SPATA2L* genes. The degree of AI appeared to vary in magnitude, direction and by tissue type (**Figure 3C** and **Supplementary Table 4**) although sample size limitations especially for Faeders ($N = 2$) need to be taken into account. Nonetheless, after correcting for multiple testing, increased expression from the inversion allele for *SPATA2L* in adrenals ($t_3 = 9.715$, Bonferroni-Dunn adjusted $p = 0.028$) and liver ($t_3 = 8.51$, Bonferroni-Dunn adjusted $p = 0.041$) were statistically clear for inversion morphs (2 Satellites and 2 Faeders).

Testicular Expression of Inversion Genes *HSD17B2* and *SDR42E1* in the Context of Sex Hormone Synthesis and Signaling Pathways

Consistent with previous reports (Küpper et al., 2016), morphs differed in GSIs [ANOVA $F_{(2, 32)} = 8.44$, $p = 0.001$] with Faeders having clearly higher GSI compared to Satellites (Tukey's $p = 0.019$) and to Independents (Tukey's $p < 0.001$), and no clear difference between Independents and Satellites (Tukey's $p = 0.63$) (**Supplementary Figure 2**). All males had enlarged gonads typical of expected sizes during the breeding season, but there were no statistically clear differences among morphs (**Supplementary Figure 2**). Because the analysis did not allow missing values and we were forced to exclude individuals that did not have the full 14 gene data set, the sample sizes were reduced to 13 Independents, seven Satellites and five Faeders. The combined gene expression of 14 genes associated with sex steroid metabolism and signaling showed clear separation of the three morphs (**Figure 5A**). LD1 explained 92% of the variance and was most heavily loaded positively by the *HSD17B2* and *LHR* genes and negatively by the *ESR2* and *SRD5A1* genes. Only the clustering of Faeders was robust to removal of individual genes (**Supplementary Figure 3**). Removal of genes with the heaviest loadings (*ESR2* and *HSD17B2*) affected this separation when *ESR2* was the only one removed, or when both were removed at the same time (**Figures 5B,D**). However, removal of the inversion gene *HSD17B2* did not affect the separation among the three morphs (**Figure 5C**). Subsequent sequential removal of two genes with the heaviest loadings at a time did not greatly disturb the Faeder cluster (**Supplementary Figure 3**). Only after the removal of the top 11 genes with the heaviest loads, morph separation was not possible. Full results of the sequential removal of genes with heaviest loadings are provided in **Supplementary Figure 3**. In

the second sequential analysis, we removed pairs of genes with the lightest loadings to discover the minimal core number of genes in the dataset that retains morph separation. This showed that at least 10 genes were necessary to retain all three clusters (**Supplementary Figure 4**).

Pairwise correlations in the expression of genes rendered significant correlations for 18, 15 and 9 pairs in Independents, Satellites and Faeders, respectively. After the FDR-adjustment, 8 of these correlations in Independents and 3 in Satellites remained significant; all were positive correlations (asterisks in **Figure 6A** matrices). Of these, only one pair comprising the inversion gene *SDR42E1* and *AROM* was shared between Independents and Satellites (**Figure 6B**). Interestingly, these two genes were also the ones with the highest correlation coefficients within these two morphs (Independents $r = 0.99$, FDR-adjusted $p = 5.36 \times 10^{-9}$; Satellites $r = 0.99$, FDR-adjusted $p = 3.3 \times 10^{-6}$). Because biologically relevant correlations may not reach statistical significance with smaller sample sizes compared to larger ones, and in our study Independents had the largest sample size, we list all pairs of genes that had correlation coefficients ≥ 0.8 (≤ -0.8), regardless of p -values (**Figure 6B**).

DISCUSSION

We used a quantitative PCR approach on candidate genes located inside and outside a prominent autosomal inversion to assess how the stable inversion polymorphism in male ruffs affects gene expression across three different tissues. Two of the tissues, adrenal glands and gonads, are important sources for hormones, and the liver is responsible for further processing steroid metabolites. Our candidate genes were either involved with steroid hormone synthesis and signaling, or major transcription factors chosen because they regulate the expression of many genes, both inside and outside of the inversion. We present evidence for decreased expression of the *CENPN* gene in inversion morphs, as well as allelic imbalance and morph-specific variation in the collective expression of genes associated with hormone synthesis and signaling.

Reduced Expression of the Breakpoint Gene *CENPN*

We first examined the expression of inversion genes across tissues in all three morphs, including the *CENPN* gene, a major candidate for the homozygous lethality of the inversion because it is interrupted by the proximal breakpoint in both inversion morphs (Küpper et al., 2016). Consistent with dosage dependent expression, we found that in Independents, transcript regions from outside and across the breakpoint showed similar levels of expression, whereas in both Satellites and Faeders, the expression of the transcript regions interrupted by the breakpoint was reduced by a factor of two or more. A qualitative assessment of *CENPN* expression across tissues showed that its expression was greatest in gonads for all morphs, compared to adrenals and liver, which is likely due to the proliferative nature of gonadal tissue for spermatogenesis during the breeding season. Among

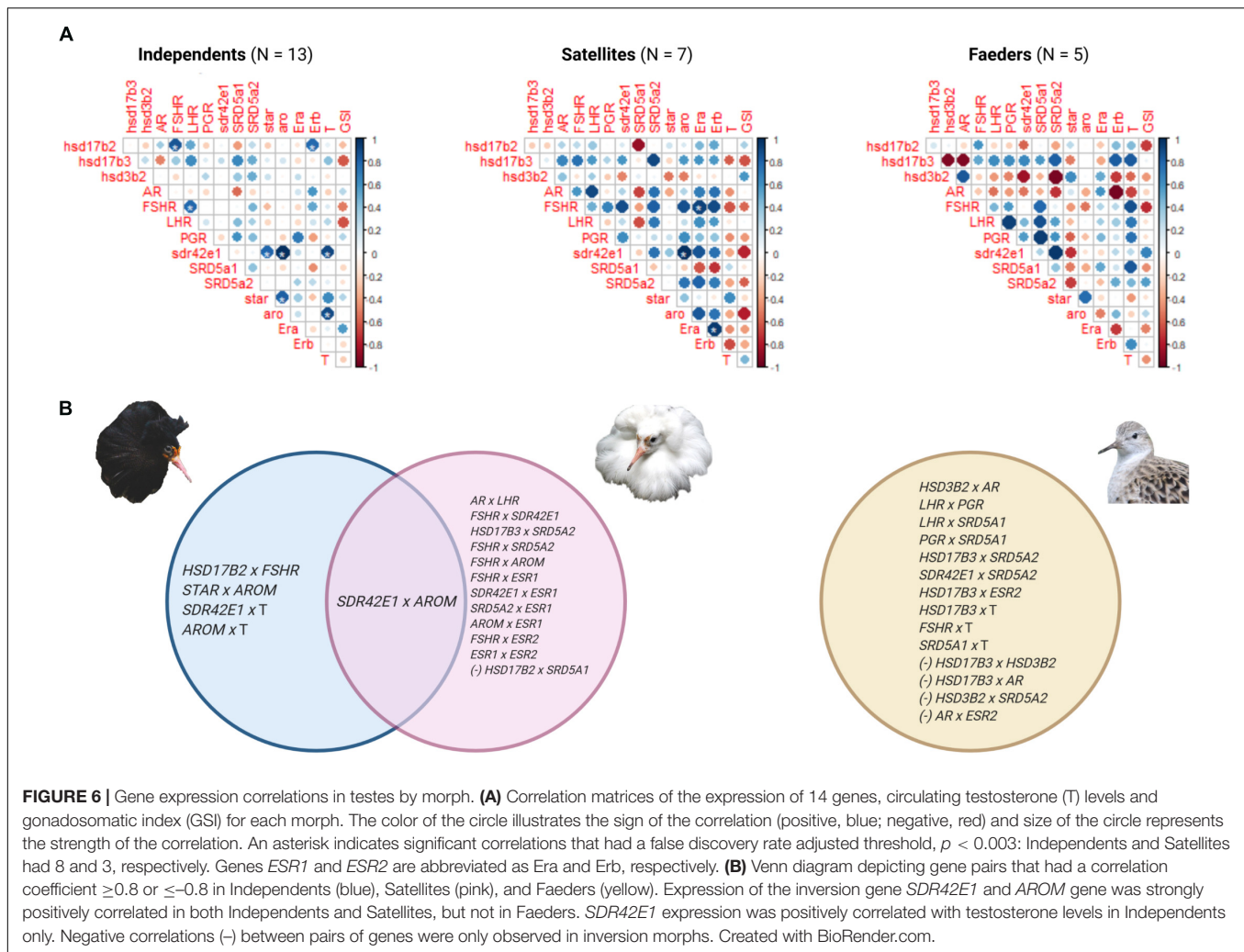


FIGURE 6 | Gene expression correlations in testes by morph. **(A)** Correlation matrices of the expression of 14 genes, circulating testosterone (T) levels and gonadosomatic index (GS) for each morph. The color of the circle illustrates the sign of the correlation (positive, blue; negative, red) and size of the circle represents the strength of the correlation. An asterisk indicates significant correlations that had a false discovery rate adjusted threshold, $p < 0.003$: Independents and Satellites had 8 and 3, respectively. Genes *ESR1* and *ESR2* are abbreviated as Era and Erb, respectively. **(B)** Venn diagram depicting gene pairs that had a correlation coefficient ≥ 0.8 or ≤ -0.8 in Independents (blue), Satellites (pink), and Faeders (yellow). Expression of the inversion gene *SDR42E1* and *AROM* gene was strongly positively correlated in both Independents and Satellites, but not in Faeders. *SDR42E1* expression was positively correlated with testosterone levels in Independents only. Negative correlations (-) between pairs of genes were only observed in inversion morphs. Created with BioRender.com.

all candidate genes investigated, *CENPN* had the strongest pattern of reduced expression in inversion morphs relative to Independents. The *CENPN* protein plays an essential role in the kinetochore, as it is responsible for making contacts with the nucleosome bound *CENPA* protein, a first step in linking centromeric chromatin to microtubules of the mitotic spindle (Carroll et al., 2009; Yan et al., 2019). These results are consistent with the hypothesis that the inversion is homozygous lethal because reduced expression from two inversion alleles would be insufficient to sustain cell division. Furthermore, our findings align with previous reports where breakpoint genes in supergene-mediated behavioral polymorphisms are disrupted in a similar manner across inversion genotypes, sexes and tissues (e.g., see LOC105193832 in Yan et al., 2020). Thus, any negative effects of inversions on the expression of genes that span across or near breakpoints represent the very first challenges that must be offset in order for inversions to have the opportunity to persist within populations. In the case of the ruff, it is clear that the ability to achieve and maintain a minimal level of *CENPN* expression from the ancestral allele is a critical element in the evolutionary trajectory of inversion morphs.

Allelic Imbalance Across Tissues in Inversion Morphs

We found evidence of allelic imbalance in all four inversion genes for which we were able to design allele specific primers that passed all quality control criteria. The magnitude and direction of the imbalance varied by tissue type. Allelic imbalance was consistent across individuals of the same morph, however, we note that sample size was limited, as we only examined four inversion morph individuals (two Satellites and two Faeders). Three genes (*ZFPM1*, *ZNF469*, *ZDHHC7*) were in recently recombined areas and therefore SNPs were exclusive to the Faeder haplotype. We therefore predicted that expression levels would be different in Faeders compared to the other two morphs. Contrary to our expectation, there were no differences in total expression between Satellites and Faeders. However, the two Faeders had reduced expression from the inversion allele for major zinc-finger transcription factors (*ZFPM1*, *ZNF469*) across all tissues. The third gene *ZDHHC7* (zinc finger DHHC-type palmitoyltransferase 7) is essential to the palmitoylation of sex steroid receptors (estrogen receptors, androgen receptor and progesterone receptors)

(Marino and Ascenzi, 2006; Pedram et al., 2012) and showed dramatically reduced expression from the inversion allele in the adrenal glands only. *ZDHHC7* is required for the ability of the estrogen receptor to exert rapid non-genomic (membrane-bound) effects in the brains of mice (Acconcia et al., 2004; Hohoff et al., 2019). Given that estrogen plays a central role in aggressive and courtship behavior in many species of birds (Soma, 2006), the expression differences in *ZDHHC7* could be biologically relevant. We also observed reduced (albeit not statistically clear) total expression of the *ZDHHC7* gene in gonads among inversion males (**Figure 3B**) and a reduced expression of Faeder allele compared to the ancestral allele (**Figure 3C**). We speculate that these differences could affect brain development and adult sexual behavior through a deficit in one or both estrogen receptors, although this requires further studies.

There are ~39 genes in the recently recombined areas that have high sequence similarity between Satellites and Independents (Küpper et al., 2016; Lamichhaney et al., 2016). Three of these (*ZFPM1*, *ZNF469*, *ZDHHC7*) we investigated here and found that gene expression differences seemed unrelated to sequence similarity as Faeder is the divergent morph according to sequence similarity but expression was more similar in Faeders and Satellites with Independents showing different expression for these genes. Follow up studies with larger sample sizes and involving all genes in the areas of recent recombination will shed a light whether this is a general pattern. Nonetheless, our results provide an interesting challenge to the view that gene expression patterns can be predicted based on sequence similarity or level of divergence alone. For the fourth gene (*SPATA2L*), located in an area where Satellites and Faeders share sequence similarity and both share many SNPs with respect to Independents, inversion morphs showed qualitatively reduced gonadal total expression compared to Independents (**Figure 3B**), but increased expression from their respective inversion alleles across all tissues assayed (**Figure 3C**). The *SPATA2L* gene is a paralog of *SPATA2*, which is involved in spermatogenesis (Graziotto et al., 1999; Onisto et al., 2000). *SPATA2L* has been reported to have enriched expression in the brains of zebrafish (*Danio rerio*) (Moro et al., 2002; Maran et al., 2009), which makes it an interesting candidate for further analysis in the ruff, as the morphs show pronounced differences in testes volume index (Küpper et al., 2016) and relative testes size (**Supplementary Figure 2**). Our allelic imbalance results add to previous research that has shown allelic imbalance for another gene within the inversion, *MC1R*, where Satellites have reduced expression from the inversion allele in dark colored feathers, compared to Independents (Schwochow-Thalmann, 2018). We show that across several tissues, expression from the inversion is often reduced, but in some cases may be increased relative to the ancestral allele, as illustrated by *SPATA2L*.

Morphs Can Be Discriminated Based on Expression Differences of Multiple Genes

We analyzed the expression of two key inversion genes, *HSD17B2* and *SDR42E1*, in the context of a dozen non-inversion genes involved in steroid synthesis pathways and signaling in testes.

The goal was to examine whether patterns of co-expression differed between Independents and inversion morphs, and then between Satellites and Faeders as they show differentiated alleles (Küpper et al., 2016). We selected these two particular inversion genes because of their potential to play direct roles in the hormonal differences among morphs (Küpper et al., 2016; Lamichhaney et al., 2016). Morphs show profound differences in androstenedione and testosterone levels during the breeding season, therefore we expected to observe differences in gene expression for *HSD17B2* because this gene's product catalyzes testosterone into androstenedione. In addition, inversion morphs share three major deletions surrounding *HSD17B2* and *SDR42E1*, with one such deletion located in between the two genes and as a consequence, has the potential to alter their expression (Küpper et al., 2016; Lamichhaney et al., 2016). Despite our expectation, neither of the two key genes showed clear expression differences between morphs. Yet, we found that the expression of *SDR42E1* and aromatase genes were positively correlated in both Independents and Satellites, but not in Faeders. This result points to a putatively novel relationship between the inversion and estrogen synthesis, which is an exciting avenue for future research given the well-documented importance of estrogen in brain organization and courtship behavior in other birds (Steimer and Hutchison, 1980; Fusani et al., 2001; Gahr, 2001; Court et al., 2020). However, because both Satellites and Faeders share the deletion that is upstream from *SDR42E1*, identifying which sequence elements (i.e., mutations unique to the Satellite allele or shared between Independents and Satellites) can account for the Satellite and Faeder differences concerning *SDR42E1* and aromatase co-expression will require further study.

Although we did not find expression differences for inversion genes besides *CENPN*, these results echo a previous study on gene expression from feathers of 5 Independent and 6 Satellite males, where none of approximately 6,000 genes assayed were differentially expressed between morphs (Ekblom et al., 2012). In our analysis of non-inversion genes, we also only found differential expression among morphs for *STAR* in the testes (reported previously in Loveland et al., 2021), but most interestingly we found clear non-overlapping morph-specific gene expression profiles when all 14 genes (2 inversion, 12 non-inversion genes) were analyzed together. Consistent with their genotype and phenotype, Satellites clustered in between Faeders and Independents. Remarkably, neither removal of *HSD17B2* nor *ESR2*, the genes with heaviest positive and negative loadings, respectively, did fully abolish clustering of the three morph groups. *ESR2* may be involved in encoding the behavioral differences between males, as a recent study showed that during embryological development the expression of *ESR2*, and not *ESR1*, mediates the organizational de-masculinization of the brain with permanent effects into adulthood in the Japanese quail (*Coturnix japonica*) (Court et al., 2020). When we sequentially removed genes according to their descending loadings, the separation of Faeders from Independents and Satellites persisted until we reached four genes: *STAR*, *HSD17B3*, *HSD3B2*, *AROM* (**Supplementary Figures 3E–H**). Subsequent removal of the *STAR* gene, which we previously reported as

differentially expressed across morphs (Loveland et al., 2021), led to a collapse of the distinct Faeder cluster (**Supplementary Figure 3I**). These results show that the inversion gene *HSD17B2* may be important in the greater context of the gene network for steroid synthesis and signaling, as it ranked among the heaviest loadings, but on its own does not show clear morph differences. Notably, even when both inversion genes *HSD17B2* and *SDR42E1* were removed from the dataset, Faeders continued to cluster separately from Independents and Satellites, indicating perhaps more pronounced trans-acting effects on genes outside of the inversion by the ancient Faeder inversion haplotype than the derived Satellite haplotype. This difference between Faeder and Satellite inversions is more likely to be due to sequence divergence that is independent of hormone levels, given the similarity in androgenic profiles between Satellites and Faeders (Loveland et al., 2021). In contrast, the stepwise removal of genes with lightest loadings led to a core of 10 genes where morphs still clustered, whereas with fewer genes the separation became gradually less clear and completely collapsed in the 2 gene dataset (**Supplementary Figure 4**).

The lack of observable morph differences in the expression of *HSD17B2* does not necessarily rule out a direct role for this gene, as we only measured gene expression and did not directly measure enzymatic activity of this protein. This gene has also accumulated several missense mutations that could affect its function (Küpper et al., 2016; Lamichhaney et al., 2016). One such mutation (A235S) is shared by Satellites and Faeders and is located immediately adjacent to the tyrosine residue (Y236) that forms part of the highly conserved ("N-S-Y-K") catalytic tetrad of these types of enzymes (Persson et al., 2009, 2003) (numbering based on accession number XP_014797711). Therefore, it is possible that the *HSD17B2* enzyme, given evidence of this missense mutation at a critical location, has modified its catalytic rate in inversion morphs compared to Independent males.

We found that the inversion gene *SDR42E1* was strongly positively correlated with aromatase in Independents and Satellites only. Two putative functions for *SDR42E1* are oxidoreductase activity and 3-beta-hydroxy-delta5-steroid dehydrogenase activity, which would give it the capacity to affect several intermediate molecules in steroid synthesis, similar to the role of *HSD3B2* (see **Figure 4**). *SDR42E1* might be involved in converting DHEA to androstenedione, as well as pregnenolone to progesterone, which would be relevant to the hormonal profiles of inversion morphs. Furthermore, in seasonal breeders, the sequential conversion in the brain of adrenally sourced DHEA into androstenedione by *HD3B2*, followed by the conversion of androstenedione to testosterone and then to estrogen by aromatase, has been proposed as one possible mechanism to regulate aggression during the non-breeding season in other bird species, when testosterone levels are very low (Heimovics et al., 2018). Given the low levels of testosterone in inversion morphs, it will be of interest to know whether such a mechanism is at play, either mediated by *HSD3B2* and/or *SDR42E1*, and whether it influences the nearly discrete differences in aggression and courtship behavior between morphs. Unfortunately, direct evidence for specific

functional roles of *SDR42E1* are so far lacking but it certainly remains an appealing candidate for more detailed studies in ruffs. Our report that it has a relationship of co-expression with aromatase, and the fact that *ESR2* ranked in the top three genes in the LDA analysis that was key to separating Satellites from Independents, suggests that estrogen should be given a focus in future studies, given its behavioral effects require the aromatization of androgens. In other bird species, the aromatization of testosterone into estrogen is essential to organizational and activational effects on the brain for adult courtship behaviors (Steimer and Hutchison, 1980; Belle et al., 2005; Soma, 2006; Court et al., 2020), so low testosterone levels in inversion morphs could also have a major consequence on estrogen levels.

Interestingly, *PGR* expression in the testes repeatedly ranked among the top genes with negative loadings. In a recent study, we reported that Faeder males had reduced pituitary expression of *PGR* compared to Independent males (Loveland et al., 2021). Progesterone has been implicated in the regulation of aggression in females in black coucals *Centropus grillii*, a species with reversed sex roles (Goymann et al., 2008), and in courtship behavior in the ring dove (*Streptopelia risoria*), specifically in the context of the effects of aromatase inhibitors (Belle et al., 2005). Thus, it seems that roles for estradiol and progesterone are emerging as potentially relevant for differences among male morphs and it will be interesting to see whether these findings extend into more comprehensive analyses of transcriptomes in neural tissue.

CONCLUSION

A chromosomal inversion has contributed to the origin of two male ARTs and their persistence over considerable evolutionary time. Among ruff males, extreme sexual selection has induced a strong mating skew where a few dominant males sire most of the offspring. The inversion enabled two new types of males to exploit a previously unoccupied social niche. In this way, the inversion profoundly altered behavioral phenotypes and the dynamics of the mating competition. Today's Satellites lack aggression whereas today's Faeders lack both aggression and courtship. Examining patterns of gene expression provides information on the molecular underpinnings of these seemingly adaptive losses of functions in inversion morphs. The ruff inversion captured several genes involved in steroid metabolism. These provide strong candidate genes for the discrete differences in the reproductive biology between morphs, as steroids are major modifiers of gene expression as their genes are highly pleiotropic. We investigated how the expression of candidate genes associated with viability and, differences in physiology and behavior located both, inside and outside of the inversion varies across tissues in male ruffs. As expected, we found that the breakpoint gene *CENPN* has reduced expression in inversion morphs and given its essential role in cell division, still stands as the most likely reason why the inversion is homozygous lethal. We found widespread evidence across tissues for allelic imbalance in inversion genes that are major transcription factors (*ZFPM1*, *ZNF469*),

required for the rapid non-genomic effects of estrogen receptors (*ZDHHC7*), and relevant to spermatogenesis (*SPATA2L*). Besides *CENPN*, no other inversion gene showed differential expression across morphs. However, analyses of the collective expression of 14 genes (combining inversion and non-inversion genes) produced clear non-overlapping clusters of three morphs. When key inversion genes were analyzed collectively with non-inversion genes related to steroid metabolism pathways and signaling, we discovered that the *SDR42E1* and *AROM* genes were positively correlated in Independents and Satellites, but not in Faeders. These results suggest that estradiol synthesis may also be affected by the inversion and could have important implications for understanding the evolution of cooperative courtship in this species. Our findings for genes in recently recombined regions showed that sequence similarity did not predict gene expression patterns, although future work should clarify whether this is true for other genes in these areas. This study provides the basis for further more detailed investigations using gene network analysis based on full transcriptomes.

DATA AVAILABILITY STATEMENT

The raw data are stored in Edmond, the Open Research Data Repository of the Max Planck Society (<https://dx.doi.org/10.17617/3.5z>).

REFERENCES

- Acconcia, F., Ascenzi, P., Bocedi, A., Spisni, E., Tomasi, V., Trentalancia, A., et al. (2004). Palmitoylation-dependent estrogen receptor α membrane localization: regulation by 17 β -estradiol. *Mol. Biol. Cell* 16, 231–237. doi: 10.1091/mbc.e04-07-0547
- Albalat, R., Brunet, F., Laudet, V., and Schubert, M. (2011). Evolution of retinoid and steroid signaling: vertebrate diversification from an amphioxus perspective. *Genome Biol. Evol.* 3, 985–1005. doi: 10.1093/gbe/evr084
- Antonarakis, S. E., Rossiter, J. P., Young, M., Horst, J., de Moerloose, P., Sommer, S. S., et al. (1995). Factor VIII gene inversions in severe hemophilia A: results of an international consortium study. *Blood* 86, 2206–2212. doi: 10.1182/blood.V86.6.2206.bloodjournal8662206
- Balthazart, J., and Ball, G. F. (1995). Sexual differentiation of brain and behavior in birds. *Trends Endocrinol. Metab.* 6, 21–29. doi: 10.1016/1043-2760(94)00098-0
- Beato, M. (1993). “Gene regulation by steroid hormones,” in *Gene Expression: General and Cell-Type-Specific, Progress in Gene Expression*, ed. M. Karin (Boston, MA: Birkhäuser), 43–75. doi: 10.1007/978-1-4684-6811-3_3
- Belle, M. D. C., Sharp, P. J., and Lea, R. W. (2005). Aromatase inhibition abolishes courtship behaviours in the ring dove (*Streptopelia risoria*) and reduces androgen and progesterone receptors in the hypothalamus and anterior pituitary gland. *Mol. Cell. Biochem.* 276, 193–204. doi: 10.1007/s11010-005-4060-6
- Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B* 57, 289–300. doi: 10.1111/j.2517-6161.1995.tb02031.x
- Bentz, A. B., Dossey, E. K., and Rosvall, K. A. (2019). Tissue-specific gene regulation corresponds with seasonal plasticity in female testosterone. *Gen. Comp. Endocrinol.* 270, 26–34. doi: 10.1016/j.ygcen.2018.10.001
- Brinke, A., Tagliavacca, L., Naylor, J., Green, P., Giangrande, P., and Giannelli, F. (1996). Two chimaeric transcription units result from an inversion breaking Intron 1 of the factor VIII gene and a region reportedly affected by reciprocal translocations in T-cell Leukaemia. *Hum. Mol. Genet.* 5, 1945–1951. doi: 10.1093/hmg/5.12.1945
- Carroll, C. W., Silva, M. C. C., Godek, K. M., Jansen, L. E. T., and Straight, A. F. (2009). Centromere assembly requires the direct recognition of CENP-A nucleosomes by CENP-N. *Nat. Cell Biol.* 11, 896–902. doi: 10.1038/ncb1899
- Cheeseman, I. M., and Desai, A. (2008). Molecular architecture of the kinetochore-microtubule interface. *Nat. Rev. Mol. Cell Biol.* 9, 33–46.
- Court, L., Vandries, L., Balthazart, J., and Cornil, C. A. (2020). Key role of estrogen receptor β in the organization of brain and behavior of the Japanese quail. *Horm. Behav.* 125:104827. doi: 10.1016/j.yhbeh.2020.104827
- Cox, R. M. (2020). Sex steroids as mediators of phenotypic integration, genetic correlations, and evolutionary transitions. *Mol. Cell. Endocrinol.* 502:110668. doi: 10.1016/j.mce.2019.110668
- Cox, R. M., Cox, C. L., McGlothlin, J. W., Card, D. C., Andrew, A. L., and Castoe, T. A. (2017). Hormonally mediated increases in sex-biased gene expression accompany the breakdown of between-sex genetic correlations in a sexually dimorphic lizard. *Am. Nat.* 189, 315–332. doi: 10.1086/690105
- Cumming, A. M. (2004). The factor VIII gene intron 1 inversion mutation: prevalence in severe hemophilia A patients in the UK. *J. Thromb. Haemost.* 2, 205–206. doi: 10.1111/j.1538-7836.2004.05621.x
- Dagilis, A. J., and Kirkpatrick, M. (2016). Prezygotic isolation, mating preferences, and the evolution of chromosomal inversions. *Evolution* 70, 1465–1472. doi: 10.1111/evo.12954
- Eklom, R., Farrell, L. L., Lank, D. B., and Burke, T. (2012). Gene expression divergence and nucleotide differentiation between males of different color morphs and mating strategies in the ruff. *Ecol. Evol.* 2, 2485–2505. doi: 10.1002/ece3.370
- Faria, R., Johannesson, K., Butlin, R. K., and Westram, A. M. (2019). Evolving Inversions. *Trends Ecol. Evol.* 34, 239–248. doi: 10.1016/j.tree.2018.12.005
- Fuller, Z. L., Haynes, G. D., Richards, S., and Schaeffer, S. W. (2016). Genomics of natural populations: how differentially expressed genes shape the evolution of

ETHICS STATEMENT

The animal study was reviewed and approved by the Animal Care Committee of Simon Fraser University.

AUTHOR CONTRIBUTIONS

JLL and CK conceived the gene expression elements of the project. DBL established and maintained the captive flock. JLL collected and processed samples, performed qPCR, and analyzed the data. JLL wrote the initial draft with input from CK. All authors edited and approved the final manuscript and designed sample collection conditions.

FUNDING

This study was funded by the Max Planck Society and the National Research Council of Canada.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fgene.2021.641620/full#supplementary-material>

- p>chromosomal inversions in
- Drosophila pseudoobscura*
- .
- Genetics*
- 204, 287–301. doi: 10.1534/genetics.116.191429
- Fusani, L., Gahr, M., and Hutchison, J. B. (2001). Aromatase inhibition reduces specifically one display of the ring dove courtship behavior. *Gen. Comp. Endocrinol.* 122, 23–30. doi: 10.1006/gcen.2001.7608
- Gahr, M. (2001). Distribution of sex steroid hormone receptors in the avian brain: functional implications for neural sex differences and sexual behaviors. *Microsc. Res. Tech.* 55, 1–11. doi: 10.1002/jemt.1151
- Goymann, W., Wittenzellner, A., Schwabl, I., and Makomba, M. (2008). Progesterone modulates aggression in sex-role reversed female African black coucals. *Proc. Biol. Sci.* 275, 1053–1060. doi: 10.1098/rspb.2007.1707
- Graziotto, R., Foresta, C., Scannapieco, P., Zeilante, P., Russo, A., Negro, A., et al. (1999). cDNA cloning and characterization of PD1: a novel human testicular protein with different expressions in various testiculopathies. *Exp. Cell Res.* 248, 620–626. doi: 10.1006/excr.1999.4449
- Heimovics, S. A., Merritt, J. R., Jalabert, C., Ma, C., Maney, D. L., and Soma, K. K. (2018). Rapid effects of 17 β -estradiol on aggressive behavior in songbirds: environmental and genetic influences. *Horm. Behav.* 104, 41–51. doi: 10.1016/j.yhbeh.2018.03.010
- Hohoff, C., Zhang, M., Ambrée, O., Kravchenko, M., Buschert, J., Kerkenberg, N., et al. (2019). Deficiency of the palmitoyl acyltransferase ZDHHC7 impacts brain and behavior of mice in a sex-specific manner. *Brain Struct. Funct.* 224, 2213–2230. doi: 10.1007/s00429-019-01898-6
- Horton, B. M., Michael, C. M., Prichard, M. R., and Maney, D. L. (2020). Vasoactive intestinal peptide as a mediator of the effects of a supergene on social behaviour. *Proc. Biol. Sci.* 287:20200196. doi: 10.1098/rspb.2020.0196
- Huang, Y.-C., Dang, V. D., Chang, N.-C., and Wang, J. (2018). Multiple large inversions and breakpoint rewiring of gene expression in the evolution of the fire ant social supergene. *Proc. Biol. Sci.* 285:20180221. doi: 10.1098/rspb.2018.0221
- Imsland, F., Feng, C., Boije, H., Bed'hom, B., Fillon, V., Dorshorst, B., et al. (2012). The Rose-comb mutation in chickens constitutes a structural rearrangement causing both altered comb morphology and defective sperm motility. *PLoS Genet.* 8:e1002775. doi: 10.1371/journal.pgen.1002775
- Jukema, J., and Piersma, T. (2006). Permanent female mimics in a lekking shorebird. *Biol. Lett.* 2, 161–164. doi: 10.1098/rsbl.2005.0416
- Kent, W. (2002). BLAT - the BLAST-like alignment tool. *Genome Res.* 12, 656–664.
- Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M., et al. (2002). The human genome browser at UCSC. *Genome Res.* 12, 996–1006.
- Kirkpatrick, M. (2010). How and why chromosome inversions evolve. *PLoS Biol.* 8:e1000501. doi: 10.1371/journal.pbio.1000501
- Küpper, C., Stocks, M., Risse, J. E., dos Remedios, N., Farrell, L. L., McRae, S. B., et al. (2016). A supergene determines highly divergent male reproductive morphs in the ruff. *Nat. Genet.* 48, 79–83. doi: 10.1038/ng.3443
- Lakich, D., Kazazian, H. H., Antonarakis, S. E., and Gitschier, J. (1993). Inversions disrupting the factor VIII gene are a common cause of severe haemophilia A. *Nat. Genet.* 5, 236–241. doi: 10.1038/ng1193-236
- Lamichaney, S., Fan, G., Widemo, F., Gunnarsson, U., Thalmann, D. S., Hoepfner, M. P., et al. (2016). Structural genomic changes underlie alternative reproductive strategies in the ruff (*Philomachus pugnax*). *Nat. Genet.* 48, 84–88. doi: 10.1038/ng.3430
- Lank, D. B., Coupe, M., and Wynne-Edwards, K. E. (1999). Testosterone-induced male traits in female ruffs (*Philomachus pugnax*): autosomal inheritance and gender differentiation. *Proc. Biol. Sci.* 266:2323. doi: 10.1098/rspb.1999.0926
- Lank, D. B., Farrell, L. L., Burke, T., Piersma, T., and McRae, S. B. (2013). A dominant allele controls development into female mimic male and diminutive female ruffs. *Biol. Lett.* 9:20130653. doi: 10.1098/rsbl.2013.0653
- Lindtke, D., Lucek, K., Soria-Carrasco, V., Villoutreix, R., Farkas, T. E., Riesch, R., et al. (2017). Long-term balancing selection on chromosomal variants associated with crypsis in a stick insect. *Mol. Ecol.* 26, 6189–6205. doi: 10.1111/mec.14280
- Loveland, J. L., Giraldo-Deck, L. M., Lank, D. B., Goymann, W., Gahr, M., and Küpper, C. (2021). Functional differences in the hypothalamic-pituitary-gonadal axis are associated with alternative reproductive tactics based on an inversion polymorphism. *Horm. Behav.* 127:104877. doi: 10.1016/j.yhbeh.2020.104877
- Maran, C., Tassone, E., Masola, V., and Onisto, M. (2009). The Story of SPATA2 (Spermatogenesis-Associated Protein 2): from Sertoli Cells to Pancreatic Beta-Cells. *Curr. Genomics* 10, 361–363. doi: 10.2174/138920209788920976
- Marino, M., and Ascenzi, P. (2006). Steroid hormone rapid signaling: the pivotal role of S-Palmitoylation. *IUBMB Life* 58, 716–719. doi: 10.1080/15216540601019485
- Merritt, J. R., Davis, M. T., Jalabert, C., Libecap, T. J., Williams, D. R., Soma, K. K., et al. (2018). Rapid effects of estradiol on aggression depend on genotype in a species with an estrogen receptor polymorphism. *Horm. Behav.* 98, 210–218. doi: 10.1016/j.yhbeh.2017.11.014
- Merritt, J. R., Davis, M. T., Jalabert, C., Libecap, T. J., Williams, D. R., Soma, K. K., et al. (2020). A supergene-linked estrogen receptor drives alternative phenotypes in a polymorphic songbird. *Proc. Natl. Acad. Sci. U.S.A.* 117:202011347. doi: 10.1073/pnas.2011347117
- Moro, E., Maran, C., Slongo, M. L., Argenton, F., Toppo, S., and Onisto, M. (2002). Zebrafish spata2 is expressed at early developmental stages. *Int. J. Dev. Biol.* 51, 241–246. doi: 10.1387/ijdb.062220em
- Naseeb, S., Carter, Z., Minnis, D., Donaldson, I., Zeef, L., and Delneri, D. (2016). Widespread impact of chromosomal inversions on gene expression uncovers robustness via phenotypic buffering. *Mol. Biol. Evol.* 33, 1679–1696. doi: 10.1093/molbev/msw045
- Newhouse, D. J., Barcelo-Serra, M., Tuttle, E. M., Gonser, R. A., and Balakrishnan, C. N. (2019). Parent and offspring genotypes influence gene expression in early life. *Mol. Ecol.* 28, 4166–4180. doi: 10.1111/mec.15205
- Onisto, M., Graziotto, R., Scannapieco, P., Marin, P., Merico, M., Slongo, M. L., et al. (2000). A novel gene (PD1) with a potential role on rat spermatogenesis. *J. Endocrinol. Invest.* 23, 605–608. doi: 10.1007/BF03343783
- Pedram, A., Razandi, M., Deschenes, R. J., and Levin, E. R. (2012). DHHC-7 and -21 are palmitoylacyltransferases for sex steroid receptors. *Mol. Biol. Cell* 23, 188–199. doi: 10.1091/mbc.E11-07-0638
- Persson, B., Kallberg, Y., Bray, J. E., Bruford, E., Dellaporta, S. L., Favia, A. D., et al. (2009). The SDR (short-chain dehydrogenase/reductase and related enzymes) nomenclature initiative. *Chem. Biol. Interact.* 178, 94–98. doi: 10.1016/j.cbi.2008.10.040
- Persson, B., Kallberg, Y., Oppermann, U., and Jörnvall, H. (2003). Coenzyme-based functional assignments of short-chain dehydrogenases/reductases (SDRs). *Chem. Biol. Interact.* 14, 271–278. doi: 10.1016/S0009-2797(02)00223-5
- Said, I., Byrne, A., Serrano, V., Cardeno, C., Vollmers, C., and Corbett-Detig, R. (2018). Linked genetic variation and not genome structure causes widespread differential expression associated with chromosomal inversions. *Proc. Natl. Acad. Sci. U.S.A.* 115, 5492–5497. doi: 10.1073/pnas.1721275115
- Schwander, T., Libbrecht, R., and Keller, L. (2014). Supergenes and complex phenotypes. *Curr. Biol.* 24, R288–R294. doi: 10.1016/j.cub.2014.01.056
- Schwochow-Thalmann, D. (2018). *Molecular identification of colour pattern genes in birds*. Uppsala: Swedish University of Agricultural Sciences.
- Soma, K. K. (2006). Testosterone and aggression: berthold, birds and beyond. *J. Neuroendocrinol.* 18, 543–551. doi: 10.1111/j.1365-2826.2006.01440.x
- Steimer, T. H., and Hutchison, J. B. (1980). Aromatization of testosterone within a discrete hypothalamic area associated with the behavioral action of androgen in the male dove. *Brain Res.* 192, 586–591. doi: 10.1016/0006-8993(80)90912-9
- Sun, D., Huh, I., Zinzow-Kramer, W. M., Maney, D. L., and Yi, S. V. (2018). Rapid regulatory evolution of a nonrecombining autosome linked to divergent behavioral phenotypes. *Proc. Natl. Acad. Sci. U.S.A.* 115, 2794–2799. doi: 10.1073/pnas.1717721115
- Tan, M. E., Li, J., Xu, H. E., Melcher, K., and Yong, E. (2015). Androgen receptor: structure, role in prostate cancer and drug discovery. *Acta Pharmacol. Sin.* 36, 3–23. doi: 10.1038/aps.2014.18
- Thierry-Mieg, D., and Thierry-Mieg, J. (2006). AceView: a comprehensive cDNA-supported gene and transcripts annotation. *Genome Biol.* 7(Suppl. 1), S12. doi: 10.1186/gb-2006-7-s1-s12
- van Rhijn, J. (1973). Behavioural dimorphism in male ruffs, *Philomachus pugnax*. *Behaviour* 47, 153–229.
- Wang, Y., Li, J., Feng, C., Zhao, Y., Hu, X., and Li, N. (2017). Transcriptome analysis of comb and testis from Rose-comb Silky chicken (R1/R1) and Beijing Fatty wild type chicken (r/r). *Poult. Sci.* 96, 1866–1873. doi: 10.3382/ps/pew447

- Wellenreuther, M., and Bernatchez, L. (2018). Eco-evolutionary genomics of chromosomal inversions. *Trends Ecol. Evol.* 33, 427–440. doi: 10.1016/j.tree.2018.04.002
- Yan, K., Yang, J., Zhang, Z., McLaughlin, S. H., Chang, L., Fasci, D., et al. (2019). Structure of the inner kinetochore CCAN complex assembled onto a centromeric nucleosome. *Nature* 574, 278–282. doi: 10.1038/s41586-019-1609-1
- Yan, Z., Martin, S. H., Gotzek, D., Arsenault, S. V., Duchon, P., Helleu, Q., et al. (2020). Evolution of a supergene that regulates a trans-species social polymorphism. *Nat. Ecol. Evol.* 4, 240–249. doi: 10.1038/s41559-019-1081-1

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2021 Loveland, Lank and Küpper. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



Impact of Genetic Variation in Gene Regulatory Sequences: A Population Genomics Perspective

Manas Joshi^{1*}, Adamandia Kapopoulou² and Stefan Laurent¹

¹Department of Comparative Development and Genetics, Max Planck Institute for Plant Breeding Research, Cologne, Germany, ²Institute of Ecology and Evolution, University of Bern, Bern, Switzerland

OPEN ACCESS

Edited by:

Deborah A. Triant,
University of Virginia, United States

Reviewed by:

Alejandra Medina-Rivera,
National Autonomous University of
Mexico, Mexico

Ann Kathrin Huylmans,
Institute of Science and Technology
Austria (IST Austria), Austria

*Correspondence:

Manas Joshi
mjoshi@mpipz.mpg.de

Specialty section:

This article was submitted to
Systems Biology Archive,
a section of the journal
Frontiers in Genetics

Received: 29 January 2021

Accepted: 31 May 2021

Published: 02 July 2021

Citation:

Joshi M, Kapopoulou A and
Laurent S (2021) Impact of Genetic
Variation in Gene Regulatory
Sequences: A Population
Genomics Perspective.
Front. Genet. 12:660899.
doi: 10.3389/fgene.2021.660899

The unprecedented rise of high-throughput sequencing and assay technologies has provided a detailed insight into the non-coding sequences and their potential role as gene expression regulators. These regulatory non-coding sequences are also referred to as cis-regulatory elements (CREs). Genetic variants occurring within CREs have been shown to be associated with altered gene expression and phenotypic changes. Such variants are known to occur spontaneously and ultimately get fixed, due to selection and genetic drift, in natural populations and, in some cases, pave the way for speciation. Hence, the study of genetic variation at CREs has improved our overall understanding of the processes of local adaptation and evolution. Recent advances in high-throughput sequencing and better annotations of CREs have enabled the evaluation of the impact of such variation on gene expression, phenotypic alteration and fitness. Here, we review recent research on the evolution of CREs and concentrate on studies that have investigated genetic variation occurring in these regulatory sequences within the context of population genetics.

Keywords: regulatory evolution, natural variation, functional non-coding elements, population genomics, selection, tests for selection

INTRODUCTION

The initial human genome sequencing project revealed that the proportion of the total genome translated into proteins is ~1.5% (International Human Genome Sequencing Consortium, 2001), while the remaining portion (~98.5%) consists of non-coding DNA. This larger proportion of non-coding DNA is a hallmark of the genomes of higher organisms (Li and Liu, 2019). Evaluating the impact of genetic variation at the coding level is facilitated by a large number of annotated gene models and the simplicity of the genetic code for protein-coding DNA sequences. However, similar studies at the functional non-coding level have suffered from the comparatively sparse annotation as well as the complex and multifarious nature of the regulatory code. In this context, a vigorous debate unfolded as to the amount of functional information carried by the non-coding genome and eventually led to the broad acceptance that while essential, non-coding functional elements amount to a modest proportion of the total non-coding DNA (Doolittle, 2013; Graur et al., 2013; Rands et al., 2014; Huang et al., 2017).

In the last decade, advances in sequencing and assay technologies have contributed to the annotation of a large number of functional non-coding elements. For example, the ENCODE and modENCODE consortia (The modENCODE Consortium et al., 2011;

The ENCODE Project Consortium, 2012) used chromatin immunoprecipitation using sequencing (ChIP-seq) and ChIP-on-chip assays to gather a comprehensive catalog of binding sites for a large number of Transcription Factors (TFs) in human, *Drosophila melanogaster*, and *Caenorhabditis elegans* based on genome-wide binding affinity profiles. The availability of such annotation data, along with genomic variation data, has enabled the exploration of non-coding regions for diversity-based signatures of functional constraint. On the other hand, variants occurring in these regions have also contributed to adaptive evolution (Zhen and Andolfatto, 2012). Hence, analyzing the patterns of constraint and variation in CREs contributes to our understanding of between-species phenotypic differences and the process of adaptation.

In this review, we introduce common approaches used to identify regulatory regions. Following this, we will list some of the statistical tools that are used to infer the action of negative and positive selection on non-coding functional regions. Finally, we list studies that presented analyses of selective forces acting at the level of non-coding genomic elements. We have sorted these studies into two sections, the first containing studies that highlight the action of negative (purifying) selection, while the other containing studies that highlight the action of positive selection on non-coding elements (Pollard et al., 2006; Prabhakar et al., 2006; Gittelman et al., 2015).

ANNOTATING NON-CODING ELEMENTS BASED ON THEIR BIOCHEMICAL SIGNATURES

Gene expression regulation is in part controlled by functional non-coding genomic elements. Annotating such elements is important to quantify their exposure to natural selection. Such elements can now be identified based on their biochemical signatures using high-throughput techniques. One of the methods to identify potential regulatory elements is DNase-seq. It allows the identification of regions in the genome at which the chromosome has lost its condensed structure and is therefore susceptible to interactions with available TFs and cleavage by the DNase I nuclease. Such loci are termed DNase I hypersensitive sites (DHSs) and are localized by sequencing the DNA fragments cleaved by the nuclease and mapping them to the reference genome (Sullivan et al., 2015). Another method to assess genome-wide chromatin accessibility is the assay of transposase accessible chromatin using sequencing (ATAC-seq), which is considered faster and more sensitive compared to DNase-seq (Buenrostro et al., 2016). Although loci identified by DNase-seq and ATAC-seq have been shown to be enriched in TF binding sites (TFBS; Calviello et al., 2019), these methods do not provide information about the nature of interacting TFs. On the other hand, ChIP-seq can be used to identify binding sites for a specific TF. In this method, the TF of interest is allowed to bind to its putative binding sites before the DNA is sheared by sonication. TF-DNA bound complexes are then extracted using a TF-specific antibody and DNA is dissociated from the TF and finally sequenced and aligned to the reference genome to identify enriched regions (ChIP-seq peaks; Park, 2009).

ANNOTATING NON-CODING ELEMENTS USING EVOLUTIONARY CONSTRAINT

The availability of whole-genome sequence data from multiple species has enabled the detection of non-coding genomic regions with extreme sequence conservation at various phylogenetic levels. Conservation at these regions is generally thought to be caused by the presence of functional non-coding elements exposed to similar levels of negative selection across a set of species (Sandelin et al., 2004; De La Calle-Mustienes et al., 2005; Pennacchio et al., 2006). Comparative genomic analysis of conserved elements is therefore an efficient approach to detect non-coding elements involved in the regulation of developmental pathways that are common to many higher organisms. Here we list studies that have attempted to identify such conserved elements using different sets of species.

Visel et al. (2007) used a combination of comparative genome analyses coupled with experimental validations to identify tissue-specific human enhancers. Conserved non-coding elements (CNEs) were identified based on conservation across large evolutionary distances (i.e., non-mammalian vertebrates) and tissue-specificity was established using transgenic mice experiments. Additionally, they also identified ultra-conserved elements (UCNEs), defined as being at least 200 bp long and sharing 100% sequence identity between human, mouse, and rat genomes. This dataset is accessible through the VISTA Enhancer Browser,¹ which is actively maintained and currently contains 3,148 *in vivo* tested elements. Woolfe et al. (2007) identified CNEs through multiple pairwise alignments of *Fugu* (pufferfish) and four mammalian genomes (human, mouse, rat, and dog), where CNEs are defined as sequences with 65% identity and are at least 40 bp long. They highlighted the association of the identified CNEs with known developmental genes. Lee et al. (2007) determined CNEs that are associated with Transcription Factors in vertebrate genomes, where CNEs from human to mouse were defined as sharing at least 70% identity and being at least 100 bp long, while CNEs from human to *Fugu* had to share at least 65% identity and being at least 50 bp long. The relaxed criteria for human and *Fugu* genome comparison account for the larger evolutionary distance separating the two species. In addition to this, varying proportions (ranging from 0.63 to 10.45%) of human-*Fugu* CNEs were also identified to be overlapping with regions that are experimentally verified TFBS for various genes, indicating the potential role of CNEs in regulating transcription. Persampieri et al. (2008) described ~73,000 CNEs with at least 50% sequence identity between humans and zebrafish and with a length of at least 50 bp. This collection is accessible through the *cne-Viewer*.² Engström et al. (2008) determined highly conserved non-coding elements (HCNEs) across multiple metazoan species using pairwise whole-genome alignments. The threshold of sequence identity used to define an HCNE for each pair of species ranged from 70 to 100%. This dataset is accessible through the

¹<https://enhancer.lbl.gov/>

²<http://bioinformatics.bc.edu/chuanglab/cneViewer/>

ANCORA database.³ Dimitrieva and Bucher (2013) highlighted UCNEs by comparing the whole genome sequences of human and chicken, where every UCNE was required to have at least 95% sequence identity and a minimal length of 200 bp. In addition to UCNEs, they also highlight ultra-conserved genomic regulatory blocks (UGRBs), which are clusters of UCNEs that show conserved synteny across different vertebrates. They also annotated a subset of their UCNEs as being putative regulatory elements for developmental genes. This collection of UCNEs and UGRBs is available through the UCNEbase website.⁴ Lomonaco et al. (2014) also determined ultra-conserved elements, where every element had to have 100% sequence identity across human, mouse, and rat, in addition to a minimal length of 200 bp.⁵ Dousse et al. (2016) identified CNEs across five clades of vertebrates, where every CNE was identified using the software *phastCons* (Siepel et al., 2005). This collection of CNEs is available in the CEGA database.⁶ Polychronopoulos et al. (2017) have compiled a list with all publicly available CNE datasets.

METHODS FOR DETECTING SELECTION

Inferring the action of selective pressure on the non-coding elements (NCEs) has been one of the central challenges for selection-based studies. One of the major limitations for such studies has been sparse annotation data for regulatory regions. Coding elements in the genome tend to be well-annotated, however the same is not true for the non-coding elements. However, this has been partly overcome due to advances in sequencing technologies, like RNA-seq, ChIP-seq, DNase-seq, etc. Comparative genomics studies have used these biochemical signatures to make an informed guess of the potentially functional NCEs. Various metrics have been introduced to detect selective pressures acting on genomic sequences and the fitness consequences of new mutations by using available regulatory annotation. *phastCons* (Siepel et al., 2005) uses multiple sequence alignment information to identify evolutionarily conserved elements by employing a phylogenetic hidden Markov model. *INSIGHT* (Arbiza et al., 2013) detects the influence of selection on TFBS (ChIP-peaks) based on polymorphism and divergence data; resembling the MacDonald Kreitman (MK) test, named after John H McDonald and Martin Kreitman, who first tested their approach on the *Adh* locus in *D. melanogaster* (McDonald and Kreitman, 1991). The starting point of the MK test is a contingency table summarizing the number of polymorphic (intra-specific) and divergent (inter-specific) variants separately for non-synonymous and synonymous sites. Variants that strongly enhance adaptation tend to fix rapidly in the population and hence contribute less to the polymorphism (within species variation) compared to divergence (between species variation). The MK framework has been used to estimate the proportion of adaptive substitutions that are driven by positive selection

within the population of species, a parameter denoted α . One of the key shortcomings of this approach is its sensitivity to the presence of slightly deleterious mutations, which can severely bias its estimates (Haller and Messer, 2017). However, *INSIGHT* overcomes this by using a probabilistic model that explicitly accounts for the presence of weak negative selection. Key quantities estimated by *INSIGHT* are the proportion of selected sites and the number of adaptive substitutions and weakly deleterious variants. *fitCons* (Gulko et al., 2015) clusters unannotated sequences based on their epigenetic markers and uses ρ metric (probability of a nucleotide within a functional non-coding element to be under selection) inferred from *INSIGHT* to estimate the probability of a new mutation having a potential fitness effect. *LINSIGHT* (Huang et al., 2017) employs neural networks to make an overall estimate of ρ for different genomic features. Here, ρ gives an estimate of which feature is most predictive of fitness for any given positional mutation in the genome. *LASSIE* (Huang and Siepel, 2019) accumulates information on all point-specific mutations within non-coding regions and estimates the selection coefficient of every mutation using a maximum likelihood algorithm. One of the central drawbacks of *fitCons* is that the clustering algorithm is dependent on the epigenomic and annotation signatures and is independent of the evolutionary properties. *fitCons2* (Gulko and Siepel, 2019) addresses this by finding clusters of sites that are distinct in evolutionary and epigenomic properties. Kircher et al. (2014) developed a metric, C-score, which predicts the deleterious effect of a new mutation and is comparable across different sites (non-synonymous, synonymous, regulatory, etc.; Racimo and Schraiber, 2014). Finally, a widely used metric to identify elements under selection is the GERP score (Davydov et al., 2010). This score reflects the decrease of substitutions in an inter-species sequence alignment compared to the neutral expectation. Liu and Robinson-Rechavi (2020) proposed a new method to infer the action of selective forces on TFBSs. This method employs Support-Vector Machines, a machine learning approach, to infer the changes in the binding affinity of the TFBS due to variants and does not necessitate a prior definition of “neutral sites.” Here, variants that aid in adaptation would be expected to improve the binding affinity model of the TFBS and will be consequently maintained under positive selection.

POPULATION GENOMICS ANALYSES OF PURIFYING SELECTION AT NON-CODING FUNCTIONAL ELEMENTS

Deleterious mutations are usually associated with some detrimental effect on the fitness of the species. These mutations are usually subjected to the force of purifying selection and are either lost or are maintained in lower frequency within the population of species. Given their low frequency, they usually do not contribute to the between-species diversity. Here, we document various studies that have highlighted the action of purifying selection in various species.

³<http://ancora.genereg.net>

⁴<https://ccg.epfl.ch/UCNEbase/>

⁵<http://ucbase.unimore.it>

⁶<http://cega.ezlab.org>

Torgerson et al. (2009) analyzed the genetic variation at Conserved non-coding sequences (CNCs) using sequencing data from 35 human samples (20 European Americans and 15 African Americans). CNCs are non-coding sequences that are conserved within a population of species/within a group species. Certain functional studies interpret conservation as a proxy for functionality, hence use conserved elements as potential candidates in their study. For this study, CNCs are defined as non-coding sequences that are conserved in both the human and mouse genome (with at least 70% sequence identity and a minimal length of 100 bp). They report a higher proportion of rare derived alleles in CNCs as compared to synonymous and intergenic sites, indicating the presence of slightly deleterious alleles in CNCs consistent with functional activity. In addition to interpreting summary statistics of genetic variation data, they also reported negative estimates of the population scaled selection coefficient ($\gamma = 2Ns$) for CNCs in the flanking regions of genes. These observations indicate that the CNCs are under a comparatively higher influence of selective constraints as compared to the intergenic and synonymous sites. Mu et al. (2011) used genomic variation data from the 1000 genomes project to analyze patterns of polymorphism on various aspects of TF-binding sites. They found that ChIP-seq peaks harbored an excess of low-frequency SNPs and structural variants (SVs) as compared to motifs not bound by TFs. Using chimpanzees as the outgroup for the divergence study, they also showed that TF-bound motifs had lower SNP divergence as compared to unbound motifs. In a typical ChIP-seq analysis, post precipitation and sequencing, the reads are mapped to the reference genome, and the areas with the highest coverage are identified as ChIP-seq peaks. These peaks typically contain a consensus binding motif for the protein of interest. The Site Frequency Spectra for polymorphic sites and structural variants showed a significant excess of rare alleles in TF-bound motifs as compared to the broader peak regions. This study also showed that regions associated with TF-binding activity are under higher purifying selection compared to non-functional regions, and that intensity of this selection increases with proximity to coding regions of genes. Vernot et al. (2012) measured nucleotide diversity in DNA binding motifs from 732 TFs that overlapped with DNase I peaks from 138 cell and tissue types, using whole-genome sequencing data of 53 human individuals from five populations available in the Complete Genomics database.⁷ They showed that while diversity varies by over seven-fold across binding motifs (from 2.67×10^{-4} to 2.0×10^{-3}), 60% of binding motifs have lower mean diversity than fourfold degenerate sites, consistent with exposure to purifying selection and hence functional constraint. Their results also highlighted an important heterogeneity in diversity levels between binding motifs, with HOX-, POU-, and FOX-domain factors, which are enriched in controllers of development and cell differentiation, displaying particularly low diversity. Diversity measured in DNase I peaks was significantly lower when peaks were shared by multiple cell types. Similar results were obtained for *Saccharomyces cerevisiae* by Connelly et al. (2013), who

quantified the strength of purifying selection acting on binding motifs using genetic variation in 37 strains by employing the metric *Neutrality Index* (NI; Rand and Kann, 1996). Out of the 133 binding motifs in their study, 63 had a value of NI larger than the one obtained for non-synonymous sites indicating a marked exposure of binding motifs to purifying selection. In this study, the authors also used the NI to measure selective constraint at individual intergenic regions. In plants, Haudry et al. (2013) used the whole genome sequence information from nine Brassicaceae species to compare selective constraints acting on CNCs and four-fold degenerate sites using *phastCons* and the folded Site Frequency Spectrum (SFS). SFS is a summary statistic, used extensively in population genetics, which summarizes the distribution of allele frequencies within a population of species. A folded SFS uses the minor allele frequency, i.e., the allele that is the least frequent, to construct the SFS. 90,000 CNCs were identified using *phastCons* analyses on a phylogeny of nine Brassicaceae, representing around 3.8% of the non-coding regions analyzed in this study. In addition to this, they used the population-level data on two of the nine species, namely *A. thaliana* and *C. grandiflora*, to check for a similar signal of conservation within populations. They highlighted that for the population of both species, CNCs displayed an excess of low-frequency minor alleles and lower nucleotide diversity as compared to four-fold degenerate sites, consistent with the action of purifying selection, although this signal was weaker than for the highly conserved zero-fold degenerate sites. De Silva et al. (2014) analyzed patterns of variation at CNCs from CONDOR, a database of developmentally associated CNEs across vertebrates (Woolfe et al., 2007), and obtained multiple alignments for seven vertebrate species (*Homo sapiens*, *Macaca mulatta*, *Mus musculus*, *Gallus gallus*, *Xenopus tropicalis*, *Danio rerio*, and *Takifugu rubripes*). They categorized CNCs into two different classes: non-variable regions (NVRs) which are invariant across all species and restricted variable regions (RVRs) which have at least one variable site across all species (excluding humans). When comparing the SFS of CNCs with synonymous regions (negative control), they observed that CNCs have an excess of rare derived alleles indicating CNCs to be under purifying selection. More specifically, they observed that NVRs have a stronger signal of purifying selection compared to RVRs suggesting that the increased substitution rate observed at RVRs in humans is due to relaxed constraint and not adaptation. They also infer that NVRs harbor a larger proportion of detrimental sites (32%) as compared to that of non-synonymous sites (21%), indicating NVRs to be at a comparatively higher level of purifying selection as compared to non-synonymous regions. Naidoo et al. (2018) used genome-wide polymorphism data from several human populations (Drmanac et al., 2010; Schlebusch et al., 2012; The 1000 Genomes Project Consortium, 2015) and non-coding annotation from the ENCODE project (The ENCODE Project Consortium, 2012). They calculated nucleotide diversity and Tajima's D as a statistical measure of constraints acting on different classes of genome sequences. Tajima's D is a summary statistic that compares the average number of pairwise differences with the average number of segregating sites within a population. Negative

⁷<https://www.completegenomics.com/public-data/>

values of Tajima's *D* indicate an excess of rare alleles and can be used to identify regions where genetic variants segregate at lower frequencies than at neutral loci, a signal consistent with the action of negative selection. They observed coding regions to experience the highest level of purifying selection, closely followed by promoters and untranslated regions (UTRs), while enhancers were the least constrained. This study highlights that on average, regulatory elements in proximity to the coding regions displayed stronger purifying selection as compared to distal ones.

To summarize, the level of constraints acting on CNCs seems to be intermediate when compared with other coding and non-coding sequences. Overall, CNCs are reported to be under higher constraints when compared with synonymous sites and other intergenic regions, and under lower constraints when compared with non-synonymous sites. Such observations could indicate that CNCs are composed of different combinations of binding affinity and non-binding affinity motifs. Hence, the intermediate levels of constraints cannot be directly translated to an overall intermediate level of selection intensities.

POPULATION GENOMIC ANALYSES OF POSITIVE SELECTION AT NON-CODING FUNCTIONAL ELEMENTS

Beneficial mutations represent a small fraction of all naturally occurring mutations, but they are important for adaptation to varying environmental conditions. Given their positive contribution to fitness, such mutations tend to rapidly increase in frequency, and eventually fix, within populations and species. When the number of beneficial mutations responsible for a new adaptive phenotype is small, their rapid increase in frequency generates a characteristic signature in polymorphism data referred to as a selective sweep (Cutter and Payseur, 2013). Several studies have documented selective sweeps in non-coding regions and demonstrated how the beneficial allele modified the expression of the target gene.

Schlenke and Begun (2004) studied the within-species diversity in North-American and African populations of *Drosophila simulans* in the 2R chromosome, a freely recombining region of the genome. The levels of heterozygosity in the 100 kb region under study were reported to be significantly reduced specifically in the North-American population; potentially indicating a recent selective sweep. In this genomic region, they identified a fixed insertion of a transposon in the 5' end of the *Cyp6g1* gene in the American population, correlated with increased transcript abundance, and which had been previously associated with insecticide resistance in *D. melanogaster*. Chan et al. (2010) studied the loss of pelvic phenotype in certain pelvic-reduced stickleback fish populations by performing F1 crosses between pelvic-complete and pelvic-reduced fishes in an experimental setting. They highlighted the loss of an enhancer, *Pel*, for the *Pitx1* gene (expressed in hindlimbs of many vertebrates) in pelvic-reduced fishes to be the driver for the loss of the pelvis. They showed that the heterozygosity at the *Pel* enhancer is significantly less in the pelvic-reduced

compared to pelvic-complete populations, could not be explained solely by population size bottlenecks, and is therefore consistent with the expected signature of a selective sweep. *LCT*, the gene coding for the lactase enzyme, is a well-described example for recent selective sweeps in humans (Enattah et al., 2002; Bersaglieri et al., 2004; Tishkoff et al., 2007). The geographical distribution of the persistence of this enzyme into adulthood is shown to be associated with dairy farming (Enattah et al., 2002), hence the ability to digest lactose during adulthood varies in different populations. This lactase persistence has been proposed to be regulated by cis-acting elements (Wang et al., 1995). Enattah et al. (2002) highlighted two alleles, that are located within the intronic regions of the *MCM6* gene, in the Northern European population that are associated with lactase persistence into adulthood. Bersaglieri et al. (2004) further highlighted the high between-population differences in the frequency of these persistence markers.

Generally, beneficial mutations are rare compared to deleterious mutations, and detecting them is difficult for multiple reasons: confounding demographic parameters can leave similar signatures in the genome, the selective sweep may be too old and the beneficial allele fixed within the population or, in the case of a polygenic adaptation model, the signal for individual loci under positive selection is too weak to be detected (Barghi et al., 2020). Hence, highlighting the effect of positive selection on a single locus is challenging. In addition to this, the effects of underlying background selection are important influencing factors in studies highlighting genome-wide scans of positive selection (see Charlesworth and Jensen, 2021). Some population genomics studies tried to overcome this problem by aggregating the signal carried by genetic variation over multiple loci.

Kudaravalli et al. (2009) tested whether SNPs with a significant association with gene expression (eQTLs) are frequent targets of selection in humans. For this, they analyzed HapMap (International HapMap Consortium, 2003) genomes from three different human populations: Asian, Central European, and Yoruban together with gene expression data from lymphoblastoid cells obtained for all 210 unrelated HapMap individuals. To detect the signature of recent or ongoing positive selection on eQTLs they used the iHS (integrated haplotype score; Voight et al., 2006), a powerful approach to detect selection when the beneficial mutation has not been fixed and is segregating at a frequency between 50 and 85%. Their results showed that SNPs surrounded by signatures of positive selection were more likely to be eQTLs compared to random SNPs, leading to the suggestion that selection on transcript levels is an important aspect of human adaptation. More broadly, this study also showed how logistic regression models can serve as an appropriate statistical approach to test for associations between signals of positive selection and molecular phenotypes. Haddrill et al. (2008) analyzed polymorphism and divergence at 67 coding and non-coding elements in *D. simulans* ($n = 20$). They observed excess of low-frequency alleles in the SFS for introns, 5' UTRs, and 3' UTRs, indicative of selective constraints in those non-coding regions. Based on a MacDonald-Kreitman (MK) analysis, they also reported that the proportion of adaptive

substitution, α , for 5' UTRs and 3' UTRs is comparable with non-synonymous sites in this species. This study highlights that UTRs in *D. simulans* have been under both positive as well as negative selection.

Torgerson et al. (2009) analyzed genetic variation at CNCs using resequencing data from 35 human samples (20 European-Americans and 15 African Americans) using the chimpanzee genome as the outgroup. In this study, CNCs were defined as non-coding sequences conserved in human and mouse. They discretized genomic data into GO functional categories to identify GO categories that are significantly associated with selection in CNCs. For this purpose, they use a modified version of the MK test, *mkprf* (Bustamante et al., 2002), which infers population scaled selection coefficient at individual loci (γ). In African-Americans, three categories associated with positive selection in CNCs were regulation of cellular processes, protein modification, and cell cycle, while categories associated with negative selection were cytosol, ribosome, extracellular region, and carrier activity [with false discovery rate (FDR) < 25%]. In European-Americans, the categories associated with positive selection were calcium ion binding, organelle organization and biogenesis, cell cycle, and behavior (with FDR < 25%), while categories associated with negative selection were proteinaceous extracellular matrix and extracellular space (with FDR < 20%). By analyzing selection pressures acting on genomic data grouped into the functional categories, this study highlights population-specific functional categories that are more likely to be targets of selective forces. He et al. (2011) analyzed TFBS in *D. melanogaster* and *D. simulans* for 30 TFs from REDfly (Rivera et al., 2019), a curated collection of known insect *cis*-regulatory modules. Using a Position Weight Matrix scoring approach, they predicted the effect of each SNP on TFBS binding affinity and measured the frequencies of TFBS-modifying alleles. The unfolded SFS of affinity-decreasing mutations is skewed towards low frequency derived alleles suggesting that negative selection acts to maintain existing TFBS. Furthermore, the results of MacDonald-Kreitman analyses on both affinity increasing and decreasing mutations indicated that positive selection enabled gains and losses of TFBS in both species. Vernot et al. (2012), in addition to evaluating the selective constraint acting on DHSs (see the previous section), used the same dataset to conduct a genome-wide scan for signatures of positive selection at the level of regulatory regions. For this, they used the LSBL metric (Shriver et al., 2004) to identify DNase I peak with significant allele frequency differences across human populations compared to the genomic background. DHS peaks falling in the top 1% of the genome-wide distribution of LSBL values were considered potential targets of positive selection and genes located within 50 kb of those candidate DHS peaks were tested for enrichment of KEGG pathways.⁸ Impressively, among the 15 enriched pathways, the authors identified the melanogenesis pathways in the European population only, suggesting that in addition to already described coding-variants, regulatory changes contribute, as well, to the evolution of adaptive pigmentation phenotypes in this population. Interestingly, the authors also

show that the proportion of highly differentiated DHS differs across the four cell types surveyed in their study.

Arbiza et al. (2013) analyzed polymorphism and divergence data at binding sites with ChIP-seq peaks from the ENCODE project using genomic data from 54 unrelated human individuals and three primate genomes. Based on their new probabilistic implementation of the MK framework (INSIGHT), they estimate the proportion of selected sites in binding sites within ChIP peaks to be 0.33 (vs. 0.80 for second codon positions). They also identify a large variation in the proportion of selected sites across TFs that is largely explained by differences in the information content of the associated binding models. Binding sites also carried a significant signal of positive selection, such that 1 out of 8,300 nucleotides in TFBS was estimated to have been fixed through positive selection and 1 out of 20 recently fixed alleles are adaptive substitutions. Among all TF analyzed in this study, binding sites for the Zinc finger TF GATA2 and GATA3 displayed the largest number of adaptive substitutions (312). Huang et al. (2017) studied cell and tissue-specific constraints acting on enhancers. They obtained a comprehensive enhancer annotation list in humans from a study by Andersson et al. (2014). They introduced LINSIGHT, a method to estimate the fitness consequence of mutation in non-coding regions. They showed that enhancers in tissues associated with sensory perception, the immune system, and the male reproductive system have low LINSIGHT scores, suggesting that these enhancers are under low constraints and could potentially be under positive selection. They also point out that enhancers active in tissues associated with the female reproductive system are under higher constraints as compared to tissues associated with the male reproductive system. Along with the introduction of LINSIGHT, this study highlighted that the fitness consequence of mutations in enhancers is dependent on many aspects, including cell and tissue specificity, and constraints acting on the promoters of the target gene.

DISCUSSION

Concerning purifying selection, polymorphism- and divergence-based studies highlighted in this review (to exemplify — Torgerson et al., 2009; De Silva et al., 2014) demonstrate that certain non-coding elements in the genome seem to be under constraint, indicating the action of purifying selection. Interestingly, De Silva et al. (2014) pointed out that highly conserved CNCs across various vertebrates appear to be under higher levels of constraints as compared to non-synonymous sites. DNA binding motifs within corresponding ChIP-seq peaks, one of the most precise annotations of functional non-coding elements across various species, have been shown to be under higher levels of purifying selection as compared to fourfold degenerate sites in humans (Vernot et al., 2012), and, as compared to non-synonymous sites in yeast (Connelly et al., 2013). However, overall, functional non-coding classes show varying patterns of purifying selection which is intermediate to synonymous and non-synonymous sites (Haddrill et al., 2008; Torgerson et al., 2009). Often regulatory modules have

⁸<https://www.genome.jp/kegg/pathway.html>

been highlighted to be under higher constraints as compared to other non-annotated and non-coding regions, as an example TFBS motifs are under higher constraints as compared to unbound motifs (Mu et al., 2011). Finally, the intensity of these constraints seems to be higher for elements that are in the proximity of the coding regions as compared to the distal elements (Mu et al., 2011; Naidoo et al., 2018).

Demonstrating the effect of positive selection on individual loci is challenging, hence several studies took the approach of pooling loci-based common non-coding functional annotation and compared the aggregated signal to a neutral reference class. For example, Torgerson et al. (2009) and Vernot et al. (2012) clustered regulatory regions associated with genes based on their functional GO terms and biological pathways that they participate in, respectively, for various human populations. Here, Vernot et al. (2012) highlighted that the pigmentation pathway seems to constitute genes whose regulatory elements could have potentially been targeted by positive selection in the European population, suggesting that regulatory changes could be responsible for adaptive phenotypic changes. Huang et al. (2017) employ their method LINSIGHT to aggregate signals of selection from different tissue types in humans. In some cases, regulatory elements could be under the influence of both positive as well as negative selection, where negative selection maintains regulatory elements and positive selection is responsible for their gain and loss within species as has been pointed out in the studies by Haddrill et al. (2008) and He et al. (2011). Additionally, in some cases, variations in CREs have also been proposed to be paving way for speciation. To exemplify, Mack and Nachman (2017) highlighted that the accumulation of variations within CREs could be linked to post-zygotic isolation which eventually leads to a reduction in inter-species gene flow and, potentially, speciation. Post-zygotic isolation is one of the mechanisms of reproductive isolation where the inter-species hybrid is either inviable or sterile, thus leading to an increase in the inter-species differences. Along similar lines, in a speciation study on *Mus musculus domesticus* and *Mus musculus musculus*, Mack et al. (2016) highlighted the potential role of the accumulation of changes within regulatory elements in speciation.

CHALLENGES AND PITFALLS

With the advent of whole-genome sequencing and assay technologies, the availability of genomic data has grown exponentially. However, one of the central questions remains open, which is — What proportion of the genome is “functional”? In the case of non-coding elements, many studies have highlighted the approach of linking conservation to functionality, however, this could potentially dilute the signature of natural selection (Arbiza et al., 2013). To make this search more precise, a proposed alternative has been to exclusively consider non-coding sequences that display some biochemical signatures. Many comparative genomics studies use these biochemical signatures as starting points to detect patterns of constraints on elements (displaying the signatures) as compared to putative neutral elements and infer functionality and forces of selection in action. Here, constraints

and the presence of a biochemical signature, both, act as a proxy for functionality. Along similar lines, in the case of humans, some studies have hypothesized the proportion of functional sites in the genome to be around 4–8% (Ward and Kellis, 2012; Rands et al., 2014), these estimates are based on the proportion of sites that are under constraint. However, changes in the non-constraint regions could also have functional consequences (Ludwig et al., 2000; Dermitzakis and Clark, 2002). Estimating the proportion of genetic variants, within CREs, that contribute to adaptive evolution is challenging, mainly due to a lack of a robust model of neutral versus adaptive evolution, specifically for regulatory regions (Liu and Robinson-Rechavi, 2020).

As compared to coding regions, functional studies are challenging within NCEs due to sparse annotation data. This has been partially overcome with biochemical assays and large-scale annotation projects like ENCODE (humans; The ENCODE Project Consortium, 2012) and modENCODE (*D. melanogaster*; The modENCODE Consortium et al., 2011). However, these assays generally highlight biochemically active regions, which is not a direct indication of functionality (Doolittle, 2013; Graur et al., 2013; Huang et al., 2017). This advocates for the need for refined functional annotations of the non-coding elements. One of the other challenges in functional studies has been to choose appropriate neutral sites, which are sites indifferent to variations. Comparing such sites against “test” sites aid in elucidating the signal of selective forces. However, as highlighted by Casillas et al. (2007), the choice of the genomic class used as neutral reference can lead to under- or over-estimations of the action of selective forces on “test” sites. In addition to selecting neutral regions, the neutral forces associated with the demographic history of the populations should also be factored in to make an informed estimate of the action of selective forces (Zhen and Andolfatto, 2012).

Such methods usually aggregate the signal of selection over multiple loci, as the signal from a single locus is sparse, estimating the marginal contribution of individual loci is difficult (Andolfatto, 2005). To exemplify, in the case of local adaptation, a certain group of loci that contribute to adaptation will be under the action of positive selection and evolve rapidly as compared to the other non-functional sequences. Methods attempting to detect selective pressure will highlight these loci. To aid in interpretation, some studies (Haygood et al., 2007; Torgerson et al., 2009) use Gene Ontology (GO) identifiers to highlight biological categories, and consequently the participating genes, which are likely subject to selection. However, as highlighted by Galtier and Duret (2007), one of the explanations for rapidly evolving elements, besides the action of selection, could also be other factors, such as biased gene conversion, making the inference of selection challenging.

Effective partitioning of the regulatory elements is one of the central challenges for performing functional studies of the non-coding elements. He et al. (2011) highlight an interesting approach of partitioning the regulatory regions into affinity increasing and affinity decreasing sites, similar to synonymous and non-synonymous sites in the coding regions. However, such partitioning is only possible for regulatory elements that have a well-characterized binding model. The new sequencing

methods and the rapid rise in sequencing data will help to fine-tune the NCE annotation and establish TF-specific binding models. In addition to this, low-affinity binding sites in NCEs have been reported to play a key role in regulatory robustness by enabling the regulatory elements to harbor multiple binding sites (Crocker et al., 2015; Hajheidari et al., 2019). However, reliable detection of such elements through ChIP-seq experiments has been challenging due to their sparse signal, making them difficult to distinguish from the genomic background noise (Crocker et al., 2015). Enabling reliable detection of such elements will be one of the major challenges for future developments in assay technologies and bioinformatics pipelines.

REFERENCES

- Andersson, R., Claudia, G., Irene, M.-E., Ilka, H., Jette, B., Mette, B., et al. (2014). An atlas of active enhancers across human cell types and tissues. *Nature* 507, 455–461. doi: 10.1038/nature12787
- Andolfatto, P. (2005). Adaptive evolution of non-coding DNA in *Drosophila*. *Nature* 437, 1149–1152. doi: 10.1038/nature04107
- Arbiza, L., Ilan, G., Bulent, A. A., Melissa, J. H., Brad, G., Alon, K., et al. (2013). Genome-wide inference of natural selection on human transcription factor binding sites. *Nat. Genet.* 45, 723–729. doi: 10.1038/ng.2658
- Barghi, N., Hermisson, J., and Schlötterer, C. (2020). Polygenic adaptation: a unifying framework to understand positive selection. *Nat. Rev. Genet.* 21, 769–781. doi: 10.1038/s41576-020-0250-z
- Bersaglieri, T., Sabeti, P. C., Patterson, N., Vanderploeg, T., Schaffner, S. F., Drake, J. A., et al. (2004). Genetic signatures of strong recent positive selection at the lactase gene. *Am. J. Hum. Genet.* 74, 1111–1120. doi: 10.1086/421051
- Buenrostro, J., Wu, B., Chang, H., and Greenleaf, W. (2016). ATAC-seq method. *Curr. Protoc. Mol. Biol.* 2015, 21–29. doi: 10.1002/0471142727.mb2129s109
- Bustamante, C. D., Nielsen, R., Sawyer, S. A., Olsen, K. M., Purugganan, M. D., and Hartl, D. L. (2002). The cost of inbreeding in *Arabidopsis*. *Nature* 416, 531–534. doi: 10.1038/416531a
- Calle-Mustienes, L., De, E., Feijóo, C. G., Manzanares, M., Tena, J. J., Rodríguez-Seguel, E., et al. (2005). A functional survey of the enhancer activity of conserved non-coding sequences from vertebrate Iroquois cluster gene deserts. *Genome Res.* 15, 1061–1072. doi: 10.1101/gr.4004805
- Calviello, K., Antje, H., Wurmus, R., Yusuf, D., and Ohler, U. (2019). Reproducible inference of transcription factor footprints in ATAC-seq and DNase-seq datasets using protocol-specific bias modeling. *Genome Biol.* 20:42. doi: 10.1186/s13059-019-1654-y
- Casillas, S., Barbadilla, A., and Bergman, C. M. (2007). Purifying selection maintains highly conserved noncoding sequences in *Drosophila*. *Mol. Biol. Evol.* 24, 2222–2234. doi: 10.1093/molbev/msm150
- Chan, Y. F., Marks, M. E., Jones, F. C., Villarreal, G., Shapiro, M. D., Brady, S. D., et al. (2010). Adaptive evolution of pelvic reduction in sticklebacks by recurrent deletion of a *pitx1* enhancer. *Science* 327, 302–305. doi: 10.1126/science.1182213
- Charlesworth, B., and Jensen, J. D. (2021). The effects of selection at linked sites on patterns of genetic variability. *AREES* (in press).
- Connelly, C. F., Skelly, D. A., Dunham, M. J., and Akey, J. M. (2013). Population genomics and transcriptional consequences of regulatory motif variation in globally diverse *Saccharomyces cerevisiae* strains. *Mol. Biol. Evol.* 30, 1605–1613. doi: 10.1093/molbev/mst073
- Crocker, J., Abe, N., Rinaldi, L., McGregor, A. P., Frankel, N., Wang, S., et al. (2015). Low affinity binding site clusters confer HOX specificity and regulatory robustness. *Cell* 160, 191–203. doi: 10.1016/j.cell.2014.11.041
- Cutter, A. D., and Payseur, B. A. (2013). Genomic signatures of selection at linked sites: unifying the disparity among species. *Nat. Rev. Genet.* 14, 262–274. doi: 10.1038/nrg3425
- Davydov, E. V., Goode, D. L., Sirota, M., Cooper, G. M., Sidow, A., and Batzoglou, S. (2010). Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLoS Comput. Biol.* 6:1001025. doi: 10.1371/journal.pcbi.1001025
- De Silva, D. R., Nichols, R., and Elgar, G. (2014). Purifying selection in deeply conserved human enhancers is more consistent than in coding sequences. *PLoS One* 9:e103357. doi: 10.1371/journal.pone.0103357
- Dermitzakis, E. T., and Clark, A. G. (2002). Evolution of transcription factor binding sites in mammalian gene regulatory regions: conservation and turnover. *Mol. Biol. Evol.* 17, 1114–1121. doi: 10.1093/oxfordjournals.molbev.a004169
- Dimitrieva, S., and Bucher, P. (2013). UCNEbase — A database of ultraconserved non-coding elements and genomic regulatory blocks. *Nucleic Acids Res.* 41, 101–109. doi: 10.1093/nar/gks1092
- Doolittle, W. F. (2013). Is junk DNA bunk? A critique of ENCODE. *Proc. Natl. Acad. Sci. U. S. A.* 110, 5294–5300. doi: 10.1073/pnas.1221376110
- Dousse, A., Junier, T., and Zdobnov, E. M. (2016). CEGA-a catalog of conserved elements from genomic alignments. *Nucleic Acids Res.* 44, D96–D100. doi: 10.1093/nar/gkv1163
- Drmanac, R., Sparks, A. B., Callow, M. J., Halpern, A. L., Burns, N. L., Kermani, B. G., et al. (2010). Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. *Science* 327, 78–81. doi: 10.1126/science.1181498
- Enattah, N. S., Sahi, T., Savilahti, E., Terwilliger, J. D., Peltonen, L., and Järvelä, I. (2002). Identification of a variant associated with adult-type hypolactasia. *Nat. Genet.* 30, 233–237. doi: 10.1038/ng826
- Engström, P. G., Fredman, D., and Lenhard, B. (2008). Ancora: a web resource for exploring highly conserved noncoding elements and their association with developmental regulatory genes. *Genome Biol.* 9, 8–11. doi: 10.1186/gb-2008-9-2-r34
- Galtier, N., and Duret, L. (2007). Adaptation or biased gene conversion? Extending the null hypothesis of molecular evolution. *Trends Genet.* 23, 273–277. doi: 10.1016/j.tig.2007.03.011
- Gittelman, R. M., Hun, E., Ay, F., Madeoy, J., Pennacchio, L., Noble, W. S., et al. (2015). Comprehensive identification and analysis of human accelerated regulatory DNA. *Genome Res.* 25, 1245–1255. doi: 10.1101/gr.192591.115
- Graur, D., Zheng, Y., Price, N., Azevedo, R. B. R., Zufall, R. A., and Elhaik, E. (2013). On the immortality of television sets: ‘function’ in the human genome according to the evolution-free gospel of encode. *Genome Biol. Evol.* 5, 578–590. doi: 10.1093/gbe/evt028
- Gulko, B., Hubisz, M. J., Gronau, I., and Siepel, A. (2015). A method for calculating probabilities of fitness consequences for point mutations across the human genome. *Nat. Genet.* 47, 276–283. doi: 10.1038/ng.3196
- Gulko, B., and Siepel, A. (2019). An evolutionary framework for measuring epigenomic information and estimating cell-type-specific fitness consequences. *Nat. Genet.* 51, 335–342. doi: 10.1038/s41588-018-0300-z
- Hadrill, P. R., Bachtrog, D., and Andolfatto, P. (2008). Positive and negative selection on noncoding DNA in *Drosophila simulans*. *Mol. Biol. Evol.* 25, 1825–1834. doi: 10.1093/molbev/msn125
- Hajheidari, M., Wang, Y., Bhatia, N., Vuolo, F., Franco-Zorrilla, J. M., Karady, M., et al. (2019). Autoregulation of RCO by low-affinity binding modulates cytokinin action and shapes leaf diversity. *Curr. Biol.* 29, 4183–4192. doi: 10.1016/j.cub.2019.10.040

AUTHOR CONTRIBUTIONS

MJ and SL designed the structure of the review. MJ, AK, and SL wrote the review. All authors contributed to the article and approved the submitted version.

FUNDING

SL and MJ acknowledge funding by a core grant of the Max-Planck Society to Miltos Tsiantis from the Department of Comparative Development and Genetics.

- Haller, B. C., and Messer, P. W. (2017). AsymptoticMK: a web-based tool for the asymptotic McDonald-Kreitman test. *G3: Genes Genome Genet.* 7, 1569–1575. doi: 10.1534/g3.117.039693
- Haudry, A., Platts, A. E., Vello, E., Hoen, D. R., Leclercq, M., Williamson, R. J., et al. (2013). An atlas of over 90,000 conserved noncoding sequences provides insight into crucifer regulatory regions. *Nat. Genet.* 45, 891–898. doi: 10.1038/ng.2684
- Haygood, R., Fedrigo, O., Hanson, B., Yokoyama, K. D., and Wray, G. A. (2007). Promoter regions of many neural- and nutrition-related genes have experienced positive selection during human evolution. *Nat. Genet.* 39, 1140–1144. doi: 10.1038/ng2104
- He, B. Z., Holloway, A. K., Maerkl, S. J., and Kreitman, M. (2011). Does positive selection drive transcription factor binding site turnover? A test with *Drosophila* cis-regulatory modules. *PLoS Genet.* 7:1002053. doi: 10.1371/journal.pgen.1002053
- Huang, Y. F., Gulko, B., and Siepel, A. (2017). Fast, scalable prediction of deleterious noncoding variants from functional and population genomic data. *Nat. Genet.* 49, 618–624. doi: 10.1038/ng.3810
- Huang, Y. F., and Siepel, A. (2019). Estimation of allele-specific fitness effects across human protein-coding sequences and implications for disease. *Genome Res.* 29, 1310–1321. doi: 10.1101/gr.245522.118
- International HapMap Consortium (2003). The international HapMap project. *Nature* 426, 789–796. doi: 10.1038/nature02168
- International Human Genome Sequencing Consortium (2001). Initial sequencing and analysis of the human genome. *Nature* 409, 860–921. doi: 10.1038/35057062
- Kircher, M., Witten, D. M., Jain, P., O’roak, B. J., Cooper, G. M., and Shendure, J. (2014). A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.* 46, 310–315. doi: 10.1038/ng.2892
- Kudaravalli, S., Veyrieras, J. B., Stranger, B. E., Dermitzakis, E. T., and Pritchard, J. K. (2009). Gene expression levels are a target of recent natural selection in the human genome. *Mol. Biol. Evol.* 26, 649–658. doi: 10.1093/molbev/msn289
- Lee, A. P., Yang, Y., Brenner, S., and Venkatesh, B. (2007). TFCONES: a database of vertebrate transcription factor-encoding genes and their associated conserved noncoding elements. *BMC Genomics* 8:441. doi: 10.1186/1471-2164-8-441
- Li, J., and Liu, C. (2019). Coding or noncoding, the converging concepts of RNAs. *Front. Genet.* 10:496. doi: 10.3389/fgenet.2019.00496
- Liu, J., and Robinson-Rechavi, M. (2020). Robust inference of positive selection on regulatory sequences in the human brain. *Sci. Adv.* 6:eabc9863. doi: 10.1126/sciadv.abc9863
- Lomonaco, V., Martoglia, R., Mandreoli, F., Anderlucci, L., Emmett, W., Biciato, S., et al. (2014). UCbase 2.0: ultraconserved sequences database. *Database* 2014:bau062. doi: 10.1093/database/bau062
- Ludwig, M. Z., Bergman, C., Patel, N. H., and Kreitman, M. (2000). Evidence for stabilizing selection in a eukaryotic enhancer element. *Nature* 403, 564–567. doi: 10.1038/35000615
- Mack, K. L., Campbell, P., and Nachman, M. W. (2016). Gene regulation and speciation in house mice. *Genome Res.* 26, 451–461. doi: 10.1101/gr.195743.115
- Mack, K. L., and Nachman, M. W. (2017). Gene regulation and speciation. *Trends Genet.* 33, 68–80. doi: 10.1016/j.tig.2016.11.003
- McDonald, J. H., and Kreitman, M. (1991). Adaptive protein evolution at the *Adh* locus in *Drosophila*. *Nature* 351, 652–654. doi: 10.1038/351652a0
- Mu, X. J., Lu, Z. J., Kong, Y., Lam, H. Y. K., and Gerstein, M. B. (2011). Analysis of genomic variation in non-coding elements using population-scale sequencing data from the 1000 genomes project. *Nucleic Acids Res.* 39, 7058–7076. doi: 10.1093/nar/gkr342
- Naidoo, T., Sjödin, P., Schlebusch, C., and Jakobsson, M. (2018). Patterns of variation in cis-regulatory regions: examining evidence of purifying selection. *BMC Genomics* 19:95. doi: 10.1186/s12864-017-4422-y
- Park, P. J. (2009). ChIP-seq: advantages and challenges of a maturing technology. *Nat. Rev. Genet.* 10, 669–680. doi: 10.1038/nrg2641
- Pennacchio, L. A., Ahituv, N., Moses, A. M., Prabhakar, S., Nobrega, M. A., Shoukry, M., et al. (2006). In vivo enhancer analysis of human conserved non-coding sequences. *Nature* 444, 499–502. doi: 10.1038/nature05295
- Persampieri, J., Ritter, D. I., Lees, D., Lehoczky, J., Li, Q., and Guo, S. (2008). CneViewer: a database of conserved non-coding elements for studies of tissue-specific gene regulation. *Bioinformatics* 24, 2418–2419. doi: 10.1093/bioinformatics/btn443
- Pollard, K. S., Salama, S. R., King, B., Kern, A. D., Dreszer, T., Katzman, S., et al. (2006). Forces shaping the fastest evolving regions in the human genome. *PLoS Genet.* 2:e168. doi: 10.1371/journal.pgen.0020168
- Polychronopoulos, D., King, J. W. D., Nash, A. J., Tan, G., and Lenhard, B. (2017). Conserved non-coding elements: developmental gene regulation meets genome organization. *Nucleic Acids Res.* 45, 12611–12624. doi: 10.1093/nar/gkx1074
- Prabhakar, S., Noonan, J. P., Pääbo, S., and Rubin, E. M. (2006). Accelerated evolution of conserved noncoding sequences in humans. *Science* 314:786. doi: 10.1126/science.1130738
- Racimo, F., and Schraiber, J. G. (2014). Approximation to the distribution of fitness effects across functional categories in human segregating polymorphisms. *PLoS Genet.* 10:1004697. doi: 10.1371/journal.pgen.1004697
- Rand, D. M., and Kann, L. M. (1996). Excess amino acid polymorphism in mitochondrial DNA: contrasts among genes from *Drosophila*, mice, and humans. *Mol. Biol. Evol.* 13, 735–748. doi: 10.1093/oxfordjournals.molbev.a025634
- Rands, C. M., Meader, S., Ponting, C. P., and Lunter, G. (2014). 8.2% of the human genome is constrained: variation in rates of turnover across functional element classes in the human lineage. *PLoS Genet.* 10:1004525. doi: 10.1371/journal.pgen.1004525
- Rivera, J., Keränen, S. V. E., Gallo, S. M., and Halfon, M. S. (2019). REDfly: the transcriptional regulatory element database for *Drosophila*. *Nucleic Acids Res.* 47, D828–D834. doi: 10.1093/nar/gky957
- Sandelin, A., Bailey, P., Bruce, S., Engström, P. G., Klos, J. M., Wasserman, W. W., et al. (2004). Arrays of ultraconserved non-coding regions span the loci of key developmental genes in vertebrate genomes. *BMC Genomics* 5:99. doi: 10.1186/1471-2164-5-99
- Schlebusch, C. M., Skoglund, P., Sjödin, P., Gattepaille, L. M., Hernandez, D., Jay, F., et al. (2012). Genomic variation in seven khoe-san complex African history. *Science* 1187, 374–379. doi: 10.1126/science.1227721
- Schlenke, T. A., and Begun, D. J. (2004). Strong selective sweep associated with a transposon insertion in *Drosophila simulans*. *Proc. Natl. Acad. Sci. U. S. A.* 101, 1626–1631. doi: 10.1073/pnas.0303793101
- Shriver, M. D., Kennedy, G. C., Parra, E. J., Lawson, H. A., Sonpar, V., Huang, J., et al. (2004). The genomic distribution of population substructure in four populations using 8,525 autosomal SNPs. *Hum. Genomics* 1, 274–286. doi: 10.1186/1479-7364-1-4-274
- Siepel, A., Bejerano, G., Pedersen, J. S., Hinrichs, A. S., Hou, M., Rosenbloom, K., et al. (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* 15, 1034–1050. doi: 10.1101/gr.3715005
- Sullivan, A. M., Bubb, K. L., Sandstrom, R., Stamatoyanopoulos, J. A., and Queitsch, C. (2015). DNase I hypersensitivity mapping, genomic footprinting, and transcription factor networks in plants. *Curr. Plant Biol.* 3, 40–47. doi: 10.1016/j.cpb.2015.10.001
- The 1000 Genomes Project Consortium (2015). A global reference for human genetic variation. *Nature* doi: 10.1038/nature15393
- The ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature* 489, 57–74. doi: 10.1038/nature11247
- The modENCODE Consortium, Roy, S., Ernst, J., Kharchenko, P. V., Kheradpour, P., Negre, N., et al. (2011). Identification of functional elements and regulatory circuits by *Drosophila* ModENCODE. *Science* 330, 1787–1797. doi: 10.1126/science.1198374.Identification
- Tishkoff, S. A., Reed, F. A., Ranciaro, A., Voight, B. F., Babbitt, C. C., Silverman, J. S., et al. (2007). Convergent adaptation of human lactase persistence in Africa and Europe. *Nat. Genet.* 39, 31–40. doi: 10.1038/ng1946
- Torgerson, D. G., Boyko, A. R., Hernandez, R. D., Indap, A., Hu, X., White, T. J., et al. (2009). Evolutionary processes acting on candidate cis-regulatory regions in humans inferred from patterns of polymorphism and divergence. *PLoS Genet.* 5:1000592. doi: 10.1371/journal.pgen.1000592
- Vernot, B., Stergachis, A. B., Maurano, M. T., Vierstra, J., Neph, S., Thurman, R. E., et al. (2012). Personal and population genomics of human regulatory variation. *Genome Res.* 22, 1689–1697. doi: 10.1101/gr.134890.111
- Visel, A., Minovitsky, S., Dubchak, I., and Pennacchio, L. A. (2007). VISTA enhancer browser — a database of tissue-specific human enhancers. *Nucleic Acids Res.* 35, 88–92. doi: 10.1093/nar/gkl822

- Voight, B. F., Kudaravalli, S., Wen, X., and Pritchard, J. K. (2006). A map of recent positive selection in the human genome. *PLoS Biol.* 4:e72 doi: 10.1371/journal.pbio.0040072
- Wang, Y., Harvay, C. B., Pratt, W. S., Sams, V., Sarner, M., Rossi, M., et al. (1995). The lactase persistence/non-persistence polymorphism is controlled by a cis-acting element. *Hum. Mol. Genet.* 4, 657–662. doi: 10.1093/hmg/4.4.657
- Ward, L. D., and Kellis, M. (2012). Evidence of abundant purifying selection in humans for recently acquired regulatory functions. *Science* 334, 1675–1678. doi: 10.1126/science.1225057
- Woolfe, A., Goode, D. K., Cooke, J., Callaway, H., Smith, S., Snell, P., et al. (2007). CONDOR: a database resource of developmentally associated conserved non-coding elements. *BMC Dev. Biol.* 7:100. doi: 10.1186/1471-213X-7-100
- Zhen, Y., and Andolfatto, P. (2012). Methods to detect selection on noncoding DNA. *Methods Mol. Biol.* 2012, 141–149. doi: 10.1007/978-1-61779-585-5_6
- Conflict of Interest:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2021 Joshi, Kapopoulou and Laurent. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



Evolutionary Perspective and Expression Analysis of Intronless Genes Highlight the Conservation of Their Regulatory Role

OPEN ACCESS

Edited by:

Katja Nowick,
Freie Universität Berlin, Germany

Reviewed by:

Jaroslav Bryk,
University of Huddersfield,
United Kingdom
Scott William Roy,
San Francisco State University,
United States
Ekaterina Shelest,
University of Portsmouth,
United Kingdom

*Correspondence:

Alfredo Varela-Echavarría
avarela@unam.mx
Maribel Hernández-Rosales
maribel.hr@cinvestav.mx

Specialty section:

This article was submitted to
Evolutionary and Population Genetics,
a section of the journal
Frontiers in Genetics

Received: 15 January 2021

Accepted: 01 June 2021

Published: 09 July 2021

Citation:

Aviña-Padilla K,
Ramírez-Rafael JA,
Herrera-Oropeza GE, Muley VY,
Valdivia DI, Díaz-Valenzuela E,
García-García A, Varela-Echavarría A
and Hernández-Rosales M (2021)
Evolutionary Perspective
and Expression Analysis of Intronless
Genes Highlight the Conservation
of Their Regulatory Role.
Front. Genet. 12:654256.
doi: 10.3389/fgene.2021.654256

Katia Aviña-Padilla^{1,2}, José Antonio Ramírez-Rafael³, Gabriel Emilio Herrera-Oropeza^{1,4}, Vijaykumar Yogesh Muley¹, Dulce I. Valdivia², Erik Díaz-Valenzuela², Andrés García-García³, Alfredo Varela-Echavarría^{1*} and Maribel Hernández-Rosales^{2*}

¹ Instituto de Neurobiología, Universidad Nacional Autónoma de México, Querétaro, Mexico, ² Centro de Investigación y de Estudios Avanzados del IPN, Unidad Irapuato, Guanajuato, Mexico, ³ Centro de Física Aplicada y Tecnología Avanzada, Universidad Nacional Autónoma de México, Querétaro, Mexico, ⁴ Centre for Developmental Neurobiology, Institute of Psychiatry, Psychology, and Neuroscience, King's College London, London, United Kingdom

The structure of eukaryotic genes is generally a combination of exons interrupted by intragenic non-coding DNA regions (introns) removed by RNA splicing to generate the mature mRNA. A fraction of genes, however, comprise a single coding exon with introns in their untranslated regions or are intronless genes (IGs), lacking introns entirely. The latter code for essential proteins involved in development, growth, and cell proliferation and their expression has been proposed to be highly specialized for neuro-specific functions and linked to cancer, neuropathies, and developmental disorders. The abundant presence of introns in eukaryotic genomes is pivotal for the precise control of gene expression. Notwithstanding, IGs exempting splicing events entail a higher transcriptional fidelity, making them even more valuable for regulatory roles. This work aimed to infer the functional role and evolutionary history of IGs centered on the mouse genome. IGs consist of a subgroup of genes with one exon including coding genes, non-coding genes, and pseudogenes, which conform approximately 6% of a total of 21,527 genes. To understand their prevalence, biological relevance, and evolution, we identified and studied 1,116 IG functional proteins validating their differential expression in transcriptomic data of embryonic mouse telencephalon. Our results showed that overall expression levels of IGs are lower than those of MEGs. However, strongly up-regulated IGs include transcription factors (TFs) such as the class 3 of POU (HMG Box), *Neurog1*, *Olig1*, and *BHLHe22*, *BHLHe23*, among other essential genes including the β -cluster of protocadherins. Most striking was the finding that IG-encoded *BHLH* TFs fit the criteria to be classified as microproteins. Finally, predicted protein orthologs in other six genomes confirmed high conservation of IGs associated with regulating neural processes and with chromatin organization and epigenetic regulation in *Vertebrata*. Moreover, this study highlights that IGs are essential modulators of regulatory processes,

such as the Wnt signaling pathway and biological processes as pivotal as sensory organ developing at a transcriptional and post-translational level. Overall, our results suggest that IG proteins have specialized, prevalent, and unique biological roles and that functional divergence between IGs and MEGs is likely to be the result of specific evolutionary constraints.

Keywords: intronless genes, exon-intron architecture, embryonic telencephalon, protocadherins, histones, transcription factors, microproteins, evolutionary histories

INTRODUCTION

Most eukaryotic genes contain introns, nucleotide DNA sequences that after transcription as part of the messenger RNA are removed by splicing during its maturation. Since the introns interrupt the multiple exonic sequences, these genes are thus termed multiple exon genes (MEGs). Eukaryotic genomes, however, also contain an important proportion of genes in which the coding sequence is contained within a single exon. Diverse studies of genes of this type have been performed over the past decades and have been variously referred to as “single-exon genes” (SEGs) and “intronless genes” (IGs), both terms carrying some ambiguity as genes containing an intron in their 5' UTR are often included among them (Sunahara et al., 1990; Gentles and Karlin, 1999; Sakharkar et al., 2002, 2005a, 2006; Tine et al., 2011; Zou et al., 2011; Yan et al., 2014). For example, a recent ontology defines SEGs as nuclear genes with functional protein-coding capacity whose coding sequence comprises only one exon, thus including genes with introns in their untranslated regions termed uiSEGs, as well as those lacking introns entirely, termed “*Intronless Genes*” (Jorquera et al., 2018). Pseudogenes, functional RNAs, tRNA, rRNA, ribozyme long non-coding RNAs, and miRNAs are excluded from this definition. To avoid any possible ambiguity, in this article we use the term “*Intronless genes*” in the narrow definition of Jorquera et al. (2018) as protein-coding nuclear genes completely devoid of introns.

Owing to their prokaryotic-like architecture, IGs in eukaryotic genomes, provide interesting datasets for computational analysis in comparative genomics and for the study of evolutionary trajectories. Comparative analysis of their sequences in different genomes could allow the identification of their unique and conserved features, thus providing insights into the role of introns in gene evolution leading to a better understanding of genome architecture and arrangement.

The abundant presence of introns in most genes of multicellular organisms entails regulatory processes associated with the generation of multiple splice variants missing in intronless genes. The absence of splicing events in IGs represents a higher transcriptional fidelity, making them even more valuable for regulatory roles. To date, more than 2000 genes with a single coding exon in the human genome have been classified (Jorquera et al., 2016). Among them, a considerable fraction of IGs encode G-protein-coupled receptors (GPCRs), core canonical histones which are integral part of nucleosomes and often confer specific structural and functional features, transcription factors, proteins involved in signal transduction, regulation of

development, growth, and cell proliferation (Sakharkar et al., 2005a; Grzybowska, 2012).

The expression of IGs has been proposed to be highly specialized for neural functions and linked to diseases such as cancer, neuropathies, and developmental disorders. Examples of IGs with clinical relevance are the *RPRM* gene related to gastric cancer which causes increased cell proliferation and possesses tumor suppression activity (Amigo et al., 2018) and the protein kinase *CK2α* gene which is up-regulated in all human cancers (Hung et al., 2010). Other IGs linked to cancer include *CLDN8* in colorectal carcinoma and renal cell tumors, *ARLTS1* in melanoma, and *PURA* and *TAL2* in leukemia (Grzybowska, 2012). IGs have also been associated with neuropathies, such as *ECDRI*, a cerebellar degeneration-related protein, and *NPBWR2*, a neuropeptide B/W receptor type (Louhichi et al., 2011).

Regarding their role in the diseases described above, IGs in humans are potential clinical biomarkers and drug targets that deserve careful consideration (Ohki et al., 2000; Grzybowska, 2012; Liu et al., 2017). Their functional role and their evolutionary conservation in other genomes, however, remains poorly understood. Furthermore, there is a current debate between related theories that place the origin of introns, early or late during the evolutionary history of eukaryotes (de Souza, 2003; Fedorova and Fedorov, 2003).

With this backdrop, our work aimed to characterize the functional role of mouse IGs and to infer their evolutionary pattern across six additional vertebrate genomes. We have analyzed their expression, particularly during brain development at early embryonic stages, and their potential as transcriptional as well as post-translational modulators.

Overall, this study sheds light on the concerted role played by this peculiar group of genes and helps contrast the functional features of intron-containing and intronless genes across vertebrate species and their collective evolutionary roadmaps.

MATERIALS AND METHODS

Data Extraction and Curation for MEGs and IGs

Data were extracted using Python scripts¹ and Ensembl APIs. Seven vertebrate genomes including *Mus musculus*, *Homo sapiens*, *Pan troglodytes*, *Monodelphis domestica*, *Rattus norvegicus*, *Gallus gallus*, and *Danio rerio* assembled at a

¹https://github.com/GEmilioHO/intronless_genes

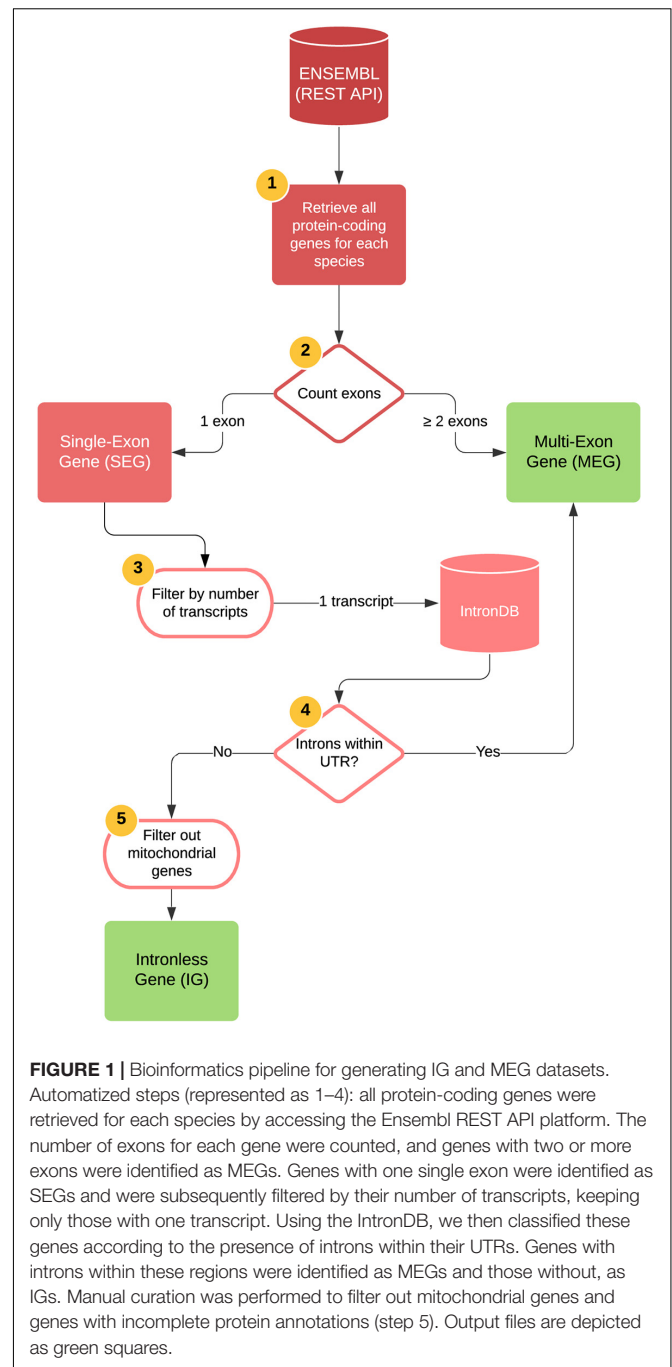
chromosome level were accessed at the Ensembl REST API platform² (using Python with the `ensembl_rest` package). For an explanation of the species choice see section “Search for Orthologs of Mouse IGs.” The pipeline process was as follows: protein-coding genes with CDS identifiers for transcripts for all chromosomes were retrieved and classified into a temporary dataset that contained genes with a single coding exon (temp-DS1) and a dataset containing “multiple exon genes” (MEGs) depending on exon and transcript count (Figure 1). The former was then submitted to the Intron DB³ to filter out genes with UTR introns. The output of the pipeline was a second temporary dataset containing protein-coding genes that did not contain introns in their entire length (temp-DS2). After data extraction, a manual curation step was performed to assess their nuclear nature and protein-coding transcript biotype, which allowed us to discard proteins encoded in the mitochondrial genome, hence yielding the final dataset containing only protein-coding nuclear genes completely devoid of introns, or “Intronless genes” (IG) (Figure 1). The final MEG dataset contained 20,694 protein-coding genes with two or more exons and the IG dataset contained 1,116 protein-coding genes with only one exon and one transcript.

Computational Prediction of Mouse Intronless Gene Function

The mouse IG and MEG datasets were used to perform an over-representation analysis of functional assignment using the following databases: SUPERFAMILY⁴ (proteins of known three-dimensional structure); Pfam⁵ (protein domains); and PROSITE⁶ (biologically meaningful signatures or motifs). All tests used MEGs as controls to determine unique, shared or overrepresented features among both types of genes. For data visualization of SUPERFAMILY and Pfam results we employed the ClusterProfiler R package (Yu et al., 2012), and Python scripts using a hypergeometric test for PROSITE enrichment.

Functional Enrichment Analysis of IG and MEG Proteins

The functional enrichment analyses were performed using Metascape⁷ for the biological process category, including KEGG and Reactome pathways. First, the functional enrichment of the 1,116 mouse IG proteins was performed using all mouse proteins as a background “universe” (selecting input as species: *M. musculus*, analysis as species: *M. musculus*). Then, in a second approach the meta-analysis workflow was used to compare enriched terms for the list of orthologs of mouse IGs regarding their grouping into five “ages,” each of them corresponding to one of the taxonomic categories: *Vertebrata*, *Tetrapoda*, *Theria*, *Eutheria*, and *Muridae*.



Out of the 1,116 IGs, 543 have orthologs across the analyzed species. Mouse IG orthologs that conserved the IG structure were compared against MEGs conserved as MEGs in other organisms. The meta-analysis workflow was used to compare enriched terms for the list of mouse orthologs in the aforementioned genomes to the pathways of three random samples of the same size of multi-exon genes to confirm that we obtained similar results. For this analysis, orthologs were clustered regarding their previously inferred “age” in five groups (selecting input as species: any species, analysis as species: *M. musculus*).

²<http://rest.ensembl.org>
³<http://www.nextgenbioinformatics.org/IntronDB/>
⁴<http://supfam.org/>
⁵<https://pfam.xfam.org/>
⁶<https://prosite.expasy.org/>
⁷<http://metascape.org/>

Finally, we performed a third approach to determine the conservation of the functional role of IGs. First, we determined the overrepresented GO terms for biological processes and molecular function of the orthologs from the seven genomes using AmiGO2⁸ were obtained. Then, GO terms with their corresponding *p*-values were clusterized using REVIGO which finds a representative subset of the terms using an algorithm that relies on semantic similarity measures (Supek et al., 2011).

Data Source and Differential Expression Analysis

Read counts from a previous transcriptomic analysis of mouse embryonic telencephalon were used to identify differentially expressed genes (Muley et al., 2020). The transcriptomes were obtained using the Illumina HiSeq RNA sequencing (RNA-seq) platform. The procedure for read-counts normalization, and to calculate differential expression analysis is described in <https://data.mendeley.com/datasets/rdt5757cbw/1>. A gene was considered expressed if its count-per-million (CPM) value was above 5.66-7. Mouse IG and MEG datasets were submitted to analysis to determine the directionality of the change in expression at developmental stage A (E.9.5) compared to stage B (E.10.5). Genes having significant *p*-values with positive log2-fold change represent an increased expression (UP), those with negative log2-fold change are considered with decreased expression (DN), while gene expression with *p*-values above 0.05 represents no change between stages (NC), and read-count lower than five in less than four samples out of eight are considered not expressed (NE).

Functional Enrichment Analysis of Differentially Expressed IGs and MEGs

The functional enrichment was assessed using the overrepresentation analysis of the functional assignment. Genes with differential expression up to two log2-fold change values were considered as up-regulated with a *p*- and *q*-value set at 0.05 and 0.10, respectively. The ClusterProfiler R package (Yu et al., 2012) was employed for data analysis and visualization.

Post-translational Modifications and Regulatory Assignment of IG Proteins

For post-translational modification assignments of IG and MEG proteins, the dbPTM⁹ was used. A two proportion *Z*-test was used to assess whether the proportions of each post-translational modification among IG and MEG proteins were similar. The *p*-value was set at 0.05. When the resulting *p*-value was not significant, meaning that the proportions of IG and MEG proteins were similar for a specific post-translational modification, this was classified as “similar.” On the other hand, when the resulting *p*-value was <0.05 the post-translational modification was classified as more abundant in “IG” or “MEG” depending on which one had a higher relative percentage of such modification. Post-translational modifications exclusive of either

IG or MEG proteins were classified as “unique.” Then, using the miPFinder program¹⁰, we determined the mouse gene candidates for IG-encoding microproteins.

Search for Orthologs of Mouse IGs

Mouse peptide sequences were submitted to Proteinortho (Lechner et al., 2011, 2014) to infer orthologous genes in the genomes of rat, human, chimp, opossum, chicken, and zebrafish. As a first step, Proteinortho performs sequence comparison between each pair of genomes and reports best bidirectional hits (BBHs) for alignments with equal or above fifty percent of sequence identity. In a second step, it represents each gene or protein as a node of a graph and places an edge between two genes if they were identified as a BBH, then, it applies a clustering algorithm and finally reports orthogroups and the orthology relations as pairs of genes in two different genomes.

Each of the species used in this study is a model organism of a different taxonomic level, and therefore, the conservation of mouse orthologs in close or distant related species reveals the “age” of the gene. The conservation was measured by gradually including more species, and the resulting groups were labeled with the name of the largest taxonomical category that includes all species within a group. Therefore, for orthologs of mouse IG genes that were identified in rat, they are said to be conserved in *Muridae*; those present in *Muridae*, in human, and in chimp are said to be conserved in the group *Eutheria*; those found in all the previous species and in opossum as well are conserved in the group *Theria*; those also present in chick are conserved in *Tetrapoda*; and finally, those also conserved in zebrafish are conserved in *Vertebrata*. This classification, however, is only used to refer to the conservation among the species analyzed in the present study.

Reconstruction of the Evolutionary History of Mouse IGs and Their Conservation in Other Organisms

From the ProteinOrtho predictions, orthology graphs were constructed, and an in-house developed method for the evolutionary reconstruction of gene families was used. This method implements the theory reported in Hernandez-Rosales et al. (2012) and Hellmuth et al. (2013), and it can be found at <https://gitlab.com/jarr.tecn/revolutionh-tl>. This tool starts by performing a modular decomposition (Tedder et al., 2008) on orthology graphs and then inferring the corresponding gene trees. Each internal node of these trees represents an evolutionary event: duplication or speciation. Subsequently, the gene trees are reconciled with the species tree to determine in which branch of the species tree duplication events occur and, at the same time, infer gene losses. This method allows us to infer how ancestral a gene is, determined by its orthologs in the other species, as well as to identify species-specific genes. Finally, we identified the orthologs of the 1,116 mouse IG proteins that were conserved as IGs, or that were identified as MEGs in the abovementioned genomes.

⁸<http://amigo.geneontology.org/amigo/landing>

⁹<http://dbptm.mbc.nctu.edu.tw>

¹⁰<https://github.com/DaStraub/miPFinder>

Syntenic Conservation of the β -Protocadherin Cluster

To determine the syntenic conservation of the mouse protocadherin IG members of the beta cluster across the selected genomes, the genomic coordinates of orthologs genes were retrieved from GTF files employing custom R scripts and plotted using the genoPlotR R package.

RESULTS

Functional Assignment of Protein Coding IGs in the Mouse Genome

The origin of IGs has been explained mostly by retrotransposition, which occurs by homologous recombination between the genomic copy of a gene and an intronless cDNA (Kaessmann et al., 2009). Mouse IGs represent 6% of the total number of one-exon genes, while retrotransposed single-exon pseudogenes with lost molecular function constitute almost half of them (coding, non-coding, pseudogenes) (**Supplementary Figure 1**).

Computational analysis was performed to identify mouse IGs. Then, based on the comparative annotation of IG and MEG datasets, a study was performed to identify their unique and shared molecular and biological features.

The grouping of IGs by protein domains that have an evolutionary relationship (SUPERFAMILY database) revealed a higher enrichment of the histone fold, 4-helical cytokine family of signal transducers, and transcription factor families including the Poxvirus and Zinc finger (POZ) domain, “Winged helix” DNA-binding domain, High Mobility Box group (HMG-box), and A DNA-binding domain in eukaryotes, as well as the transmembrane protein families Cadherin-like, and Frizzled cysteine-rich domain. MEG-encoded proteins, in contrast, are enriched in protein kinase-like, immunoglobulin, Krüppel associated box (KRAB) domain, and Armadillo repeat motifs (ARM) repeat families. The top enriched structural families of IG and MEG groups are shown in **Supplementary Figure 2**.

The analysis of the conserved functional domains (Pfam database) among the enriched protein families encoded by mouse IGs, revealed 598 hits of three main classes: 253 were transmembrane protein receptors, 101 core histones, 84 transcription factors, and 160 that belong to other classes (**Figure 2A**). Among the transmembrane protein receptors, the most enriched domains were Taste 2 receptors (TAS2R), Vomer-nasal 1 receptors (V1R), and seven transmembrane group 1 (7tm_1), common in GPCR and vomeronasal receptors (**Figure 2B**). Other domains identified were cadherin, Pheripheral myelin protein 22 (PMP22-claudin), (Desintegrin and metalloproteinase domain (ADAM), Disintegrin, and Frizzled (FZ). In the transcription factor group, Broad-Complex, tramtrack, and bric-à-brac (BTB), Myb DNA-binding, HMG-Box, and forkhead were enriched protein domains in the mouse IGs compared to MEGs (**Figure 2C**). Meanwhile, in the histone group, four domains were observed: Histone, Histone H2A 1363

C-terminal (H2AC), Centromere kinetochore component CENP-T histone (CENPTC), and Linker histone (**Figure 2D**). Finally, in the other groups we found among others Keratin, Interferon, Ubiquitin, Actin, and FYTT enriched domains (**Figure 2E**). The classification of IGs in functional groups mentioned above was then used for the transcriptional analysis (**Figure 2**).

Analysis of biologically significant motifs (PROSITE database) among MEG and IG proteins identified a total of 1,239 (12,546 hits) and 144 (634 hits) distinct protein signatures, respectively. Interestingly, among the most abundant motifs in the mouse IGs are GPCR, leucine-rich repeat, histone, transcription factor Forkhead domain, Myc-type basic helix-loop-helix (bhlh) motifs, ankyrins, and cadherin domains which were found to be infrequent in MEG proteins (**Figure 3A**). It is noteworthy, however, that among the top motifs that were unique to IG proteins H2B signature was the largest group (**Figure 3B**). Hence, these results show that most of the top predictions of IGs signatures are characteristic of transmembrane receptors, histones, and specific transcription factors, having a unique signature for histones.

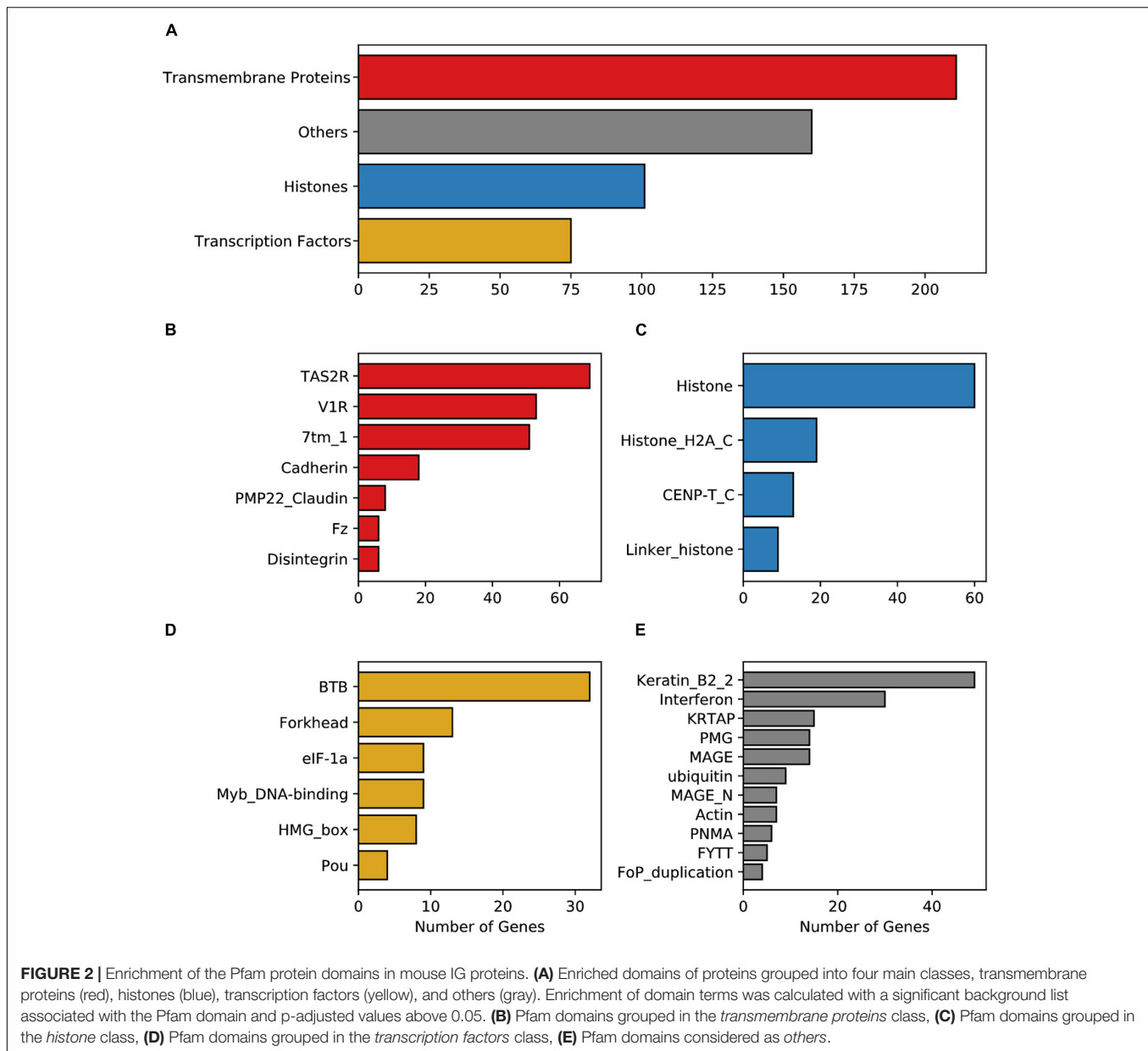
Finally, the functional enrichment of mouse IGs revealed biological pathways associated with genetic and protein regulatory processes including detection of chemical stimulus involved in sensory perception of the bitter taste, chromatin silencing, positive regulation of peptidyl-serine phosphorylation of STAT proteins, and nucleosome positioning ($-\log_{10} 34.23 > -3.27$). Other functions detected were immune, neuro-specific, and development processes such as mmu05322-Systemic lupus erythematosus, R-MMU-6805567 Keratinization, R-MMU-500792 GPCR ligand binding, R-MMU-1266695 Interleukin 7 signaling, hard palate development, and noradrenergic neuron differentiation ($-\log_{10} 29.94 > -3.75$) (**Figure 4**).

Altogether, functional assignment analysis suggests that IGs have distinct biological roles in comparison to MEGs.

Up-Regulation of IGs Reveal Their Regulatory Role on Neural Functions Through Mouse Development

Previous studies detected enrichment of neural-related functions among IGs (Grzybowska, 2012). Moreover, since the expression of MEGs is modulated by the balance between the rate of transcription elongation and the alternative splicing of exons (Fong et al., 2014), we hypothesized that the natural absence of splicing on IG mRNAs could confer them differential regulatory roles in complex biological processes. Therefore, it was our interest to identify and analyze IGs that are expressed in mice during brain development. For that purpose, we analyzed expression data from the developing mouse telencephalon at stages in which its patterning is taking place (E9.5 and E10.5) (Muley et al., 2020).

Overall, the expression of IGs was lower than that of MEGs (**Figure 5A**), which is consistent with previous *in silico* observations (Sakharkar et al., 2005a). Out of 1,116 transcripts, differential expression analysis was performed for 1,087, with 37 of them (3.4%) showing up-regulation and nine down-regulation (0.82%) from gestational day E9.5 to E10.5. Moreover,

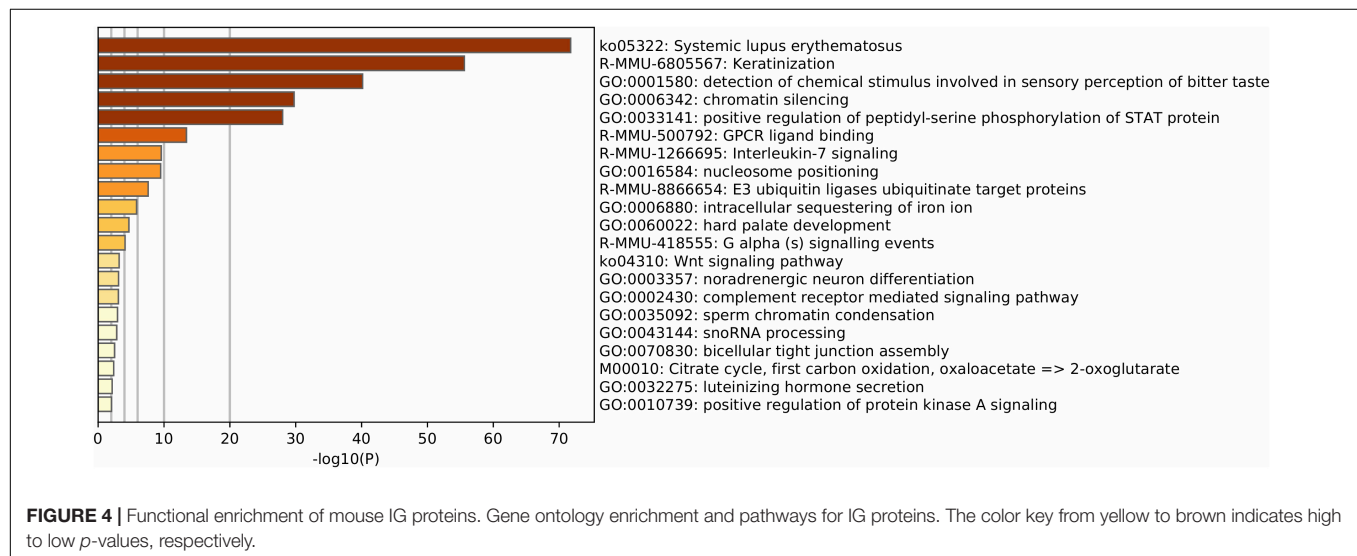
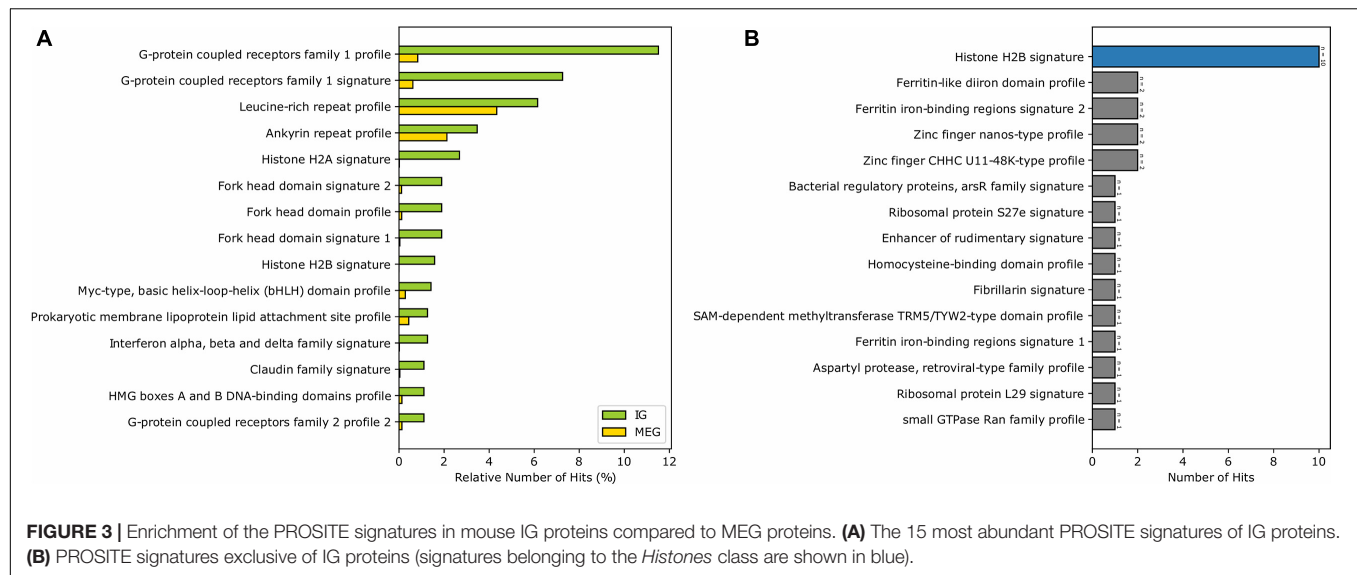


387 (35.63%) did not change expression, and 653 (60.12%) were not expressed during the analyzed stages (**Figure 5B**). Meanwhile, among MEGs, 1247 were up-regulated (6.13%), 789 were down-regulated (3.88%), 13,198 had no expression changes (64.93%), and 5090 did not show expression (25.04%) (**Figure 5B**). It is noteworthy that an inverse expression pattern of genes with no expression changes (a higher percentage of MEGs than of IGs) and those not expressed (a higher percentage of IGs than of MEGs) was found in this comparison.

Our analysis revealed that all up-regulated IGs are exclusively enriched in biological pathways in eye and sensory organ development processes compared to MEGs also involved in other developmental and neural function pathways (**Figures 5C,D**). Moreover, significantly enriched terms in molecular function found for up-regulated IGs are consistent with their regulatory

role, including rRNA methyltransferase and DNA-binding transcription repressor activities. In contrast, in the up-regulated MEGs the molecular function terms are highly enriched for transmembrane transporters and channel voltage activities (**Figure 5D**).

From the IG transmembrane protein group, transcripts for *Tram111*, *Cdk5r2*, *Nrarp*, *Kcnf1*, *Fzd7*, *Fzd8*, *Fzd10*, and *Cldn5* were up-regulated. Strikingly, from the cluster of 22 β -protocadherins (*pcdhbs*), which contains 18 IGs, 9 of these were among those up-regulated in the E10.5 telencephalon (*Pcdh3*, *Pcdh4*, *Pcdh7*, *Pcdh10*, *Pcdh11*, *Pcdh17*, *Pcdh19*, *Pcdh20*, and *Pcdh21*) (**Figures 6A, 7A**). It is important to note that all but one of all protocadherins of this cluster were expressed in the developing telencephalon. Our expression analysis additionally revealed up-regulation of *Olig1*, *Bhlhe22*,

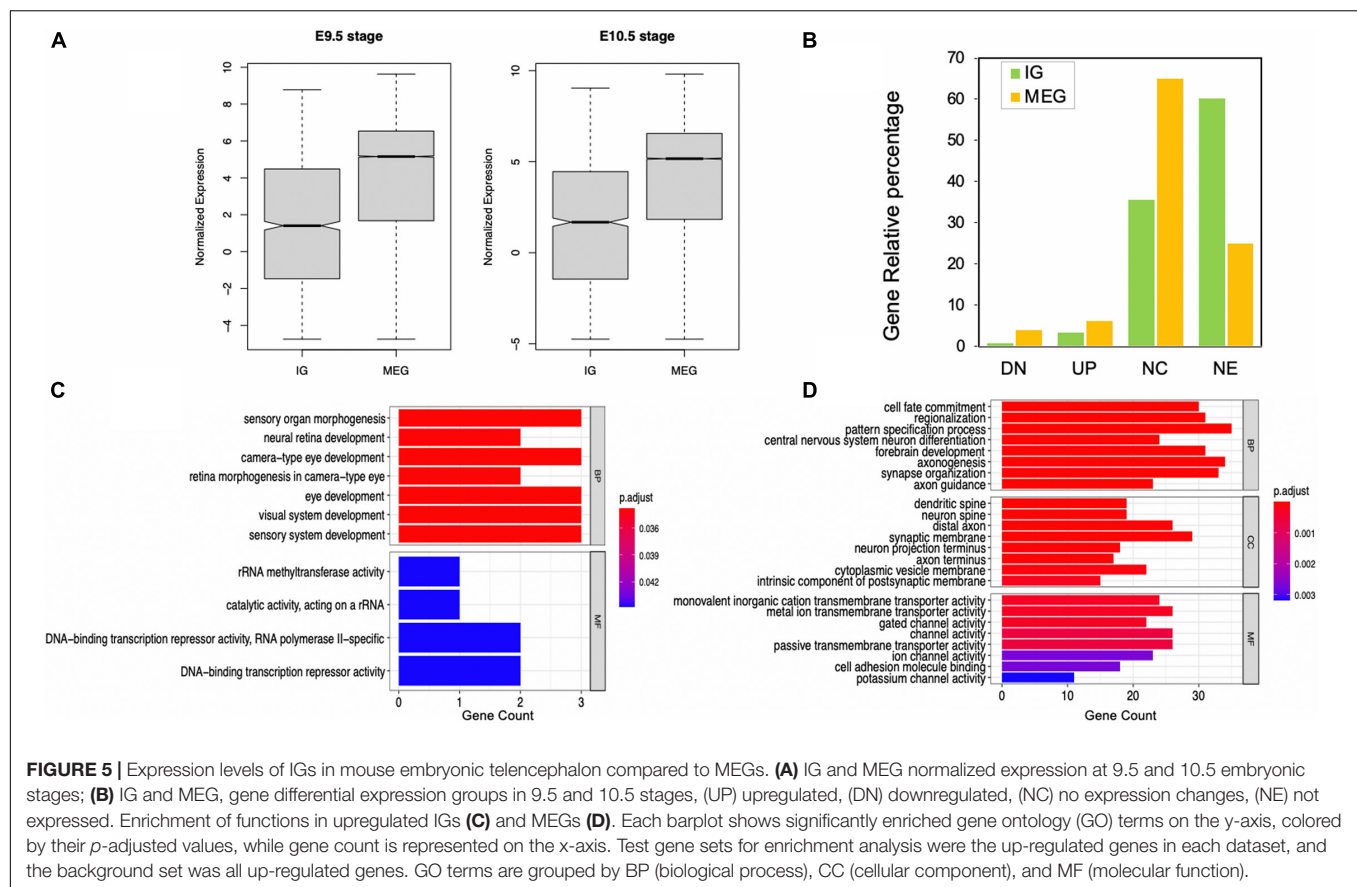


Bhlhe23, *Pou3f1*, *Pou3f2*, *Pou3f4*, *Foxq1*, and *Neurog1*, most of which are BHLH transcription factors crucial for the regulation of brain development and neuro-specific functions (Figure 6B). Moreover, regarding IGs within the histone group, *H2bc21*, *H2bu2*, *H2aw* were up-regulated during the mouse embryonic stages (Figure 6C). Finally, IGs from other groups with up-regulation were also observed (Figure 6D).

The β -Protocadherin Gene Cluster Displays a High Degree of Syntenic Conservation Across Mammalian Genomes

To gain further insight on the evolutionary conservation of IGs with a functional role in telencephalon patterning, we studied the syntenic conservation of the ortholog genes of the 18 mouse IG β -protocadherins (Figure 7A) in our set of seven species. We

determined that human, chimp, rat, opossum, and chick contain orthologs of the mouse single exon *pcdhh* genes, which are absent in zebrafish. Overall, all the orthologous genes of the cluster are located in a single locus in their respective genomes, with varying lengths ranging from ~128 to 310 kb, displaying a few local inversions (Figure 7B). These results are consistent with previous studies that have explored the syntenic conservation of *pcdhh* genes across other vertebrate species (Noonan et al., 2004; Yu et al., 2008). Even though we found syntenic conservation of some members of the β -protocadherin cluster, we observed slight disruptions of the order of genes due to gene expansions, which can be either gene duplications or *de novo* formation. These gains are most notorious in the mammalian genomes (Figure 7B), suggesting gene expansion of the intronless β -protocadherins could be relevant for their role in neurogenesis, as well as other neuro-specific functions associated with the Wnt canonical pathway.



We looked in more detail at the evolutionary histories of β -protocadherins, by reconciling the gene trees of this gene cluster with the taxonomic species tree (Figure 7C). First, we observed that none of these genes is conserved in zebrafish. Moreover, some genes are gained in specific lineages, for example, *Pcdhb17* is only observed in mice and rats, while eight genes are shared across the mammals in the study. Three β -protocadherins appear to be shared among primates and the marsupial opossum, while only *Pcdhb7*, *Pcdhb15*, and *Pcdhb19* are shared between mouse and chick, and across other intermediate species, suggesting that these are the oldest β -protocadherins that give origin to the rest of them.

Then, by assessing syntenic conservation in a pair-wise fashion, we found relevant lineage-specific gene losses (Figure 7D). For instance, *Pcdhb18* is absent in rats while it is present in mice and duplicated in primates. This evidence suggests that the complexity of nervous system characteristic of mammals could also be associated with the duplication of single exon genes besides splicing-derived protein isoform diversity.

Characterization and Functional Role of Post-translational Modifications in Mouse IG Proteins

In addition to alternative splicing, and mRNA editing, post-translational modifications (PTMs) constitute a defining factor

of the complexity of proteomes by increasing structural and functional diversity of each proteoform, the set of multiple protein molecules encoded by one gene. Hence, protein PTMs have an essential role in protein structure-function, including activity, stability, folding, and turnover (Uversky, 2015). Since IGs fit the “one gene—one protein” concept we aimed to determine whether PTMs represent exclusive mechanisms of regulation for these genes. In our analysis, we observed that Succinylation and S-nitrosylation had similar prevalence in IG and MEG groups. These were followed with much lower frequency by Glutathionylation, Glutarylation, Palmitoylation, and Oxidation (Figure 8A). In contrast, PTMs with a unique presence in MEGs were Nitration, Myristoylation, Sulfation, Carboxylation, GPI-anchor, and Pyrrolidone carboxylic acid (Figure 8B).

Notably, in accordance with their protein assignment, a differential enrichment for IG proteins was observed of Acetylation, Crotonylation, Methylation, Malonylation, and Hydroxylation (Figure 8C) which are characteristic features of the core histones. Other PTMs of IG proteins, albeit at lower frequency, were Citrullination, Sumoylation, and Amidation (Figure 8C). The complementary group of PTMs more enriched in MEG proteins included Phosphorylation, followed by O-linked Glycosylation, Ubiquitination, and N-linked Glycosylation (Figure 8D). PTMs classified based on the modification-enabled functionality for membrane localization

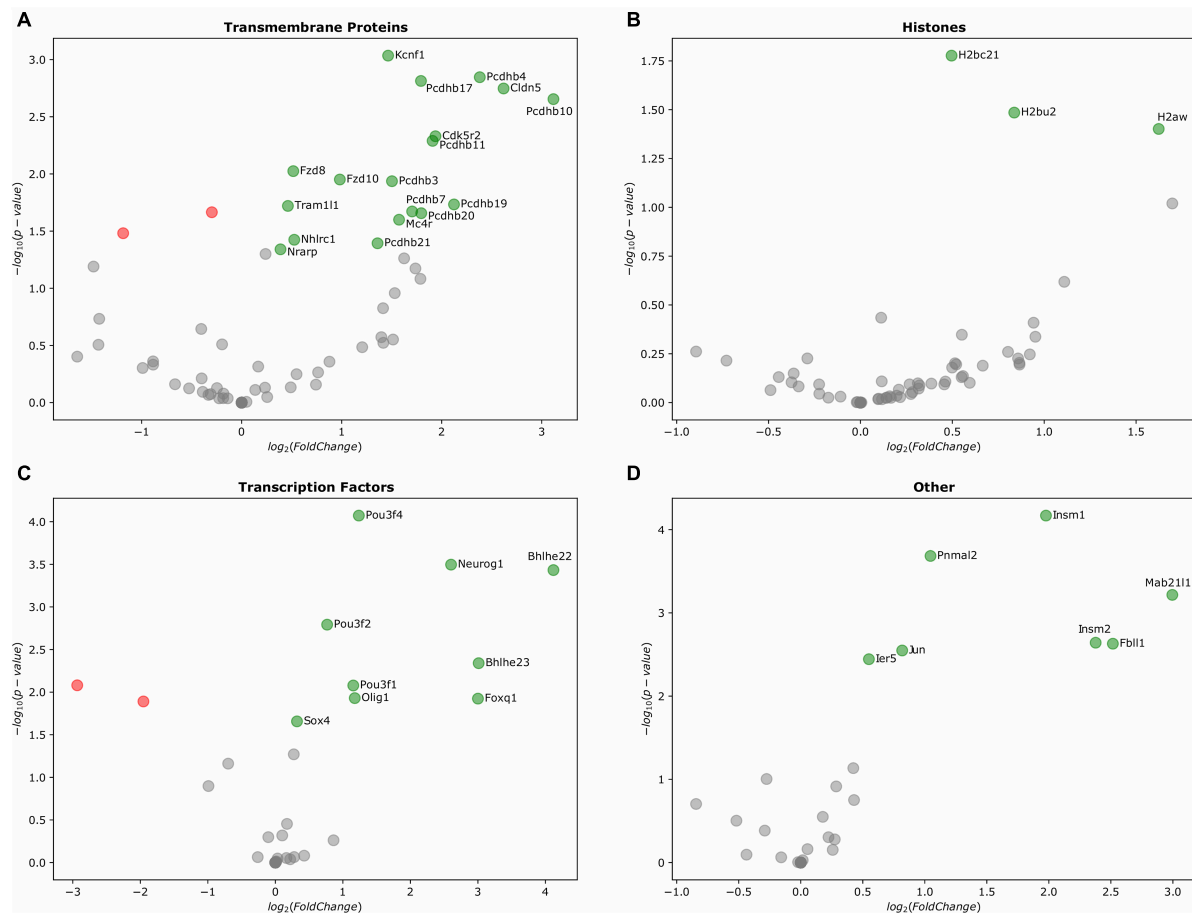


FIGURE 6 | Differentially expressed IGs in the mouse embryonic telencephalon. Expression of genes grouped by their Pfam assignment, determined by their log2-fold change values. Up-regulated genes are highlighted in green, while downregulated genes are colored in red. **(A)** Gene expression of transmembrane proteins, **(B)** histones, **(C)** transcription factors, and **(D)** other protein families.

such as Myristoylation and GPI-anchor were found to be more frequent for MEGs (**Figure 8D**).

Functional Assignment of IGs as Microproteins

The regulation of multidomain proteins at the post-translational level can be mediated by microproteins (miPs) (Staudt and Wenkel, 2011) which are small proteins containing a single domain that form heterodimers with their targets and exert dominant-negative regulatory effects (de Klein et al., 2015; Eguen et al., 2015). In *Eukarya*, microproteins have been found to have a remarkable influence on diverse biological processes.

Aware of the differential occurrence of PTMs on IG proteins, the DNA binding repressor activity molecular function of up-regulated IGs during mouse brain development, and due to the remarkable role of miPs, we assessed whether this group of genes encoded proteins fitting the miP definition. Characteristic features of miPs are the short length of their primary structure, a homodimer domain, and negative modulating activity of protein multi-complexes. Our first approach was to analyze the

peptide length of IG and MEG proteins. The highest length-frequency for IG peptides was in the range of 200–400 amino acids, compared to that of MEG peptides which was 300–500. Then, using the miPfinder tool (Straub and Wenkel, 2017), we identified the following IGs as microprotein candidates: the BHLH transcription factors *Bhlha9*, *Msg1*, *Ferd3l*, *Bhlhe23*, and *Ascl5* (*e*-value 4.6E–30), as well as the histones *H1f0*, *H1f1*, *Hils1*, *H1f2* *H1f3*, *H1f4*, *H1f5*, *H1f6*, and *H1f10* (*e*-value 7.4E–09), corresponding to the H1 linker histone group.

Conservation and Evolution of IGs Across Vertebrata

To infer the evolutionary age of genes we implemented a bioinformatics method to assess the extent and patterns of distribution of each gene's orthologs and paralogs in different species. The rationale of this approach is that widespread conservation of the orthologs of a gene in the different vertebrate taxa is an indication of old age for that particular gene. This approach allowed us to determine the conservation of IGs across 7 genomes, as well as to identify species-specific mouse IGs

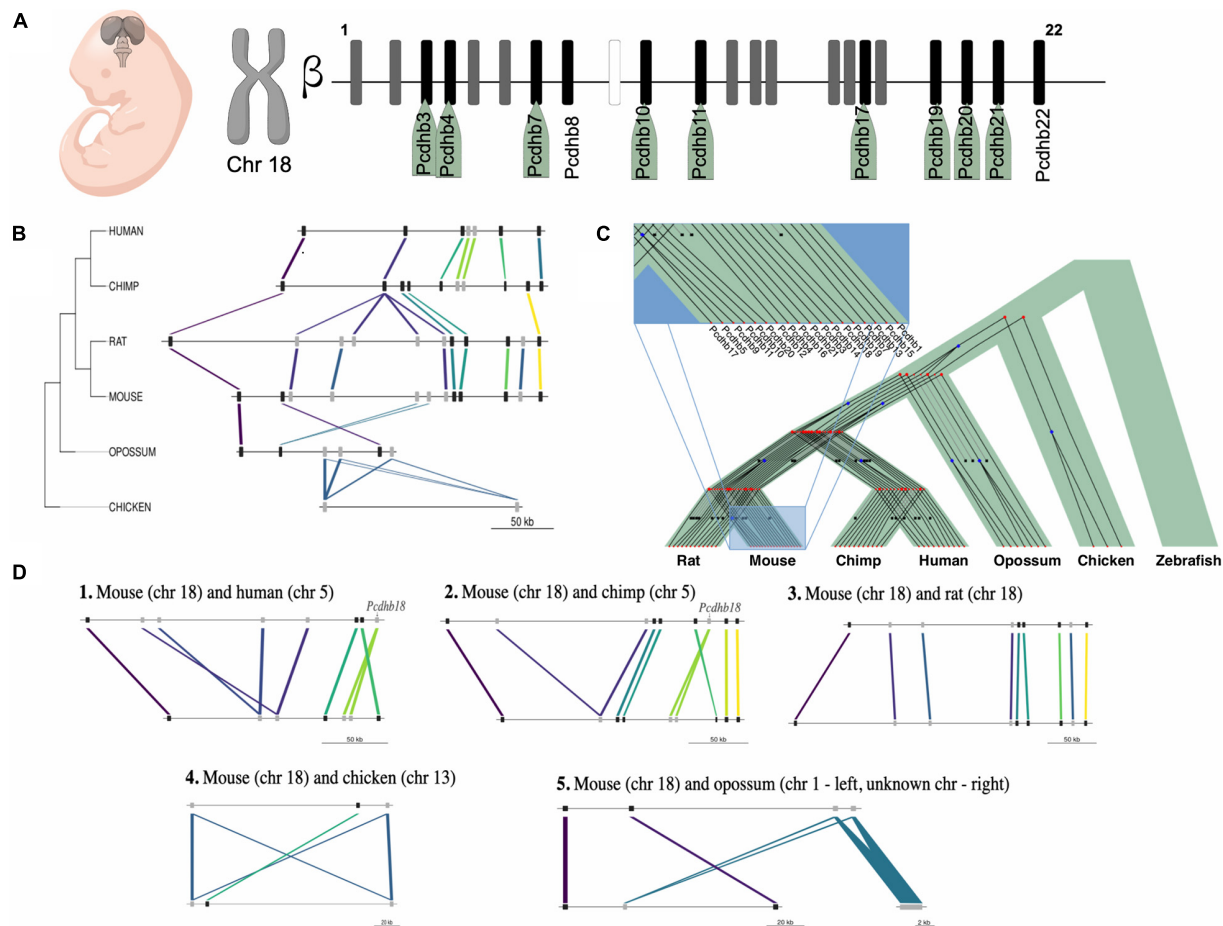
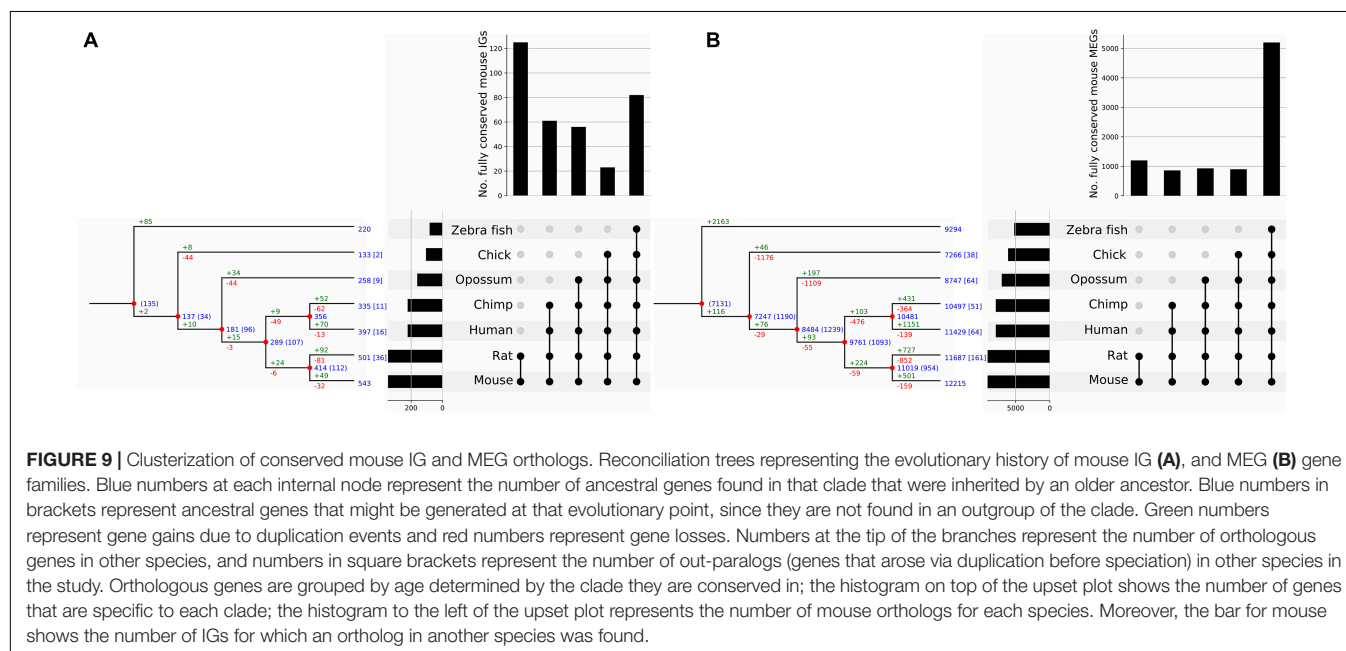
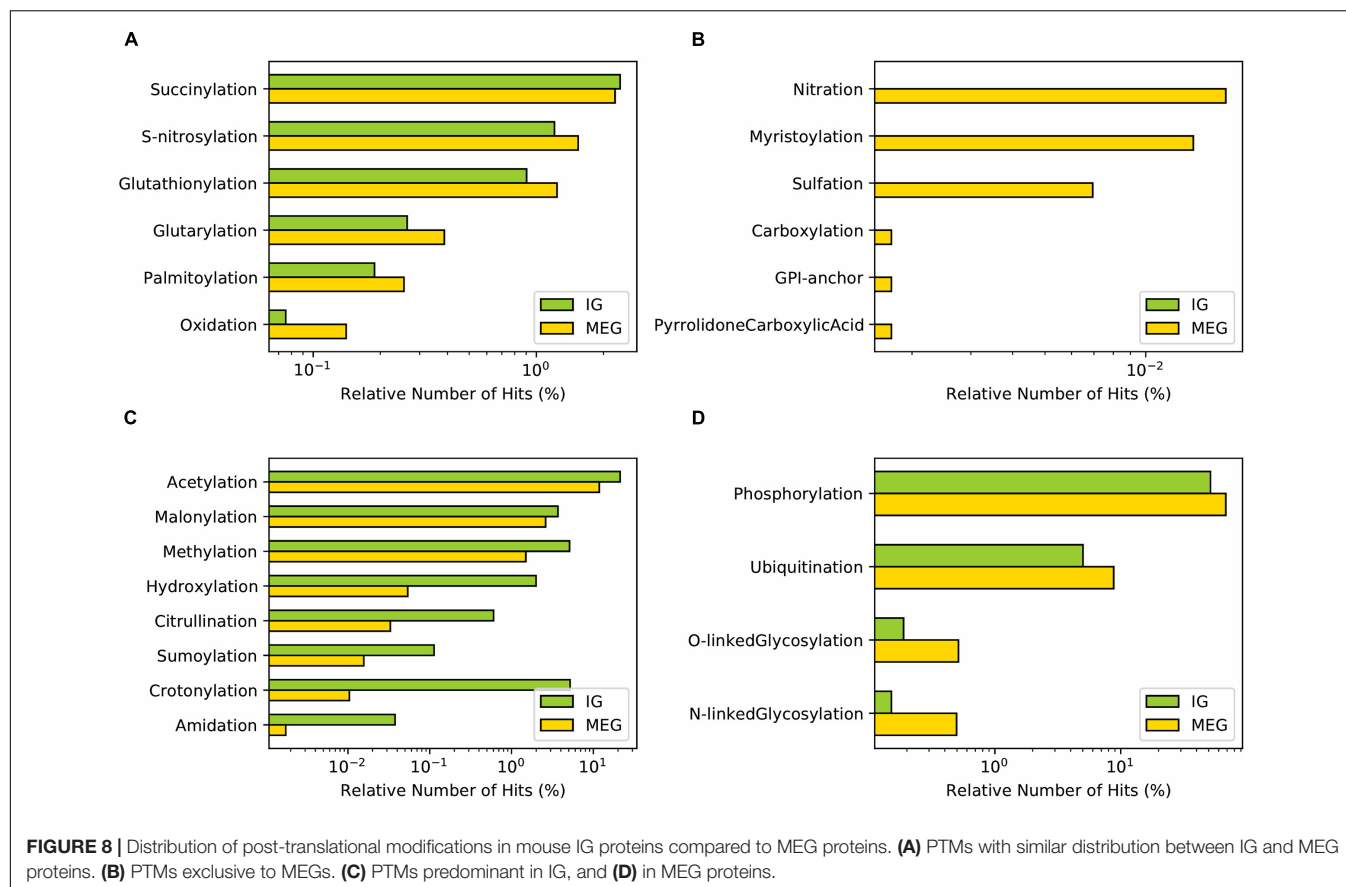


FIGURE 7 | Expression and evolution analysis of the mouse cluster of β -protocadherins. **(A)** Cluster of the 22 β -protocadherins in the chromosome 18 depicting in black up-regulated genes (names highlighted in green) in telencephalon between embryonic stages 9.5 and 10.5, in gray expressed genes with no relative changes between stages, and in white not expressed genes; **(B)** Syntenic map of the approximately 300 kb β -protocadherin locus across five mammalian lineages and chicken as an outgroup. Colored lines depict orthology relationships across the phylogenetic tree. Genes connected by the same color belong to the same orthogroup identified by ProteinOrtho. Genes are shown as black squares for single-copy orthologs and gray squares for expanded genes; **(C)** Reconciliation tree of the protocadherin gene trees and the species tree. Each gene tree represents an orthogroup, internal nodes represent evolutionary events (blue squares represent duplications, red bullets represent speciations) and black crosses in leaves represent inferred gene loss events; **(D)** Pairwise syntenic comparisons of the mouse β -protocadherin locus to four mammalian genomes and chick highlighting the lineage-specific loss and expansions of the *Pcdhb18* gene.

(Figure 9). In this analysis, we found that 543 out of the 1,116 mouse IGs have orthologs in at least one of the other species. For the mammalian genomes, we found 442 genes conserved as IGs out of 501 orthologs in the rat genome, 335 orthologs in chimp with 250 conserved as IGs, 397 with 262 IGs in human, and 258 with 167 IGs in opossum (Table 1). Meanwhile, we found 133 orthologs in chick with 78 conserved as IGs, and 220 in zebrafish with 91 conserved as IGs (Table 1). We also identified out-paralogs of mouse IGs (genes that arose via duplication before a speciation) that are conserved in the other species: 36 in rat, 16 in human, 11 in chimp, 9 in opossum, 2 in chick and none in zebrafish. Finally, we identified 573 IGs with no orthologs in the other species, suggesting that these are species-specific mouse IGs.

Overall, we found that 70% of the IG orthologs are IGs as well, and 30% are MEGs (Table 1). As for MEGs (Figure 9),

less than 5% of their orthologs are IGs and the rest are MEGs (Tables 1, 2). As expected, due to its evolutionary closeness with the mouse, the genome with the highest conservation in gene architecture is the rat, with approximately 88% of conserved IGs, while the largest difference was found for the zebrafish genome with only 41% of conserved IGs. Furthermore, at the superfamily level as identified by SUPERFAMILY, we found that 24% of the IG superfamilies are conserved as IG-only, while 76% are predominantly IGs but contain at least one MEG ortholog in another species. Similarly, for MEG superfamilies, 35% were conserved as MEG-only, while 65% were predominantly MEGs with at least one IG ortholog. Hence, these analyses revealed that most of the IGs identified in the mouse genome remained with this genetic structure in other species thus supporting their high conservation across vertebrate genomes.



From the previous analysis we clusterized IGs and MEGs into five age-groups named by the taxonomic category that includes all the species of each group and thus represents the most recent common ancestor (MRCA) of each ortholog as

inferred from the extant species analyzed. These groups were: *Vertebrata*, *Tetrapoda*, *Theria*, *Eutheria*, and *Muridae* (Figure 9). From the reconstruction of the evolutionary history of the mouse IGs, our results revealed that their conservation is more marked

TABLE 1 | Summary of mouse IG orthologs in selected genomes.

Genome	IGs	MEGs	Others	Total
Zebrafish	91	90	39	220
Chick	78	55	0	133
Opposum	167	87	4	258
Chimp	250	59	26	335
Human	262	97	38	397
Rat	442	57	2	501

TABLE 2 | Summary of mouse MEG orthologs in selected genomes.

Genome	IGs	MEGs	Others	Total
Zebrafish	108	7,602	1,584	9,294
Chick	122	7,137	7	7,266
Opposum	159	10,304	966	11,429
Chimp	523	8,040	184	8,747
Human	603	9,520	374	10,497
Rat	1,171	10,457	59	11,687

in the *Muridae* as it contains the largest number of orthologs common to its members (**Figure 9A**) followed in abundance by *Vertebrata*. This indicates that a large number of IGs are sufficiently old to have orthologs in all the vertebrates analyzed, and that the clades that include the closest relatives to mice have increasing IG ortholog abundance. In contrast, the highest conservation of MEGs in gene numbers is among *Vertebrata* thereby revealing a much older age than that of *Muridae* IGs (**Figure 9B**). For both IG and MEG orthologs, the number of paralog-related genes increases with gene age consistently with the rate of duplication of the edges of each clade (**Figure 9**). Moreover, a significant number of in-paralogous genes in the zebrafish genome, generated via duplication after speciation, have an ortholog in the mouse genome.

With the purpose of determining whether there was a differential functional enrichment of IGs according to their evolutionary age, we analyzed the enrichment of molecular pathway GO terms in both IGs and MEGs. In agreement with a specialized role of IGs, our results show that IG and MEG orthologs are involved in different biological pathways although some shared pathways were detected as well (**Figure 10**). Conserved IG proteins with the MRCA among *Vertebrata* are histones highly enriched in negative regulation of megakaryocyte differentiation ($-\log_{10}$, -20.82). Other orthologs conserved to this group are linked in a lower level to thermogenesis, basal cell carcinoma, positive regulation of protein kinase A signaling, ribosomal large subunit assembly, wound healing, and vascular process in the circulatory system, platelet aggregation and development process such as cell-fate specification, negative regulation of animal organ morphogenesis, pituitary gland development, and regulation of bicellular tight junction assembly ($-\log_{10}$, $-9.33 > -2.81$). For *Theria* we found G alpha signaling events ($-\log_{10}$, -8.86), while in the *Eutheria* group peptidyl-serine phosphorylation of STAT protein, and chromatin silencing were enriched ($-\log_{10}$, -6.47 ; -5.69). Noticeably, the most recent genes which belong to the *Muridae* group are exclusively enriched

in intracellular sequestering of iron, complement receptor-mediated signaling pathway, and histone deubiquitination ($-\log_{10}$, $-8.64 > -3.63$).

Our analysis also identified IG proteins with enriched pathways shared among the various age groups. Detection of chemical stimulus involved in sensory perception of bitter taste, and keratinization are GO terms shared among *Muridae* ($-\log_{10}$, -22.91 ; -7.07), *Eutheria* ($-\log_{10}$, -13.36 ; -2.10) and *Theria* ($-\log_{10}$, -12.37 ; -2.39). Meanwhile, GPCR ligand binding is shared among *Vertebrata* ($-\log_{10}$, -5.59), *Theria* ($-\log_{10}$, -6.11), and *Eutheria* ($-\log_{10}$, -3.05) (**Figure 10A**).

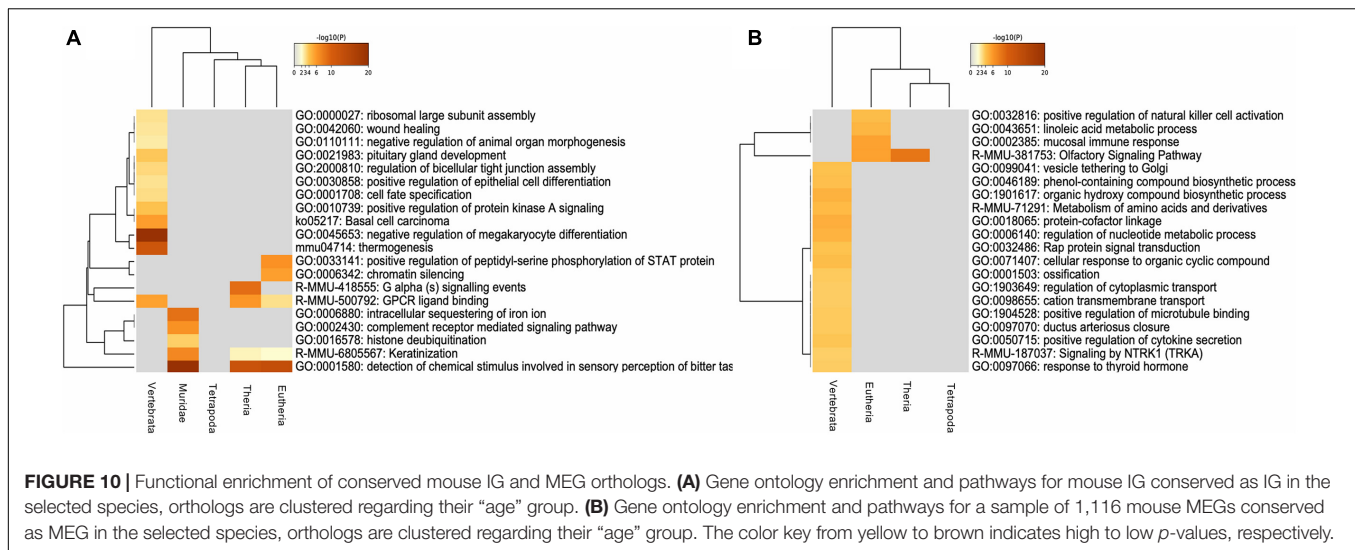
Then we focused on determining the conservation of the biological role of IG orthologs among the different genomes. GO terms that are highly enriched in the seven genomes analyzed were detection of chemical stimulus involved in sensory perception of smell, organic substance metabolism, DNA packaging, signaling, multicellular organismal process, cell communication, transport, and localization, while GO molecular function terms enriched are olfactory receptor activity, Wnt protein binding, odorant binding, protein binding, catalytic, and molecular transducer activity.

DISCUSSION

The mechanism of alternative splicing is a pivotal contributor to the diversity of proteins and the functional complexity of eukaryotic genomes. Intron-containing genes are capable of generating multiple protein isoforms by this process by which exons can be removed, lengthened, or shortened (Sakharkar et al., 2005b). In contrast, protein coding genes lacking introns produce a single peptide of predictable sequence which may undergo posttranslational fine tuning. The availability of detailed annotation of sequenced genomes for many organisms contributes toward a better understanding of their structure which has been shaped by flexible evolutionary pressure (Bult et al., 2019). Studying the evolutionary dynamics of exon-intron patterns at the genomic level is likely to shed light on their role in genome structure and gene architecture.

To further our insight into the structure and the evolution of mouse IGs, we examined their function, differential expression in the developing brain, the signatures for post-translational modifications of their encoded proteins, their potential as modulators of multiprotein complexes, as well as their evolutionary dynamics in comparison to their orthologs in other vertebrates.

Our work revealed that, in accordance with previous studies, IGs and MEGs appear to specialize in different functions which is supported by their enrichment in distinct biological pathways and differential abundance of post-translational modifications. As an additional indication of this specialization, IGs that are up-regulated during the development of the mouse telencephalon, are associated with specific developmental programs in this structure and display a functional enrichment profile that differs from that of up-regulated MEGs. Moreover, mouse IGs, some of which fit the criteria to be defined as regulatory microproteins, appear to be of more recent origin than MEGs in vertebrates.



Consistent with this notion, about half of IGs do not appear to have orthologs in other genomes thus suggesting a relevant role in mouse evolution. The synteny of the β -protocadherins, however, points out to a mammalian conserved function of these IGs although some species-specific changes in this gene cluster were observed for the various species analyzed.

Functional Assignment of IGs Highlights Prevalent and Unique Biological Roles

In the mouse genome, coding genes are predominantly of MEG type¹¹ (80% of a total of 22,481). However, a considerable number of IGs are present in this genome, and the conservation of this fraction has been reported for other mammalian genomes (Sakharkar et al., 2006).

Our comparative analysis of the types of proteins encoded by IGs and MEGs revealed that these two populations have very divergent functional profiles. The most abundant types of proteins found among the former were the chromatin components histones and centromere proteins, transmembrane proteins of the G-protein coupled receptor family 1 and cadherins, and transcription factors containing BTB, forkhead, and HLH domains. In contrast, among MEGs, the most abundant proteins were those containing zinc finger, Pkinase (Protein-kinase), and PH (Pleckstrin homology) domains. These observations are consistent with previous findings that IGs are highly enriched in GPCRs, and seven transmembrane domain proteins and reveal a functional divergence between IGs and MEGs likely to be the result of differential evolutionary constraints (Sakharkar et al., 2006).

Among the transmembrane IG proteins, vomeronasal and taste receptors stand out as the most abundant. These proteins play a highly relevant role in chemoreception which is the most salient means of the interaction of *Muridae* with their environment, with conspecifics, with potential prey or predators. These findings suggest that some olfaction and taste receptors are required to be constantly transcribed in an efficient

and rapid process, which may be a factor that favors their overrepresentation among IGs. Moreover, the taste receptor cells of vertebrates are continually renewed throughout the organism's life which suggests a high demand for the housekeeping expression of these genes.

A considerable enrichment was also observed among IGs of BTB, Forkhead, HLH, HMG-Box transcription factors, and the chromatin components core histones H2AC, CENP-TC, and the histone Linker cluster. Thus, this indicates that IGs are playing also an important role in packaging, transcription and chromatin assembly.

Altogether, these results suggest that intronless genes have specialized roles, and a strong link to gene expression regulation and chromatin structure. Overall, our results suggest that IG proteins have specialized, prevalent, and unique biological roles.

Differential Expression of IGs in Telencephalon Development, Key Genes for Wnt Signaling

Embryonic development relies on the complex interplay of fundamental cellular processes, including proliferation, differentiation, and apoptosis. Regulation of these events is essential for the establishment of structures and organ development. The formation of the telencephalic architecture results from the interaction of the signaling centers located on the edges of the pallium. In this process, the Wnt signaling pathway plays an essential role in the dorsomedial pattern, where signals from the cortical hem direct the morphogenesis of the hippocampus, the corpus callosum, and the generation of migratory Cajal-Retzius cells. From the lateral pallium, the anti-hem signals, EGF, FGF, Frizzled, and Sfrp determine the development of the olfactory cortex.

In a previous study, we highlighted the up-regulation of Wnt signaling genes of the canonical pathway in the early stages of the developing telencephalon in mice (Muley et al., 2020). The main receptors of the Wnt/beta-catenin signaling

¹¹<https://www.ensembl.org/index.html>

pathway are Frizzled domain proteins (Fzd), a family of seven-transmembrane G-protein coupled receptors that also possess a large extracellular cysteine-rich domain.

In this work, we found *Fzd7*, *Fzd8*, and *Fzd10* among the IG transmembrane proteins that were developmentally regulated in the embryonic telencephalon, as well as a group of 11 IGs of the protocadherin β -cluster. Previous studies have described *Fzd8* as an essential receptor of the Wnt pathway implicated in brain development and size (Boyd et al., 2015). This gene is also highly expressed in two human cancer cell lines, indicating that it may play a role in tumorigenesis (Li et al., 2017; Murillo-Garzón et al., 2018). *Fzd10* functions in the canonical Wnt/beta-catenin signaling pathway which may be involved in signal transduction during tissue morphogenesis (Wang et al., 2005). In keeping with this, protocadherins have also been described as regulators in the Wnt signaling pathway (Mah and Weiner, 2017). More specifically, protocadherins of the β -cluster, along with those of the α and γ clusters, act cooperatively in mice in olfactory-axon targeting, in the formation of diverse neural circuits, and in neuronal survival (Hasegawa et al., 2016, 2017). These functions, however, correspond to developmental stages that occur later than the one addressed in this study. Hence, our striking finding that half of the 22 β -protocadherins are up-regulated along with Wnt receptors during the development of the telencephalon, suggest that this group of mostly IGs has a differential function thus far unknown related to Wnt signaling at this early stage. Consistent with this idea, the Wnt binding molecular function GO term is one of the most conserved among IGs in the genomes analyzed in this study.

Our expression analysis additionally revealed up-regulation of *Olig1*, *Bhlhe22*, *Bhlhe23*, *Pou3f1*, *Pou3f2*, *Pou3f4*, *Foxq1*, and *Neurog1*. Notably, this represents the up-regulation of three of the four members of the *Pou3* class of transcription factors present in mouse. The *Pou* genes encode a broad family of 6 classes (*Pou1f–Pou6f*) which are involved in developmental processes, mainly cell fate determination and differentiation (Tantin, 2013). Among those, the four members of the *Pou3f* class are preferentially expressed in ectodermal derivatives such as the developing mammalian nervous system (Bally-Cuif and Hammerschmidt, 2003). The human *Pou3f3* is an intronless gene also named *Brain-1*, which is a well-known transcription factor involved in the development of the central nervous system and its variant alleles have been associated with intellectual disability and language neurodevelopmental disorders (Snijders Blok et al., 2019). Furthermore, an important role of *Neurog1* is as a promoter of proliferation or neuronal differentiation, while *Olig1* is involved in the generation and maturation of specific neural cells during the development of the spinal cord (Qi et al., 2016; Song et al., 2017). *Bhlhe22* and *Bhlhe23* in turn, are among those that were up-regulated the most in mice during telencephalon development. In humans, *Bhlhe22* has been identified as a highly methylated gene in endometrial cancer with potential epigenetic biomarkers in cervical scrapings (Liew et al., 2019), while *Bhlhe23* has been linked to mammalian retinal development (Woods et al., 2018).

Finally, among the histone group, *H2bc21* (*Hist2h2be*), *H2bu2* (*Hist3h2ba*), *H2aw* (*Hist3h2a*) were also up-regulated. The *H2b* histone family members are responsible for the chromosomal

fiber nucleosome structure in eukaryotes. *H2bc21/Hist2h2be* has been described in mouse as expressed in olfactory epithelium, while *H2bu2* (*Hist3h2ba*) in neocortex and lens of camera type-eye, and *H2aw* (*Hist3h2a*) in retina¹². In humans, *Hist2h2be* is a hub gene related to poor prognosis in rhabdomyosarcoma tumors in pediatric patients (Li et al., 2019).

Summarizing, IGs appear to play crucial roles in the mouse telencephalon involved in gliogenesis, eye, and sensory organ development, canonical Wnt signaling, nucleosome organization, and have molecular regulatory roles. Therefore, in accordance with the functional assignment, our expression analysis supports that IGs play a critical role during mammalian brain development.

IG Proteins in the Histone Category Have Unique Signatures and Undergo Specific PTMs

In agreement with the link of IGs to chromatin structure found in this and previous works, we also identified unique and highly represented PTM signatures in the histone protein category. Proteins encoded by mouse IGs have enriched signatures for histone *H2A* and *H2B*, characteristic of key core histones involved in chromatin structure in eukaryotic cells, as well as linker histone *H1/H5* and the *CENB-type HTH*. In addition to the identification of exclusive signatures, we compared potential regulatory mechanisms of IG and MEG PTMs. Although the variability of PTMs is high, these modifications are typically very specific and, altogether, 300 types are known to occur in proteins (Witze et al., 2007). Among all PTMs, we found that the most abundant (Phosphorylation and Acetylation) are differentially represented among IG and MEG proteins.

As it could be expected from the observed enrichment in chromatin remodeling protein domains, our results show that proteins encoded by IGs undergo specific PTMs for histones such as crotonylation, methylation, sumoylation, citrullination, and sumoylation. However, our results suggest that IG-encoded histones have high specificity for Lysine-crotonylation, which is a recently identified post-translational modification associated with active promoters to directly stimulate transcription. Moreover, PTMs with changes in the physicochemical properties of amino acids like citrullination and amidation, are a characteristic feature highly enriched in IG proteins.

Potential Role of IGs as miPs in Neural Development and Function

When we assessed the potential role of IGs as microproteins we found proteins with strong potential to be modulators of multi-protein complexes. The targets or microproteins are mostly transcription factors that bind DNA as dimers. In this study, we found potential miPs encoded by intronless genes that are *bHLH* transcription factors with a regulatory role during critical events such as neural development and function. For example, *Ferd3l*, an evolutionarily conserved *bHLH* protein, is expressed in the developing central nervous system and functions as a transcriptional inhibitor. Other examples

¹²<https://bgee.org/>

are *bHLHe23*, a transcriptional regulator in the pancreas and brain that marks the dimesencephalic boundary (Bramblett et al., 2002), *Bhlha9* a regulator of apical ectodermal ridge formation during limb development (Kataoka et al., 2018), *Msg1* which is predominantly expressed in nascent mesoderm, the heart tube, limb bud, and sclerotome during mouse embryogenesis (Dunwoodie et al., 1998), and *Ascl5* member of the ASCL family of proneural transcription factors that control the development of the nervous system, particularly neuroblast cell fate determination (Guillemot et al., 1993). Moreover, its potential role in tumorigenesis has been described with up-regulation in lung cancer and down-regulation in brain tumors such as glioblastoma, anaplastic oligoastrocytoma, anaplastic oligodendroglioma, and oligodendroglioma (Wang et al., 2017). Additionally, consistent with the potential role in the development of IG-encoded miPs, we identified members of the *H1* linker histone group that fit the criteria to be classified as miPs. These histone proteins belong to a complex family with distinct specificity for tissues, developmental stages, and organisms in which they are expressed (Izzo et al., 2008).

Patterns of Evolution of IGs Differ From Those of MEGs in Vertebrates

According to earlier studies that reported the high conservation of mouse IG orthologs among other eukaryotic genomes, our analysis across seven species belonging to three classes of vertebrates revealed that the most numerous orthologs in each species were also IGs. This high rate of genetic structure conservation has been previously associated with their essential role in cell housekeeping functions, particularly those functionally pivotal proteins involved in molecular and biological roles such as transcription, translation, energy metabolism, amino-acid biosynthesis, and binding, which must be highly conserved (Sakharkar et al., 2006). Individually, eutherians (rat, chimp, human) are the species with the most orthologs. Moreover, *Muridae*, the clade that includes the mouse and rat, has the largest number of conserved orthologs and this number decreases as the clades gradually include the more distantly related species. Moreover, a distinctly large number was also found common to all species analyzed (*Vertebrata*), which is consistent with previous findings that identified functional and evolutionary conservation of eukaryotic IGs with highly distant genomes such as bacteria (Sakharkar et al., 2006). The higher abundance of conserved IG orthologs among species more closely related to mouse and lower in groups including more distant species, could be the result of the gradual loss of IG orthologs during the divergence of the diverse vertebrate branches. This abundance, however, could also be due to an increased rate of IG generation among mammals. Evidence supporting the latter possibility comes from our finding that in stark contrast, an increase of conserved MEG orthologs among species more closely related to mouse was not observed.

The clade *Vertebrata* contains the older IGs that are involved in diverse functions, among which nucleosome structure stands out. Histone IGs are conserved among all species, with some losses in opossum and chick. As histones are basic proteins

known to be conserved across eukaryotes, it is not surprising that they are found to be some of the oldest IGs. In some cases, histones are conserved in the genome as clusters, and in some others, they appear to have been generated in a specific lineage, due to multiple gene duplications. *Muridae*, in contrast to *Vertebrata*, contains the more recent IGs which are involved in keratinization, GPCRs, and chemodetection by GPCRs. We can conclude that *Muridae* IGs are younger and have more specific functions, whereas IGs conserved in all vertebrates are more ancient and have more general or basic functions, as expected due to the high conservation of sequence and function among vertebrates.

Gene duplication is an important mechanism for the acquisition of new genes, frequently providing specialized or new gene functions (Magadum et al., 2013). Known mechanisms of gene duplication include retroposition, tandem duplication, and genome duplication (Pan and Zhang, 2008). Our analysis shows that the vast majority (48%) of one exon genes in the mouse genome are a consequence of retroposition. Moreover, regarding the duplication events, we found clear examples of IG tandem repeat cluster organization. For example, the syntenic conservation of the tandem cluster of β -protocadherins and their neural tissue-specific expression suggest that some aspects of the nervous system characteristic of mammals could be associated with the duplication of intronless genes, such as olfactory-axon targeting, the formation of neural circuits, neuronal survival, or neurite self-avoidance during development (Dennis et al., 2016; Hasegawa et al., 2017; Brasch et al., 2019). Similar to the single exon β -cluster of protocadherins, we also observed that IG histones in the mouse genome are present as tandem families with a tendency to cluster in their chromosome organization. An example of this is the *H2A* histone family member *LIJ*, a family of ten IG members in the mouse X chromosome, with only one ortholog (*H2ALIRP*) in human and one (*H2A-beta*) in opossum. Almost all of the tandem repeat genes have parallel transcription orientation, which means they are encoded on the same strand.

The disrupted gene structure of most eukaryotic genes has led to a long-lasting debate regarding the origin of introns. The “exon theory of genes” also known as “introns-early,” proposed the presence of introns in prokaryotic primordial genes (Gilbert, 1987; Roy and Gilbert, 2006), while in the “insertional theory of introns” or “introns-late theory,” introns are a eukaryotic innovation (Doolittle and Stoltzfus, 1993; Stoltzfus, 1994; Stoltzfus et al., 1994; Mattick, 1994; Logsdon, 1998). Recent genomic evidence supports a view that combines aspects of both theories but still placing the invasion of eukaryotic genes by introns at the emergence of eukaryotic cells (Koonin, 2007). In accordance with this combined view, the comparison of mouse intron-bearing and intron-lacking IG orthologs among the analyzed organisms, suggests that IGs are more recent than MEGs. This is also consistent with findings that have revealed that intron-exon gene structure is highly stable among vertebrates and that individual intron losses outnumber intron gains in diverse vertebrate lineages (Roy et al., 2003; Coulombe-Huntington and Majewski, 2007; Loh et al., 2008; Venkatesh et al., 2014). This evolutionary stability also suggests that the observed

increase in IG abundance in *Muridae* is due to gene duplication of IGs rather than intron loss of MEG orthologs. Our findings, however, also revealed some interesting gene superfamilies of a few members each, containing only IGs which were restricted to one or two species. These IGs are also likely to have been generated recently in the branches leading to the species analyzed herein but the mechanism involved remains to be studied further.

The present study aimed to identify the conservation of the role of intronless genes in mammals and other vertebrate genomes. A comprehensive understanding of their biological function is essential to compare and contrast their evolution with that of intron-containing genes. Hence, we studied the complex regulatory role of intronless genes and their conservation in cellular environments using computational functional assignment, gene expression analysis, and evolutionary reconstruction. First, we determined that the functions associated with IGs are very different from those associated with MEGs. Expression analysis of the developing telencephalon also revealed specific up-regulation of IGs that encode genes involved in Wnt signaling, *bHLH* and *Pou* transcription factors, as well as chromatin proteins. Among Wnt signaling-related proteins, it was striking to detect up-regulation of half of all protocadherins of the β -cluster. Moreover, some IG transcription factors meet the criteria to be considered microproteins and thus appear to have modulatory properties of protein complex formation. Overall, our results highlight a role for IGs as essential modulators of diverse biological processes as pivotal as cortical development, chemosensory functions, chromatin condensation, and gene silencing. In fact, specific modifications of IG proteins indicate that their regulatory roles extend to the post-translational level. Notably, some of the IGs highlighted in this study also have potential clinical relevance in humans. For example, *Fzd8* and *pcdhs* which are associated to Wnt signaling, an evolutionarily conserved regulatory pathway related to cell fate determination and proliferation during development, have also been identified as part of a key mechanism in cancer biology. Other IGs discussed in this study and linked with cancer and neurodevelopmental disorders were *Pou3f3*, *bHLHE22*, *ASCL5*, and *Hist2h2be*.

Furthermore, the analysis of the evolutionary patterns of IGs revealed a large fraction of genes that appear to be of more recent generation as compared to the older and more conserved MEGs. Overall, this analysis reveals specific functions of IGs that distinguish them from MEGs and therefore strengthen the notion suggested by previous observations that these two groups are under differential evolutionary constraints.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: <https://data.mendeley.com/datasets/rdt5757cbw/1>.

AUTHOR CONTRIBUTIONS

KA-P: project design, performed data collection, manuscript writing, proofreading, carried out bioinformatic analyses,

prepared figures, and their interpretation. JR-R: expertise in bioinformatic analysis methods, performed evolutionary reconstruction, prepared figures. GH-O: literature search, bioinformatic analyses, and prepared figures. DV: writing, bioinformatic analysis, and prepared figures. ED-V: writing, bioinformatic analysis, and prepared figures. AG-G: expertise in data analysis methods for the API-REST and data collection, prepared figures. VM: proofreading, performed DEG analysis, prepared figures. AV-E: supervised the study, provided advice on the research strategy, and participated in manuscript writing. MH-R: co-director of the study and project development, performed bioinformatic analysis and interpretation, writing, and proofreading. All authors contributed to the article and approved the submitted version.

FUNDING

This project was supported by research funding provided by CONACYT grants QRO-2018-01-01-88344, 314869, 254206, 267749, and 315802 as well as DGAPA IN229620. KA-P received financial support from the DGAPA program for a postdoctoral fellowship at the INB UNAM and is a current holder of support from CONACyT (CVU: 227919).

ACKNOWLEDGMENTS

KA-P acknowledges the CABANA program for training in bioinformatics. For technical support we thank Luis Alberto Aguilar Bautista, Alejandro de León Cuevas, Carlos Sair Flores Bautista, and Jair García of the Laboratorio Nacional de Visualización Científica Avanzada (LAVIS). Critical comments and suggestions to this project development were received from Roddy Jorquera, Carolina Gonzalez, Michael Jerzioski, and Carlos Lozano Flores.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fgene.2021.654256/full#supplementary-material>

Supplementary Code | https://github.com/GEmilioHO/intronless_genes.

Supplementary Tool | <https://gitlab.com/jarr.tecn/revolutionh-tl>.

Supplementary Figure 1 | Prevalence of intronless protein-coding genes among single-exon genes in the mouse genome. All mouse genes having one exon are classified regarding their gene biotype, proportion of protein-coding intronless genes (IGs) is highlighted in pink.

Supplementary Figure 2 | Enrichment of SUPERFAMILY assignments of mouse IG and MEG proteins. **(A)** Enriched scop families in IG proteins: The scop families with the largest gene ratios are plotted in order of gene ratio. The size of the dots represents the number of genes in the significant background list associated with the scop family, while the color of the dots represents the adjusted p-values, **(B)** Enriched scop families in MEG proteins: The scop families with the largest gene ratios are plotted in order of gene ratio. The size of the dots represents the number of genes in the significant background list associated with the scop family, while the color of the dots represents the p-adjusted values.

REFERENCES

- Amigo, J. D., Opazo, J. C., Jorquera, R., Wichmann, I. A., Garcia-Bloj, B. A., Alarcon, M. A., et al. (2018). The reprimo gene family: a novel gene lineage in gastric cancer with tumor suppressive properties. *Int. J. Mol. Sci.* 19:1862. doi: 10.3390/ijms19071862
- Bally-Cuif, L., and Hammerschmidt, M. (2003). Induction and patterning of neuronal development, and its connection to cell cycle control. *Curr. Opin. Neurobiol.* 13, 16–25. doi: 10.1016/s0959-4388(03)00015-1
- Boyd, J. L., Skove, S. L., Rouanet, J. P., Pilaz, L. J., Bepler, T., Gordán, R., et al. (2015). Human-chimpanzee differences in a FZD8 enhancer alter cell-cycle dynamics in the developing neocortex. *Curr. Biol.* 25, 772–779. doi: 10.1016/j.cub.2015.01.041
- Bramblett, D. E., Copeland, N. G., Jenkins, N. A., and Tsai, M. J. (2002). BHLHB4 Is a BHLH transcriptional regulator in pancreas and brain that marks the dimesencephalic boundary. *Genomics* 79, 402–412. doi: 10.1006/geno.2002.6708
- Brasch, J., Goodman, K. M., Noble, A. J., Micah, R., Seetha, M., Fabiana, B., et al. (2019). Visualization of clustered protocadherin neuronal self-recognition complexes. *Nature* 569, 280–283. doi: 10.1038/s41586-019-1089-3
- Bult, C. J., Blake, J. A., Smith, C. L., Kadin, J. A., and Richardson, J. E. (2019). Mouse Genome Database (MGD) 2019. *Nucleic Acids Res.* 47, D801–D806.
- Coulombe-Huntington, J., and Majewski, J. (2007). Characterization of intron loss events in mammals. *Genome Res.* 17, 23–32. doi: 10.1101/gr.5703406
- de Klein, N., Magnani, E., Banf, M., and Rhee, S. Y. (2015). MicroProtein Prediction Program (MiP3): a software for predicting microproteins and their target transcription factors. *Int. J. Genomics* 2015:734147.
- de Souza, S. J. (2003). “The emergence of a synthetic theory of intron evolution,” in *Origin and Evolution of New Gene Functions*, ed. M. Long (Dordrecht: Springer), 117–121. doi: 10.1007/978-94-010-0229-5_2
- Dennis, D., Picketts, D., Slack, R. S., and Schuurmans, C. (2016). Forebrain neurogenesis: from embryo to adult. *Trends Dev. Biol.* 9:77.
- Doolittle, W. F., and Stoltzfus, A. (1993). Genes-in-pieces revisited. *Nature* 361:403. doi: 10.1038/361403a0
- Dunwoodie, S. L., Rodriguez, T. A., and Beddington, R. S. (1998). *Msg1* and *Mrg1*, founding members of a gene family, show distinct patterns of gene expression during mouse embryogenesis. *Mech. Dev.* 72, 27–40. doi: 10.1016/s0925-4773(98)00011-2
- Eguen, T., Straub, D., Graeff, M., and Wenkel, S. (2015). MicroProteins: small size–big impact. *Trends Plant Sci.* 20, 477–482. doi: 10.1016/j.tplants.2015.05.011
- Fedorova, L., and Fedorov, A. (2003). “Introns in gene evolution,” in *Origin and Evolution of New Gene Functions*, ed. M. Long (Berlin: Springer), 123–131. doi: 10.1007/978-94-010-0229-5_3
- Fong, N., Kim, H., Zhou, Y., Ji, X., Qiu, J., Saldi, T., et al. (2014). Pre-mRNA splicing is facilitated by an optimal RNA polymerase II elongation rate. *Genes Dev.* 28, 2663–2676. doi: 10.1101/gad.252106.114
- Gentles, A. J., and Karlin, S. (1999). Why are human G-Protein-coupled receptors predominantly intronless? *Trends Genet.* 15, 47–49. doi: 10.1016/s0168-9525(98)01648-5
- Gilbert, W. (1987). The exon theory of genes. *Cold Spring Harb. Symp. Quant. Biol.* 52, 901–905.
- Grzybowska, E. A. (2012). Human intronless genes: functional groups, associated diseases, evolution, and mRNA processing in absence of splicing. *Biochem. Biophys. Res. Commun.* 424, 1–6. doi: 10.1016/j.bbrc.2012.06.092
- Guillemot, F., Lo, L. C., Johnson, J. E., Auerbach, A., Anderson, D. J., and Joyner, A. L. (1993). Mammalian achaete-scute homolog 1 is required for the early development of olfactory and autonomic neurons. *Cell* 75, 463–476. doi: 10.1016/0092-8674(93)90381-y
- Hasegawa, S., Kobayashi, H., Kumagai, M., Nishimaru, H., Tarusawa, E., Kanda, H., et al. (2017). Clustered protocadherins are required for building functional neural circuits. *Front. Mol. Neurosci.* 10:114. doi: 10.3389/fnmol.2017.00114
- Hasegawa, S., Kumagai, M., Hagihara, M., Nishimaru, H., Hirano, K., Kaneko, R., et al. (2016). Distinct and cooperative functions for the protocadherin- α - β and γ clusters in neuronal survival and axon targeting. *Front. Mol. Neurosci.* 9:155. doi: 10.3389/fnmol.2016.00155
- Hellmuth, M., Hernandez-Rosales, M., Huber, K. T., Moulton, V., Stadler, P. F., and Wieseke, N. (2013). Orthology relations, symbolic ultrametrics, and cographs. *J. Math. Biol.* 66, 399–420. doi: 10.1007/s00285-012-0525-x
- Hernandez-Rosales, M., Hellmuth, M., Wieseke, N., Huber, K. T., Moulton, V., and Stadler, P. F. (2012). From event-labeled gene trees to species trees. *BMC Bioinformatics* 13:S6.
- Hung, M. S., Lin, Y. C., Mao, J. H., Kim, I. J., Xu, Z., Yang, C. T., et al. (2010). Functional polymorphism of the CK2 α intronless gene plays oncogenic roles in lung cancer. *PLoS One* 5:e11418. doi: 10.1371/journal.pone.0011418
- Izzo, A., Kamieniarz, K., and Schneider, R. (2008). The histone H1 family: specific members, specific functions? *Biol. Chem.* 389, 333–343. doi: 10.1515/bc.2008.037
- Jorquera, R., González, C., Clausen, P., Petersen, B., and Holmes, D. S. (2018). Improved ontology for eukaryotic single-exon coding sequences in biological databases. *Database* 2018, 1–6. doi: 10.1002/9783527678679.dg08413
- Jorquera, R., Ortiz, R., Ossandon, F., Cardenas, J. P., Sepulveda, R., Gonzalez, C., et al. (2016). SinEx DB: a database for single exon coding sequences in mammalian genomes. *Database* 2016:baw095. doi: 10.1093/database/baw095
- Kaessmann, H., Vinckenbosch, N., and Long, M. (2009). RNA-based gene duplication: mechanistic and evolutionary insights. *Nat. Rev. Genet.* 10, 19–31. doi: 10.1038/nrg2487
- Kataoka, K., Matsushima, T., Ito, Y., Sato, T., Yokoyama, S., and Asahara, H. (2018). *Bhlha9* regulates apical ectodermal ridge formation during limb development. *J. Bone Miner. Metab.* 36, 64–72. doi: 10.1007/s00774-017-0820-0
- Koonin, E. V. (2007). The biological big bang model for the major transitions in evolution. *Biol. Direct* 2:21. doi: 10.1186/1745-6150-2-21
- Lechner, M., Findeiß, S., Steiner, L., Marz, M., Stadler, P. F., and Prohaska, S. J. (2011). Proteinortho: detection of (Co-) orthologs in large-scale analysis. *BMC Bioinformatics* 12:124.
- Lechner, M., Hernandez-Rosales, M., Doerr, D., Wieseke, N., Thévenin, A., Stoye, J., et al. (2014). Orthology detection combining clustering and synteny for very large datasets. *PLoS One* 9:e105015. doi: 10.1371/journal.pone.0105015
- Li, Q., Ye, L., Zhang, X., Wang, M., Lin, C., Huang, S., et al. (2017). FZD8, a Target of P53, promotes bone metastasis in prostate cancer by activating canonical Wnt/ β -Catenin signaling. *Cancer Lett.* 402, 166–176. doi: 10.1016/j.canlet.2017.05.029
- Li, Q., Zhang, L., Jiang, J., Zhang, Y., Wang, X., Zhang, Q., et al. (2019). CDK1 and CCNB1 as potential diagnostic markers of rhabdomyosarcoma: validation following bioinformatics analysis. *BMC Med. Genomics* 12:198.
- Liew, P. L., Huang, R. L., Wu, T. I., Liao, C. C., Chen, C. W., Su, P. H., et al. (2019). Combined genetic mutations and DNA-methylated genes as biomarkers for endometrial cancer detection from cervical scrapings. *Clin. Epigenetics* 11:170.
- Liu, X. Y., Fan, Y. C., Gao, S., Zhao, J., Chen, L. Y., Li, F., et al. (2017). Methylation of SOX1 and VIM promoters in serum as potential biomarkers for hepatocellular carcinoma. *Neoplasma* 64, 745–753. doi: 10.4149/neo_2017_513
- Logsdon, J. M. Jr. (1998). The recent origins of spliceosomal introns revisited. *Curr. Opin. Genet. Dev.* 8, 637–648. doi: 10.1016/s0959-437x(98)80031-2
- Loh, Y. H., Brenner, S., and Venkatesh, B. (2008). Investigation of Loss and gain of introns in the compact genomes of pufferfishes (Fugu and Tetraodon). *Mol. Biol. Evol.* 25, 526–535. doi: 10.1093/molbev/msm278
- Louhichi, A., Fourati, A., and Rebaï, A. (2011). IGD: a resource for intronless genes in the human genome. *Gene* 488, 35–40. doi: 10.1016/j.gene.2011.08.013
- Magadum, S., Banerjee, U., Murugan, P., Gangapur, D., and Ravikesavan, R. (2013). Gene duplication as a major force in evolution. *J. Genet.* 92, 155–161. doi: 10.1007/s12041-013-0212-8
- Mah, K. M., and Weiner, R. A. (2017). Regulation of Wnt signaling by protocadherins. *Semin. Cell Dev. Biol.* 69, 158–171. doi: 10.1016/j.semcdb.2017.07.043
- Mattick, J. S. (1994). Introns: evolution and function. *Curr. Opin. Genet. Dev.* 4, 823–831. doi: 10.1016/0959-437x(94)90066-3
- Muley, V. Y., López-Victorio, C. J., Ayala-Sumano, J. T., González-Gallardo, A., González-Santos, L., Lozano-Flores, C., et al. (2020). Conserved and divergent expression dynamics during early patterning of the telencephalon in mouse and chick embryos. *Prog. Neurobiol.* 186:101735. doi: 10.1016/j.pneurobio.2019.101735
- Murillo-Garzon, V., Gorroño-Etxebarria, I., Åkerfelt, M., Puustinen, M. C., Sistonen, L., Nees, M., et al. (2018). Frizzled-8 integrates Wnt-11 and transforming growth factor- β signaling in prostate cancer. *Nat. Commun.* 9:1747.
- Noonan, J. P., Grimwood, J., Schmutz, J., Dickson, M., and Myers, R. M. (2004). Gene conversion and the evolution of protocadherin gene cluster diversity. *Genome Res.* 14, 354–366. doi: 10.1101/gr.2133704

- Ohki, R., Nemoto, J., Murasawa, H., Oda, E., Inazawa, J., Tanaka, N., et al. (2000). Reprimin, a new candidate mediator of the P53-mediated cell cycle arrest at the G2 phase. *J. Biol. Chem.* 275, 22627–22630. doi: 10.1074/jbc.c000235200
- Pan, D., and Zhang, L. (2008). Tandemly arrayed genes in vertebrate genomes. *Comp. Funct. Genomics* 2008:545269.
- Qi, Q., Zhang, Y., Shen, L., Wang, R., Zhou, J., Lü, H., et al. (2016). Olig1 Expression pattern in neural cells during rat spinal cord development. *Neuropsychiatr. Dis. Treat.* 12, 909–916. doi: 10.2147/ndt.s99257
- Roy, S. W., and Gilbert, W. (2006). The evolution of spliceosomal introns: patterns, puzzles and progress. *Nat. Rev. Genet.* 7, 211–221. doi: 10.1038/nrg1807
- Roy, S. W., Fedorov, A., and Gilbert, W. (2003). Large-Scale comparison of intron positions in mammalian genes shows intron loss but no gain. *Proc. Natl. Acad. Sci. U.S.A.* 100, 7158–7162. doi: 10.1073/pnas.1232297100
- Sakharkar, K. R., Sakharkar, M. K., Culiat, C. T., Chow, V. T., and Pervaiz, S. (2006). Functional and evolutionary analyses on expressed intronless genes in the mouse genome. *FEBS Lett.* 580, 1472–1478. doi: 10.1016/j.febslet.2006.01.070
- Sakharkar, M. K., Chow, V. T., Ghosh, K., Chaturvedi, I., Lee, P. C., Bagavathi, S. P., et al. (2005a). Computational prediction of SEG (Single Exon Gene) function in humans. *Front. Biosci.* 10:1382–1395. doi: 10.2741/1627
- Sakharkar, M. K., Kanguane, P., Petrov, D. A., Kolaskar, A. S., and Subbiah, S. (2002). “SEGE: a database on ‘intron less/single exonic’ genes from eukaryotes. *Bioinformatics* 18, 1266–1267. doi: 10.1093/bioinformatics/18.9.1266
- Sakharkar, M. K., Perumal, B. S., Lim, Y. P., Chern, L. P., Yu, Y., and Kanguane, P. (2005b). Alternatively spliced human genes by exon skipping—a database (ASHESdb). *In Silico Biol.* 5, 221–225.
- Snijders Blok, L., Kleefstra, T., Venselaar, H., Maas, S., Kroes, H. Y., Lachmeijer, A. M. A., et al. (2019). De novo variants disturbing the transactivation capacity of POU3F3 cause a characteristic neurodevelopmental disorder. *Am. J. Hum. Genet.* 105, 403–412. doi: 10.1016/j.ajhg.2019.06.007
- Song, Z., Jadali, A., Fritsch, B., and Kwan, K. Y. (2017). NEUROG1 regulates CDK2 to promote proliferation in otic progenitors. *Stem Cell Rep.* 9, 1516–1529. doi: 10.1016/j.stemcr.2017.09.011
- Staudt, A. C., and Wenkel, S. (2011). Regulation of protein function by ‘MicroProteins.’ *EMBO Rep.* 12, 35–42. doi: 10.1038/embor.2010.196
- Stoltzfus, A. (1994). Origin of introns—early or late? *Nature* 369, 526–527. doi: 10.1038/369526b0
- Stoltzfus, A., Spencer, D. F., Zuker, M., Logsdon, J. M., and Doolittle, W. F. (1994). Testing the exon theory of genes: the evidence from protein structure. *Science* 265, 202–207. doi: 10.1126/science.8023140
- Straub, D., and Wenkel, S. (2017). Cross-species genome wide identification of evolutionary conserved microproteins. *Genome Biol. Evol.* 9, 777–789. doi: 10.1093/gbe/evx041
- Sunahara, R. K., Niznik, H. B., Weiner, D. M., Stormann, T. M., Brann, M. R., Kennedy, J. L., et al. (1990). Human dopamine D1 receptor encoded by an intronless gene on chromosome 5. *Nature* 347, 80–83. doi: 10.1038/347080a0
- Supek, F., Bošnjak, M., Škunca, N., and Šmuc, T. (2011). REVIGO Summarizes and visualizes long lists of gene ontology terms. *PLoS One* 6:e21800. doi: 10.1371/journal.pone.0021800
- Tantin, D. (2013). Oct transcription factors in development and stem cells: insights and mechanisms. *Development* 140, 2857–2866. doi: 10.1242/dev.095927
- Tedder, M., Corneil, D., Habib, M., and Paul, C. (2008). “Simpler linear-time modular decomposition via recursive factorizing permutations,” in *International Colloquium on Automata, Languages, and Programming*, eds L. Aceto, I. Damgård, L. A. Goldberg, M. M. Halldórsson, A. Ingólfssdóttir, and I. Walukiewicz (Berlin: Springer), 634–645. doi: 10.1007/978-3-540-70575-8_52
- Tine, M., Kuhl, H., Beck, A., Bargelloni, L., and Reinhardt, R. (2011). Comparative Analysis of intronless genes in teleost fish genomes: insights into their evolution and molecular function. *Mar. Genomics* 4, 109–119. doi: 10.1016/j.margen.2011.03.004
- Uversky, V. N. (2015). Functional roles of transiently and intrinsically disordered regions within proteins. *FEBS J.* 282, 1182–1189. doi: 10.1111/febs.13202
- Venkatesh, B., Lee, A. P., Ravi, V., Maurya, A. K., Lian, M. M., Swann, J. B., et al. (2014). Elephant shark genome provides unique insights into gnathostome evolution. *Nature* 505, 174–179. doi: 10.1038/nature12826
- Wang, C. Y., Shahi, P., Huang, J. T., Phan, N. N., Sun, Z., Lin, Y. C., et al. (2017). Systematic analysis of the achaete-scute complex-like gene signature in clinical cancer patients. *Mol. Clin. Oncol.* 6, 7–18. doi: 10.3892/mco.2016.1094
- Wang, Z., Shu, W., Lu, M. M., and Morrissey, E. E. (2005). Wnt7b activates canonical signaling in epithelial and vascular smooth muscle cells through interactions with Fzd1, Fzd10, and LRP5. *Mol. Cell. Biol.* 25, 5022–5030. doi: 10.1128/mcb.25.12.5022-5030.2005
- Witke, E. S., Old, W. M., Resing, K. A., and Ahn, N. G. (2007). Mapping protein post-translational modifications with mass spectrometry. *Nat. Methods* 4, 798–806.
- Woods, S. M., Mountjoy, E., Muir, D., Ross, S. E., and Atan, D. (2018). A comparative analysis of rod bipolar cell transcriptomes identifies novel genes implicated in night vision. *Sci. Rep.* 8:5506.
- Yan, H., Zhang, W., Lin, Y., Dong, Q., Peng, X., Jiang, H., et al. (2014). Different evolutionary patterns among intronless genes in maize genome. *Biochem. Biophys. Res. Commun.* 449, 146–150. doi: 10.1016/j.bbrc.2014.05.008
- Yu, G., Wang, L. G., Han, Y., and He, Q. Y. (2012). ClusterProfiler: an R package for comparing biological themes among gene clusters. *Omics* 16, 284–287. doi: 10.1089/omi.2011.0118
- Yu, W. P., Rajasegaran, V., Yew, K., Loh, W. L., Tay, B. H., Amemiya, C. T., et al. (2008). Elephant shark sequence reveals unique insights into the evolutionary history of vertebrate genes: a comparative analysis of the protocadherin cluster. *Proc. Natl. Acad. Sci. U.S.A.* 105, 3819–3824. doi: 10.1073/pnas.0800398105
- Zou, M., Guo, B., and He, S. (2011). The roles and evolutionary patterns of intronless genes in deuterostomes. *Comp. Funct. Genomics* 2011:680673.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2021 Aviña-Padilla, Ramírez-Rafael, Herrera-Oropeza, Muley, Valdivia, Díaz-Valenzuela, García-García, Varela-Echavarría and Hernández-Rosales. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Advantages of publishing in Frontiers



OPEN ACCESS

Articles are free to read
for greatest visibility
and readership



FAST PUBLICATION

Around 90 days
from submission
to decision



HIGH QUALITY PEER-REVIEW

Rigorous, collaborative,
and constructive
peer-review



TRANSPARENT PEER-REVIEW

Editors and reviewers
acknowledged by name
on published articles

Frontiers

Avenue du Tribunal-Fédéral 34
1005 Lausanne | Switzerland

Visit us: www.frontiersin.org

Contact us: frontiersin.org/about/contact



REPRODUCIBILITY OF RESEARCH

Support open data
and methods to enhance
research reproducibility



DIGITAL PUBLISHING

Articles designed
for optimal readership
across devices



FOLLOW US

@frontiersin



IMPACT METRICS

Advanced article metrics
track visibility across
digital media



EXTENSIVE PROMOTION

Marketing
and promotion
of impactful research



LOOP RESEARCH NETWORK

Our network
increases your
article's readership