GENOMIC COLOCALIZATION AND ENRICHMENT ANALYSES

EDITED BY: Geir Kjetil Sandve, Subhajyoti De, Ryan Matthew Layer and Eivind Hovig

PUBLISHED IN: Frontiers in Genetics

frontiers Research Topics

ISSN 1664-8714
ISBN 978-2-88966-566-2
DOI 10.3389/978-2-88966-566-2

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

GENOMIC COLOCALIZATION AND ENRICHMENT ANALYSES

Topic Editors:
Geir Kjetil Sandve, University of Oslo, Norway
Subhajyoti De, The State University of New Jersey, United States
Ryan Matthew Layer, University of Colorado Boulder, United States
Eivind Hovig, University of Oslo, Norway

Topic Editor Ryan Matthew Layer is a co-founder of Base2 Genomics. The rest of Topic Editors declare no competing interests with regards to the Research Topic.

Citation: Sandve, G. K., De, S., Layer, R. M., Hovig, E., eds. (2021). Genomic Colocalization and Enrichment Analyses. Lausanne: Frontiers Media SA.
doi: 10.3389/978-2-88966-566-2

Table of Contents

frontiers in Genetics

EDITORIAL
published: 26 January 2021
doi: 10.3389/fgene.2020.617876

Editorial: Genomic Colocalization and Enrichment Analyses

Chakravarthi Kanduri[1], *Geir Kjetil Sandve*[1], *Eivind Hovig*[1,2], *Subhajyoti De*[3] and *Ryan M. Layer*[4,5]*

[1] Department of Informatics, University of Oslo, Oslo, Norway, [2] Department of Tumor Biology, Institute for Cancer Research, Radium Hospital, Oslo University Hospital, Oslo, Norway, [3] Rutgers Cancer Institute of New Jersey, New Brunswick, GA, United States, [4] Computer Science Department, University of Colorado, Boulder, CO, United States, [5] BioFrontiers Institute, University of Colorado, Boulder, CO, United States

Keywords: colocalization analyses, bioinformatics, genomics, genome annotation, computational biology, enrichment analyses, co-occurrence analyses, genomic overlap

Editorial on the Research Topic

Genomic Colocalization and Enrichment Analyses

INTRODUCTION

To decipher the molecular basis of health and disease, profiling multiple molecular modalities is a common practice [e.g., genetic variation, transcription, chromatin accessibility, epigenomic marks, binding sites, and three dimensional (3D) genome architecture]. Most of these molecular assays generate lists of genomic loci that are relevant to the trait/phenotype under investigation. Functional interpretation of these lists is often carried out through colocalization and enrichment analyses (Kanduri et al., 2018), which is akin to gene ontology/pathway analysis for lists of genes. A wide range of tools and methodologies have been developed over the past decade to perform colocalization and enrichment analyses of genomic regions. Given the availability and continuous generation of massive high resolution, cell-specific public datasets (e.g., ENCODE, RoadMap Epigenomics, GTEx, and BLUEPRINT), both existing and novel colocalization/enrichment analysis strategies will continue to generate new knowledge in our understanding of the molecular basis of health and disease. To highlight current research demonstrating the utility of colocalization/enrichment analysis, we invited contributions for a special Research Topic. The received contributions in this article collection include a comprehensive literature review, tools that extend the state-of-the-art methodology and enhance the user convenience in performing colocalization/enrichment analyses, and applied work that demonstrates the utility of colocalization/enrichment analyses.

LITERATURE REVIEW SUMMARIZING HOW COLOCALIZATION/ENRICHMENT ANALYSES HAVE AIDED THE FUNCTIONAL INTERPRETATION OF GWAS FINDINGS

Cano-Gamez and Trynka provide a detailed overview of how various strategies, especially enrichment and colocalization analysis, have aided in the interpretation of the findings of genome-wide association studies (GWAS). Specifically, the authors summarized single nucleotide polymorphism (SNP) enrichment analysis and statistical colocalization analysis. SNP enrichment analysis is one way to identify the tissue/cell types that are relevant for a disease by integrating either genome-wide-significant or a full set of assayed SNPs with molecular annotation tracks (e.g., either gene expression or chromatin accessibility). Once the relevant tissue/cell types are identified,

Edited and reviewed by:
Richard D. Emes,
University of Nottingham,
United Kingdom

****Correspondence:***
Ryan M. Layer
ryan.layer@colorado.edu

Specialty section:
This article was submitted to Computational Genomics, a section of the journal Frontiers in Genetics

***Received:** 15 October 2020*
***Accepted:** 24 November 2020*
***Published:** 26 January 2021*

Citation:
Kanduri C, Sandve GK, Hovig E, De S and Layer RM (2021) Editorial: Genomic Colocalization and Enrichment Analyses. Front. Genet. 11:617876. doi: 10.3389/fgene.2020.617876

further refined analysis using similar statistical analysis methods could disentangle the enrichments in highly similar cell types (e.g., to differentiate between cell states). Statistical colocalization analysis is one way to interpret novel GWAS findings by linking GWAS findings with likely target genes. This can be achieved by integrating GWAS signal with eQTL data to evaluate whether the same variant is causal in both GWAS and eQTL studies. In addition to summarizing the knowledge and strategies of the SNP enrichment and colocalization analysis, the authors have also provided perspectives on how the state-of-the-art technologies (e.g., single cell sequencing, genome editing) could be utilized in the future for the interpretation of GWAS findings.

TOOL THAT EXTENDS THE STATE-OF-THE-ART

One of the applications of colocalization analysis is for the interpretation of the functions of non-coding genomic regions. GREAT (McLean et al., 2010) and many other similar tools assign a regulatory domain for each gene that extends user-customizable distance both upstream and downstream to the transcription start site (TSS) of that gene. The regions of DNA binding events (both proximal and distal) are then assigned to genes, and subsequent statistical testing akin to traditional gene ontology analysis is performed to aid the functional interpretation. In this special collection, a novel method titled ProxReg (Lee et al.) complements the current state-of-the-art methods to aid the functional interpretation of non-coding regions by extending the methodology to not only test the proximity to TSS, but also to enhancers. The authors show that ProxReg provides additional insights into the regulatory mechanisms and binding tendencies of transcription factors (e.g., cell-specific regulatory mechanisms of the same TF by binding at promoters in one cell type and binding at enhancers in another cell type).

TOOLS THAT ENHANCE THE USER CONVENIENCE IN PERFORMING COLOCALIZATION/ENRICHMENT ANALYSES

EpiColoc

One of the arduous tasks when using colocalization analysis tools to test/generate hypotheses is the need to carefully curate a collection of reference genomic tracks that are annotated thoroughly. Existing tools provide carefully curated collections of reference track collections (e.g., see Sheffield and Bock, 2016; Simovski et al., 2017; Layer et al., 2018); but epiColoc (Zhou et al.) published in this special issue takes a step further in this direction, and provides large collections of curated genomic tracks (44,385 bulk/single cell genomic tracks across 53 human cell/tissue types). The curated data span across transcriptional regulators, histone modifications, chromatin accessibility, transcriptional events, and chromatin segmentation data.

LD-Annot

To perform any colocalization or enrichment analysis that involves SNPs, it is desirable to include statistically significant SNPs and all SNPs that are in tight linkage disequilibrium (LD) with them. Often, subsequent enrichment analyses are carried out on reference genome annotations that are overlapping the LD blocks. LD-annot provides a convenient wrapper around the popular PLINK tool (Chang et al., 2015) that computes LD between the genotypes of a given dataset and uses that information to intersect and extract the reference genome annotations overlapping the LD blocks.

APPLIED WORK

The study by Cresswell and Dozmorov, which includes a novel method titled TADcompare, demonstrates the utility of colocalization/enrichment analyses in aiding the functional interpretation of genomic regions with unknown biological significance. TADcompare is a method specifically developed to identify the changes in interacting domains (one of the features of three-dimensional genome architecture) and compare them across different conditions. One of the main challenges for TADcompare (as noted by the authors) was that no ground truth exists for boundaries of interacting domains, making it difficult to quantify the identified boundaries' biological relevance. To tackle this challenge, the authors of TADcompare used a range of colocalization analyses of epigenomic annotations and also a colocalization-based gene ontology enrichment analysis to determine whether the known genomic features that are characteristic of interacting domains and boundaries are enriched proximal to the identified boundaries and if that is different than background (non-boundaries).

The study by Ronzio et al. presents a new pipeline based on colocalization/enrichment analyses to identify regulatory modules of transcription factors (TFs) and TF recruitment rules. Instead of requiring overlap between a pair of ChIP-seq tracks, a proximity-based test statistic is suggested to quantify colocalization. The significance (*p*-value) is computed according to either hypergeometric or Poisson distribution. One possibility in the pipeline is to convert the *p*-values into scores and perform clustering analysis between the scores of multiple experiments to visualize potential regulatory modules. Further, motif enrichment analysis either relative to the whole accessible DNA or selected windows (e.g., upstream/downstream regions) could be performed for a pair of TFs. One could draw inferences on the recruitment patterns based on the observed pattern of motif enrichment (motifs for both TFs enriched or only one of them or none). Construction of the background sets in both colocalization analysis and motif enrichment analysis using alternative definitions (e.g., focusing only relative to enhancers/promoters) would allow one to identify specific regulatory modules.

The study by Tan et al. extends the traditional GSEA approach to establish associations between chemical-associated gene sets and gene expression in colorectal and rectal cancers. In the absence of a reliable tool to simulate CNVs from whole-exome sequencing data, Xing et al. developed the SECNVs tool which can generate CNVs with multiple customizable parameter options to mimic realistic CNVs from experimental data. The simulated CNV datasets could be utilized to explore the patterns of enrichment of CNVs in various contexts. For example, earlier others (Alexandrov et al., 2020; Singh et al., 2020) analyzed the patterns of enrichment of somatic mutations in tumor genomes and associated mutational signatures in their (epi)genomic contexts to infer their likely etiologies during tumorigenesis.

Overall, this Research Topic summarizes and showcases some of the existing and novel ways of utilizing genome colocalization/enrichment analyses to study a wide range of genetics and genomic research questions. The methods and tools published in this Research Topic extend the state of the art and enhance user convenience in performing genomic colocalization/enrichment analysis. With the continuous increase in the generation of genomic/epigenomic datasets, the interpretation of the resulting genomic regions becomes vital; we expect that the methodological principles of genomic colocalization/enrichment analysis will be utilized in many innovative ways in the future to further aid the functional interpretation of genomics datasets.

AUTHOR CONTRIBUTIONS

CK wrote the manuscript. GS, EH, SD, and RL edited the manuscript. All authors contributed to the article and approved the submitted version.

REFERENCES

Alexandrov, L. B., Kim, J., Haradhvala, N. J., Huang, M. N., Tian Ng, A. W., Wu, Y., et al. (2020). The repertoire of mutational signatures in human cancer. *Nature* 578, 94–101. doi: 10.1038/s41586-020-1943-3

Chang, C. C., Chow, C. C., Tellier, L. C., Vattikuti, S., Purcell, S. M., and Lee, J. J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience* 4:7. doi: 10.1186/s13742-015-0047-8

Kanduri, C., Bock, C., Gundersen, S., Hovig, E., and Sandve, G. K. (2018). Colocalization analyses of genomic elements: approaches, recommendations and challenges. *Bioinformatics* 35, 1615–1624. doi: 10.1093/bioinformatics/bty835

Layer, R. M., Pedersen, B. S., DiSera, T., Marth, G. T., Gertz, J., and Quinlan, A. R. (2018). GIGGLE: a search engine for large-scale integrated genome analysis. *Nat. Methods* 15, 123–126. doi: 10.1038/nmeth.4556

McLean, C. Y., Bristor, D., Hiller, M., Clarke, S. L., Schaar, B. T., Lowe, C. B., et al. (2010). GREAT improves functional interpretation of cis-regulatory regions. *Nat. Biotechnol.* 28, 495–501. doi: 10.1038/nbt.1630

Sheffield, N. C., and Bock, C. (2016). LOLA: enrichment analysis for genomic region sets and regulatory elements in R and Bioconductor. *Bioinformatics* 32, 587–589. doi: 10.1093/bioinformatics/btv612

Simovski, B., Vodák, D., Gundersen, S., Domanska, D., Azab, A., Holden, L., et al. (2017). GSuite HyperBrowser: integrative analysis of dataset collections across the genome and epigenome. *GigaScience* 6, 1–12. doi: 10.1093/gigascience/gix032

Singh, V. K., Rastogi, A., Hu, X., Wang, Y., and De, S. (2020). Mutational signature SBS8 predominantly arises due to late replication errors in cancer. *Commun. Biol.* 3:421. doi: 10.1038/s42003-020-01119-5

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.

METHODS
published: 26 November 2019
doi: 10.3389/fgene.2019.01192

LD-annot: A Bioinformatics Tool to Automatically Provide Candidate SNPs With Annotations for Genetically Linked Genes

Julien Prunier[1,2], *Audrey Lemaçon*[1], *Alexandre Bastien*[3], *Mohsen Jafarikia*[4,5], *Ilga Porth*[2], *Claude Robert*[2] *and Arnaud Droit*[1]*

[1] Genomics Center, Centre Hospitalier Universitaire de Québec–Université Laval Research Center, Quebec, QC, Canada, [2] Forestry Research Centre, Forestry Department, Université Laval, Quebec, QC, Canada, [3] Faculty of Agricultural and Food Science, Université Laval, Quebec, QC, Canada, [4] Canadian Centre for Swine Improvement, Ottawa, ON, Canada, [5] Department of Animal Biosciences, University of Guelph, Guelph, ON, Canada

A multitude of model and non-model species studies have now taken full advantage of powerful high-throughput genotyping advances such as SNP arrays and genotyping-by-sequencing (GBS) technology to investigate the genetic basis of trait variation. However, due to incomplete genome coverage by these technologies, the identified SNPs are likely in linkage disequilibrium (LD) with the causal polymorphisms, rather than be causal themselves. In addition, researchers could benefit from annotations for the identified candidate SNPs and, simultaneously, for all neighboring genes in genetic linkage. In such case, LD extent estimation surrounding the candidate SNPs is required to determine the regions encompassing genes of interest. We describe here an automated pipeline, “LD-annot,” designed to delineate specific regions of interest for a given experiment and candidate polymorphisms on the basis of LD extent, and furthermore, provide annotations for all genes within such regions. LD-annot uses standard file formats, bioinformatics tools, and languages to provide identifiers, coordinates, and annotations for genes in genetic linkage with each candidate polymorphism. Although the focus lies upon SNP arrays and GBS data as they are being routinely deployed, this pipeline can be applied to a variety of datasets as long as genotypic data are available for a high number of polymorphisms and formatted into a vcf file. A checkpoint procedure in the pipeline allows to test several threshold values for linkage without having to rerun the entire pipeline, thus saving the user computational time and resources. We applied this new pipeline to four different sample sets: two breeding populations GBS datasets, one within-pedigree SNP set coming from whole genome sequencing (WGS), and a very large multi-varieties SNP dataset obtained from WGS, representing variable sample sizes, and numbers of polymorphisms. LD-annot performed within minutes, even when very high numbers of polymorphisms are investigated and thus will efficiently assist research efforts aimed at identifying biologically meaningful genetic polymorphisms underlying phenotypic variation. LD-annot tool is available under a GPL license from https://github.com/ArnaudDroitLab/LD-annot.

Keywords: linkage disequilibrium, candidate SNP, SNP annotation, bioinformatics tool, variant call format (VCF), SNP chip analyses

OPEN ACCESS

Edited by:
Geir Kjetil Sandve,
University of Oslo,
Norway

Reviewed by:
Padma Nimmakayala,
West Virginia State University,
United States
Michael Poidinger,
Murdoch Childrens Research
Institute (MCRI), Australia

Correspondence:
Arnaud Droit
Arnaud.Droit@crchudequebec.ulaval.ca

Specialty section:
This article was submitted to
Bioinformatics and
Computational Biology,
a section of the journal
Frontiers in Genetics

Received: *02 July 2019*
Accepted: *28 October 2019*
Published: *26 November 2019*

Citation:
Prunier J, Lemaçon A, Bastien A, Jafarikia M, Porth I, Robert C and Droit A (2019) LD-annot: A Bioinformatics Tool to Automatically Provide Candidate SNPs With Annotations for Genetically Linked Genes. Front. Genet. 10:1192. doi: 10.3389/fgene.2019.01192

INTRODUCTION

The progress in molecular technologies enabled the study of genetic variants at the genome level, in both model and non-model species, such as Genome-Wide Association Studies (GWAS) identifying genetic variants likely involved in variation of interesting quantitative traits or in adaptation to environmental stress. Among those molecular techniques, SNP genotyping chips and genotyping-by-sequencing (GBS) approaches [also addressing the related reduction site-associated DNA sequencing (RADseq) in this paper] are often deployed to efficiently screen genomes at the population level and test for relationships between genetic polymorphisms and either quantitative characteristics or environmental conditions (i.e. Keller et al., 2013; Narum et al., 2013; Sonah et al., 2015; Carter et al., 2018; Torkamaneh et al., 2018;). GBS is based on sequencing genome subparts using restriction enzymes and insert size selection (Elshire et al., 2011) and yields thousands of genetic variants randomly distributed over the genome. SNP genotyping chips are based on allele-specific hybridization and traditionally include SNPs previously identified and selected to be regularly distributed across the genome (Carvalho et al., 2007; Bai et al., 2018). Both techniques usually result in thousands of SNPs successfully genotyped.

Research projects based on either of these variant detection approaches often investigate the genomic basis of trait variations related to agronomic performance in cultivated plants or animals (Carter et al., 2018; Torkamaneh et al., 2018;), the dispersion of invasive species (White et al., 2013; Roe et al., 2018), or species' adaptation (Hess et al., 2012; Keller et al., 2013), for instance. Such studies typically use regression models to select candidate SNPs presenting significant trait variations between distinct genotypic classes. However, these polymorphisms might not be directly responsible for phenotypic variations but in linkage disequilibrium (LD) with larger genomic regions encompassing untested genetic variants that might be truly causal for the studied phenotypic variation.

LD is the non-random assortment of alleles between neighboring loci due to the short physical distance limiting recombination between them during meiosis. This phenomenon results in a systemic association between alleles of the same parental origin. For biallelic loci, LD is often estimated using the correlation coefficient (denoted r^2) between two alleles at two different loci (Hill and Robertson 1968). This estimate varies with the recombination coefficient which is a function of physical distance between markers (Hill and Robertson 1968). However, the recombination coefficient actually fluctuates along the genome, with regions known to present lower recombination coefficients than others, such as centromeric regions for instance (Smith et al., 2005). In addition, r^2 is also impacted by inbreeding which results in lower genetic diversity that in turn leads to homozygosity hiding recombination events. Hence, r^2 also varies between populations according to population demographic history (Reich et al., 2001), even within species. Similarly, the r^2 estimator presents a variability related to allele frequencies (minor allele frequency, MAF) (VanLiere and Rosenberg 2008) or sample size effect (Jorgenson and Witte 2006). Despite its limitations, the r^2 estimate remains largely used and most interesting when scanning GWAS results, for instance, since the correlation between two SNPs is still indicative of a mathematical link (Bush and Moore 2012), either reflecting a true low recombination rate between them or not.

Candidate polymorphisms, identified from GWAS or F_{ST}-based outlier analyses for instance, most often need to be further studied with additional approaches such as gene expression profiling among individuals with contrasting trait expression or genetic engineering for instance, to corroborate these variants' involvement in trait variation (Ermann and Glimcher 2012). In these regards, annotations of genes encompassing or overlapping DNA segments harboring SNPs in LD with these candidate ones (referred as genes in genetic linkage with candidate SNPs in this paper) are crucial to support their biological significance and help prioritize subsequent investigations. Given the r^2 variability among populations and markers subsets, estimating an experiment-specific LD on both sides of one candidate SNP is an adequate procedure to find the nearby genes that are genetically linked to this candidate and select significant annotations. Even though a number of softwares and packages dedicated to genomic polymorphisms annotation already exist (Wang et al., 2010; Rope et al., 2011; Cingolani et al., 2012), they either only consider the sequences encompassing the candidate SNPs (Wang et al., 2010; Cingolani et al., 2012) or use LD estimates from a different population, usually a population of reference such HapMapII or the 1000 Genomes Project in Humans (Johnson et al., 2008; Machiella and Chanock 2015), thus leading to limited or biased results. Furthermore, candidate polymorphisms found lying outside gene sequence boundaries are often annotated using the closest gene annotation in non-human organisms, without estimating in the specific experiment the genomic regions in genetic linkage with those (e.g. Stanton-Geddes et al., 2013). Thus, we developed a new bioinformatics annotation tool that estimates LD in order to gather annotations from regions genetically linked to candidate polymorphisms, thus strengthening their potential and help prioritizing them for further analyses.

MATERIALS AND METHODS

Tested Datasets

When studying relationships between genetic markers and quantitative traits, research efforts usually involve testing and genotyping (1) hundreds to thousands of outbred individuals from natural populations, or (2) the progeny of a controlled cross between two individuals differing widely (i.e. segregating) for the trait of interest. In the first approach, individuals are sampled and later phenotyped in controlled and uniform conditions to perform a GWAS identifying candidate polymorphisms. In the second approach, a progeny is also assessed in controlled and uniform conditions, and the co-segregation of alleles and trait values allows to identify candidate SNPs. Both approaches have different assumptions regarding the levels of LD; average LD is usually moderate to low in association tests while very high in F1 progenies where many candidate SNPs are found in complete or nearly complete LD. Here, we tested our annotation pipeline with four different datasets to investigate a wide range of expected LD levels, originating from: (1) a domesticated animal, (2) a domesticated plant, and (3) a wild insect. These sets also varied in sampling size, numbers of tested SNPs, and candidate SNPs, thus further allowing to evaluate the pipeline's performance.

Domesticated Species Datasets

We applied our tool to annotate GWAS results in *Sus scrofa domesticus* which is characterized by high LD levels due to hundreds of years of selection to improve performance. This GWAS tested GBS data for association with meat quality (Prunier, Droit, Robert et al. unpublished) and was based on the genotyping of 196 individuals coming from two different breeding companies selecting sires and dams after each generation to improve meat quality in the Duroc pig breed (**Figure 1A**). The association tests yielded 199 candidate SNPs spread over the 18 autosomal chromosomes.

Even though the main focus of the present study is on GBS and SNP-array datasets, we also tested a dataset of 14,374,088 SNPs obtained from whole genome sequencing of the plant model *Medicago truncatula* varieties. These were investigated using GWAS for candidate genes involved in agronomic trait variations based on 226 accessions and representing as many inbred lines (Stanton-Geddes et al., 2013) (**Figure 1B**). The association study led to the identification of 1,537 candidate SNPs likely involved in variation of plant height or flowering timing, among other traits, and distributed over *Medicago*'s eight chromosomes. In order to run our pipeline, this publicly available dataset (www.medicagohapmap.org) was converted into a vcf file using bash commands and we tested both the entire set of SNPs and a set of SNPs with a minor allele frequency higher than 5%, yielding a total of 593,614 SNPs.

Wild Species Dataset

While three previous datasets were related to organisms with well described genomes, we finally assessed LD-annot capability to annotate candidate SNPs in a non-model, namely *Lymantria dispar* spp. This moth is an invasive species in North American forests as their caterpillars can successfully feed on foliage of numerous tree species (polyphagy) and therefore can damage vast tree plantations and natural forests. The co-segregation of SNP alleles and flying capabilities was followed over four generations (F2–F5) in this line resulting from the mating between a fully flying individual and a flightless individual in this species complex (**Figure 1C**). This analysis yielded a total of 250 SNPs possibly related to the moth's ability to fly.

Implementation

The LD-annot pipeline efficiently integrates a public package as well as new bash and python scripts to import SNP-array data, estimate SNP-specific genomic regions genetically linked to candidate SNP and extract corresponding gene annotations (**Figure 2**). It can be deployed on any Unix-based (or bash developer mode on Windows OS) following installation steps described here: https://github.com/ArnaudDroitLab/LD-annot/blob/master/README.md.

LD-annot uses the public package PLINK1.9 to calculate LD (r^2) levels. The user must define an r^2 threshold for limiting the region surrounding a candidate SNP in which annotations will be extracted, i.e. only polymorphisms linked to one candidate polymorphism with

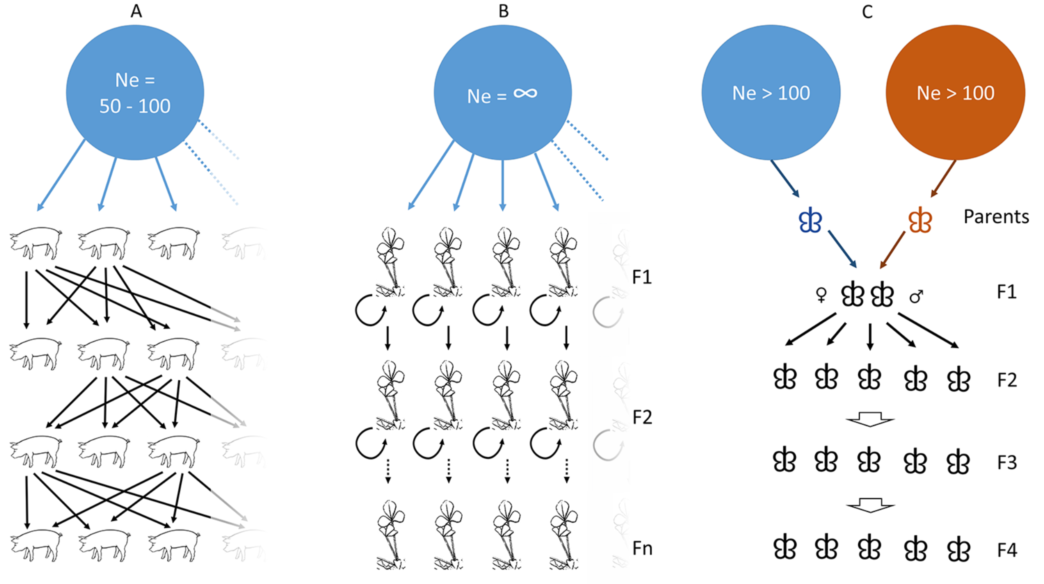
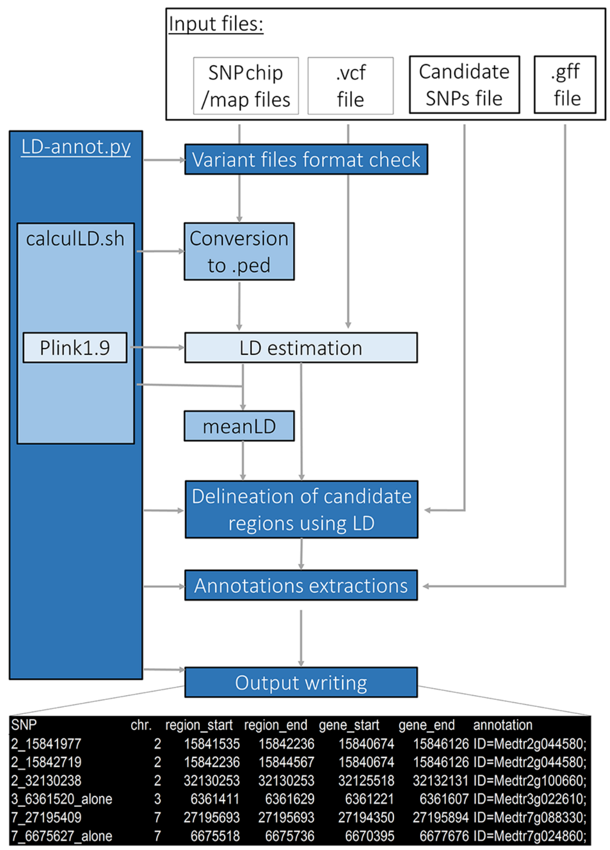


FIGURE 1 | Population and kinship history for the three types of datasets used as study cases. **(A)** the pig case in which trait-based genetic selection has been performed for centuries from a large ancestral population many generations ago; **(B)** the *Medicago* case in which inbred lines have been obtained from self-crossing of individuals originating from a very large population; **(C)** the Asian gypsy moth case where an introgressed progeny was obtained from mating between a flying individual and a non-flying individual, repeated over few generations.



```
SNP              chr.  region_start  region_end  gene_start  gene_end  annotation
2_15841977          2      15841535    15842236    15840674  15846126 ID=Medtr2g044580;
2_15842719          2      15842236    15844567    15840674  15846126 ID=Medtr2g044580;
2_32130238          2      32130253    32130253    32125518  32132131 ID=Medtr2g100660;
3_6361520_alone     3       6361411     6361629     6361221   6361607 ID=Medtr3g022610;
7_27195409          7      27195693    27195693    27194350  27195894 ID=Medtr7g088330;
7_6675627_alone     7       6675518     6675736     6670395   6677676 ID=Medtr7g024860;
```

FIGURE 2 | LD-annot overview. The LD-annot.py script is the master script that checks file format and calls a bash script for format conversion and PLINK LD estimation, and afterward calculates average LD and linked regions boundaries and gathers annotations for linked genes. At the bottom, an example header of the output file is presented.

a LD value superior to this threshold will be considered to delineate the region of interest (**Figure 2**). The pipeline includes a format check of input files and a checkpoint procedure. The latter allows to restart the analysis with different thresholds for r^2 for instance, without rerunning the format checks nor pairwise LD calculations, thus avoiding to run all steps and reducing the time for the analysis.

Command and Parameters

The pipeline is launched using only a single command line containing the parameters and paths for input files. In addition, LD-annot.py calls a bash script (calculLD.sh) that must be placed in the same folder. The command using vcf format input file is:

```
python3 LD-annot.py geno.vcf annot.gff3 candidate\
type thr output
```

while the command using SNP-array input file is:

```
python3 LD-annot.py PathToSnpFiles annot.gff3\
candidate type thr output SNP_Map
```

where "type" is the feature (mRNA, CDS, gene), "thr" is the threshold for r^2, and "SNP_Map" is a txt file providing chromosome and position identifiers for each SNP included on the SNP-array.

Inputs

The LD-annot pipeline is based on three different inputs.

The first input contains all genotypes for the studied population; this file is usually in vcf format obtained from a variant caller [Haplotypecaller or Platypus, for instance (DePristo et al., 2011; Rimmer et al., 2014)] for next-generation genotyping such as GBS data, or a folder including all individuals' genotypes in the case of SNP-array genotyping. In the latter case, genotyping is usually spread over txt files, one for each individual, which contain polymorphisms names and genotypes after 12 lines of comments and headers. In the case of GBS data, the vcf file is directly converted by PLINK1.9 before running LD calculations. In the case of SNP-array data, a formatting step is performed before LD calculations using PLINK1.9. This bash script gathers all individuals' genotypes included in the designated folder and converts this information into a .ped, .map, and .fam files for PLINK1.9 by making use of an additional input file providing the chromosome and position for each SNP on the SNP-array. Afterward, .ped files are converted to .bed files to save memory space and running time for both types of data, and r^2 are then calculated using PLINK1.9 (**Figure 2**).

The annotation file is a text file respecting a gff-like format (gff, gtf, or gff3) including the chromosome number/name in the first column, the feature in the third column (CDS, mRNA, exon), the starting and ending positions in respectively the fourth and fifth columns, and the annotation (= attributes) in the last column.

Finally, the third file contains the list of candidate SNPs with chromosome name in the first column, position in the second column, and SNP_ID in the third column (not required).

Note that the chromosome identification should be consistent among the various files; the number may often be prefixed with a "chr" or not. As this is the most likely source of errors and incompatibility, the format checking step generates error messages pointing at corrupt files and probable causes.

Linkage Calculation and Annotation Extraction

Linkage disequilibrium is estimated using the r^2 correlation score calculated using PLINK for genotyped SNPs located on the same chromosome in linkage for $r^2 > 0.4$. This low threshold is defined as the lowest one that a user may select. The threshold defined by the user is used later in the pipeline when estimating an average distance in linkage with candidate SNPs according to this threshold, and during delineation of genomic regions in linkage with each candidate SNP for annotations extraction.

Based on the LD calculations previously computed and the r^2 threshold set by the user, annotations from a .gff/.gff3/.gtf-type file are then gathered to create an annotation file for each candidate variant. A ".gff/.gff3/.gtf" file usually includes annotations for different features (mRNA, CDS, exon, gene) which represents a hierarchical classification of the same genomic regions and thus results in some repetition of the information. According to the approaches deployed to annotate the reference genome, the level of its completeness or the biological question asked in the research, one might favor one over the other features. Thus, LD-annot offers an option to select the feature of interest and avoid redundancy of the information at the various levels (i.e. gene, mRNA, and exon), which also make it flexible to any feature that may be indicated in the annotations file.

After input format checking and r^2 calculations, the python script gathers chromosome, position, and annotation for the designated feature. Afterward, it makes a dictionary of "candidate" regions (chr, start, and end) around candidate SNPs by using the position of the foremost upward and downward SNPs in linkage with each one of those candidates according to r^2 threshold chosen by the user. However, a candidate SNP might not be surrounded by other genotyped SNPs because of true absence of polymorphisms (possibly in a specific sampling set) or low quality genotyping. In such cases, the average distance calculated earlier in the pipeline is used to delineate the region of interest around such candidates and an "alone" flag is added to the candidate SNP name in the output file. It should be noted that this average is a broad estimate and those results should be interpreted with caution given the r^2 variability along the genome, and the possibility of the non-Gaussian distribution of distances between SNPs in LD.

Finally, all annotated regions with the selected feature in .gff/.gff3/.gtf file that overlap the "candidate" region are included into an output file that provides: chromosome, candidate SNP position, region start and end positions, annotation start and end positions, and the annotation *per se*. According to the number of annotations overlapping the candidate region, a candidate SNP can be found several times in the output file.

RESULTS AND DISCUSSION

LD-annot Performances

We assessed the performance of our tool through the analysis of the four datasets previously described and covering a large distribution in numbers of genotyped and candidate SNPs, and a variety of r^2 thresholds. The goal being to make this procedure amenable to researchers without coding skills nor access to high-performance infrastructures, we ran the pipeline using a common laptop computer with 4CPU cores and 8 Gbytes of RAM.

As expected, there was a significant correlation between the number of variants included in the analysis and the processing time (ANOVA, $p < 2e\text{-}16$; **Figure 3**). However, a single analysis never exceeded 16.1 min despite the very large SNP set (> 14M SNPs) originating from *Medicago* (**Table 1**). In such case, making use of the checkpoint feature allowed to reduce the computational time from 16.1 min to less than 10 (**Figure 3A**). As datasets are always increasing in size with technological progress and the usual need to test several r^2 thresholds, we believe the checkpoint procedure will be beneficial to the genomics research community.

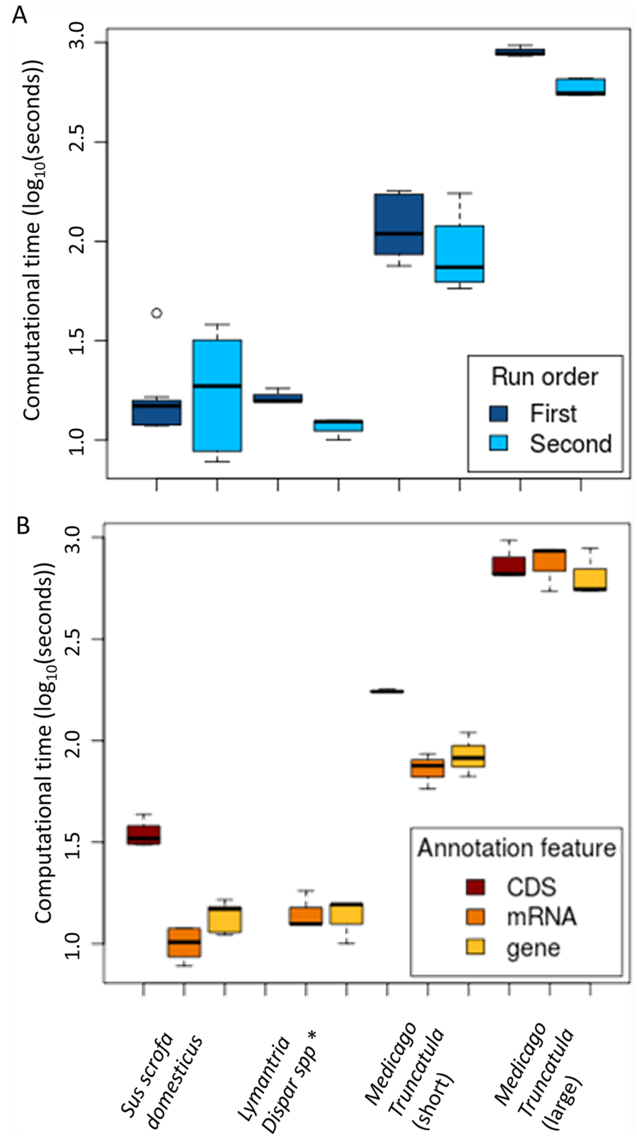


FIGURE 3 | Pipeline performances according to the run number **(A)** and the type of annotated features **(B)**. **(A)** LD-annot involves a checkpoint procedure that does not require rerunning each step when testing several LD thresholds, which results in shorter turnover of analysis after its first run. **(B)** The type of feature has an impact on the time for analysis since mRNA and CDS features are usually more complex than gene features in an annotation file. *Note that no CDS annotations were available for the *Lymantria dispar* genome.

Another factor impacting the analysis time is the size of the annotation file and particularly the type of feature specified by the user in the command line. Annotation files (.gff/.gff3/.gtf) typically harbor more annotation lines in the "CDS" feature than for "gene" or "mRNA." As a result, the analyses were significantly longer when searching for "CDS" feature annotations (ANOVA, $p = 0.0137$; **Figure 3B**). In line with this trend, regions linked to candidate SNPs extended when the r^2 threshold increased, resulting in an increasing number of annotations and time length for the analysis, although the difference was not significant.

TABLE 1 | LD-annot time analysis according to the sizes of SNP sets and candidate SNP sets.

Dataset*	Total SNPs set size	Candidate SNP number	Time (s)	r^2 threshold	Average distance (bp)†
Sus1	54,712	199	18.3	0.7	50494
Sus1	54,712	199	19.3	0.9	18000
Sus2	54,712	199	20.0	0.7	53614
Sus2	54,712	199	21.0	0.9	17430
Lymantria	321,868	250	13.5	0.7	6191
Lymantria	321,868	250	14.0	0.9	4620
Medicago	593,614	1,536	109.7	0.7	706
Medicago	593,614	1,536	110.6	0.9	601
Medic-large	14,374,089	1,536	581.6	0.7	44
Medic-large	14,374,089	1,536	692.5	0.9	33

**Sus1 and Sus2: the two pig genotyping-by-sequencing datasets; Lymantria: the gypsy moth SNP set; Medicago: the public Medicago dataset after filtering for low minor allele frequencies; Medic-large: the entire SNP set for Medicago (Stanton-Geddes et al., 2013).*

†Average distance between a pair of SNPs in linkage disequilibrium according to the threshold for r^2 estimated from all SNPs in the dataset.

Average Distance

The LD-annot pipeline calculates an average distance (in bp) separating two SNPs in LD according to the specified r^2 threshold across the whole dataset. This distance is later used to delineate a linked region around a candidate SNP (the average distance on both sides) when there is no surrounding genotyped SNPs. This distance is a function of inbreeding as illustrated by our datasets where the higher the original effective population size, the shorter is the distance in LD. Even within the pig species, the pedigree denoted *Sus1* generally presented shorter distances than *Sus2* pedigree which was developed from a smaller effective population of sires and dams.

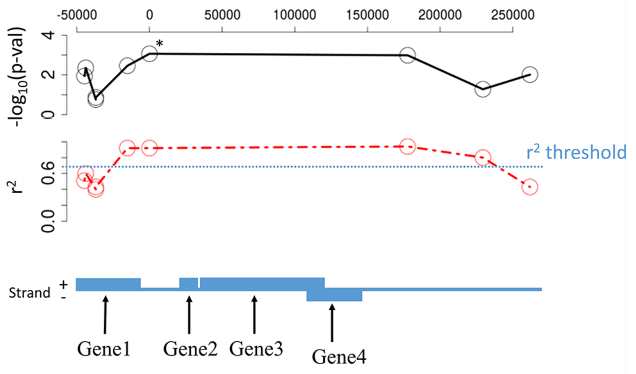
This distance is also varying according to the number of genotyped SNPs which is related to the occurrence of rare SNPs that tend to present lower r^2 values than more common SNPs (Pritchard and Przeworski 2001; Pe'er et al., 2006). As a result, removing SNPs with minor allele frequency <0.05 resulted in a sizable increase in distances (up to 18-fold) when testing the *Medicago* SNP set.

When genotyping a sample set using GBS approach, the SNP distribution over the genome is not controlled and the proportion of the genome interrogated by the genotyping is often an important question for researchers. The average distance provided by the tool can further be used to broadly estimate the genome coverage given the r^2 thresholds. For instance, using 54,712 SNPs in the *Sus1* pedigree allowed to investigate the entire 2.4Gb *Sus scrofa* genome with $r^2 > 0.7$, but 82% and only 40% of this genome with $r^2 > 0.8$ and 0.9, respectively. The same SNP set in the *Sus2* pedigree allowed to investigate 100, 87, and 38% of the genome with $r^2 > 0.7$, 0.8, and 0.9, respectively. However, these coverage values should be seen as broad estimates and, therefore, interpreted with caution given r^2 variability across the genome.

Why Not Consider Only the Closest Gene?

Selecting annotations associated with a candidate polymorphism is usually accomplished using the proximity criteria, in other words, the gene including the SNP in its sequence or the closest gene for non-coding SNP is often seen as the relevant one (e.g. Stanton-Geddes et al., 2013). However, other remote genes might be in genetic linkage with the candidate SNP while not presenting SNP in the studied SNP set, which does not allow to test their association *per se*. Even when presenting SNPs, these genes may have been missed because of too many missing genotypes or too low minor allele frequency for a specific locus which, in turn, did not permit to significantly detect them as candidate SNPs. For instance, when using LD-annot in *Sus scrofa*, we found a total of 334 genes in genetic linkage with only 176 of the candidate SNPs while the remaining candidate SNPs were not linked to any genes using an r^2 threshold >0.7. We even observed six cases of annotations for distant genes (second or third order of the closest genes and still in LD with the candidate SNP using $r^2 > 0.9$) that were in fact more informative with regards to the trait of interest than the closest one (**Figure 4**).

Contrastingly, the closest gene might be far away and not genetically linked with the candidate SNP which could lead to biased interpretation, particularly when performing enrichment



Annotations
Gene1: collagen and fibril organization and catabolism
Gene2: cell differentiation, multicellular development
Gene3: **regulation of growth, skeletal development**
Gene4: phenylalanine catabolism

FIGURE 4 | Illustration of one candidate SNP likely involved in pig meat quality that is genetically linked to four different genes; note that the most biologically meaningful is not the closest one but of the third order. The candidate SNP is at the position "0" upon the chromosome and marked with an asterisk; $-\log_{10}$(p-val) is the p-value for the association test between allelic variation and meat quality; r^2 is the correlation coefficient calculated in the dataset (red line) using PLINK and the specified threshold for linkage was 0.7 (blue line).

analyses. In *Medicago*, over the 1,536 candidate SNPs that were annotated using the closest gene (Stanton-Geddes et al., 2013), only 541 SNPs were actually genetically linked with their target gene ($r^2 > 0.7$). On the other hand, 40 candidate SNPs were genetically linked with two genes, and 62 annotated genes were linked to more than one candidate SNP (**Supplementary Table 1**), hence showing the importance of taking into account the LD when looking at annotations supporting the importance of a candidate SNP.

In the case of progenies study (gypsy moth case), the LD level is very high which resulted in blocks of several candidate SNPs genetically linked together, thus defining large regions possibly encompassing several genes. However, only 100 SNPs were in linkage with 64 genes ($r^2 > 0.9$) among the 250 candidate SNPs spread over 103 contigs. Despite the high level of LD and that all scaffolds harboring a candidate SNP were also encompassing one gene at the very least (2.39 genes in average), some candidate SNPs were not found in genetically linked with any gene. The distribution of recombination rates was not continuous as expected given the low number of individuals and generations, and LD breakpoints were observed along scaffolds. Thus, a SNP might be relatively close to a gene but still not representing it. Altogether, these results illustrate the need to evaluate the experiment-specific LD surrounding candidate SNPs when employing genes to annotate and prioritize these for further investigations, and understand the mechanisms underlying their association with trait variation.

CONCLUSION

The LD-annot tool yields supporting lines of evidence to help identify biologically meaningful genetic polymorphisms underlying phenotypic variation. It can be used with any sort of annotations and polymorphism data as long as the input format matches either SNP-chips or vcf files. One can obtain annotations for repeats or specific methylation sites, for instance, and use this tool to identify those features that are statistically linked to candidate SNPs for a given sampling.

DATA AVAILABILITY STATEMENT

Medicago data can be found in Stanton-Geddes et al. 2013. Data generated in this study are included in the article/**Supplementary Material**. Scripts are available at: https://github.com/ArnaudDroitLab/LD-annot/.

AUTHOR CONTRIBUTIONS

JP developed and coded the bioinformatics tool with help from AL and AB, and tested it using the various datasets. MJ gathered the pig meat quality measurements. IP obtained the funding allowing to sequence the gypsy moth pedigree and JP identified candidate SNPs for flight in this pedigree. CR and AD obtained the funding to sequence pig individuals and support the bioinformatics tool development. All co-authors read and edited the manuscript.

FUNDING

This work, including genotyping-by-sequencing data in pig, has been funded by the "chips-for-better-chops" project financially supported by Genome Canada (Genomic Applications Partnership Program), the Canadian Centre for Swine Improvement, FastGenetics Inc. and Olymel Inc. Asian gypsy moth sequence data originated from the BioSafe project financially supported by Genome Canada, Genome BC and Genome Quebec and the Canadian Food Inspection Agency.

ACKNOWLEDGMENTS

We thank Simon Gotty, Gwenaëlle Lemoine, and Charles Joly-Beauparlant (Ph.D.) from Arnaud Droit's team for their assistance with Python, Windows deployment, and Github public release, as well as Gwylim Blackburn, Ph.D., from the Pacific Forestry Centre (Natural Resources Canada) for his support and testing of the first versions of the tool. We also want to thank Daniela Grossi from FastGenetics and Brian Sullivan from the Canadian Centre for Swine Improvement Inc. for comments regarding the first draft of the manuscript. We finally thank reviewers for their comments and suggestions regarding the submitted manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.01192/full#supplementary-material

REFERENCES

Bai, B., W, Le, Zhang, Y. J., Lee, M., Yuzer Alfiko, R. R., and Ye, B. Q. (2018). Developing genome-wide SNPs and con-structing an ultrahigh-density linkage map in oil palm. *Sci. Rep.* 8 (1), 691. doi: 10.1038/s41598-017-18613-2

Bush, W. S., and Moore, J. H. (2012). Chapter 11: genome-wide association studies. *PloS Comput. Biol.* 8 (12), e1002822. doi: 10.1371/journal.pcbi.1002822

Carter, A., Tenuta, A., Rajcan, I., Welacky, T., Woodrow, L., and Eskandari, M. (2018). Identification of quantitative trait loci for seed isoflavone concentration in soybean (Glycine Max) against soybean cyst nematode stress. *Plant Breed. = Z. Fur Pflanzenzuchtung* 137 (5), 721–729. doi: 10.1111/pbr.12627

Carvalho, B., Bengtsson, H., Speed, T. P., and Irizarry, R. A. (2007). Exploration, normalization, and genotype calls of high-density oli-gonucleotide SNP array data. *Biostatistics* 8 (2), 485–499. doi: 10.1093/biostatistics/kxl042

Cingolani, P., Platts, A., Wang, L. L., Coon, M. T., Nguyen, et al. (2012). A program for annotating and predicting the effects of single nucleotide polymor-phisms, SnpEff: SNPs in the genome of drosophila melanogaster strain w1118; Iso-2; Iso-3. *Fly* 6 (2), 80–92. doi: 10.4161/fly.19695

DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., et al. (2011). A frame-work for variation discovery and genotyping using next-generation dna sequencing data. *Nat. Genet.* 43 (5), 491–498. doi: 10.1038/ng.806

Elshire, R. J., Glaubitz, J. C., Sun, Q., Poland, J. A., Kawamoto, K., . Buckler, E. S., et al. (2011). A robust, simple geno-typing-by-sequencing (GBS) approach for high diversity species. *PloS One* 6 (5), e19379. doi: 10.1371/journal.pone.0019379

Ermann, J., and Glimcher, L. H. (2012). After GWAS: Mice to the Rescue?. *Curr. Opin. In Immunol.* 24 (5), 564–570. doi: 10.1016/j.coi.2012.09.005

Hess, J. E., Campbell, N. R., Close, D. A., Docker, M. F., and Narum, S. R. (2012). Population genomics of pacific lamprey: adaptive variation in a highly dispersive species. *Mol. Ecol.* 22 (11), 2898–2916. doi: 10.1111/mec.12150

Hill, W. G., and Robertson, A. (1968). Linkage disequilibrium in finite populations. *TAG. Theor. Appl. Genet. Theoretische Und Angewandte Genetik* 38 (6), 226–231. doi: 10.1007/BF01245622

Johnson, A. D., Handsaker, R. E., Pulit, S. L., Nizzari, M. M., O'Donnell, C. J., and De Bakker, P. I. (2008). SNAP: a web-based tool for identification and annotation of proxy SNPs using HapMap. *Bioinformatics* 24 (24), 2938–2939.

Jorgenson, E., and Witte, J. S. (2006). Coverage and power in genomewide association studies. *Am. J. Hum. Genet.* 78 (5), 884–888. doi: 10.1086/503751

Keller, I., Wagner, C. E., Greuter, L., Mwaiko, S., Selz, O. M., Sivasundar, A., et al. (2013). Population genomic signatures of diver-gent adaptation, gene flow and hybrid speciation in the rapid radiation of lake victoria cichlid fishes. *Mol. Ecol.* 22 (11), 2848–2863. doi: 10.1111/mec.12083

Machiela, M. J., and Chanock, S. J. (2015). LDlink: a web-based application for exploring population-specific haplotype structure and linking correlated alleles of possible functional variants. *Bioinformatics* 31 (21), 3555–3557.

Narum, S. R., Buerkle, C. A., Davey, J. W., Miller, M. R., and Hohenlohe, P. A. (2013). Genotyping-by-sequencing in ecological and conser-vation genomics. *Mol. Ecol.* 22 (11), 2841–2847. doi: 10.1111/mec.12350

Pe'er, I., Chretien, Y. R., de Bakker, P. I. W., Barrett, J. C., Daly, M. J., and Altshuler, D. M. (2006). Biases and reconciliation in estimates of linkage disequilibrium in the human genome. *Am. J. Hu-man Genet.* 78 (4), 588–603. doi: 10.1086/502803

Pritchard, J. K., and Przeworski, M. (2001). Linkage disequilibrium in humans: models and data. *Am. J. Hum. Genet.* 69 (1), 1–14. doi: 10.1086/321275

Reich, D. E., Cargill, M., Bolk, S., Ireland, J., Sabeti, P. C., Richter, D. J., et al. (2001). Linkage Disequilibrium in the Human Genome. *Nature* 411 (6834), 199–204. doi: 10.1038/35075590

Rimmer, A., Phan, H., Mathieson, I., Iqbal, Z., Twigg, S. R. F., Wilkie, A. O. M., et al. (2014). Integrating mapping-, assembly- and haplotype-based ap-proaches for calling variants in clinical sequencing applications. *Nat. Genet.* 46 (8), 912–918. doi: 10.1038/ng.3036 WGS500 Consortium.

Roe, A. D., Torson, A. S., Bilodeau, G., Bilodeau, P., Blackburn, G. S., Cui, M., et al. (2018). Biosurveillance of forest insects: part i—integration and application of genomic tools to the surveillance of non-native forest insects. *J. Pest Sci.* 92 (1), 51–70. doi: 10.1007/s10340-018-1027-4

Rope, A. F., Wang, K., Evjenth, R., Xing, J., Johnston, J. J., et al. (2011). Using VAAST to identify an X-linked disorder resulting in lethality in male infants due to n-terminal acetyltransferase deficiency. *Am. J. Hum. Genet.* 89 (1), 28–43. doi: 10.1016/j.ajhg.2011.05.017

Smith, A. V., Thomas, D. J., Munro, H. M., and Abecasis, G. R. (2005). Sequence features in regions of weak and strong linkage disequilib-rium. *Genome Res.* 15 (11), 1519–1534. doi: 10.1101/gr.4421405

Sonah, H., O'Donoughue, L., Cober, E., Rajcan, I., and Belzile, F. (2015). Identification of loci governing eight agronomic traits using a GBS-GWAS approach and validation by QTL mapping in soya bean. *Plant Biotechnol. J.* 13 (2), 211–221. doi: 10.1111/pbi.12249

Stanton-Geddes, J., Paape, T., Epstein, B., Briskine, R., Yoder, J., Mudge, J., et al. (2013). Candidate genes and genetic architecture of symbiotic and agronomic traits revealed by whole-genome, sequence-based association genetics in medicago truncatula. *PloS One* 8 (5), e65688. doi: 10.1371/journal.pone.0065688

Torkamaneh, D., Boyle, B., and Belzile, F. (2018). Efficient ge-nome-wide genotyping strategies and data integration in crop plants. *TAG. Theor. Appl. Genet. Theoretische Und Angewandte Genetik* 131 (3), 499–511. doi: 10.1007/s00122-018-3056-z

VanLiere, J. M., and Rosenberg, N. A. (2008). Mathematical properties of the r2 measure of linkage disequilibrium. *Theor. Population Biol.* 74 (1), 130–137. doi: 10.1016/j.tpb.2008.05.006

Wang, K., Li, M., and Hakonarson, H. (2010). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data.". *Nucleic Acids Res.* 38 (16), e164–e164. doi: 10.1093/nar/gkq603

White, T. A., Perkins, S. E., Heckel, G., and Searle, J. B. (2013)Adaptive evolution during an ongoing range expansion: the invasive bank vole (Myodes Glareolus) in ireland. *Mol. Ecol.* 22 (11), 2971–2985. doi: 10.1111/mec.12343

Conflict of Interest: The authors declare that this study received funding from FastGenetics Inc. and Olymel Inc. which also provided meat quality data for the pigs' use-cases. 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.

frontiers in Genetics

ORIGINAL RESEARCH
published: 12 February 2020
doi: 10.3389/fgene.2020.00053

epiCOLOC: Integrating Large-Scale and Context-Dependent Epigenomics Features for Comprehensive Colocalization Analysis

*Yao Zhou [1], Yongzheng Sun [1], Dandan Huang [1] and Mulin Jun Li [1,2]**

[1] Department of Pharmacology, Tianjin Key Laboratory of Inflammation Biology, School of Basic Medical Sciences, National Clinical Research Center for Cancer, Tianjin Medical University Cancer Institute and Hospital, Tianjin Medical University, Tianjin, China, [2] Collaborative Innovation Center of Tianjin for Medical Epigenetics, Tianjin Key Laboratory of Medical Epigenetics, Tianjin Medical University, Tianjin, China

OPEN ACCESS

Edited by:
Geir Kjetil Sandve,
University of Oslo,
Oslo, Norway

Reviewed by:
Mikhail Dozmorov,
Virginia Commonwealth University,
Richmond, United States
Enrique Medina-Acosta,
Universidade Estadual do Norte
Fluminense Darcy Ribeiro,
Brazil

Correspondence:
Mulin Jun Li
mulinli@connect.hku.hk

Specialty section:
This article was submitted to
Bioinformatics and
Computational Biology,
a section of the journal
Frontiers in Genetics

Received: *08 September 2019*
Accepted: *17 January 2020*
Published: *12 February 2020*

Citation:
Zhou Y, Sun Y, Huang D and Li MJ (2020) epiCOLOC: Integrating Large-Scale and Context-Dependent Epigenomics Features for Comprehensive Colocalization Analysis. Front. Genet. 11:53. doi: 10.3389/fgene.2020.00053

High-throughput genome-wide epigenomic assays, such as ChIP-seq, DNase-seq and ATAC-seq, have profiled a huge number of functional elements across numerous human tissues/cell types, which provide an unprecedented opportunity to interpret human genome and disease in context-dependent manner. Colocalization analysis determines whether genomic features are functionally related to a given search and will facilitate identifying the underlying biological functions characterizing intricate relationships with queries for genomic regions. Existing colocalization methods leveraged diverse assumptions and background models to assess the significance of enrichment, however, they only provided limited and predefined sets of epigenomic features. Here, we comprehensively collected and integrated over 44,385 bulk or single-cell epigenomic assays across 53 human tissues/cell types, such as transcription factor binding, histone modification, open chromatin and transcriptional event. By classifying these profiles into hierarchy of tissue/cell type, we developed a web portal, epiCOLOC (http://mulinlab.org/epicoloc or http://mulinlab.tmu.edu.cn/epicoloc), for users to perform context-dependent colocalization analysis in a convenient way.

Keywords: colocalization, epigenomics and epigenetics, functional annotation analysis, genetic variants, cell type specific, web server

INTRODUCTION

The epigenome, beyond genome sequence, has been increasingly recognized as key component in the gene regulation to drive certain biological processes and associate with many human diseases (Lawrence et al., 2016; Dor and Cedar, 2018; Feinberg, 2018). In the past decades, high-throughput epigenomic sequencing assays have profiled large numbers of functional elements across numerous human tissues/cell types, such as histone modification, DNA methylation, open chromatin, transcription factor binding site (TFBS), etc. The International Human Epigenome Consortium (IHEC) project (Bujold et al., 2016) have been initialized, across different countries and consortiums, to coordinate the production of reference maps of human epigenomes for key cellular states relevant to health and diseases. These unprecedented growths of epigenetic profiles and following comprehensive analysis of tissue/cell type-specific epigenomes will ultimately lead to a better understanding of how human population and genome function are shaped in response to the environment (Egtex, 2017).

To facilitate convenient and accurate utilization of increasing volume of epigenomic data, several commonly-used resources have uniformly processed raw profiles and made them easily accessible, including ENCODE (Consortium, 2012), Roadmap Epigenomics (Roadmap Epigenomics et al., 2015), Blueprint Epigenome (Stunnenberg et al., 2016) and CistromeDB (Mei et al., 2017; Zheng et al., 2019). Furthermore, comprehensive epigenomics accumulation has motivated novel computational methods of modelling functional elements across many tissues/cell types, such as ChromHMM (Roadmap Epigenomics et al., 2015) and Segway (Libbrecht et al., 2019). Therefore, integrating such large-scale and context-dependent epigenomics features for novel biological findings is in urgent demand (Dozmorov, 2017; Cazaly et al., 2019). To this end, colocalization analysis was frequently used to study the interplay of various functional elements in different biological processes and conditions, where potential enrichment of a given genomic/epigenomic profile in pre-defined dataset could be drawn from the global perspective (Kanduri et al., 2019). Integrated with large-scale tissue/cell type-specific epigenomics data, colocalization analysis provides a powerful avenue to investigate biological relations and cell type specificities, such as identifying co-occurrence of transcription regulators (Yan et al., 2013) and inferring causal tissues/cell types from disease-associated variants identified by genome-wide association study (GWAS) (Farh et al., 2015).

Many colocalization tools have been developed by holding diverse assumptions and background models to assess the significance of enrichment. For instances, GSuite HyperBrowser is a web-based tool that performs colocalization analysis using either analytical approaches or Monte Carlo simulations (Simovski et al., 2017). LOLA utilizes Fisher's exact test based on universe regions to inspect enrichment and provides a web-based portal LOLAweb (Sheffield and Bock, 2016; Nagraj et al., 2018). GoShifter (Trynka et al., 2015) and GARFIELD (Iotchkova et al., 2019), which were implemented into standalone tools, specifically quantify enrichment of overlaps between GWAS variants and genomic annotations by considering linkage disequilibrium (LD). To overcome the discordant enrichment among exiting methods, Coloc-stats integrates multiple colocalization analysis tools in a single web interface (Simovski et al., 2018). This integrated system serves as a one-stop shop for performing comprehensive colocalization analysis and asseses the consistency of the conclusions across seven different methods. However, some critical issues remain unaddressed. First, existing tools only provide limited pre-defined sets for genomic features in different biological domains. Current web-based tools, such as GSuite HyperBrowser, GenomeRunner (Dozmorov et al., 2016) and LOLAweb, only incorporate a small number of epigenomic profiles from ENCODE, Cistrome and other specific annotation datasets, which restrict the broader applications of online colocalization analysis. Second, the descriptions of tissue and cell type information are disordered and only based on free text, making current tools unable to properly classify or group tissues/cell types to inspect the specificity of enrichment. Therefore, a uniform human tissue/cell-type definition is needed. Furthermore, the growing volume of epigenomic profiles on extensive tissues/cell types, collection and integration of these genomic features require a great effort to download. Most colocalization web tools are time-consuming for features intersection and background generation when dealing with such accumulating data scale. To ease the comprehensive colocalization analysis for biologists and geneticists, a faster and versatile online platform would be welcome.

For this study we comprehensively collected and integrated over 44,385 bulk or single cell epigenomic profiles across 53 human tissues/cell types. By classifying and mapping these profiles into hierarchy of tissue/cell type, we developed a web portal, epiCOLOC, for users to perform context-dependent colocalization analysis in a convenient way. We leveraged a recent ultrafast genomics search engine, GIGGLE, to identify and prioritize the enrichment of genomic loci shared between query features and our pre-defined epigenomic interval files (Layer et al., 2018). epiCOLOC equips many visualization functions and is freely available at http://mulinlab.org/epicoloc or http://mulinlab.tmu.edu.cn/epicoloc.

EPIGENOMIC PROFILES INTEGRATION AND PROCESSING

Data Collection

We collected human genomic and epigenomic data from various public resources including ENCODE (Consortium, 2012), Roadmap Epigenomics (Roadmap Epigenomics et al., 2015), Cistrome (Mei et al., 2017), ReMap (Cheneby et al., 2018), ChIP-Atlas (Oki et al., 2018), DeepBlue (Albrecht et al., 2017), BOCA (Fullard et al., 2018), TCGA (Corces et al., 2018) and HACER (Wang et al., 2019) (**Supplementary Table 3**). According to data sources and corresponding attributes, we classified collected features into following categories: 1) Transcriptional regulator, which incorporates ChIP-seq profiles of large number of transcriptional factors and chromatin remodelers; 2) Histone modification, which incorporates ChIP-seq profiles of different histone modifications; 3) Chromatin accessibility, which contains DNase-seq, ATAC-seq and FAIRE-seq profiles of open chromatin; We also curated several single cell ATAC-seq assays in this category; 4) Transcriptional event, which contains CAGE-seq, GRO-seq and PRO-seq profiles of nascent transcription signals; 5) Chromatin segmentation, which introduces tissue/cell type-specific chromatin states predicted by ChromHMM and Segway (**Figure 1A** and **Supplementary Table 1**). In order to improve accuracy and robustness of epiCOLOC backend database, we removed low-quality profiles according to the quality control scheme provided in the original resource. For example, we removed ChIP-seq data not passing two Cistrome quality metrics, including fraction of reads in peaks, and sufficient number of peaks with good enrichment. We also

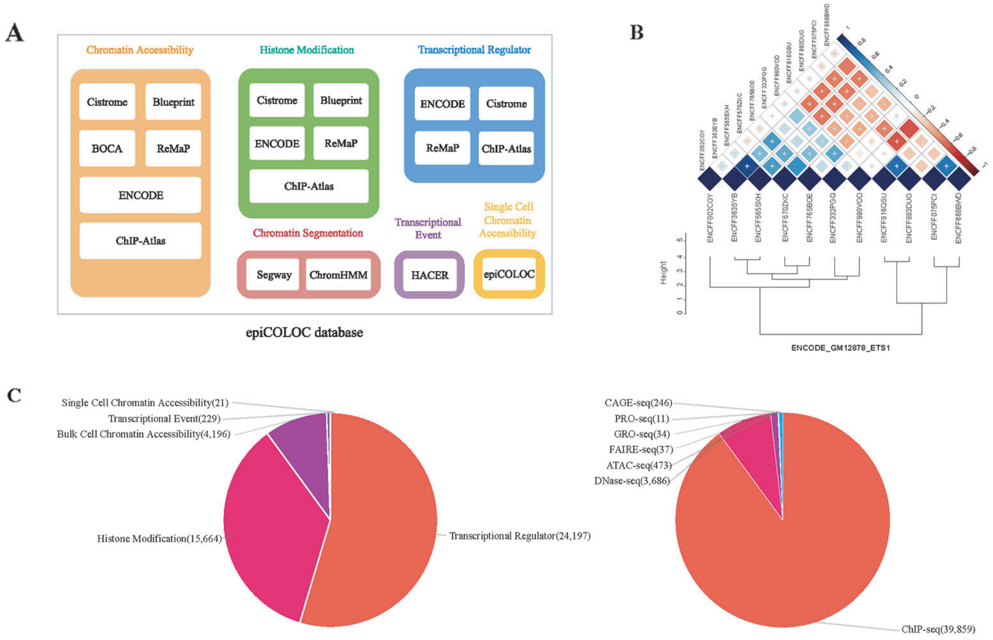


FIGURE 1 | The overview of epiCOLOC design and datasets. **(A)** The source schema of epiCOLOC data collection. **(B)** An example to illustrate outlier profiles removing. **(C)** The summary of data types in the current version of epiCOLOC.

excluded ENCODE profiles with error audit flags, such as extremely low read length, not tagged antibody, etc. Current epiCOLOC database covers 1,631 chromatin markers, which comprises 88 histone modifications, 1,538 transcriptional regulators, open chromatin and transcriptional event.

Data Processing

Tissue Organization and Mapping

We mapped cell lines to tissues by accounting for some auxiliary information from original epigenomic studies and several standards from GTEx (Consortium et al., 2017), Expression Atlas (Papatheodorou et al., 2018), Cellosaurus (Bairoch, 2018), ATCC (www.atcc.org), and BRENDA Tissue Ontologies (www.ebi.ac.uk/ols/ontologies/bto), yielding 53 main human tissues in total. For some main tissues that contain multiple well characterized components or some cell lines that cannot simply map to specific main tissues, we set independent terms in tissue set and finally generated 137 sub-tissues (**Supplementary Table 2**). We then manually mapped tissue/cell type name of each profile to our uniformly defined tissue set.

Cell Type Mapping

To reduce the complexity of cell type description in our collected epigenomic profiles, we performed cell type mapping using Cellosaurus that collected almost all cell line synonyms in a reference database (Bairoch, 2018). We acquired the Cellosaurus accession numbers and corresponding synonyms for all recorded cell lines, and assigned uniform synonyms identifiers to epigenomic profiles, which greatly reduces the heterogeneity of cell type descriptions. For cancer cell types mapping, we borrowed DepMap which provides standard terms for over thousands of cancer cell lines and organoid models (Van Der Meer et al., 2019). Since DepMap provides Cellosaurus accession numbers, we were able to easily map cancer cell lines to consistent reference.

Profile Grouping

Since the epigenomic data were generated by different laboratories or produced using different protocols, replicates and analysis methods among collected sources, we sought to identify profiles describing similar biological processes in each source. We grouped all collected profiles according to source + assay type + tissue/cell type + biological target, and assigned unique group identifiers to them.

Outlier Profiles Removal

To further ensure informative profiles in each group, we designed a strategy to eliminate potential outlier profiles that may deviate from underlying biological process of the group (**Supplementary Methods**). For each group with at least three profiles, we first constructed a pair-wise similarity matrix for all profiles based on GIGGLE combo score (Layer et al., 2018). Then, hierarchical clustering was used to cluster these profiles based on Euclidean distance and the optimal number of clusters was automatically determined by inconsistency coefficient method (Zahn, 1971). Furthermore, we only retained profiles within the largest cluster as representatives in this group. For example, we identified that four outlier profiles among 11 ETS1 ChIP-seq peak profiles in GM12878 cell line, and excluded them in the colocalization analysis (**Figure 1B**).

epiCOLOC Web Tool Implementation

The current version of epiCOLOC incorporates 44,385 tissue/cell type-specific functional profiles from 44,364 bulk-cell studies and 21 single-cell studies after quality control (**Supplementary Table 4**). Most of these profiles (89.8%) are derived from ChIP-seq for transcription regulators and histone modifications, while, 9.5% profiles came from DNase-seq and ATAC-seq for chromatin accessibility (**Figure 1C**).

Colocalization Method

To achieve a fast and efficient colocalization based on high volume epigenomic features, we embedded a genomic feature search engine, GIGGLE, into epiCOLOC web server (Layer et al., 2018). GIGGLE uses Fisher's exact test and odds ratio of “observed” versus “expected” to measure enrichment between query features and pre-indexed genomic intervals. It also creates a combination score called GIGGLE combo score, which is the product of -log10(Fisher's exact test *P*-value) and log2(odds ratio). Given thousands of epigenomic profiles in epiCOLOC database, GIGGLE can significantly reduce the running time from hours to minutes. For example, epiCOLOC takes about 6 minutes to finish colocalization analysis on transcriptional regulator profiles of all blood cells for a set of 10k intervals (randomly generated genomic intervals with varying length). For each profile group, we calculated median score to represent group-level enrichment. With the aid of efficient colocalization strategy, epiCOLOC tries to provide powerful context-specific epigenomic evidences, leading to novel biological problems identification, such as “Are two transcription factors (TFs) colocalized and forming cooperation” or “Are the query variants/intervals enriched in chromatin open regions of specific tissues?” or “Are the query variants/intervals overlap with transcribed enhancers regions more than would be expected by chance?” More biological examples can be found in our website http://mulinlab.org/epicoloc/Introduction/#Biological-examples.

Web Interface and Usage

epiCOLOC was implemented in a web-based tool with built-in large-scale and context-dependent epigenomic annotations. The epigenomic profiles were indexed using GIGGLE. The web server was developed by Python, jQuery, igv.js, amcharts.js and related JavaScript modules.

Querys

epiCOLOC accepts two types of genomic format: BED-like format and VCF-like format. Both plain text and uploaded file of regions of interest (ROIs) or variant positions are well supported. Uploaded file can be BED or VCF text file or compressed gzip file (<20Mb).

Options

epiCOLOC provides several options for users to customize colocalization analysis, including 1) select tissues (53 tissues/137 sub-tissues); 2) select profile categories (Transcriptional regulator, Histone modification, Chromatin accessibility, Transcriptional event, Chromatin segmentation); 3) change human genome assembly (GRCh37 and GRCh38); 4) define background genome size (3,095,677,412 for GRCh37 and 3,088,269,832 for GRCh38 in default); 5) set maximal interval length (500bp in default, and ROIs which exceed maximum length will be removed); 6) set extended length on both sides (no extension by default); 7) set central window size (cut the central area of genomic intervals, no central window by default).

Job Submission

Once submitted, the job will be sent to the backend of the web server for colocalization analysis. epiCOLOC displays a progress bar to track the execution status. It allows job retrieval by searching for the job ID in the home page, or by using a fixed URL (http://mulinlab.org/epicoloc/<jobid>) to check results directly, or through email notification.

Results Visualization

We used GIGGLE combo scores to prioritize colocalization results. Higher combo score indicates better enrichment on a specific profile, while negative combo scores suggest depleted enrichment (**Supplementary Figure 1**). Users can inspect and visualize the results in four different manners: 1) Prioritization table, which shows statistics metrics of colocalization including combo score, Fisher's exact *P*-value, odds ratio, the number of overlaps and extra information of enriched profiles (**Figure 2A**); 2) Tissue-wise pie charts for enrichment and depletion, which depict the per tissue proportion in all enriched (positive combo score) or depleted (negative combo score) profiles (**Figure 2B**). Users can click the slice of each tissue in the pie chart to see detailed sub-tissue results; 3) Tissue-wise bar plots, which display the representative enriched or depleted profiles in each tissue (**Figure 2C**). The user can search, scroll, zoom and hover over the bar plot to get detailed information of enrichment (only assay IDs for the best profiles in each group are displayed in hover tooltip). Once the label under the tissue-wise bar plotsis clicked, cell type-wise bars which depict enrichment patterns for the top 20 enriched cell types appear in a pop-up window. 4) The IGV dashboard displays relative genomic location for queries genomic intervals and top five enriched profiles in colocalization analysis.

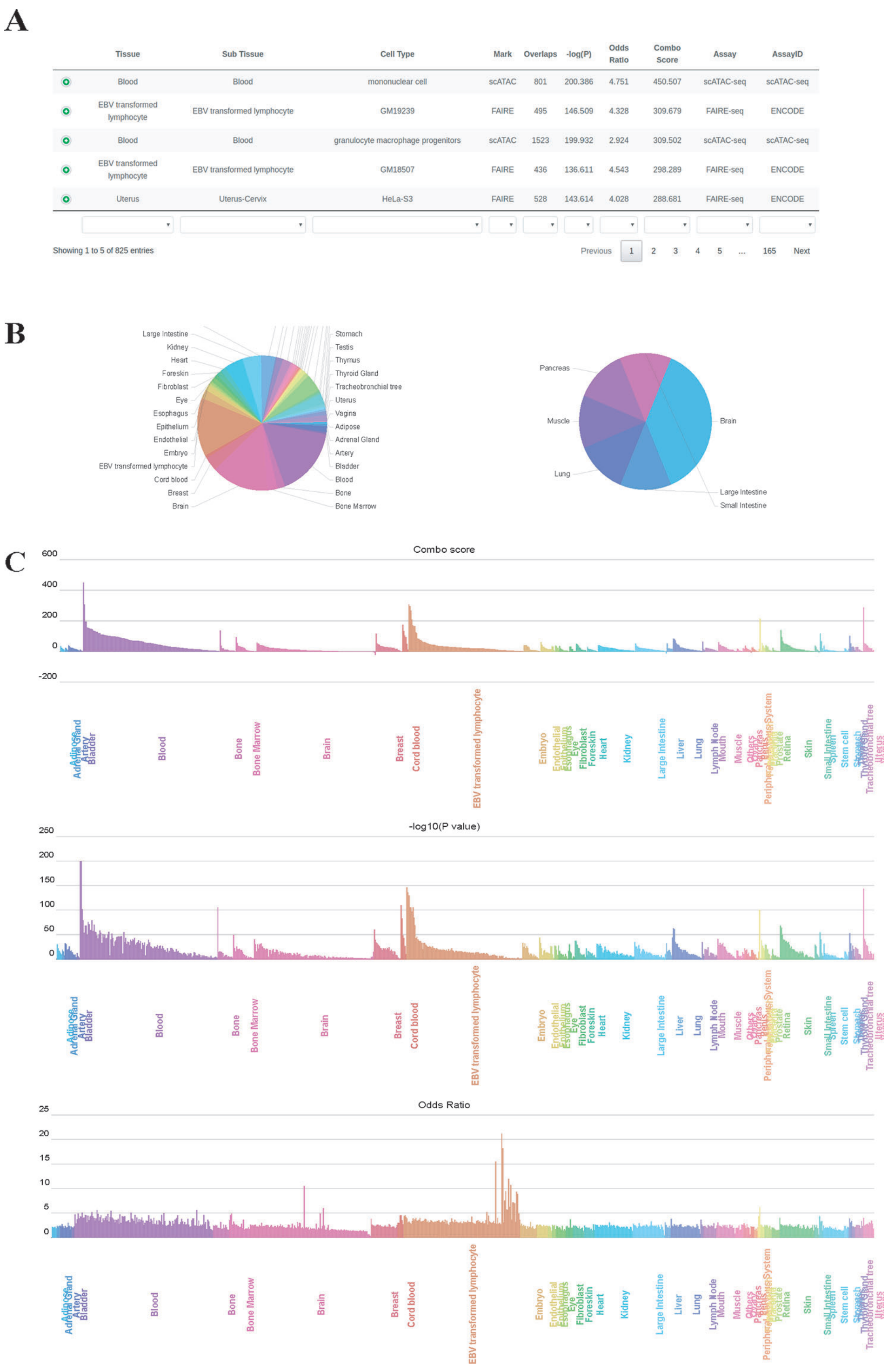
Download

epiCOLOC allows users to download colocalization results in csv format and result figures in png, jpg or pdf formats.

CASE STUDIES AND EVALUATIONS

By integrating large-scale tissue/cell type-specific epigenomic profiles, epiCOLOC could be used to investigate many biological questions. Here, we used several examples to demonstrate the performances and potential usages of epiCOLOC.

To identify potential disease-relevant genomic features and tissues using GWAS variants, we first performed colocalization analysis on disease-associated variants for inflammatory bowel



	Tissue	Sub Tissue	Cell Type	Mark	Overlaps	-log(P)	Odds Ratio	Combo Score	Assay	AssayID
⊕	Blood	Blood	mononuclear cell	scATAC	801	200.386	4.751	450.507	scATAC-seq	scATAC-seq
⊕	EBV transformed lymphocyte	EBV transformed lymphocyte	GM19239	FAIRE	495	146.509	4.328	309.679	FAIRE-seq	ENCODE
⊕	Blood	Blood	granulocyte macrophage progenitors	scATAC	1523	199.932	2.924	309.502	scATAC-seq	scATAC-seq
⊕	EBV transformed lymphocyte	EBV transformed lymphocyte	GM18507	FAIRE	436	136.611	4.543	298.289	FAIRE-seq	ENCODE
⊕	Uterus	Uterus-Cervix	HeLa-S3	FAIRE	528	143.614	4.028	288.681	FAIRE-seq	ENCODE

Showing 1 to 5 of 825 entries

FIGURE 2 | Results page of epiCOLOC. Colocalization result for IBD GWAS variants in open chromatin regions, **(A)** Prioritization table. **(B)** Pie chart that depicts the number of significant enriched or depleted profiles in each tissue. **(C)** Bar plots that display ordered combo score, *P*-value, odds ratio in tissue-wise manner.

disease (IBD) (Liu et al., 2015) to test the tissue-specific enrichment. Using chromatin accessibility features, we found that IBD GWAS variants (*P*-value < 5E-8) were significantly enriched in blood tissue, where open chromatin profiles on monocyte, lymphocyte and granulocyte macrophage progenitor received highest enrichment scores. (**Figure 2**, and also see colocalization result from: http://mulinlab.org/epicoloc/results/bc2fa49a-6dfa-40f1-bb61-1349c9118168). This result was consistent with GARFIELD results using functional annotations from ENCODE and Roadmap Epigenomics (Iotchkova et al., 2019). We then used coronary artery disease (CAD) GWAS variants (*P*-value < 5E-8) to perform colocalization in open chromatin regions (Van Der Harst and Verweij, 2018). Consistent with GARFIELD reports, we observed that most of tissues showed similar enrichment patterns, without distinct tissue specificity at open chromatin (http://mulinlab.org/epicoloc/results/63b0cd1b-f22f-43dd-9452-fdea114f6c3d). However, when using fine-mapped CAD variants, we observed several highly enriched signals in tissues like liver and artery blood vessel (http://mulinlab.org/epicoloc/results/04bf79a8-f7cd-4960-913e-5c5c84c05753), implying that the importance of selecting informative ROIs before colocalization analysis.

Next we sought to demonstrate that whether epiCOLOC could be used to identify potential cooperative factors for given TF. Transcription factor 7-like 2 (TCF7L2), a TF in the Wnt-signaling pathway, has been proven to play a central role in coordinating the expression of proinsulin and forming mature insulin (Zhou et al., 2014). TCF7L2 binding sites had been reported to colocalize with HNF4alpha and FOXA2 in HepG2 cell (Frietze et al., 2012). We hence used TCF7L2 ChIP-seq in HepG2 to perform colocalization analysis using epiCOLOC. In our colocalization results, TCF7L2 ChIP-seq peaks were significantly enriched in EP300, CREM, SP1, FOXA2 and HNF4alpha ChIP-seq profiles in various tissues/cell types (http://mulinlab.org/epicoloc/results/d736578a-59a4-4160-a6fe-1a9c420c4adf). Furthermore, we used two motif finding tools, PscanChIP (Zambelli et al., 2013) and HOMER (Heinz et al., 2010), with the same query input to investigated enriched TF motifs. We found that TF motifs including HNF4alpha, FOXA2, TCF7, GATA4, FOXP1, FOXA1, FOXK2 and FOXO3 can be simultaneously identified among two motif finding tools and our epiCOLOC, which also validates the efficacy of our tool.

DISCUSSION

In this study, we have integrated a comprehensive and tissue/cell type-specific epigenomics profiles database. With strict pre-processing, quality control and tissue mapping, we established a user-friendly web portal, epiCOLOC, which to perform fast and context-dependent colocalization analysis; and provide a series of visualization functions to interpret results; and significantly distinguish between existing web-based tools (**Supplementary Table 5**). In the applied examples, we demonstrated the accuracy and practicality of epiCOLOC in identifying causal tissues/cell types from GWAS disease-associated variants and inferring co-occurrence of transcription regulators.

There are some limitations in this work which deserve optimization in our future works. First, the statistical assumption of GIGGLE is simple and could be sub-optimal in several cases. We strongly recommend users to prioritize results by combo score and set stringent thresholds. As observed from the combo scores distribution when $P< = 0.05$ using query intervals that randomly generated in genome (**Supplementary Figure 2**), we propose to use an empirical combo score cutoff, 5 for enrichment and -2 for depletion, as advisable criteria to further filter enrichment or depletion results. Although GIGGLE can greatly speed up colocalization analysis, as compared with GenomeRunner (Dozmorov et al., 2016) and LOLAweb (Nagraj et al., 2018), it limits the usage of user-specific background of genomic regions and the analysis of multiple genomic intervals. Second, although epiCOLOC is applicable to perform colocalization analysis using genetic variants, but it cannot account for LD and allele frequency. Third, there are uneven epigenomic profiles for different tissues/cell types. It may potentially affect the robustness of colocalization when applying epiCOLOC to the tissues/cell types having fewer data available, and it also cannot determine the missing enrichment for tissues/cell types lacking sufficient data. In addition, single-cell technologies, such as single-cell ATAC-seq and single-cell ChIP-seq (Grosselin et al., 2019), have been developed to analyze genome-wide epigenomic features. Such approaches pave the way to study the role of epigenetic heterogeneity in many biological conditions and will be largely incorporated into epiCOLOC in the next stage. Recently, a novel algorithm named Augmented Interval List (AIList) (Feng et al., 2019), which introduces a new data structure and provides a significantly improved fundamental operation for highly scalable genomic data analysis. This method together with upcoming large-scale genomic features will be added in the epiCOLOC future updates.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found from ENCODE, Roadmap Epigenomics, etc and also related sources has been listed here: http://mulinlab.org/epicoloc/Introduction/.

AUTHOR CONTRIBUTIONS

ML designed and guided the study, YZ, YS and DH developed the tool, YZ and ML wrote the manuscript.

FUNDING

This work was supported by grants from the National Natural Science Foundation of China 31871327, 31701143 (ML), Natural Science Foundation of Tianjin 18JCZDJC34700, 19JCJQJC63600 (ML). We also appreciate all tool and resource providers.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.00053/full#supplementary-material

REFERENCES

Albrecht, F., List, M., Bock, C., and Lengauer, T. (2017). DeepBlueR: large-scale epigenomic analysis in R. *Bioinformatics* 33, 2063–2064. doi: 10.1093/bioinformatics/btx099

Bairoch, A. (2018). The cellosaurus, a cell-line knowledge resource. *J. Biomol. Tech.* 29, 25–38. doi: 10.7171/jbt.18-2902-002

Bujold, D., Morais, D., Gauthier, C., Cote, C., Caron, M., Kwan, T., et al. (2016). The international human epigenome consortium data portal. *Cell Syst.* 3, 496–499. e492. doi: 10.1016/j.cels.2016.10.019

Cazaly, E., Saad, J., Wang, W., Heckman, C., Ollikainen, M., and Tang, J. (2019). Making sense of the epigenome using data integration approaches. *Front. Pharmacol.* 10, 126. doi: 10.3389/fphar.2019.00126

Cheneby, J., Gheorghe, M., Artufel, M., Mathelier, A., and Ballester, B. (2018). ReMap 2018: an updated atlas of regulatory regions from an integrative analysis of DNA-binding ChIP-seq experiments. *Nucleic Acids Res.* 46, D267–D275. doi: 10.1093/nar/gkx1092

Consortium, G. T., Laboratory, D. A. Coordinating Center -Analysis Working, G. Statistical Methods Groups-Analysis Working, G., Enhancing, G. G., and Fund, N. I. H. C., et al. (2017). Genetic effects on gene expression across human tissues. *Nature* 550, 204–213. doi: 10.1038/nature24277

Consortium, E. P. (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature* 489, 57–74. doi: 10.1038/nature11247

Corces, M. R., Granja, J. M., Shams, S., Louie, B. H., Seoane, J. A., Zhou, W., et al. (2018). The chromatin accessibility landscape of primary human cancers. *Science* 362, eaav1898. doi: 10.1126/science.aav1898

Dor, Y., and Cedar, H. (2018). Principles of DNA methylation and their implications for biology and medicine. *Lancet* 392, 777–786. doi: 10.1016/S0140-6736(18)31268-6

Dozmorov, M. G., Cara, L. R., Giles, C. B., and Wren, J. D. (2016). GenomeRunner web server: regulatory similarity and differences define the functional impact of SNP sets. *Bioinformatics* 32, 2256–2263. doi: 10.1093/bioinformatics/btw169

Dozmorov, M. G. (2017). Epigenomic annotation-based interpretation of genomic data: from enrichment analysis to machine learning. *Bioinformatics* 33, 3323–3330. doi: 10.1093/bioinformatics/btx414

Egtex, G. P. (2017). Enhancing GTEx by bridging the gaps between genotype, gene expression, and disease. *Nat. Genet.* 49, 1664–1670. doi: 10.1038/ng.3969

Farh, K. K., Marson, A., Zhu, J., Kleinewietfeld, M., Housley, W. J., Beik, S., et al. (2015). Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature* 518, 337–343. doi: 10.1038/nature13835

Feinberg, A. P. (2018). The key role of epigenetics in human disease prevention and mitigation. *N. Engl. J. Med.* 378, 1323–1334. doi: 10.1056/NEJMra1402513

Feng, J., Ratan, A., and Sheffield, N. C. (2019). Augmented interval list: a novel data structure for efficient genomic interval search. *Bioinformatics.* 35, 4907–4911. doi: 10.1093/bioinformatics/btz407

Frietze, S., Wang, R., Yao, L. J., Tak, Y. G., Ye, Z. Q., Gaddis, M., et al. (2012). Cell type-specific binding patterns reveal that TCF7L2 can be tethered to the genome by association with GATA3. *Genome Biol.* 13, R52. doi: 10.1186/gb-2012-13-9-r52

Fullard, J. F., Hauberg, M. E., Bendl, J., Egervari, G., Cirnaru, M. D., Reach, S. M., et al. (2018). An atlas of chromatin accessibility in the adult human brain. *Genome Res.* 28, 1243–1252. doi: 10.1101/gr.232488.117

Grosselin, K., Durand, A., Marsolier, J., Poitou, A., Marangoni, E., Nemati, F., et al. (2019). High-throughput single-cell ChIP-seq identifies heterogeneity of chromatin states in breast cancer. *Nat. Genet.* 51, 1060–1066. doi: 10.1038/s41588-019-0424-9

Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y. C., Laslo, P., et al. (2010). Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol. Cell* 38, 576–589. doi: 10.1016/j.molcel.2010.05.004

Iotchkova, V., Ritchie, G. R. S., Geihs, M., Morganella, S., Min, J. L., Walter, K., et al. (2019). GARFIELD classifies disease-relevant genomic features through integration of functional annotations with association signals. *Nat. Genet.* 51, 343–34+. doi: 10.1038/s41588-018-0322-6

Kanduri, C., Bock, C., Gundersen, S., Hovig, E., and Sandve, G. K. (2019). Colocalization analyses of genomic elements: approaches, recommendations and challenges. *Bioinformatics* 35, 1615–1624. doi: 10.1093/bioinformatics/bty835

Lawrence, M., Daujat, S., and Schneider, R. (2016). Lateral thinking: how histone modifications regulate gene expression. *Trends Genet.* 32, 42–56. doi: 10.1016/j.tig.2015.10.007

Layer, R. M., Pedersen, B. S., Disera, T., Marth, G. T., Gertz, J., and Quinlan, A. R. (2018). GIGGLE: a search engine for large-scale integrated genome analysis. *Nat. Methods* 15, 123–126. doi: 10.1038/nmeth.4556

Libbrecht, M. W., Rodriguez, O. L., Weng, Z., Bilmes, J. A., Hoffman, M. M., and Noble, W. S. (2019). A unified encyclopedia of human functional DNA elements through fully automated annotation of 164 human cell types. *Genome Biol.* 20, 180. doi: 10.1186/s13059-019-1784-2

Liu, J. Z., Van Sommeren, S., Huang, H., Ng, S. C., Alberts, R., Takahashi, A., et al. (2015). Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. *Nat. Genet.* 47, 979–986. doi: 10.1038/ng.3359

Mei, S., Qin, Q., Wu, Q., Sun, H., Zheng, R., Zang, C., et al. (2017). Cistrome data browser: a data portal for ChIP-Seq and chromatin accessibility data in human and mouse. *Nucleic Acids Res.* 45, D658–D662. doi: 10.1093/nar/gkw983

Nagraj, V. P., Magee, N. E., and Sheffield, N. C. (2018). LOLAweb: a containerized web server for interactive genomic locus overlap enrichment analysis. *Nucleic Acids Res.* 46, W194–W199. doi: 10.1093/nar/gky464

Oki, S., Ohta, T., Shioi, G., Hatanaka, H., Ogasawara, O., Okuda, Y., et al. (2018). ChIP-Atlas: a data-mining suite powered by full integration of public ChIP-seq data. *EMBO Rep.* 19, e46255. doi: 10.15252/embr.201846255

Papatheodorou, I., Fonseca, N. A., Keays, M., Tang, Y. A., Barrera, E., Bazant, W., et al. (2018). Expression Atlas: gene and protein expression across multiple studies and organisms. *Nucleic Acids Res.* 46, D246–D251. doi: 10.1093/nar/gkx1158

Roadmap Epigenomics, C., Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., et al. (2015). Integrative analysis of 111 reference human epigenomes. *Nature* 518, 317–330. doi: 10.1038/nature14248

Sheffield, N. C., and Bock, C. (2016). LOLA: enrichment analysis for genomic region sets and regulatory elements in R and Bioconductor. *Bioinformatics* 32, 587–589. doi: 10.1093/bioinformatics/btv612

Simovski, B., Vodak, D., Gundersen, S., Domanska, D., Azab, A., Holden, L., et al. (2017). GSuite HyperBrowser: integrative analysis of dataset collections across the genome and epigenome. *Gigascience* 6, 1–12. doi: 10.1093/gigascience/gix032

Simovski, B., Kanduri, C., Gundersen, S., Titov, D., Domanska, D., Bock, C., et al. (2018). Coloc-stats: a unified web interface to perform colocalization analysis of genomic features. *Nucleic Acids Res.* 46, W186–W193. doi: 10.1093/nar/gky474

Stunnenberg, H. G.International Human Epigenome, C., , Hirst, M. (2016). The international human epigenome consortium: a blueprint for scientific collaboration and discovery. *Cell* 167, 1145–1149. doi: 10.1016/j.cell.2016.11.007

Trynka, G., Westra, H. J., Slowikowski, K., Hu, X. L., Xu, H., Stranger, B. E., et al. (2015). Disentangling the effects of colocalizing genomic annotations to functionally prioritize non-coding variants within complex-trait loci. *Am. J. Hum. Genet.* 97, 139–152. doi: 10.1016/j.ajhg.2015.05.016

Van Der Harst, P., and Verweij, N. (2018). Identification of 64 novel genetic loci provides an expanded view on the genetic architecture of coronary artery disease. *Circ. Res.* 122, 433–443. doi: 10.1161/CIRCRESAHA.117.312086

Van Der Meer, D., Barthorpe, S., Yang, W., Lightfoot, H., Hall, C., Gilbert, J., et al. (2019). Cell model passports-a hub for clinical, genetic and functional datasets of preclinical cancer models. *Nucleic Acids Res.* 47, D923–D929. doi: 10.1093/nar/gky872

Wang, J., Dai, X., Berry, L. D., Cogan, J. D., Liu, Q., and Shyr, Y. (2019). HACER: an atlas of human active enhancers to interpret regulatory variants. *Nucleic Acids Res.* 47, D106–D112. doi: 10.1093/nar/gky864

Yan, J., Enge, M., Whitington, T., Dave, K., Liu, J., Sur, I., et al. (2013). Transcription factor binding in human cells occurs in dense clusters formed around cohesin anchor sites. *Cell* 154, 801–813. doi: 10.1016/j.cell.2013.07.034

Zahn, C. T. (1971). Graph-theoretical methods for detecting and describing gestalt clusters. *IEEE Trans. Comput.* 20, 68–86. doi: 10.1109/T-C.1971.223083

Zambelli, F., Pesole, G., and Pavesi, G. (2013). PscanChIP: finding over-represented transcription factor-binding site motifs and their correlations in sequences from ChIP-Seq experiments. *Nucleic Acids Res.* 41, W535–W543. doi: 10.1093/nar/gkt448

Zheng, R., Wan, C., Mei, S., Qin, Q., Wu, Q., Sun, H., et al. (2019). Cistrome data browser: expanded datasets and new tools for gene regulatory analysis. *Nucleic Acids Res.* 47, D729–D735. doi: 10.1093/nar/gky1094

Zhou, Y. D., Park, S. Y., Su, J., Bailey, K., Ottosson-Laakso, E., Shcherbina, L., et al. (2014). TCF7L2 is a master regulator of insulin production and processing. *Hum. Mol. Genet.* 23, 6419–6431. doi: 10.1093/hmg/ddu359

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.

frontiers in Genetics

TECHNOLOGY AND CODE
published: 21 February 2020
doi: 10.3389/fgene.2020.00082

SECNVs: A Simulator of Copy Number Variants and Whole-Exome Sequences From Reference Genomes

Yue Xing[1,2,3*], *Alan R. Dabney*[2], *Xiao Li*[4], *Guosong Wang*[5], *Clare A. Gill*[5*] *and Claudio Casola*[6*]

[1] Interdisciplinary Program in Genetics, Texas A&M University, College Station, TX, United States, [2] Department of Statistics, Texas A&M University, College Station, TX, United States, [3] Department of Veterinary Integrative Biosciences, Texas A&M University, College Station, TX, United States, [4] Department of Molecular and Cellular Medicine, Texas A&M University, College Station, TX, United States, [5] Department of Animal Science, Texas A&M University, College Station, TX, United States, [6] Department of Ecosystem Science and Management, Texas A&M University, College Station, TX, United States

OPEN ACCESS

Edited by:
Jie Chen,
Augusta University,
United States

Reviewed by:
Joseph Glessner,
Children's Hospital of Philadelphia,
United States
Milind B. Ratnaparkhe,
ICAR Indian Institute of Soybean
Research, India

***Correspondence:**
Yue Xing
inumia@tamu.edu
Clare A. Gill
clare-gill@tamu.edu
Claudio Casola
ccasola@tamu.edu

Specialty section:
This article was submitted to
Bioinformatics and
Computational Biology,
a section of the journal
Frontiers in Genetics

Received: 05 November 2019
Accepted: 24 January 2020
Published: 21 February 2020

Citation:
Xing Y, Dabney AR, Li X, Wang G,
Gill CA and Casola C (2020) SECNVs:
A Simulator of Copy Number Variants
and Whole-Exome Sequences From
Reference Genomes.
Front. Genet. 11:82.
doi: 10.3389/fgene.2020.00082

Copy number variants are duplications and deletions of the genome that play an important role in phenotypic changes and human disease. Many software applications have been developed to detect copy number variants using either whole-genome sequencing or whole-exome sequencing data. However, there is poor agreement in the results from these applications. Simulated datasets containing copy number variants allow comprehensive comparisons of the operating characteristics of existing and novel copy number variant detection methods. Several software applications have been developed to simulate copy number variants and other structural variants in whole-genome sequencing data. However, none of the applications reliably simulate copy number variants in whole-exome sequencing data. We have developed and tested Simulator of Exome Copy Number Variants (SECNVs), a fast, robust and customizable software application for simulating copy number variants and whole-exome sequences from a reference genome. SECNVs is easy to install, implements a wide range of commands to customize simulations, can output multiple samples at once, and incorporates a pipeline to output rearranged genomes, short reads and BAM files in a single command. Variants generated by SECNVs are detected with high sensitivity and precision by tools commonly used to detect copy number variants. SECNVs is publicly available at https://github.com/YJulyXing/SECNVs.

Keywords: copy number variation, simulation, software, whole-exome sequencing, read depth

INTRODUCTION

Copy number variants (CNVs) represent DNA duplications and deletions ranging from a few dozen base pairs to several million bases that have been associated with phenotypic changes and human disease (Feuk et al., 2006). There is no precise definition for the minimum length of CNVs in research, although a minimum length of 1 kb is commonly used for clinical applications. Initially discovered by array-based methods (Pinkel et al., 1998), CNVs have been increasingly detected using next-generation sequencing (NGS) data (Shen et al., 2019). A substantial proportion of CNVs encompass protein-coding genes (Zmienko et al., 2014). Many software applications have been developed to detect CNVs using either whole-genome sequencing (WGS) (Bartenhagen and Dugas, 2013; Pattnaik et al., 2014; Qin et al., 2015; Faust, 2017; Xia et al., 2017) or whole-exome sequencing (WES) (Sathirapongsasuti et al., 2011; Fromer et al., 2012; Klambauer et al., 2012; Koboldt et al., 2012a; Koboldt et al., 2012b; Krumm et al., 2012; Plagnol et al., 2012; Magi et al., 2013) data.

WES is based on the capture and sequencing of transcribed regions (exons) of protein coding sequences, which combined represent approximately 1% of the human genome. Thus, WES offers a significant benefit in terms of the sequencing costs compared to WGS. Additionally, WES data are an increasingly important source to identify genetic variants in non-model organisms (Lu et al., 2016; Kaur and Gaikwad, 2017). In species with very large genomes and limited opportunities for WGS experiments, WES data are expected to represent a critical source of information to detect CNVs (Hirsch et al., 2014).

Detection strategies for CNVs from next-generation sequencing data consist of four different approaches based on read depth, physical distance between read pairs (or paired-end mapping), detection of split reads, and comparison of *de novo* and reference genome assemblies (Alkan et al., 2011; Pirooznia et al., 2015). Because each of these approaches have limitations, programs that combine multiple strategies to detect CNVs based on WGS datasets have also been developed [see (Pirooznia et al., 2015)]. In WES data, the approaches based on the distance between read pairs and detection of split reads have limited efficacy because the boundaries of the CNV region must fall completely within a target region for a CNV to be detected (Fromer et al., 2012; Alkodsi et al., 2014). However, the target regions only span a sparse 1% of the whole genome, therefore most of the breakpoints of CNVs are not located in the captured target regions (Tan et al., 2014; Yao et al., 2017). In addition, read pair and split read methods both rely on paired-end reads across a CNV region or reads mapped across CNV breakpoints (Tan et al., 2014). In read pair-based methods, shorter insert size compared to diploid individuals indicate deletions, whereas longer insert size compared to diploid individuals indicate duplications; in split read based methods, the split of reads is used to identify CNV and CNV breakpoints (Pirooznia et al., 2015). Because of the average size of exons and introns, most target regions in WES data fall between 100 and 300 bps, which makes detection of CNVs from WES data using read pair and split read methods practically impossible. Therefore, most available WGS-based CNV detection methods relying on paired-end reads cannot be successfully applied to WES data (Tan et al., 2014). Conversely, read depth-based methods rely on the number of sequenced reads aligned to each target region to calculate the average read depth over each base (Fromer et al., 2012), and it is assumed that read depth signal is proportional to copy number. Thus, read depth represents the only effective strategy to detect CNVs from WES datasets and has been implemented in several programs (reviewed in Tan et al., 2014). Because these programs are built using different implementations and statistical models, they tend to produce datasets of CNVs with relatively little overlap (Magi et al., 2013; Kadalayil et al., 2014; Tan et al., 2014; Nam et al., 2016; Yao et al., 2017; Zare et al., 2017; Pounraja et al., 2019). WES data tend to have higher levels of noise and specific biases compared to WGS data (Zare et al., 2017), making detection of CNVs from WES data less accurate overall. In addition, there are limitations of the read depth method that make CNV detection in WES data less accurate (Tan et al., 2014). These limitations include poor resolution, systematic group effects, GC bias and difficulty in prediction of breakpoints in WES datasets (Tan et al., 2014). Therefore, benchmark analyses are necessary to evaluate the performance of CNV detection programs that utilize exome-sequencing datasets. Both simulated exome data and data from either arrays or WGS have enabled the assessment of CNV detection tools for WES datasets (Magi et al., 2013; Kadalayil et al., 2014; Tan et al., 2014; Nam et al., 2016; Yao et al., 2017; Zare et al., 2017). Simulations allow a more comprehensive assessment of the accuracy and power of these tools.

Most software applications developed to simulate CNVs fall short of generating the required outputs for WES datasets and are difficult to implement or cannot be applied to certain datasets (**Table 1**). Here, we introduce Simulator of Exome Copy Number Variants (SECNVs), a fast, robust and customizable software application for simulating CNV datasets using WES data. It relies upon a completely new approach to simulate test genomes and target regions to overcome some of the limitations of other WGS CNV simulation tools, and is the first ready-to-use WES CNV simulator. The simulator can be easily installed and used on Linux and MAC OS systems to facilitate comparison of the performance of different CNV detection methods and to test the most appropriate parameter settings for CNV identification.

METHODS

Characteristics Needed for a Simulator of Copy Number Variants

To generate WES reads, specific regions of a reference genome, called "target regions," are captured and sequenced (Goh and Choi, 2012). To reproduce a realistic distribution of structural variants, a CNV simulator for WES data should generate variants that overlap partly or entirely with one or more target regions (**Figure 1**). The WES CNV detection tools require a list of target regions (exons) (Sathirapongsasuti et al., 2011; Fromer et al., 2012; Klambauer et al., 2012; Koboldt et al., 2012a; Koboldt et al., 2012b; Krumm et al., 2012; Plagnol et al., 2012; Magi et al., 2013), which can be obtained from public databases, and so this list could be used as the input to simulate short reads for those regions (Koboldt et al., 2009; Sathirapongsasuti et al., 2011; Koboldt et al., 2012a; Plagnol et al., 2012; Tan et al., 2014). Short reads would be simulated from a control genome (same as the reference genome) and test genomes (with simulated CNVs,

TABLE 1 | Comparison of current simulators to SECNVs.

Simulator Type	Name	Language	Steps	Extra Files? *	Output Format	Short Reads Simulated?
WGS	RSVSim (Bartenhagen and Dugas, 2013)	R	Multiple	No	Test genome (fasta) and CNV information	No
	SCNVSim (Qin et al., 2015)	JAVA	2	Samtools index of reference genome; Chromosome length file; Repeat mask file	Test genome (fasta) and CNV information	No
	Pysim-sv (Xia et al., 2017)	Python	Multiple	No	Fastq and SAM	Some
	SVsim (Faust, 2017)	Python	1	Samtools index of reference genome	Test genome (fasta) and bedpe	No
	SInC (Pattnaik et al., 2014)	C	2	No	Test genome (fasta) and short reads	Some
WES	VarSimLab[1] (CNV-Sim[2])	Python	1	No	Short reads or SAM, CNV information	Some
	SECNVs	Python	1	No	Test genome (fasta), short reads and/or BAM files, and CNV information	Yes

**Extra files mean any files required in addition to the reference genome and target regions files.*
CNV, copy number variant; WES, whole-exome sequencing; SECNVs, Simulator of Exome Copy Number Variants.

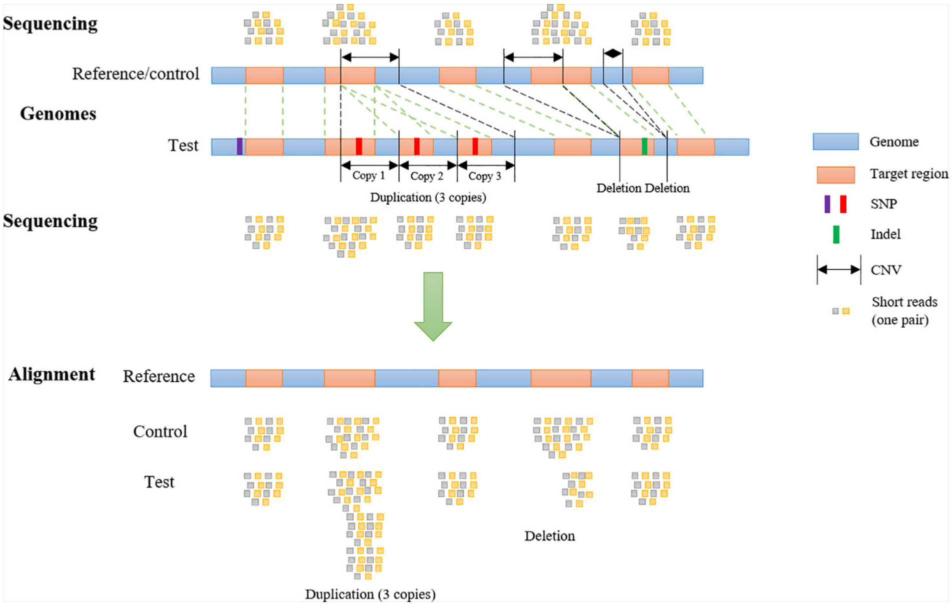


FIGURE 1 | Copy Number Variant detection by alignment of whole-exome sequencing reads to a reference genome. Whole-exome sequencing data are obtained by sequencing target regions in genomes of interest. If the test genome contains duplications and/or deletions overlapping target regions, these regions will be rearranged (duplicated, deleted or shifted in their genomic coordinates) compared to control and reference genome. Reads from the test and control genomes are aligned to the original target regions in the reference genome. Copy number variants are detected according to the alignment.

SNPs, and indels) and aligned back to the control genome. In the read alignment file for the test genome, simulated CNVs (duplications and deletions) would ideally appear as increased read coverage or reduced read coverage, respectively (**Figure 1**). Options to generate short reads rearranged according to customized length, type (duplication or deletion) and copy number of CNVs within the genomic coordinates of target regions (whole exons and, potentially, regions upstream and downstream of exons) should be available in such a program (**Figure 1**). To mimic real data, it would be desirable to introduce SNPs and indels during this step as well.

These characteristics were incorporated into the program SECNVs, which we designed to solve the issue of how to reliably simulate CNVs for WES datasets. The Python-based SECNVs pipeline copies the FASTA reference genome (control) and a list of start and end coordinates for exons to a working directory. From the command line, the user can choose to expand or connect regions to specify targets for sequencing and define the type, total number, copy number, and length of CNVs to simulate. SECNVs makes a list of randomly generated CNVs and using that information creates a file of rearranged target regions. Next, FASTA-formatted test genome sequence(s) and FASTQ-formatted short sequence reads for the target regions from the control and single or pooled test genome(s) are simulated, and BAM file(s) and index(es) for them are all generated in a single command. These files can then be used as the input to compare various CNV detection tools.

Simulation of Rearranged Genomes and Rearranged Target Regions

Before simulating WES, test genomes containing simulated CNVs that overlap with target regions are produced. First, the reference sequence is preprocessed based on how the user wants to handle gaps in the sequence. Next, a list of coordinates for CNVs are generated. Then SNPs and indels are simulated to create test genomes that mimic real data. Finally, CNVs are created in the FASTA-formatted test genome files.

Preprocessing

First, SECNVs reads in a FASTA reference genome file and a file of target regions, and checks which option the user chose to handle ambiguous nucleotides (N) or assembly gaps (collectively referred to herein as “gaps”). An assembly gap is a stretch of 50 (default) or more “Ns” in the sequence. The user can choose to replace ambiguous nucleotides or gaps with random nucleotides, to avoid simulating CNVs in regions containing gaps, or to ignore the presence of gaps (default). If the user chose replacement, SECNVs finds gaps in the reference genome and fills them with random nucleotides. Instead, if the user chose to avoid them, after finding the gaps, SECNVs stores the genomic coordinates that demarcate each gap for the following steps.

Creating a List of Coordinates for Copy Number Variant Regions

Before actually simulating a FASTA test genome and WES reads that contain CNVs, a list of sites where the CNVs will be placed is generated by the software. Placement of CNVs in the sequence depends on many user defined parameters: proportion of each type (duplication, deletion), total number, range of copy number, range and distribution of lengths (random, Gaussian, Beta, user-supplied), spacing (random, Gaussian), and minimum spacing between CNVs. Unless the user specifies a number of CNVs per chromosome, the application considers the proportion of CNVs that would be expected on each chromosome based on the length of the chromosome. The software randomly allocates whether each CNV is a duplication or deletion and the number of copies will be simulated within the user-defined range for copy number, and the length is also assigned randomly within the user-defined range and length distribution. Once the length of the CNVs have been determined, for each CNV, the software randomly chooses the start point of that CNV based on CNV spacing and calculates the coordinate for the end point. At this stage, the software stores the coordinates for the beginning and end of the CNV region. Next, if the user specified that CNVs should not overlap with any gaps, SECNVs checks the coordinates of the CNV region against the coordinates for gaps. If an overlap is found, the CNV is discarded; otherwise it is kept for the next step. SECNVs then compares the start and end coordinate of the CNV region to the list of target regions. If there is partial (default minimum overlap is 50 bp) or complete overlap with targets, the region is retained, otherwise it is discarded, and the loop starts again. Before writing the coordinates for the CNV regions to the file, SECNVs checks for overlap with previously generated regions or a user-defined buffer region. Only non-overlapping CNV regions are recorded in the final list from SECNVs. The loop is repeated until the total number of CNVs is reached, unless the chromosome is too small and/or the number of target regions is too limited to simulate enough CNVs. In this situation, SECNVs outputs a warning message and the number of CNVs simulated on that chromosome is printed instead of the user specified number, and the program continues for other chromosomes. Users can also choose to simulate CNVs outside of target regions. The process is very similar to simulating CNVs overlapping with target regions. The only difference is that if a CNV does not overlap with any target region, it will be kept; otherwise it will be discarded. SECNVs can also work with a list of predefined CNV regions. In this case SECNVs will read in the CNV list and use it as the final output of this step.

The Gaussian distribution of CNV spacing is generated by random selection from a symmetrically truncated Gaussian distribution mapped to the length of the chromosome, with distribution parameters (mean, SD) supplied by the user. Likewise, Gaussian distribution of CNV length is generated by random selection from a symmetrically truncated Gaussian distribution mapped to the range of user specified CNV length given distribution parameters (mean, SD) supplied by the user. The Beta distribution of CNV length, which is more realistic for CNV length distribution (Bartenhagen and Dugas, 2013), is generated by random selection from a Beta distribution mapped to the range of user-specified CNV lengths, again given distribution parameters (alpha, beta) from the user. Default values for alpha and beta are those used in Bartenhagen and Dugas, 2013. Otherwise, the user must estimate the parameters for the Beta distribution using a collection of sample CNV lengths from their own data, which can easily be done using R (Team, 2016). Detailed instructions for this are included in the manual.

The final product of this step is a list of coordinates for CNV regions that are used in the following step to produce test genomes that are rearranged from the reference genome and adjusted coordinates for target regions.

Simulation of Test Genomes and Adjusted Target Regions

The target regions are duplicated, deleted or shifted as a result of the simulated CNVs, as shown in **Figure 1**. Short reads are generated based on the rearranged genome and target regions, and aligned to the original reference/control genome in this “simulation of short reads” step.

Before introducing CNVs into the FASTA genome sequence, SNPs and indels are simulated as requested by the user. First, SNPs are randomly generated in the target regions plus a user-supplied buffer region upstream and downstream of the target regions (default is 0), based on the SNP rate specified by the user.

In this step, SECNVs randomly extracts n positions from these regions to simulate SNPs, where n equals the total length of the regions multiplied by the SNP rate. Then, nucleotides for these positions are randomly changed to another nucleotide in the test genome using the weights assigned by (Park, 2009) to represent the known mutation rate for SNP in human. Users can modify the mutation rates for other organisms. Detailed instructions are in the manual.

Next, indels are randomly generated in the target regions based on the indel rate (default is 0) specified by the user. In this step, m start points of indels are randomly generated in the target regions, where m equals the total length of the target regions multiplied by the indel rate. The length of each indel is then assigned by randomly choosing a number between 1 and the maximum indel length specified by the user. Type of indel (insertion or deletion) is randomly assigned to each indel as well. Next, SECNVs sorts the indels by their start points, and generates them one by one. If an indel is an insertion, SECNVs will make a random string of nucleotides of the previously assigned length and insert it at the assigned start point of that indel in the test genome sequence. Then, SECNVs recalculates the genomic coordinates of the target regions. If the start and/or end of the target regions are greater than the start point of the indel, their coordinates are increased by the length of that indel. The start point of the remaining indels is iteratively changed as well: coordinates of subsequent indels are increased by the length of that indel.

If an indel is a deletion, SECNVs will first check if the length of that indel is smaller than the target region it is in. If not, the length of that indel is reduced to ensure that at least one base pair of the target region remains. Then the sequence between the coordinates defining the indel is deleted from the test genome. Next, if the start and/or end of the target regions are greater than the start point of the indel, their coordinates are adjusted by subtracting the length of that indel. The start point of the rest of the indels are adjusted in the same way.

Finally, after the SNP and indels are created, the simulated CNVs are generated in the FASTA test genome files. In general, users would simulate CNVs that overlap with targets. The list of CNVs is sorted by coordinate and processed one by one so that the coordinates for subsequent CNVs are adjusted, similar to the process used for indels.

For each CNV, the genomic start and end coordinates and length are extracted. Then SECNVs loops iteratively through the genomic coordinates for all the target regions on a chromosome. If a target region is completely inside the CNV, it is categorized as "inside the CNV." If a target region partially overlaps with a CNV, it will be split into at least two parts: the parts outside of the CNV and the part overlapping with the CNV, and then categorized (upstream, inside, downstream), as shown in **Figure 2**. Sometimes users will choose to simulate CNV completely outside of target regions, even though those CNV will be undetectable in the WES. If a target region is completely before the CNV, it is categorized as "upstream of the CNV" and if a target region is completely after the CNV, it is categorized as "downstream of the CNV."

The next step is to adjust the coordinates to take into account the placement of the CNV relative to the target regions. As shown in **Supplementary Figure 1**, the coordinates of target regions categorized as "upstream of the CNV" remain unchanged. Coordinates of target regions categorized as "inside the CNV" must be adjusted. For duplications, the new start and end positions of these target regions will be: new position = length * (number of copies – 1) + old position, where the number of copies loops from 1 to the total copy number of that duplication, thus creating a tandem CNV duplication event in the test sequence. For deletions, when the CNV and target region overlap, the coordinates for that part of the target are deleted from the file. Coordinates of target regions categorized as "downstream of the CNV" will be altered as follows. For duplications, new position = length * (total copy number of duplication –1) + old position. For deletions, new position = old position – length of the CNV.

Finally, the CNV sequence will be copied to or deleted from the FASTA test genome accordingly. All the genomic coordinates of CNVs subsequent to this CNV in the list are adjusted in the same manner as the target regions categorized as

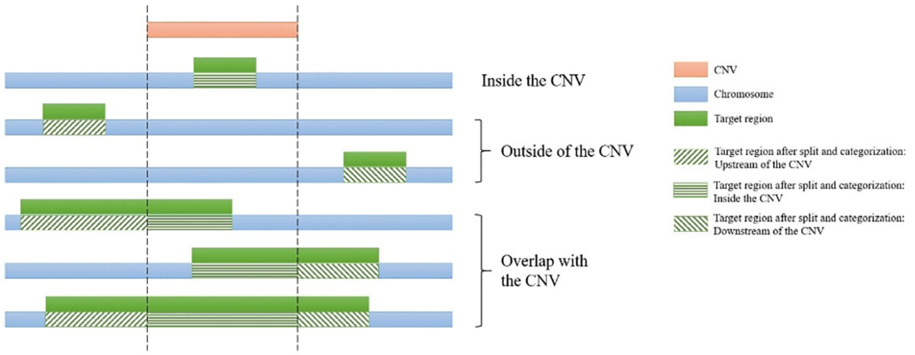

Inside the CNV
Outside of the CNV
Overlap with the CNV
CNV
Chromosome
Target region
Target region after split and categorization: Upstream of the CNV
Target region after split and categorization: Inside the CNV
Target region after split and categorization: Downstream of the CNV

FIGURE 2 | Categorization of target regions in the test genome. For each of the simulated copy number variants (CNVs), all target regions on a chromosome are assigned as "upstream of the CNV," "inside the CNV," and/or "downstream of the CNV." If the region partially overlaps it is also split. Afterwards genomic coordinates for the targets are recalculated and split regions are reconnected.

"downstream of the CNV." Finally, if a target region was previously split, it will be reconnected.

The software loops iteratively through all CNVs to create the rearranged test genome sequence and target regions for short read simulation. If the user chose to simulate multiple test genomes, the steps after preprocessing will be repeated to simulate each test genome.

The output files generated by this step include: 1. Test genome (s) (FASTA) with non-overlapping CNVs; 2. Target regions for test genome(s) (.bed); 3. Control genome (FASTA, optional); 4. Target regions for control genome (.bed, always generated in case the target regions are modified by user in sequencing steps); 5. List(s) of CNVs overlapping with target regions (.bed); 6. List(s) of CNVs outside of target regions (.bed, optional).

This is the core step of SECNVs. Users can choose to continue the pipeline within SECNVs to simulate short reads or use another short-read simulator and the files produced from this step as input.

Simulation of Short Reads

Users have the option to generate short read files with SECNVs by simulating single- or paired-end sequences from the test and control genomes. During this step, if the spacing between target regions is less than the spacing selected by user (default 0), the target regions are connected to form a single region (called a combined target region) to simulate the sequences. Users can also choose to expand the target regions by including additional nucleotides (default 0) upstream and downstream of the target regions (called an extended target region) for sequencing. The number of reads, type of reads (paired-end or single-end), fragment size, standard deviation of fragment size, read length, quality score offset, and error model can also be specified. A default error model: Illumina HighSeq 2500 for WES paired end sequencing is provided. This default error model was generated using a modified GemSIM script (McElroy et al., 2012) which fixed a bug to make the error profile generation function work. The dataset used for generating this error model was a human WES dataset from the Sequence Read Archive at the National Center for Biotechnology Information: run number ERR3385637. Users can also generate their own error model from real data using this modified GemSIM script, to keep the error profiles up to date as sequencing technology changes over time. Detailed instructions on how to use it to make new error profiles are included in the manual.

Instead, SECNVs reads in the headers of the input file as keys of a dictionary and reads the sequences line by line and combines them as values of that dictionary for the corresponding keys. Short read sequences are generated within the combined and extended target regions, which match just the target regions when default settings are used. Reads passing GC filtering are synthesized using a modification to the Wessim1 (Kim et al., 2013) algorithm (ideal target approach). Wessim1 only simulates reads at the start and the end of each target region (**Supplementary File 1**). Custom codes were written to modify Wessim1's scripts to correct this shortcoming of the program. Now fragments across the entire target regions based on fragment size and standard deviation of fragment size are produced and saved as FASTQ sequence, better mimicking real-world WES sequencing data. Output files from this step are the short reads for test genome(s) (FASTQ) and the short reads for the control genome (FASTQ, optional).

Creating BAM Files and Indexes From the Simulated Short Read Files

BAM files and indexes can be generated from the short read files for the test and control genomes through a standard pipeline that implements the widely-used tools BWA (Li and Durbin, 2009), samtools (Li et al., 2009; Li, 2011), Picard^3, and GATK (McKenna et al., 2010):

1. The Burrows–Wheeler Aligner of BWA is used to align the FASTQ reads to create a SAM file.
2. Samtools is used to convert the file format to a BAM file, sort the BAM file, and remove potential PCR duplicates.
3. Picard is used to add read groups to the samples.
4. GATK is used to locally realign reads, to minimize the number of mismatching bases across all the reads.

The output files in this step include: 1. Indexes for the control genome (.dict,.fai,.sa, etc., if BAM files are to be generated and no indexes exist in the output directory); 2. BAM file(s) and index (es) for the test genome(s) (.bam and.bai); 3. BAM file(s) and index(es) for control genome (.bam and.bai, optional).

Validation of Method

To confirm that the code for the algorithm implemented in SECNVs is correctly simulating the test genome and target regions, a small pseudo-genome was used as input to illustrate the process in **Supplementary Figure 1**.

Example Command Lines

1. Simulate 10 CNVs overlapping with target regions, and one CNV outside of target regions randomly on each chromosome using default lengths, copy numbers, minimum distance between each of the 2 CNVs and proportion of duplications. For each CNV overlapping with target regions, the overlapping length is no less than 90 bps. CNV break points follow a Gaussian(1, 2) distribution, and CNV lengths follow a Beta(2, 5) distribution. CNVs are not generated in gaps. A total of five test and control samples are built. Short reads (fastq) files are generated using default settings, paired-end sequencing.

   ```
   SECNVs/SECNVs.py -G < input_fasta> -T < target_region> -o < output_dir> \-e_chr 10 -o_chr 1 -ol 90 -ms gauss -as 1 -bs 2 -ml beta -al 2 -bl 5 -eN gap -n 5 -sc -pr -ssr
   ```

2. Simulate CNVs overlapping with target regions from a provided CNV list. Twenty CNVs are to be simulated outside of target regions randomly on the whole genome with default settings. CNVs are not to be generated on any stretches of "N"s. A pair of test and control genome are built.

   ```
   SECNVs/SECNVs.py -G < input_fasta> -T < target_region> -o < output_dir> \-e_cnv < list_of_CNV_overlapping_with_target_regions> -o_tol 20 -eN all -sc
   ```

3. Simulate 20 CNVs overlapping with target regions on the whole genome and have at least 100 bps between any two CNVs. CNVs are not generated outside of target regions. Gaps (50 or more consecutive "N"s) are replaced by random nucleotides. SNP rate is 0.001 and indel rate is 0.00001, and the maximum indel length is 100 bps. Paired-end sequencing reads with quality offset 35 are then produced. For a pair of test and control genomes BAM files are generated.

   ```
   SECNVs/SECNVs.py -G < input_fasta> -T < target_region> -o < output_dir> \-e_tol 20 -f 100 -rN gap -sc -pr -q 35 -ssr -sb \-s_r 0.001 -i_r 0.00001 -i_mlen 100 \-picard < absolute_path_to_picard> -GATK < absolute_path_to_GATK>
   ```

4. Simulate CNVs overlapping with target regions and outside of target regions from provided files of CNV lengths. Combined single regions are formed from two or more regions originally separated by less than 100 bps. CNVs are not generated on gaps (60 or more consecutive "N"s). A total of 10 test and control samples are built. The paired-end sequencing must include sequences 50 bp upstream and downstream of the target regions. The final output consists of short reads (fastq) files with 100,000 reads.

   ```
   SECNVs/SECNVs.py -G < input_fasta> -T < target_region> -o < output_dir> \-ml user -e_cl < length_file_1> -o_cl < length_file_2> \-clr 100 -eN gap -n_gap 60 -n 10 -sc -pr -tf 50 -nr 100000 -ssr
   ```

Simulation of Mouse and Human Whole-Exome Sequencing Datasets

To evaluate the performance of SECNVs, we used mouse (mm10) and human (hg38) chromosome 1 (downloaded from UCSC genome browser: https://genome.ucsc.edu/) as control genomes. Target regions were exons, which were also downloaded from the UCSC genome browser. We simulated 20 test genomes for each species that included 100 randomly distributed CNVs that overlapped at least 100 bp of target regions and ranged from 1,000 to 100,000 bp in length. Another 10 CNVs outside of target regions were also generated for each species. For each test genome, all sequences with "Ns" (gaps) were excluded, the SNP rate was set at 10^{-3}, and the indel rate was set at 10^{-5} (Mills et al., 2006). The minimum distance between any two CNVs was 1000 bp. For the synthesis of short reads, target regions less than 100 bp apart were connected and 50 bp upstream and downstream of the connected target regions were also sequenced. Paired-end sequencing was set to a base quality offset of 33. A total of one million reads were generated for each sample, with a fragment size of 200 bp and a read length of 100 bp. The rearranged fasta genome files with target regions, fastq short read files, BAM files and indexes for the 20 samples and control were simulated in one command for each species as follows:

```
Mouse: python SECNVs/SECNVs.py -G mouse/mouse.1.fa -T mouse/mouse.1.bed -e_chr 100 -o_chr 10 -o test_mous -rn mouse -ssr -sb -f 1000 -ol 100 -tf 50 -clr 100 -sc -eN all -pr -n 20 -q 33 -s_r 0.001 -i_r 0.00001 -nr 1000000 -picard < absolute_path_to_picard> -GATK < absolute_path_to_GATK>
```

```
Human: python SECNVs/SECNVs.py -G hg38/hg38.1.fa -T hg38/hg38.1.bed -e_chr 100 -o_chr 10 -o test_human_nn -rn human -ssr -sb -f 1000 -ol 100 -tf 50 -clr 100 -sc -eN all -pr -n 20 -q 33 -s_r 0.001 -i_r 0.00001 -nr 1000000 -picard < absolute_path_to_picard> -GATK < absolute_path_to_GATK>
```

Validation of Read Generation and Alignment

To confirm SECNVs is reliably simulating short reads and BAM files, read alignments in the target regions for human and mouse chromosome 1 against the respective reference genome were visualized with IGV (Thorvaldsdóttir et al., 2013).

Evaluation of Performance of Simulator of Exome Copy Number Variants

To demonstrate the utility of simulated datasets generated by SECNVs in performance testing of CNV detection software, we chose three commonly used WES CNV detection tools: ExomeDepth (Plagnol et al., 2012), CODEX2 (Jiang et al., 2018), and CANOES (Backenroth et al., 2014). Performance was evaluated for sensitivity, precision and false discovery rate. Sensitivity is the number of true CNVs that are correctly detected divided by the total number of true CNVs. Precision is the number of CNVs correctly detected by tools, divided by the total number of CNVs detected by tools. False discovery rate (FDR), which equals to 1—precision, is the number of CNVs incorrectly detected by tools, divided by the total number of CNVs detected by tools. During this evaluation, we found that CNV transition probability in ExomeDepth and CNV occurrence in CANOES influenced the test results the most, so we evaluated different values for these parameters as well. All other parameters were either left as default or set to fit the characteristics of CNVs we expected to detect. For example, in ExomeDepth, length of expected CNVs was set to 50000, which was about the average CNV length expected. A CNV was considered detected if at least 80% of the detected CNV overlap with a simulated CNV. We also used the best application with optimized parameter settings to test if any CNVs outside of target regions were detected.

RESULTS

In this study, we presented a fast, reliable and highly-customizable software application, SECNVs, which takes in a reference genome and target regions to simulate SNPs, indels and CNVs in one or multiple test genomes, as well as the control, and outputs fasta formatted genome files with target regions, short read files, BAM files and indexes in a single command.

Computational Speed

SECNVs is a fast software application for CNV simulation. The detailed approximate computation time to generate CNVs for each sample on human and mouse chromosome 1 is shown in **Table 2**.

Validation of Method

Simulation of SNPs, indels and CNVs in the tiny pseudo-genome established that the method of random replacement of gaps, simulation of SNPs, indels and CNVs in the test genome is

TABLE 2 | Computation time and memory usage of SECNVs.

	Mouse chromosome 1	Human chromosome 1
1. Read in genome and target region files, exclude all "N" sequences	<15 s	<25 s
2. Generate list of CNVs overlapping with target regions	<4 min	<8 min
3. Generate list of CNVs outside of target regions	<50 s	<3 min
4. Generate rearranged genome: make SNPs, indels and CNVs in the genome	<50 s	<80 s
5. Generate short reads	<7.5 min	<7.5 min
6. Create indexes for the control genome	<3.5 min	<4.5 min
7. Generate BAM file and index	<2 min	<2 min
Max memory	5,630 MB	6,516 MB
Average memory	1,885.78 MB	2,521.26 MB

CNV, copy number variant; SECNVs, Simulator of Exome Copy Number Variants.

accurate; and the rearrangement of target regions for the test genome is accurate as well (**Supplementary File 2**; **Supplementary Figure 1**).

The simulated short reads and BAM files generated using the modified Wessim1 code align across the whole target regions (**Figure 3**). Differences in read coverage at combined and extended target regions in test and control BAM files are characteristic of CNVs spanning these target regions (**Figures 3A, C**). For target regions without CNVs, there was no obvious difference in read coverage at combined and extended target regions in test and control BAM files (**Figures 3B, D**). No reads were aligned outside of target regions, regardless of whether there were CNVs or not, except for a few alignment errors. The reliability of the BAM files ensured that CNVs overlapping with target regions could be readily detected.

Sensitivity and Precision of Copy Number Variant Detection From Simulated WES Datasets

Average sensitivity, precision, FDR and the number of CNVs detected by ExomeDepth, CANOES and CODEX2 using simulated reads from human and mouse genomes are summarized in **Table 3** and **Figure 4**. Regardless of the species, as the transition probability from ExomeDepth increased (**Figures 4A, B**) and until the number of CNV detected matched the number of CNV simulated, sensitivity increased, and precision was high. Beyond 100 CNV, the number of detected CNV rapidly inflated and precision rapidly declined. A similar profile was observed for occurrence of CNV in CANOES (**Figures 4C, D**). The overall performance of ExomeDepth in terms of precision and sensitivity was better than CANOES or CODEX (**Figure 5**). False discovery rate was higher for the human data than the mouse data. CODEX was not able to detect all of the simulated CNV. Although the parameters of transition probability in ExomeDepth and occurrence of CNV in CANOES are similar in concept, the comparison in **Figure 5** shows that as each of these parameters was changed, performance of the two software tools was very different. We also confirmed that CNVs simulated outside of exomes were never detected.

DISCUSSION

CNVs represent an important source of genetic variation and have been associated with disease and other important phenotypic traits in humans, domesticated animals and crops (Zhang et al., 2009;

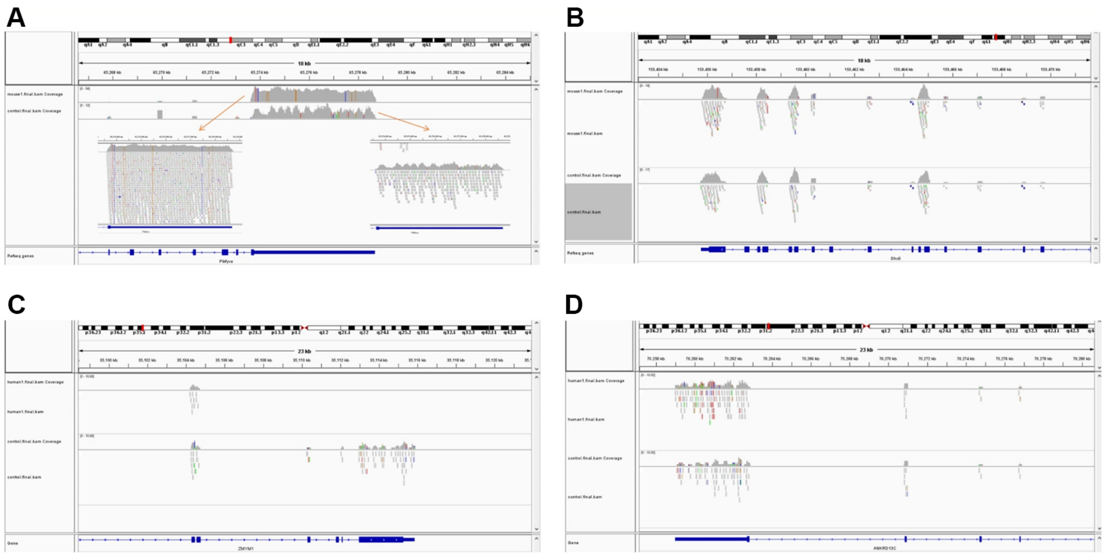


FIGURE 3 | Exemplar simulated output BAM files visualized in IGV. **(A)** A 10 copy duplication at mouse chr1:65272798-65339955, which partially overlap with exons of the Pikfyve gene. Because of the read depth in this region, the reads tracks are shown side-by-side; **(B)** A region of Dhx9 gene of the mouse genome, showing no copy number variants in this region; **(C)** A deletion at human chr1: 35106158-35150376, which partially overlap with exons of the Zmym1 gene; **(D)** A region of ANKRD13C gene of the human genome, showing no copy number variants in this region. In each image, the top track is a region of the test genome and the middle track shows the same region of the control genome. The bottom track is the exons and introns of genes.

TABLE 3 | Average sensitivity, precision, FDR, and number of CNV detected using simulated WES datasets.

Application	Parameter value	Mouse				Human			
		Sensitivity	Precision	FDR	Total # of CNVs detected	Sensitivity	Precision	FDR	Total # of CNVs detected
Exome-Depth	0.0005	0.34	0.99	0.01	34.2	0.20	0.98	0.02	20.1
	0.001	0.37	0.99	0.01	37.3	0.22	0.98	0.02	22.7
	0.010	0.48	0.99	0.01	49.3	0.32	0.98	0.02	32.8
	0.020	0.53	0.99	0.01	55.0	0.36	0.98	0.02	38.1
	0.030	0.55	0.98	0.02	58.4	0.39	0.98	0.02	41.5
	0.050	0.60	0.99	0.01	64.3	0.44	0.98	0.02	47.1
	0.080	0.65	0.98	0.02	70.4	0.48	0.98	0.02	52.7
	0.100	0.67	0.98	0.02	73.1	0.50	0.97	0.03	56.0
	0.150	0.70	0.96	0.04	78.0	0.55	0.93	0.07	66.1
	0.200	0.72	0.94	0.06	83.3	0.61	0.85	0.15	80.6
	0.250	0.74	0.90	0.10	89.5	0.64	0.69	0.31	106.1
	0.300	0.76	0.82	0.18	101.6	0.68	0.49	0.51	157.4
	0.350	0.77	0.65	0.35	127.9	0.71	0.31	0.69	262.1
	0.400	0.78	0.43	0.57	193.7	0.73	0.17	0.83	482.5
CANOES	0.0005	0.34	0.90	0.10	37.7	0.32	0.91	0.09	35.6
	0.001	0.36	0.90	0.10	40.1	0.35	0.90	0.10	38.7
	0.010	0.45	0.89	0.11	50.7	0.44	0.88	0.12	50.7
	0.050	0.55	0.79	0.21	72.3	0.54	0.79	0.21	73.7
	0.080	0.56	0.65	0.35	95.2	0.55	0.64	0.36	100.8
	0.100	0.55	0.53	0.47	120.4	0.54	0.48	0.52	136.6
	0.120	0.52	0.39	0.61	158.0	0.50	0.31	0.69	198.1
	0.125	0.51	0.35	0.65	169.5	0.49	0.27	0.73	218.0
	0.150	0.46	0.20	0.80	248.2	0.41	0.13	0.87	356.3
	0.200	0.35	0.09	0.91	404.4	0.27	0.05	0.95	574.7
	0.250	0.28	0.06	0.94	460.8	0.20	0.04	0.96	611.6
	0.300	0.25	0.06	0.94	473.9	0.19	0.03	0.97	631.9
	0.350	0.22	0.05	0.95	479.9	0.17	0.03	0.97	655.6
CODEX2	–	0.47	0.72	0.28	65.9	0.31	0.71	0.29	44.9

CNV, copy number variant; WES, whole-exome sequencing; FDR, false discovery. The parameter for ExomeDepth is transition probability and the parameter for CANOES is CNV occurrence.

Alkan et al., 2011). WES projects represent an increasingly common source of genomic data that can be harvested to detect CNVs. Often applied in the detection of mutations associated with cancer, Mendelian and complex diseases in humans, WES data have also been generated for multiple non-model organisms to identify genetic variants, including CNVs (Prunier et al., 2017; Low et al., 2019). However, previous studies have shown that detection of CNVs from WES data is inconsistent across the tools designed to detect these variants (Guo et al., 2013; Nam et al., 2016; Yao et al., 2017). These evaluations have largely relied on datasets of well-characterized CNVs obtained using array-based experiments or WGS data from human samples, a benchmarking approach that presents several limitations. First, known variants tend to occupy the higher end of the spectrum of lengths for CNVs. Second, known CNVs are often derived from cancer tissues and are expected to show different features than germline CNVs. Third, the characteristics of CNVs might differ significantly between humans and other organisms. Therefore, flexible CNV simulators would allow more rigorous testing of the efficacy of these tools.

Previously developed simulators fall short of producing realistic CNVs and present a variety of operational issues that make them challenging or impossible to use. Most of the applications for simulating CNVs and other structural variants from WGS data (**Table 1**) (Bartenhagen and Dugas, 2013; Pattnaik et al., 2014; Qin et al., 2015; Faust, 2017; Xia et al., 2017), require commands be entered in several steps, and require further processing to use their outputs. Among them, only RSVSim (Bartenhagen and Dugas, 2013) and SVsim (Faust, 2017) allow users to specifically generate CNVs in user-selected regions of the genome. However, they cannot calculate rearranged coordinates of the target regions in the test genome after simulating CNVs, which makes it impossible to use their outputs to generate accurate sets of short reads if the user only knows the original target regions but not the probe sequences in sequencing step. Additionally, RSVSim runs into an infinite loop when there are too many gaps in the genome.

To the best of our knowledge, VarSimLab[1], previously released as CNV-Sim[2], is the only other program specifically designed to simulate CNVs from WES data (Zare et al., 2017), but the website for the software indicates it is currently not usable. We found that CNV-Sim had a problem generating short reads. The main issue with these programs appears to be their reliance on the ideal target approach implemented in the application Wessim1 (Kim et al., 2013) to generate short reads, but this approach does not provide coverage across target regions (see **Supplementary File 1**). Furthermore, the majority of CNVs simulated by CNV-Sim overlap with each other when target regions are nearby on the reference genome. A third issue is that this program creates temporary "genomes" by deleting the segments between target regions. Additional copies for the test

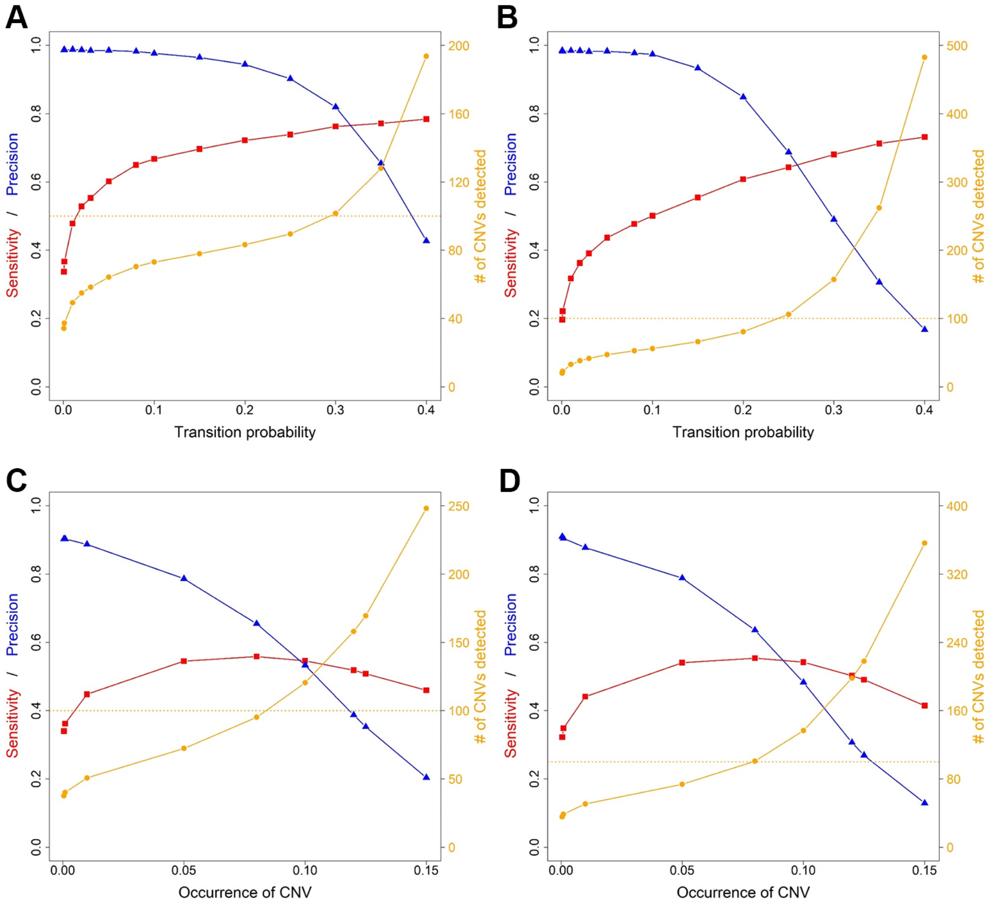


FIGURE 4 | Sensitivity, precision and number of copy number variants (CNVs) detected for ExomeDepth and CANOES. The sensitivity, precision and number of CNVs detected in **(A)** simulated mouse data for ExomeDepth, **(B)** simulated human data for ExomeDepth, **(C)** simulated mouse data for CANOES and **(D)** simulated human data for CANOES are displayed. Red lines show sensitivity, blue lines show precision and orange lines show the number of CNVs detected. Solid triangles, squares and circles represent the actual data points.

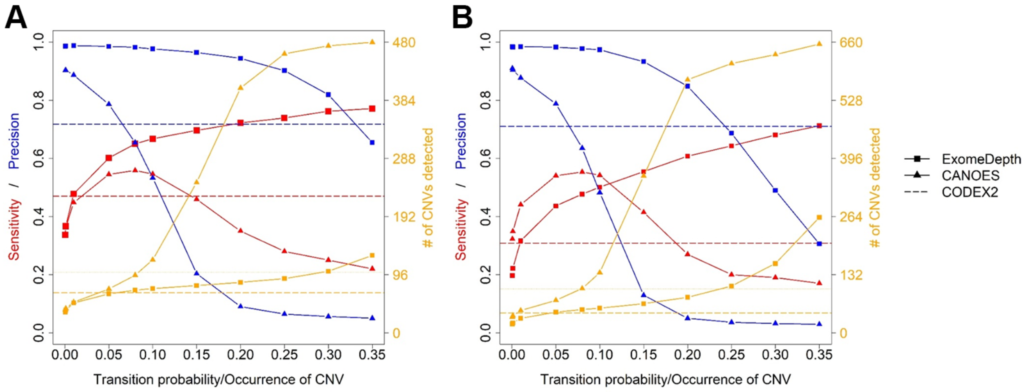


FIGURE 5 | Comparison of sensitivity, precision and number of copy number variants (CNVs) detected by the three software applications. **(A)** simulated mouse data and **(B)** simulated human data. Red lines show sensitivity, blue lines show precision and orange lines show the number of CNVs detected. Because CODEX2 does not have the parameter “transition probability” or “CNV occurrence,” a single value for sensitivity, precision and number of CNVs is displayed.

genome are generated in the case of duplications, and additional copies for the control genome are created in the case of deletions. Short read files are then generated using these test and control genomes (called “tumor” and “normal” in CNV-Sim). This makes it impossible to generate pooled samples with a common control, whereas many CNV detection applications for WES data require pooled samples as input. In addition, it cannot simulate the realistic scenario of CNVs with different degrees of overlap with target regions. Finally, CNV-Sim only accepts one chromosome at a time.

Here, we described SECNVs, a novel software application that fills this gap by simulating realistic CNVs from WES data. First, it uses a completely new method to accurately and reliably simulate test genome(s) and target regions with SNPs, indels and CNVs. Second, it incorporates a modified version of the Wessim1 algorithm to simulate short reads, which effectively mimics real-world WES sequencing, including GC filtration. Third, to keep the sequencing error profile up to date, SECNVs provides a recent error profile for short read simulation and includes detailed instructions on how to make user-specified error profiles from real data. Finally, the options for CNV simulation are highly customizable. In this paper, SECNVs was applied to human and mouse data and the results showed that CNVs simulated by the software application were successfully detected by various WES CNV detection software applications, demonstrating that output from SECNVs can be used to test these applications and their parameters.

Sensitivity and FDR were similar to previous reports using real data (Seiser and Innocenti, 2015; Zare et al., 2017). It is known that the sensitivity is low and FDR is high for CNV detection in WES datasets (Tan et al., 2014; Yao et al., 2017). This is because read depth approaches are the only reliable method for WES CNV detection, but they have many limitations (Tan et al., 2014). In addition, the high SNP and indel rate introduced, as well as GC filtration and sequencing errors affect the alignment and reduce sensitivity and increase FDR. However, compared to the real human CNV detection from Illumina genotyping microarrays by other Hidden Markov Model-Based CNV detection methods mentioned in Seiser and Innocenti (2015), and real human CNV detection from WES data using various software applications (Zare et al., 2017), the sensitivity, precision, and FDR all suggest that CNVs generated by SECNVs are reliable and can be easily detected.

After reaching the true CNV number, the number of detected CNVs tends to inflate. When this happens, sensitivity either increases very slowly or begins to drop, and precision decreases rapidly. Therefore, the user can choose the transition probability, CNV occurrence or similar parameter to detect CNVs approximating the number of real CNVs, and determine if the sensitivity and precision are acceptable. Alternatively, the user can sacrifice some sensitivity and detect fewer CNVs to get higher precision. Of course, users can test other parameters as well to find out the most suitable software application and parameter settings for their data.

Using datasets simulated with SECNVs, we were also able to characterize the performance of CNV detection software applications under a wide range of parameters. We showed that simulations are critical to assess the effect of key parameters on the sensitivity and accuracy of such applications. Thus, SECNVs can simulate highly customized WES datasets to mimic real-world data and enable users to identify the most appropriate software application and parameter settings for their real data.

SECNVs is suitable for the analysis of large, complex genomes. In our tests on mouse and human chromosome 1 (~195.5 and ~250 Mbp), the program simulated a new chromosome for each species with 110 CNVs/chromosome, removed gaps, and generated one million reads/sample, BAM files and indexes in less than 30 min. The run time for scaled-up simulations using complete mammalian genomes (~3–3.5 Gbp) should therefore require less than a day. Longer computational times are likely to occur for incomplete genomes with more extensive gap regions, although the gaps-exclusion component of SECNVs is computationally fast (**Table 2**). Because there is no limitation in the number of input chromosome/scaffolds/contigs, SECNVs can be applied to highly fragmented assemblies of nonmodel organisms. For instance, CNVs have been simulated using SECNVs on the ~21 Gbp assembly of the loblolly pine, which consists of 1,755,249 contigs and scaffolds. In general, for very large assemblies such as those of wheat, conifer and some amphibian genomes, generating CNVs is likely to be computationally demanding. Using only scaffolds and contigs containing the target regions is advised in order to accelerate the simulation of CNVs.

We identified two main limitations in the current version of SECNVs. First, because SNPs and indels are simulated in the test genome and then CNVs are simulated, there is no variability among the duplicated sequences. Second, all CNVs are tandem duplications or deletions in SECNVs. However, these limitations do not affect CNV detection from WES data, because most WES CNV detection methods are read depth based, which cannot distinguish between tandem duplication and insertion elsewhere in the genome (Tan et al., 2014). The nature of WES datasets makes methods other than read depth ineffective for WES CNV detection (Tan et al., 2014).

One caution is that most diploid reference genomes report a consensus sequence for each pair of chromosomes. Therefore, by default the SECNVs simulator adds or deletes two copies at a time. If the user wanted to extend the simulator to an odd number of duplication or deletion events, the bam files for the reference and test genomes could be merged.

Currently, SECNVs only simulates indels in target regions to increase speed. For SNPs, buffer regions upstream and downstream of target regions are allowed for simulation, because SNPs downstream of CNVs may affect the detection of that CNV (Bartenhagen and Dugas, 2013). Buffer regions upstream and downstream of target regions are also allowed for short read simulation. In future version of SECNVs, SNPs and indels will be simulated for the whole genome.

The time used for each step implemented in SECNVs strongly depends on the parameter settings supplied by the user and increases in an approximately linear manner. For instance, run time is positively correlated with the number of CNVs, SNPs and indels that are generated. In addition, the run time tends to increase significantly when the length of all CNVs combined exceeds the length of the genome, but we expect that this model will rarely be implemented when simulating realistic CNVs even in small genomes. Given its flexibility, precision and variety of unique features, SECNVs represents a reliable application to study CNVs using WES data for various species and under a variety of conditions.

DATA AVAILABILITY STATEMENT

The genome sequence and target region (exon) files analyzed for this study can be found in the UCSC genome database at https://genome.ucsc.edu/cgi-bin/hgGateway (Consortium, 2001). The datasets simulated and analyzed during the current study are available from the corresponding authors on request.

AUTHOR CONTRIBUTIONS

YX designed, created and tested the software application, evaluated its performance, wrote the draft of the manuscript and contributed to its major revisions. AD, CC, and CG provided suggestions on the design and update of the software application, contributed to major revisions of the manuscript, and read and approved the final manuscript. XL participated in designing and testing the software application and contributed to revisions of the manuscript. GW provided suggestions on the update of the software application and solved technical problems.

FUNDING

This work has been supported by the National Institute of Food and Agriculture, U.S. Department of Agriculture, under award number TEX0-1-9599, Texas A&M AgriLife Research, and the Texas A&M Forest Service to C.C. and by Texas A&M AgriLife Research through an Enhancing Research Capacity for Beef Production Systems grant awarded to C.A.G.

ACKNOWLEDGMENTS

The high-performance research computing system Ada at Texas A&M University was used to test the software application. This manuscript has been released as a Pre-Print at (Xing et al., 2019).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.00082/full#supplementary-material

REFERENCES

Alkan, C., Coe, B. P., and Eichler, E. E. (2011). Genome structural variation discovery and genotyping. *Nat. Rev. Genet.* 12 (5), 363–376. doi: 10.1038/nrg2958

Alkodsi, A., Louhimo, R., and Hautaniemi, S. (2014). Comparative analysis of methods for identifying somatic copy number alterations from deep sequencing data. *Briefings Bioinf.* 16 (2), 242–254. doi: 10.1093/bib/bbu004

Backenroth, D., Homsy, J., Murillo, L. R., Glessner, J., Lin, E., Brueckner, M., et al. (2014). CANOES: detecting rare copy number variants from whole exome sequencing data. *Nucleic Acids Res.* 42 (12), e97–e97. doi: 10.1093/nar/gku345

Bartenhagen, C., and Dugas, M. (2013). RSVSim: an R/Bioconductor package for the simulation of structural variations. *Bioinformatics* 29 (13), 1679–1681. doi: 10.1093/bioinformatics/btt198

Consortium, I. H. G. S. (2001). Initial sequencing and analysis of the human genome. *Nature* 409 (6822), 860. doi: 10.1038/35057062

Faust, G. (2017). SVsim: a tool that generates synthetic Structural Variant calls as benchmarks to test/evaluate SV calling pipelines. Available at: https://github.com/GregoryFaust/SVsim

Feuk, L., Carson, A. R., and Scherer, S. W. (2006). Structural variation in the human genome. *Nat. Rev. Genet.* 7 (2), 85. doi: 10.1038/nrg1767

Fromer, M., Moran, J. L., Chambert, K., Banks, E., Bergen, S. E., Ruderfer, D. M., et al. (2012). Discovery and statistical genotyping of copy-number variation from whole-exome sequencing depth. *Am. J. Hum. Genet.* 91 (4), 597–607. doi: 10.1016/j.ajhg.2012.08.005

Goh, G., and Choi, M. (2012). Application of whole exome sequencing to identify disease-causing variants in inherited human diseases. *Genomics Inf.* 10 (4), 214–219. doi: 10.5808/GI.2012.10.4.214

Guo, Y., Sheng, Q., Samuels, D. C., Lehmann, B., Bauer, J. A., Pietenpol, J., et al. (2013). Comparative study of exome copy number variation estimation tools using array comparative genomic hybridization as control. *BioMed. Res. Int.* 2013. doi: 10.1155/2013/915636

Hirsch, C. D., Evans, J., Buell, C. R., and Hirsch, C. N. (2014). Reduced representation approaches to interrogate genome diversity in large repetitive plant genomes. *Briefings In Funct. Genomics* 13 (4), 257–267. doi: 10.1093/bfgp/elt051

Jiang, Y., Wang, R., Urrutia, E., Anastopoulos, I. N., Nathanson, K. L., and Zhang, N. R. (2018). CODEX2: full-spectrum copy number variation detection by high-throughput DNA sequencing. *Genome Biol.* 19 (1), 202. doi: 10.1186/s13059-018-1578-y

Kadalayil, L., Rafiq, S., Rose-Zerilli, M. J. J., Pengelly, R. J., Parker, H., Oscier, D., et al. (2014). Exome sequence read depth methods for identifying copy number changes. *Briefings Bioinf.* 16 (3), 380–392. doi: 10.1093/bib/bbu027

Kaur, P., and Gaikwad, K. (2017). From genomes to GENE-omes: exome sequencing concept and applications in crop improvement. *Front. In Plant Sci.* 8, 2164–2164. doi: 10.3389/fpls.2017.02164

Kim, S., Jeong, K., and Bafna, V. (2013). Wessim: a whole-exome sequencing simulator based on in silico exome capture. *Bioinformatics* 29 (8), 1076–1077. doi: 10.1093/bioinformatics/btt074

Klambauer, G., Schwarzbauer, K., Mayr, A., Clevert, D. A., Mitterecker, A., Bodenhofer, U., et al. (2012). cn.MOPS: mixture of Poissons for discovering copy number variations in next-generation sequencing data with a low false discovery rate. *Nucleic Acids Res.* 40 (9), e69. doi: 10.1093/nar/gks003

Koboldt, D. C., Chen, K., Wylie, T., Larson, D. E., McLellan, M. D., Mardis, E. R., et al. (2009). VarScan: variant detection in massively parallel sequencing of individual and pooled samples. *Bioinf. (Oxford England)* 25 (17), 2283–2285. doi: 10.1093/bioinformatics/btp373

Koboldt, D. C., Zhang, Q., Larson, D. E., Shen, D., McLellan, M. D., Lin, L., et al. (2012a). VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res.* 22 (3), 568–576. doi: 10.1101/gr.129684.111

Koboldt, D. C., Zhang, Q., Larson, D. E., Shen, D., McLellan, M. D., Lin, L., et al. (2012b). VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res.* 22 (3), 568–576. doi: 10.1101/gr.129684.111

Krumm, N., Sudmant, P. H., Ko, A., O'Roak, B. J., Malig, M., Coe, B. P., et al. (2012). Copy number variation detection and genotyping from exome sequence data. *Genome Res.* 22 (8), 1525–1532. doi: 10.1101/gr.138115.112

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics* 25 (16), 2078–2079. doi: 10.1093/bioinformatics/btp352

Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* 27 (21), 2987–2993. doi: 10.1093/bioinformatics/btr509

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler Transform. *Bioinformatics* 25 (14), 1754–1760. doi: 10.1093/bioinformatics/btp324

Low, T. Y., Mohtar, M. A., Ang, M. Y., and Jamal, R. (2019). Connecting proteomics to next-generation sequencing: proteogenomics and its current applications in biology. *Proteomics* 19 (10), e1800235. doi: 10.1002/pmic.201800235

Lu, M., Krutovsky, K. V., Nelson, C. D., Koralewski, T. E., Byram, T. D., and Loopstra, C. A. (2016). Exome genotyping, linkage disequilibrium and population structure in loblolly pine (Pinus taeda L.). *BMC Genomics* 17 (1), 730. doi: 10.1186/s12864-016-3081-8

Magi, A., Tattini, L., Cifola, I., D'Aurizio, R., Benelli, M., Mangano, E., et al. (2013). EXCAVATOR: detecting copy number variants from whole-exome sequencing data. *Genome Biol.* 14 (10), R120. doi: 10.1186/gb-2013-14-10-r120

McElroy, K. E., Luciani, F., and Thomas, T. (2012). GemSIM: general, error-model based simulator of next-generation sequencing data. *BMC Genomics* 13 (1), 74. doi: 10.1186/1471-2164-13-74

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., et al. (2010). The genome analysis toolkit: a mapreduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 20 (9), 1297–1303. doi: 10.1101/gr.107524.110

Mills, R. E., Luttig, C. T., Larkins, C. E., Beauchamp, A., Tsui, C., Pittard, W. S., et al. (2006). An initial map of insertion and deletion (INDEL) variation in the human genome. *Genome Res.* 16 (9), 1182–1190. doi: 10.1101/gr.4565806

Nam, J.-Y., Kim, N. K. D., Kim, S. C., Joung, J.-G., Xi, R., Lee, S., et al. (2016). Evaluation of somatic copy number estimation tools for whole-exome sequencing data. *Briefings In Bioinf.* 17 (2), 185–192. doi: 10.1093/bib/bbv055

Park, L. (2009). Relative mutation rates of each nucleotide for another estimated from allele frequency spectra at human gene loci. *Genet. Res. (Camb.)* 91 (4), 293–303. doi: 10.1017/S0016672309990164

Pattnaik, S., Gupta, S., Rao, A. A., and Panda, B. (2014). SInC: an accurate and fast error-model based simulator for SNPs, Indels and CNVs coupled with a read generator for short-read sequence data. *BMC Bioinf.* 15 (1), 40. doi: 10.1186/1471-2105-15-40

Pinkel, D., Segraves, R., Sudar, D., Clark, S., Poole, I., Kowbel, D., et al. (1998). High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nat. Genet.* 20 (2), 207–211. doi: 10.1038/2524

Pirooznia, M., Goes, F. S., and Zandi, P. P. (2015). Whole-genome CNV analysis: advances in computational approaches. *Front. Genet.* 6, 138. doi: 10.3389/fgene.2015.00138

Plagnol, V., Curtis, J., Epstein, M., Mok, K. Y., Stebbings, E., Grigoriadou, S., et al. (2012). A robust model for read count data in exome sequencing experiments and implications for copy number variant calling. *Bioinf. (Oxford England)* 28 (21), 2747–2754. doi: 10.1093/bioinformatics/bts526

Pounraja, V. K., Jayakar, G., Jensen, M., Kelkar, N., and Girirajan, S. (2019). A machine-learning approach for accurate detection of copy number variants from exome sequencing. *Genome Res.* 29 (7), 1134–1143. doi: 10.1101/gr.245928.118

Prunier, J., Caron, S., and MacKay, J. (2017). CNVs into the wild: screening the genomes of conifer trees (Picea spp.) reveals fewer gene copy number variations in hybrids and links to adaptation. *BMC Genomics* 18 (1), 97. doi: 10.1186/s12864-016-3458-8

Qin, M., Liu, B., Conroy, J. M., Morrison, C. D., Hu, Q., Cheng, Y., et al. (2015). SCNVSim: somatic copy number variation and structure variation simulator. *BMC Bioinf.* 16 (1), 66. doi: 10.1186/s12859-015-0502-7

Sathirapongsasuti, J. F., Lee, H., Horst, B. A. J., Brunner, G., Cochran, A. J., Binder, S., et al. (2011). Exome sequencing-based copy-number variation and loss of heterozygosity detection: ExomeCNV. *Bioinformatics* 27 (19), 2648–2654. doi: 10.1093/bioinformatics/btr462

Seiser, E. L., and Innocenti, F. (2015). Hidden Markov model-based CNV detection algorithms for illumina genotyping microarrays. *Cancer Inf.* 13 (Suppl 7), 77–83. doi: 10.4137/CIN.S16345

Shen, W., Szankasi, P., Durtschi, J., Kelley, T. W., and Xu, X. (2019). Genome-wide copy number variation detection using NGS: data analysis and interpretation. *Methods Mol. Biol.* 1908, 113–124. doi: 10.1007/978-1-4939-9004-7_8

Tan, R., Wang, Y., Kleinstein, S. E., Liu, Y., Zhu, X., Guo, H., et al. (2014). An evaluation of copy number variation detection tools from whole-exome sequencing data. *Hum. Mutat.* 35 (7), 899–907. doi: 10.1002/humu.22537

Team, R. C. (2016). R: a language and environment for statistical computing (Vienna, Austria: R Foundation for Statistical Computing).

Thorvaldsdóttir, H., Robinson, J. T., and Mesirov, J. P. (2013). Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Briefings In Bioinf.* 14 (2), 178–192. doi: 10.1093/bib/bbs017

Xia, Y., Liu, Y., Deng, M., and Xi, R. (2017). Pysim-sv: a package for simulating structural variation data with GC-biases. *BMC Bioinf.* 18 (3), 53. doi: 10.1186/s12859-017-1464-8

Xing, Y., Dabney, A. R., Li, X., Wang, G., Gill, C. A., and Casola, C. (2019). SECNVs: a simulator of copy number variants and whole-exome sequences from reference genomes. *bioRxiv* 824128. doi: 10.1101/824128

Yao, R., Zhang, C., Yu, T., Li, N., Hu, X., Wang, X., et al. (2017). Evaluation of three read-depth based CNV detection tools using whole-exome sequencing data. *Mol. Cytogenet.* 10, 30–30. doi: 10.1186/s13039-017-0333-5

Zare, F., Dow, M., Monteleone, N., Hosny, A., and Nabavi, S. (2017). An evaluation of copy number variation detection tools for cancer using whole exome sequencing data. *BMC Bioinf.* 18 (1), 286. doi: 10.1186/s12859-017-1705-x

Zhang, F., Gu, W., Hurles, M. E., and Lupski, J. R. (2009). Copy number variation in human health, disease, and evolution. *Annu. Rev. Genomics Hum. Genet.* 10, 451–481. doi: 10.1146/annurev.genom.9.081307.164217

Zmienko, A., Samelak, A., Kozlowski, P., and Figlerowicz, M. (2014). Copy number polymorphism in plant genomes. *Theor. Appl. Genet.* 127 (1), 1–18. doi: 10.1007/s00122-013-2177-7

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.

frontiers in Genetics

METHODS
published: 21 February 2020
doi: 10.3389/fgene.2020.00072

Integrating Peak Colocalization and Motif Enrichment Analysis for the Discovery of Genome-Wide Regulatory Modules and Transcription Factor Recruitment Rules

*Mirko Ronzio, Federico Zambelli, Diletta Dolfini, Roberto Mantovani and Giulio Pavesi**

Dipartimento di Bioscienze, Università di Milano, Milan, Italy

OPEN ACCESS

Edited by:
Geir Kjetil Sandve,
University of Oslo, Norway

Reviewed by:
Endre Barta,
National Agricultural Research and Innovation Centre, Hungary
Chenlong Li,
Agriculture and Agri-Food Canada (AAFC), Canada

****Correspondence:***
Giulio Pavesi
giulio.pavesi@unimi.it

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Genetics

Received: *07 October 2019*
Accepted: *22 January 2020*
Published: *21 February 2020*

Citation:
Ronzio M, Zambelli F, Dolfini D, Mantovani R and Pavesi G (2020) Integrating Peak Colocalization and Motif Enrichment Analysis for the Discovery of Genome-Wide Regulatory Modules and Transcription Factor Recruitment Rules. Front. Genet. 11:72. doi: 10.3389/fgene.2020.00072

Chromatin immunoprecipitation followed by next-generation sequencing (ChIP-Seq) has opened new avenues of research in the genome-wide characterization of regulatory DNA-protein interactions at the genetic and epigenetic level. As a consequence, it has become the *de facto* standard for studies on the regulation of transcription, and literally thousands of data sets for transcription factors and cofactors in different conditions and species are now available to the scientific community. However, while pipelines and best practices have been established for the analysis of a single experiment, there is still no consensus on the best way to perform an integrated analysis of multiple datasets in the same condition, in order to identify the most relevant and widespread regulatory modules composed by different transcription factors and cofactors. We present here a computational pipeline for this task, that integrates peak summit colocalization, a novel statistical framework for the evaluation of its significance, and motif enrichment analysis. We show examples of its application to ENCODE data, that led to the identification of relevant regulatory modules composed of different factors, as well as the organization on DNA of the binding motifs responsible for their recruitment.

Keywords: ChIP-seq, colocalization analysis, transcription factor (TF), transcriptional regulation, transcription factor binding sites (TFBS)

INTRODUCTION

Next-generation sequencing based assays have opened novel avenues of investigation in every aspect of research in genomics. In particular, they have become the standard in the genome-wide characterization of all the elements concurring to the regulation of gene transcription, like nucleosome positioning (Buenrostro et al., 2013; Pajoro et al., 2018), DNA accessibility (Giresi et al., 2007), DNA methylation, histone modifications and chromatin states (Roadmap Epigenomics Consortium et al., 2015), transcription factor binding (Johnson et al., 2007), and long-distance enhancer-promoter interactions (Fullwood and Ruan, 2009; Lieberman-Aiden et al., 2009).

As a consequence, literally thousands of experiments have been published on one or more of the above aspects in different species and conditions, and large scale projects like ENCODE (Gerstein et al., 2012) and Roadmap Epigenomics (Roadmap Epigenomics Consortium et al., 2015) have been launched. It is now standard practice also for small or midsize labs to produce several datasets with different experiments, and to merge the results obtained into a single overall picture of the regulatory landscape of the condition studied.

Genome-wide NGS-based assays usually produce as a main result a list of genomic regions, where base pairs included in these regions satisfy the condition being tested, e.g., they are nucleosome-free, bound by a transcription factor, occupied by a nucleosome carrying a given histone modification, and so on. Further information can be associated with each region, as for example, its enrichment in the sequenced sample, that can be expressed according to different measures. Key factors for the reliability of the results produced are both wet lab protocols and the downstream bioinformatic analysis of the data, and indeed, as of today, a consensus has been reached for which are to be considered the best practices for both (Landt et al., 2012). However, we are far from having de facto standards for the integrative analysis of the results of different experiments. In principle, a single base pair on the genome appearing in the output of two or more experiments can be considered to satisfy simultaneously the different conditions that have been tested. How this information can be interpreted depends on the experiments producing the data to be analyzed. For example, ChIP-Seq assays for different histone modifications in the same condition can be processed with approaches like segmentation (Ernst and Kellis, 2017), in order to identify their most relevant combinations on the genome, and to produce a genome-wide map of chromatin states each characterized by a different combination of modifications, mapping the location of active or repressed promoters, enhancers, transcribed regions, and so on. However, the integrative analysis of different ChIP-Seq experiments for DNA binding proteins like transcription factors is usually performed with different criteria and principles, often designed *ad hoc* for the protein or condition studied.

In this work we present a simple pipeline for the integrative analysis of any number of ChIP-Seq experiments for transcription factors (TFs) or cofactors. Each ChIP-Seq experiment returns a genome-wide map of the binding locations on DNA of the protein studied. While this phenomenon is usually represented as a single protein interacting with DNA, in reality different factors and cofactors form large protein complexes, binding DNA at distal and proximal regions, that recruit RNA polymerase and initiate transcription. Thus, it is of the utmost importance for obtaining a complete understanding of the mechanisms behind the regulation of transcription not to treat each factor as a separate entity, but to identify combinations of different factors binding DNA as a complex at the same loci of the genome, and evaluate if these coassociations can constitute widespread regulatory modules.

Given the results of ChIP-Seq experiments for any number of different TFs or cofactors, our pipeline has been designed to answer the following questions: (1) Which are the combinations of factors and cofactors that are found with higher frequency on the genome? (2) Are the combinations found actually significant, that is, not resulting from random associations between different proteins but indeed found with high frequency on the genome, thus signaling a higher level of organization in transcriptional regulation? (3) Which are the recruiting rules on DNA, that is, which are the factors actually bound to DNA, and are there specific combinations (e.g., distance or orientation requirements) for their DNA-binding sites?

These questions have become more and more relevant over the years, once large datasets, like the assays performed by the ENCODE project, have been released. Indeed, curated databases containing thousands of ChIP-Seq datasets for TFs in different species and conditions are now publicly available, like ChIP-Atlas (Oki et al., 2018) or ReMap (Chèneby et al., 2018). An important feature of these repositories is that, like in the ENCODE project, all datasets included have been uniformly reprocessed, in order to make their comparison as less biased as possible by different choices in data analysis.

TF colocalization on the genome has been already defined and tacked with different approaches, ever since the introduction of the first NGS-based assays, as for example in (Chen et al., 2008). Several works have addressed the problem by starting from the position of TF binding peaks on the genome [see among many others (Chen et al., 2008; Gerstein et al., 2012)]. Colocalization, and its significance, is then assessed starting from the number of overlapping peaks, and evaluated with explorative or correlation measures like Pearson correlation (Chen et al., 2008), z-scores (Gerstein et al., 2012), the Jaccard index (Salvatore et al., 2019), or with machine learning based techniques like self-organizing maps (Xie et al., 2013).

An orthogonal approach is to analyze regions resulting from a single ChIP-Seq experiment for enrichment of sequence motifs known to represent sites be bound by other TFs, as for example in (Wang et al., 2012; Levitsky et al., 2019). Candidate TFs thus identified can be likely members of the same regulatory module. The pipeline we introduce here is indeed a combination of these two approaches, peak colocalization and motif enrichment analysis, with the additional introduction of a statistical framework designed *ad hoc* to assess both.

METHODS

Transcription Factor Colocalization

Defining Overlapping Peaks and Cobinding Regions

The overlap among two or more genome-wide datasets can formalized in different ways, both in the definition of common features and in the evaluation of the significance of the overlap found, as reviewed for example in (Kanduri et al., 2019; Salvatore et al., 2019).

ChIP-Seq experiments for DNA binding proteins like TFs produce as output enriched regions usually called “peaks.” This is due to the experimental protocol preparing the DNA to be sequenced (**Figure 1**). The fragments, produced by random DNA sonication, are usually of about 200 bps. Once mapping on the genome of the sequenced reads has been performed, the actual fragment size can be reestimated according to the distance between clusters of reads mapping on opposite strands, (Zhang et al., 2008; Bailey and MacHanick, 2012; Mathelier et al., 2016) in order to improve the resolution obtained by the experiment. The result is a coverage “signal” map, giving an estimate of how many times each base pair of the genome was present in the sequenced DNA sample. Since the actual point of interaction between the protein studied and DNA is present in each of the fragments, enriched regions show a typical “peak” shape in the signal map (**Figure 1**).

Algorithms for “peak calling” thus return regions where the observed enrichment and respective signal is not considered to be due to random experimental noise (Thomas et al., 2017). A typical region is reported to be a few hundreds of base pairs long, while the sites actually bound by TFs are much smaller, usually no more than 10–12 base pairs. However, the local maxima of the peaks correspond, or at least are not too distant from, the actual binding site of the protein studied (Zhang et al., 2008). Indeed, ChIP-Seq peaks usually show a good “centrality,” that is, the likely binding site of DNA returned by sequence analysis is usually found to be within a few dozen base pairs from the summit (Zhang et al., 2008; Bailey and MacHanick, 2012; Mathelier et al., 2016). For proteins like cofactors, not directly in contact with DNA, the argument still holds, with the only difference that summits and binding sites are related to the protein(s) of the complex tethering the cofactor on DNA.

The above considerations should be kept in mind when performing colocalization analyses for TF binding. Simply defining a DNA locus as “cobound” by two different transcription factors if two peak regions overlap might correspond to cases in which the actual binding sites of the factors are hundreds of base pairs apart. Thus, our approach to defining two (or more) TFs or cofactors as “colocalizing” on the genome is based instead on peak summits coordinates. In other words, we do not require two peak regions just to overlap, but we consider the location of the respective summits. We then define two peaks as “overlapping,” and the TFs to bind DNA in close proximity, if the respective summits are within d_s base pairs from

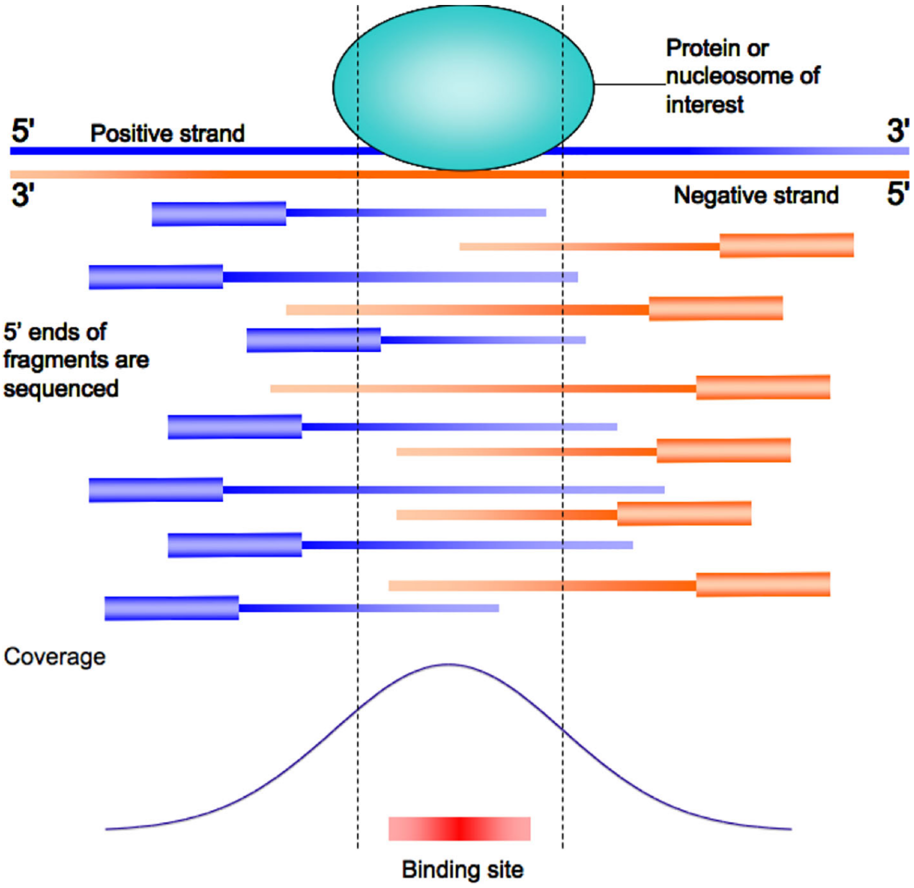

Protein or nucleosome of interest
5'
Positive strand
3'
3'
Negative strand
5'
5' ends of fragments are sequenced
Coverage
Binding site

FIGURE 1 | The typical peak shaped enrichment plot for a ChIP-Seq experiment resulting from read mapping on the genome. The actual point of contact of the protein studied on DNA is usually close to the point of maximum local enrichment (peak summit).

one another, with an approach similar to (Chen et al., 2008), where the "center" of peak regions was employed to assess colocalization. With respect to (Chen et al., 2008), however, we introduce also a statistical assessment of the significance of overlaps, as detailed in the next section. As a default threshold for this step we set as maximum distance d_s = 150 bps, a distance commonly employed in studies of this kind (Wang et al., 2012), which also makes the calculation of the statistical significance of overlap straightforward as shown in the next section.

We define pairs of peaks satisfying this criterion as *cobinding* peaks. This, in turn, usually corresponds to having the binding sites on DNA of the two factors located within a number of base pairs (d_{bs}) not too much different from the d_s distance. Or, alternatively, summit proximity could be due to only one of the two factors in contact with DNA, with the other one being anyway part of the same complex (**Figure 2**).

Assessing Statistical Significance of Peak Overlaps

Let TF1 and TF2 be two TFs on which the cobinding analysis is performed; let n and m the respective number of peaks, and k the number of cobinding peaks defined as at the previous point. We want to assess the probability of finding by chance k cobinding peaks (hence, regions bound by both TFs), given n and m.

In our approach, we also define a constant N, denoting the number of regions across the genome available for TF binding, whose size equals our "cobinding" region size of 150 bps. A straightforward way to estimate N would be to count the overall number of regions bound by all the TFs active in the condition studied. This approach, however, would have the effect of underestimating N if binding data were available only for a limited number of TFs, i.e. several regions would not be included in the count simply because the ChIP-Seq experiments for the TFs binding them had not been performed.

However, since a region bound by a TF usually corresponds to accessible DNA, estimate for N can be obtained from the number of nucleosome-free regions, and their respective size. Thus, we took advantage of maps of accessible DNA produced in several different cell lines in the framework of the ENCODE project, through digital genomic footprinting (Sabo et al., 2004; Vierstra and Stamatoyannopoulos, 2016). The advantage of these datasets (retrieved from the UCSC genome browser track "UW DNaseI DGF" on the GRCh37 assembly) is that the genome is split into regions of exactly 150 bps, that corresponds to the maximum d_s distance between summits we allow for peak overlap. Thus, cobinding peaks can be seen as two peak summits falling exactly within the same accessible region. The value of N is naturally cell- and condition-specific, ranging roughly from 200,000 to 250,000 for most of the ENCODE cell lines on which this assay has been performed. In case this number is available for the condition studied it can be thus employed in a straightforward way. If not, we advise to employ N = 250,000, a value we consider to be reliable enough for all different conditions in human, and also in other mammalian genomes like mouse. The only exception to this rule are embryonic stem cell lines, in which nucleosome occupancy has been shown to be significantly lower (Celona et al., 2011; Harwood et al., 2019), with an average number of genomic loci available for binding almost doubled. Thus, for these the suggested value is N = 500,000.

Indeed, the vast majority of the accessible regions results to be bound by TFs (80%–90% in the different ENCODE cell lines for which digital footprinting data are available). Once again, the sole exception are stem cell lines, where the percentage is lower (70% in ENCODE H7-ESC cells), also because less TF ChIP-Seq experiments are available for this condition. The number of regions actually bound (more than 300,000) is anyway larger, nearly twice as much as the other cell lines.

Similar estimates can be derived for other species and taxa, since nucleosome occupancy and DNA accessibility data are available for all the most widely studied species, as for example in Drosophila (Thomas et al., 2011), or Arabidopsis thaliana (Tannenbaum et al., 2018). In these two species the smaller genome size (hundreds of millions of base pairs) in turn results in a proportionally lower number of estimated accessible regions (tens of thousands).

An alternative, if data are available, could be to focus on regions annotated as active promoters or enhancers, as revealed by presence of specific histone marks or resulting from a genome segmentation approach like ChromHMM (Ernst and Kellis, 2017).

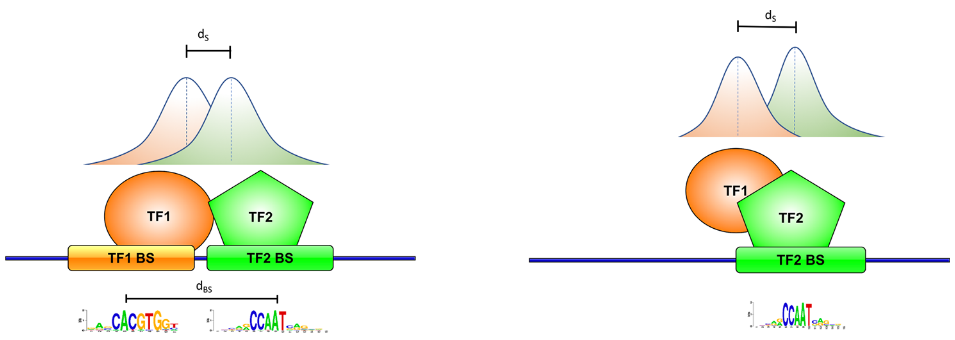


FIGURE 2 | Definition of cobinding peaks. Peak summits are usually close of the point of contact of the corresponding transcription factor (TF) with DNA. Two peak summits within a given number of base pairs d_s (150 in this work) should thus correspond to two TFs binding DNA in close proximity with one another (left, with the respective binding sites within d_{bs} base pairs), or to one of the two TF tethering the other on DNA (right).

An estimate for N can be thus derived by the number of regions of size d_s found annotated as active promoter or enhancer. Also, the overlap between two TFs can be assessed in either subset of regions, thus identifying promoter- or enhancer-specific modules.

Once an estimate for N has been produced, the probability of finding k cobinding peaks for two TFs by chance given n and m (the number of peaks for the two TFs, respectively) can be computed with different approaches, for example with the hypergeometric distribution:

$$p(k; n, m, N) = \frac{\binom{n}{k}\binom{N-n}{m-k}}{\binom{N}{m}}$$

or the Poisson distribution:

$$p(k; n, m, N) = \frac{e^{-\lambda}\lambda^{k}}{k!}$$

where $\lambda = \frac{mn}{N}$. In our experiments we employed the latter, since the p-values returned are more conservative. Since both distributions are two-tailed, low p-values can point to significant colocalization across the genome (k higher than the expected value), or vice versa if k is lower than the expected value than the two TFs considered tend to avoid one another on the genome.

This analysis is performed on every pair of experiments available. If several pairwise comparisons are performed, then the p-values should also be corrected for multiple testing. For example, in the results we present here we analyzed 329 ENCODE datasets for TFs in the K562 cell line, thus with 329 × 329 pairwise comparisons. We employed once again the most conservative choice, the Bonferroni correction, multiplying the p-values by $329^2 = 108{,}241$.

Building Modules With More Than Two Factors

The results of the pairwise comparisons described at the previous step can be further extended to modules composed by more than two TFs or cofactors.

An initial explorative analysis (see Results) is to define a colocalization score for each pair of experiments i and j, starting from the corresponding number of cobinding peaks k, and the respective p-value p_{ij}, as $-\log_{10}p_{ij}$ if the observed overlap is higher than the expected value, $\log_{10}p_{ij}$ (and hence a negative number) otherwise. The resulting values can be employed to represent the results as a heatmap, and clustering the heatmap can in turn highlight groups of TFs with significant pairwise overlaps, hence likely to be found together in the same regulatory module.

Another approach we introduce is to choose a "base" $\mathrm{TF_b}$, and determine whether other TFs tend to colocalize within its peaks. For every pair of TFs ($\mathrm{TF_i}$ and $\mathrm{TF_j}$) different from $\mathrm{TF_b}$, this step is formalized as follows:

- Let k_b the number of peaks for the base $\mathrm{TF_b}$;
- Let k_i and k_j the number of cobinding regions with $\mathrm{TF_b}$ of the two other TFs ($\mathrm{TF_i}$ and $\mathrm{TF_j}$);
- Let k_{ij} be the number of cobinding regions for both $\mathrm{TF_i}$ and $\mathrm{TF_j}$ with $\mathrm{TF_b}$

At this point, the significance of the cobinding of $\mathrm{TF_i}$ and $\mathrm{TF_j}$ in correspondence to $\mathrm{TF_b}$ binding sites can be assessed again with a statistical test as in normal pairwise comparison. That is, we can compute the probability of finding k_{ij} cobinding regions for $\mathrm{TF_i}$, $\mathrm{TF_j}$ and $\mathrm{TF_b}$, given k_i, k_j, and k_b. The p-value can be computed again with a Poisson distribution:

$$p(k_{ij}; k_i, k_j, k_b) = \frac{e^{-\lambda}\lambda^{k_{ij}}}{k_{ij}!}$$

where $\lambda = \frac{k_i k_j}{k_b}$.

The resulting p-values can in turn be converted again into scores, with the respective clustering highlighting groups of TFs colocalizing, but this time with $\mathrm{TF_b}$ as "tether" on DNA. Once the base $\mathrm{TF_b}$ has been chosen, this step can be performed by selecting only TFs that had a significant overlap with $\mathrm{TF_b}$ at the previous step in their pairwise comparison with it.

If necessary, this step can be iterated any number of times, e.g. assessing the significance of the overlap of a fourth TF given the cobinding regions of $\mathrm{TF_b}$, $\mathrm{TF_i}$, $\mathrm{TF_j}$, and so on.

Defining Binding and Recruitment Rules Through Motif Analysis

Once a list of genomic regions bound simultaneously by two (or more) TFs has been produced, the next step is to determine if the respective binding sites are actually present on DNA, and if so if they present any arrangement, e.g., are found at a precise distance, further hinting at co-operative binding and interactions between the respective proteins.

The binding specificity of a TF is usually defined with a *position specific frequency matrix*, or *profile*, obtained by the alignment of a collection of binding sites for the TF (Stormo, 2000; Zambelli et al., 2013a), defining its nucleotide preference on DNA. Several collections of profiles are freely available, derived from large-scale *in vitro* assays like SELEX or by the application of motif discovery tools to ChIP-Seq peak regions (Wingender, 2008; Mathelier et al., 2016; Khan et al., 2018; Wingender et al., 2018). For example, the latest version of the JASPAR database (Khan et al., 2018) includes for human and mouse profiles derived from the analysis of the ENCODE datasets.

This step can be formalized as a *motif enrichment analysis*, that is, the regions are analyzed in order to determine whether the motif representing the binding specificity of each of the TFs involved can be considered to be enriched in them, both in number and quality of instances found. Different tools have been introduced for this task, including a tool we developed called PscanChIP (Zambelli et al., 2013b). Since, as we previously discussed, the region more likely to correspond to the actual point of contact of the TF on DNA is located near peak summits, PscanChIP requires as input a list of one base pair peak summit coordinates, and scans the region of 150 base pairs centered on each one employing a collection of motifs like the JASPAR database or defined by the user.

Actual motif enrichment is evaluated by the tool in different ways:

- Global enrichment: enrichment is assessed with respect to a genomic background, that is, motifs are overrepresented in the selected regions with respect to the rest of the genome accessible to TF binding. Hence, a motif with significant global enrichment could correspond to the actual binding site of the TF (usually the most significant one), or binding sites of other TFs which show a clear genome-wide tendency to bind in association with it.
- Local enrichment: the enrichment of the motif in the peak summit regions is compared to the regions immediately upstream and downstream of the summit regions themselves;
- Positional bias: the localization of the most likely instance of the motif in each summit regions is identified, and the resulting distribution is compared to a theoretical uniform distribution.

Thus, the results of motif enrichment analysis can be interpreted as follows: if a motif corresponding to one of the TFs binding the regions selected is found to be significantly enriched according to the global p-value reported by PscanChIP, then the corresponding TF can be assumed to be in contact with DNA. Since the regions submitted as input are bound *in vivo* by one or more TFs, then the corresponding motifs should be the ones with the lowest global p-values among those employed in the analysis. Also, given the centrality of ChIP-Seq peaks, they should present a positional bias towards the middle of the regions. Otherwise, TFs for which no significant motif enrichment is found can be considered not to be directly binding DNA, although part of a complex in contact with DNA (**Figure 2**).

There are a few main differences between PscanChIP and other methods for the same task. The presence/absence of a motif instance in a region is evaluated with a score, ranging from 0 to 1, instead of a yes/no decision (binding motif present/absent) as for example in recent works (Dergilev et al., 2017; Czipa et al., 2020; Levitsky et al., 2019), which are also focused on the analysis of regions surrounding ChIP-Seq summits. Mean and variance of scores of best motif instances in each of the summit regions are in turn employed by PscanChIP to assess motif enrichment not only with respect to regions flanking peaks (local enrichment), as in similar tools (Zhang et al., 2011; Bailey and MacHanick, 2012), but also with respect to the rest of the genome, providing a more accurate evaluation of their significance.

PscanChIP also permits to perform a "motif centered" analysis. Once the first round of motif enrichment analysis has been completed in the neighborhood of peak summits, users can select one of the motifs resulting to be significantly enriched, and rerun the analysis centered this time on the most likely instance of the motif in each of the input regions. Regions containing a low quality instance for the motif chosen are automatically discarded. The idea is to replace the peak summit for the TF of interest with the most likely location of its actual binding site on DNA. Thus, if two or more TFs have their respective binding sites enriched in the regions, then the motif centered analysis is meant to highlight if there is also a preferential arrangement of their sites in the regions, signaled by the "positional bias" p-value output by PscanChIP. This fact is in turn a strong indicator that the corresponding TFs are likely to interact, require a precise arrangement for their binding sites on DNA, and thus influence the respective recruitment on DNA. Thus, by submitting to PscanChIP cobinding peak regions for two of more TFs, we can assess the enrichment and relative positions of the respective binding sites.

Given a set of ChIP-Seq experiments for TFs and cofactors, and the corresponding peaks and summits, our pipeline can be thus summarized in the following steps:

1. Compute the summit neighborhood overlap for each pair of TFs, and the corresponding p-values;
2. convert the p-values into scores, and cluster the experiments according to the scores; this step is optional, but provides a quick overview of the results obtained;
3. for selected pairs of TFs, define the recruitment and binding rules on DNA by submitting the list of peak summits of either one falling in cobinding regions to PscanChIP:
   a. If motifs for both TFs are found to be significantly enriched according to the global p-value, assess whether there is a preferential arrangement or spacing of the corresponding binding sites through motif centered analysis on either one, by checking whether PscanChIP reports a significant positional bias p-value (< 0.01) for the other; if so, the distribution of the distances between the two motifs can be further analyzed, starting from the relative motif position in each of the input regions reported by PscanChIP.
   b. If only one motif is found to be enriched, then the corresponding TF can be considered to be recruiting the other on DNA.
   c. If neither motif is enriched, then either the motifs employed are not correct for the TFs studied, or there might exist a third factor responsible for the recruitment of the two factors considered.
4. The previous steps can be iterated in order to find significant triplets, quadruplets, and so on, of TFs, and the corresponding binding sites on DNA.

Peak cobinding analysis can be easily implemented with in-house scripts, or with utilities like bedtools (Quinlan and Hall, 2010). A shell script making use of the bedtools "intersect" function (bedtools version 2.29) is provided as **Supplementary File 1**.

PscanChIP is available both through a dedicated web interface, or as a standalone software package. Both already contain the latest release of the JASPAR database. Users can anyway add to the already present collection their own profiles, e.g., the result of a motif discovery analysis on the regions of ChIP-Seq experiments with tools like MEME (Machanick and Bailey, 2011), HOMER (Heinz et al., 2010) or Weeder (Zambelli et al., 2014). Histograms of motif distance distributions presented here were produced by plotting the relative distance between two motifs as output by the motif centered analysis on one of the two of PscanChIP.

MATERIALS

ChIP-Seq data (“optimal thresholded” peak and summit coordinates) for 492 experiments of transcription factors of cofactors in the K562 cell line were retrieved from the ENCODE data repository (Davis et al., 2018) (www.encodeproject.org) as of 31st December of 2018. Each experiment has been performed in at least two replicates, whose consistency has been checked according to different metrics. Only experiments with consistent replicates have been released by ENCODE, with replicates merged into a single list of consensus peak and summit coordinates (Landt et al., 2012).

Since in some cases for the same TF data contained more than one experiment (e.g., with different tagging or antibodies, with or without stimulation of the cells), we filtered the datasets as follows: (1) Experiments on stimulated cells were not considered. (2) In case for the same TF experiments were performed with antibodies against both the wild-type protein and a tagged protein (e.g., with flag or GFP), only the former was kept. Finally, in case of redundant experiments for the same TF not satisfying any of the above conditions we proceeded as follows: (a) if an experiment contained less than 10,000 peaks, and less than half of the peaks of the other(s), it was discarded; (b) if the overlap among the remaining experiments was above 66% we kept the one with the highest number of peaks; otherwise all the experiments for the TF were discarded. Peak overlap was defined as for cobinding peaks, that is, the respective summits had to be located within 150 bps.

After filtering, we obtained non redundant experiments for 329 TFs and other DNA binding proteins. The resulting list, with the respective ENCODE identifiers, is available as **Supplementary Table 1**.

Sequence analysis was performed with PscanChIP version 1.3 (Zambelli et al., 2013b) using the JASPAR 2018 collection of binding sites profiles (Khan et al., 2018), and the K562 background.

RESULTS

A preliminary version of the pipeline we present had been applied to a comprehensive analysis of ENCODE ChIP-Seq data for transcription factors and cofactors in three different cell lines, focused on modules containing transcription factor NF-Y (Dolfini et al., 2016). NF-Y is a trimeric TF composed of two histone-like subunits (NF-YB and NF-YC) and a sequence-specific subunit (NF-YA) binding to the CCAAT motif (CCAAT box). The main difference of our previous work with the pipeline we present here is that, since the original study was focused on NF-Y, motif enrichment analysis was performed as a preliminary step, and cobinding peaks and binding rules were further investigated only for those TF whose binding regions were enriched for the CCAAT box motif. Here, instead, motif enrichment is assessed as a final step, so to include in the pipeline the analysis of colocalization and recruitment for factors not directly contacting DNA.

We consider NF-Y an excellent case study for several reasons. Its binding sites are functionally very well characterized from the genetic point of view, are in general important, and in some cases their presence in promoters is outright essential for the transcription of the corresponding genes. The binding site motif has a high information content, spanning 5 base pairs flanking the central CCAAT, for a total of 10 discriminatory base pairs. The motif is specific for only one complex, hence avoiding the daunting task of disentangling subtle differences in binding preferences among members of large TFs families.

ENCODE data contain experiments for two of the subunits of the complex (NF-YA and NF-YB) in three cell lines. In each cell line, the number of peaks for NF-YA is always lower than NF-YB, and virtually all peaks for NF-YA overlap a peak for NF-YB. A more detailed analysis revealed that the peaks identified for NF-YB only indeed correspond to “quasi-peaks” for NF-YA, that present enrichment levels below detection thresholds for the bioinformatic tools employed. The conclusion was thus that the NF-YB antibody is more efficient than the one for NF-YA, and that the two subunits can be assumed to be found together bound on DNA, as further discussed in (Fleming et al., 2013). We thus employed peaks for NF-YB for our analysis as representative of the whole NF-Y trimer.

Since the original release, new datasets have considerably expanded the ENCODE repository, for new TFs or new cell lines. Furthermore, while the initial ENCODE release contained datasets processed with different tools and strategies, all ChIP-Seq datasets have been reprocessed with a unique bioinformatic pipeline, applying also more stringent quality controls for experiments to be included in the official release. The result is that some of the TFs originally included in the early ENCODE releases -and in our study- have been removed, or the original peak lists changed, both in number of peaks, peak size, or genomic coordinates of peak regions. We thus reprocessed the new datasets, focusing on the K562 cell line, with our pipeline (see also Materials).

An updated version of the results is summarized in **Figure 3**. The heatmap shows the significance of pairwise cobinding between the ENCODE experiments available for the K562 cell line (result of step 1 of the pipeline). The values represent the log10 of the p-value resulting by the statistical assessment of the overlap significance. Blue colors represent overlap higher than expected (-log10 of the p-value), vice versa for black (log10 of the p-value). Four main large clusters are clearly identifiable in the center of the heatmap, formed by general transcription factors as well as promoter-binding TFs. Several smaller clusters however emerge, composed by proteins binding DNA at distal regions away from genes. The complete results are available as **Supplementary Table 2**.

Figure 4 shows the significance of the number of cobinding peaks between pairs of TFs within NF-YB peaks (result of step 2), restricted only to those TFs that had a significant overlap with NF-YB at the first step (enriched with p-value < 10^{-10}). It can be observed how several small clusters emerge, clearly identifiable along the main diagonal, each corresponding to a potential genome-wide regulatory module composed by NF-Y and other factors and cofactors.

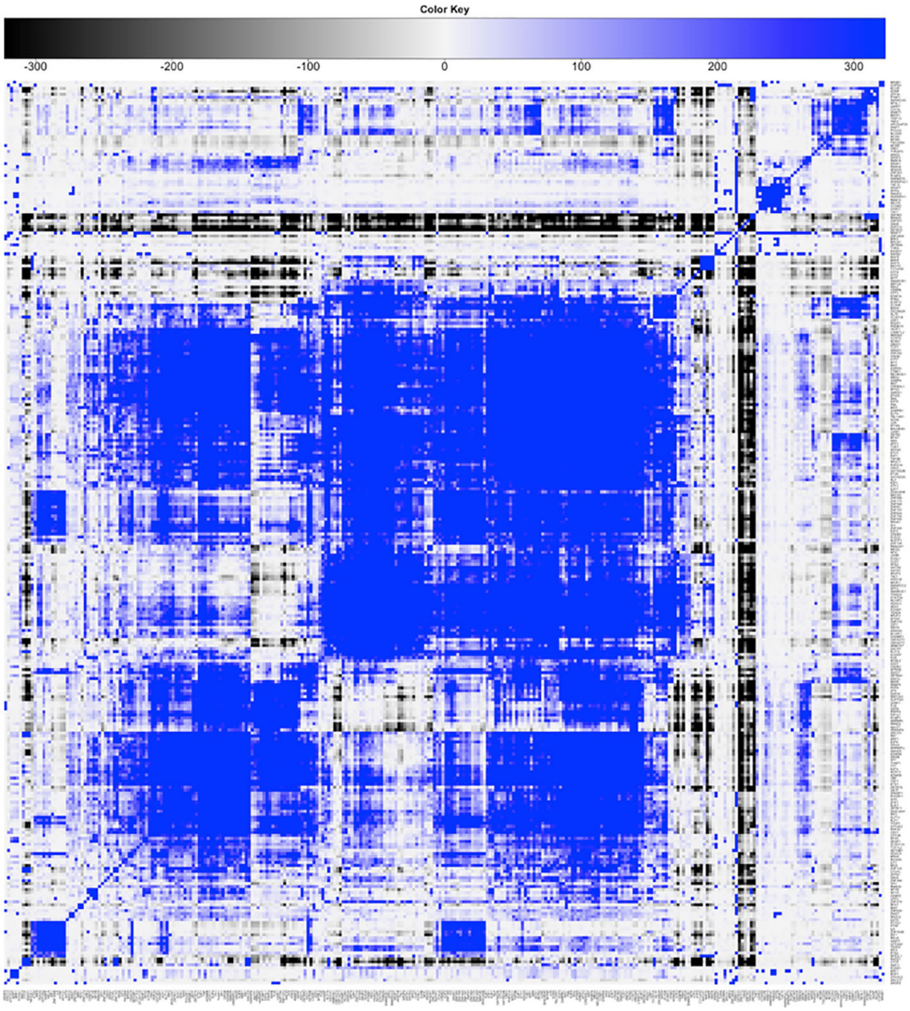


FIGURE 3 | Result of step 1 of the pipeline. Clustered heatmap of pairwise coassociation scores among 329 ENCODE ChIP-Seq experiments in the K562 cell line. Coassociation scores are defined as –log10 of the p-value if the overlap is higher than expected, log10 of the p-value otherwise. Pearson correlation was employed as distance for clustering.

The already identified module containing NF-Y, FOS, and other factors (Dolfini et al., 2016) was confirmed by the new analysis on the reprocessed data (cluster in orange in **Figure 4**). The presence of NF-YA, which colocalizes with NF-YB as a rule, highlights the significance of this cluster, that is, it covers a significant fraction of the NF-Y binding sites on the genome. FOS is known to form a dimer with JUN, and to bind DNA on the AP1 motif. The surprising result of our analysis was that in the regions of FOS/NF-Y overlap the AP1 motif is not enriched, but indeed seemed to be avoided (under-represented according to PscanChIP), with the CCAAT box bound by NF-Y as the most enriched one. Vice versa, FOS summits not overlapping with NF-Y had the expected AP1 as the most enriched motif. Motif centered analysis on the NF-Y/FOS cobinding peaks identified a second binding motif for NF-Y, with the two CCAAT boxes located with precise spacing on DNA, hinting at two NF-Y molecules forming a complex with FOS (Dolfini et al., 2016; Zambelli and Pavesi, 2017). The same conclusion has been confirmed by independent studies, leading to the interesting hypothesis of a single complex connecting enhancers bound by JUN/FOS to a promoter bound by NF-Y (Haubrock et al., 2016).

To further substantiate these results we repeated the analysis of step 2 computing the significance of cobinding peaks between pairs of TFs within FOS peaks, shown in **Figure 5**. Two distinct clusters are easily identifiable, one (highlighted red in the figure) composed by NF-YA/NF-YB and the other factors clustering with NF-Y and FOS in the previous analysis. The second one (green in **Figure 5**) is composed by factors forming the canonical AP1 complex (JUN/JUNB/JUND). The two clusters and clearly separated, and, more interestingly, the members of each show a significant under-representation of their overlap with the others. In other words, the number of cobinding peaks between pairs of

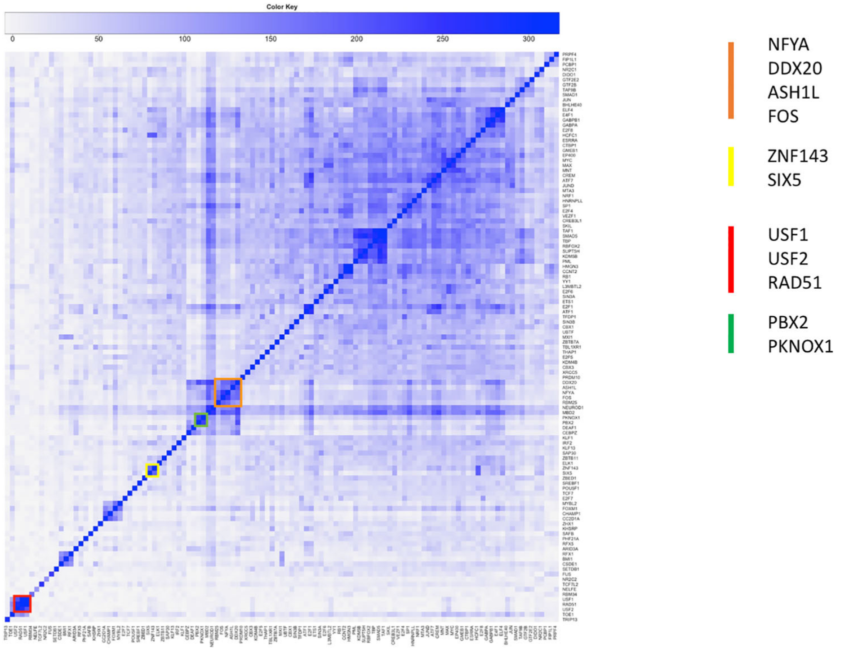


FIGURE 4 | Clustered heatmap of pairwise coassociation scores restricted to peaks cobinding with NF-YB for transcription factors (TFs) with significant overlap with NF-YB (p-value lower than 10^{-10}). Pearson correlation was employed as distance for clustering.

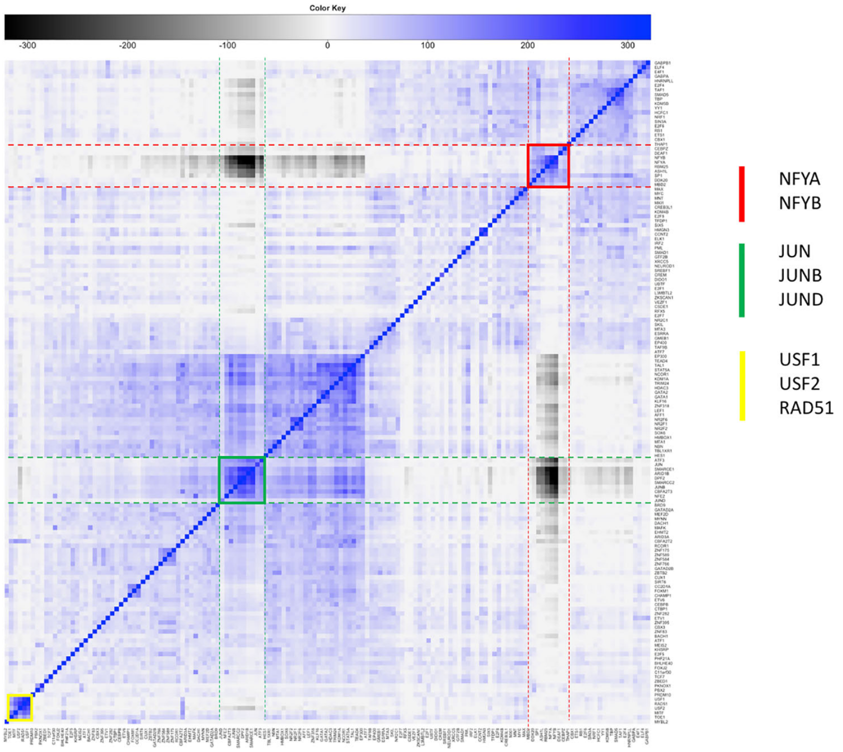


FIGURE 5 | Clustered heatmap of pairwise coassociation scores restricted to peaks cobinding with FOS for transcription factors (TFs) with significant overlap with FOS (p-value lower than 10^{-100}). Pearson correlation was employed as distance for clustering.

members belonging to different clusters is significantly lower than expected, converted into negative scores represented in grayscale in the heatmap. The overall message thus becomes clear: FOS is recruited on DNA by forming a complex either with NF-Y or with JUN factors, but never by both. In fact, when members of either cluster are found with FOS the others are avoided, and vice versa.

Another cluster (in red in **Figure 4**) shows the overlap of NF-Y with both USF1 and USF2, generalizing to the whole genome previous observations (Zhu et al., 2003). USF factors in turn show a significant colocalization within NF-Y peaks together with RAD51. In this case, the motif enrichment analysis for both the NF-Y/USF1 and NF-Y/USF2 cobinding regions returns both the CCAAT-box and the expected E-box as significantly enriched, with a strikingly precise spacing and orientation between the two (shown in **Figure 6** for NF-Y/USF1 cobinding peaks). In most of the cobinding regions, the CCAAT is located downstream of the E-box, at 17 or 18 bps of distance.

Interestingly, the USF1/USF2 cluster emerges in coassociation with FOS as well (highlighted in yellow in **Figure 5**). Indeed, the interactions of FOS with USF1/2 has been known ever since the discovery of the latter (Blanar and Rutter, 1992; Aperlo et al., 1996). The USF cluster in **Figure 5** does not shows significant overlap with either the NF-Y or the AP1 cluster. Thus, to determine whether FOS colocalizes with USF1/2 with or without NF-Y, we performed another cobinding analysis centered on USF1 peaks (**Figure 7**). Here several clusters emerge, and, strikingly, one small cluster composed exactly bby NF-YA, NF-YB, and FOS (highlighted in red in **Figure 7**). However, the cobinding of FOS with JUN in USF1 peaks is also significant, although JUN clusters elsewhere with members of the AP1 complex (green in **Figure 7**).

By combining the results obtained from the three different points of view just studied, the overall picture emerges. FOS can be recruited either by NF-Y or as a member of the AP1 complex with JUN factors, and the two modes are mutually exclusive. When USF1 is found on DNA together with NF-Y or FOS, it is in general with USF2; when FOS is bound on DNA with USF1/2, it is mainly found in the NF-Y complex, but not exclusively; that is, USF1/2 can be found in a smaller, but significant number of genomic loci also in association with the AP1 complex containing FOS. Complete cobinding statistics for the three TF centered analyses are available as **Supplementary Tables 3–5**.

Another example of combinations of factors colocalizing with NF-Y is the SIX5/ZNF143 pair (yellow cluster in **Figure 4**). The CCAAT box and the ZNF143 binding motifs show evident

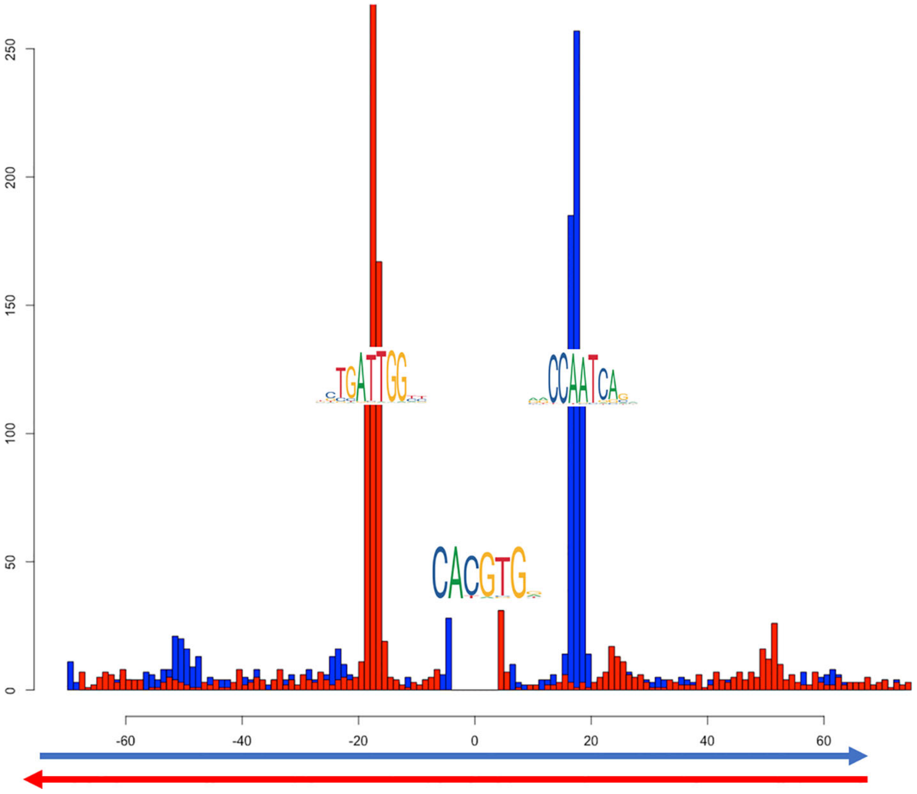


FIGURE 6 | Distribution of distances between the most likely instances of the CCAAT box and the E-box in cobinding peaks of NF-YB with USF1, as reported by PscanChIP from in the motif centered analysis on the USF1 motif (**Supplementary Table 6**). The blue histogram shows the distribution of the position of the CCAAT box when found on the positive strand of the genome; red when on the negative strand. The origin of the x-axis corresponds to the center of the USF1 binding sites (E-box). The analysis has been performed on 2748 cobinding regions for NF-YB and USF1.

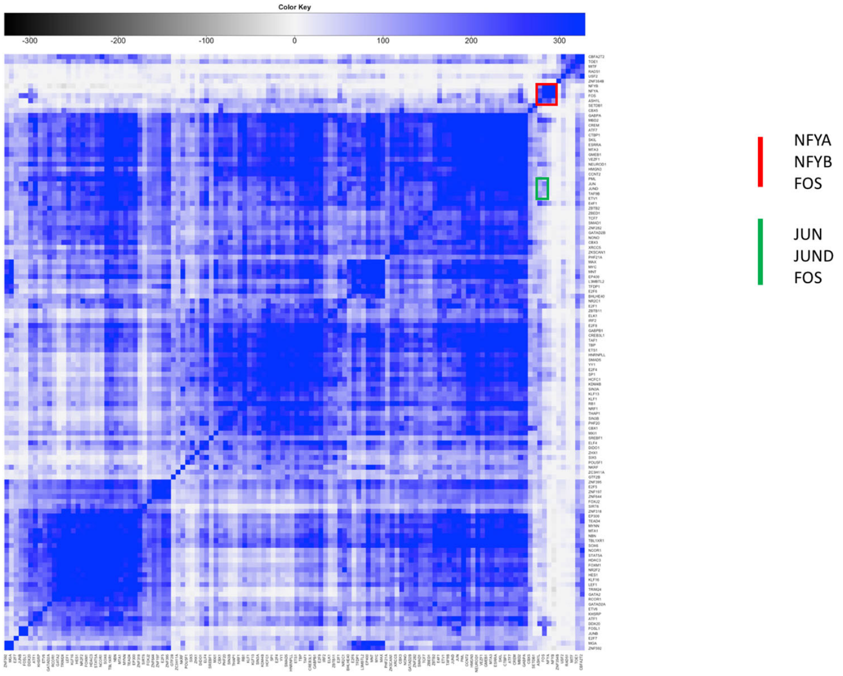


FIGURE 7 | Clustered heatmap of pairwise coassociation scores restricted to peaks cobinding with USF1 for transcription factors (TFs) with significant overlap with USF1 (p-value lower than 10^{-100}). Pearson correlation was employed as distance for clustering.

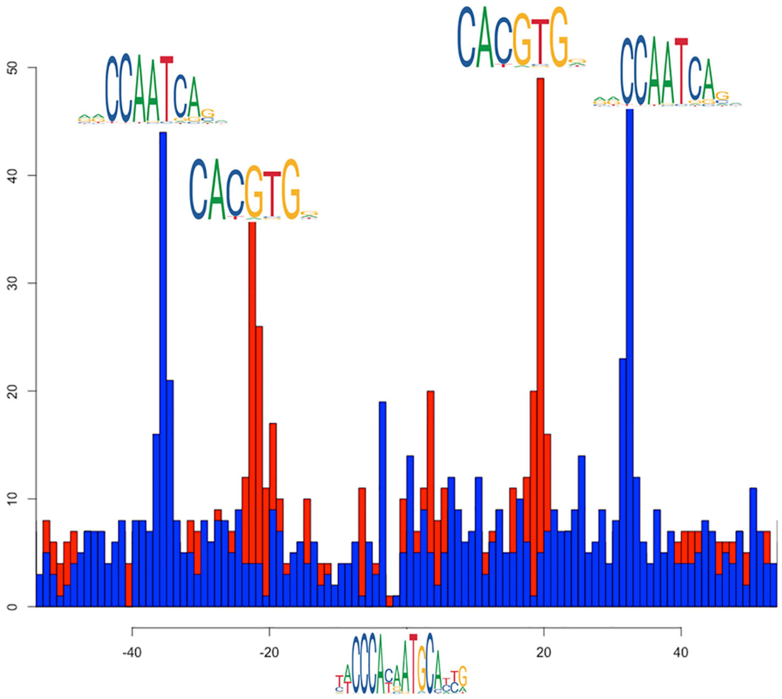


FIGURE 8 | Distribution relative positions of the CCAAT box (blue histogram) and the E-box (red histogram) around the most likely instances of the SIX5/ZNF143 binding site in the cobinding peaks of NF-YB with ZNF143, as reported by PscanChIP from in the motif centered analysis on the ZNF143 motif (**Supplementary Table 7**). The origin of the x-axis corresponds to the center of the ZNF143 binding sites. The analysis has been performed on 1424 cobinding regions for NF-YB and ZNF143.

preferential spacing (**Figure 8**). Sequence analysis also returned significant enrichment and positional bias for an additional E-box motif, also plotted in **Figure 8**, located in between the ZNF143 and CCAAT motifs, once again with a strong positional preference. Thus, in this case, the preferential arrangement of binding sites on the genome seems to be ZNF143/E-BOX/CCAAT, on either strand, with a precise spacing. Since none of the known E-box binding TFs so far included in the K562 datasets shows relevant cobinding with ZNF143 inside NF-Y peaks, it remains to be determined what could be the actual TF binding the E-Boxes, or if there are different TFs of the same family binding each a subset of them.

A final example is colocalizing peaks with precise motif arrangement of NF-YB with PBX2 (in turn with significant overlap with PKNOX1, green cluster in **Figure 4**): the respective binding sites on the genome can be found, once again with a clear distance preference (**Figure 9**). The interaction of NF-Y with TALE transcription factors, including PBX2, and the arrangement of their binding sites on DNA has indeed been recently reported as for example in zebrafish (Ladam et al., 2018). In this case, however, PscanChIP motif analysis reports that in about 20% of the cobinding peaks the CCAAT box motif is returned to be the most likely candidate also for the binding of PBX2, since its consensus motif (CTGTCAATCA) in turn contains a CAAT subsequence (see **Supplementary Table 8**). Also, the p-value associated by PscanChIP to the PBX2 motif is only marginally significant. Thus, it remains to be ascertained whether the binding motifs found on DNA are actually bound by the respective transcription factors in all the cobinding regions, or, as more likely, there are instances where a single or double CCAAT box bound by NF-Y is the motif tethering the complex on DNA.

DISCUSSION

We presented a computational pipeline that, starting from a collection of peak regions resulting from the analysis of different TFs and cofactors, is able to single out the most relevant TF combinations and modules in the condition studied. The integration of peak and summit overlap with a sequence analysis method developed specifically for the analysis of ChIP-Seq regions also permits the characterization of the recruitment rules on DNA for the complex and the organization of the respective binding sites on the genome.

A preliminary version of this pipeline has been applied to the analysis of the complete collection of ENCODE ChIP-Seq experiments in three different cell lines, focusing on modules containing transcription factor NF-Y. In this work, we reanalyzed the updated ENCODE data for K562, essentially

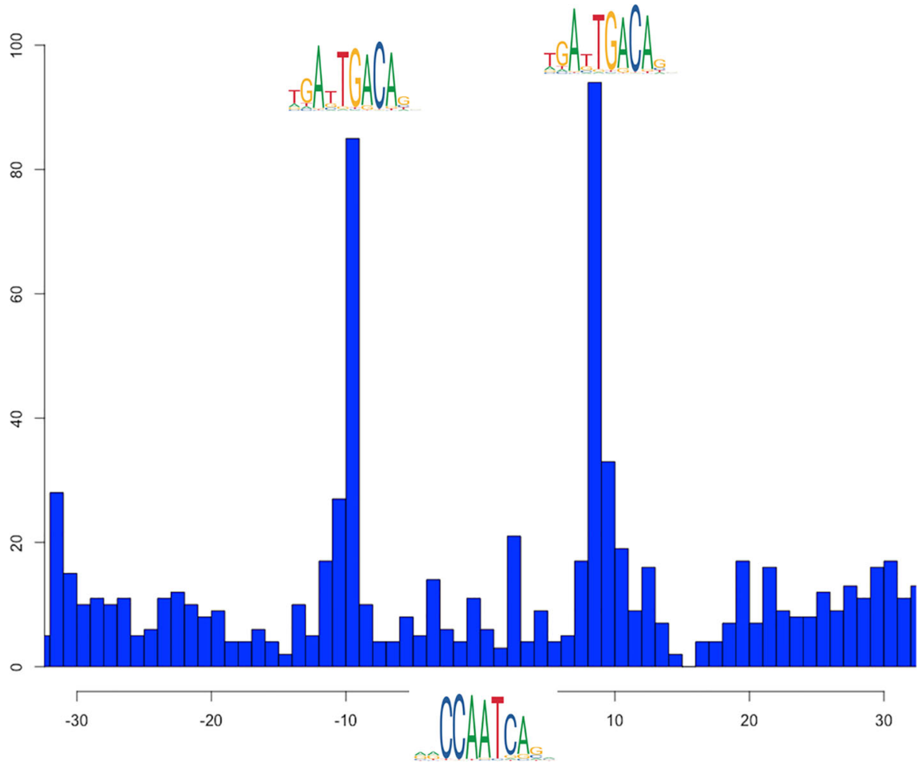


FIGURE 9 | Distribution relative positions of PBX2 binding site with respect to the CCAAT box motif in cobinding peaks of NF-YB with PBX2, as reported by PscanChIP from in the motif centered analysis on the CCAAT box motif (**Supplementary Table 8**). The origin of the x-axis corresponds to the center of the CCAAT box binding sites. The analysis has been performed on 1782 cobinding regions for NF-YB and ZNF143.

confirming the previous results for NF-Y, as well as finding novel candidate interactors and genome-wide coassociations involving also FOS, USF1, and USF2. We are now working on manuscripts detailing the results obtained also on additional cell lines, on different TFs and cofactors, as well as linking these findings to functionality.

Our approach permits to build a picture of the regulatory landscape of a given condition, highlighting the TF coassociations found more frequently, and assessing their significance as well as the corresponding organization of binding sites on the genome. It can be integrated with additional sources of information. For example, one could focus on active promoters or enhancers, as revealed by presence of specific histone marks, and restrict the cobinding peak analysis only to those regions that carry a precise chromatin annotation, resulting for example from a genome segmentation approach (Hoffman et al., 2009; Ernst and Kellis, 2017). In this way separate maps of regulatory modules specific for enhancers and/or promoters can be built. The data can be complemented with RNA-Seq experiments performed after inactivation of the single TF, so that the functionality—positive, negative, or neutral—of the single modules can be inferred. Finally, the exact pattern of binding in a single selected region can be further analyzed by employing more sophisticated sequence analysis approaches (Gheorghe et al., 2019).

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found at the Encode Project (www.encodeproject.org) and the accession numbers are listed in **Supplementary Material.**

AUTHOR CONTRIBUTIONS

GP and RM devised the original pipeline, that has been fine tuned and modified with input from DD, MR, and FZ. MR and FZ implemented different parts of the pipeline, supervised by GP. MR run the analyses presented in the article and collected the results. All authors examined and discussed the results.

FUNDING

Consiglio Nazionale delle Ricerche, Flagship Research Project "EPIGEN."

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.00072/full#supplementary-material

SUPPLEMENTARY FILE 1 | Shell script implementing the key steps of the pipeline through the "bedtools" utility (version 2.29).

SUPPLEMENTARY TABLE 1 | List of ENCODE identifiers of the ChIP-Seq datasets employed in this study.

SUPPLEMENTARY TABLE 2 | Results of the peak/summit overlap step between all pairs of ENCODE datasets in the K562 cell line.

SUPPLEMENTARY TABLE 3 | Results of the peak/summit overlap step between all pairs of ENCODE datasets in the K562 cell line within NF-YB peaks.

SUPPLEMENTARY TABLE 4 | Results of the peak/summit overlap step between all pairs of ENCODE datasets in the K562 cell line within FOS peaks.

SUPPLEMENTARY TABLE 5 | Results of the peak/summit overlap step between all pairs of ENCODE datasets in the K562 cell line within USF1 peaks.

SUPPLEMENTARY TABLE 6 | List of co-binding summit coordinates for NF-YB and USF1, complete PscanChIP output, and relative position of NF-YB binding sites with respect to USF1 sites as output by PscanChIP.

SUPPLEMENTARY TABLE 7 | List of co-binding summit coordinates for NF-YB and ZNF143, complete PscanChIP output, and relative position of NF-YB and E-Box binding sites with respect to ZNF143 sites as output by PscanChIP.

SUPPLEMENTARY TABLE 8 | List of co-binding summit coordinates for NF-YB and PBX2, complete PscanChIP output, and relative position of NF-YB with respect to PBX2 sites as output by PscanChIP.

REFERENCES

Aperlo, C., Boulukos, K. E., and Pognonec, P. (1996). The basic region/helix-loop-helix/leucine repeat transcription factor USF interferes with Ras transformation. *Eur. J. Biochem.* 241, 249–253. doi: 10.1111/j.1432-1033.1996.0249t.x

Bailey, T. L., and MacHanick, P. (2012). Inferring direct DNA binding from ChIP-seq. *Nucleic Acids Res.* 40 (17), e128. doi: 10.1093/nar/gks433

Blanar, M. A., and Rutter, W. J. (1992). Interaction cloning: identification of a helix-loop-helix zipper protein that interacts with c-Fos. *Science* (80-), 1014–1018. doi: 10.1126/science.1589769

Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y., and Greenleaf, , W. J. (2013). Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods* 10, 1213–1218. doi: 10.1038/nmeth.2688

Celona, B., Weiner, A., Di Felice, F., Mancuso, F. M., Cesarini, E., Rossi, R. L., et al. (2011). Substantial Histone reduction modulates Genomewide nucleosomal occupancy and global transcriptional output. *PloS Biol.* 9 (6), e1001086. doi: 10.1371/journal.pbio.1001086

Chèneby, J., Gheorghe, M., Artufel, M., Mathelier, A., and Ballester,, B. (2018). ReMap 2018: an updated atlas of regulatory regions from an integrative analysis of DNA-binding ChIP-seq experiments. *Nucleic Acids Res.* 46, D267–D275. doi: 10.1093/nar/gkx1092

Chen, X., Xu, H., Yuan, P., Fang, F., Huss, M., Vega, V. B., et al. (2008). Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. *Cell* 133, 1106–1117. doi: 10.1016/j.cell.2008.04.043

Czipa, E., Schiller, M., Nagy, T., Kontra, L., Steiner, L., Koller, J., et al. (2020). ChIPSummitDB: a ChIP-seq-based database of human transcription factor binding sites and the topological arrangements of the proteins bound to them. *Database* 2020, baz141. doi: 10.1093/database/baz141

Davis, C. A., Hitz, B. C., Sloan, C. A., Chan, E. T., Davidson, J. M., Gabdank, I., et al. (2018). The encyclopedia of DNA elements (ENCODE): data portal update. *Nucleic Acids Res.* 46, D794–D801. doi: 10.1093/nar/gkx1081

Dergilev, A. I., Spitsina, A. M., Chadaeva, I. V., Svichkarev, A. V., Naumenko, F. M., Kulakova, E. V., et al. (2017). Computer analysis of colocalization of the TFs' binding sites in the genome according to the ChIP-seq data. *Russ. J. Genet. Appl. Res.* 7, 513–522. doi: 10.1134/S2079059717050057

Dolfini, D., Zambelli, F., Pedrazzoli, M., Mantovani, R., and Pavesi,, G. (2016). A high definition look at the NF-Y regulome reveals genome-wide associations with selected transcription factors. *Nucleic Acids Res.* 44, 4684–4702. doi: 10.1093/nar/gkw096

Ernst, J., and Kellis, M. (2017). Chromatin-state discovery and genome annotation with ChromHMM. *Nat. Protoc.* 12, 2478–2492. doi: 10.1038/nprot.2017.124

Fleming, J. D., Pavesi, G., Benatti, P., Imbriano, C., Mantovani, R., and Struhl, K. (2013). NF-Y coassociates with FOS at promoters, enhancers, repetitive elements, and inactive chromatin regions, and is stereo-positioned with growth-controlling transcription factors. *Genome Res.* 23, 1195–1209. doi: 10.1101/gr.148080.112

Fullwood, M. J., and Ruan, Y. (2009). ChIP-based methods for the identification of long-range chromatin interactions. *J. Cell Biochem.* 107, 30–39. doi: 10.1002/jcb.22116

Gerstein, M. B., Kundaje, A., Hariharan, M., Landt, S. G., Yan, K. K., Cheng, C., et al. (2012). Architecture of the human regulatory network derived from ENCODE data. *Nature* 489, 91–100. doi: 10.1038/nature11245

Gheorghe, M., Sandve, G. K., Khan, A., Chèneby, J., Ballester, B., and Mathelier, A. (2019). A map of direct TF-DNA interactions in the human genome. *Nucleic Acids Res.* 47, e21. doi: 10.1093/nar/gky1210

Giresi, P. G., Kim, J., McDaniell, R. M., Iyer, V. R., and Lieb, J. D. (2007). FAIRE (formaldehyde-assisted isolation of regulatory elements) isolates active regulatory elements from human chromatin. *Genome Res.* 17, 877–885. doi: 10.1101/gr.5533506

Harwood, J. C., Kent, N. A., Allen, N. D., and Harwood, A. J. (2019). Nucleosome dynamics of human iPSC during neural differentiation. *EMBO Rep.* 20 (6), e46960. doi: 10.15252/embr.201846960

Haubrock, M., Hartmann, F., and Wingender, E. (2016). NF-Y binding site architecture defines a C-Fos targeted promoter class. *PloS One* 11 (8), e0160803. doi: 10.1371/journal.pone.0160803

Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y. C., Laslo, P., et al. (2010). Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol. Cell* 38, 576–589. doi: 10.1016/j.molcel.2010.05.004

Hoffman, M. M., Buske, O., Bilmes, J., and Noble, W. (2009). Segway: simultaneous segmentation of multiple functional genomics data sets with heterogeneous patterns of missing data. NobleGsWashingtonEdu 2–5.

Johnson, D. S., Mortazavi, A., Myers, R. M., and Wold, B. (2007). Genome-wide mapping of in vivo protein-DNA interactions. *Science* (80-), 1497–1502. doi: 10.1126/science.1141319

Kanduri, C., Bock, C., Gundersen, S., Hovig, E., and Sandve, G. K. (2019). Colocalization analyses of genomic elements: approaches, recommendations and challenges. *Bioinformatics* 35, 1615–1624. doi: 10.1093/bioinformatics/bty835

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, D260–D266. doi: 10.1093/nar/gkx1126

Ladam, F., Stanney, W., Donaldson, I. J., Yildiz, O., Bobola, N., and Sagerström, C. G. (2018). TALE factors use two distinct functional modes to control an essential zebrafish gene expression program. *Elife* 7, e36144. doi: 10.7554/eLife.36144

Landt, S. G., Marinov, G. K., Kundaje, A., Kheradpour, P., Pauli, F., Batzoglou, S., et al. (2012). ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Res.* 22, 1813–1831. doi: 10.1101/gr.136184.111

Levitsky, V., Zemlyanskaya, E., Oshchepkov, D., Podkolodnaya, O., Ignatieva, E., Grosse, I., et al. (2019). A single ChIP-seq dataset is sufficient for comprehensive analysis of motifs co-occurrence with MCOT package. *Nucleic Acids Res.* 47(21), e139. doi: 10.1093/nar/gkz800

Lieberman-Aiden, E., Van Berkum, N. L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., et al. (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* (80-), 289–293. doi: 10.1126/science.1181369

Machanick, P., and Bailey, T. L. (2011). MEME-ChIP: Motif analysis of large DNA datasets. *Bioinformatics* 27, 1696–1697. doi: 10.1093/bioinformatics/btr189

Mathelier, A., Fornes, O., Arenillas, D. J., Chen, C. Y., Denay, G., Lee, J., et al. (2016). JASPAR 2016: A major expansion and update of the open-access database of transcription factor binding profiles. *Nucleic Acids Res.* 44, D110–D115. doi: 10.1093/nar/gkv1176

Oki, S., Ohta, T., Shioi, G., Hatanaka, H., Ogasawara, O., Okuda, Y., et al. (2018). ChIP -Atlas: a data-mining suite powered by full integration of public ChIP -seq data. *EMBO Rep.* 19 (12), e46255. doi: 10.15252/embr.201846255

Pajoro, A., Muiño, J. M., Angenent, G. C., and Kaufmann, K. (2018). "Profiling nucleosome occupancy by MNase-seq: experimental protocol and computational analysis," in *Methods in molecular biology*, (New York, NY: Humana Press), 167–181.

Quinlan, A. R., and Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26, 841–842. doi: 10.1093/bioinformatics/btq033

Roadmap Epigenomics Consortium,, Kundaje, , A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., et al. (2015). Integrative analysis of 111 reference human epigenomes. *Nature* 518, 317–330. doi: 10.1038/nature14248

Sabo, P. J., Hawrylycz, M., Wallace, J. C., Humbert, R., Yu, M., Shafer, A., et al. (2004). Discovery of functional noncoding elements by digital analysis of chromatin structure. *Proc. Natl. Acad. Sci. U. S. A.* 101, 16837–16842. doi: 10.1073/pnas.0407387101

Salvatore, S., Dagestad Rand, K., Grytten, I., Ferkingstad, E., Domanska, D., Holden, L., et al. (2019). Beware the Jaccard: the choice of similarity measure is important and non-trivial in genomic colocalisation analysis. *Brief Bioinform*. doi: 10.1093/bib/bbz083

Stormo, G. D. (2000). DNA binding sites: representation and discovery. *Bioinformatics* 16, 16–23. doi: 10.1093/bioinformatics/16.1.16

Tannenbaum, M., Sarusi-Portuguez, A., Krispil, R., Schwartz, M., Loza, O., and Benichou, , J. I. C. (2018). Regulatory chromatin landscape in Arabidopsis thaliana roots uncovered by coupling INTACT and ATAC-seq. *Plant Methods* 14, 113. doi: 10.1186/s13007-018-0381-9

Thomas, S., Li, X. Y., Sabo, P. J., Sandstrom, R., Thurman, R. E., Canfield, T. K., et al. (2011). Dynamic reprogramming of chromatin accessibility during Drosophila embryo development. *Genome Biol.* 12. doi: 10.1186/gb-2011-12-5-r43

Thomas, R., Thomas, S., Holloway, A. K., and Pollard, K. S. (2017). Features that define the best ChIP-seq peak calling algorithms. *Brief Bioinform.* 18, 441–450. doi: 10.1093/bib/bbw035

Vierstra, J., and Stamatoyannopoulos, J. A. (2016). Genomic footprinting. *Nat. Methods* 13, 213–221. doi: 10.1038/nmeth.3768

Wang, J., Zhuang, J., Iyer, S., Lin, X. Y., Whitfield, T. W., Greven, M. C., et al. (2012). Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. *Genome Res.* 22, 1798–1812. doi: 10.1101/gr.139105.112

Wingender, E., Schoeps, T., Haubrock, M., Krull, M., and Dönitz, J. (2018). TFClass: expanding the classification of human transcription factors to their mammalian orthologs. *Nucleic Acids Res.* 46, D343–D347. doi: 10.1093/nar/gkx987

Wingender, E. (2008). The TRANSFAC project as an example of framework technology that supports the analysis of genomic regulation. *Brief Bioinform.* 9, 326–332. doi: 10.1093/bib/bbn016

Xie, D., Boyle, A. P., Wu, L., Zhai, J., Kawli, T., and Snyder, M. (2013). Dynamic trans-acting factor colocalization in human cells. *Cell* 155, 713. doi: 10.1016/j.cell.2013.09.043

Zambelli, F., and Pavesi, G. (2017). Genome wide features, distribution and correlations of NF-Y binding sites. *Biochim. Biophys. Acta* 1860, 581–589. doi: 10.1016/j.bbagrm.2016.10.007

Zambelli, F., Pesole, G., and Pavesi, G. (2013a). Motif discovery and transcription factor binding sites before and after the next-generation sequencing era. *Brief Bioinform.* 14, 225–237. doi: 10.1093/bib/bbs016

Zambelli, F., Pesole, G., and Pavesi, G. (2013b). PscanChIP: finding over-represented transcription factor-binding site motifs and their correlations in sequences from ChIP-Seq experiments. *Nucleic Acids Res.* 41, W535–W543. doi: 10.1093/nar/gkt448

Zambelli, F., Pesole, G., and Pavesi, G. (2014). Using weeder, Pscan, and PscanChIP for the discovery of enriched transcription factor binding site motifs in nucleotide sequences. *Curr. Protoc. Bioinform.* 47, 2.11.1–2.1131. doi: 10.1002/0471250953.bi0211s47

Zhang, Y., Liu, T., Meyer, C. A., Eeckhoute, J., Johnson, D. S., Bernstein, B. E., et al17. (2008). Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* 9, R137. doi: 10.1186/gb-2008-9-9-r137

Zhang, Z., Chang, C. W., Goh, W. L., Sung, W. K., and Cheung, E. (2011). CENTDIST: discovery of co-associated factors by motif distribution. *Nucleic Acids Res.* 39, W391–W399. doi: 10.1093/nar/gkr387

Zhu, J., Giannola, D. M., Zhang, Y., Rivera, A. J., and Emerson, S. G. (2003). NF-Y cooperates with USF1/2 to induce the hematopoietic expression of HOXB4. *Blood* 102, 2420–2427. doi: 10.1182/blood-2003-01-0251

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.

Testing Proximity of Genomic Regions to Transcription Start Sites and Enhancers Complements Gene Set Enrichment Testing

*Christopher Lee[1,2], Kai Wang[1], Tingting Qin[1] and Maureen A. Sartor[1,2]**

[1] Department of Computational Medicine and Bioinformatics, School of Medicine, University of Michigan, Ann Arbor, MI, United States, [2] Department of Biostatistics, School of Public Health, University of Michigan, Ann Arbor, MI, United States

OPEN ACCESS

Edited by:
Geir Kjetil Sandve,
University of Oslo, Norway

Reviewed by:
Shenglin Mei,
Harvard Medical School,
United States
Zhaohui Steve Qin,
Emory University, United States

***Correspondence:**
Maureen A. Sartor
sartorma@umich.edu;
sartorma@med.umich.edu

Specialty section:
This article was submitted to
Bioinformatics and Computational
Biology,
a section of the journal
Frontiers in Genetics

Received: 08 October 2019
Accepted: 20 February 2020
Published: 06 March 2020

Citation:
Lee C, Wang K, Qin T and
Sartor MA (2020) Testing Proximity
of Genomic Regions to Transcription
Start Sites and Enhancers
Complements Gene Set Enrichment
Testing. Front. Genet. 11:199.
doi: 10.3389/fgene.2020.00199

Large sets of genomic regions are generated by the initial analysis of various genome-wide sequencing data, such as ChIP-seq and ATAC-seq experiments. Gene set enrichment (GSE) methods are commonly employed to determine the pathways associated with them. Given the pathways and other gene sets (e.g., GO terms) of significance, it is of great interest to know the extent to which each is driven by binding near transcription start sites (TSS) or near enhancers. Currently, no tool performs such an analysis. Here, we present a method that addresses this question to complement GSE methods for genomic regions. Specifically, the new method tests whether the genomic regions in a gene set are significantly closer to a TSS (or to an enhancer) than expected by chance given the total list of genomic regions, using a non-parametric test. Combining the results from a GSE test with our novel method provides additional information regarding the mode of regulation of each pathway, and additional evidence that the pathway is truly enriched. We illustrate our new method with a large set of ENCODE ChIP-seq data, using the *chipenrich* Bioconductor package. The results show that our method is a powerful complementary approach to help researchers interpret large sets of genomic regions.

Keywords: gene set enrichment test, ChIP-seq data analysis, non-parametric test, pathway analysis, genomic regions

INTRODUCTION

Cell development and differentiation depend on complex gene expression patterns which are precisely and spatiotemporally controlled. The complex process of gene regulation involves many different mechanisms, including regulation of transcription (Berger, 2007; Deaton and Bird, 2011), post-transcriptional regulation (Roundtree et al., 2017), and regulation of translation (Sonenberg and Hinnebusch, 2009). Transcription is the first step to decode the genetic information from

DNA to functional elements, and this process is regulated by many *cis*-regulatory elements across the genome (Wittkopp and Kalay, 2011). *Cis*-regulatory elements include promoters, enhancers, silencers, and insulators, with promoters and enhancers being two important ones that can initiate transcription and are the most well-studied (Andersson, 2015). Both promoters and enhancers are regions of DNA sequences that typically are a few hundred base pairs in length (Nguyen et al., 2016). Promoters are usually located immediately upstream of the transcription start sites (TSSs) on the 5′ end of target genes (Sanyal et al., 2012) and recruit transcription factors (TFs) and RNA polymerase II (RNAPII) to instruct the direction and initiation of transcription (Schoenfelder and Fraser, 2019). Conversely, enhancers can be located upstream, downstream, or in the intron of the target gene or another unrelated gene (Shlyueva et al., 2014) and bound by TFs and cofactors to activate or increase the transcription rate of their target genes (Li et al., 2016). The protein sequences and regulatory motifs of many TFs are well conserved across living organisms, indicating that genome-wide gene regulatory mechanisms have important conserved properties (Lambert et al., 2018). However, some TFs such as ESR1 bind to different sets of target genes in a cell type specific manner (Gertz et al., 2012), resulting in complex and dynamic TF regulatory programs. Thus, deciphering the rules of TF binding events is a key step to understanding gene expression patterns and associated biological pathways.

A diverse collection of sequence-based approaches exist to probe the gene regulome (Pinsach-Abuin et al., 2016). For instance, ChIP-seq can provide genome-wide information about gene regulation for specific TFs or chromatin marks by identifying thousands of genomic regions (i.e., peaks, which we will refer to for simplicity) across the genome (Schmidt et al., 2009). ATAC-seq and copy number variation (CNV) sequencing are also popular for studying genome-wide regulation (Xie and Tammi, 2009; Buenrostro et al., 2015). Through the aforementioned sequencing data, we can identify significant peaks that were bound by a particular TF or modified chromatin mark (ChIP-seq), open chromatin regions (ATAC-seq), or regions with a CNV. We can further infer their underlying regulatory functions by associating the identified regions with target genes, whether predicted or verified. Since biological processes involve many genes and pathways, gene-centered analysis on regulome data may not be as informative as Gene Set Enrichment (GSE) testing (Subramanian et al., 2005).

Most GSE methods were developed for gene expression data, do not adjust for the varying lengths of genes or regulatory space between them, and thus are not generally appropriate for GSE testing with large sets of peaks. However, several GSE methods have been developed to specifically test sets of peaks, including GREAT (McLean et al., 2010), ChIP-Enrich (Welch et al., 2014), Broad-Enrich (Cavalcante et al., 2014), and Poly-Enrich (Lee et al., 2018). Among these, Poly-Enrich is the only method that counts genomic regions (which we will refer to as peaks for simplicity) for each gene, adjusts for the varying lengths of genes and regulatory space between them, and provides a flexible approach with the ability to assign weights to peaks.

Current methods for GSE testing of peaks focus mainly on the relationship between peaks and TSSs (promoters). However, although some TFs [e.g., E2F1 (Ertosun et al., 2016)] preferentially bind to promoters, others [e.g., FOXA1 (Pristera et al., 2015)] tend to bind enhancers, while still other TFs bind to both enhancers and promoters depending on context (e.g., master regulators, such as Serum response factor). Therefore, it is of great interest to know the patterns of TF binding with respect to promoters and enhancers of the target genes and pathways. Although GREAT (McLean et al., 2010), ChIP-Enrich (Welch et al., 2014), and Poly-Enrich (Lee et al., 2018) incorporate distal binding events in their GSE testing, no method has been established for answering the question of whether a TF is binding closer to TSSs, near enhancers, both, or neither for a specific gene set.

Other methods such as ChIPseeker (Yu et al., 2015) and Seq2pathway (Wang et al., 2015) also perform GSE testing for genomic regions. Different from previous GSE testing methods that assign peaks to nearest TSS (NTSS), ChIPseeker applies a max distance cutoff for assigning peaks to genes. Seq2pathway incorporates the significance of each genomic region and both coding and non-coding regions in GSE testing. Methods such as Cistrome-GO (Li et al., 2019) and TREG (Chen et al., 2013) incorporate the distance between ChIP-seq peaks and a gene's TSS into the GSE testing itself. Cistrome-GO integrates the peak distance to TSS and the peak number together to estimate the gene regulation potential. TREG collects the peak distances to a gene's TSS within a 2Mb window around each TSS into the GSE test. However, since these methods embed the information about binding proximity to a TSS within the test itself, it is difficult for the user to interpret the results with respect to this information, or separate the effect of proximity from that of enrichment. A recently published, unique tool called loci2path (Xu et al., 2019) links a set of genomic regions to key pathways by testing for enrichment of expression quantitative trait loci (eQTLs) in the genomic regions, including tissue-specific analyses. By using eQTL target genes, loci2path does not rely on assigning genomic regions to the nearest gene, and thus it is a complementary method to a proximity test. No method, to our knowledge, incorporates enhancer proximity.

Here, we propose a new method, Proximity Regulation (ProxReg) to address this shortcoming of current methods. By measuring the distance between each peak and the closest TSS (or enhancer) and then performing a modified two-sided Wilcoxon rank-sum test, we test whether the peaks in a gene set are significantly closer to TSSs or enhancers than expected by chance. Our method, in combination with a GSE test, is able to provide additional evidence that a pathway is truly enriched and information on the regulatory mechanism for that enrichment. After validating the Type I error rate of our method, we test ProxReg by applying it, in combination with Poly-Enrich (implemented in the *chipenrich* Bioconductor package) to 90 ENCODE ChIP-seq datasets (Sloan et al., 2015), including 35 TFs. In many cases, this led to a significant improvement in the ability to pinpoint the known biological processes in which a TF functions. In summary, we show the power and benefits of ProxReg, which is available in five species (fruit fly, zebrafish, mouse, rat, and human) for promoters and in human for enhancers, to complement GSE testing of large sets of peaks.

MATERIALS AND METHODS

Datasets Used

We used a total of 90 human ChIP-seq datasets from the Encyclopedia of DNA Elements (ENCODE) at University of California, Santa Cruz (ENCODE Project Consortium, 2004; Qu and Fang, 2013; Sloan et al., 2015) that consists of 35 TFs over the three Tier 1 cell lines [embryonic stem cells (H1-hESC), B-Lymphocyte (GM12878), and myelogenous leukemia cell (K562)] (**Supplementary Table 1**).

Gene sets tested were Gene Ontology: Biological Processes (GO BP) from *GO.db* Bioconductor package version 3.4.2 (The Gene Ontology Consortium, 2018). We filtered gene sets to only use those with more than 15 and less than 2000 genes, as small gene sets have very little statistical power and large gene sets tend to be too vague to have meaningful biological interpretation. This resulted in 5159 GO BP gene sets.

Measuring Peak Distances to Nearest Transcription Start Site or Enhancer Midpoint

Each peak's "regulatory proximity" was defined as the distance, in base pairs, between the peak's midpoint and either the closest TSS or the midpoint of the closest enhancer region. Human gene TSS locations were obtained from the *chipenrich* package, which for hg19 version 3.5.0 are from Bioconductor packages *TxDb.Hsapiens.UCSC.hg19.knowngene* version 3.2.2 (Carlson and Maintainer, 2015) and *org.Hs.eg.db* version 3.5.0 (Carlson, 2018). Enhancer regions were defined by the union of DNase hypersensitive sites (DNase DHSs) found in at least two of the 125 cell and tissue types processed by ENCODE (Thurman et al., 2012) and distal and non-promoter DHS within 500 kb of the correlated promoter DHSs from 32 cell types (Thurman et al., 2012). The minimum of two cell types was used to reduce false positives. Unions were calculated using the *expand_and_resect2* function in the *granges* R package with min.gapwidth = 0, and distal and non-promoter elements were defined as those >5 kb from a TSS. That is, we removed only the portion of an enhancer that was <5 kb from a TSS. This resulted in a total set of 1,616,520 regions >5 kb from a TSS composed of enhancers, silencers, and insulators, although for simplicity we refer to the total set as enhancers. Finally, all peaks are then assigned to the gene with the NTSS.

ProxReg Step 1: Normalizing for Gene Locus Length and Average Distance to Enhancer

Identical to our previous work, we define a gene's locus length (in bps) as the length of the region on the genome such that a peak binding in the region is assigned to that target gene (Cavalcante et al., 2014; Welch et al., 2014). As genes with larger locus lengths (i.e., longer distances to neighboring genes) are more likely to have peaks binding farther away from the gene's TSS, gene locus length is associated with average peak distance to TSS, and thus gene locus length is a potential confounding variable. To empirically normalize for gene locus length, we used the combined set of peaks from all 90 ENCODE ChIP-seq peak datasets and computed a cubic smoothing spline for log locus length (x-axis) vs. log peak distances (y-axis) using the *gam* function in the *mgcv* package. The spline provides the expected, global average binding distance for each gene, which we then used to obtain the normalized adjusted binding distance as:

$$D_{tss}^{adj} = \log D_{tss} - \log D_{spline}$$

Thus, peaks that are closer to a TSS than expected based on the spline fit will contribute to significant promoter proximity for a gene set.

Similar to how a gene with a longer locus length tends to have peaks farther from its TSS, gene loci with farther spaced enhancers tend to have peaks farther from them. More specifically, the distance to an enhancer region is associated with how far apart a gene's enhancers are spread, which is dependent on both the gene locus length and the number and distribution of enhancers associated with the locus region. Therefore, the average (or expected) enhancer density for each gene is a potentially confounding variable. To normalize for this, we first determined every gene's empirical average distance to an enhancer with our list of 90 ENCODE ChIP-seq datasets, and then calculated each peak's distance to the nearest enhancer, and finally averaged this distance for each individual gene. As these 90 experiments do not cover every gene, if a dataset happens to have a peak assigned to a gene not covered, the average distance to enhancer will be set as the predicted mean of a linear estimation using the log gene locus length of the known genes. Similar to the locus length normalization, we have the adjusted enhancer distance:

$$D_{enh}^{adj} = \log D_{enh} - \log AvgD_{enh}$$

Thus, peaks closer to an enhancer than expected by chance will contribute to significant enhancer proximity for a gene set.

ProxReg Step 2: Testing for Proximal Regulatory Binding

For a gene set of interest, the peaks assigned to genes in the gene set are placed in one group while all other peaks assigned to other genes, called the background genes, are placed in another. We let any gene that has the potential of a peak being assigned to it and annotated in the gene set database to be a background gene, which is equivalent to the procedure of gene expression tools such as DAVID (Da Wei Huang et al., 2007). The goal is to test whether the peaks in the gene set are significantly closer to TSSs (or enhancers) than expected by chance, given the adjusted distances described above. We use a two-sided Wilcoxon rank-sum test, with positive values denoting the distances in the gene set are *smaller* than those not in the gene set, to test if peaks in the gene set tend to be closer or farther from regulatory regions than those not in the gene set. To account for multiple testing, we use the Benjamini–Hochberg method to calculate FDR values (Benjamini and Hochberg, 1995).

Gene Set Enrichment Testing Using Poly-Enrich

We tested all 90 ENCODE datasets using the *polyenrich* method in the *chipenrich* Bioconductor package, using the "nearest_tss" gene locus definition and GO biological processes for the gene sets. Poly-Enrich performs GSE on sets of peaks by testing if the number of peaks regulating a gene set is greater or less than that not in the gene set, taking into account the number of peaks assigned to each gene (Lee et al., 2018). The statistical model uses a negative binomial *glm* with an adjustment for gene locus length. Significantly enriched gene sets have more peaks, while depleted ones have fewer.

Permutations to Assess Type I Error Rate

To test Type I error rate of the ProxReg method, we simulated a null set of peak distances (i.e., with no gene sets having significant proximal binding) in three ways: (1) by reassigning every peak to a random gene, where all genes are equally likely to be assigned (*Unif*). (2) To test for correct normalization of gene locus length, we randomized peaks to a gene as above, except genes were first binned with other genes of similar locus length as defined by their TSSs. Specifically, we ranked genes by locus length, binned them into sets of 100 genes, and then reassigned every peak to a random gene within the same bin (*ByLocusLength*). (3) To test the normalization of average distance to enhancers, we ranked genes by expected distance to enhancer by chance, and then binned genes into sets of 100. Again, we then reassigned every peak to a random gene within the same bin (*ByAvgDEnh*). We performed 10 randomizations per ChIP-seq experiment and chose α-levels of 0.05 and 0.001 to test for a controlled Type I error rate.

Simulations to Estimate Power

We simulated significant proximal gene sets by starting from a null set of peaks using the *ByLocusLength* permutation strategy. We then added peaks near the TSSs of genes from a gene set, with the choice of a small (471 genes) or a large (1717 genes) gene set. The number of peaks added was equal to 0.01, 0.05, or 0.1% of the total number of starting peaks (4839) in the null set. The distance was chosen from an exponential distribution with mean d_0, and an equal chance for upstream or downstream. We chose values of 100, 500, 1000 for d_0 to simulate scenarios of closer and farther binding. For each scenario, 200 simulated gene sets were ran.

Clustering for TF Regulatory Patterns

To investigate the regulatory patterns among all 90 ENCODE ChIP-seq data sets, we performed clustering to classify them. We first applied a p-value cut off (<0.001) for both ProxReg (promoter and enhancer) results and Poly-Enrich results. We counted the numbers of points (GO BP terms) in each of four regions, defined by the different colored regions shown in **Figure 3**, for both promoter and enhancer results in all 90 data sets. Then, a hierarchical clustering heat map was generated based on the log 2 value of counts from each region. The Euclidean distance metric was used with Ward's minimum variance method for clustering. In addition, we also calculated the Pearson correlation between ProxReg promoter results and enhancer results. Since we propose our method as a complementary method for GSE testing, only signed negative log p-values of significant GO terms (FDR < 0.05) from Poly-Enrich were used for correlation calculations.

Test for the Ability of ProxReg to Reduce False Positives From GSE Results

To test the ability of our method to reduce false positives from GSE results, we compared the results of ProxReg and Poly-Enrich together to Poly-Enrich alone, using gene sets for each TF that the TF is likely to regulate. Since no gold standard is available for this, we used the GO BP terms that our 35 TFs were assigned to in the human annotation Bioconductor package *org.Hs.eg.db* (Carlson, 2018). Motivation for this derives from the fact that TFs do not regulate random sets of genes, but rather a well-coordinated set of genes in order to fulfill a cellular biological goal. Indeed, it's been shown that genes in a GO biological process term tend to be regulated by a common TF (Allocco et al., 2004; Qian et al., 2005; Roider et al., 2008; O'Connor et al., 2016). The cellular biological goal is precisely what GO biological process terms aim to describe, as it is defined as "The larger processes, or 'biological programs' accomplished by multiple molecular activities" (The Gene Ontology Consortium, 2004), which for TFs in DNA binding. Based on these two facts, the TFs that are assigned to a GO biological process term relate closely to this biological process, and since the function of TFs is to regulate genes, it follows logically that TFs tend to regulate genes in the biological processes to which they belong. As an example, the NCBI Gene website, an authoritative source for the properties of genes, states in the main summary of E2F family genes that "the E2F family plays a crucial role in the control of cell cycle". This family includes members E2F1, E2F2, E2F3a, E2F3b, E2F4, E2F5, E2F6, E2F7, and E2F8. In each case, we can also find at NCBI Gene that these TFs are assigned to the GO BP terms related to cell cycle. To further validate our approach, we tested whether the TFs actually do tend to target the promoters of genes in their assigned GO terms. Indeed, we found a strong overall trend to targeting more genes in the assigned GO terms versus the non-assigned GO terms (**Supplementary Figure 1** and **Supplementary Table 2**). Although TFs may not regulate all of their target gene sets in every cell type, we conclude that the degree of overlap between a method's predictions and a TF's assigned GO BP terms represents a useful benchmarking tool.

To assess this, we first counted all significantly enriched gene sets from Poly-Enrich for all 90 ENCODE ChIP-seq data sets and found their overlap with the GO BP terms each TF was assigned to in *org.Hs.eg.db*. These GO terms were used to count significant GO terms from ProxReg promoter and enhancer results. Fisher's exact test was used to determine whether ProxReg further enriched the resulting GO terms to those assigned to by the TF, beyond what GSE testing accomplished. We used datasets for TFs that are assigned to at least five GO BP terms that were also significant with GSE testing alone. Fisher's exact test results demonstrated whether ProxReg was able to increase the odds ratio of identifying GO BP terms assigned to the TF, compared to GSE testing alone.

Website Implementation and Bioconductor Availability

Proximity Regulation is available in the *chipenrich* Bioconductor package with the *proxReg()* function, and at the ChIP-Enrich website[1], as an additional option following any of our current GSE tests. To run ProxReg, the user uploads a file of peaks, which can be in narrowPeak or BED format. They then select to test for proximity to either NTSS or enhancers. Currently, we have only implemented testing for enhancer proximity in human (hg19 genome), but others will be added as enhancers are sufficiently defined in other species and newer genome versions. Finally, the user selects what gene sets to test from any of our included gene set databases (including KEGG, Panther, MSigDB gene sets, and several others; details in *chipenrich* package and on website), or a user-generated set. An example of the *proxReg()* function outputs four files:

Opts: the options that the user input into the function.
Peaks: a peak-level summary showing the peak-to-gene assignment for each peak, as well as their distances to TSS or enhancer.
Results: the results of the proximity tests. Lists the tested gene sets along with their descriptions, the test effect, closer/farther status, *p*-value, and FDR. Also included is the list of Entrez gene IDs with contributing signal for each proximity test.
Qcplot: a histogram showing the distribution of peak distances.

All R code for recreating analysis and figures can be found at: https://github.com/sartorlab/proxReg. An example for the use of ProxReg can be found in the *chipenrich* Bioconductor vignette.

RESULTS

Overview of ProxReg Method

We developed a new method, ProxReg, to test the proximity of peaks to TSSs or enhancers in a gene set of interest. The motivation for our new method is illustrated in **Figure 1**. The goal is to test whether the enrichment of a GO term or pathway is driven by regulation via promoters or distal regions (i.e., enhancers). To accomplish this, we firstly measure the distances from the midpoints of the peaks to the nearest regulatory regions (either TSSs or enhancers), and assign each peak to its target gene according to the gene with the NTSS (Welch et al., 2014). Specifically, for each gene we defined its gene locus to be the region between the upstream and downstream midpoints of its TSS and the neighboring gene's TSSs. However, one cannot simply directly test whether the distances are smaller within a gene set versus other genes, due to potentially confounding variables that first need to be taken into account. Since a gene locus with a large length was observed to have farther peaks from its TSS on average (**Figure 1A**), we first normalize for the gene locus length before testing the proximity to TSSs (see section "Materials and Methods"). For enhancers, we observed that the distance to an enhancer was dependent on the average distance from each enhancer to peaks in a gene locus (**Figure 1B**). Thus, we normalized the raw peak to enhancer distances using the average enhancer density for each gene. Finally, a two-sided Wilcoxon rank sum test was used for testing the proximity of peaks in a gene set to TSSs (or enhancers) compared to peaks outside the gene set. Generally, this test would be performed on all of the enriched gene sets identified by a GSE test, to understand whether the enrichment of each gene set was due to regulatory activity near promoters or enhancers.

Recommended Workflow for ProxReg

To test our new method, 90 ENCODE ChIP-seq data sets (36 TFs in three Tier 1 cell lines) (ENCODE Project Consortium, 2004; Qu and Fang, 2013; Sloan et al., 2015) were used in this study. The recommended workflow for implementing our new method is summarized in **Figure 2**. We begin with a gene definition file containing gene locus definitions (provided by our software, or uploaded custom by the user) and a set of peaks of interest (provided by the user). The distance between the midpoint of peaks and NTSSs (or midpoint of enhancers) are measured and adjusted for all background genes. The ProxReg non-parametric test is ran for the chosen gene sets (e.g., GO). In parallel to this, a standard GSE test is performed using the same gene sets. In this article, we applied the *polyenrich* method for the GSE test (Lee et al., 2018), but others may be used. Result files contain the proximity results with test direction (enriched/depleted from GSE, and closer/farther from ProxReg), *p*-values and FDR values. Combined with the *p*-values from GSE, the gene set proximity and enrichment patterns can be easily visualized (**Figure 2**; see section "Results").

Controlled Type 1 Error Rate and Ability to Detect True Positive Results

We validated the Type 1 error rate (rate of false positives) of ProxReg using randomizations of real datasets to simulate null datasets with no significant proximities to TSSs or enhancers. We performed three types of permutations: the "*Unif*" permutation, which takes every peak and reassigns another gene to it with each gene having the same probability, the "*ByLocusLength*" permutation, which tests the effectiveness of the locus length normalization in the distance to TSS test, and the "*ByAvgDEnh*" permutation, which tests the effectiveness of the normalization to average distance to enhancer in the distance to enhancer test (see section "Materials and Methods" for details). For a p-value < 0.05 cutoff, we expect a Type I error rate of approximately 5%. For a p-value < 0.001 cutoff, we expect a Type I error rate of approximately 0.1%. Results indicate that for each permutation (*Unif* and *ByLocusLength* for TSS proximity tests, and *Unif* and *ByAvgDEnh* for enhancer proximity tests), the

[1] http://chip-enrich.med.umich.edu

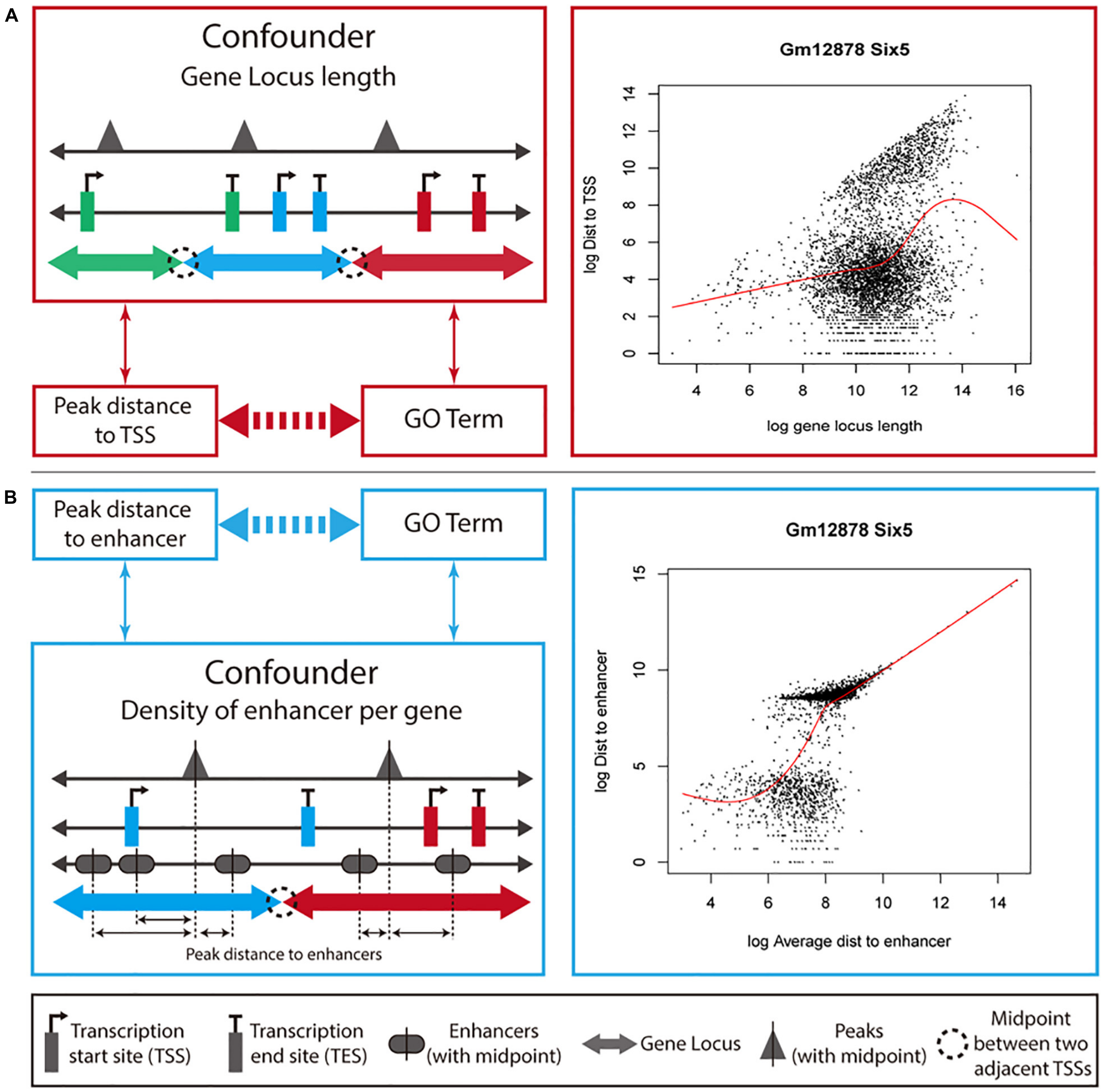

A
Confounder
Gene Locus length
Peak distance to TSS
GO Term
Gm12878 Six5
log Dist to TSS
0
2
4
6
8
10
12
14
4
6
8
10
12
14
16
log gene locus length
B
Peak distance to enhancer
GO Term
Confounder
Density of enhancer per gene
Peak distance to enhancers
Gm12878 Six5
log Dist to enhancer
0
5
10
15
4
6
8
10
12
14
log Average dist to enhancer
Transcription start site (TSS)
Transcription end site (TES)
Enhancers (with midpoint)
Gene Locus
Peaks (with midpoint)
Midpoint between two adjacent TSSs

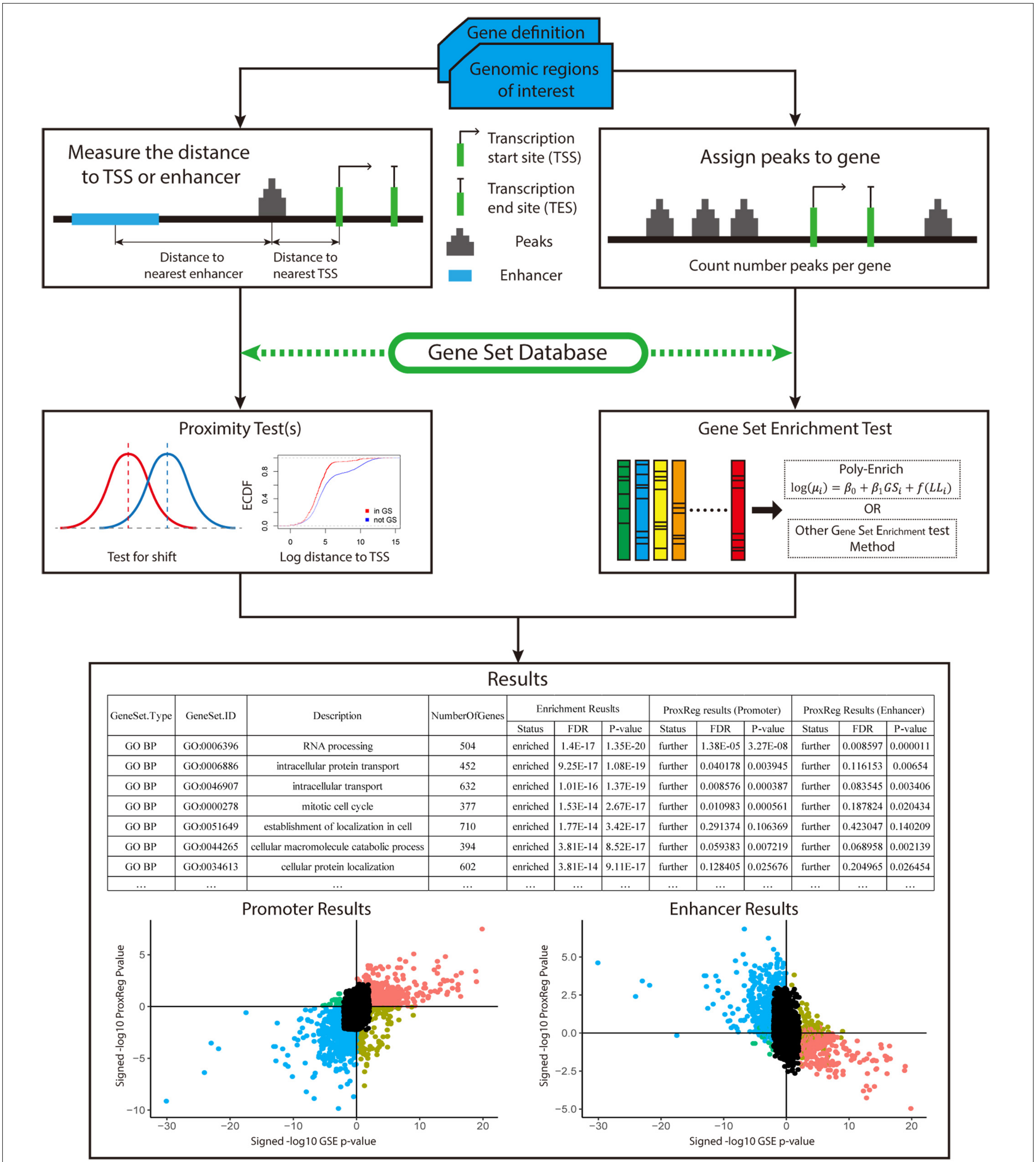
FIGURE 1 | Overview of how ProxReg adjusts for confounding variables. We describe the ProxReg adjustments in two parts. **(A)** When testing proximity to TSSs, we normalize the peak distances to TSSs according to their relationship with gene locus lengths. **(B)** When testing proximity to enhancer, we normalize the peak distances to enhancers according to their relationship with enhancer density, modeled by the average distance of any peak to an enhancer. In both cases, we avoid a potential confounding effect, as shown by the arrows between variables on the left-hand side.

Type 1 error rate is reasonably controlled at the expected level (**Supplementary Figure 2**).

To ensure that our method is able to identify gene sets with true cases of TSS or enhancer proximity, we generated artificial peak datasets starting with a randomized data set using the *ByLocusLength* permutation, and then adding peaks with TSS distances following a specified distribution. We added peaks by varying the number of peaks and the distance of peaks to assess a wide range of scenarios. We also used two gene sets of different sizes (see section "Materials and Methods" for details). We expected the following changes in parameters to increase power: a smaller gene set used (easier to influence average distance), more peaks added, and a smaller average distance. We can see that all three of these scenarios increased power to detect the true positive gene sets as expected (**Supplementary Figure 3**).

Integration of GSE and ProxReg Results Reveals Different Regulatory Patterns of TFs

We clustered the 90 ENCODE ChIP-seq datasets into three groups based on the hierarchical clustering heat map illustrated


Gene definition
Genomic regions of interest
Measure the distance to TSS or enhancer
Distance to nearest enhancer
Distance to nearest TSS
Transcription start site (TSS)
Transcription end site (TES)
Peaks
Enhancer
Assign peaks to gene
Count number peaks per gene
Gene Set Database
Proximity Test(s)
Test for shift
ECDF
0.0
0.4
0.8
in GS
not GS
0
5
10
15
Log distance to TSS
Gene Set Enrichment Test
Poly-Enrich
$\log(\mu_i) = \beta_0 + \beta_1 GS_i + f(LL_i)$
OR
Other Gene Set Enrichment test Method
Results
Promoter Results
Signed -log10 ProxReg Pvalue
5
0
−5
−10
−30
−20
−10
0
10
20
Signed -log10 GSE p-value
Enhancer Results
Signed -log10 ProxReg Pvalue
5.0
2.5
0.0
−2.5
−5.0
−30
−20
−10
0
10
20
Signed -log10 GSE p-value

GeneSet.Type	GeneSet.ID	Description	NumberOfGenes	Enrichment Reuslts			ProxReg results (Promoter)			ProxReg Results (Enhancer)		
				Status	FDR	P-value	Status	FDR	P-value	Status	FDR	P-value
GO BP	GO:0006396	RNA processing	504	enriched	1.4E-17	1.35E-20	further	1.38E-05	3.27E-08	further	0.008597	0.000011
GO BP	GO:0006886	intracellular protein transport	452	enriched	9.25E-17	1.08E-19	further	0.040178	0.003945	further	0.116153	0.00654
GO BP	GO:0046907	intracellular transport	632	enriched	1.01E-16	1.37E-19	further	0.008576	0.000387	further	0.083545	0.003406
GO BP	GO:0000278	mitotic cell cycle	377	enriched	1.53E-14	2.67E-17	further	0.010983	0.000561	further	0.187824	0.020434
GO BP	GO:0051649	establishment of localization in cell	710	enriched	1.77E-14	3.42E-17	further	0.291374	0.106369	further	0.423047	0.140209
GO BP	GO:0044265	cellular macromolecule catabolic process	394	enriched	3.81E-14	8.52E-17	further	0.059383	0.007219	further	0.068958	0.002139
GO BP	GO:0034613	cellular protein localization	602	enriched	3.81E-14	9.11E-17	further	0.128405	0.025676	further	0.204965	0.026454
...	...	...	...	...	...	...	...	...	...	...	...	...

FIGURE 2 | Overview of how of ProxReg fits in with the overall workflow of gene set enrichment testing with genomic regions. The peak distances to TSSs or enhancers are calculated for the proximity test. In parallel, all peaks are assigned to genes for gene set enrichment testing. The same gene set database is used for both proximity and gene set enrichment testing. Combining the gene set enrichment and proximity tests, the results can be visualized as shown in section "Results." The left scatter plot is an example of the combination of enrichment and promoter results. The right scatter plot is an example of enhancer results combined with enrichment results. The *x*-axis of these two scatter plots represent the gene set enrichment test result. A larger signed –log *p*-value indicates more enrichment, while negative values indicate depletion. The Y-axis represents the proximity results. Larger signed –log *p*-values indicate GO terms having genomic regions closer to the TSSs or enhancers.

in **Figure 3**. The first and largest group (47 datasets) is characterized by a strong positive correlation between GSE and promoter (TSS) ProxReg signed significance levels, and a strong negative correlation between GSE and enhancer ProxReg signed significance levels, indicating that the majority of enriched gene sets are due to binding near TSSs (**Figure 3** blue cluster; many genes in regions p1, p2, e3, and e4). TFs like SIX5 (SIX homeobox 5), SP1 (Specificity Protein 1*), and GABP (Nuclear Respiratory Factor 2) are included in this group. The second largest group (32 datasets) is more interesting because the datasets consist of some enriched gene sets with significant proximity to promoters, and other enriched gene sets with significant proximity to enhancers (**Figure 3** red cluster; genes spread out across mainly p1, p4, e1, and e4). The results for these TFs enable understanding the different regulatory mechanisms used for different biological processes. MEF2A (Myocyte-specific enhancer factor 2A) in K562 cells, a member of this group, was observed to regulate GTPase activity and translational initiation-related GO terms from TSSs, and transmission of nerve impulse and multicellular organismal signaling GO terms from enhancers. Similarly, P300 (Histone acetyltransferase p300), a well-known marker of enhancers, was found to regulate chromatin organization from TSSs, while regulating phosphatidylinositol dephosphorylation and phosphatidylinositol-mediated signaling-related GO terms from enhancers (Fryer et al., 2002; De Luca et al., 2003). The smallest group included only 11 ChIP-seq datasets. This group was characterized mainly by enriched gene sets with many having significant proximity to enhancer regions and/or far from promoters (**Figure 3** purple cluster; many genes in p4 and e1). Members of this group included Pol II in all three cell lines and EGR1 in K562 cells and Gm12878 cells, indicative of Pol II binding along entire gene lengths and not just at promoters. In addition, we examined the Pearson correlation between promoter results and enhancer results for all 90 ENCODE ChIP-seq data sets. Eighty-eight of them show a negative correlation between the promoter results and enhancer results (**Figure 4**). This negative correlation indicates that overall, GO terms are significantly enriched either by the TF binding closer to promoters or closer to enhancers. Among these 90 data sets, most of them (67 out of 88 data sets) show a strong negative correlation as shown in **Figure 4A**. Several of them have weak correlations as shown in **Figure 4B**. The two datasets that did not show negative correlations are neuron restrictive silencer factor (NRSF) and CMYC in H1-hESC cells. After removing non-significant GO BP terms from Poly-Enrich results, NRSF data set shows a weak positive correlation based on the remaining GO BP terms and no significant GO BP terms in CMYC data set.

ProxReg Identifies Known Associations With Promoter and Enhancer Binding, Using SIX5 and NRSF Peaks

To further illustrate our method, we assess ProxReg results for two TFs known to have a very strong tendency to bind either in proximal promoters or enhancers. We first selected SIX5 in GM12878 cells as an example, which is involved in determination and maintenance of retina formation that proposed binding to promoter regions of related genes (e.g., myogenin and IGFBP5) (Spitz et al., 1998; Sato et al., 2002). The results of SIX5 are shown in **Figure 5**.

In **Figure 5A**, we can see that the majority of the ChIP-seq peaks (67.4%) are near TSSs. Through the combination of ProxReg results and Poly-Enrich results, a great majority of gene sets are enriched by the TF binding near TSSs (positive correlation in **Figure 5B**) instead of near enhancers (negative correlation in **Figure 5C**). Using two particular GO terms from the scatter plots, we show the distribution of distances from peaks to TSSs or enhancers (bottom part of **Figures 4C**, **5B**). Combining the locations of these two GO terms (GS1 and GS2 in the scatter plots), illustrates how our method is able to provide additional information for interpreting GSE testing results.

We also selected NRSF in the K562 cells as an example. NRSF, also known as RE1-Silencing Transcription factor (REST), is a TF known to silence neuronal genes in non-neuronal cells, it can act as a transcriptional repressor or enhancer of target genes, often regulating from enhancer regions (Schoenherr and Anderson, 1995; Seth and Majzoub, 2001). Almost half of NRSF ChIP-seq peaks (51.2%) are far from TSSs (**Figure 5D**). A similar strategy was used for illustration of ProxReg with the transcriptional repressor NRSF in K562 cells. Consistent with previous observations that this TF tends to bind to silencers/enhancers instead of promoters, there is a relative strong positive correlation shown in the enhancer scatter plot (**Figure 5F**) but not for TSSs. Thus the results confirm that most enriched GO terms were enriched due to the TF binding in or near enhancer regions. These results validate our new method, ProxReg, is a powerful tool that can be used as a complementary approach for interpreting GSE test results.

ProxReg Enriches GSE Findings for Likely True Positives

We assessed whether ProxReg can be used to not only estimate the proximity effects but also help users to remove possible misleading or false positive gene sets from GSE results. To accomplish this, we compared the significantly enriched gene sets to a set of GO biological process (BP) terms from *org.Hs.eg.db* for each TF before versus after taking into account their ProxReg results. The GO BP terms from *org.Hs.eg.db* consists of the TFs and the assigned GO BP terms for the gene that encodes them (see section "Materials and Methods" for more detail).

We used ChIP-seq datasets with at least five significantly enriched GO BP terms in their *org.Hs.eg.db* set (to ensure sufficient power), which resulted in 28 datasets with ProxReg enhancer results and 36 datasets with ProxReg promoter results. We then tested whether requiring a significant ProxReg test resulted in a higher odds ratio of detecting the *TF-assigned* GO BP terms. Of the 28 enhancer dataset results, 18 (64%) had an odds ratio greater than 1. Among these, 11 (61%) of them were significant. Conversely, only two enhancers' results had an odds ratio significantly less than 1. These two results were from EGR1 and ATF3 in K562 cells. Previous research (Cullen et al., 2010) suggests that EGR1 recognizes and binds to promoter regions of target genes, so it is possible that the

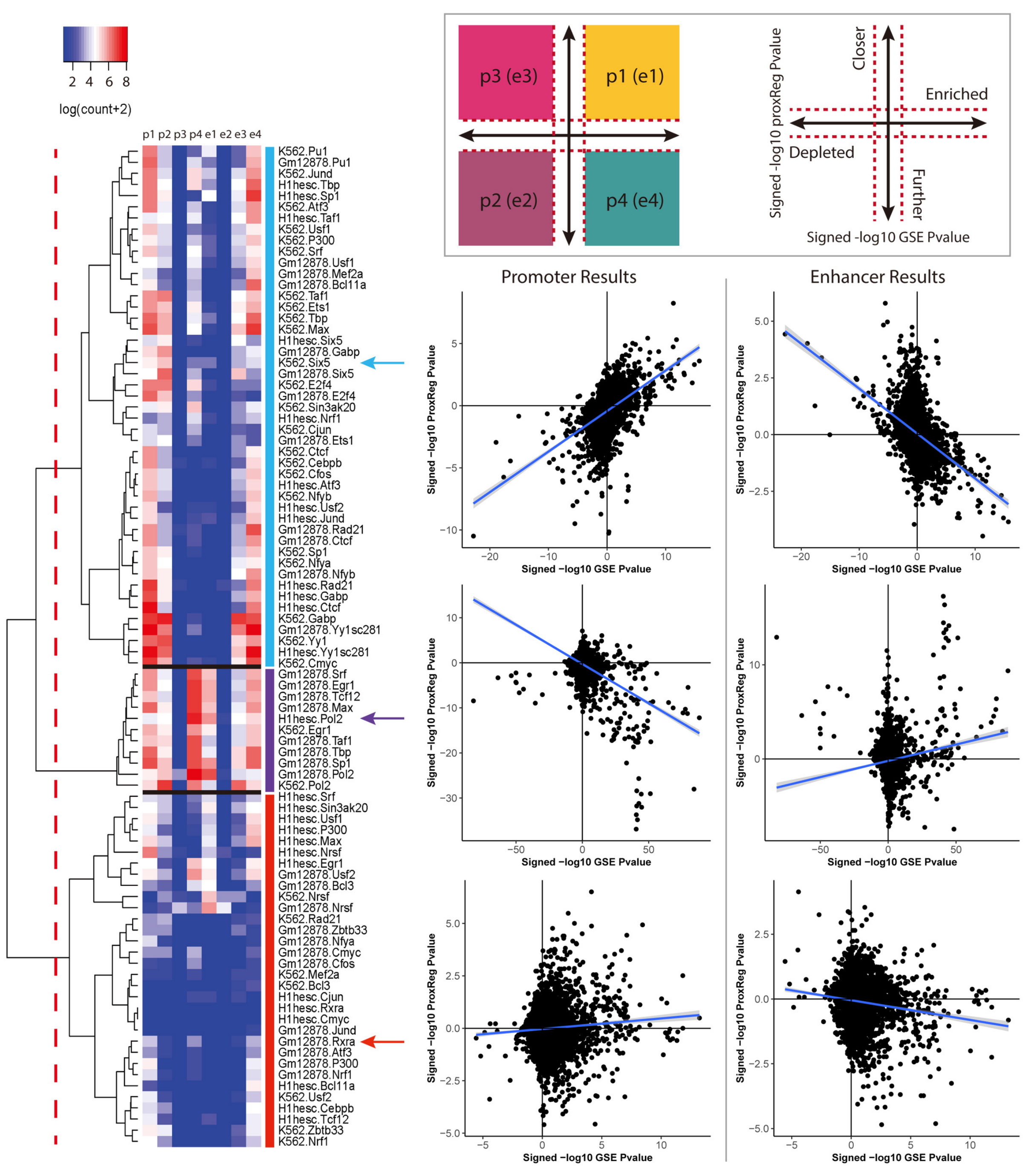


FIGURE 3 | The regulation patterns of the 90 ENCODE ChIP-seq datasets. A *p*-value cutoff (<0.001) was applied to define the four regions as illustrated in the top panel. The cutoffs are represented by the red dash lines. For each data set, the points count of the combination of ProxReg promoter results and Poly-Enrich results are labeled as p1, p2, p3, and p4. Similarly, the combination of enhancer results and Poly-Enrich results are labeled as e1, e2, e3, and e4. Based on our analyses, 47 data sets show a clear positive correlation in promoter results and a clear negative correlation in enhancer results. 32 datasets show no strong correlation in either promoter or enhancer results. The remaining 11 data sets show a clear positive correlation in the enhancer results. For each group, the promoter and enhancer results of one data set are illustrated as an example.

GO BP terms from *org.Hs.eg.db* we compared to is incomplete, with previous data mainly being focused on biological processes that EGR1 regulates from promoter regions. A similar case may be true for ATF3.

Among 36 ProxReg promoter results, 25 (69%) had an odds ratio greater than 1. Among these 25 results, 15 (60%) of them were significant. Conversely, only two promoter results were significant with an odds ratio smaller than 1. One of them was

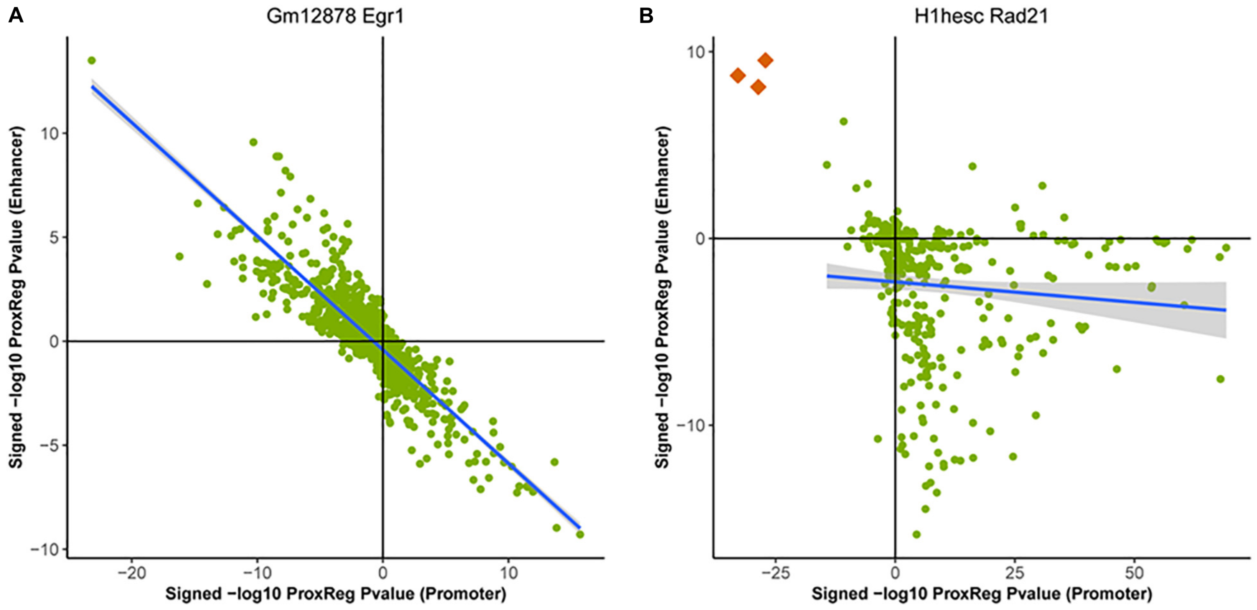


FIGURE 4 | Examples of the correlation between ProxReg promoter *p*-values and enhancer *p*-values. Majority of the 90 ENCODE ChIP-seq data sets show a strong negative correlation as shown in **(A)**. A small portion of these data sets show a pattern as shown in **(B)**. The three orange dots in **(B)** are GO terms related detection of chemical stimulus (GO:0050907, GO:0009593, and GO:0050911).

PU.1 in K562 cells. A previous study (Heinz et al., 2015) indicated that PU.1 usually binds to a PU-box found on enhancers of target genes, consistent with the ProxReg promoter results of PU.1 peaks having an odds ratio less than 1. Although we only found five significant GO BP terms from our results that are also assigned to PU.1, some other significant GO BP terms that we identified were biologically related to the remaining GO terms assigned to PU.1. For instance, some GO terms assigned to PU.1 were related to response to toxic substances, drugs, and antibiotics, and many immune response-related GO terms were significant. Overall, these results demonstrate that ProxReg can be used as a powerful supplemental method to remove misleading or false positive GSE test results (**Supplementary Table 3**), and provide additional evidence for novel regulated processes initially identified by GSE testing.

ProxReg Analysis Identified NRSF Regulatory Pattern Switching in Different Cell Types

The ProxReg results can guide and refine the biological interpretation of GSE results by identifying whether each enriched gene set is regulated mainly via binding close to promoters or enhancers. We exemplified this using the findings of NRSF, which was shown to regulate neuron development mostly via binding to enhancers in K562 cells (see details above). To further investigate the regulation patterns of NRSF in different cell lines, we utilized ENCODE NRSF ChIP-seq experiments from three cell types (GM12878, H1-hESC, and K562), and performed and integrated the Poly-Enrich and ProxReg analyses for each cell type. In GM12878, almost all significant GO terms identified by both Poly-Enrich and ProxReg were found to be closer to enhancers, except one GO term "establishment of localization in cell", which was significantly closer to promoters (FDR = 2.04 × 10^{-6}) and farther from enhancers (FDR = 9.60 × 10^{-7}) (**Figures 6A,B** and **Supplementary Table 4**). Most of them were related to neuron development, including "neurological system process," "regulation of nervous system development," and "synapse organization." In H1-hESC cells, however, NRSF binding sites were significantly enriched in GO terms which were significantly closer to promoters, and mostly related to neuron development and regulation, such as "synapse organization," "neuron projection guidance," and "neurotransmitter secretion" (**Figures 6A,C** and **Supplementary Table 5**). Less than 1% GO terms were closer to enhancers ("cell morphogenesis involved in differentiation," "regulation of cell projection organization," and "positive regulation of nervous system development"). The pattern observed in K562 was similar to that in GM12878: the majority of enriched GO terms were significantly closer to enhancers, and again most of them were related to neuron regulation (e.g., "axon guidance," "synapse maturation," and "regulation of synapse assembly") (**Figures 6A,D** and **Supplementary Table 6**), whereas only one was closer to promoters ("regulation of alternative mRNA splicing, via spliceosome"). These findings point to a fundamental shift in the binding patterns of NRSF to regulate neuronal genes during neuron development and organization processes: closer to promoters of genes in H1-hESC, while closer to enhancers in differentiated cells (GM12878 and K562). Taken together, we demonstrate that ProxReg analysis complements the GSE results by distinguishing where a TF binds to regulate genes, which is key to understanding the mechanisms of gene regulation and guiding

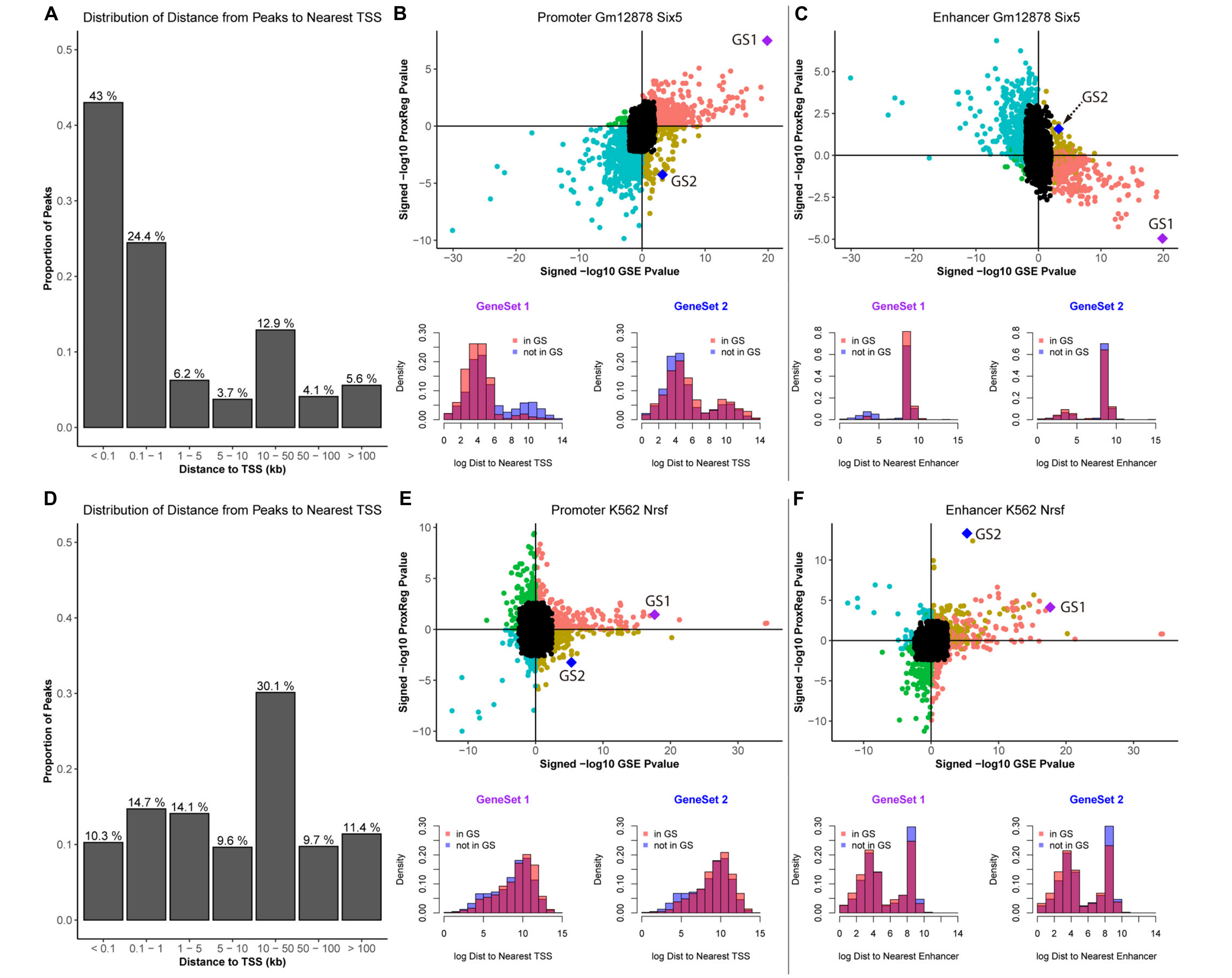


FIGURE 5 | Illustration of ProxReg results. The results of SIX5 in GM12878 cell lines are shown in **(A–C)**. **(A)** The distribution of distances from peaks to nearest TSSs. **(B)** Scatter plot of the combination of enrichment results and promoter results. Two gene sets were selected to show the distance distribution to nearest TSSs for genes in the gene set and not in the gene set. **(C)** Enhancer results combined with the enrichment results. The same gene sets were used in this scatter plot. The distribution of distances to the nearest enhancers of these two gene sets are shown in the bottom of **(C)**. Similar to SIX5 results, **(D–F)** show the results of NRSF in K562 cells. For SIX5, GeneSet 1: RNA processing. GeneSet 2: Positive regulation of nitrogen compound metabolic process. For, NRSF, GeneSet 1: Neuron differentiation. GeneSet 2: System process.

potential targeted gene therapy. ProxReg is incorporated in the *chipenrich* Bioconductor package and ChIP-Enrich website, and can be used with many additional databases of gene sets.

DISCUSSION

We introduced a genomic region proximity test called ProxReg that can be used as a complement for GSE tests, and can be used with various types of genomic regions, including ChIP-seq, ATAC-seq, GWAS SNPs, DNA methylation, and repetitive element families. The standard GSE tests for sets of genomic regions (e.g., ChIP-seq peak sets) usually only consider the relationship between the genomic regions and TSSs (McLean et al., 2010). However, it is of great interest to know whether a gene set is significantly enriched through regulatory activity near promoters or enhancers. Our new method, ProxReg, is able to find gene sets with regions that bind significantly closer to (or farther from) either promoters or enhancers. Furthermore, we validated that it has an appropriate Type I error rate, and that the statistical power of the test behaves as expected when varying the relevant variables. ProxReg uses a two-sided Wilcoxon rank-sum test for the proximity test while adjusting for important confounding variables. On its own, it provides insight into particular regulatory patterns. Integrated with GSE testing, it serves as a powerful complementary approach to enhance understanding of regulatory behavior across cell types, time points, disease stages, and more.

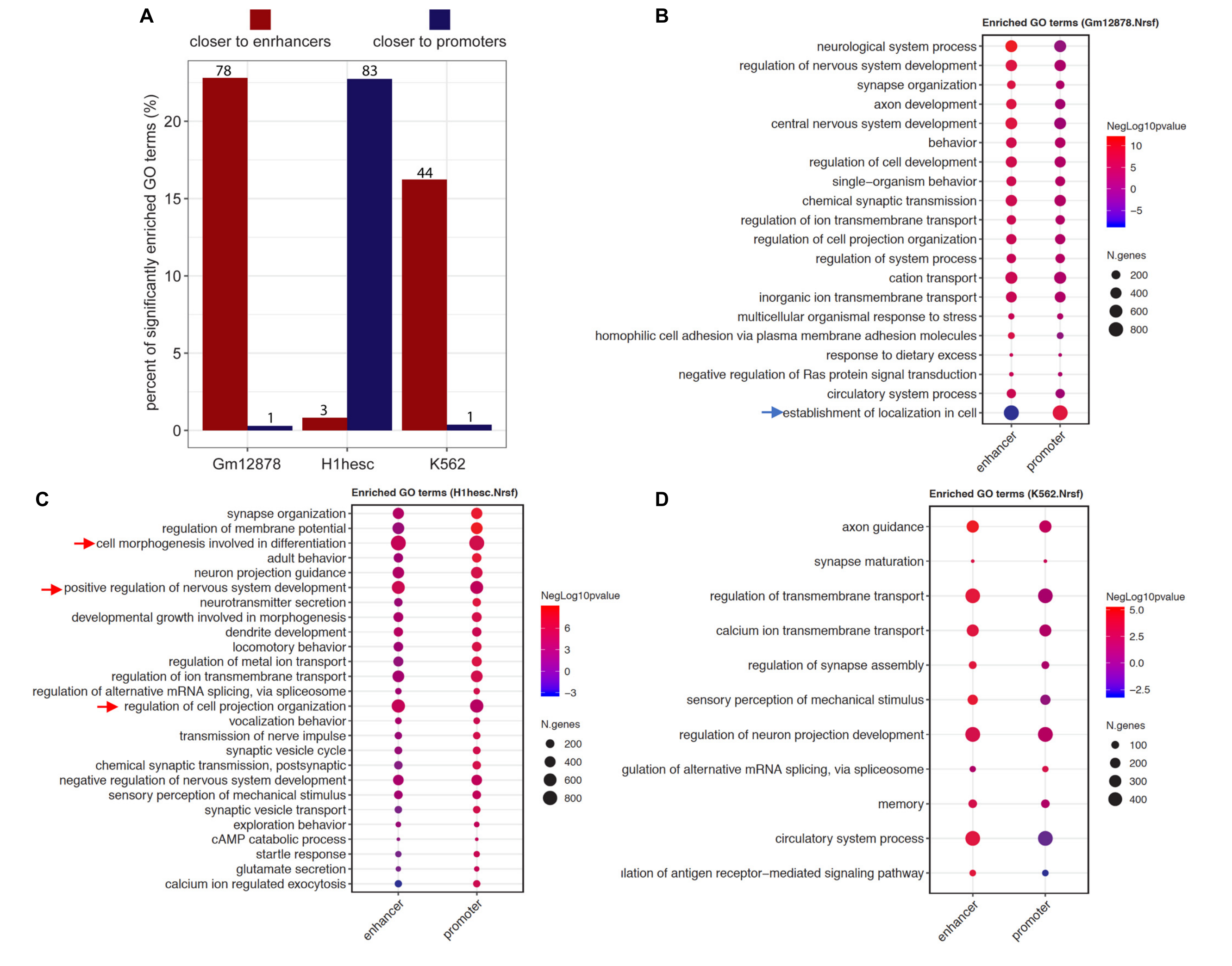


FIGURE 6 | The different regulatory patterns of NRSF in three cell lines. **(A)** The bar plots show the percentage of significantly enriched GO terms that were closer to enhancer (dark red) or promoter (dark blue) in each cell line (x-axis). The numbers of terms were marked on the top of each bar. **(B–D)** The dots represent the ProxReg enhancer or promoter significance levels (signed negative log *p*-values, resulting in positive values for proximal regions, and negative values for more distal regions) of the enriched GO terms in GM12878 **(B)**, H1-hESC **(C)**, and K562 **(D)** cell lines. In a particular cell line, the arrows point to the GO terms closer to promoters (blue arrows) while most of the terms are closer to enhancers, or point to the GO terms closer to enhancers (red arrows) while most of the terms are closer to promoters. For visualization, the redundant GO terms were removed from the list (Koneva et al., 2018).

When performing pathway analyses with current tools, the method may detect significance from regulation coming from different regions, but the underlying details are often left unknown. Standard GSE tests either do not take proximity to regulatory regions into account, or embed the proximity to TSSs within the test, still ignoring enhancers. In this way, it is difficult to interpret the results without the proximity effects. For example, when GREAT or Poly-Enrich finds a significant gene set from a ChIP-seq experiment, it is known that the gene set is enriched with peaks compared to genes not in the gene set, but we do not know if the peaks reside in promoter or enhancer regions any more than expected by chance. ProxReg is able to further show if the binding sites are closer to (or farther from) TSSs or enhancers, giving more insight into a TF's binding tendencies. We showed with real world ChIP-seq datasets from ENCODE that ProxReg was able to identify tendencies of TFs known to most often bind in proximal promoter regions (SIX5) (Spitz et al., 1998; Sato et al., 2002) or distal regions (NRSF) (Schoenherr and Anderson, 1995; Seth and Majzoub, 2001). Additionally, significantly enriched gene sets that were not found to be significant by ProxReg may have resulted from distal peaks being misassigned to incorrect target genes.

To illustrate the usefulness of ProxReg, we performed GSE and ProxReg testing on three ChIP-seq datasets of the TF NRSF in embryonic stem cells (H1-hESC) and two differentiated cell lines (K562 and GM12878). We showed how NRSF tends to regulate certain neuronal-related gene sets in differentiated cells by binding closer to enhancer regions, while regulating similar gene sets via binding to promoters in embryonic stem cells. Furthermore, we identified other non-neuronal GO terms that

NRSF regulates via binding mainly in promoter (or enhancer) regions. It is interesting to note that the enhancer binding, which is more cell-type specific and generally evolved later than regulation from promoters (Nord et al., 2013; Cai et al., 2019), was identified for the complex neuron development and related terms, while more basic processes such an establishment of location in cell and mRNA splicing, were regulated from closer to TSSs. Only in embryonic stem cells was even the neuronal-related terms regulated via promoters.

Proximity Regulation does have multiple limitations. Currently, we have implemented distance to enhancers for human (hg19 and hg38), and are planning to soon provide support for mouse (mm9 and mm10) (Haeussler et al., 2018). Since the enhancer landscape for other organisms lags the comprehensiveness of that for humans and mice, we currently only offer the promoter proximity test for other species. As other organisms' enhancer locations become more accurately defined, we plan to add support for more enhancer proximity tests.

An ongoing question is the identity of the targeted genes of enhancers binding events (Rubtsov et al., 2006; Sanyal et al., 2012; Melamed et al., 2016), which remains challenging due to long-range chromosome interactions. By analyzing TFs that tend to bind far from TSSs, we found that there are gene sets that tend to be regulated by TFs binding significantly farther from gene TSSs while also binding closer to enhancer locations. However, ProxReg assumes that each peak is associated with the gene with the NTSS, whereas this is often not true. It has been estimated that 79–95% of TF binding actually regulates a gene interceded by one or more other genes (Van Heyningen and Bickmore, 2013; Aldrup-Macdonald and Sullivan, 2014; de Sotero-Caio et al., 2017). Additionally, we used one general set of enhancer locations across the entire genome, whereas in reality, this method may benefit from allowing different tissues to have different sets of defined enhancer locations. Further research is required to understand how the comprehensiveness of the enhancer database affects the results of ProxReg, as well as of GSE tests. We are currently undergoing research on the differences in enhancer locations and their target genes in relation to GSE testing.

DATA AVAILABILITY STATEMENT

We used a total of 90 human ChIP-seq datasets from the Encyclopedia of DNA Elements (ENCODE) at University of California, Santa Cruz (details found in **Supplementary Table 1**).

AUTHOR CONTRIBUTIONS

MS conceived of the study. CL, KW, and TQ carried out the analysis. CL, KW, TQ, and MS wrote the manuscript.

FUNDING

This work is supported by the Michigan Lifestage Environmental Exposures and Disease (M-LEEaD) NIEHS Core Center grant (P30 ES017885), and the Michigan Rogel Cancer Center Support Grant (P30 CA046592) by NCI.

ACKNOWLEDGMENTS

We thank Dr. Raymond Cavalcante for help with data preparation and Bioconductor package updates.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.00199/full#supplementary-material

REFERENCES

Aldrup-Macdonald, M. E., and Sullivan, B. A. (2014). The past, present, and future of human centromere genomics. *Genes* 5, 33–50. doi: 10.3390/genes5010033

Allocco, D. J., Kohane, I. S., and Butte, A. J. (2004). Quantifying the relationship between co-expression, co-regulation and gene function. *BMC Bioinformatics* 5:18. doi: 10.1186/1471-2105-5-18

Andersson, R. (2015). Promoter or enhancer, what's the difference? Deconstruction of established distinctions and presentation of a unifying model. *Bioessays* 37, 314–323. doi: 10.1002/bies.201400162

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

Berger, S. L. (2007). The complex language of chromatin regulation during transcription. *Nature* 447, 407–412. doi: 10.1038/nature05915

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.

Cai, W., Huang, J., Zhu, Q., Li, B. E., Seruggia, D., Zhou, P., et al. (2019). Enhancer-dependence of gene expression increases with developmental age. *bioRxiv [Preprint]*

Carlson, M. (2018). *org.Hs.eg.db: Genome Wide Annotation for Human. R package version 3.4.1.*

Carlson, M., and Maintainer, B. (2015). *"TxDb. Hsapiens. UCSC. hg19. Knowngene: Annotation package for TxDb Object (s)," in R package version 3, 0, 0.*

Cavalcante, R. G., Lee, C., Welch, R. P., Patil, S., Weymouth, T., Scott, L. J., et al. (2014). Broad-enrich: functional interpretation of large sets of broad genomic regions. *Bioinformatics* 30, i393–i400. doi: 10.1093/bioinformatics/btu444

Chen, J., Hu, Z., Phatak, M., Reichard, J., Freudenberg, J. M., Sivaganesan, S., et al. (2013). Genome-wide signatures of transcription factor activity: connecting transcription factors, disease, and small molecules. *PLoS Comput. Biol.* 9:e1003198. doi: 10.1371/journal.pcbi.1003198

ENCODE Project Consortium, (2004). The ENCODE (encyclopedia of DNA elements) project. *Science* 306, 636–640. doi: 10.1126/science.1105136

The Gene Ontology Consortium, (2004). The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res.* 32, D258–D261.

The Gene Ontology Consortium, (2018). The gene ontology resource: 20 years and still going strong. *Nucleic Acids Res.* 47, D330–D338. doi: 10.1093/nar/gky1055

Cullen, E. M., Brazil, J. C., and O'connor, C. M. (2010). Mature human neutrophils constitutively express the transcription factor EGR-1. *Mol. Immunol.* 47, 1701–1709. doi: 10.1016/j.molimm.2010.03.003

Da Wei Huang, B. T. S., Tan, Q., Collins, J. R., Alvord, W. G., Roayaei, J., Stephens, R., et al. (2007). The DAVID gene functional classification tool: a novel biological module-centric algorithm to functionally analyze large gene lists. *Genome Biol.* 8:R183.

De Luca, A., Severino, A., De Paolis, P., Cottone, G., De Luca, L., De Falco, M., et al. (2003). p300/cAMP-response-element-binding-protein ('CREB')-binding protein (CBP) modulates co-operation between myocyte enhancer factor 2A (MEF2A) and thyroid hormone receptor-retinoid X receptor. *Biochem. J.* 369, 477–484. doi: 10.1042/bj20020057

de Sotero-Caio, C. G., Cabral-De-Mello, D. C., Calixto, M. D. S., Valente, G. T., Martins, C., Loreto, V., et al. (2017). Centromeric enrichment of LINE-1 retrotransposons and its significance for the chromosome evolution of Phyllostomid bats. *Chromosome Res.* 25, 313–325. doi: 10.1007/s10577-017-9565-9

Deaton, A. M., and Bird, A. (2011). CpG islands and the regulation of transcription. *Genes Dev.* 25, 1010–1022. doi: 10.1101/gad.2037511

Ertosun, M. G., Hapil, F. Z., and Osman Nidai, O. (2016). E2F1 transcription factor and its impact on growth factor and cytokine signaling. *Cytokine Growth Factor Rev.* 31, 17–25. doi: 10.1016/j.cytogfr.2016.02.001

Fryer, C. J., Lamar, E., Turbachova, I., Kintner, C., and Jones, K. A. (2002). Mastermind mediates chromatin-specific transcription and turnover of the Notch enhancer complex. *Genes Dev.* 16, 1397–1411. doi: 10.1101/gad.991602

Gertz, J., Reddy, T. E., Varley, K. E., Garabedian, M. J., and Myers, R. M. (2012). Genistein and bisphenol A exposure cause estrogen receptor 1 to bind thousands of sites in a cell type-specific manner. *Genome Res.* 22, 2153–2162. doi: 10.1101/gr.135681.111

Haeussler, M., Zweig, A. S., Tyner, C., Speir, M. L., Rosenbloom, K. R., Raney, B. J., et al. (2018). The UCSC genome browser database: 2019 update. *Nucleic Acids Res.* 47, D853–D858.

Heinz, S., Romanoski, C. E., Benner, C., and Glass, C. K. (2015). The selection and function of cell type-specific enhancers. *Nat. Rev. Mol. Cell Biol.* 16, 144–154. doi: 10.1038/nrm3949

Koneva, L. A., Zhang, Y., Virani, S., Hall, P. B., Mchugh, J. B., Chepeha, D. B., et al. (2018). HPV integration in HNSCC correlates with survival outcomes, immune response signatures, and candidate drivers. *Mol. Cancer Res.* 16, 90–102. doi: 10.1158/1541-7786.MCR-17-0153

Lambert, S. A., Jolma, A., Campitelli, L. F., Das, P. K., Yin, Y., Albu, M., et al. (2018). The human transcription factors. *Cell* 175, 598–599.

Lee, C. T., Cavalcante, R. G., Lee, C., Qin, T., Patil, S., Wang, S., et al. (2018). Poly-enrich: count-based methods for gene set enrichment testing with genomic regions and updates to ChIP-enrich. *bioRxiv [Preprint]*

Li, S., Wan, C., Zheng, R., Fan, J., Dong, X., Meyer, C. A., et al. (2019). Cistrome-GO: a web server for functional enrichment analysis of transcription factor ChIP-seq peaks. *Nucleic Acids Res.* 47, W206–W211. doi: 10.1093/nar/gkz332

Li, W. B., Notani, D., and Rosenfeld, M. G. (2016). Enhancers as non-coding RNA transcription units: recent insights and future perspectives. *Nat. Rev. Genet.* 17, 207–223. doi: 10.1038/nrg.2016.4

McLean, C. Y., Bristor, D., Hiller, M., Clarke, S. L., Schaar, B. T., Lowe, C. B., et al. (2010). GREAT improves functional interpretation of cis-regulatory regions. *Nat. Biotechnol.* 28, 495–501. doi: 10.1038/nbt.1630

Melamed, P., Yosefzon, Y., Rudnizky, S., and Pnueli, L. (2016). Transcriptional enhancers: transcription, function and flexibility. *Transcription* 7, 26–31. doi: 10.1080/21541264.2015.1128517

Nguyen, T. A., Jones, R. D., Snavely, A. R., Pfenning, A. R., Kirchner, R., Hemberg, M., et al. (2016). High-throughput functional comparison of promoter and enhancer activities. *Genome Res.* 26, 1023–1033. doi: 10.1101/gr.204834.116

Nord, A. S., Blow, M. J., Attanasio, C., Akiyama, J. A., Holt, A., Hosseini, R., et al. (2013). Rapid and pervasive changes in genome-wide enhancer usage during mammalian development. *Cell* 155, 1521–1531. doi: 10.1016/j.cell.2013.11.033

O'Connor, T., Bodén, M., and Bailey, T. L. (2016). CisMapper: predicting regulatory interactions from transcription factor ChIP-seq data. *Nucleic Acids Res.* 45:e19. doi: 10.1093/nar/gkw956

Pinsach-Abuin, M. L., Mates, J., Del Olmo, B., Allegue, C., Brugada, R., Garcia-Bassets, I., et al. (2016). Regulome-seq: searching for single nucleotide variants (SNVs) associated with disease beyond protein-coding regions. *FASEB J.* 30:1180.4.

Pristera, A., Lin, W., Kaufmann, A. K., Brimblecombe, K. R., Threlfell, S., Dodson, P. D., et al. (2015). Transcription factors FOXA1 and FOXA2 maintain dopaminergic neuronal properties and control feeding behavior in adult mice. *Proc. Natl. Acad. Sci. U.S.A.* 112, E4929–E4938. doi: 10.1073/pnas.150391112

Qian, J., Esumi, N., Chen, Y., Wang, Q., Chowers, I., and Zack, D. J. (2005). Identification of regulatory targets of tissue-specific transcription factors: application to retina-specific gene regulation. *Nucleic Acids Res.* 33, 3479–3491. doi: 10.1093/nar/gki658

Qu, H., and Fang, X. (2013). A brief review on the human encyclopedia of DNA elements (ENCODE) project. *Genomics Proteomics Bioinformatics* 11, 135–141. doi: 10.1016/j.gpb.2013.05.001

Roider, H. G., Manke, T., O'keeffe, S., Vingron, M., and Haas, S. A. (2008). PASTAA: identifying transcription factors associated with sets of co-regulated genes. *Bioinformatics* 25, 435–442. doi: 10.1093/bioinformatics/btn627

Roundtree, I. A., Evans, M. E., Pan, T., and He, C. (2017). Dynamic RNA modifications in gene expression regulation. *Cell* 169, 1187–1200. doi: 10.1016/j.cell.2017.05.045

Rubtsov, M. A., Polikanov, Y. S., Bondarenko, V. A., Wang, Y.-H., and Studitsky, V. M. (2006). Chromatin structure can strongly facilitate enhancer action over a distance. *Proc. Natl. Acad. Sci. U.S.A.* 103, 17690–17695. doi: 10.1073/pnas.0603819103

Sanyal, A., Lajoie, B. R., Jain, G., and Dekker, J. (2012). The long-range interaction landscape of gene promoters. *Nature* 489, 109–113. doi: 10.1038/nature11279

Sato, S., Nakamura, M., Cho, D. H., Tapscott, S. J., Ozaki, H., and Kawakami, K. (2002). Identification of transcriptional targets for Six5: implication for the pathogenesis of myotonic dystrophy type 1. *Hum. Mol. Genet.* 11, 1045–1058. doi: 10.1093/hmg/11.9.1045

Schmidt, D., Wilson, M. D., Spyrou, C., Brown, G. D., Hadfield, J., and Odom, D. T. (2009). ChIP-seq: using high-throughput sequencing to discover protein–DNA interactions. *Methods* 48, 240–248. doi: 10.1016/j.ymeth.2009.03.001

Schoenfelder, S., and Fraser, P. (2019). Long-range enhancer-promoter contacts in gene expression control. *Nat. Rev. Genet.* 20, 437–455. doi: 10.1038/s41576-019-0128-0

Schoenherr, C. J., and Anderson, D. J. (1995). The neuron-restrictive silencer factor (NRSF): a coordinate repressor of multiple neuron-specific genes. *Science* 267, 1360–1363. doi: 10.1126/science.7871435

Seth, K. A., and Majzoub, J. A. (2001). Repressor element silencing transcription factor/neuron-restrictive silencing factor (REST/NRSF) can act as an enhancer as well as a repressor of corticotropin-releasing hormone gene transcription. *J. Biol. Chem.* 276, 13917–13923. doi: 10.1074/jbc.m007745200

Shlyueva, D., Stampfel, G., and Stark, A. (2014). Transcriptional enhancers: from properties to genome-wide predictions. *Nat. Rev. Genet.* 15, 272–286. doi: 10.1038/nrg3682

Sloan, C. A., Chan, E. T., Davidson, J. M., Malladi, V. S., Strattan, J. S., Hitz, B. C., et al. (2015). ENCODE data at the ENCODE portal. *Nucleic Acids Res.* 44, D726–D732. doi: 10.1093/nar/gkv1160

Sonenberg, N., and Hinnebusch, A. G. (2009). Regulation of translation initiation in eukaryotes: mechanisms and biological targets. *Cell* 136, 731–745. doi: 10.1016/j.cell.2009.01.042

Spitz, F., Demignon, J., Porteu, A., Kahn, A., Concordet, J.-P., Daegelen, D., et al. (1998). Expression of myogenin during embryogenesis is controlled by Six/sine oculis homeoproteins through a conserved MEF3 binding site. *Proc. Natl. Acad. Sci. U.S.A.* 95, 14220–14225. doi: 10.1073/pnas.95.24.14220

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., et al. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U.S.A.* 102, 15545–15550. doi: 10.1073/pnas.0506580102

Thurman, R. E., Rynes, E., Humbert, R., Vierstra, J., Maurano, M. T., Haugen, E., et al. (2012). The accessible chromatin landscape of the human genome. *Nature* 489, 75–82.

Van Heyningen, V., and Bickmore, W. (2013). Regulation from a distance: long-range control of gene expression in development and disease. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 368:20120372.

Wang, B., Cunningham, J. M., and Yang, X. (2015). Seq2pathway: an R/Bioconductor package for pathway analysis of next-generation sequencing data. *Bioinformatics* 31, 3043–3045. doi: 10.1093/bioinformatics/btv289

Welch, R. P., Lee, C., Imbriano, P. M., Patil, S., Weymouth, T. E., Smith, R. A., et al. (2014). ChIP-Enrich: gene set enrichment testing for ChIP-seq data. *Nucleic Acids Res.* 42:e105. doi: 10.1093/nar/gku463

Wittkopp, P. J., and Kalay, G. (2011). Cis-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence. *Nat. Rev. Genet.* 13, 59–69. doi: 10.1038/nrg3095

Xie, C., and Tammi, M. T. (2009). CNV-seq, a new method to detect copy number variation using high-throughput sequencing. *BMC Bioinformatics* 10:80. doi: 10.1186/1471-2105-10-80

Xu, T., Jin, P., and Qin, Z. S. (2019). Regulatory annotation of genomic intervals based on tissue-specific expression QTLs. *Bioinformatics* 36, 690–697.

Yu, G., Wang, L.-G., and He, Q.-Y. (2015). ChIPseeker: an R/Bioconductor package for ChIP peak annotation, comparison and visualization. *Bioinformatics* 31, 2382–2383. doi: 10.1093/bioinformatics/btv145

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.

frontiers in Genetics

ORIGINAL RESEARCH
published: 10 March 2020
doi: 10.3389/fgene.2020.00158

TADCompare: An R Package for Differential and Temporal Analysis of Topologically Associated Domains

Kellen G. Cresswell† and Mikhail G. Dozmorov*†

Department of Biostatistics, Virginia Commonwealth University, Richmond, VA, United States

OPEN ACCESS

Edited by:
Geir Kjetil Sandve,
University of Oslo, Norway

Reviewed by:
Thomas Sexton,
INSERM U964 Institut de Génétique et de Biologie Moléculaire et Cellulaire (IGBMC), France
Ivan Merelli,
Institute of Biomedical Technologies, Italian National Research Council, Italy
Maxim V. Imakaev,
Massachusetts Institute of Technology, United States

***Correspondence:**
Mikhail G. Dozmorov
mikhail.dozmorov@vcuhealth.org

†These authors have contributed equally to this work

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Genetics

Received: 08 September 2019
Accepted: 11 February 2020
Published: 10 March 2020

Citation:
Cresswell KG and Dozmorov MG (2020) TADCompare: An R Package for Differential and Temporal Analysis of Topologically Associated Domains. Front. Genet. 11:158. doi: 10.3389/fgene.2020.00158

Recent research using chromatin conformation capture technologies, such as Hi-C, has demonstrated the importance of topologically associated domains (TADs) and smaller chromatin loops, collectively referred hereafter as "interacting domains." Many such domains change during development or disease, and exhibit cell- and condition-specific differences. Quantification of the dynamic behavior of interacting domains will help to better understand genome regulation. Methods for comparing interacting domains between cells and conditions are highly limited. We developed TADCompare, a method for differential analysis of boundaries of interacting domains between two or more Hi-C datasets. TADCompare is based on a spectral clustering-derived measure called the eigenvector gap, which enables a loci-by-loci comparison of boundary differences. Using this measure, we introduce methods for identifying differential and consensus boundaries of interacting domains and tracking boundary changes over time. We further propose a novel framework for the systematic classification of boundary changes. Colocalization- and gene enrichment analysis of different types of boundary changes demonstrated distinct biological functionality associated with them. TADCompare is available on https://github.com/dozmorovlab/TADCompare and Bioconductor (submitted).

Keywords: Hi-C, chromosome conformation capture, topologically associated domains (TADs), differential analysis, TADCompare

1. INTRODUCTION

Recent research indisputably proves the importance of the three-dimensional (3D) genome organization in regulating gene expression and other genomic processes (Osborne et al., 2004; Schoenfelder et al., 2010a,b; Tanizawa et al., 2010; Steensel, 2011; Li et al., 2012; Papantonis and Cook, 2013; Shavit and Lio, 2014; Symmons et al., 2014; Mifsud et al., 2015; Sexton and Cavalli, 2015; Franke et al., 2016; Mora et al., 2016). The 3D genomic structures consists of chromosome territories (Cremer and Cremer, 2010), A/B compartments corresponding to active/repressed chromatin (Lieberman-Aiden et al., 2009; Rao et al., 2014), topologically associated domains (TADs) (Jackson and Pombo, 1998; Ma et al., 1998; Dekker et al., 2002; Dixon et al., 2012; Nora et al., 2012; Sexton et al., 2012; Bonev et al., 2017), smaller sub-TADs (Phillips-Cremins and Corces, 2013; Rao et al., 2014) and chromatin loops (Dowen et al., 2014; Rao et al., 2014; Denker and Laat, 2016; Ji et al., 2016). These structures help to regulate global gene expression (de Laat and Grosveld, 2003; Osborne et al., 2004; Schoenfelder et al., 2010a,b; Tanizawa et al., 2010; Steensel, 2011; Li et al., 2012; Papantonis and Cook, 2013; Shavit and Lio, 2014; Symmons et al., 2014; Mifsud et al., 2015; Sexton and Cavalli, 2015; Franke et al., 2016; Mora et al., 2016). Consequently, coordinated changes in the 3D structures (Yaffe and Tanay, 2011; Dai and Dai, 2012; Symmons et al., 2014) determine cell type-specific gene expression and identity (Schoenfelder et al., 2010b; Dekker et al., 2013; Jin et al., 2013; Phillips-Cremins and Corces, 2013; Dowen et al., 2014; Rao et al., 2014; Vietri Rudan et al., 2015; Ji et al., 2016), guide recombination (Jhunjhunwala et al., 2009), X chromosome inactivation (Nora et al., 2012; Crane et al., 2015). Many 3D structures are largely invariant between different cell types, and even conserved between mammalian species (Dixon et al., 2012; Nora et al., 2012; Naumova et al., 2013; Pope et al., 2014; Rao et al., 2014; Vietri Rudan et al., 2015), indicating their high biological importance during genome evolution.

Despite the high level of conservation, recent research uncovered the dynamic nature of the 3D genomic structures, and this plasticity accompanies various biological functions and phenomena (Yu and Ren, 2017). In Drosophila, exposure to heat-shock caused local changes in certain TAD boundaries resulting in TAD merging (Li et al., 2015). Another recent study showed that during motor neuron (MN) differentiation in mammals, TAD and sub-TAD boundaries in the Hox cluster are not rigid, and their plasticity is linked to changes in gene expression during differentiation (Narendra et al., 2016). The global organization of the 3D genomic structure is found in mitosis (Nagano et al., 2017), the earliest stages of mammalian lineage development (Dixon et al., 2015; Bonev et al., 2017; Du et al., 2017; Ke et al., 2017), and somatic cell reprogramming of pluripotent stem cells (Novo et al., 2018; Zhang et al., 2018). Fusion of TADs (Nora et al., 2012; Dowen et al., 2014; Guo et al., 2015; Sanborn et al., 2015; Tang et al., 2015; Flavahan et al., 2016; Fudenberg et al., 2016), creation or destruction of sub-TADs within existing TAD boundaries (Lupiáñez et al., 2016; Taberlay et al., 2016), and/or switching TAD states between active and inactive conformations (Lieberman-Aiden et al., 2009; Dixon et al., 2012) has been associated with a variety of phenotypes (Misteli, 2010; Krijger and Laat, 2016; Spielmann et al., 2018), ranging from limb malformation (Lupiáñez et al., 2016), congenital disorders (Ibn-Salem et al., 2014), to cancer (Mitelman, 2000; Rickman et al., 2012; Gr̀oschel et al., 2014; Barutcu et al., 2015; Corces and Corces, 2016; Flavahan et al., 2016; Hnisz et al., 2016; Krijger and Laat, 2016; Lupiáñez et al., 2016; Valton and Dekker, 2016). Chromatin loops are even more dynamic and change during the cell cycle and other cellular events (Sanborn et al., 2015; Fudenberg et al., 2016; Golfier et al., 2019). These observations highlight the importance of studying changes in the boundaries of interacting domains as a means to understand genomic regulation. However, methods for identifying these changes remain underdeveloped.

To our knowledge, there are only three methods that can be adapted for detecting changes in boundaries of interacting domains; the majority have been developed for the detection of TAD-specific boundary changes. Among the three methods, localtadsim (Sauerwald et al., 2020), HiCDB (Chen et al., 2018), and DiffTAD (Zaborowski and Wilczynski, 2016), none provide an intuitive, easy to use way of calling differential boundaries. Both localtadsim and DiffTAD are two-step procedures requiring separately defined TADs and comparing them using a command-line utility. HiCDB has a built-in TAD caller but does not allow for comparisons of chromosome-specific contact matrices. All three methods require highly specific data types and file names to be able to run. The lack of usability is compounded with issues, such as a lack of upkeep, slow runtimes, and lack of statistical rigor (**Supplementary Methods**).

As the costs of Hi-C data continue to drop, several studies started to investigate the dynamics of 3D changes over time. The most notable applications include cell differentiation studies (Bonev et al., 2017), embryonic development (Du et al., 2017; Hug et al., 2017; Ke et al., 2017), cancer progression (Zhou et al., 2019). Typically, TAD boundary changes over time are quantified by overlap (Du et al., 2017; Hug et al., 2017) and classified into distinct patterns (Zhou et al., 2019). However, general-purpose methods for systematic analysis of boundary changes over time do not exist.

The number of replicates for Hi-C experiments continue to rise, requiring methods for defining consistent boundaries of interacting domains across replicates of Hi-C data. Two primary approaches have been developed to identify TAD boundaries across multiple replicates. The first approach involves merging all replicates into a consensus contact matrix and then calling interacting domains [e.g., Arrowhead (Rao et al., 2014)]. The second is to call domains on individual replicates and aggregate them. A third approach available in the TADBit tool (Serra et al., 2017) allows for the alignment of TAD boundaries to a reference set of boundaries. This method relies on the reference set being "true boundaries" and is potentially sensitive to the selection of reference boundaries. Altogether, methods for detecting consensus boundaries of interacting domains across Hi-C datasets remain underdeveloped.

We developed TADCompare, an R package aimed at providing a fast, accurate, user-friendly, and well-documented approach to differential analysis of boundaries of TADs and chromatin loops. We introduce a method based on the boundary score statistic (Cresswell et al., 2019) and use it to identify five types of boundary changes. The method is extended to allow for calling consensus boundaries and comparing them between groups of Hi-C replicates. We further demonstrate how the boundary score statistic may be used to analyze the dynamics of boundaries of interacting domains over the time course. For both differential boundary detection and time course analysis, we provide novel terminology for the classification of boundary changes. We demonstrated the robustness of TADCompare using simulated data with pre-defined interacting domains (Forcato et al., 2017) and its ability to reveal distinct biological roles of different boundary changes. In summary, TADCompare provides an all-in-one pipeline from consensus boundary calling to differential boundary detection, including time course. The output is formatted in a commonly used BED format that allows for flexible downstream analyses and visualization.

2. METHODS

2.1. Representation of Hi-C Data as a Graph

For a given Hi-C experiment, Hi-C data is represented by a chromosome-specific contact matrix C of non-overlapping regions (aka bins) of size r (resolution of the data). Each entry C_{ij} corresponds to the number of contacts between region i and region j. Previous work has shown that this contact matrix is essentially an analog of the adjacency matrix found in graph theory and Hi-C data can be thought of as a naturally occurring graph where edges are contacts and vertices are genomic regions (Boulos et al., 2013; Wang et al., 2013, 2019; Cresswell et al., 2019), or genes associated with them (Merelli et al., 2013). The graph representation of Hi-C data is the foundation of our method and allows us to use a graph-clustering based approach to identify and analyze TADs.

2.2. Calculating the Graph Spectrum

The first step of our method is to calculate the graph spectrum, defined as the eigenvectors of the Laplacian of an adjacency matrix. Using the interpretation of the contact matrix as a naturally occurring adjacency matrix, we calculate the Laplacian directly from the contact data. Briefly, the graph spectrum for a given contact matrix is calculated as follows:

1. Calculate the normalized Laplacian $\bar{L}$:

$$\bar{L} = D^{-\frac{1}{2}} C D^{-\frac{1}{2}}$$

where $D = diag(\mathbf{1}^T C)$, where $\mathbf{1}$ is a column vector of size C where each entry is 1. D can be thought of as a vector containing the sum of the degrees for a given node.

2. Perform an eigendecomposition of the Laplacian:

$$\bar{L}v = \lambda v$$

In practice, we calculate the first two eigenvectors with the largest absolute values of eigenvalues and organize them into a matrix $\bar{V}$ with dimensions $i \times 2$, where i is the number of regions in the contact matrix. $\bar{V}$ is referred to as the graph spectrum of the contact matrix.

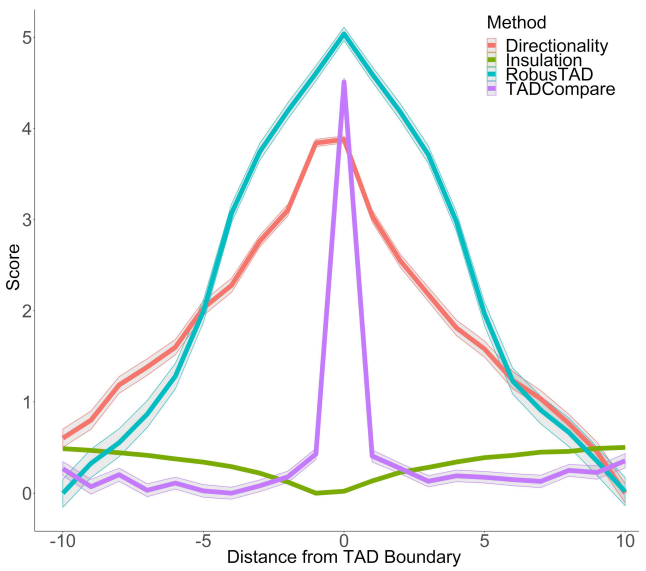


FIGURE 1 | Boundary score distinguishes boundaries better than monotonic metrics. Boundary scores calculated with four methods: directionality index, insulation score, RobusTAD, and TADCompare are shown. X-axis—distance from the boundary, measured in bins (40 kb each), Y-axis—score (signed log10 values centered at zero). Results from five simulated contact matrices, 40 kb resolution, with manually annotated boundaries (Forcato et al., 2017) are shown.

2.3. Eigenvector Gap as a Measure of Pattern Change

We can think of each row of the matrix $\bar{V}$ as a quantification of the pattern of contacts in each region of the contact matrix. Previous work (Cresswell et al., 2019) has demonstrated that by taking the Euclidean distance between row $V_{i.}$ and its neighboring row $V_{(i+1).}$, one can measure the similarity in the pattern of contacts between region i and region $i + 1$ of the chromosome, termed "eigenvector gap." A boundary between interacting regions manifests itself as a sudden break in the pattern of contacts. This pattern is reflected in the eigenvector gap by a spike in gap size followed by and preceded by smaller gaps (**Figure 1**). The eigenvector gap quantifies the degree of this break, acting as a proxy for TAD boundary likelihood. To calculate the eigenvector gaps, we perform the following procedure:

1. Normalize columns of $\bar{V}$ to sum to 1:

$$\hat{V}_{ij} = \frac{\bar{V}_{ij}}{\|\bar{V}_{.j}\|}$$

where the subscript $.j$ corresponds to column j.

2. Normalize $\hat{V}$ and project onto a unit circle:

$$\tilde{Z} = Diag(diag^{-\frac{1}{2}}(\hat{V}_{i.}\hat{V}_{i.}{}^{T}))\hat{V}_{i.}$$

3. Calculate the distance between neighboring regions (rows i and $i - 1$ of $\tilde{Z}$) and store in a vector D_i:

$$D_i = \sqrt{(\tilde{Z}_{i1} - \tilde{Z}_{(i-1)1})^2 + (\tilde{Z}_{i2} - \tilde{Z}_{(i-1)2})^2}$$

We refer to D as the vector where each entry D_i is referred to as an eigenvector gap. Formally, an eigenvector gap is the Euclidean distance between each successive row of the first two eigenvectors. In practical terms, the eigenvector gap for a given locus is a measure of how likely that loci is a boundary.

To maintain the association of each entry of the vector with its corresponding matrix region, a placeholder is used in the first entry of the vector. This is necessary because we cannot calculate an eigenvector gap for the first entry of the contact matrix due to a lack of left-bound neighbor. In mathematical terms, this means that for a matrix of size n the total number of eigenvector gaps is $n - 1$.

2.4. Converting Eigenvector Gaps to Boundary Scores

We showed that the distribution of eigenvector gaps can be approximated by a log-normal distribution (**Supplementary Figure 1**). The log-normality allows us to convert the eigenvector gap values into boundary scores:

$$B_i = \frac{(ln(D_i) - \mu)}{\sigma^2}$$

where $ln(D) \sim N(\mu, \sigma^2)$ where μ and σ^2 are the mean and variance of the distribution of the natural log of the eigenvector gaps, respectively, and B is a vector of boundary scores with a N(0,1) distribution. In practice, this value is simply the Z-score for the natural log of eigenvector gaps.

2.5. Sliding Window Eigenvector Gap Calculation

The frequency of interactions decays following power law as the distance between the interacting regions increases (Lajoie et al., 2015). This decay leads to noisy and non-informative interactions farther off-diagonal of the contact matrix. To alleviate the effect of noisy distant interactions, we perform spectral decomposition within a fixed-size window that moves along the diagonal of the matrix. For instance, a window size of 15 bins (default setting, **Supplementary Figure 2**) means that only values within 15 bins of the diagonal will be used to calculate the eigenvector gap. The sliding window approach improves the performance of the eigenvector gap calculation (Cresswell et al., 2019). Additionally, it provides for faster calculations, operating on many small matrices instead of one large matrix. In general, we found that the results are robust to window size (**Supplementary Figure 2**). At higher levels of noise and sparsity, we found that larger windows tend to perform marginally better (**Supplementary Figure 2**). This is likely due to the fact that more data points are needed to capture pattern change in these scenarios. To achieve a good compromise on performance, we used a window size of 15 for each resolution.

2.6. Handling of Non-informative Bins

Non-informative bins refer to bins with <20% of non-zero interactions. This percentage is calculated based on regions within our sliding window. Such bins can introduce instability in the algorithm and lack important information. To counter this, we remove these bins before the analysis. This is done for both contact matrices such that, if one contact matrix contains a non-informative bin at a given location and the other does not, we remove it from both. This allows us to make a one-to-one comparison of bins.

2.7. Differential Analysis Using Boundary Scores

To define the differences between two contact matrices, P and R, we compare their eigenvector gaps D_P and D_R, respectively. Given that $ln(D_P) \sim N(\mu_P, \sigma_P^2)$ and $ln(D_R) \sim N(\mu_R, \sigma_R^2)$, it follows that $ln(D_P) - ln(D_R) \sim N(\mu_P - \mu_R, \sigma_P^2 + \sigma_R^2)$. These results allow us to calculate a vector of differential boundary scores:

$$DB_i = \frac{(ln(D_{Pi}) - ln(D_{Ri})) - (\mu_P - \mu_R)}{\sigma_P^2 + \sigma_R^2}$$

or more simply,

$$DB_i = \frac{\sigma_P^2 B_P - \sigma_R^2 B_R}{\sigma_P^2 + \sigma_R^2}$$

where B_P and B_R are the boundary scores for the P and R matrices, respectively. This score can be thought of as the difference in boundary likelihood for a given locus in two data sets. Due to the aforementioned normality of the difference in log eigenvector gaps, DB_i can be thought of as a simple z-score where $DB \sim N(0, 1)$.

Boundary differences may be visualized using the package's TADcompare::DiffPlot function (**Supplementary Figure 3C**), or by external tools [e.g., HiCexplorer (Ramirez et al., 2018)].

2.8. Time Course Boundary Changes

Boundary scores provide a convenient method for modeling the change of boundaries over time. For a given boundary, or, any region of the genome, we can monitor the trajectory of the boundary score. Over time, we can define boundary score changes based on their deviation from a baseline level (typically, the first time point). It is expected that these scores will be relatively constant over time except in regions where a boundary appears or disappears. The trend across time points can be recorded and the pattern of change classified accordingly. Our implementation of time course boundary analysis allows for the usage of multiple replicates for a given time point. Briefly, at each region of the genome, the consensus boundary score is calculated, defined as the median of consensus scores across all replicates, and is then used to identify boundaries.

2.9. Gene Enrichment Testing

All gene enrichment testing was performed using the GREAT method (McLean et al., 2010) implemented in the rGREAT (Version 2.0) R package. Briefly, we detect genes within 5 kb upstream and 1 kb downstream of each type of boundary change, similar to the work of others (Chen et al., 2018). For each Gene Ontology (GO) and pathways, a hypergeometric test is then performed to determine the over-representation of boundary-associated genes. For all figures, we report results for GO Biological Processes. Results for GO Molecular Function, GO Cellular Component, MSigDB, and PANTHER pathways are reported in tables.

2.10. Colocalization Enrichment Testing

A permutation test was used to quantify the enrichment of colocalization of boundaries of interest with genomic annotations. Briefly, we flank each type of boundary change (differential or time course) by 50 kb on each side and calculate the mean number of genomic annotations across those regions (observed enrichment). Next, we generate two sets of bins, one the size of the boundaries which we are testing (considering the flanking) and another the size of all other bins. The difference in the mean number of genomic annotations colocalized with boundaries of interest was calculated for each set (expected enrichment). We repeat this procedure 10,000 times. We calculate the permutation p-value by taking the number of times the expected enrichment was greater than the observed enrichment, and dividing by 10000. $\alpha = 0.05$ was set to assess statistical significance.

2.11. Data and Code Availability

All simulated data were downloaded from the HiCToolsCompare repository (Forcato et al., 2017). In total, we used 25 simulated matrices with varying levels of noise. For sparsity and downsampling analysis matrices were manually created based on matrices from HiCToolsCompare matrices with the minimum noise level (see Cresswell et al., 2019 for methods description). Data for comparisons across cell lines, replicates, and tissues were taken from (Schmitt et al., 2016), generated at 40 kb resolution (**Supplementary Table 1**). Time course data was taken from (Rao et al., 2017), HCT-116 human colon cancer cell-line at four time points after auxin-treatment withdrawal (20, 40, 60, 180 min). Contact matrices were generated at 25, 50, and 100 kb using the straw tool from Juicer (Durand et al., 2016). Chromatin state data were taken from chromHMM (Ernst and Kellis, 2010). Histone modifications and transcription factor binding sites were downloaded from the Encyclopedia of DNA Elements (ENCODE) (Davis et al., 2018) (**Supplementary Table 2**). Scripts to recreate the results presented in the paper are available at https://github.com/cresswellkg/TADCompare_Paper. The TADCompare R package is freely available on GitHub (https://github.com/dozmorovlab/TADCompare) and on Bioconductor (submitted).

3. RESULTS

3.1. A Modified Spectral Clustering Approach Is Better Suited for Boundary Detection Than Other Approaches

Our previous work on TAD detection using spectral clustering, implemented as a SpectralTAD R package (Cresswell et al., 2019), introduced the concept of the boundary score statistic, adapted here for differential boundary detection. Briefly, the boundary score is calculated for each bin by sliding a window across the diagonal of the contact matrix, calculating the eigenvectors of the Laplacian matrix, finding the distance between consecutive eigenvectors (eigenvector gap) and converting them into Z-scores (boundary score, see Methods). The boundary score is a continuous measure of the likelihood of a given region being a boundary between interacting domains.

In contrast to other metrics for boundary identification that rely on finding inflection points of monotonic functions, such as directionality index (Dixon et al., 2012), insulation score (Crane et al., 2015), RobusTAD score (Dali and Blanchette, 2017) (**Supplementary Material**), our boundary score spikes at the boundary (**Figure 1**). This unique behavior enables easy distinction between boundaries and non-boundaries. An additional advantage of the boundary score is that its magnitude is directly interpretable as a "boundary strength." This is in contrast to other methods which are only interpretable relative to neighboring points. We can use this interpretability for parametric modeling of boundary behavior. Our previous work has shown that the boundary score is robust to noise, sparsity, and changes in sequencing depth of Hi-C data (Cresswell et al., 2019). Thus, the boundary score is well-suited for finding differences in boundaries between interacting domains.

3.2. Differential Boundary Scores Translate to Five Types of Boundary Changes

Differential boundary score is a measure of the difference between boundaries between two samples. This score follows a standard normal centered at 0 (see Methods, **Supplementary Figure 1**). Differential boundaries are detected by finding regions with the absolute differential boundary score is >2 (**Supplementary Figure 2**), which intuitively corresponds to differences with a p-value smaller than 0.05.

We divide boundary changes into five categories (complex, split, merge, shifted, strength change; **Figure 2**, **Supplementary Figure 3**). A similar strategy was used in Ke et al. (2017). An interacting domain can be **split** between the datasets, meaning it exists as a continuous domain in one and is split into two or more domains in another. In practice, this situation requires two shared boundaries and a differential domain between them. **Merging** is the opposite of splitting and arises when a boundary surrounded by two non-differential boundaries disappears in one of the contact matrices. Classification of boundary change as merged and split depends on the reference contact matrix being compared to. Finally, domains can be split in a **complex** way, meaning they are neither split or merged but instead taking on an entirely new structure. Merged and split boundaries represent the structural change of the same domain as opposed to complex boundaries, which we consider to be part of a completely different domain. The "complex," "merge," and "split" boundaries are considered to be the most disruptive changes in the 3D structure of the genome.

A **shifted** boundary is defined as the non-overlapping boundary that lies within five bins (or another user-defined threshold) of a boundary in the contact matrix in which it is being compared to. A **strength change** occurs when a boundary is present in both contact matrices, but its differential boundary score magnitude is greater than the differential threshold of 2. The other cases are considered to be non-differential boundaries. This framework allows us to systematically compare and classify boundary changes.

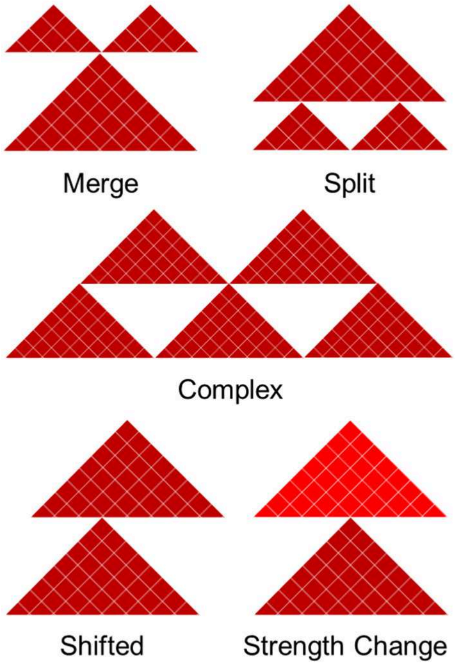


FIGURE 2 | Five types of boundary changes. Complex, split, and merge boundary changes are considered as the major differences, while shifted and strength changes are considered as the minor differences.

3.3. Boundaries Are Highly Consistent in Both Technical and Biological Replicates

Previous studies have shown that the overlap between TAD boundaries in replicate data ranges from around 60 to 70% (Dixon et al., 2012; Rao et al., 2014; Sauerwald et al., 2020). Additionally, technical replicates have been shown to have a slightly higher proportion of shared TAD boundaries (∼65%) than biological replicates (∼60%) (Sauerwald et al., 2020). We have tested and confirmed these observations by showing that significantly more boundaries were non-differential in technical replicates than in biological replicates (73 vs. 65.7%). Similarly, 9.3/8.1% of boundaries showed significant strength change, while 7.8/6.1% were shifted in the biological/technical replicates, respectively. A similar trend was observed for complex and merge-split boundaries. In summary, only 17.2/12.8% of boundaries were differential in biological/technical replicates, respectively (**Figure 3A**), confirming the higher stability of the 3D structures in technical replicates.

3.4. Boundaries Are More Similar Within Cells Than Tissues

Previous research showed that TADs are largely invariant across cell lines and, to a lesser extent, tissue types (Pope et al., 2014; Rao et al., 2014; Schmitt et al., 2016). However, the types of boundary changes remained undefined. We compared Hi-C matrices of seven different cell-lines and 18 different tissue types (Schmitt et al., 2016) (**Supplementary Table 3**). In total, the average percentage of differential boundaries was significantly less in cell lines (22.5%) than tissue samples (39.7%, **Figure 3B**). As expected, these percentages were higher than those for biological (17.2%) and technical replicates (12.8%). These results suggest that the variability of boundaries mirrors the homogeneity of data types (technical replicates, biological replicates, cell lines, and tissues, in that order).

3.5. Each Type of Differential Boundaries Is Associated With Different Levels of Epigenomic Enrichment

To understand the biological relevance of the types of boundary changes, we identified changes between the GM12878 and IMR90 cell lines [chr 1–22, 40 kb resolution (Schmitt et al., 2016)] and categorized them according to the type of change. For each change type, we assessed the number of overlapping peaks and calculated the enrichment of four genome annotation marks known to co-locate with TAD boundaries—CTCF, RAD21, insulators, and heterochromatin states.

We found that non-differential boundaries had a higher average number of overlapping peaks for all four marks, followed by "strength change" boundaries (**Figure 4A**). Similarly, enrichment of non-differential boundaries was the most significant (**Figure 4B**). Notably, the number of peaks for each mark was highly variable on "strength change" boundaries (**Figure 4A**), suggesting their biological relevance is less certain. Similarly, "shifted" boundaries had the lowest average number of peaks, suggesting that they may be detected due to noise and, consequently, be less biologically significant. In contrast, "complex" and "merge-split" boundaries had a moderate number of overlapping peaks and were moderately enriched in them (**Figure 4**). These results highlight the varied biological relevance of different types of boundary changes and suggests "complex" and "merge-split" changes are biologically important alterations of the 3D structure.

3.6. Each Type of Differential Boundaries Is Associated With Distinct Biological Functionality

To test the biological significance of different types of boundary changes, we compared neural progenitor cells (NPC) against mesenchymal stem cells (MSC) (Schmitt et al., 2016) (**Figure 5A**, **Supplementary Figure 3C**). Altogether, we found that the vast majority of boundaries are either complex (38.6%) or non-differential (32.6%). Shifted (17.5%), merge-split (7.7%) and strength change (3.5%) were less common (**Figure 5B**). Under the hypothesis that differential boundaries may be enriched in genes driving relevant biological processes (Chen et al., 2018), we investigated enrichment of genes in proximity of each type of differential TAD boundary in biological processes and other gene ontology- and pathway types using GREAT (McLean et al., 2010) (see Methods). As NPCs are more advanced on differentiation path than MSCs, we expected that boundaries changed between them would be associated with genes responsible for neural development-related processes. Indeed, genes around "merge" and "complex" boundary changes, as well as the "non-differential" boundaries were enriched in a

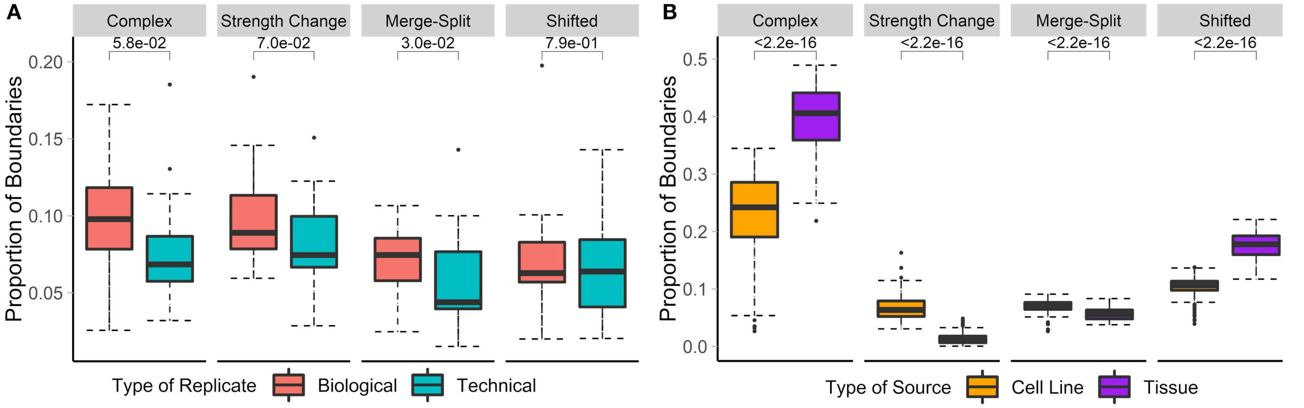


FIGURE 3 | Biological replicates and cell lines have more differential boundaries than technical replicates and tissues, respectively. Differential boundaries were calculated between Hi-C datasets of biological and technical replicates [**A**, HCT-116 cell line, 50 kb resolution, chr 1–22 (Rao et al., 2017)] and between cell lines and tissues [**B**, various cell lines, 40 kb resolution, chr 1–22 (Schmitt et al., 2016)]. Types of boundary changes were recorded, and the proportions of boundary differences for each type were summarized across chromosomes.

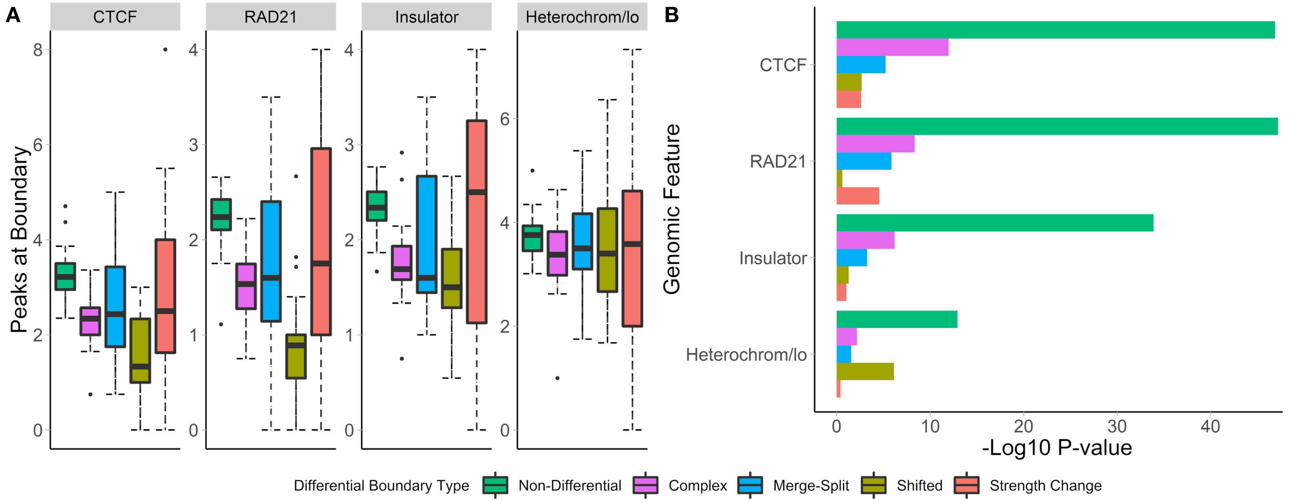


FIGURE 4 | Non-differential boundaries are more enriched for selected genome annotation marks than other types of differential boundaries. Differential boundaries were called between GM12878 and IMR90 cell lines and categorized based on differential boundary type. **(A)** The number of peaks at boundaries and **(B)** permutation *p*-values (−log10) are shown. Data from Schmitt et al. (2016), 40 kb resolution, chr 1–22.

variety of developmental processes (e.g., "cellular developmental process," etc.), including neural-specific ("nervous system development," **Figure 5B**). Notably, "split" boundary changes were not enriched in these processes, indicating the importance of the directionality of boundary changes. Genes around "merge" and "non-differential," but not "complex," boundaries were enriched in differentiation-related processes (e.g., "positive regulation of cell differentiation"), while "forebrain radial glial cell differentiation" and "neural tube development" processes were exclusively enriched in genes around "merged" boundaries (**Figure 5B**). In this case, "merge" indicates boundaries enriched in the NPC cell-line, causing a separation of interacting domains in MSC and "split" indicates a split in NPC caused by a boundary enriched in MSC. As expected, genes around "noisy" boundary changes ("shifted" and "strength change") lacked enrichment in any biological processes (**Figure 5B**, **Supplementary Table 4**). These results emphasize the importance of classifying boundary changes into distinct patterns that tend to be associated with distinct biological functionality.

To further test whether different types of boundary changes reflect biology of an experimental system, we used post-auxin treatment time course experiment from Rao et al. (2017) study (HCT-116 cell line, 40 kb resolution, 20, 40, 60, and 180 min following auxin withdrawal, 4 replicates at each time point) (Rao et al., 2017). Auxin treatment eliminates CTCF binding genome-wide; consequently, the majority of boundaries should be absent and gradually re-appear following auxin withdrawal. To identify biological processes

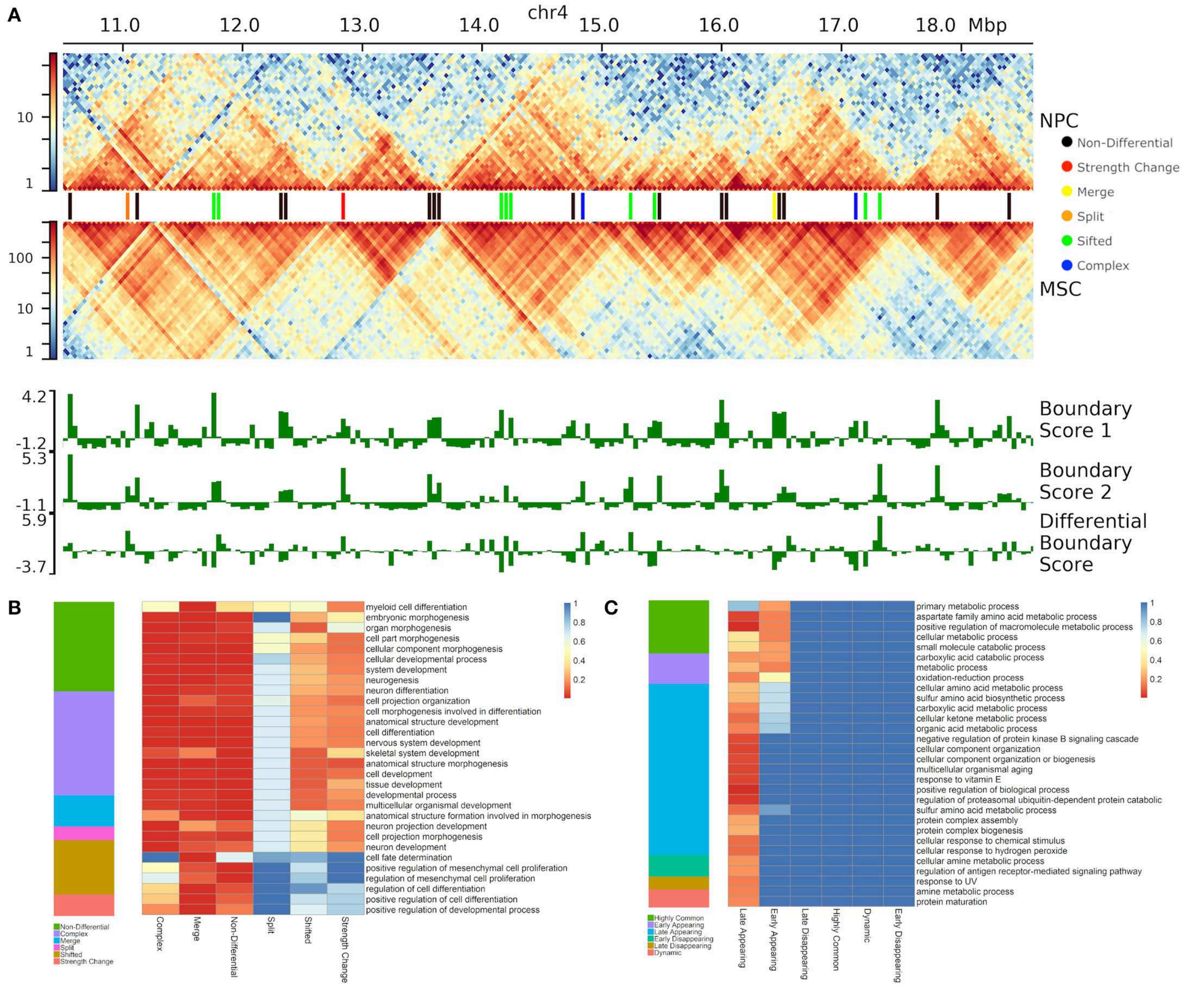


FIGURE 5 | Differential boundaries and their gene enrichment analysis. **(A)** An example of differential boundaries called between neural progenitor cell (NPC) and mesenchymal stem cells (MSC) (Schmitt et al., 2016) (chr4:10500000–18600000 region, 40 kb resolution); outlined TADs were called using SpectralTAD (Cresswell et al., 2019). **(B,C)** The top 30 gene ontologies most enriched **(B)** in NPC vs. MSC boundary comparison, and **(C)** across the time-course of boundary changes in auxin-treated cells from the HCT-116 cell-line (Rao et al., 2017) (chr 1–22, 40 kb resolution). For each type of boundary change, enrichment *p*-values (rGREAT, see Methods) are shown as heatmaps.

associated with re-appearing of boundaries, we compared first and last time points (20 and 180 min) following auxin withdrawal. As boundaries were reported to be enriched in housekeeping genes (Jin et al., 2013), we expected genes around appearing boundaries to be enriched in general cellular processes. Indeed, the vast majority of boundaries were complex (41.4%) and non-differential (34.7%) (**Supplementary Figure 4**). We found that only genes around "non-differential" and "complex" TAD boundary changes showed some level of enrichment (**Supplementary Figure 4**, **Supplementary Table 5**). As expected, "metabolic processes" and various developmental and housekeeping processes were specifically enriched in genes around complex boundary changes, while cyclic AMP synthesis and metabolic processes were enriched in genes around "non-differential" boundaries. From these results, we show that TADCompare can correctly classify less-essential boundary changes ("shifted," "strength change") and detect distinct boundary changes associated with shared and unique biological processes.

3.7. Time Course Analysis Framework

Time course analysis of boundaries refers to the analysis of boundary dynamics over time. The quantitative nature of boundary score allows us to monitor its changes at boundaries across any number of time points. We recommend taking a union of boundaries detected at each time point and monitor boundary score changes for each boundary. Monitoring boundary scores

TABLE 1 | Six patterns of temporal boundary changes.

Temporal boundary type	Time point 1	Time point 2	Time point 3	Time point 4	Total (% occurrence)
Highly common	1	1	1	1	326 (17.35%)
	1	0	1	1	
Early appearing	0	1	1	1	184 (9.79%)
Early disappearing	1	0	0	0	133 (7.08%)
Late appearing	0	0	1	1	1,047 (55.72%)
	0	0	0	1	
Late disappearing	1	1	0	0	79 (4.20%)
	1	1	1	0	
Dynamic	1	0	1	0	110 (5.86%)
	1	0	0	1	

Each column corresponds to a point in time.
"1" refers to the presence of a boundary, and "0" refers to the absence of a boundary. The "Total" column shows the percentage of occurrences in the CTCF degradation-recovery time course, HCT-116 cell line, chr 1–22 (Rao et al., 2017).

across time points provides an opportunity to quantify patterns of boundary changes.

Using the boundary score cutoff of 3 for boundary definition, we define six patterns of temporal boundary changes (adapted from Zhou et al., 2019, **Table 1**, **Figure 6**). *Highly common* boundaries refer to boundaries present across all time points or in three out of four time points. *Early appearing* boundaries switch from non-boundary to boundary at second time points and stay as boundaries for the rest of the time points. Conversely, *early disappearing* boundaries switch from boundary to non-boundary at the second time point and stay as non-boundaries. *Late appearing* boundaries switch from non-boundaries to boundaries at the last or the second to last time point. Conversely, *late disappearing* switch from boundaries to non-boundaries at the last of the second to last time point. Finally, *dynamic* boundaries are those which have inconsistent boundary status and do not follow any of the aforementioned patterns (**Figure 6**). These six patterns of temporal changes can be easily adapted for a larger number of time points.

3.8. Temporal Boundary Types Are Associated With Different Levels of Epigenomic Enrichment

To evaluate the biological relevance of temporal patterns of boundaries, we used post-auxin treatment time course experiment introduced above. Briefly, HCT-116 cells were treated with auxin to eliminate boundaries, and Hi-C measures were obtained at 20, 40, 60, and 180 min following auxin withdrawal and subsequent boundary reappearance (Rao et al., 2017). Accordingly, we expected to detect some number of highly common boundaries (already existing at 20 min) and boundaries appearing at different stages of post-auxin withdrawal (early/late appearing). Conversely, dynamic and early/late disappearing boundaries should be rare and may potentially constitute noise in TAD boundary detection.

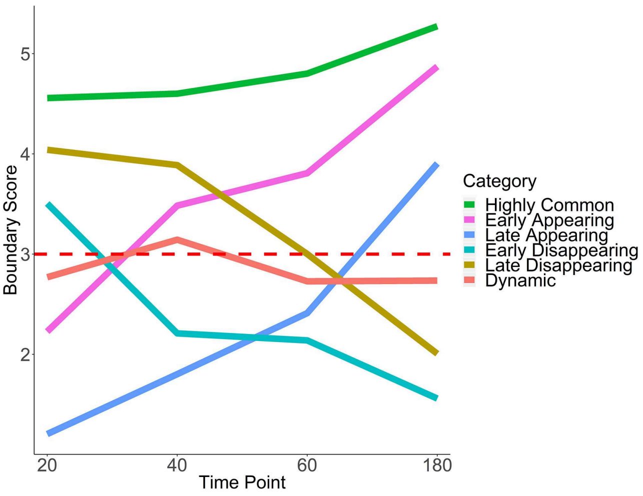


FIGURE 6 | Six patterns of boundary score change across time. Average trajectories for each pattern of boundary score change are shown. The red horizontal line indicates the cutoff for boundary detection. HCT-116 cell line, 40 kb resolution, chr 1–22.

TABLE 2 | Consensus (aka median) boundary score is supported by high boundary scores from multiple replicates.

Boundary score 1	Boundary score 2	Boundary score 3	Consensus boundary score	Union boundary?	Consensus boundary?
1	2	1	1	No	No
3	2	1	2	Yes	No
5	5	4	5	Yes	Yes
3	3	3	3	Yes	Yes
6	0	0	0	Yes	No

Examples of boundary scores across five regions in three replicates, and the corresponding consensus boundary score. Both union and consensus boundaries are calculated using a cutoff of 3.

Boundary scores were calculated for auxin-treated cells 20, 40, 60, and 180 min after withdrawal. Taking the union of boundaries (boundaries detected at one or more time points), we calculated temporal patterns for each boundary. We found that the vast majority of boundaries were late appearing (55.7%) (**Table 2**, **Figure 5C**). Early appearing (9.8%) and highly common (17.3%) made up most of the other boundaries present at the end of the time course. Approximately 20% of boundaries were highly common, i.e., resistant to auxin treatment, a number similar to previous works (Nora et al., 2017). Meanwhile, 5.9% of boundaries were dynamic, 7.1% were early disappearing, and 4.2% were late disappearing, highlighting potential errors in boundary detection. In summary, some boundaries can be detected at 20 min post-auxin treatment and remain present through all time points; however, the timing of boundary reappearance varies.

To test whether boundaries associated with different temporal patterns have different functional roles, we investigated their overlap with and enrichment in the common marks of TAD

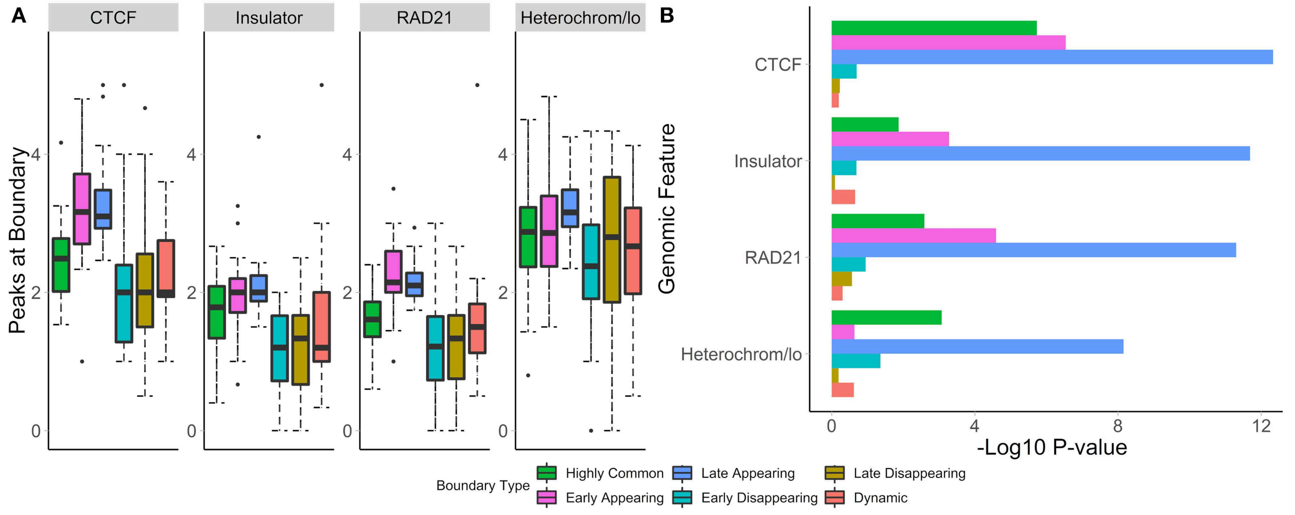


FIGURE 7 | Common and appearing boundaries show stronger enrichment in known epigenomic marks. The number of peaks at boundaries **(A)**, and permutation *p*-values **(B)** within 50 kb of boundaries in each temporal classification are shown. Hi-C data from Rao et al. (2017), 50 kb resolution, HCT-116 cell-line, chr 1–22.

boundaries (CTCF, RAD21, insulators, heterochromatin, **Figure 7A**). For highly common, early- and late-appearing boundaries, we observed more overlaps with CTCF and RAD21 sites, insulator, and heterochromatin states (**Supplementary Table 6**). Similarly, these types of boundaries were highly enriched in the aforementioned genomic annotations (**Figure 7B**). Conversely, dynamic, early, and late disappearing boundaries showed less overlap with CTCF, RAD21, insulator, and heterochromatin marks, and were less enriched in them. These observations suggest that disappearing and dynamic boundaries are likely detected due to noise in the data, while boundaries appearing after auxin treatment expectedly represent the biologically relevant signal.

3.9. Temporal Boundary Types Are Associated With Distinct Biological Functionality

Using gene enrichment analysis, we further investigated whether boundaries associated with different temporal patterns may be enriched in genes driving relevant biological processes (Chen et al., 2018) (**Supplementary Table 7**). We found that, with a few exceptions, all significant GO Biological pathways were enriched in late or early appearing boundaries (**Figure 5C**, **Supplementary Table 7**), which make up the majority of boundaries (**Table 2**, **Figure 5C**). Both early and late appearing boundaries were enriched in metabolism-related processes, such as "cellular metabolic process," "oxidation-reduction process." Late appearing boundaries, on the other hand, were enriched in "cellular component organization," "protein complex biogenesis" and the like processes (**Figure 5C**). These results are expected as cells may be activating metabolic and biogenesis pathways to recover after destruction of boundaries by auxin. These results confirm that TADCompare can accurately classify biologically relevant temporal boundary changes and discern them from noisy changes.

3.10. Consensus Boundary Score for Defining Robust Boundaries Across Multiple Hi-C Datasets

The sizeable proportion of noisy "shifted" and "strength change" boundary changes across Hi-C datasets (**Figure 3**) highlights the need to identify boundaries that are robustly detected. The consensus boundary score, defined as the median of boundary scores across replicates, addresses this challenge. Intuitively, higher consensus boundary scores correspond to boundaries supported by evidence from multiple replicates (**Table 2**). This is in contrast to a union of boundaries, where boundaries detected in at least one Hi-C dataset are pooled together. Consensus boundary scores allow us to filter out boundaries with insufficient support from multiple replicates, thus "denoise" the detected boundaries. Given the fact that boundary scores are log-normally distributed (**Supplementary Figure 1**, **Supplementary Methods**), the consensus boundary scores will also be asymptotically normal. The consensus boundary score can be used as a proxy for the normal boundary score for the analysis of replicated Hi-C datasets. Consequently, the consensus boundary scores may be compared to define boundary changes between groups of replicated Hi-C datasets.

3.11. Consensus Boundaries Are Supported by Strong Biological Evidence

To investigate the biological relevance of boundaries defined using consensus boundary score, we defined consensus boundaries across seven cell-lines (17 matrices total) (Schmitt et al., 2016). These boundaries represent cell type-invariant boundaries supported by evidence from multiple datasets. Bins of the genome were separated into three categories based on

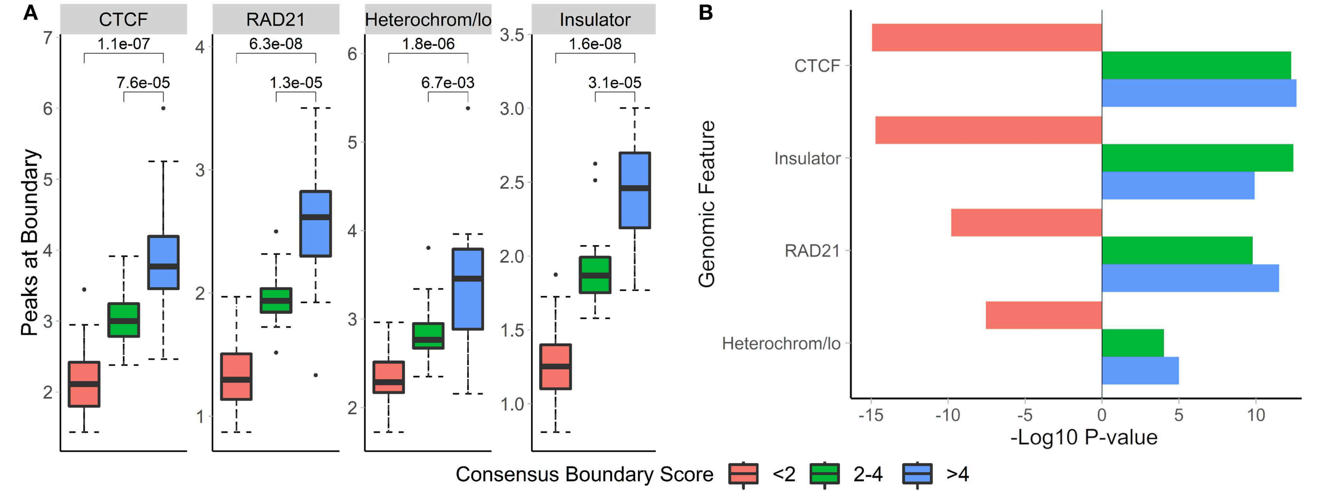


FIGURE 8 | Boundaries defined at higher consensus boundary score thresholds show stronger overlap with and enrichment in known epigenomic marks. Boundaries were classified based on the range of their consensus boundary score. Enrichment of genomic factors known to occur near TAD boundaries was calculated. **(A)** The number of peaks within 40 kb of boundaries with the corresponding consensus score range and **(B)** the −log10-transformed permutation *p*-values for each score range are shown. Negative *p*-values indicate depletion. Data from seven cell lines, chr 1–22, 40 kb resolution (Schmitt et al., 2016).

the level of their consensus boundary score (<2, 2–4 and >4). In total, there were 65,336 bins (40 kb resolution). Expectedly, the majority (62,791 bins, 96.1% of all bins) were in the <2 category, 2,032 (3.1%) bins were in the 2-4 category, and 513 (0.8%) bins were in the >4 category. We assessed the number of overlapping peaks and the enrichment of CTCF, RAD21, insulators, and heterochromatin states in different categories of bins. Expectedly, we observed increasing average number of peaks overlapping bins selected at more stringent consensus boundary score thresholds (**Figure 8**, **Supplementary Table 8**). Similarly, bins with higher consensus boundary scores have stronger enrichment in genome annotations, while bins with score <2 were significantly depleted. These results suggest that bins with higher consensus boundary scores (i.e., supported by evidence from multiple Hi-C datasets) are more biologically relevant. Therefore, to define consensus boundaries, we use a consensus boundary score cutoff of 3.

3.12. The Union of Boundaries Is Supported by Weaker Biological Evidence Than Consensus Boundaries

The union of boundaries called in individual Hi-C datasets represents an alternative method of defining boundaries across multiple datasets (**Table 1**). The union method may be useful for analysis of time course data, where boundaries are expected to change across individual datasets. We hypothesized that the union method would select for the less biologically relevant set of boundaries because many may be detected due to noise in Hi-C data.

To evaluate the biological relevance of boundaries called using both methods, we call consensus and union boundaries on a set of replicates (four cell lines, 40 kb resolution, three replicates each, data from Schmitt et al., 2016). Consensus scores were calculated separately for each cell line among the three replicates. Expectedly, the consensus method filtered out 38% of boundaries (4,906 vs. 3,059, **Supplementary Figure 5**), suggesting that many boundaries are detected in single datasets. We found that boundaries called using consensus boundary score overlapped significantly more with CTCF sites (P = 0.0006) and RAD21 (P = 0.0002) than those called using the union method (**Figure 9A**). While the enrichment results were similar for consensus- and union-defined boundaries, consensus boundaries were more significantly enriched in "heterochromatin" (**Figure 9B**). Together with previous observations (**Figure 6**), these results strengthen our conclusion that consensus boundary scores are more effective in removing "noisy" boundaries that otherwise would be captured using the union method.

3.13. Runtime Performance of TADCompare

When run on data from (Rao et al., 2014), without parallelization, both consensus boundary calling and differential boundary detection were exceptionally fast. In total, for the entire genome, differential boundary detection took ~6 s on 100 kb data, ~9 s on 50 kb data, ~17 s on 25 kb data, and ~312 s on 10 kb data. In the case of consensus boundary calling, TADCompare took ~17 s to run on 50 kb data for 4 matrices, ~32 s for 8 matrices, and ~45 s for 12 matrices. On 10 kb data, it took ~611 s to run for 4 matrices, ~1,152 s for 8 matrices, and ~1,680 s for 12 matrices. For a full summary of runtimes across all resolutions (see **Supplementary Figure 6**).

4. DISCUSSION

The initial development of Hi-C technologies focused on investigating individual genomes. While several key properties have been discovered (chromosome territories, A/B

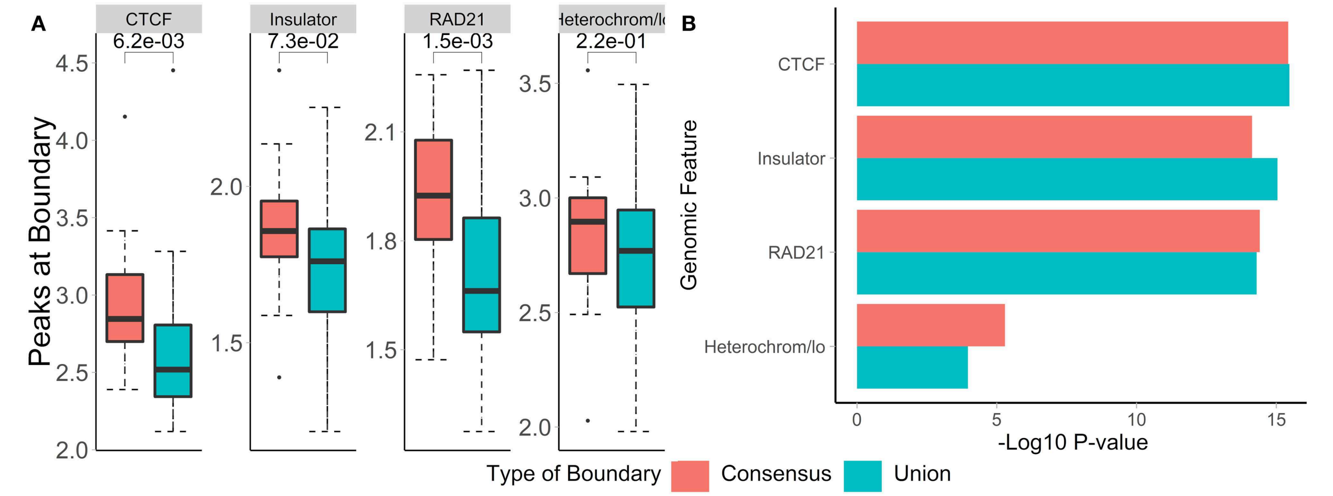


FIGURE 9 | Consensus boundaries show stronger overlap with and enrichment in known epigenomic marks than the union of boundaries. **(A)** Number of peaks at boundaries and **(B)** permutation *p*-values (−log10) are shown. Data from Schmitt et al. (2016), four cell lines, 40 kb resolution, chr 1–22.

compartments, TADs, chromatin loops, collectively referred to as "interacting domains"), the next steps include investigating changes in the 3D structure across multiple conditions. We (Stansfield et al., 2018, 2019) and others (Lun and Smyth, 2015; Djekidel et al., 2018) started to develop methods for comparative analysis of the 3D structures. However, to our knowledge, no methods are available for differential analysis of boundaries demarcating interacting domains. In this work, we introduce a method for differential boundary analysis, including a time course, that supports replicated Hi-C data. The method is based on a novel boundary score metric that provides a continuous measure of boundary likelihood (Cresswell et al., 2019). We introduce unique terminology for classifying differential and temporal boundary changes. We show that our approach is robust and effective at identifying distinct biology associated with different types of boundary changes. Our method is implemented in the TADCompare R package available on Bioconductor, filling a vital gap in intuitive R-based software for boundary detection and comparison.

The boundary score concept developed in our work addresses three main problems: differential boundary detection, time course analysis of boundary changes, and consensus boundary calling. Yet, it has a broader scope of applications. Future work will expand the utility of boundary score by developing a similarity/reproducibility score to measure the agreement between (multiple) Hi-C matrices, in the same vein as HiCRep (Yang et al., 2017), Selfish (Ardakany et al., 2019), GenomeDISCO (Ursu et al., 2018), HiC-Spector (Yan et al., 2017), QuASAR-Rep (Sauria et al., 2015). Furthermore, for differential boundary detection, our method is still limited to the comparison of two profiles of (consensus) boundary scores. This approach will eventually be expanded to include comparisons of many contact matrices, similar to the concept of comparing groups of multiple replicates in RNA-seq data. Finally, there is still room for expansion of time course boundary analysis. The continuous nature of boundary score allows for adopting time course analysis methods developed for gene expression studies (Bar-Joseph et al., 2012). More flexible classification of temporal trends may be considered, such as 24 temporal patterns proposed by Zhou et al. (2019), or fuzzy clustering techniques that do not require a pattern to belong to a specific cluster (Abu-Jamous and Kelly, 2018). In summary, our work enables further development of various aspects of 3D genome analysis.

One difficulty in our work is how to accurately quantify the biological relevance of boundaries (differential, time-varying, and consensus) that we detect. There is no natural gold standard for boundaries, but there are known genomic features that form the building blocks of TADs (CTCF, RAD21). In practice, we can use colocalization and/or signal enrichment of these marks near boundaries as a proxy for "true boundaries." To test whether enrichment is different than random (non-boundaries), we use a permutation test and present these *p*-values. In the current work, we used colocalization enrichment analysis, and plan to address changes in signal enrichment associated with changes in boundaries in future work.

The goal of the TADCompare package is to provide a practical implementation of our statistical framework for differential boundary detection. It outputs genomic coordinates of differential boundaries, type of the differences, and the associated boundary score measures. The downstream analysis options may be gene enrichment analysis in the proximity of (different types of) differential boundaries using rGREAT, epigenomic enrichment analysis [GenomeRunner (Dozmorov et al., 2012, 2016), LOLA (Sheffield and Bock, 2016)], and visual exploration of differential boundaries. Although TADCompare provides simultaneous visualization of two Hi-C matrices and the associated boundary differences and boundary scores, external tools for visualizing multiple datasets may be explored (reviewed in Yardimci et al., 2019). Tools like the HiCBricks R package (Pal et al., 2019) and the HiCexplorer Python software (Ramirez et al., 2018) start enabling the users to visualize two Hi-C matrices and the associated annotations. We continue exploring visualization options to improve exploration and interpretation of boundary differences.

Our results in this manuscript demonstrate the ability of TADCompare to provide accurate, biologically relevant results. The methods implemented span differential, time-course, and consensus analysis. To date, TADCompare is the only actively maintained and publicly available tool to provide any of this functionality. We intend for TADCompare to be a one-stop tool for comparison of HiC datasets, providing simple, easy-to-interpret results in a timely manner. As a one-of-a-kind tool, TADCompare will increase the ability of researchers to extract important biological insights from the structure of TAD boundaries.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: All URLs are listed in the **Supplementary Tables 1**, **2**. Main datasets include: ftp://ftp.ncbi.nlm.nih.gov/geo/series/GSE87nnn/GSE87112/suppl/GSE87112_file.tar.gz, https://bitbucket.org/mforcato/hictoolscompare/get/406ee2349566.zip, https://www.ncbi.nlm.nih.gov/geo/download/?acc=GSE104333, https://www.ncbi.nlm.nih.gov/geo/download/?acc=GSE63525.

AUTHOR CONTRIBUTIONS

MD and KC conceived the project. KC implemented the TADCompare and wrote the analysis scripts. MD and KC wrote the manuscript.

FUNDING

This work was supported in part by the PhRMA Foundation Research Informatics Award and the George and Lavinia Blick Research Fund scholarship to MD.

ACKNOWLEDGMENTS

The authors would like to acknowledge Dr. Katarzyna Tyc for helpful feedback.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.00158/full#supplementary-material

Supplementary Figure 1 | Log-normal distribution of eigenvector gaps converted to boundary Z-scores. Eigenvector gaps were calculated for contact matrices across three resolutions [10, 25, and 50 kb, Hi-C data from Rao et al. (2014), GM12878 cell line, chr 1–22]. Density plots are shown for the **(A)** Natural log of the eigenvector gaps and **(B)** Boundary scores derived from the same data, separated by resolution. Regions of non-TADs are highlighted by a yellow bar, moderate strength boundaries (2 < boundary score cutoff < 3) are highlighted by a red bar, and strong boundaries (cutoff > 3) are shown using a green bar. We see a slightly right-skewed distribution due to the filtering of gaps for plotting purposes.

Supplementary Figure 2 | Window size of 15 units of Hi-C data resolution and boundary score cutoff of 2 yields consistent boundary detection. Differential boundaries were compared between two simulated data sets with window size sizes ranging from 10 to 25, and boundary score cutoff ranging from 1.5 to 4. Youden index (balanced sensitivity and specificity metric) was calculated for each combination and plotted to show agreement with ground-truth annotations. Results are shown for noise-injected matrices **(A)** and sparsity-injected matrices **(B)**.

Supplementary Figure 3 | Visualization of different types of boundary score patterns. **(A)** Patterns of raw boundary scores are shown for five different types of differential boundaries (Merge, split, complex, shifted, and strength change). The red horizontal line corresponds to the user-adjustable cutoff for a boundary. Human neural progenitor cells (NPCs), chr22, most representative examples are shown. **(B)** TADCompare::DiffPlot differential boundary visualization between NPCs and mesenchymal stem cells (MSC), chr4:10500000–18600000. 40 kb resolution data from Schmitt et al. (2016).

Supplementary Figure 4 | Heatmap of gene ontology enrichment at the first and last time point in auxin-treated data. Differential boundary identification was performed on auxin-treated data at the time of application (first time point) and complete withdrawal (last time point) [HCT-116 cell line, chr 1–22, 40 kb resolution (Schmitt et al., 2016)]. A barplot of the proportion of each differential boundary type and FDR-adjusted hypergeometric *p*-values obtained from gene ontology enrichment analysis using rGREAT (see Methods) are shown. The top 30 pathways, in terms of average enrichment, are shown and clustered using the Ward method.

Supplementary Figure 5 | Venn diagram of union and consensus boundary counts. Consensus and union boundaries were called across four different cell lines (hesc, mesynchymal, npc, trophectoderm), and the number of union and consensus boundaries was recorded. The Venn diagram shows the complete overlap of consensus boundaries within union boundaries (40 kb resolution, data from Schmitt et al., 2016).

Supplementary Figure 6 | Runtime of TADCompare. Plot containing the runtime of two-way comparison **(A)** and consensus boundaries called on 4, 8, 12, and 16 replicates **(B)**. Each point represents the runtime for a specific chromosome. X-axis—chromosome, Y-axis—runtime in seconds. Hi-C data from Rao et al. (2014), chr 1–22, 10, 25, 50, and 100 kb resolution.

Supplementary Table 1 | Contact matrix data sources. The source of all contact matrices, experimental, and simulated, used in this paper are provided. Experimental data are separated based on the study and cell line.

Supplementary Table 2 | Genomic annotation data sources. The sources, with download links, for all genomic annotations used in this paper are included.

Supplementary Table 3 | Summary of differential boundary types across tissues and cell lines. The percentage of each type of differential boundary for all tissue-tissue and cell line-cell line comparisons is reported. Results are aggregated over all chromosomes. Hi-C data from Schmitt et al. (2016), 40 kb resolution, chr 1–22.

Supplementary Table 4 | Gene ontology enrichment for differential boundary types. Differential boundaries were identified between the neural progenitor cells (NPC) and mesenchymal stem cells (MSC) (Schmitt et al., 2016). Pathway analysis was performed using rGREAT (Methods), and results are separated by ontology. Boundaries with an FDR adjusted *p*-value of <0.3 are shown. 40 kb resolution, chr 1–22.

Supplementary Table 5 | Gene ontology enrichment between the first and last time points in auxin-treated data. Differential boundaries were identified between the first and last time points of auxin-treated data (Rao et al., 2017). Pathway analysis was performed using rGREAT (Methods), and results are separated by ontology. Boundaries with an FDR adjusted *p*-value of <0.3 are shown. 50 kb resolution, chr 1–22.

Supplementary Table 6 | Enrichment across different temporal boundary types. Temporal boundary types were identified across four time points in auxin-treated data (Rao et al., 2017). Results are shown for four types of temporal boundaries (Early Appearing, Late Appearing, Highly Common, Dynamic). Permutation *p*-values, along with enrichment or depletion designations, are reported. HCT-116 cell line, 40 kb resolution, chr 1–22.

Supplementary Table 7 | Gene ontology enrichment for different temporal boundary types. Temporal boundary types were identified across four time points in auxin-treated data (Rao et al., 2017). For each temporal boundary type, pathway analysis was performed using rGREAT (Methods), and results are separated by ontology. Boundaries with an FDR adjusted *p*-value of <0.3 are shown. HCT-116 cell line, 50 kb resolution, chr 1–22.

Supplementary Table 8 | Enrichment across different consensus scores. Consensus scores were called across 17 contact matrices representing seven different cell lines. Results were dichotomized into three groups (<2, 2–4, >4) based on consensus boundary scores. Permutation *p*-values, along with enrichment or depletion designations, are reported. Hi-C data from Schmitt et al. (2016), 40 kb resolution, chr 1–22.

REFERENCES

Abu-Jamous, B., and Kelly, S. (2018). Clust: automatic extraction of optimal co-expressed gene clusters from gene expression data. *Genome Biol.* 19:172. doi: 10.1186/s13059-018-1536-8

Ardakany, A. R., Ay, F., and Lonardi, S. (2019). Selfish: discovery of differential chromatin interactions via a self-similarity measure. *Bioinformatics.* 35, i145–i153. doi: 10.1093/bioinformatics/btz362

Bar-Joseph, Z., Gitter, A., and Simon, I. (2012). Studying and modelling dynamic biological processes using time-series gene expression data. *Nat. Rev. Genet.* 13, 552–64. doi: 10.1038/nrg3244

Barutcu, A. R., Lajoie, B. R., McCord, R. P., Tye, C. E., Hong, D., Messier, T. L., et al. (2015). Chromatin interaction analysis reveals changes in small chromosome and telomere clustering between epithelial and breast cancer cells. *Genome Biol.* 16:214. doi: 10.1186/s13059-015-0768-0

Bonev, B., Mendelson Cohen, N., Szabo, Q., Fritsch, L., Papadopoulos, G. L., Lubling, Y., et al. (2017). Multiscale 3D genome rewiring during mouse neural development. *Cell* 171, 557–572.e24. doi: 10.1016/j.cell.2017.09.043

Boulos, R. E., Arneodo, A., Jensen, P., and Audit, B. (2013). Revealing long-range interconnected hubs in human chromatin interaction data using graph theory. *Phys. Rev. Lett.* 111:118102. doi: 10.1103/PhysRevLett.111.118102

Chen, F., Li, G., Zhang, M. Q., and Chen, Y. (2018). HiCDB: a sensitive and robust method for detecting contact domain boundaries. *Nucleic Acids Res.* 46, 11239–11250. doi: 10.1093/nar/gky789

Corces, M. R., and Corces, V. G. (2016). The three-dimensional cancer genome. *Curr. Opin. Genet. Dev.* 36, 1–7. doi: 10.1016/j.gde.2016.01.002

Crane, E., Bian, Q., McCord, R. P., Lajoie, B. R., Wheeler, B. S., Ralston, E. J., et al. (2015). Condensin-driven remodelling of X chromosome topology during dosage compensation. *Nature* 523, 240–244. doi: 10.1038/nature14450

Cremer, T., and Cremer, M. (2010). Chromosome territories. *Cold Spring Harb. Perspect. Biol.* 2:a003889. doi: 10.1101/cshperspect.a003889

Cresswell, K. G., Stansfield, J. C., and Dozmorov, M. G. (2019). SpectralTAD: an R package for defining a hierarchy of topologically associated domains using spectral clustering. *bioRxiv* 549170. doi: 10.1101/549170

Dai, Z., and Dai, X. (2012). Nuclear colocalization of transcription factor target genes strengthens coregulation in yeast. *Nucleic Acids Res.* 40, 27–36. doi: 10.1093/nar/gkr689

Dali, R., and Blanchette, M. (2017). A critical assessment of topologically associating domain prediction tools. *Nucleic Acids Res.* 45, 2994–3005. doi: 10.1093/nar/gkx145

Davis, C. A., Hitz, B. C., Sloan, C. A., Chan, E. T., Davidson, J. M., Gabdank, I., et al. (2018). The encyclopedia of dna elements (encode): data portal update. *Nucleic Acids Res.* 46, D794–D801. doi: 10.1093/nar/gkx1081

de Laat, W., and Grosveld, F. (2003). Spatial organization of gene expression: the active chromatin hub. *Chromosome Res.* 11, 447–459. doi: 10.1023/a:1024922626726

Dekker, J., Marti-Renom, M. A., and Mirny, L. A. (2013). Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. *Nat. Rev. Genet.* 14, 390–403. doi: 10.1038/nrg3454

Dekker, J., Rippe, K., Dekker, M., and Kleckner, N. (2002). Capturing chromosome conformation. *Science* 295, 1306–1311. doi: 10.1126/science.1067799

Denker, A., and Laat, W. de (2016). The second decade of 3C technologies: detailed insights into nuclear organization. *Genes Dev.* 30, 1357–1382. doi: 10.1101/gad.281964.116

Dixon, J. R., Jung, I., Selvaraj, S., Shen, Y., Antosiewicz-Bourget, J. E., Lee, A. Y., et al. (2015). Chromatin architecture reorganization during stem cell differentiation. *Nature* 518, 331–336. doi: 10.1038/nature14222

Dixon, J. R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., et al. (2012). Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* 485, 376–380. doi: 10.1038/nature11082

Djekidel, M. N., Chen, Y., and Zhang, M. Q. (2018). FIND: differential chromatin interactions detection using a spatial poisson process. *Genome Res.* 28, 412–422. doi: 10.1101/gr.212241.116

Dowen, J. M., Fan, Z. P., Hnisz, D., Ren, G., Abraham, B. J., Zhang, L. N., et al. (2014). Control of cell identity genes occurs in insulated neighborhoods in mammalian chromosomes. *Cell* 159, 374–387. doi: 10.1016/j.cell.2014.09.030

Dozmorov, M. G., Cara, L. R., Giles, C. B., and Wren, J. D. (2012). GenomeRunner: automating genome exploration. *Bioinformatics* 28, 419–420. doi: 10.1093/bioinformatics/btr666

Dozmorov, M. G., Cara, L. R., Giles, C. B., and Wren, J. D. (2016). GenomeRunner web server: regulatory similarity and differences define the functional impact of SNP sets. *Bioinformatics* 32, 2256–2263. doi: 10.1093/bioinformatics/btw169

Du, Z., Zheng, H., Huang, B., Ma, R., Wu, J., Zhang, X., et al. (2017). Allelic reprogramming of 3D chromatin architecture during early mammalian development. *Nature* 547, 232–235. doi: 10.1038/nature23263

Durand, N. C., Shamim, M. S., Machol, I., Rao, S. S. P., Huntley, M. H., Lander, E. S., et al. (2016). Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell Syst.* 3, 95–98. doi: 10.1016/j.cels.2016.07.002

Ernst, J., and Kellis, M. (2010). Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nat. Biotechnol.* 28, 817–825. doi: 10.1038/nbt.1662

Flavahan, W. A., Drier, Y., Liau, B. B., Gillespie, S. M., Venteicher, A. S., Stemmer-Rachamimov, A. O., et al. (2016). Insulator dysfunction and oncogene activation in IDH mutant gliomas. *Nature* 529, 110–114. doi: 10.1038/nature16490

Forcato, M., Nicoletti, C., Pal, K., Livi, C. M., Ferrari, F., and Bicciato, S. (2017). Comparison of computational methods for Hi-C data analysis. *Nat. Methods* 14, 679–685. doi: 10.1038/nmeth.4325

Franke, M., Ibrahim, D. M., Andrey, G., Schwarzer, W., Heinrich, V., Sch´opflin, R., et al. (2016). Formation of new chromatin domains determines pathogenicity of genomic duplications. *Nature* 538, 265–269. doi: 10.1038/nature19800

Fudenberg, G., Imakaev, M., Lu, C., Goloborodko, A., Abdennur, N., and Mirny, L. A. (2016). Formation of chromosomal domains by loop extrusion. *Cell Rep.* 15, 2038–2049. doi: 10.1016/j.celrep.2016.04.085

Golfier, S., Quail, T., Kimura, H., and Brugués, J. (2019). Cohesin and condensin extrude loops in a cell-cycle dependent manner. *bioRxiv* 821306. doi: 10.1101/821306

Gr`oschel, S., Sanders, M. A., Hoogenboezem, R., Wit, E. de, Bouwman, B. A. M., Erpelinck, C., et al. (2014). A single oncogenic enhancer rearrangement causes concomitant evi1 and gata2 deregulation in leukemia. *Cell* 157, 369–381. doi: 10.1016/j.cell.2014.02.019

Guo, Y., Xu, Q., Canzio, D., Shou, J., Li, J., Gorkin, D. U., et al. (2015). CRISPR inversion of CTCF sites alters genome topology and enhancer/promoter function. *Cell* 162, 900–910. doi: 10.1016/j.cell.2015.07.038

Hnisz, D., Weintraub, A. S., Day, D. S., Valton, A.-L., Bak, R. O., Li, C. H., et al. (2016). Activation of proto-oncogenes by disruption of chromosome neighborhoods. *Science* 351, 1454–1458. doi: 10.1126/science.aad9024

Hug, C. B., Grimaldi, A. G., Kruse, K., and Vaquerizas, J. M. (2017). Chromatin architecture emerges during zygotic genome activation independent of transcription. *Cell* 169, 216–228.e19. doi: 10.1016/j.cell.2017.03.024

Ibn-Salem, J., Ḱohler, S., Love, M. I., Chung, H.-R., Huang, N., Hurles, M. E., et al. (2014). Deletions of chromosomal regulatory boundaries are associated with congenital disease. *Genome Biol.* 15:423. doi: 10.1186/s13059-014-0423-1

Jackson, D. A., and Pombo, A. (1998). Replicon clusters are stable units of chromosome structure: evidence that nuclear organization contributes to the efficient activation and propagation of s phase in human cells. *J. Cell Biol.* 140, 1285–1295.

Jhunjhunwala, S., Zelm, M. C. van, Peak, M. M., and Murre, C. (2009). Chromatin architecture and the generation of antigen receptor diversity. *Cell* 138, 435–448. doi: 10.1016/j.cell.2009.07.016

Ji, X., Dadon, D. B., Powell, B. E., Fan, Z. P., Borges-Rivera, D., Shachar, S., et al. (2016). 3D chromosome regulatory landscape of human pluripotent cells. *Cell Stem Cell* 18, 262–275. doi: 10.1016/j.stem.2015.11.007

Jin, F., Li, Y., Dixon, J. R., Selvaraj, S., Ye, Z., Lee, A. Y., et al. (2013). A high-resolution map of the three-dimensional chromatin interactome in human cells. *Nature* 503, 290–294. doi: 10.1038/nature12644

Ke, Y., Xu, Y., Chen, X., Feng, S., Liu, Z., Sun, Y., et al. (2017). 3D chromatin structures of mature gametes and structural reprogramming during mammalian embryogenesis. *Cell* 170, 367–381.e20. doi: 10.1016/j.cell.2017.06.029

Krijger, P. H. L., and Laat, W. de (2016). Regulation of disease-associated gene expression in the 3D genome. *Nat. Rev. Mol. Cell Biol.* 17, 771–782. doi: 10.1038/nrm.2016.138

Lajoie, B. R., Dekker, J., and Kaplan, N. (2015). The Hitchhiker's guide to Hi-C analysis: practical guidelines. *Methods* 72, 65–75. doi: 10.1016/j.ymeth.2014.10.031

Li, G., Ruan, X., Auerbach, R. K., Sandhu, K. S., Zheng, M., Wang, P., et al. (2012). Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. *Cell* 148, 84–98. doi: 10.1016/j.cell.2011.12.014

Li, L., Lyu, X., Hou, C., Takenaka, N., Nguyen, H. Q., Ong, C.-T., et al. (2015). Widespread rearrangement of 3D chromatin organization underlies polycomb-mediated stress-induced silencing. *Mol. Cell* 58, 216–231. doi: 10.1016/j.molcel.2015.02.023

Lieberman-Aiden, E., Berkum, N. L. van, Williams, L., Imakaev, M., Ragoczy, T., Telling, A., et al. (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* 326, 289–293. doi: 10.1126/science.1181369

Lun, A. T. L., and Smyth, G. K. (2015). DiffHiC: a bioconductor package to detect differential genomic interactions in Hi-C data. *BMC Bioinformatics* 16:258. doi: 10.1186/s12859-015-0683-0

Lupiáñez, D. G., Spielmann, M., and Mundlos, S. (2016). Breaking tads: how alterations of chromatin domains result in disease. *Trends Genet.* 32, 225–237. doi: 10.1016/j.tig.2016.01.003

Ma, H., Samarabandu, J., Devdhar, R. S., Acharya, R., Cheng, P. C., Meng, C., et al. (1998). Spatial and temporal dynamics of DNA replication sites in mammalian cells. *J. Cell Biol.* 143, 1415–1425.

McLean, C. Y., Bristor, D., Hiller, M., Clarke, S. L., Schaar, B. T., Lowe, C. B., et al. (2010). GREAT improves functional interpretation of cis-regulatory regions. *Nat. Biotechnol.* 28, 495–501. doi: 10.1038/nbt.1630

Merelli, I., Liò, P., and Milanesi, L. (2013). NuChart: an R package to study gene spatial neighbourhoods with multi-omics annotations. *PLoS ONE* 8:e75146. doi: 10.1371/journal.pone.0075146

Mifsud, B., Tavares-Cadete, F., Young, A. N., Sugar, R., Schoenfelder, S., Ferreira, L., et al. (2015). Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. *Nat. Genet.* 47, 598–606. doi: 10.1038/ng.3286

Misteli, T. (2010). Higher-order genome organization in human disease. *Cold Spring Harb. Perspect. Biol.* 2:a000794. doi: 10.1101/cshperspect.a000794

Mitelman, F. (2000). Recurrent chromosome aberrations in cancer. *Mutat. Res.* 462, 247–253. doi: 10.1016/s1383-5742(00)00006-5

Mora, A., Sandve, G. K., Gabrielsen, O. S., and Eskeland, R. (2016). In the loop: promoter-enhancer interactions and bioinformatics. *Brief Bioinform.* 17, 980–995. doi: 10.1093/bib/bbv097

Nagano, T., Lubling, Y., Várnai, C., Dudley, C., Leung, W., Baran, Y., et al. (2017). Cell-cycle dynamics of chromosomal organization at single-cell resolution. *Nature* 547, 61–67. doi: 10.1038/nature23001

Narendra, V., Bulajić, M., Dekker, J., Mazzoni, E. O., and Reinberg, D. (2016). CTCF-mediated topological boundaries during development foster appropriate gene regulation. *Genes Dev.* 30, 2657–2662. doi: 10.1101/gad.288324.116

Naumova, N., Imakaev, M., Fudenberg, G., Zhan, Y., Lajoie, B. R., Mirny, L. A., et al. (2013). Organization of the mitotic chromosome. *Science* 342, 948–953. doi: 10.1126/science.1236083

Nora, E. P., Goloborodko, A., Valton, A.-L., Gibcus, J. H., Uebersohn, A., Abdennur, N., et al. (2017). Targeted degradation of CTCF decouples local insulation of chromosome domains from genomic compartmentalization. *Cell* 169, 930–944.e22. doi: 10.1016/j.cell.2017.05.004

Nora, E. P., Lajoie, B. R., Schulz, E. G., Giorgetti, L., Okamoto, I., Servant, N., et al. (2012). Spatial partitioning of the regulatory landscape of the x-inactivation centre. *Nature* 485, 381–385. doi: 10.1038/nature11049

Novo, C. L., Javierre, B.-M., Cairns, J., Segonds-Pichon, A., Wingett, S. W., Freire-Pritchett, P., et al. (2018). Long-range enhancer interactions are prevalent in mouse embryonic stem cells and are reorganized upon pluripotent state transition. *Cell Rep.* 22, 2615–2627. doi: 10.1016/j.celrep.2018.02.040

Osborne, C. S., Chakalova, L., Brown, K. E., Carter, D., Horton, A., Debrand, E., et al. (2004). Active genes dynamically colocalize to shared sites of ongoing transcription. *Nat. Genet.* 36, 1065–1071. doi: 10.1038/ng1423

Pal, K., Tagliaferri, I., Livi, C. M., and Ferrari, F. (2019). HiCBricks: building blocks for efficient handling of large Hi-C datasets. *Bioinformatics*. doi: 10.1093/bioinformatics/btz808

Papantonis, A., and Cook, P. R. (2013). Transcription factories: genome organization and gene regulation. *Chem. Rev.* 113, 8683–8705. doi: 10.1021/cr300513p

Phillips-Cremins, J. E., and Corces, V. G. (2013). Chromatin insulators: linking genome organization to cellular function. *Mol. Cell* 50, 461–474. doi: 10.1016/j.molcel.2013.04.018

Pope, B. D., Ryba, T., Dileep, V., Yue, F., Wu, W., Denas, O., et al. (2014). Topologically associating domains are stable units of replication-timing regulation. *Nature* 515, 402–405. doi: 10.1038/nature13986

Ramirez, F., Bhardwaj, V., Arrigoni, L., Lam, K. C., Gruning, B. A., Villaveces, J., et al. (2018). High-resolution TADs reveal DNA sequences underlying genome organization in flies. *Nat. Commun.* 9:189. doi: 10.1038/s41467-017-02525-w

Rao, S. S. P., Huang, S.-C., Glenn St Hilaire, B., Engreitz, J. M., Perez, E. M., Kieffer-Kwon, K.-R., et al. (2017). Cohesin loss eliminates all loop domains. *Cell* 171, 305–320.e24. doi: 10.1016/j.cell.2017.09.026

Rao, S. S. P., Huntley, M. H., Durand, N. C., Stamenova, E. K., Bochkov, I. D., Robinson, J. T., et al. (2014). A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* 159, 1665–1680. doi: 10.1016/j.cell.2014.11.021

Rickman, D. S., Soong, T. D., Moss, B., Mosquera, J. M., Dlabal, J., Terry, S., et al. (2012). Oncogene-mediated alterations in chromatin conformation. *Proc. Natl. Acad. Sci. U.S.A.* 109, 9083–9088. doi: 10.1073/pnas.1112570109

Sanborn, A. L., Rao, S. S. P., Huang, S.-C., Durand, N. C., Huntley, M. H., Jewett, A. I., et al. (2015). Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes. *Proc. Natl. Acad. Sci. U.S.A.* 112, E6456–E665. doi: 10.1073/pnas.1518552112

Sauerwald, N., Singhal, A., and Kingsford, C. (2020). Analysis of the structural variability of topologically associated domains as revealed by Hi-C. *NAR Genom Bioinf.* 2:lqz008. doi: 10.1093/nargab/lqz008

Sauria, M. E., Phillips-Cremins, J. E., Corces, V. G., and Taylor, J. (2015). HiFive: a tool suite for easy and efficient HiC and 5C data analysis. *Genome Biol.* 16:237. doi: 10.1186/s13059-015-0806-y

Schmitt, A. D., Hu, M., Jung, I., Xu, Z., Qiu, Y., Tan, C. L., et al. (2016). A compendium of chromatin contact maps reveals spatially active regions in the human genome. *Cell Rep.* 17, 2042–2059. doi: 10.1016/j.celrep.2016.10.061

Schoenfelder, S., Clay, I., and Fraser, P. (2010a). The transcriptional interactome: gene expression in 3D. *Curr. Opin. Genet. Dev.* 20, 127–133. doi: 10.1016/j.gde.2010.02.002

Schoenfelder, S., Sexton, T., Chakalova, L., Cope, N. F., Horton, A., Andrews, S., et al. (2010b). Preferential associations between co-regulated genes reveal a transcriptional interactome in erythroid cells. *Nat. Genet.* 42, 53–61. doi: 10.1038/ng.496

Serra, F., Baù, D., Goodstadt, M., Castillo, D., Filion, G. J., and Marti-Renom, M. A. (2017). Automatic analysis and 3D-modelling of Hi-C data using tadbit reveals structural features of the fly chromatin colors. *PLoS Comput. Biol.* 13:e1005665. doi: 10.1371/journal.pcbi.1005665

Sexton, T., and Cavalli, G. (2015). The role of chromosome domains in shaping the functional genome. *Cell* 160, 1049–1059. doi: 10.1016/j.cell.2015.02.040

Sexton, T., Yaffe, E., Kenigsberg, E., Bantignies, F., Leblanc, B., Hoichman, M., et al. (2012). Three-dimensional folding and functional organization principles of the drosophila genome. *Cell* 148, 458–472. doi: 10.1016/j.cell.2012.01.010

Shavit, Y., and Lio, P. (2014). Combining a wavelet change point and the bayes factor for analysing chromosomal interaction data. *Mol. Biosyst.* 10, 1576–1085. doi: 10.1039/c4mb00142g

Sheffield, N. C., and Bock, C. (2016). LOLA: Enrichment analysis for genomic region sets and regulatory elements in R and bioconductor. *Bioinformatics* 32, 587–589. doi: 10.1093/bioinformatics/btv612

Spielmann, M., Lupiáñez, D. G., and Mundlos, S. (2018). Structural variation in the 3D genome. *Nat. Rev. Genet.* 19, 453–467. doi: 10.1038/s41576-018-0007-0

Stansfield, J. C., Cresswell, K. G., and Dozmorov, M. G. (2019). MultiHiCcompare: joint normalization and comparative analysis of complex Hi-C experiments. *Bioinformatics* 35, 2916–2923. doi: 10.1093/bioinformatics/btz048

Stansfield, J. C., Cresswell, K. G., Vladimirov, V. I., and Dozmorov, M. G. (2018). HiCcompare: an R-package for joint normalization and comparison of Hi-C datasets. *BMC Bioinformatics* 19:279. doi: 10.1186/s12859-018-2288-x

Steensel, B. van (2011). Chromatin: constructing the big picture. *EMBO J.* 30, 1885–1895. doi: 10.1038/emboj.2011.135

Symmons, O., Uslu, V. V., Tsujimura, T., Ruf, S., Nassari, S., Schwarzer, W., et al. (2014). Functional and topological characteristics of mammalian regulatory domains. *Genome Res.* 24, 390–400. doi: 10.1101/gr.163519.113

Taberlay, P. C., Achinger-Kawecka, J., Lun, A. T. L., Buske, F. A., Sabir, K., Gould, C. M., et al. (2016). Three-dimensional disorganization of the cancer genome occurs coincident with long-range genetic and epigenetic alterations. *Genome Res.* 26, 719–731. doi: 10.1101/gr.201517.115

Tang, Z., Luo, O. J., Li, X., Zheng, M., Zhu, J. J., Szalaj, P., et al. (2015). CTCF-mediated human 3D genome architecture reveals chromatin topology for transcription. *Cell* 163, 1611–1627. doi: 10.1016/j.cell.2015.11.024

Tanizawa, H., Iwasaki, O., Tanaka, A., Capizzi, J. R., Wickramasinghe, P., Lee, M., et al. (2010). Mapping of long-range associations throughout the fission yeast genome reveals global genome organization linked to transcriptional regulation. *Nucleic Acids Res.* 38, 8164–8177. doi: 10.1093/nar/gkq955

Ursu, O., Boley, N., Taranova, M., Wang, Y. R., Yardimci, G. G., Stafford Noble, W., et al. (2018). GenomeDISCO: a concordance score for chromosome conformation capture experiments using random walks on contact map graphs. *Bioinformatics.* 34, 2701–2707. doi: 10.1093/bioinformatics/bty164

Valton, A.-L., and Dekker, J. (2016). TAD disruption as oncogenic driver. *Curr. Opin. Genet. Dev.* 36, 34–40. doi: 10.1016/j.gde.2016.03.008

Vietri Rudan, M., Barrington, C., Henderson, S., Ernst, C., Odom, D. T., Tanay, A., et al. (2015). Comparative Hi-C reveals that CTCF underlies evolution of chromosomal domain architecture. *Cell Rep.* 10, 1297–1309. doi: 10.1016/j.celrep.2015.02.004

Wang, H., Duggal, G., Patro, R., Girvan, M., Hannenhalli, S., and Kingsford, C. (2013). "Topological properties of chromosome conformation graphs reflect spatial proximities within chromatin," in *Proceedings of the International Conference on Bioinformatics, Computational Biology and Biomedical Informatics, BCB'13*. (New York, NY: ACM), 306–315. doi: 10.1145/2506583.2506633

Wang, Y. R., Sarkar, P., Ursu, O., Kundaje, A., and Bickel, P. J. (2019). Network modelling of topological domains using Hi-C data. *Ann. Appl. Stat.* 13, 1511–1536. doi: 10.1214/19-AOAS1244

Yaffe, E., and Tanay, A. (2011). Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture. *Nat. Genet.* 43, 1059–1065. doi: 10.1038/ng.947

Yan, K.-K., Yardimci, G. G., Yan, C., Noble, W. S., and Gerstein, M. (2017). HiC-spector: a matrix library for spectral and reproducibility analysis of Hi-C contact maps. *Bioinformatics* 33, 2199–2201. doi: 10.1093/bioinformatics/btx152

Yang, T., Zhang, F., Yardımcı, G. G., Song, F., Hardison, R. C., Noble, W. S., et al. (2017). HiCRep: assessing the reproducibility of Hi-C data using a stratum-adjusted correlation coefficient. *Genome Res.* 27, 1939–1949. doi: 10.1101/gr.220640.117

Yardimci, G. G., Ozadam, H., Sauria, M. E., Ursu, O., Yan, K. K., Yang, T., et al. (2019). Measuring the reproducibility and quality of Hi-C data. *Genome Biol.* 20:57. doi: 10.1186/s13059-019-1658-7

Yu, M., and Ren, B. (2017). The three-dimensional organization of mammalian genomes. *Annu. Rev. Cell Dev. Biol.* 33, 265–289. doi: 10.1146/annurev-cellbio-100616-060531

Zaborowski, R., and Wilczynski, B. (2016). DiffTAD: detecting differential contact frequency in topologically associating domains Hi-C experiments between conditions. *bioRxiv* 093625. doi: 10.1101/093625

Zhang, Y., Xiang, Y., Yin, Q., Du, Z., Peng, X., Wang, Q., et al. (2018). Dynamic epigenomic landscapes during early lineage specification in mouse embryos. *Nat. Genet.* 50, 96–105. doi: 10.1038/s41588-017-0003-x

Zhou, Y., Gerrard, D. L., Wang, J., Li, T., Yang, Y., Fritz, A. J., et al. (2019). Temporal dynamic reorganization of 3D chromatin architecture in hormone-induced breast cancer and endocrine resistance. *Nat. Commun.* 10:1522. doi: 10.1038/s41467-019-09320-9

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.

frontiers in Genetics

ORIGINAL RESEARCH
published: 24 April 2020
doi: 10.3389/fgene.2020.00385

Integrating Genome-Wide Association Studies and Gene Expression Profiles With Chemical-Genes Interaction Networks to Identify Chemicals Associated With Colorectal Cancer

*Xinyue Tan, Hanmin Tang, Liuyun Gong, Lina Xie, Yutiantian Lei, Zhenzhen Luo, Chenchen He, Jinlu Ma and Suxia Han**

Department of Oncology, The First Affiliated Hospital, Xi'an Jiaotong University, Xi'an, China

OPEN ACCESS

Edited by:
Ryan Matthew Layer,
University of Colorado Boulder,
United States

Reviewed by:
Shihao Shen,
University of California, Los Angeles,
United States
Lei Zhang,
Soochow University, China

****Correspondence:***
Suxia Han
shan87@mail.xjtu.edu.cn

Specialty section:
This article was submitted to Bioinformatics and Computational Biology,
a section of the journal Frontiers in Genetics

Received: *18 July 2019*
Accepted: *27 March 2020*
Published: *24 April 2020*

Citation:
Tan X, Tang H, Gong L, Xie L, Lei Y, Luo Z, He C, Ma J and Han S (2020) Integrating Genome-Wide Association Studies and Gene Expression Profiles With Chemical-Genes Interaction Networks to Identify Chemicals Associated With Colorectal Cancer.
Front. Genet. 11:385.
doi: 10.3389/fgene.2020.00385

Colorectal cancer (CRC) is the third most common cancer and has the second highest mortality rate in global cancer. Exploring the associations between chemicals and CRC has great significance in prophylaxis and therapy of tumor diseases. This study aims to explore the relationships between CRC and environmental chemicals on genetic basis by bioinformatics analysis. The genome-wide association study (GWAS) datasets for CRC were obtained from the UK Biobank. The GWAS data for colon cancer (category C18) includes 2,581 individuals and 449,683 controls, while that of rectal cancer (category C20) includes 1,244 individuals and 451,020 controls. In addition, we derived CRC gene expression datasets from the NCBI-GEO (GSE106582). The chemicals related gene sets were acquired from the comparative toxicogenomics database (CTD). Transcriptome-wide association study (TWAS) analysis was applied to CRC GWAS summary data and calculated the expression association testing statistics by FUSION software. We performed chemicals related gene set enrichment analysis (GSEA) by integrating GWAS summary data, mRNA expression profiles of CRC and the CTD chemical-gene interaction networks to identify relationships between chemicals and genes of CRC. We observed several significant correlations between chemicals and CRC. Meanwhile, we also detected 5 common chemicals between colon and rectal cancer, including methylnitronitrosoguanidine, isoniazid, PD 0325901, sulindac sulfide, and importazole. Our study performed TWAS and GSEA analysis, linked prior knowledge to newly generated data and thereby helped identifying chemicals related to tumor genes, which provides new clues for revealing the associations between environmental chemicals and cancer.

Keywords: colorectal cancer, genome-wide association study, transcriptome-wide association study, gene set enrichment analysis, comparative toxicogenomics database

INTRODUCTION

Colorectal cancer (CRC) is the third most common cancer worldwide and has the second highest mortality rate in global cancer (Bray et al., 2018; Ferlay et al., 2019). In western countries, CRC accounts for about 10% of cancer deaths (Kuipers et al., 2015). The accepted view is that genetic, lifestyle, and environmental factors are closely related to CRC (Dekker et al., 2019). Current research shows that environmental chemicals play important roles in the etiology of CRC. Several chemicals have been suggested to promote the tumorigenesis and development of CRC. For instance, analysis of an Iowa Women's Health Study cohort suggested that exposure to TTHM in drinking water is associated with increased risk of rectal cancer (Jones et al., 2019). In addition, another case-control study observed that organochlorine and organophosphorus pesticides may induce CRC (Abolhassani et al., 2019). In contrast, numerous chemicals have been identified that inhibit CRC. Metastatic CRC (mCRC) often indicates a poor prognosis. The 5-year overall survival (OS) rate of patients with mCRC is less than 15% (Siegel et al., 2017; Bray et al., 2018), and the median OS of unresectable mCRC patients who received only supportive therapy was only 5 months (Lucas et al., 2011). However, the 5-years OS rate increased to 10% in such patients receiving 5-fluorouracil (5-FU)/leucovorin (LV) plus oxaliplatin (FOLFOX) (Gustavsson et al., 2015). Thus, FOLFOX chemotherapy regimen is still the standard first-line therapy for unresectable mCRC (Giacchetti et al., 2000; Goldberg et al., 2004; Kouhara et al., 2007; Bokemeyer et al., 2011). Recent studies have found that some non-chemotherapeutic chemicals also have an inhibitory effect on CRC, such as semisynthetic retinoid, lidocaine, and beta-carotene (Mattingly et al., 2003; Pham et al., 2013; Qu et al., 2018).

Therefore, it has great significance to clarify the relationship between chemicals in environmental and CRC for the treatment and prevention of diseases. But obtaining the entire life-time exposure of an individual is difficult and complex, for lacking sensitive methods to measure specific exposures. Although the exposure is known to have occurred, the transient character of the exposure indicators increases the difficulty of measuring the specific exposure (Messerlian et al., 2017). With the help of new technologies, such as genome-wide association research (GWAS), more convenient and efficient analyses have been produced to identify interactions between multiple environmental exposures with genes (Khoury et al., 2005). Studies of gene-environment interactions have been widely applied in psychological research, however, lack in the field of oncology (Manuck and McCaffery, 2014). The Comparative Toxicogenomics Database (CTD) is a public repository, aims to advance people's understanding of how environmental exposures affects human health (Mattingly et al., 2003). This database provides information regarding chemical-gene/protein interactions as well as chemical- and gene-disease relationships that is organized by individual genes, gene sets, organisms, chemicals, sequence type (DNA, mRNA, and protein), gene ontology annotations and sequences (Mattingly et al., 2006).

Genome-wide association studies (GWAS) analyze DNA sequence variations to provide associations for complex human traits and diseases efficiently (Tam et al., 2019). Transcriptome-wide association studies (TWAS) is further developed on this basis, which can evaluate the association of each gene to diseases by integrating tissue-related gene expression measurements with GWAS summary data (Gong et al., 2018). Currently, TWAS has been proved with high efficiency in determining the genetic mechanism of complex diseases (Gusev et al., 2018; Wu et al., 2018). The Gene Expression Omnibus (GEO) is a worldwide resource which distributes a large number of high-throughput microarray and next-generation sequence functional genomic data sets (Barrett et al., 2013). Different from the traditional GWAS to explain the relationship between DNA and external phenotype, we simultaneously used the GEO to obtain the gene expression profile (mRNA expression profile chip data) of colorectal cancer, that is, a comprehensive analysis at the DNA and mRNA level. This is helpful to narrow the range of chemicals related candidate genes on the basis of traditional GWAS analysis.

Briefly, in this work, the CTD chemical-gene interaction networks, GWAS summary datasets and gene expression profiles were integrated. TWAS analysis was performed by FUSION software to evaluate the expression association testing statistics. The gene set enrichment analysis (GSEA) with the running-sum statistic and weighted Kolmogorov-Smirnov–like statistic were applied to detect the correlation between environmental chemicals and CRC (Charmpi and Ycart, 2015). Firstly, we obtained the empirical distributions of GSEA statistics for each chemical for statistical tests. Subsequently, the *P*-value of each chemical was conducted from the permuted empirical distribution of GSEA statistics. Finally, we summarized and analyzed the obtained chemicals associated with CRC.

MATERIALS AND METHODS

GWAS Summary Dataset for CRC

GeneATLAS[1], a huge resource storing the information of hundreds of traits and millions of related gene variants based on the UK Biobank cohort, provides a convenient way for researchers to acquire data from the UK Biobank (Lin et al., 2019). To be specific, it allows researchers to query genome-wide association results for 9,113,133 genetic variants and download over 30 million genetic variants (>23 billion phenotype-genotype pairs) for GWAS summary statistics (Canela-Xandri et al., 2018).

A large-scale GWAS summary data of colon cancer and rectal cancer in our study were downloaded from the GeneATLAS in UK Biobank. In the cancer register category, 103,470 data items are available from 84,726 participants. In brief, our GWAS summary data, which contained 5,899 available data items, were from categories C18 (malignant neoplasm of colon) and C20 (malignant neoplasm of rectum). Detailed information regarding the methods, process, and approaches were described in the previous studies (Hammerschlag et al., 2017).

[1] http://geneatlas.roslin.ed.ac.uk/

Gene Expression Datasets of CRC

NCBI-GEO[2] is an international public repository with next-generation sequencing and microarray/gene profiles which was used in this study to obtain the mRNA expression profiles of mucosa and colorectal tumor tissues (GSE106582). CRC patients were recruited at the University Hospital of Heidelberg, from whom the gene expression profiles of 77 tumor and 117 mucosa tissues were obtained using an Illumina HumanHT-12 V4.0 expression beadchip. Using GEO2R, a web tool based on the GEO database, differential gene expression was assessed by comparing the expression of genes from colorectal tumor tissues to those of respective mucosa tissues.

Transcriptome-Wide Association Study (TWAS) Methodology

TWAS analysis utilizes disease GWAS summary statistics combining with pre-computed gene expression weights to calculate the association of every gene with known diseases (Gusev et al., 2018). In other words, TWAS can integrate the associations between GWAS and gene expression measurements to identify genes associated with traits. In this study, A TWAS for CRC was conducted using functional summary-based imputation (FUSION) software and the gene expression weight references of whole blood, rectum, and colon tissues were acquired from the FUSION website[3]. Specifically, the gene expression weights of whole blood were collected from 1,264 subjects of the Young Finns Study (Raitakari et al., 2008; Nuotio et al., 2014).

Firstly, based on FUSION software we performed prediction models to calculate the gene expression weights of different tissues (Huang et al., 2019). Then we conducted the correlation statistics between gene expressed level and CRC combining the gene expression weights and summary-level GWAS results. Z_{TWAS} = w'Z/(w'Lw)$^{1/2}$ was used to calculate the association statistics. Z denotes the scores of CRC while w denotes the weights. L means the SNP-correlation linkage disequilibrium (LD) matrix. A TWAS *p*-value was calculated for each gene within whole blood, rectum and colon tissues, respectively (Qi et al., 2019). The genes with $p < 0.05$ were considered as significant. Detailed information can be found in the published study (Gusev et al., 2018).

Chemical-Gene Expression Interaction Database of the CTD Database

The Comparative Toxicogenomics Database (CTD)[4] is a publicly and accessible database for toxicogenomic information (Zhang et al., 2019). The CTD currently includes more than 30.5 million toxicogenomic relationships associated with chemicals, proteins, etc. (Davis et al., 2017) and provides information regarding chemical, gene, phenotype, and disease relationships to advance our understanding of the effects of environmental toxin exposure on public health (Grondin et al., 2018). A unique and powerful feature of the CTD is knowledge transfer with respect to any information that is directly annotated to chemicals, genes and diseases (Davis et al., 2013). This study download 11,190 chemicals related gene sets from the CTD. The process of retrieving information using CTD was described in the study previously (Mattingly et al., 2006).

Identification of Environmental Chemicals Elements Associated With Colorectal Cancer

The GSEA algorithm was originally used for microarray study and GWAS-based GSEA was developed subsequently (He et al., 2018). At present, it is utilized to identify abnormally expressed gene sets for target diseases, and has been applied in etiology researches of multiple diseases (Wang et al., 2007). Firstly, for the *j*th (j = 1,2,3…*N*) gene, the most significant GWAS association test statistics of the SNPs was assigned to *j*th gene according to the score r_j of the given gene. Secondly, all genes G = (G_{1^*}, G_{2*},…,G_{N^*}) were ranked by their scores from the highest to the lowest (He et al., 2018), which was expressed as U = (j_{1^*}, j_{2^*},…,j_{N^*}). Thirdly, for a chemicals related gene set *S*, an enrichment score *ES* was calculated for CRC by the running sum statistic and weighted Kolmogorov-Smirnov-like statistic (Subramanian et al., 2005; Charmpi and Ycart, 2015). Gene set S independently derived from N_H genes. *ES* represents the overrepresentation of CRC associated genes in chemicals related gene set S. *ES* was calculated as:

$$\boldsymbol{ES(S)} = \max_{1\le j\le N}\left\{\sum_{G_{j^*}\in S, j^*\le j}\frac{|r_{j^*}|^p}{N_R} - \sum_{G_{j^*}\notin S, j^*\le j}\frac{1}{N-N_H}\right\},$$

where

$$N_R = \sum_{G_j\in S}|r_{j^*}|^p.$$

Finally, after L time permutations, we can obtain the null distribution of $ES^{null} = (ES_1^{null}, ES_2^{null}, \ldots, ES_l^{null})$. To control the effect of the gene sets with varying sizes, the observed *ES*(*S*) is normalized by the average value and standard deviation of the permutated ES_S^{null}, defined by $NES^S = \frac{ES^S - mean(ES_S^{null})}{SD(ES_S^{null})}$. The *P*-values were finally calculated from the NES for each chemicals related gene set.

This study conducted a total of 5,000 permutations to calculate the empirical distributions of GSEA statistics of each chemical. And the chemicals related gene sets with $P < 0.05$ are considered statistically significant. Previous research provides the detailed descriptions regarding this approach (Zhao et al., 2018). Similarly, all mRNA expression profile from GEO were analyzed using the same approach (Weng et al., 2011).

RESULTS

Environmental Chemicals Associated With Colorectal Cancer

From the CRC GWAS summary datasets, we identified 175 chemicals that were significantly associated with colon cancer

[2] https://www.ncbi.nlm.nih.gov/gds/

[3] http://gusevlab.org/projects/fusion/

[4] http://ctdbase.org/

(including 34 for colon tissue and 141 for whole blood) as well as 103 chemicals significantly associated with rectal cancer (including 20 for rectal tissue and 83 for whole blood) ($P < 0.05$; **Supplementary Tables S1, S2**). For the expression profile of CRC, we identified 1,198 significant chemicals ($P < 0.05$; **Supplementary Table S3**).

After a comparative analysis of the TWAS and mRNA expression profile GSEA results, we significantly detected several chemicals associated with the colon cancer and rectal cancer ($P < 0.05$). For colon cancer, 104 common chemicals were detected, including 83 in colon tissue and 24 in blood tissue, and 3 in both tissues (**Supplementary Table S4**),such as Antirheumatic Agents (*P*-value1 = 0.0244, *P*-value2 = 0.0230), Chenodeoxycholic Acid (*P*-value1 = 0.0002, *P*-value2 = 0.0002) and Trientine (*P*-value1 = 0.0464, *P*-value2 = 0.0314; Supplementary note: In this paragraph, *P*-value1 is *P*-value in GWAS dataset and *P*-value2 is *P*-value in mRNA expression profile). For rectal cancer, 51 common chemicals were discovered, including 12 in rectum tissue and 39 in blood tissue (**Supplementary Table S5**). **Tables 1**, **2** summarized the top 10 chemicals identified for the colon cancer and rectal cancer separately.

TABLE 1 | List of top ten chemicals identified for colon cancer after a comparative of GWAS and mRNA GSEA results.

Chemical Name	*P*-value1[a]	*P*-value2[b]
Antirheumatic Agents	0.0002	0.0002
LG 100815	0.0004	0.0004
Zinc Acetate	0.0010	0.0004
Aerosols	0.0016	0.0002
Titanium dioxide	0.0026	0.0002
Motexafin gadolinium	0.0046	0.0006
Clofibric Acid	0.0052	0.0002
Vitallium	0.0052	0.0002
Raloxifene Hydrochloride	0.0066	0.0002
Soman	0.0094	0.0002

[a]P-value1: P-value in GWAS dataset. [b]P-value2: P-value in mRNA expression profile.

TABLE 2 | List of top ten chemicals identified for rectal cancer after a comparative of GWAS and mRNA GSEA results.

Chemical name	*P*-value1[a]	*P*-value2[b]
NAD	0.0020	0.0004
Sulindac sulfide	0.0052	0.0002
Casticin	0.0086	0.0002
Benz(a)anthracene	0.0124	0.0002
Methylnitronitrosoguanidine	0.0132	0.0002
Afimoxifene	0.0134	0.0002
4-phenylbutyric acid	0.0150	0.0004
Nickel	0.0178	0.0002
Ochratoxin A	0.0180	0.0002
Promethazine	0.0196	0.0006

[a]P-value1: P-value in GWAS dataset. [b]P-value2: P-value in mRNA expression profile.

TABLE 3 | The common significant chemicals between colon cancer and rectal cancer GSEA results.

Chemical name	*P*-value1[a]	*P*-value2[b]	*P*-value3[c]
Methylnitronitrosoguanidine	0.0394	0.0132	0.0002
Isoniazid	0.0164	0.0262	0.0068
PD 0325901	0.0348	0.0406	0.0012
Sulindac sulfide	0.0374	0.0052	0.0002
Importazole	0.0378	0.0450	0.0224

[a]P-value1: P-value in colon GWAS dataset. [b]P-value2: P-value in rectal GWAS dataset. [c]P-value3: P-value in mRNA expression profile.

Meanwhile, **Table 3** shows the common significant environmental chemicals between colon cancer and rectal cancer. We detected 5 chemicals, including methylnitronitrosoguanidine (*P*-value1 = 0.0394, *P*-value2 = 0.0132, *P*-value3 = 0.0002), isoniazid (*P*-value1 = 0.0164, *P*-value2 = 0.0262, *P*-value3 = 0.0068), PD 0325901 (*P*-value1 = 0.0348, *P*-value2 = 0.0406, *P*-value3 = 0.0012), sulindac sulfide (*P*-value1 = 0.0374, *P*-value2 = 0.0052, *P*-value2 = 0.0002), importazole (*P*-value1 = 0.0378, *P*-value2 = 0.0450, *P*-value3 = 0.0224; Supplementary note: In this paragraph, *P*-value1 is *P*-value in colon GWAS dataset, *P*-value2 is *P*-value in rectal GWAS dataset, *P*-value3 is *P*-value in mRNA expression profile). The specific technology roadmap and Venn diagram are shown in **Figure 1**.

DISCUSSION

CRC is the fourth deadliest cancer lead to 900,000 deaths worldwide annually (Dekker et al., 2019). It has become a global public health problem due to its high morbidity and mortality worldwide. Both genetic and environmental factors play significant roles in the etiology of colorectal cancer. Cancer risk factors include biological agents (infection), exposure to synthetic chemicals, and lifestyle factors, which together contribute to the development of 70–95% of cancers (Wu et al., 2016). Several chemicals have been reported promote the tumorigenesis and tumor development of CRC (Abolhassani et al., 2019; Cernigliaro et al., 2019; Jones et al., 2019). This provides a new clue for us to prevent the occurrence of colorectal cancer. Meanwhile, except for the standard treatment, many chemicals have been reported to inhibit CRC in recent years. For example, the anti-colorectal cancer effect of awsonaringenin (LSG), a flavonoid compound, has been demonstrated in previous research (Anwar et al., 2018). Environmental chemicals are related to various malignant tumors besides CRC. For example, acrylamide, benzo(a)pyrene and polychlorinated biphenyls can induce carcinogenesis for cytotoxicity and DNA damage to hepatic cells (Erkekoglu et al., 2017). The discovery for the active substance in chemicals related cancer is of great significance for the treatment to tumor patients. Since the chemicals environmental exposure is usually complex and accurately measuring exposure levels *in vivo* is still with many objective problems, we try to explore the relationships between chemicals and cancer in an easier way.

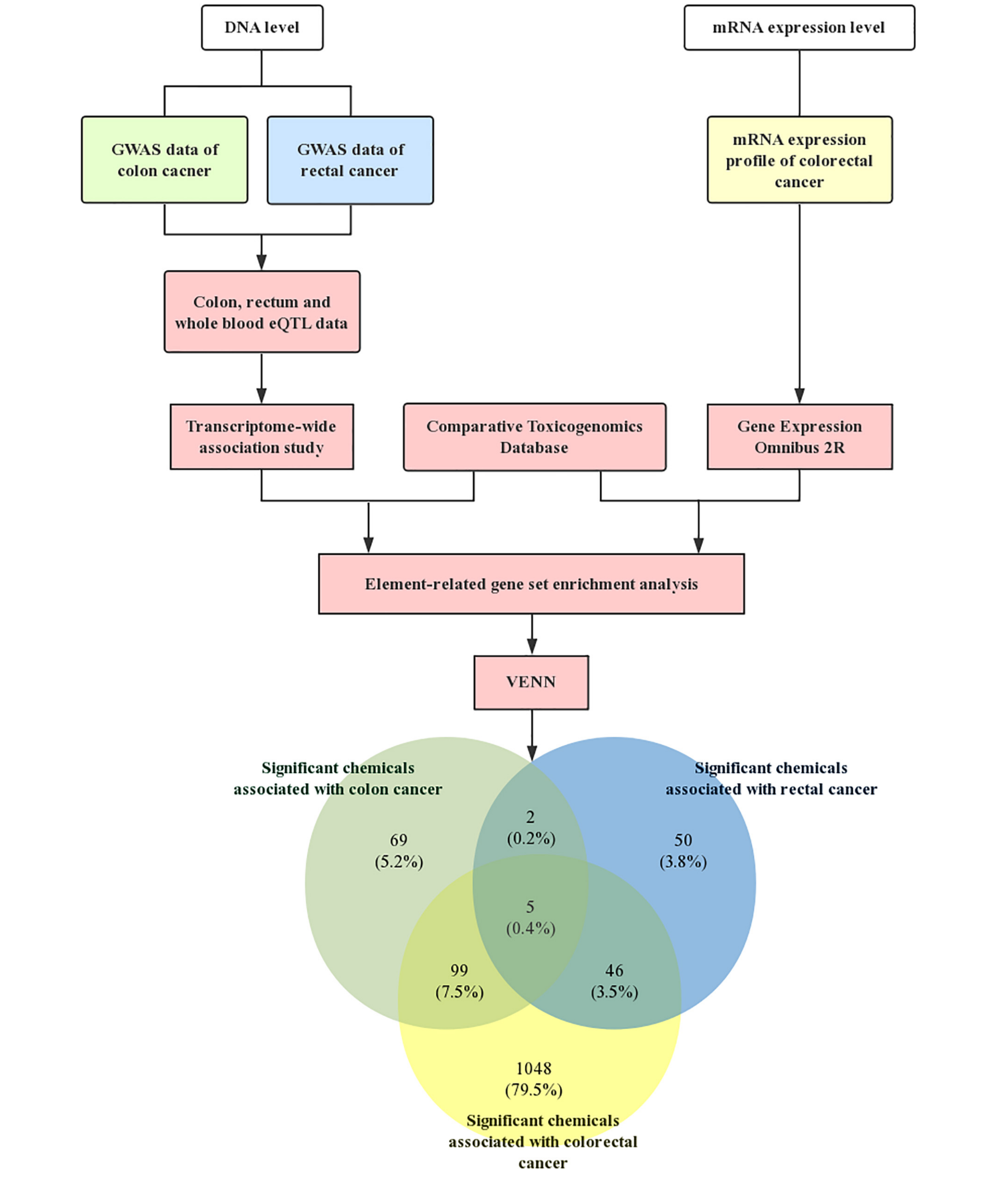

DNA level
mRNA expression level
GWAS data of colon cacner
GWAS data of rectal cancer
mRNA expression profile of colorectal cancer
Colon, rectum and whole blood eQTL data
Transcriptome-wide association study
Comparative Toxicogenomics Database
Gene Expression Omnibus 2R
Element-related gene set enrichment analysis
VENN
Significant chemicals associated with colon cancer
Significant chemicals associated with rectal cancer
69 (5.2%)
2 (0.2%)
50 (3.8%)
5 (0.4%)
99 (7.5%)
46 (3.5%)
1048 (79.5%)
Significant chemicals associated with colorectal cancer

FIGURE 1 | Technology roadmap. First, the GWAS dataset of colon cancer and rectal cancer were downloaded from GeneATLAS, a large database based on the UK Biobank cohort. Meanwhile, we obtained mRNA expression profiles of CRC from NCBI-GEO. The software FUSION was used to assess the CRC GWAS summary data for tissue-related TWAS analysis. The chemicals related gene sets were then generated by the CTD. Subsequently, chemical-related gene set enrichment analysis (GSEA) was conducted to detect the association between chemicals and CRC. Finally, the Venn diagram showed the significant chemicals associated with colorectal cancer.

In this study, we extended the classical GSEA approach to detect associations between chemicals and CRC using TWAS data and gene expression datasets. We identified several chemicals showing genetic correlation evidence with the CRC.

We identified several significant chemicals for the colon cancer, such as aspirin and titanium dioxide, which have been reported by previous study. Aspirin, a well-known antirheumatic drug, is proved that can prolong the survival of patients with colorectal cancer and activate T cell-mediated antitumor immunity (Hamada et al., 2017). Bettini, Boutet-Robinet et al. has reported that daily oral food-grade titanium dioxide (TiO2) intake is related to an chronic intestinal inflammation and will increase the risk of carcinogenesis (Bettini et al., 2017).

NAD and Nickel are two remarkable chemicals associated with rectal cancer. A recent study revealed that increased nicotinamide adenine dinucleotide pool suppressed reactive oxygen species level to promote progression of colon cancer (Hong et al., 2019). In a previous study, trace elements in normal and cancerous tissue which obtained from 18 patients suffering from colon and rectum cancer were quantitatively determined by X-ray fluorescence, and the result showed that Nickel elevated in cancerous tissues (Gregoriadis et al., 1983).

Five overlapped chemicals have been identified associated with CRC, including the carcinogens methylnitronitrosoguanidine, isoniazid. And PD 0325901, sulindac sulfide and importazole have the ability to inhibit the carcinogenesis and development of cancer.

Methylnitronitrosoguanidine (MNNG) is anticipated to be declared a human carcinogen based on sufficient evidence of its carcinogenicity from investigations involving animal models. MNNG caused tumors at different tissue sites in several animal model species by several different exposure routes. Research indicated that the intrarectal infusion of MNNG into large intestine of rats can cause tumors (Tsukamoto et al., 2015; U.S. Department of Health and Human Services, 2016).

Isoniazid (INH) is an irreversible inhibitor of Monoamine oxidase A (MAOA) that is widely regarded as a major anti-tuberculosis drug (Zareifopoulos and Panayiotakopoulos, 2017). MAOA is a mitochondrial-bound enzyme. It was confirmed that MAOA may promote the progression of prostate cancer by mediating EMT (Wu et al., 2014; Lv et al., 2018). However, because conflicting results have been reported for the importance of MAOA in HCC and cholangiocarcinoma (Huang et al., 2012; Li et al., 2014), the role of MAOA may vary across cancer types. Lee et al. demonstrated that Monoamine Oxidase Inhibitors (MAOIs) are associated with increased colorectal cancer risk (adjusted OR = 1.22, 95% CI = 1.06-1.41; Lee et al., 2017).

PD 0325901 is an MEK inhibitor. Interestingly, Roper et al. (2014) have shown PI3K/MEK inhibition combined with NVP-BKM120 and PD-0325901 treatment can induce tumor progression in a wild-type PIK3CA mouse model, KRAS mutant CRC, based on the inhibition of mTORC1 and MCL-1 and the activation of BIM. Moreover, PD0325901 was reported to inhibit oxaliplatin-induced neuropathy and enhance oxaliplatin efficacy (Tsubaki et al., 2015).

Liggett et al. observed that the non-steroidal anti-inflammatory drug sulindac sulfide inhibits the expression of the potential oncogene structural protein nesprin-2 in CRC cells (Liggett et al., 2014). The results of another study suggested the inhibition of sulindac sulfide on the growth of colon cancer cells and down-regulation of specific transcription factors (Li et al., 2015). Furthermore, the inhibitory effects of 5-fluorouracil and oxaliplatin on human CRC cell survival were demonstrated to be synergistically enhanced by sulindac sulfide (Flis and Splwinski, 2009).

Importazole is a small molecule inhibitor of the transport receptor importin-β (Soderholm et al., 2011) that can inhibit the proliferation and induce apoptosis of multiple myeloma cells by blocking the NF-KB signaling pathway (Yan et al., 2015). Moreover, intravenous administration of the specific KPNB1 inhibitor importazole was effective in reducing the volume and weight of prostate cancer tumor in mice inoculated with PC3 PCa cells (Yang et al., 2019). Thus, the results of the above studies show that importazole can inhibit tumors.

We conducted a large scale correlation study between colorectal cancer and environmental chemicals and explored the associations between chemicals and colorectal cancer systematically. Our analysis approach has two advantages. Firstly, we identified interaction between chemicals and genes directly. From the perspective of genome, the result is more stable to overcome the shortcomings of traditional exposure measurement methods. From the perspective of benefit, genome-wide summary data usually can be obtained online conveniently. Secondly, our research analyzed summaries of TWAS and mRNA expression profiles, in other words, we made a comprehensive analysis in the DNA and mRNA expression levels. This is helpful to narrow the range of chemicals related candidate genes on the basis of traditional GWAS analysis and make the results more reliable. Current research shows that chemicals in environmental factors have great significance in the etiology of multiple cancers (Thompson et al., 2015). However, we only researched the colon cancer and rectal cancer. As cancer sequencing gene data sets increasing, we will apply our method to large-scale studies of cancer gene-environment interactions.

In summary, we conducted an integrative analysis of GWAS summary data, mRNA expression profiles and chemical-gene interaction networks. Tools such as TWAS and GSEA helped linking these datasets and identifying several chemicals associated with CRC. The results of our study evaluate the associations between CRC and chemicals systematically, and provide new clues for revealing the association between chemicals and genes and their effects on cancer. Furthermore, our method can be used to analyze other chemicals and complex malignant disease, which is helpful for assessing the relationship between environmental exposure and cancer.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: https://www.ncbi.nlm.nih.gov/gds/

(GSE106582), http://geneatlas.roslin.ed.ac.uk/ (categories C18, categories C20), http://gusevlab.org/projects/fusion/, http://ctdbase.org/.

AUTHOR CONTRIBUTIONS

XT and SH designed experiments. XT, HT, ZL, and YL reviewed and downloaded the original data. XT, LG, and LX processed and analyzed the data. XT, CH, and JM analyzed experimental results. XT, HT, and SH wrote the manuscript.

FUNDING

This work was supported by grants from the Innovation Capacity Support Plan of Shaanxi Province 2018TD−002.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.00385/full#supplementary-material

REFERENCES

Abolhassani, M., Asadikaram, G., Paydar, P., Fallah, H., Aghaee-Afshar, M., Moazed, V., et al. (2019). Organochlorine and organophosphorous pesticides may induce colorectal cancer; A case-control study. *Ecotoxicol. Environ. Saf.* 178, 168–177. doi: 10.1016/j.ecoenv.2019.04.030

Anwar, A., Uddin, N., Siddiqui, B. S., Siddiqui, R. A., Begum, S., and Choudhary, M. I. (2018). A natural flavonoid lawsonaringenin induces cell cycle arrest and apoptosis in HT-29 colorectal cancer cells by targeting multiple signalling pathways. *Mol. Biol. Rep.* 45, 1339–1348. doi: 10.1007/s11033-018-4294-5

Barrett, T., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., Tomashevsky, M., et al. (2013). NCBI GEO: archive for functional genomics data sets–update. *Nucleic Acids Res.* 41, D991–D995. doi: 10.1093/nar/gks1193

Bettini, S., Boutet-Robinet, E., Cartier, C., Comera, C., Gaultier, E., Dupuy, J., et al. (2017). Food-grade TiO2 impairs intestinal and systemic immune homeostasis, initiates preneoplastic lesions and promotes aberrant crypt development in the rat colon. *Sci. Rep.* 7:40373. doi: 10.1038/srep40373

Bokemeyer, C., Bondarenko, I., Hartmann, J., De Braud, F., Schuch, G., Zubel, A., et al. (2011). Efficacy according to biomarker status of cetuximab plus FOLFOX-4 as first-line treatment for metastatic colorectal cancer: the OPUS study. *Ann. Oncol.* 22, 1535–1546. doi: 10.1093/annonc/mdq632

Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R. L., Torre, L. A., and Jemal, A. (2018). Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J. Clin.* 68, 394–424. doi: 10.3322/caac.21492

Canela-Xandri, O., Rawlik, K., and Tenesa, A. (2018). An atlas of genetic associations in UK Biobank. *Nat. Genet.* 50, 1593–1599. doi: 10.1038/s41588-018-0248-z

Cernigliaro, C., D'Anneo, A., Carlisi, D., Giuliano, M., Marino Gammazza, A., Barone, R., et al. (2019). Ethanol-mediated stress promotes autophagic survival and aggressiveness of colon cancer cells via activation of Nrf2/HO-1 pathway. *Cancers* 11:505. doi: 10.3390/cancers11040505

Charmpi, K., and Ycart, B. (2015). Weighted kolmogorov smirnov testing: an alternative for gene set enrichment analysis. *Stat. Appl. Genet. Mol. Biol.* 14, 279–293. doi: 10.1515/sagmb-2014-0077

Davis, A. P., Grondin, C. J., Johnson, R. J., Sciaky, D., King, B. L., McMorran, R., et al. (2017). The comparative toxicogenomics database: update 2017. *Nucleic Acids Res.* 45, D972–D978. doi: 10.1093/nar/gkw838

Davis, A. P., Murphy, C. G., Johnson, R., Lay, J. M., Lennon-Hopkins, K., Saraceni-Richards, C., et al. (2013). The comparative toxicogenomics database: update 2013. *Nucleic Acids Res.* 41, D1104–D1114. doi: 10.1093/nar/gks994

Dekker, E., Tanis, P. J., Vleugels, J. L. A., Kasi, P. M., and Wallace, M. B. (2019). Colorectal cancer. *Lancet* 394, 1467–1480. doi: 10.1016/S0140-6736(19)32319-0

Erkekoglu, P., Oral, D., Chao, M. W., and Kocer-Gumusel, B. (2017). Hepatocellular carcinoma and possible chemical and biological causes: a review. *J. Environ. Pathol. Toxicol. Oncol.* 36, 171–190. doi: 10.1615/JEnvironPatholToxicolOncol.2017020927

Ferlay, J., Colombet, M., Soerjomataram, I., Mathers, C., Parkin, D. M., Pineros, M., et al. (2019). Estimating the global cancer incidence and mortality in 2018: GLOBOCAN sources and methods. *Int. J. Cancer* 144, 1941–1953. doi: 10.1002/ijc.31937

Flis, S., and Splwinski, J. (2009). Inhibitory effects of 5-fluorouracil and oxaliplatin on human colorectal cancer cell survival are synergistically enhanced by sulindac sulfide. *Anticancer Res.* 29, 435–441.

Giacchetti, S. P. B., Zidani, R., Le Bail, N., Faggiuolo, R., and Focan, C. (2000). Phase III multicenter randomized trial of oxaliplatin added to chronomodulated fluorouracil–leucovorin as first-line treatment of metastatic colorectal cancer. *J. Clin. Oncol.* 18, 136–147. doi: 10.1200/JCO.2000.18.1.136

Goldberg, R. M., Sargent, D. J., Morton, R. F., Fuchs, C. S., Ramanathan, R. K., Williamson, S. K., et al. (2004). A randomized controlled trial of fluorouracil plus leucovorin, irinotecan, and oxaliplatin combinations in patients with previously untreated metastatic colorectal cancer. *J. Clin. Oncol.* 22, 23–30. doi: 10.1200/JCO.2004.09.046

Gong, L., Zhang, D., Lei, Y., Qian, Y., Tan, X., and Han, S. (2018). Transcriptome-wide association study identifies multiple genes and pathways associated with pancreatic cancer. *Cancer Med.* 7, 5727–5732. doi: 10.1002/cam4.1836

Gregoriadis, G. C., Apostolidis, N. S., Romanos, A. N., and Paradellis, T. P. (1983). A comparative study of trace elements in normal and cancerous colorectal tissues. *Cancer* 52, 508–519. doi: 10.1002/1097-0142(19830801)52:3<508::AID-CNCR2820520322>3.0.CO;2-8

Grondin, C. J., Davis, A. P., Wiegers, T. C., Wiegers, J. A., and Mattingly, C. J. (2018). Accessing an expanded exposure science module at the comparative toxicogenomics database. *Environ. Health Perspect.* 126:014501. doi: 10.1289/EHP2873

Gusev, A., Mancuso, N., Won, H., Kousi, M., Finucane, H. K., Reshef, Y., et al. (2018). Transcriptome-wide association study of schizophrenia and chromatin activity yields mechanistic disease insights. *Nat. Genet.* 50, 538–548. doi: 10.1038/s41588-018-0092-1

Gustavsson, B., Carlsson, G., Machover, D., Petrelli, N., Roth, A., Schmoll, H. J., et al. (2015). A review of the evolution of systemic chemotherapy in the management of colorectal cancer. *Clin. Colorectal Cancer* 14, 1–10. doi: 10.1016/j.clcc.2014.11.002

Hamada, T., Cao, Y., Qian, Z. R., Masugi, Y., Nowak, J. A., Yang, J., et al. (2017). Aspirin use and colorectal cancer survival according to tumor CD274 (Programmed Cell Death 1 Ligand 1) expression status. *J. Clin. Oncol.* 35, 1836–1844. doi: 10.1200/JCO.2016.70.7547

Hammerschlag, A. R., Stringer, S., de Leeuw, C. A., Sniekers, S., Taskesen, E., Watanabe, K., et al. (2017). Genome-wide association analysis of insomnia complaints identifies risk genes and genetic overlap with psychiatric and metabolic traits. *Nat. Genet.* 49, 1584–1592. doi: 10.1038/ng.3888

He, A., Wang, W., Prakash, N. T., Tinkov, A. A., Skalny, A. V., Wen, Y., et al. (2018). Integrating genome-wide association study summaries and element-gene interaction datasets identified multiple associations between elements and complex diseases. *Genet. Epidemiol.* 42, 168–173. doi: 10.1002/gepi.22106

Hong, S. M., Hwang, S. W., Wang, T., Park, C. W., Ryu, Y. M., Jung, J. H., et al. (2019). Increased nicotinamide adenine dinucleotide pool promotes colon cancer progression by suppressing reactive oxygen species level. *Cancer Sci.* 110, 629–638. doi: 10.1111/cas.13886

Huang, H., Cheng, S., Ding, M., Wen, Y., Ma, M., Zhang, L., et al. (2019). Integrative analysis of transcriptome-wide association study and mRNA expression profiles identifies candidate genes associated with autism spectrum disorders. *Autism Res.* 12, 33–38. doi: 10.1002/aur.2048

Huang, L., Frampton, G., Rao, A., Zhang, K. S., Chen, W., Lai, J. M., et al. (2012). Monoamine oxidase A expression is suppressed in human cholangiocarcinoma via coordinated epigenetic and IL-6-driven events. *Lab. Invest.* 92, 1451–1460. doi: 10.1038/labinvest.2012.110

Jones, R. R., DellaValle, C. T., Weyer, P. J., Robien, K., Cantor, K. P., Krasner, S., et al. (2019). Ingested nitrate, disinfection by-products, and risk of colon and rectal cancers in the Iowa Women's Health Study cohort. *Environ. Int.* 126, 242–251. doi: 10.1016/j.envint.2019.02.010

Khoury, M. J., Davis, R., Gwinn, M., Lindegren, M. L., and Yoon, P. (2005). Do we need genomic research for the prevention of common diseases with environmental causes? *Am. J. Epidemiol.* 161, 799–805. doi: 10.1093/aje/kwi113

Kouhara, J., Yoshida, T., Nakata, S., Horinaka, M., Wakada, M., Ueda, Y., et al. (2007). Fenretinide up-regulates DR5/TRAIL-R2 expression via the induction of the transcription factor CHOP and combined treatment with fenretinide and TRAIL induces synergistic apoptosis in colon cancer cell lines. *Int. J. Oncol.* 30, 679–687. doi: 10.3892/ijo.30.3.679

Kuipers, E. J., Grady, W. M., Lieberman, D., Seufferlein, T., Sung, J. J., Boelens, P. G., et al. (2015). Colorectal cancer. *Nat. Rev. Dis. Primers* 1:15065. doi: 10.1038/nrdp.2015.65

Lee, H. C., Chiu, W. C., Wang, T. N., Liao, Y. T., Chien, I. C., Lee, Y., et al. (2017). Antidepressants and colorectal cancer: A population-based nested case-control study. *J. Affect. Disord.* 207, 353–358. doi: 10.1016/j.jad.2016.09.057

Li, J., Yang, X. M., Wang, Y. H., Feng, M. X., Liu, X. J., Zhang, Y. L., et al. (2014). Monoamine oxidase A suppresses hepatocellular carcinoma metastasis by inhibiting the adrenergic system and its transactivation of EGFR signaling. *J. Hepatol.* 60, 1225–1234. doi: 10.1016/j.jhep.2014.02.025

Li, X., Pathi, S. S., and Safe, S. (2015). Sulindac sulfide inhibits colon cancer cell growth and downregulates specificity protein transcription factors. *BMC Cancer* 15:974. doi: 10.1186/s12885-015-1956-8

Liggett, J. L., Choi, C. K., Donnell, R. L., Kihm, K. D., Kim, J. S., Min, K. W., et al. (2014). Nonsteroidal anti-inflammatory drug sulindac sulfide suppresses structural protein Nesprin-2 expression in colorectal cancer cells. *Biochim. Biophys. Acta* 1840, 322–331. doi: 10.1016/j.bbagen.2013.09.032

Lin, B. M., Nadkarni, G. N., Tao, R., Graff, M., Fornage, M., Buyske, S., et al. (2019). Genetics of chronic kidney disease stages across ancestries: the PAGE study. *Front. Genet.* 10:494. doi: 10.3389/fgene.2019.00494

Lucas, A. S., O'Neil, B. H., and Goldberg, R. M. (2011). A decade of advances in cytotoxic chemotherapy for metastatic colorectal cancer. *Clin. Colorectal Cancer* 10, 238–244. doi: 10.1016/j.clcc.2011.06.012

Lv, Q., Wang, D., Yang, Z., Yang, J., Zhang, R., Yang, X., et al. (2018). Repurposing antitubercular agent isoniazid for treatment of prostate cancer. *Biomater. Sci.* 7, 296–306. doi: 10.1039/C8BM01189C

Manuck, S. B., and McCaffery, J. M. (2014). Gene-environment interaction. *Annu. Rev. Psychol.* 65, 41–70. doi: 10.1146/annurev-psych-010213-115100

Mattingly, C. J., Colby, G. T., Forrest, J. N., and Boyer, J. L. (2003). The comparative toxicogenomics database (CTD). *Environ. Health Perspect.* 111, 793–795. doi: 10.1289/ehp.6028

Mattingly, C. J., Rosenstein, M. C., Colby, G. T., Forrest, J. N. Jr., and Boyer, J. L. (2006). The Comparative Toxicogenomics Database (CTD): a resource for comparative toxicological studies. *J. Exp. Zool. A Comp. Exp. Biol.* 305, 689–692. doi: 10.1002/jez.a.307

Messerlian, C., Martinez, R. M., Hauser, R., and Baccarelli, A. A. (2017). 'Omics' and endocrine-disrupting chemicals - new paths forward. *Nat. Rev. Endocrinol.* 13, 740–748. doi: 10.1038/nrendo.2017.81

Nuotio, J., Oikonen, M., Magnussen, C. G., Jokinen, E., Laitinen, T., Hutri-Kahonen, N., et al. (2014). Cardiovascular risk factors in 2011 and secular trends since 2007: the cardiovascular risk in young finns study. *Scand. J. Public Health* 42, 563–571. doi: 10.1177/1403494814541597

Pham, D. N., Leclerc, D., Levesque, N., Deng, L., and Rozen, R. (2013). beta,beta-carotene 15,15'-monooxygenase and its substrate beta-carotene modulate migration and invasion in colorectal carcinoma cells. *Am. J. Clin. Nutr.* 98, 413–422. doi: 10.3945/ajcn.113.060996

Qi, X., Wang, S., Zhang, L., Liu, L., Wen, Y., Ma, M., et al. (2019). An integrative analysis of transcriptome-wide association study and mRNA expression profile identified candidate genes for attention-deficit/hyperactivity disorder. *Psychiatry Res.* 282:112639. doi: 10.1016/j.psychres.2019.112639

Qu, X., Yang, L., Shi, Q., Wang, X., Wang, D., and Wu, G. (2018). Lidocaine inhibits proliferation and induces apoptosis in colorectal cancer cells by upregulating mir-520a-3p and targeting EGFR. *Pathol. Res. Pract.* 214, 1974–1979. doi: 10.1016/j.prp.2018.09.012

Raitakari, O. T., Juonala, M., Ronnemaa, T., Keltikangas-Jarvinen, L., Rasanen, L., Pietikainen, M., et al. (2008). Cohort profile: the cardiovascular risk in Young Finns Study. *Int. J. Epidemiol.* 37, 1220–1226. doi: 10.1093/ije/dym225

Roper, J., Sinnamon, M. J., Coffee, E. M., Belmont, P., Keung, L., Georgeon-Richard, L., et al. (2014). Combination PI3K/MEK inhibition promotes tumor apoptosis and regression in PIK3CA wild-type, KRAS mutant colorectal cancer. *Cancer Lett.* 347, 204–211. doi: 10.1016/j.canlet.2014.02.018

U.S. Department of Health and Human Services (2016). *14th Report on Carcinogens*. Washington, DC: U.S. Department of health and human services.

Siegel, R. L., Miller, K. D., and Jemal, A. (2017). Cancer statistics, 2017. *CA Cancer J. Clin.* 67, 7–30. doi: 10.3322/caac.21387

Soderholm, J. F., Bird, S. L., Kalab, P., Sampathkumar, Y., Hasegawa, K., Uehara-Bingen, M., et al. (2011). Importazole, a small molecule inhibitor of the transport receptor importin-beta. *ACS Chem. Biol.* 6, 700–708. doi: 10.1021/cb2000296

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., et al. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U.S.A.* 102, 15545–15550. doi: 10.1073/pnas.0506580102

Tam, V., Patel, N., Turcotte, M., Bosse, Y., Pare, G., and Meyre, D. (2019). Benefits and limitations of genome-wide association studies. *Nat. Rev. Genet.* 20, 467–484. doi: 10.1038/s41576-019-0127-1

Thompson, P. A., Khatami, M., Baglole, C. J., Sun, J., Harris, S. A., Moon, E. Y., et al. (2015). Environmental immune disruptors, inflammation and cancer risk. *Carcinogenesis* 36(Suppl. 1), S232–S253. doi: 10.1093/carcin/bgv038

Tsubaki, M., Takeda, T., Tani, T., Shimaoka, H., Suzuyama, N., Sakamoto, K., et al. (2015). PKC/MEK inhibitors suppress oxaliplatin-induced neuropathy and potentiate the antitumor effects. *Int. J. Cancer* 137, 243–250. doi: 10.1002/ijc.29367

Tsukamoto, H., Mizoshita, T., Katano, T., Hayashi, N., Ozeki, K., Ebi, M., et al. (2015). Preventive effect of rebamipide on N-methyl-N'-nitro-N-nitrosoguanidine-induced gastric carcinogenesis in rats. *Exp. Toxicol. Pathol.* 67, 271–277. doi: 10.1016/j.etp.2015.01.003

Wang, K., Li, M., and Bucan, M. (2007). Pathway-based approaches for analysis of genomewide association studies. *Am. J. Hum. Genet.* 81, 1278–1283. doi: 10.1086/522374

Weng, L., Macciardi, F., Subramanian, A., Guffanti, G., Potkin, S. G., Yu, Z., et al. (2011). SNP-based pathway enrichment analysis for genome-wide association studies. *BMC Bioinformatics* 12:99. doi: 10.1186/1471-2105-12-99

Wu, J. B., Shao, C., Li, X., Li, Q., Hu, P., Shi, C., et al. (2014). Monoamine oxidase A mediates prostate tumorigenesis and cancer metastasis. *J. Clin. Invest.* 124, 2891–2908. doi: 10.1172/JCI70982

Wu, L., Shi, W., Long, J., Guo, X., Michailidou, K., Beesley, J., et al. (2018). A transcriptome-wide association study of 229,000 women identifies new candidate susceptibility genes for breast cancer. *Nat. Genet.* 50, 968–978. doi: 10.1038/s41588-018-0132-x

Wu, S., Powers, S., Zhu, W., and Hannun, Y. A. (2016). Substantial contribution of extrinsic risk factors to cancer development. *Nature* 529, 43–47. doi: 10.1038/nature16166

Yan, W., Li, R., He, J., Du, J., and Hou, J. (2015). Importin beta1 mediates nuclear factor-kappaB signal transduction into the nuclei of myeloma cells and affects their proliferation and apoptosis. *Cell. Signal.* 27, 851–859. doi: 10.1016/j.cellsig.2015.01.013

Yang, J., Guo, Y., Lu, C., Zhang, R., Wang, Y., Luo, L., et al. (2019). Inhibition of Karyopherin beta 1 suppresses prostate cancer growth. *Oncogene* 38, 4700–4714. doi: 10.1038/s41388-019-0745-2

Zareifopoulos, N., and Panayiotakopoulos, G. (2017). Neuropsychiatric effects of antimicrobial agents. *Clin. Drug Investig.* 37, 423–437. doi: 10.1007/s40261-017-0498-z

Zhang, Y. F., Huang, Y., Ni, Y. H., and Xu, Z. M. (2019). Systematic elucidation of the mechanism of geraniol via network pharmacology. *Drug Des. Devel. Ther.* 13, 1069–1075. doi: 10.2147/DDDT.S189088

Zhao, Y., He, A., Zhu, F., Ding, M., Hao, J., Fan, Q., et al. (2018). Integrating genome-wide association study and expression quantitative trait locus study identifies multiple genes and gene sets associated with schizophrenia. *Prog. Neuropsychopharmacol. Biol. Psychiatry* 81, 50–54. doi: 10.1016/j.pnpbp.2017.10.003

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.

frontiers in Genetics

REVIEW
published: 13 May 2020
doi: 10.3389/fgene.2020.00424

From GWAS to Function: Using Functional Genomics to Identify the Mechanisms Underlying Complex Diseases

*Eddie Cano-Gamez[1] and Gosia Trynka[1,2]**

[1] Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, United Kingdom, [2] Open Targets, Wellcome Genome Campus, Cambridge, United Kingdom

Genome-wide association studies (GWAS) have successfully mapped thousands of loci associated with complex traits. These associations could reveal the molecular mechanisms altered in common complex diseases and result in the identification of novel drug targets. However, GWAS have also left a number of outstanding questions. In particular, the majority of disease-associated loci lie in non-coding regions of the genome and, even though they are thought to play a role in gene expression regulation, it is unclear which genes they regulate and in which cell types or physiological contexts this regulation occurs. This has hindered the translation of GWAS findings into clinical interventions. In this review we summarize how these challenges have been addressed over the last decade, with a particular focus on the integration of GWAS results with functional genomics datasets. Firstly, we investigate how the tissues and cell types involved in diseases can be identified using methods that test for enrichment of GWAS variants in genomic annotations. Secondly, we explore how to find the genes regulated by GWAS loci using methods that test for colocalization of GWAS signals with molecular phenotypes such as quantitative trait loci (QTLs). Finally, we highlight potential future research avenues such as integrating GWAS results with single-cell sequencing read-outs, designing functionally informed polygenic risk scores (PRS), and validating disease associated genes using genetic engineering. These tools will be crucial to identify new drug targets for common complex diseases.

Keywords: GWAS, SNP enrichment, colocalization analysis, TWAS, single-cell RNA seq, eQTL, QTL

OPEN ACCESS

Edited by:
*Xiyin Wang,
North China University of Science
and Technology, China*

Reviewed by:
*Dusanka Savic Pavicevic,
University of Belgrade, Serbia
Eivind Hovig,
University of Oslo, Norway*

****Correspondence:***
*Gosia Trynka
gosia@sanger.ac.uk*

Specialty section:
*This article was submitted to
Bioinformatics and Computational
Biology,
a section of the journal
Frontiers in Genetics*

Received: *16 October 2019*
Accepted: *06 April 2020*
Published: *13 May 2020*

Citation:
*Cano-Gamez E and Trynka G
(2020) From GWAS to Function:
Using Functional Genomics to Identify
the Mechanisms Underlying Complex
Diseases. Front. Genet. 11:424.
doi: 10.3389/fgene.2020.00424*

INTRODUCTION

Common non-communicable diseases such as autoimmunities, neurodegeneration, and cardiovascular disease are among the most pressing challenges in present day healthcare. These conditions are influenced by the interaction between a genetic predisposition and environmental or lifestyle factors (Smith et al., 2005). As opposed to rare diseases, which are often caused by the dysfunction of a single gene, common diseases are complex traits, i.e., they are influenced by the added contribution of thousands of common genetic variants, each having a small individual effect on the phenotype (Hindorff et al., 2011). This makes studying complex diseases challenging, as their genetic architecture follows a polygenic rather than a Mendelian model (Visscher and Goddard, 2019).

Genome-wide association studies (GWAS) are designed to map the polygenic architecture of common diseases by identifying genetic variants present at a significantly higher frequency in individuals with disease than in the healthy population (Wellcome Trust Case Control Consortium, 2007). Over the last 12 years, GWAS have grown significantly both in sample size and in the number of investigated traits (Visscher et al., 2017), with 128,550 associations and over 4,000 publications reported in the GWAS catalog to date (MacArthur et al., 2017).

Despite the success of GWAS, the clinical insights derived from their results have been limited. This is due to the difficulty of interpreting GWAS associations. Firstly, neighboring genetic variants are often correlated with one another, as they tend to be inherited together due to co-segregation during meiotic recombination, a phenomenon referred to as linkage disequilibrium (LD) [for a more detailed discussion of LD, refer to the review by Slatkin (2008)]. LD results in multiple variants in a locus being present in the same individual purely due to this correlation. This makes it difficult to distinguish the causal variants underpinning the association. Secondly, it is unclear which cell types are causal to the disease, as the pathophysiology of complex diseases often implicates interactions of multiple cell types. For example, the development of atherosclerotic plaques involves monocytes, lymphocytes, mast cells, neutrophils and smooth muscle (Insull, 2009). It is unclear which cell types are the true drivers of a disease (i.e., in which cell type GWAS variants act) and which are the consequence of the disease pathogenic processes. Finally, over 90% of GWAS variants fall in non-coding regions of the genome and thus do not directly affect the coding sequence of a gene. The accumulation of these variants in DNA regulatory elements (Maurano et al., 2012) and the observation that they can disrupt binding sites for transcription factors (TFs) (Musunuru et al., 2010) suggests that these variants act by regulating the expression levels of genes. However, disease-associated loci often contain multiple genes, making it challenging to distinguish the affected ones. In summary, follow-up studies are necessary to interpret GWAS results and to infer the exact disease-causal variants, the genes they regulate and the cell types in which they act (**Figure 1**).

Statistical methods designed to tackle these challenges integrate GWAS results with functional genomics data such as gene expression or chromatin activity profiles assayed across a range of cell types and tissues. In particular, fine-mapping aims to define causal variants, SNP enrichment methods prioritize disease relevant cell types and colocalization nominates likely target genes (**Figure 1**). Here, we review a selection of methods that facilitate translation of GWAS results, focusing on SNP enrichment and colocalization approaches, and we highlight some biological conclusions derived from these studies. We also discuss transcriptome-wide association studies which directly associate genes with diseases. For a detailed analysis of fine-mapping methods, we refer the reader to a previous review (Schaid et al., 2018). Finally, we reflect on some of the challenges and opportunities of post-GWAS research, such as the availability of high-throughput single-cell sequencing platforms, the identification of relevant intermediate phenotypes, the development of polygenic risk scores (PRS), and the systematic application of genetic engineering for GWAS validation.

IDENTIFYING CELL TYPES RELEVANT TO COMPLEX DISEASES

The variants mapped through GWAS provide a strong genetic anchor to complex disease biology and therefore to the development of new therapies. However, going from genetics to function requires robust model systems in which disease-causal cells and tissues can be probed and manipulated. For example, tumor-derived human cell lines have been relevant for the systematic identification of novel drug targets in cancer (Behan et al., 2019). Such model systems provide valuable clues for drug target validation, as they enable us to elucidate the molecular mechanisms of disease, to identify relevant genes and to screen compounds with therapeutic potential at high-throughput. However, for many complex diseases, it is unclear which cells are causal. For instance, independent studies have proposed that rheumatoid arthritis is caused by cells as diverse as T cells (Cope et al., 2007), B cells (Bugatti et al., 2014), macrophages (Udalova et al., 2016), and synoviocytes (Beatrix Bartok, 2010). Psychiatric traits, which involve dysregulation of the central nervous system, pose a similar challenge due to the complex histological structure of the brain. For example, over 20 different cellular models have been used to study bipolar disorder (Viswanath et al., 2015). The lack of ground truth causal cell types makes the functional validation of GWAS variants challenging, as dozens of tissues could be involved in the development of a trait. Statistical methods that integrate GWAS variants either with transcriptome or chromatin annotations assayed across a range of different tissues can help nominate the most disease-relevant cell types.

Snp Enrichment Analysis Based on Genome-Wide Significant Gwas Variants

Identification of disease-relevant cell types assumes that GWAS variants are overrepresented in genomic regions specifically active in the pathogenic cell types (SNP enrichment). SNP enrichment methods integrate GWAS results with different genomic annotations and prioritize the cell types in which associated variants overlap annotations more frequently than expected by chance. For example, cell type specific activity of a genomic region (e.g., a GWAS locus) can be defined by the expression levels of genes within the region. An approach proposed by Hu et al. (2011) (*SNPsea*) defines as highly cell type specific those genes with high expression in individual cell types as compared to all other cell types. If, for a given trait, GWAS loci are overrepresented (enriched) for genes specifically expressed in a given cell type, that cell type is prioritized. The statistical significance is derived from a permutation-based test in which disease-associated loci are compared with random loci of similar properties (e.g., distance to TSS and gene density) (Slowikowski et al., 2014). The authors used this approach for three different immune-mediated diseases (Crohn's disease, systemic lupus erythematosus and rheumatoid arthritis), testing

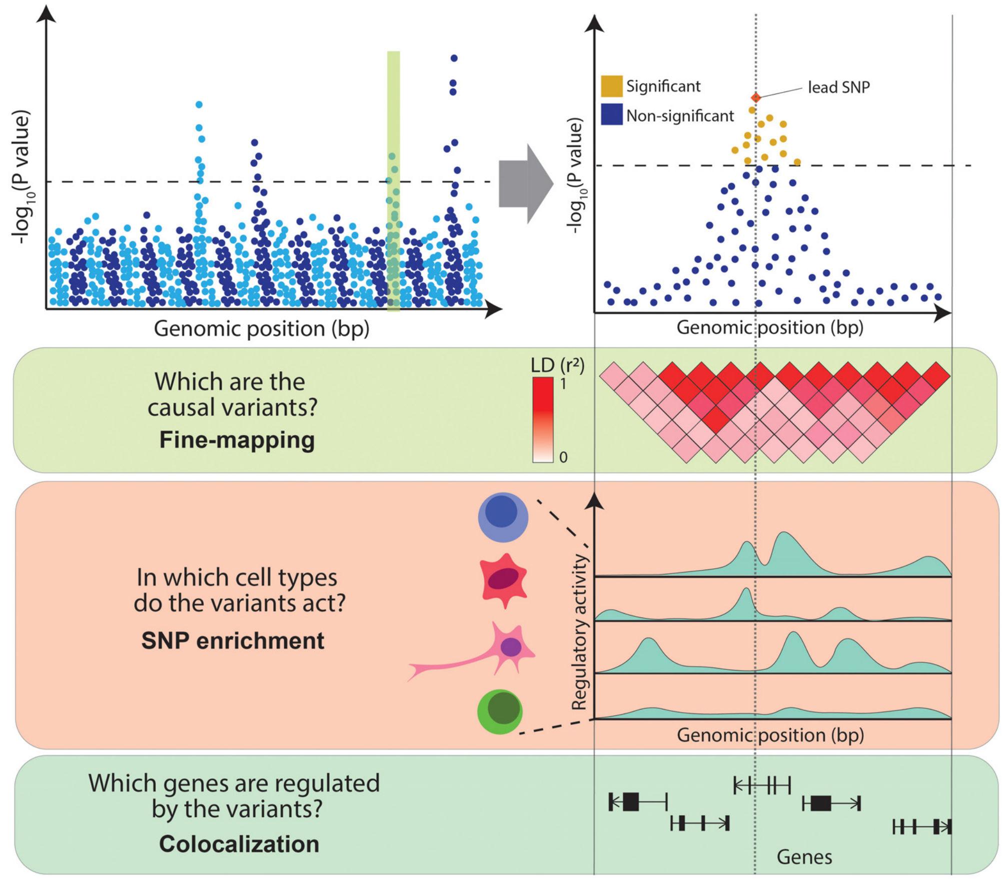


FIGURE 1 | Challenges in interpreting GWAS associations. From the top: Manhattan plot illustrates the association between genetic variants and a trait (e.g., a disease) at a genome-wide level (left panel) and within an example locus (right panel). Variants above the dotted line represent genome-wide significant associations. The panels below illustrate the main challenges in interpreting GWAS associations: high LD between variants (encoded in shades of red), variable levels of regulatory activity of the genomic regions across cell types (peaks of different heights represent different levels of activity of chromatin marks) and multiple genes within the associated locus.

for enrichment in gene expression across 79 human and 223 mouse tissues. While lupus-associated variants were enriched in genes specifically expressed in B cells, rheumatoid arthritis variants were enriched in genes specific to CD4+ memory T cells (Hu et al., 2011). This demonstrated that SNP enrichment is a valid approach for cell type prioritization and suggested that variants associated with immune-mediated diseases result in dysfunction of the adaptive immune system.

However, gene expression-based methods use an arbitrary definition of which genes contribute to the SNP enrichment score at each locus and either select a single gene with the highest cell type specific gene expression or include all the genes within the locus (Hu et al., 2011). The caveat of this is that the first approach can select the wrong gene and does not account for the effects of multiple causal genes, while the second approach can dilute the signal by including many genes which are likely not relevant to the tested trait.

Alternatively, GWAS variants can be integrated with chromatin annotations such as open chromatin regions (assayed by DNase-hypersensitivity or ATAC-seq) (Boyle et al., 2008; Buenrostro et al., 2013), histone modifications (e.g., H3K4me1, H3K4me3, H3K27ac, and H3K27me3) (Bannister and Kouzarides, 2011) or DNA methylation (Frommer et al., 1992). These annotations are profiled using sequencing-based approaches which identify genomic elements with high levels of regulatory activity (i.e., peaks). For example, DNA accessibility peaks indicate regions available for transcription factor (TF) binding, H3K4me3 peaks highlight gene promoters (Barski et al., 2007) and H3K27ac peaks mark active enhancer and promoter regions (Creyghton et al., 2010). As opposed to gene expression, chromatin marks can be physically overlapped with GWAS variants and therefore enrichment analysis can be estimated directly from the SNPs located within the annotations (**Figure 2**). Initiatives like the Encyclopedia of DNA Elements (ENCODE) (ENCODE Project Consortium, 2012), Roadmap Epigenomics (Roadmap Epigenomics Consortium et al., 2015), and the BLUEPRINT project (Chen et al., 2016) have profiled tens of epigenetic marks across dozens of human tissues, providing rich resources for these type of SNP enrichment analyses.

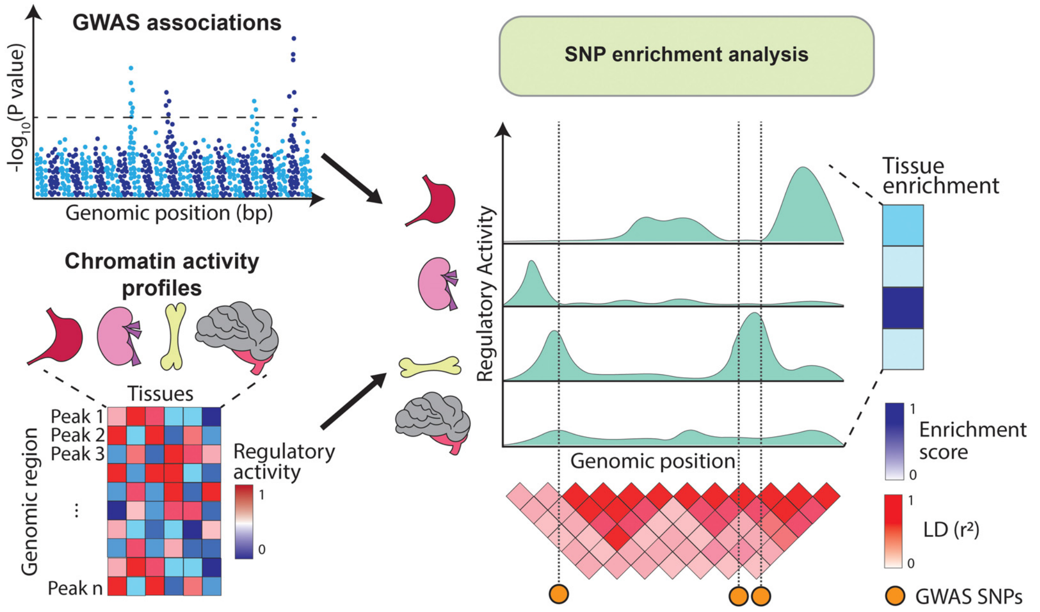


FIGURE 2 | Overview of SNP enrichment analysis using chromatin annotations. SNP enrichment analysis integrates association signals from GWAS (Manhattan plot on the top left) with functional genomics data such as chromatin annotations (heatmap on the bottom left). GWAS SNPs are overlapped with regulatory elements (right panel) and if in a given tissue the overlap occurs more frequently than expected by chance, the tissue is assigned a high enrichment score.

An early example of SNP enrichment analysis with chromatin annotations overlapped GWAS variants for 447 traits with DNase-hypersensitive (DHS) regions from 348 tissues (Maurano et al., 2012). Using a simple binomial test, this study found that GWAS SNPs were enriched in DHS regions compared to a background set of common SNPs from the HapMap project (International HapMap Consortium, 2003). These SNP enrichment results were tissue-specific, for example, variants for coronary heart disease and body mass index were enriched in DHS regions active in fetal cells. Conversely, variants associated with age-related diseases (e.g., cancer and immune-mediated diseases) were significantly depleted from fetal DHS regions. These findings suggest that GWAS variants could modify the regulatory activity of non-coding elements in a cell-type specific manner.

However, GWAS loci reside in regions of high gene density, which also include higher density of chromatin regulatory elements, which can confound enrichment estimates if not accounted for. To address this issue, enrichment of disease variants in DHS regions can compare GWAS SNPs to random sets of SNPs with similar properties (i.e., LD, gene density and distance to TSS) in a permutation-based approach (*GREGOR*) (Schmidt et al., 2015). By matching SNPs, this approach is robust to both gene and annotation density. Results from this study confirmed that GWAS SNPs are generally enriched in active regulatory regions compared to random SNPs.

In addition to the binary overlap between SNPs and annotations, SNP enrichment analysis can also take into account other peak properties, such as the position of a variant within a peak and the height of the peak (reflecting the levels of regulatory activity). Moreover, SNP enrichment analysis can be extended to chromatin marks other than DHS. For example, *epiGWAS* tests for the accumulation of GWAS variants in chromatin regions defined using ChIP-seq for histone modifications (Trynka et al., 2013). In this approach, variants within each GWAS locus are scored for their distance to the summit of the nearest peak and for the height of the peak i.e., the height (h) to distance (d) ratio (h/d). The contribution to the final enrichment score is determined by a single variant per locus with the highest h/d score, and statistical significance of the enrichment is inferred by comparison to a matched set of random SNPs sampled from the GWAS catalog (MacArthur et al., 2017). This approach is suitable for narrow histone marks, where peak summits can be reliably defined. The authors confirmed that variants associated with LDL cholesterol levels were enriched in gene promoters active in the liver, and that type 2 diabetes variants were enriched in gene promoters active in both liver cells and pancreatic islets. In both cases, the tissues are well understood to play a role in disease biology. The authors also used this approach across immune-mediated traits where pathogenic cell types are less well characterized. This revealed an enrichment of rheumatoid arthritis and type 1 diabetes variants in CD4+ T cell subsets, particularly in regulatory T cells.

One limitation of the above methods, which all rely on random sampling of SNPs to derive a null distribution, is that they make assumptions on the SNP parameters that need to be controlled for in random sampling (e.g., proximity to transcription start site, minor allele frequency, gene density, etc.). However, the presence of hidden confounders could bias the enrichment statistics if uncontrolled for. For example, high LD in a given genomic region can result in inflated SNP enrichment estimates (Trynka et al., 2015). One approach to address this, the *GoShifter* method (Trynka et al., 2015), derives statistical significance by shifting the location of functional annotations within the tested regions while preserving the distance between them. The result is a null distribution of SNP-annotation overlaps due to chance. This approach maintains the local genomic architecture, including the number of tested SNPs in LD, the number of annotations and the distance between the features, therefore controlling for hidden confounders. *GoShifter* confirmed a significant enrichment of rheumatoid arthritis variants in promoter regions specific to CD4+ memory T cells and also detected an enrichment of breast cancer variants in human mammary epithelial cells (Trynka et al., 2015). Both of these cell types are known to be involved in disease.

Given a well powered GWAS, SNP enrichment analysis can provide important insights into disease pathogenic tissues from leveraging the genetic signals. For example, Onengut-Gumuscu et al. (2015) asked if credible sets of type 1 diabetes SNPs (defined with a Bayesian fine-mapping approach) were enriched in functional annotations from the ENCODE (ENCODE Project Consortium, 2012) and Roadmap (Roadmap Epigenomics Consortium et al., 2015) projects (Onengut-Gumuscu et al., 2015). They did so by comparing the proportion of disease-associated SNPs and non-disease SNPs which overlapped functional elements, stratifying variants by their minor allele frequency. Interestingly, type 1 diabetes credible sets were strongly enriched in immune cell enhancers, particularly enhancers active in CD4+ and CD8+ T cells. Conversely, there was no detectable enrichment in enhancers active in pancreatic islets, in agreement with type 1 diabetes being an immune-mediated pathology. In contrast, a separate study profiled open chromatin, TF binding and gene expression in human pancreatic islets and integrated these profiles with GWAS loci for type 2 diabetes and fasting glycemia (Pasquali et al., 2014). The authors used a permutation-based test to estimate enrichments and concluded that glycemia and type 2 diabetes SNPs were strongly enriched in pancreatic islet enhancers, where they disrupted DNA binding by key islet TFs. This illustrates how SNP enrichment can distinguish different disease etiologies based solely on genetic associations, despite the traits sharing similar physiological manifestations.

Once the disease-relevant cell types are identified, subsequent experiments can be carried out to further refine the observed enrichments to the most relevant cell states. For example, we recently followed up the previously reported enrichment of immune disease variants in naive and memory CD4+ T cells, and macrophages (Hu et al., 2011; Fairfax et al., 2014; Trynka et al., 2013, 2015) by stimulating these cell types in the presence of different cytokine cocktails and profiling chromatin landscape with ATAC-seq and H3K27ac ChIP-seq across 55 cell states (Soskic et al., 2019). We observed that, in closely related cell types, the induction of different cell states results in quantitative changes in ATAC-seq and H3K27ac peaks, rather than in the induction of new cell state specific peaks. The broadly applied SNP enrichment methods, which rely on binary SNP-peak overlaps, failed to distinguish disease SNP enrichment between the different cell states. Therefore, we developed a new method (*CHEERS*) to tease apart enrichments in closely related cell types or cell states (Soskic et al., 2019). *CHEERS* asks whether GWAS variants tend to accumulate in regions with highly cell type-specific regulatory activity. SNPs are first intersected with chromatin elements (e.g., chromatin accessibility or ChIP-seq peaks) and are then assigned a score reflecting cell type specific regulatory activity of the region (i.e., how many sequencing reads exist within that region in one cell type as compared to the other cell types). Because this approach is based on cell type-specificity rather than absolute regulatory activity, it can disentangle enrichment patterns across highly similar cell types. We applied this approach to GWAS variants for 11 diseases, using chromatin annotations from our cytokine-stimulated dataset. Variants associated with different subtypes of inflammatory bowel disease (IBD) were enriched in chromatin elements specifically active in the Th1 cell state. For the remaining immune diseases, the strongest enrichment was in early stages of memory T cell activation. This enrichment pattern is important, as it not only nominates T cells as a relevant cell type, but also begins to explain which specific cellular processes are altered in disease. Additionally, a separate study performed SNP enrichment analysis for nine immune diseases using gene expression and chromatin accessibility profiles of 25 immune cell types in resting and activated states (Calderon et al., 2019). Here too, the strongest enrichment was observed among stimulated T cells.

Genome-Wide Snp Enrichment Analysis

The approaches described so far leverage the signal from genome wide significant variants as shown in **Table 1**. However, complex traits result from thousands of risk alleles and the majority of trait-associated variants remain undiscovered (Visscher et al., 2017). Thus, restricting the analysis to genome-wide significant variants could limit statistical power to detect biologically important enrichments. This has motivated the development of a number of methods which use all the common variants to estimate enrichments.

In a method called *fGWAS,* Pickrell reasoned that if GWAS variants were enriched in a given functional category, then SNPs belonging to that category would be more likely to have an effect on the trait (Pickrell, 2014). Using whole genome variants from imputation (Pasaniuc et al., 2014), he modeled the probability of a locus being associated with a disease as a function of its annotations using a hierarchical Bayesian model. When applied to chromatin regulatory maps from 402 tissues and 18 complex traits, *fGWAS* identified enrichment of HDL-associated variants in enhancers specifically active in the liver. Moreover, variants were generally depleted from repressed chromatin regions across all traits. By integrating functional annotations with GWAS statistics, *fGWAS* can also "re-weigh" and discover association signals for variants which did not originally reach genome-wide significance (Pickrell, 2014). An example is the SNP rs6659176, upweighted by *fGWAS* and confirmed to be associated with HDL through an independent study (Global Lipids Genetics Consortium et al., 2013).

In another study, Iotchkova et al. used a logistic regression framework to assess SNP enrichment (*GARFIELD*) and modeled

TABLE 1 | Methods for SNP enrichment analysis.

Method	Publications	Hypothesis tested	Input data
SNPsea	Hu et al., 2011; Slowikowski et al., 2014	Accumulation of GWAS variants near genes with high tissue specificity	Gene expression, GWAS index variants
EpiGWAS	Trynka et al., 2013	Accumulation of GWAS variants near highly active regulatory elements	Chromatin marks, GWAS index variants
GREGOR	Schmidt et al., 2015	Accumulation of GWAS variants in regulatory elements	Chromatin marks, GWAS index variants
GoShifter	Trynka et al., 2015	Intersection of GWAS variants with regulatory annotations (based on local-shifting of annotations)	Functional annotations, GWAS index variants
fGWAS	Pickrell, 2014	Higher GWAS effect sizes observed if a loci and a SNP overlap a functional annotation	Functional annotations, GWAS summary statistics
CHEERS	Soskic et al., 2019	Accumulation of GWAS variants in regulatory elements with high tissue specificity	Chromatin marks (quantitative), GWAS index variants
GARFIELD	Iotchkova et al., 2019	Higher GWAS effect sizes observed in variants that overlap regulatory annotations	Chromatin annotations, full GWAS summary statistics
RolyPoly	Calderon et al., 2017	Higher GWAS effect sizes observed near highly expressed genes	Gene expression, full GWAS summary statistics
LDSC	Finucane et al., 2015	Accumulation of heritability in variants overlapping a functional annotation	Chromatin annotations, full GWAS summary statistics
LDSC-SEG	Finucane et al., 2018	Accumulation of heritability near tissue specific genes	Gene expression, full GWAS summary statistics

Selected approaches and methods for enrichment testing of GWAS SNPs in functional annotations included in this review.

the trait association status of each SNP as a probability (Iotchkova et al., 2019), defined as a function of the variant's features (i.e., overlap with a functional annotation, distance to the nearest TSS and number of LD proxies). The significant association of a SNP (a binary variable) was tested at several significance thresholds, thus allowing more SNPs to be included in the calculation. The authors applied *GARFIELD* to DHS regions and functional annotations from ENCODE (Ernst and Kellis, 2012) and found that variants associated with height were enriched in DHS elements across all tissues, while ulcerative colitis variants showed tissue-specific enrichment mostly in blood cell types. Interestingly, the authors observed some of the enrichments only at lower significance thresholds. For example, variants associated with beta cell activity index were enriched in pancreatic islets enhancers only at lower significance thresholds (P value $< 1 \times 10^{-5}$). This suggests that including more trait-associated variants can improve enrichment estimates.

Enrichment Analysis Based on Snp Heritability

Heritability is the proportion of a trait's variance that is due to genetic variation. In particular, SNP heritability is the amount of phenotypic variance explained by a given set of SNPs (Yang et al., 2017). A number of methods have been developed to estimate the SNP heritability of a trait using either individual-level genotypes or summary statistics (Yang et al., 2010; Bulik-Sullivan et al., 2015) from GWAS. This gave rise to partitioning heritability approaches, which test for a significant accumulation of trait heritability in different functional categories of the genome. The authors of stratified LD-score regression (*LDSC*) (Finucane et al., 2015) argue that if GWAS variants are enriched in a functional category, then variants falling within that category will explain more trait heritability than other variants. To test for this, Finucane et al., 2015 partitioned all common SNPs into categories based on the functional elements that they overlapped. These categories included 24 unspecific annotations (coding regions, promoters, enhancers, introns, conserved elements and DHSs, among others) as well as histone modification profiles acquired from a variety of cell types. The authors calculated the SNP heritability of variants in each category using GWAS data for 17 traits and defined an enrichment score as the proportion of SNP heritability in a category divided by the proportion of SNPs in that category (Finucane et al., 2015). The authors found that, in general, conserved regions of the genome explained more heritability. Moreover, variants within enhancers specific to disease-relevant cell types also explained a substantial proportion of heritability. For example, liver-specific enhancers were enriched for HDL heritability and enhancers active in the central nervous system captured more SNP heritability of psychiatric traits (e.g., schizophrenia and bipolar disorder) than variants residing in enhancers present in other cell types.

However, one limitation of the *LDSC* method is its dependency on chromatin activity profiles, which are not always available. In contrast, gene expression profiles are available for a far greater number of cell types, including the less abundant ones. LD-score regression applied to specifically expressed genes (*LDSC-SEG*) extends the *LDSC* framework to partition heritability using gene expression profiles (Finucane et al., 2018). If first identifies the top 10% most specific genes expressed in each tissue and extends the regions on each side of the genes by 100 kb. The resulting regions are used as tissue-specific annotations in which variants are partitioned. Because gene expression is available for a wider set of tissues than epigenetic data, this enabled the analysis of less common cell types. The authors used *LDSC-SEG* to integrate expression profiles form GTEx with GWAS data for psychiatric traits and showed evidence of differential heritability enrichment across brain regions. For example, while only cells from the cortex were enriched for schizophrenia SNP heritability, both the cortex and the cerebellum were enriched for bipolar disorder SNP heritability. Subsequent application of

LDSC-SEG to brain expression data from the PsychENCODE project (The PsychENCODE Consortium et al., 2015) revealed that schizophrenia SNP heritability enrichment was driven by glutamatergic neurons, while bipolar disorder SNP heritability enrichment was driven by GABAergic neurons. Importantly, these psychiatric traits had not been analyzed for SNP enrichment before because of the insufficient number of GWAS-significant variants. This highlights the increased statistical power enabled by including all common variants in the analysis.

Finally, the *RolyPoly* method models the polygenic architecture of complex traits to estimate SNP enrichment (Calderon et al., 2017). In brief, the authors reasoned that variants with higher GWAS effect sizes would tend to be close to genes with higher expression in the causal tissues. Using a regression model, *RolyPoly* estimates the influence of cell type specific gene expression on the variance of GWAS effect sizes in each tissue. The authors applied *RolyPoly* to tissue-specific expression data from GTEx and confirmed a significant enrichment of variants affecting cholesterol levels in genes expressed in the liver and the small intestine. Moreover, they integrated GWAS data with single-cell gene expression profiles from brain tissue (Darmanis et al., 2015) and found a significant enrichment of risk variants for Alzheimer's disease in genes specific to microglia (Calderon et al., 2017). This agrees with increasing evidence suggesting the immune system is involved in Alzheimer's pathology (Gosselin et al., 2017).

In summary, SNP enrichment analysis leverages GWAS signals and functional annotations to pinpoint disease-relevant cell types. Multiple approaches have been proposed to estimate enrichment, such as integrating genome-wide significant variants with chromatin or gene expression profiles, as well as partitioning the SNP heritability of a trait based on the functional annotations of the genome. The increasing availability of expression and chromatin data for more cell types and states is expected to improve the granularity of these enrichment signals. This will allow us to confidently nominate the specific cell types and states causally involved in disease.

PRIORITIZING CAUSAL GENES AT GWAS LOCI

Once the most relevant cell types are identified, the next step is to prioritize genes causally involved in disease. Identification of candidate genes is most straightforward for coding variants, which directly disrupt the structure of a protein. One notable example is a locus containing the *TYK2* gene, as well as several gene members of the ICAM family. Variants at this locus have been associated with a number of immune diseases such as rheumatoid arthritis, ankylosing spondylitis, multiple sclerosis and IBD (Franke et al., 2010; Jostins et al., 2012; International Genetics of Ankylosing Spondylitis Consortium et al., 2013; Okada et al., 2014). Importantly, a number of these SNPs are *TYK2* missense variants. Of three independent signals at this locus, at least one is entirely explained by a single coding SNP which confers disease protection (Diogo et al., 2015). This SNP induces a proline to alanine substitution in the catalytic domain of *TYK2*, a kinase that mediates signal transduction downstream of various cytokine receptors (Dendrou et al., 2016). This substitution significantly impairs cytokine signaling, thus altering the communication between immune cells. Surprisingly, even though this variant protects against more than 10 different autoimmune diseases, complete knock-out of *TYK2* causes severe susceptibility to infections (Kreins et al., 2015). This led to the theory that *TYK2* function constitutes a spectrum, with complete abrogation causing immunodeficiency and augmented function increasing susceptibility to autoimmunity (Dendrou et al., 2016). Thus, a compound able to modulate the kinase activity of *TYK2* could be a successful drug candidate for autoimmune disorders.

However, 90% of the variants identified by GWAS are non-coding (Farh et al., 2015) and cannot be easily linked to a candidate causal gene. In contrast, these variants are thought to regulate gene expression via mechanisms such as modification of promoter and enhancer activity or disruption of binding sites for TFs. An example is the 1q13 locus, which contains a variant significantly associated with LDL cholesterol levels and myocardial infarction (Myocardial Infarction Genetics Consortium et al., 2009; Teslovich et al., 2010). This variant was shown to create a new TF binding site, which in turn causes the recruitment of an enhancer-binding protein, sharply increasing the expression of the nearby gene *SORT1* (Musunuru et al., 2010), a regulator of lipoprotein levels in plasma. *SORT1* in turn downregulates the levels of LDL. This makes *SORT1* an interesting drug target in myocardial infarction.

Most disease-associated variants are thought to act by mechanisms analogous to those at the *SORT1* locus. However, GWAS loci often contain multiple genes and identifying the causal genes is challenging. Profiling molecular traits (e.g., gene expression, DNA methylation, TF binding) and integrating them with GWAS results can be useful in linking non-coding variants to their target genes and unveiling the underlying regulatory events.

Colocalization Analysis

The quantification of molecular traits such as gene expression across thousands of individuals with different genotypes enables the association of genetic variants with intermediate traits (quantitative trait loci mapping, QTL) (**Figure 3A** and **Table 2**). The decreasing costs of high-throughput sequencing have resulted in dozens of QTL-mapping studies, profiling traits as diverse as gene expression (eQTLs) (Nica and Dermitzakis, 2013), protein expression (pQTLs) (Melzer et al., 2008; Yao et al., 2018), exon splicing (sQTLs) (Monlong et al., 2014; Ongen and Dermitzakis, 2015; Li et al., 2018), DNA methylation (mQTLs) (Banovich et al., 2014; Hannon et al., 2016), chromatin acetylation (acQTLs) (Sun et al., 2016; Pelikan et al., 2018), and chromatin accessibility (caQTLs) (Degner et al., 2012; Kumasaka et al., 2016). Of these, eQTLs are the most common, partly because of the robustness of RNA-sequencing technologies. One of the most comprehensive eQTL resources is the Genotype-Tissue expression project (GTEx), which profiled 53 tissues across nearly 1,000 individuals (GTEx Consortium, 2013; Melé et al., 2015). Another initiative, the BLUEPRINT project, measured the transcriptome, together with DNA methylation and histone

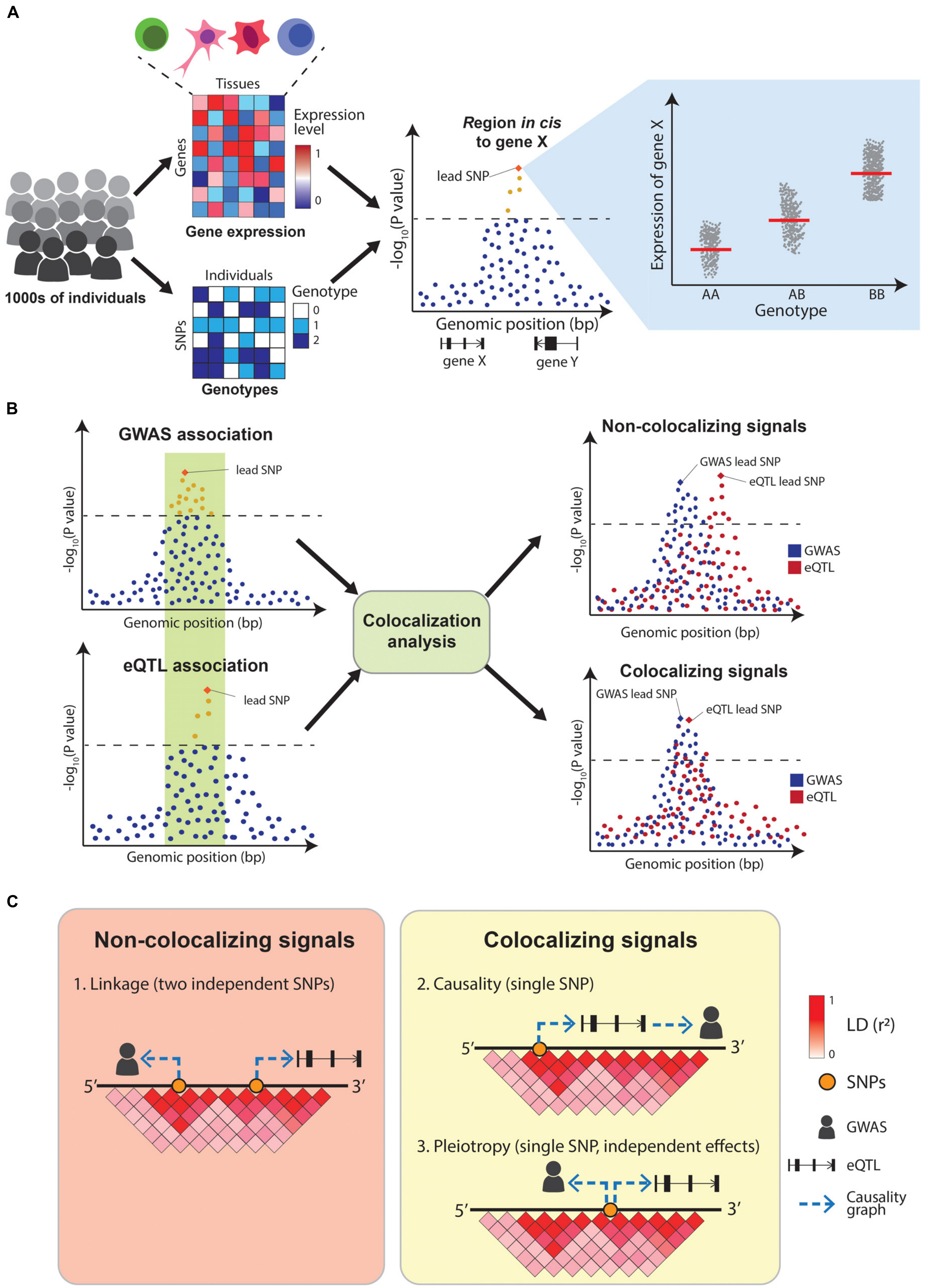


FIGURE 3 | Overview of eQTL-mapping and colocalization. **(A)** In eQTL-mapping gene expression is profiled in thousands of individuals and the expression level of each gene is tested for association with genotypes at nearby (*cis*) SNPs. **(B)** Colocalization compares the association patterns of GWAS and eQTLs at a locus to find if both signals are driven by the same causal variants. **(C)** GWAS and eQTL signals can overlap for three reasons: two independent causal variants in LD (linkage), a single causal variant affecting the GWAS trait via gene expression modulation (causality) or a single causal variant affecting both traits independently (pleiotropy). A positive colocalization supports causality or pleiotropy in favor of linkage.

TABLE 2 | Methods for colocalization analysis.

Method	Publication	Approach	Input data
Regulatory trait concordance (RTC)	Nica et al., 2010	Conditional regression	Individual genotypes
Proportionality test	Wallace et al., 2012	Test for concordance of effects	Individual genotypes
Sherlock	He et al., 2013	Genome-wide comparison of association "signatures"	Summary statistics
COLOC	Giambartolomei et al., 2014	Bayesian test	Summary statistics
gwas-pw	Pickrell et al., 2016	Bayesian test	Summary statistics
eCAVIAR	Hormozdiari et al., 2016	Bayesian fine-mapping and colocalization	Summary statistics
enloc	Wen et al., 2017	Bayesian test for enrichment, fine-mapping and colocalization	Summary statistics
MOLOC	Giambartolomei et al., 2018	Bayesian test for multiple traits	Summary statistics

Selected approaches and methods used to test for colocalization between GWAS and QTL signals included in this review.

modifications, in the most abundant cell types in peripheral blood from 197 individuals (Chen et al., 2016).

Integrating QTL maps with GWAS can identify potential molecular mechanisms underlying disease associations. Early examples of this simply assessed whether GWAS variants were also significant eQTLs. A study by Nicolae et al. (2010) combined GWAS results with eQTLs from human lymphoblastoid cell lines and concluded that GWAS SNPs are almost twice as likely to be eQTLs than random sets of SNPs. Similarly, a study by Dubois et al. (2010) concluded that 20 out of 38 (52%) risk loci for celiac disease were eQTLs in primary immune cells. However, these early approaches did not sufficiently control for the genetic architecture underlying GWAS and eQTL signals, resulting in high numbers of false positives findings. In particular, linkage disequilibrium between SNPs makes it challenging to identify which variants within a GWAS and a QTL locus are causally driving the associations. Overlapping eQTL and GWAS signals can be explained by three possible scenarios: (1) two independent causal SNPs in LD with each other (linkage), (2) a single-causal SNP which affects the trait by modulating the expression of a gene (causality), or (3) a single-causal SNP with independent effects on trait and gene expression (pleiotropy). Distinguishing between these scenarios is crucial to appropriately interpret GWAS results (**Figures 3B,C**). Additionally, eQTLs are abundant (Lappalainen et al., 2013) with 48% of common genetic variants estimated to act as eQTLs for at least one gene (Liu B. et al., 2019), making the overlap between GWAS and eQTL signals likely to happen due to chance. This motivated the development of formal statistical tests that estimate the probability of the overlaps between the two signals being due to chance. These methods are called colocalization tests.

A study by Plagnol et al. (2009) focused on a potentially causal relationship between the 12q13 locus, associated with type 1 diabetes, and the nearby gene *RPS26*. The authors reasoned that if the locus in question increased disease susceptibility via regulation of *RPS26* expression, then the effect sizes inferred from the GWAS and the *RPS26* eQTL (i.e., odds ratios and regression coefficients, respectively) should be proportional. In other words, the SNPs with the highest effects on type 1 diabetes would tend to also have the highest effects on *RPS26* expression, and the direction of effects would be consistent. The authors developed a statistical test for this proportionality (QTLmatch) and concluded that there was no evidence for colocalization at the 12q13 locus. Subsequently, Wallace et al. (2012) revisited this approach and implemented a generalized version into a more robust statistical framework.

An alternative approach described by Nica et al. (2010) first identifies loci with potential colocalizations and next regresses from the eQTL effect the most significant GWAS SNP in a locus. The eQTL association is then re-tested using the residuals from regression. To account for LD in the region, the procedure is repeated for all the SNPs in the region and the impact of the top GWAS SNP is compared to that of other variants. In the presence of a true colocalization, the regression coefficient of the top GWAS SNP results in a significantly larger impact than that of any other variant in the region. This process was implemented into a method called *Regulatory-Trait Concordance* (*RTC*).

Despite the usefulness of these approaches, neither of the two formally compares the odds of colocalization versus a null hypothesis. Instead, they are based on the proportionality of effects or the conditional association between two traits, which can be biased by LD and variable selection (Wallace, 2013). This can result in a large proportion of false positives. Additionally, both approaches require individual-level genotype data, which is seldom available. This motivated the development of methods which could be applied to GWAS summary statistics. Giambartolomei et al. (2014) proposed a colocalization test (*COLOC*) which computes the odds of colocalization compared to the null hypothesis using GWAS summary statistics. The authors identified five mutually exclusive scenarios at any given locus: either (1) the locus is not associated with any of the traits (the null hypothesis, H_0), (2) the locus is only significant in the GWAS (H_1), (3) the locus is only a significant eQTL (H_2), (4) the locus is associated with both traits due to two independent signals (linkage, H_3) or (5) the locus is associated with both traits due to a single colocalizing SNP (colocalization, H_4). The probability of each of these scenarios is estimated using a Bayesian framework and any locus where the probability of H_4 is significantly higher than that of H_3 (and of any other scenario) is said to colocalize.

Since its release, *COLOC* has become a reference method for colocalization testing. However, a limitation is that it only tests for two traits at a time. Elucidating the full chain of events that connects sequence variation to organismal phenotypes involves more than one molecular trait. For example, a variant can increase DNA methylation, in turn reducing the expression of a nearby gene, impairing cell function and increasing disease risk. Disentangling these effects requires a joint colocalization test for signals from DNA methylation, gene expression and cell function. *MOLOC* expanded the original formulation of *COLOC* to include multiple traits (Giambartolomei et al., 2018). These traits can be independent GWAS, molecular traits or a combination of both. To show the utility of their framework, the authors considered an example case with three traits: GWAS variants for schizophrenia, gene expression and DNA methylation (mQTLs) in the human brain. They showed that adding a third trait significantly increased the power to link variants to genes, as evidenced by 39 new candidate target genes which could only be identified when combining mQTLs and eQTLs. However, these improvements come at the expense of interpretability, increasing the number of possible hypotheses at a locus to 15. Further increases in the number of traits would make the interpretation of colocalization results even more challenging.

Importantly, a trait association signal can result from multiple causal variants (allelic heterogeneity, AH) and recent studies estimate that 20% of the loci identified by GWAS or eQTL-mapping could show AH (Hormozdiari et al., 2017). Methods which assume a single causal variant could potentially misclassify AH cases as colocalizations (Giambartolomei et al., 2014). One method that accounts for multiple causal SNPs per locus is *eCAVIAR* (Hormozdiari et al., 2016) a modified version of the Bayesian method *CAVIAR*, originally designed to perform statistical fine-mapping (Hormozdiari et al., 2014) by estimating the posterior probability of causality for each variant at a GWAS locus (Schaid et al., 2018). Hormozdiari et al. (2016) proposed that fine-mapping could be applied independently to GWAS and QTL associations, and then integrated. Specifically, they defined the probability of a colocalization as the product of the probabilities that the variant was causal in the GWAS and in the eQTL (i.e., the product of the posterior probabilities derived from fine-mapping). Because this approach estimates a posterior probability for each SNP, it does not assume a single causal variant per locus. Instead, *eCAVIAR* can be extended to find colocalizations under the assumption of any number of causal SNPs while accounting for LD.

Colocalization can also be combined with SNP-enrichment, as demonstrated by the statistical method *ENLOC* (Wen et al., 2017). In brief, the authors reasoned that if the majority of GWAS SNPs for a trait are also eQTLs in a given cell type (i.e., if GWAS SNPs are enriched in eQTLs), then most overlaps between the two traits will be driven by true colocalizations. In contrast, if GWAS SNPs are not enriched in eQTLs in that cell type, more of the overlaps are expected to be due to chance. Thus, the authors first estimate an SNP enrichment score and then weigh the priors of their Bayesian model by the identified scores. The authors argue that this approach significantly improves the performance of both fine-mapping and colocalization.

Finally, the effects of GWAS variants are not restricted locally to the genes in close proximity and could have more distal effects (*trans* eQTLs). For example, a GWAS variant could affect the expression of a TF, which would result in a cascade of effects on downstream genes. *Trans* eQTLs are located far away from their target genes and tend to have small effect sizes, which makes them extremely challenging to map at moderate sample sizes due to the burden imposed by multiple testing. In addition, *trans* eQTLs are estimated to be substantially more numerous than *cis* eQTLs (Liu X. et al., 2019), potentially leading to many false positive colocalizations. However, He et al. (2013) reasoned that, while a colocalization between one *trans* eQTL and one GWAS SNP is very likely to be a false positive, the presence of colocalizations between multiple *trans* eQTLs for the same gene and multiple SNPs from the same GWAS is unlikely to be due to chance. Thus, they proposed that the association signals for two traits (e.g., a complex trait and the expression of a gene) could be compared not locally but genome-wide, analogously to comparing two "fingerprints" or "signatures." If two traits tend to have the same signature, they are said to colocalize. The authors applied their method (*Sherlock*) to integrate summary statistics from a GWAS for type 2 diabetes (T2D) with 3,210 *cis* and 242 *trans* eQTLs specific to the liver (Schadt et al., 2008). This analysis identified four candidate genes regulated by T2D variants, two of which acted *in trans* and would have thus been missed by traditional colocalization approaches. Importantly, three of these four genes (*TSPAN8*, *GNB5*, and *JAZF1*) were supported by previous functional studies. The increasing sample sizes of gene expression studies are allowing us to systematically map *trans* eQTLs (Westra et al., 2013) and will provide more statistical power to detect meaningful colocalizations between GWAS and *trans* eQTLs.

Application of Colocalization to Complex Diseases

One of the areas where colocalization analysis has been particularly informative is in identifying the mechanisms underlying immune-mediated diseases. A study by Fortune et al. (2015) used colocalization to investigate the shared etiology of complex immune diseases. The authors investigated 126 GWAS loci associated with type 1 diabetes, rheumatoid arthritis, celiac disease and multiple sclerosis and identified 33 to be shared across these four diseases. Colocalization revealed that at 14 of these regions the causal variants were likely to be different. In contrast, the remaining loci showed evidence of a single causal variant affecting all traits. For example, the associations at the *CTLA4* locus colocalized between the three tested diseases. Interestingly, the authors also found three significant colocalizations between type 1 and type 2 diabetes loci, suggesting that these diseases could share certain aspects of their etiology, despite type 1 diabetes having an immune origin.

Colocalization has also pointed to genes and functional elements involved in these diseases. A study by Huang et al. (2017) fine-mapped variants associated with IBD and integrated them with eQTLs mapped in immune cells. The authors found that a large number of IBD variants colocalized with eQTLs in CD4+ T cells (Huang et al., 2017). However, in a separate study immune disease risk variants (including IBD variants) were tested for colocalization with eQTLs across three immune cell types (lymphoblastoid cells, CD4+ T cells and monocytes) (Chun et al., 2017) and it was found that the majority of loci did not colocalize with eQTLs. The authors concluded that GWAS variants could act via more complicated mechanisms and regulate other molecular traits rather than gene expression. A study by Bossini-Castillo et al. (2019) mapped QTLs for gene expression and chromatin traits (histone modifications and chromatin accessibility) in regulatory CD4+ T cells, a rare cell type that plays a central role in regulating the immune response. The authors integrated chromatin and gene expression QTLs with GWAS loci for 14 immune-mediated diseases and identified 253 colocalizations, the majority of which implicated histone acetylation (H3K27ac) QTLs (acQTL). Interestingly, over 70% of these acQTLs were not linked to any eQTL effects, i.e., the loci were associated with local chromatin regulatory activity but not with the expression of nearby genes. A proportion of these colocalizations could represent context-specific eQTLs, which would only be detected upon exposure of the cells to the correct environmental cues. This is known to be the case for other immune cells such as human macrophages, where exposure to cytokines or pathogens has been shown to induce context-specific chromatin accessibility and expression QTLs (Alasoo et al., 2018).

Another area where colocalization has been particularly informative is cardiovascular disease. Franceschini et al. (2018) performed GWAS meta-analyses of two cardiovascular traits (carotid plaque burden and carotid artery thickness) and tested the variants for colocalization with vascular tissue eQTLs, with the aim of investigating the molecular mechanisms underlying cardiac phenotypes. This analysis prioritized two candidate genes (*CCDC71L* and *PRKAR2B*) which colocalized with both traits, suggesting potential disease mechanisms in which regulation of gene expression in arterial smooth muscle impacts artery thickness and plaque formation, ultimately leading to atherosclerosis. In a separate study Liu et al. (2018) integrated GWAS loci for coronary artery disease (CAD) with expression and splicing QTLs mapped in smooth muscle cells from 52 individuals. The authors identified five significant colocalizations (*FES, SMAD3, TCF21, PDGFRA*, and *SIPA1*) and found that increased levels of *TCF21* and *FES* were associated with reduced risk of CAD. Importantly, all of the genes were involved in vascular remodeling, strengthening the hypothesis that gene expression in arterial smooth muscle could have an important impact in local tissue architecture, thus modifying the risk of several correlated cardiovascular traits.

Finally, colocalization analysis can also inform about the relationship between shared genetic architectures across complex traits. A study by Pickrell et al. (2016) used results from 43 GWAS for 42 traits, including neurological phenotypes, anthropometric traits, social traits, immune-mediated disease, metabolic phenotypes, and hematopoietic traits. The authors developed a method (*gwas-pw*) which tested for colocalization between all possible pairwise combinations of these 42 traits and then grouped together those for which there was substantial evidence of colocalization across multiple loci (Pickrell et al., 2016). Most of the traits showed few colocalizations with each other. Nonetheless, the analysis identified two groups of traits (10 traits in total) with a higher number of colocalizations with each other than expected by chance. The first group contained metabolic phenotypes (triglycerides, HDL cholesterol, LDL cholesterol, and CAD), while the second group contained hematopoietic traits (red blood cell volume, hemoglobin concentration and platelet count, among others). The large number of colocalizations in the second group suggests pleiotropic effects across the associated variants, which could indicate that the same variants are able to regulate the differentiation of several independent hematopoietic lineages.

Twas: Direct Association of Genes and Traits

The examples outlined so far rely on colocalization analyses using genome-wide significant SNPs to nominate causal genes for complex traits. However, the majority of variants contributing to complex phenotypes have not yet been identified, as their effect sizes are too small to be detected at current GWAS sample sizes (Visscher et al., 2017). Another way to gain insights into the biology of complex traits is by directly testing for association between a trait and gene expression (i.e., identifying which genes are expressed at a significantly different level in cases compared to controls in disease-relevant cell types). Given that the number of genes is substantially lower than the number of common variants, using gene expression rather than genotypes for association benefits from a reduced multiple testing burden. Nonetheless, carrying out such a study is currently unfeasible, as it would require profiling gene expression across hundreds of thousands of individuals in both cases and controls, and across dozens of tissues. Alternatively, cell type-specific gene expression profiles can be predicted (i.e., imputed) based on genotypes, thus obviating the need to perform costly RNA-sequencing experiments. *Transcriptome-wide association studies* (TWAS) leverage information from GWAS and eQTL catalogs to predict the transcriptome of cases and controls, thus allowing the direct association of traits and genes without directly profiling gene expression in every individual included in the GWAS (Wainberg et al., 2019).

Predicting expression of a gene based on genotypes is possible because gene expression is highly heritable (Wright et al., 2014) and most of the gene expression heritability is attributable to variants in proximity (*in cis*) to the genes (Lloyd-Jones et al., 2017). TWAS uses tissue-specific eQTL maps as reference datasets to train predictors that take an individual's genotype as an input and estimate their transcriptome levels (Gamazon et al., 2015; Gusev et al., 2016; **Figure 4A**). These predictors use only information from SNPs *in cis* to the genes and are restricted to genes with highly heritable expression. This prediction process is analogous to genotype imputation and allows for direct association between a trait and the expression of each gene (**Figure 4B**). Moreover, by focusing on the heritable component of gene expression, it minimizes the confounding by disease-caused changes in gene expression.

PrediXcan (Gamazon et al., 2015), an implementation of TWAS, uses an elastic net model to predict gene expression from eQTL catalogs. The authors applied this approach to data from the Wellcome Trust Case Control Consortium (WTCCC) (Wellcome Trust Case Control Consortium, 2007) and identified 41 genes associated with five complex diseases. The majority of these genes were known candidates from GWAS, while others

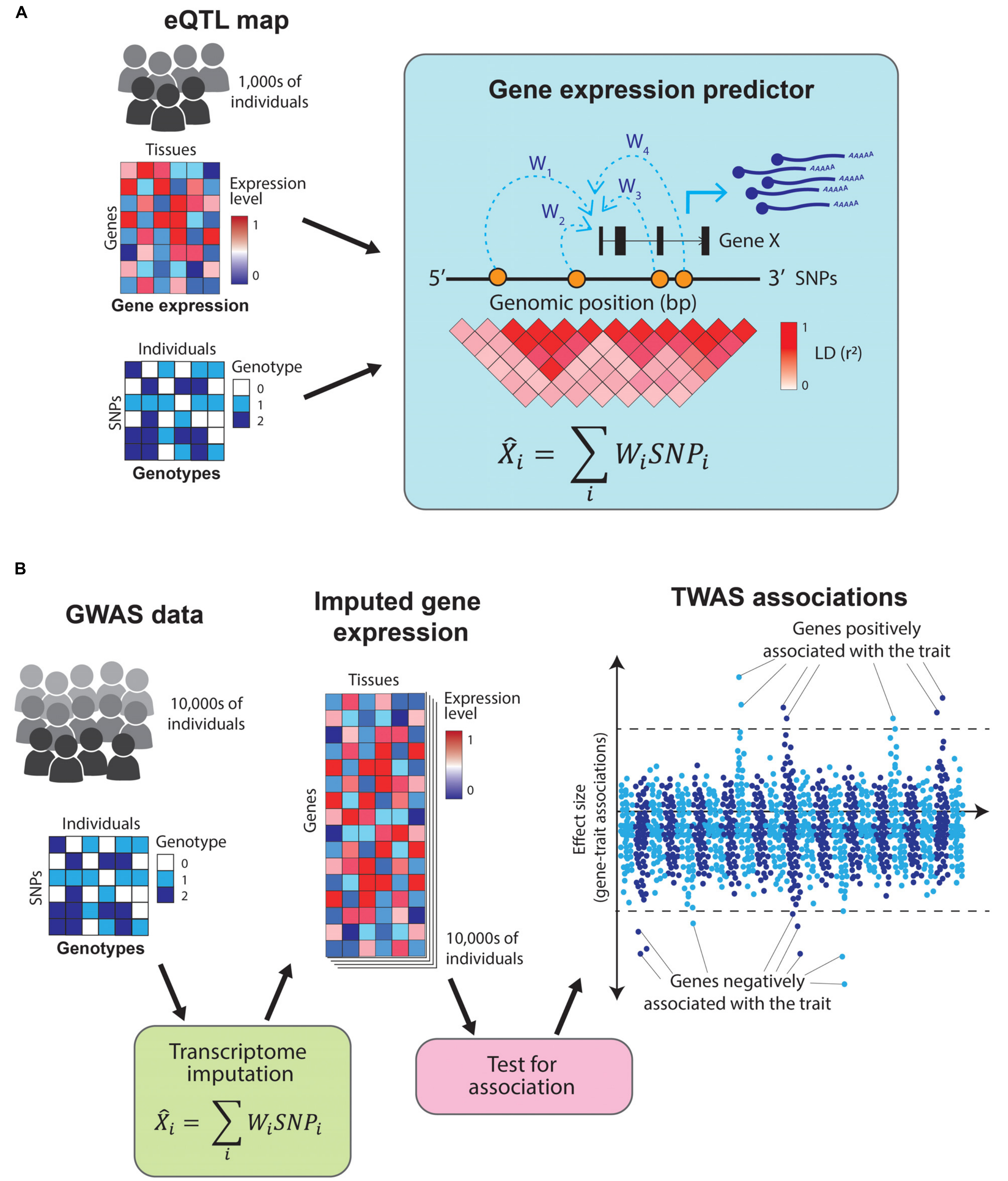


FIGURE 4 | Overview of transcriptome-wide association studies TWAS leverage information from eQTL catalogs and GWAS studies to directly associate traits to genes. **(A)** TWAS use eQTL maps (which contain tissue-specific gene expression and genotypes for thousands of individuals) as a training set to build gene expression predictors. These predictors take the SNPs *in cis* to a gene and estimate its expression levels. **(B)** The resulting predictors are used to impute gene expression values across the hundreds of thousands of individuals in a GWAS study (which contains genotypes but no gene expression data). Finally, the imputed gene expression values are directly tested for association with the GWAS trait, resulting in a set of genes which positively or negatively influence it.

(e.g., *KCNN4* and *PTPRE*) had not been implicated in the diseases before. Importantly, because TWAS directly associates traits to genes, the associations have a clear directionality of effects. As an illustration, a SNP nearby *ERBB3* had been previously associated with type 1 diabetes (Hakonarson et al., 2008). *PrediXcan* confirmed the association between *ERBB3* and type 1 diabetes and found that low *ERBB3* expression increased disease risk (Gamazon et al., 2015). Defining the directionality of effects of GWAS variants, and particularly identifying risk variants which increase gene expression, can nominate effective drug targets and accelerate the development of new therapies.

To overcome the requirement for individual-level genotypes, the authors of *PrediXcan* subsequently derived a mathematical formulation (*S-PrediXcan*) which achieves comparable results using GWAS summary statistics (Barbeira et al., 2018). The authors applied *S-PrediXcan* to over 100 phenotypes across 44 GTEx tissues and found that most of the associations detected were tissue-specific, highlighting the need to profile gene expression in disease-relevant cell types. For example, LDL levels were positively associated with *SORT1* expression only in the liver and negatively associated with *PCSK9* only in tibial nerve. In contrast, schizophrenia was negatively associated with *C4A* expression across 42 of the 44 tissues tested (Barbeira et al., 2018).

Because most of the SNPs used to predict gene expression in TWAS are enriched in regulatory DNA (Trynka and Raychaudhuri, 2013), including epigenetic annotations in the model can improve transcriptome imputation. *EpiXcan* is an implementation of *PrediXcan* which takes into account annotations such as DNA methylation or histone modifications (Zhang et al., 2019). The contribution of each SNP in the prediction is weighted by its overlap with regulatory elements in a Bayesian hierarchical model. When applied to 58 traits and 14 eQTL data sets, *EpiXcan* increased the number of gene-trait associations by over 18% compared to *PrediXcan*. Most of these associations were tissue-specific. For example, TWAS associations with CAD were only detected in arterial tissue, while schizophrenia associations were specific to the brain (Zhang et al., 2019). Moreover, integrating *EpiXcan* with a catalog of chemical perturbations revealed drug repurposing opportunities. An example is ursolic acid, which can reverse the gene expression changes associated with BMI. This compound is currently under investigation for the treatment of obesity (Kunkel et al., 2012).

Another TWAS approach proposed by Gusev et al. (2016) uses a Bayesian predictor to impute gene expression from genotypes. First, the method determines the weights of the Bayesian predictor based on a reference eQTL catalog. The contributions of each variant to the predictions are proportional to its eQTL effects on each gene. Next, gene expression is imputed directly from the GWAS summary statistics. To do this, the authors first use the summary statistics to impute the GWAS effect sizes of all common variants (Pasaniuc et al., 2014) and then multiply these effect sizes by the Bayesian weight of each variant (determined from the eQTL catalog as previously described). Each variant is then re-weighed by its LD with other variants in the locus. Finally, the contribution of all variants proximal to a gene is combined into a single expression-trait association estimate. The authors used this approach to find genes involved in the regulation of circulating lipid levels (HDL, LDL, total cholesterol, and triglycerides). This analysis nominated 665 lipid-associated genes, of which 66 had not been previously identified by any of the independent GWAS (Gusev et al., 2016). The majority of these novel genes showed additional functional evidence from mouse studies. For example, *FTSJ3* expression correlated with fat mass and glucose-to-insulin ratio in mice, while *ITIH4* correlated with LDL levels.

Gusev et al. (2018) subsequently extended their approach to epigenetic data. The authors performed a TWAS to test for association between gene expression in brain tissue and risk for schizophrenia, including as an additional layer of information chromatin marks (i.e., H3K27ac, H3K4me1, and H3K4me3) assayed in 76 lymphoblastoid cell lines. This allowed them to nominate both genes and regulatory elements involved in disease. For example, the authors found two chromatin elements associated with *MAPK3* expression, which was in turn associated with schizophrenia risk. They then functionally validated this association, showing that *MAPK3* is involved in a neuro-proliferation phenotype in zebrafish (Gusev et al., 2018).

Finally, *summary data-based Mendelian randomization* (*SMR*) uses a Mendelian randomization (MR) framework to perform a TWAS analysis (Zhu et al., 2016). MR takes advantage of the fact that an individual's genotype is independent of confounding factors such as nurture or environmental covariates. In traditional MR, genotypes are used as an instrumental variable to infer causal relationships between an exposure (e.g., the levels of a metabolite or protein) and a trait (e.g., a disease) (Evans and Davey Smith, 2015). In *SMR*, an analogous approach is used to infer associations between gene expression and a trait. In brief, the authors use genetic variants as instrumental variables and estimate the effect size of a gene in a trait as the ratio of the GWAS effect size to the eQTL effect size of a variant affecting the expression of the gene (Zhu et al., 2016). Traditional TWAS approaches impute gene expression from genotypes and then associate genes to traits. However, because imputation is based on the combined effects of multiple proximal variants, TWAS cannot directly point to the individual variants underlying gene-trait associations. In contrast, *SMR* estimates a separate gene-trait effect size from each individual SNP in a locus, thus making it possible to link variants to genes. By comparing the effect-sizes derived from all the SNPs in a locus, *SMR* is able to identify cases in which a single variant affects both gene expression and a complex trait. This test (*HEIDI*) is a form of colocalization analysis (Zhu et al., 2016). However, since most gene-trait effects are small due to polygenicity (Boyle et al., 2017), *SMR* requires eQTL catalogs of very large sample size. The authors applied *SMR* to a large peripheral blood eQTL study (5,311 samples) (Westra et al., 2013) and identified 289 genes associated with body-mass index, waist-hip ratio, rheumatoid arthritis and schizophrenia. Of these, 104 genes showed evidence of a single causal variant. An interesting example includes a locus associated with rheumatoid arthritis which contains the genes *TRAF1* and *C5*. Based on its function, *TRAF1* had been prioritized as the most likely target gene. *SMR* confirmed the prioritization of *TRAF1* and provided evidence of a single causal variant in the region (Zhu et al., 2016).

In summary, colocalization and TWAS prioritize the genes causally involved in complex diseases. Colocalization analysis integrates association signals from GWAS and QTLs in a locus by locus basis to identify instances in which both traits share a causal variant. In contrast, TWAS leverages information from eQTL catalogs to impute gene expression values and directly associate genes to traits. The availability of QTL catalogs from a wider variety of cell types, as well as of larger sample sizes, will improve gene prioritization and translate GWAS results to refined sets of disease-causal genes.

FUTURE PERSPECTIVES IN INTERPRETING GWAS ASSOCIATIONS

Enrichment and colocalization analyses have prioritized tissues and genes involved in complex diseases. However, these approaches are largely limited by the availability of comprehensive reference functional data sets. For example, enrichment and colocalization mostly rely on gene expression data from bulk tissues. However, gene expression profiles from bulk tissue are dominated by the most abundant cell types and do not capture information about cell composition and cell type frequencies (Trapnell, 2015). Moreover, colocalization methods are purely observational and cannot establish causality. For example, a SNP could affect both a gene and a trait via independent mechanisms (i.e., pleiotropy), and colocalization cannot conclusively distinguish this scenario from a single causal variant. Thus, candidate genes require additional experimental validation to unambiguously establish causality, for example, by integrating GWAS variants with single-cell assays, or validating candidate genes with gene-editing technologies.

Integration of Gwas With Single-Cell Genomics

Single-cell genomic assays enable quantification of molecular traits at the single-cell level. For example, multiple existing methods allow profiling gene expression (Picelli et al., 2013; Macosko et al., 2015; Kimmerling et al., 2016; Zheng et al., 2017), chromatin accessibility (Buenrostro et al., 2015), and TF occupancy (Rotem et al., 2015; Grosselin et al., 2019) with single-cell resolution. These assays can resolve the cellular composition of complex organs and tissues, and are used to assemble cells into reference tissue atlases (Regev et al., 2017). Moreover, they can order differentiating cells into time-course trajectories that span different stages of differentiation, an approach called pseudotime ordering (Saelens et al., 2019).

The high resolution of single-cell genomic maps makes them a promising resource for SNP enrichment analysis. This is illustrated by a recent GWAS of hematological traits like hematocrit, hemoglobin and blood cell counts (Ulirsch et al., 2019). In this study, the authors integrated fine-mapped GWAS variants with bulk and single-cell chromatin accessibility profiles spanning a large number of hematopoietic and progenitor cell lineages. The authors developed a SNP enrichment test (*g-chromVAR*) which integrates the quantitative levels of chromatin accessibility in each single cell with the posterior probabilities of causality of each variant inferred from fine-mapping. Enrichment estimates varied throughout the differentiation trajectory and concentrated at specific stages of hematopoiesis. For example, variants associated with platelet counts were progressively more enriched as cells differentiated into megakaryocytes, the precursors of platelets. Conversely, enrichment decreased along differentiation toward the lymphoid lineage. With the rapid increase in the number, depth and size of single-cell datasets, more studies like this will soon be possible and applicable to a whole range of complex traits. However, single-cell genomic approaches introduce new challenges to the current statistical methods, such as data size, sparsity, and high dropout rates (Lähnemann et al., 2020). Thus, it will be essential to develop new statistical methods designed to deal with the intricacies of single-cell data.

Single-cell technologies can also expand the current scope of colocalization. Because the throughput of these assays is growing at an unprecedented scale, it is now possible to profile single-cell transcriptomes in large scale populations of individuals, allowing to map single-cell eQTLs (sc-eQTLs). One such study profiled gene expression in 45,000 single-cells isolated from peripheral blood of 45 healthy individuals (van der Wijst et al., 2018) and identified eQTLs with opposite effects in different cell types in blood. For example, rs4804315 increased the expression of *ZNF414* in NK cells but decreased it in T cells. Moreover, the authors also recapitulated two previously reported monocyte eQTLs for the *HLA-DQA1* and *CTSC* genes and showed that they were specific to the classical monocyte subpopulation (van der Wijst et al., 2018). These results would be difficult to obtain from bulk gene expression measurements. This study serves as a proof of concept and shows how single-cell eQTL associations could rapidly become available for integration with GWAS.

An additional advantage of single-cell sequencing is the possibility of ordering cells into time-course trajectories, thus adding a temporal component to the association models used for eQTL-mapping. This permits the identification of eQTLs with different effect sizes at different stages of differentiation (dynamic eQTLs). Two studies mapped dynamic eQTLs during the differentiation of human induced pluripotent stem cells (iPSCs). The first study investigated iPSC differentiation into endoderm (Cuomo et al., 2020). The authors profiled single-cell gene expression at four time points across 125 iPS cell lines and ordered cells into a time-course trajectory spanning distinct cell states. This uncovered 785 dynamic eQTLs. Interestingly, this study was able to map eQTLs with a cell cycle-dependent effect size. The second study focused on cardiomyocyte differentiation and mapped eQTLs at 16 time points across 19 iPS cell lines (Strober et al., 2019). Here, the authors ordered cells in time-course trajectories based on bulk RNA expression profiles and identified modules of genes which increase or decrease along differentiation. Next, they performed eQTL-mapping using a Gaussian model which accounted for the interaction between genotypes and differentiation time. This resulted in the identification of 550 genes with linear and 693 genes with non-linear dynamic eQTL effects. Interestingly, two dynamic eQTLs which regulated the expression of *SCN5A* (a gene altered in dilated cardiomyopathy) were also GWAS variants for QRS and QT interval duration, thus suggesting that dysregulation of gene expression dynamics could have important phenotypic consequences. Until now, colocalization has not been applied to this type of data. However, as the sample sizes of sc-eQTL and dynamic eQTL catalogs grow, they will become an increasingly important resource for identifying subtle changes in gene expression dynamics which lead to disease.

Integration of Polygenic Risk Scores With Functional Annotations

Genome-wide association studies variants can be used to identify individuals at high risk of disease. This can be achieved by combining hundreds of disease associated-variants carried by an individual into a single score that reflects their overall genetic risk, a polygenic risk score (PRS) (Chatterjee et al., 2016). The integration of PRSs with epidemiological risk factors such as age, sex, smoking status, diet, or family history of disease could improve the stratification of individuals, potentially resulting in more effective clinical interventions (Torkamani et al., 2018). To build a PRS, a subset of variants is selected based on their GWAS association. Next, each variant is assigned a weight, which corresponds to its standardized effect size (i.e., the odds ratio from the GWAS multiplied by the effect direction). Finally, the genetic dosage of each individual variant (i.e., 0, 1, and 2 according to the number of risk alleles carried) is multiplied by its weight, and all loci across the genome are added into a single score. PRSs are often normally distributed and individuals can be grouped by PRS decile, with those in the top deciles being at highest risk (for a detailed discussion refer to the review by Chatterjee et al., 2016).

Polygenic risk scores performance has increased as GWAS studies increased in sample sizes and larger validation cohorts became available, as shown in CAD (Ripatti et al., 2010; Mega et al., 2015; Abraham et al., 2016; Khera et al., 2016) and cancer (Garcia-Closas et al., 2013; Mavaddat et al., 2015; Maas et al., 2016). The availability of large-scale biobanks (Gaziano et al., 2016; Nagai et al., 2017; Bycroft et al., 2018) has enabled unparalleled improvements in this area by linking genetic information with electronic health records for hundreds of thousands of individuals. Two of the largest PRS studies leveraged UK BioBank data to estimate CAD risk using up to 6.6 million SNPs (Abraham et al., 2016; Khera et al., 2018). Khera et al. (2018) demonstrated that individuals at the highest PRS percentiles were at a risk equivalent to that of carrying a monogenic mutation for familial hypercholesterolemia. Another study used 2.1 million SNPs to build an obesity PRS (Khera et al., 2019) and demonstrated that PRSs can stratify individuals before phenotypic differences appear. While the authors observed no differences in birthweight of individuals at different PRS deciles, these became apparent when individuals reached puberty.

Despite these advancements, polygenic scores face severe challenges. Firstly, prediction accuracy remains low. Secondly, PRSs are based on European GWASs and their transferability between populations is low (Martin et al., 2017, 2019). This is alarming, as it could result in misdiagnosis of individuals in underrepresented populations (Manrai et al., 2016). Finally, little is known about the functional mechanisms underlying PRSs. Some of these challenges are now being tackled using functional annotations.

Prediction accuracy is dependent on the SNPs used to build the PRS. In particular, GWAS effect sizes can be confounded by LD (Bulik-Sullivan et al., 2015). To minimize this, SNPs are pruned by LD and thresholded by P value, but this can eliminate causal SNPs in LD with each other. To circumvent this, *LDpred* uses a Bayeseian model to shrink the effect sizes of each variant (Vilhjálmsson et al., 2015) based on a prior that models the effect sizes with an LD-informed normal distribution. The PRS constructed in this way outperformed other methods. *LDpred-func* extended *LDpred* by including the overlap between variants and functional elements in the Bayesian prior (Márquez-Luna et al., 2018). By segmenting the genome into coding, conserved, and regulatory elements, *LDpred-func* improved prediction estimates for height. An equivalent method, *AnnoPred*, also uses a Bayesian model to create functionally informed polygenic scores, outperforming traditional PRSs for breast cancer (Hu et al., 2017). A further study leveraged gene co-expression networks in the brain to identify modules of genes with a common regulation (Hari Dass et al., 2019). Based on these modules, the authors identified genes co-expressed with the insulin receptor and used SNPs in proximity to build a PRS that incorporates known disease biology. Nonetheless, using prior knowledge to design PRSs can introduce bias and requires further evaluation.

Functional annotations can also improve transferability of PRS across populations. Despite the LD difference between populations, most causal variants are thought to be shared (Marigorta and Navarro, 2013). Moreover, they often overlap functional annotations which are also shared between populations (Tehranchi et al., 2019). Thus, overlapping GWAS signals with functional annotations (i.e., functional fine-mapping) can increase the chance of including the functional SNPs in a PRS regardless of the population. A recent study leveraged cell type-specific binding of TFs and epigenetic marks in 245 cell types to identify the annotations most enriched for disease heritability. SNPs overlapping these annotations were used to build PRSs for 29 traits (Amariuta et al., 2020). Using the UK and Japan BioBanks, the authors demonstrated that population transferability improved when incorporating functional annotations.

Biobanks can also help in functionally interpreting PRSs. Richardson et al. (2019) used GWAS variants and UK BioBank data to build 162 PRSs spanning traits as varied as anthropometric measurements, cardiovascular traits, and ICD10 codes. They identified traits correlated with each other based on their polygenic scores and used MR to infer causality. Polygenic scores for triglyceride levels, urate levels, LDL, and gout were significantly correlated with each other. MR analysis revealed evidence that elevated triglycerides cause higher urate production, which in turn increases risk of gout. A similar study derived PRSs for blood traits such as hematocrit and cell counts (Xu et al., 2020) and correlated them with disease PRSs. This pinpointed disease-relevant traits, e.g., the PRS for eosinophil counts was highly correlated with the PRS for allergies.

Finally, gene expression is also beginning to be integrated with PRSs. Võsa et al. (2018) mapped *cis* and *trans* eQTLs in a meta-analysis of 31,684 samples from 37 cohorts. They subsequently identified genes affected by dozens of *trans* eQTLs and proposed that such genes could act as hubs where biological processes converge, potentially accumulating a disproportionate amount of genetic risk for complex diseases. These genes are roughly equivalent to the *core genes* proposed by the omnigenic model (Boyle et al., 2017; Liu X. et al., 2019). To identify these hubs, the authors defined quantitative trait scores (QTS) as the associations between the expression of a gene and the PRS of a disease. They mapped 2,658 eQTS genes, including a group of IFN-regulated genes which were correlated with lupus PRS. In the future, increases in the sample sizes of eQTL studies may enable systematic mapping of cell-type specific eQTSs.

Validation of Gwas Findings Using Gene Editing

Recent years have seen a rapid expansion in the number and efficacy of gene-editing tools. In particular, CRISPR/Cas9 allows the deletion of specific sections of the genome with high accuracy (Wang et al., 2016). CRISPR-based approaches have been used to systematically knock down genes genome-wide, an approach referred to as CRISPR screening (Koike-Yusa et al., 2014). The applications of CRISPR screening are numerous. For example, it can be used to investigate which genes are essential for cancer growth, which in turn provides a platform for drug target identification (Behan et al., 2019).

Coupling CRISPR-editing platforms with informative functional readouts could be a powerful approach to validate GWAS results. For example, a recent study asked which genes are essential for T cell activation by systematically knocking-down all genes in primary human T cells and measuring proliferation upon stimulation (Shifrut et al., 2018). A second study used a similar approach to investigate T helper cell differentiation in mice (Henriksson et al., 2019). These studies are relevant in the context of complex immune diseases, for which GWAS variants are thought to act during T cell activation and differentiation (Calderon et al., 2019; Soskic et al., 2019). Nonetheless, using CRISPR to follow-up candidate genes requires previous knowledge regarding which functional assays are the most disease-relevant. For example, neuronal cell types are thought to be implicated in psychiatric traits (Finucane et al., 2018), but it is not known which specific neuronal functions are compromised in disease, and thus it is uncertain what the best readout for a CRISPR-screen would be. Selecting informative assays may require mapping the genetic architecture of cellular and intermediate traits. A recent study showed that variants which modulate secretion of monocyte cytokines (cytokine-QTLs) tend to be associated with susceptibility to infection (Li et al., 2016). Thus, a CRISPR-screen to validate infection susceptibility genes should probably assess cytokine secretion. Alternatively, single-cell gene expression can also be used as a readout for CRISPR-screens. Due to its high resolution, single-cell sequencing can match the transcriptome of cells with their corresponding guide RNAs. This is the basis of methods like CROP-seq and Perturb-seq (Dixit et al., 2016; Datlinger et al., 2017) that have been used to investigate which genes are essential in processes such as dendritic cell response with single-cell resolution. In the future, high-throughput phenotyping of human cells will be crucial for identifying the best assays to validate candidate GWAS genes.

Gene-editing approaches can also be used to study the non-coding genome. For example, CRISPR-interference (CRISPRi) uses guide RNAs and a defective version of the Cas9 enzyme to prevent regulatory elements from contacting their target genes (Qi et al., 2013). In contrast, CRISPR-activation (CRISPRa) uses a transcriptional activator fused to the Cas9 protein to enhance transcription (Bikard et al., 2013). These tools can be used to map the function of disease-associated regulatory elements. Moreover, deep mutagenesis employs error-prone PCR to randomly mutate all the nucleotides in a regulatory sequence one at a time (McCullum et al., 2010). Mutagenesis is often coupled either with the expression of a reporter gene like luciferase or with a sequencing-based readout. A recent study used deep-mutagenesis followed by sequencing to study the function of each nucleotide in 20 regulatory elements associated with rare and common diseases (Kircher et al., 2019), including the well-known LDL-associated locus near *SORT1* (Musunuru et al., 2010). This enabled the systematic identification of clusters of nucleotides for which mutation significantly alters gene expression. Importantly, these sites often contained known GWAS SNPs and corresponded to TF binding sites, thus suggesting a molecular mechanism for the implicated variants. Another study investigated loci associated with hematological traits using fine-mapping followed by deep mutagenesis (Ulirsch et al., 2016). The authors found strong regulatory effects for 32 variants (corresponding to 23 lead SNPs from GWAS) of which three had a clear molecular mechanism. These approaches could transform our understanding of how genetic variants affect organismal phenotypes.

Ideally, gene-editing should be performed in disease-relevant cell types (for example, in cells prioritized by SNP enrichment). However, current gene-editing approaches are mostly limited to cell lines. The reasons for this are varied. The application of mutagenesis to primary cells is hindered by the large numbers of cells required and the need to keep cells in culture for prolonged periods of time. CRISPR-editing is further limited by the p53-dependent cellular toxicity which accompanies Cas9-induced double-strand breaks (Ihry et al., 2018). Methodological advances such as better systems for Cas9 delivery (DeWitt et al., 2017; Shifrut et al., 2018) will likely overcome some of these limitations. However, further technological development is needed to routinely apply gene-editing as a follow up strategy for GWAS.

CONCLUSION

The integration of GWAS associations with cell type-specific functional data has significantly furthered our understanding of how genetic variation leads to disease. On the one hand, SNP enrichment approaches have enabled the prioritization of cell types and tissues based on their disease-relevance. These methods work by testing for the accumulation of variants in regulatory elements specific to a given cell type. They can either be restricted to genome-wide significant variants or estimate enrichments based on the contributions of all common SNPs. On the other hand, colocalization analysis integrates eQTL and GWAS associations to identify the target genes of GWAS loci, leveraging LD information and association patterns. Moreover, TWAS allows the direct association of genes with phenotypes via transcriptome imputation. These approaches are beginning to reveal the tissues and genes affected in complex diseases like autoimmunity, schizophrenia and coronary heart disease. However, they are limited by the resolution of current functional datasets and cannot establish causality. In the future, we anticipate that the integration of GWAS with single-cell data and the validation of candidate genes via gene-editing and cellular phenotyping will help us translate GWAS findings into clinically actionable gene sets.

AUTHOR CONTRIBUTIONS

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

FUNDING

GT is supported by the Wellcome Trust (grant WT206194) and by Open Targets. EC-G was supported by a Gates Cambridge Scholarship (OPP1144).

REFERENCES

Abraham, G., Havulinna, A. S., Bhalala, O. G., Byars, S. G., De Livera, A. M., Yetukuri, L., et al. (2016). Genomic prediction of coronary heart disease. *Eur. Heart J.* 37, 3267–3278.

Alasoo, K., Rodrigues, J., Mukhopadhyay, S., Knights, A. J., Mann, A. L., Kundu, K., et al. (2018). Shared genetic effects on chromatin and gene expression indicate a role for enhancer priming in immune response. *Nat. Genet.* 50, 424–431. doi: 10.1038/s41588-018-0046-7

Amariuta, T., Ishigaki, K., Sugishita, H., Ohta, T., Matsuda, K., Murakami, Y., et al. (2020). In silico integration of thousands of epigenetic datasets into 707 cell type regulatory annotations improves the trans-ethnic portability of polygenic risk scores. *bioRxiv [Preprint]*

Bannister, A. J., and Kouzarides, T. (2011). Regulation of chromatin by histone modifications. *Cell Res* 21:381. doi: 10.1038/cr.2011.22

Banovich, N. E., Lan, X., McVicker, G., van de Geijn, B., Degner, J. F., Blischak, J. D., et al. (2014). Methylation QTLs are associated with coordinated changes in transcription factor binding, histone modifications, and gene expression levels. *PLoS Genet.* 10:e1004663. doi: 10.1371/journal.pgen.1004663

Barbeira, A. N., Dickinson, S. P., Bonazzola, R., Zheng, J., Wheeler, H. E., Torres, J. M., et al. (2018). Exploring the phenotypic consequences of tissue specific gene expression variation inferred from GWAS summary statistics. *Nat. Commun.* 9:1825.

Barski, A., Cuddapah, S., Cui, K., Roh, T.-Y., Schones, D. E., Wang, Z., et al. (2007). High-resolution profiling of histone methylations in the human genome. *Cell* 129, 823–837. doi: 10.1016/j.cell.2007.05.009

Beatrix Bartok, G. S. F. (2010). Fibroblast-like synoviocytes: key effector cells in rheumatoid arthritis. *Immunol. Rev.* 233:233. doi: 10.1111/j.0105-2896.2009.00859.x

Behan, F. M., Iorio, F., Picco, G., Gonçalves, E., Beaver, C. M., Migliardi, G., et al. (2019). Prioritization of cancer therapeutic targets using CRISPR-Cas9 screens. *Nature* 568, 511–516. doi: 10.1038/s41586-019-1103-9

Bikard, D., Jiang, W., Samai, P., Hochschild, A., Zhang, F., and Marraffini, L. A. (2013). Programmable repression and activation of bacterial gene expression using an engineered CRISPR-Cas system. *Nucleic Acids Res.* 41, 7429–7437. doi: 10.1093/nar/gkt520

Bossini-Castillo, L., Glinos, D. A., Kunowska, N., Golda, G., Lamikanra, A., Spitzer, M., et al. (2019). Immune disease variants modulate gene expression in regulatory CD4+ T cells and inform drug targets. *bioRxiv [Preprint]*

Boyle, A. P., Davis, S., Shulha, H. P., Meltzer, P., Margulies, E. H., Weng, Z., et al. (2008). High-resolution mapping and characterization of open chromatin across the genome. *Cell* 132, 311–322. doi: 10.1016/j.cell.2007.12.014

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

Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y., and Greenleaf, W. J. (2013). Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods* 10, 1213–1218. doi: 10.1038/nmeth.2688

Buenrostro, J. D., Wu, B., Litzenburger, U. M., Ruff, D., Gonzales, M. L., Snyder, M. P., et al. (2015). Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature* 523, 486–490. doi: 10.1038/nature14590

Bugatti, S., Vitolo, B., Caporali, R., Montecucco, C., and Manzo, A. (2014). B Cells in rheumatoid arthritis: from pathogenic players to disease biomarkers. *Biomed Res. Int.* 2014:681678. doi: 10.1155/2014/681678

Bulik-Sullivan, B. K., Loh, P.-R., Finucane, H. K., Ripke, S., Yang, J., Schizophrenia Working Group of the Psychiatric Genomics Consortium, et al. (2015). LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* 47, 291–295. doi: 10.1038/ng.3211

Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L. T., Sharp, K., et al. (2018). The UK Biobank resource with deep phenotyping and genomic data. *Nature* 562, 203–209. doi: 10.1038/s41586-018-0579-z

Calderon, D., Bhaskar, A., Knowles, D. A., Golan, D., Raj, T., Fu, A. Q., et al. (2017). Inferring relevant cell types for complex traits by using single-cell gene expression. *Am. J. Hum. Genet.* 101, 686–699. doi: 10.1016/j.ajhg.2017.09.009

Calderon, D., Nguyen, M. L. T., Mezger, A., Kathiria, A., Müller, F., Nguyen, V., et al. (2019). Landscape of stimulation-responsive chromatin across diverse human immune cells. *Nat. Genet.* 51, 1494–1505. doi: 10.1038/s41588-019-0505-9

Chatterjee, N., Shi, J., and García-Closas, M. (2016). Developing and evaluating polygenic risk prediction models for stratified disease prevention. *Nat. Rev. Genet.* 17, 392–406. doi: 10.1038/nrg.2016.27

Chen, L., Ge, B., Casale, F. P., Vasquez, L., Kwan, T., Garrido-Martín, D., et al. (2016). Genetic Drivers of Epigenetic and Transcriptional Variation in Human Immune Cells. *Cell* 167, 1398–1414.e24.

Chun, S., Casparino, A., Patsopoulos, N. A., Croteau-Chonka, D. C., Raby, B. A., De Jager, P. L., et al. (2017). Limited statistical evidence for shared genetic effects of eQTLs and autoimmune-disease-associated loci in three major immune-cell types. *Nat. Genet.* 49, 600–605. doi: 10.1038/ng.3795

Cope, A. P., Schulze-Koops, H., and Aringer, M. (2007). The central role of T cells in rheumatoid arthritis. *Clin. Exp. Rheumatol.* 25, S4–S11.

Creyghton, M. P., Cheng, A. W., Welstead, G. G., Kooistra, T., Carey, B. W., Steine, E. J., et al. (2010). Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc. Natl. Acad. Sci. U.S.A.* 107, 21931–21936. doi: 10.1073/pnas.1016071107

Cuomo, A. S. E., Seaton, D. D., McCarthy, D. J., Martinez, I., Bonder, M. J., Garcia-Bernardo, J., et al. (2020). Single-cell RNA-sequencing of differentiating iPS cells reveals dynamic genetic effects on gene expression. *Nat. Commun.* 11:810.

Darmanis, S., Sloan, S. A., Zhang, Y., Enge, M., Caneda, C., Shuer, L. M., et al. (2015). A survey of human brain transcriptome diversity at the single cell level. *Proc. Natl. Acad. Sci. U.S.A.* 112, 7285–7290. doi: 10.1073/pnas.1507125112

Datlinger, P., Rendeiro, A. F., Schmidl, C., Krausgruber, T., Traxler, P., Klughammer, J., et al. (2017). Pooled CRISPR screening with single-cell transcriptome readout. *Nat. Methods* 14, 297–301. doi: 10.1038/nmeth.4177

Degner, J. F., Pai, A. A., Pique-Regi, R., Veyrieras, J.-B., Gaffney, D. J., Pickrell, J. K., et al. (2012). DNase I sensitivity QTLs are a major determinant of human expression variation. *Nature* 482, 390–394. doi: 10.1038/nature10808

Dendrou, C. A., Cortes, A., Shipman, L., Evans, H. G., Attfield, K. E., Jostins, L., et al. (2016). Resolving TYK2 locus genotype-to-phenotype differences in autoimmunity. *Sci. Transl. Med.* 8:363ra149. doi: 10.1126/scitranslmed.aag1974

DeWitt, M. A., Corn, J. E., and Carroll, D. (2017). Genome editing via delivery of Cas9 ribonucleoprotein. *Methods* 12, 9–15. doi: 10.1016/j.ymeth.2017.04.003

Diogo, D., Bastarache, L., Liao, K. P., Graham, R. R., Fulton, R. S., Greenberg, J. D., et al. (2015). TYK2 protein-coding variants protect against rheumatoid arthritis and autoimmunity, with no evidence of major pleiotropic effects on non-autoimmune complex traits. *PLoS ONE* 10:e0122271. doi: 10.1371/journal.pone.0122271

Dixit, A., Parnas, O., Li, B., Chen, J., Fulco, C. P., Jerby-Arnon, L., et al. (2016). Perturb-Seq: dissecting molecular circuits with scalable single-cell RNA profiling of pooled genetic screens. *Cell* 167, 1853–1866.e17. doi: 10.1016/j.cell.2016.11.038

Dubois, P. C. A., Trynka, G., Franke, L., Hunt, K. A., Romanos, J., Curtotti, A., et al. (2010). Multiple common variants for celiac disease influencing immune gene expression. *Nat. Genet.* 42, 295–302.

Ernst, J., and Kellis, M. (2012). ChromHMM: automating chromatin-state discovery and characterization. *Nat. Methods* 9, 215–216. doi: 10.1038/nmeth.1906

Evans, D. M., and Davey Smith, G. (2015). Mendelian randomization: new applications in the coming age of hypothesis-free causality. *Annu. Rev. Genomics Hum. Genet.* 16, 327–350. doi: 10.1146/annurev-genom-090314-050016

Fairfax, B. P., Humburg, P., Makino, S., Naranbhai, V., Wong, D., Lau, E., et al. (2014). Innate immune activity conditions the effect of regulatory variants upon monocyte gene expression. *Science* 343:1246949. doi: 10.1126/science.1246949

Farh, K. K.-H., Marson, A., Zhu, J., Kleinewietfeld, M., Housley, W. J., Beik, S., et al. (2015). Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature* 518, 337–343. doi: 10.1038/nature13835

Finucane, H. K., Bulik-Sullivan, B., Gusev, A., Trynka, G., Reshef, Y., Loh, P.-R., et al. (2015). Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat. Genet.* 47, 1228–1235.

Finucane, H. K., Reshef, Y. A., Anttila, V., Slowikowski, K., Gusev, A., Byrnes, A., et al. (2018). Heritability enrichment of specifically expressed genes identifies disease-relevant tissues and cell types. *Nat. Genet.* 50, 621–629. doi: 10.1038/s41588-018-0081-4

Fortune, M. D., Guo, H., Burren, O., Schofield, E., Walker, N. M., Ban, M., et al. (2015). Statistical colocalization of genetic risk variants for related autoimmune diseases in the context of common controls. *Nat. Genet.* 47, 839–846. doi: 10.1038/ng.3330

Franceschini, N., Giambartolomei, C., de Vries, P. S., Finan, C., Bis, J. C., Huntley, R. P., et al. (2018). GWAS and colocalization analyses implicate carotid intima-media thickness and carotid plaque loci in cardiovascular outcomes. *Nat. Commun.* 9:5141.

Franke, A., McGovern, D. P. B., Barrett, J. C., Wang, K., Radford-Smith, G. L., Ahmad, T., et al. (2010). Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. *Nat. Genet.* 42, 1118–1125.

Frommer, M., McDonald, L. E., Millar, D. S., Collis, C. M., Watt, F., Grigg, G. W., et al. (1992). A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. *Proc. Natl. Acad. Sci. U.S.A.* 89, 1827–1831. doi: 10.1073/pnas.89.5.1827

Gamazon, E. R., Wheeler, H. E., Shah, K. P., Mozaffari, S. V., Aquino-Michaels, K., Carroll, R. J., et al. (2015). A gene-based association method for mapping traits using reference transcriptome data. *Nat. Genet.* 47, 1091–1098. doi: 10.1038/ng.3367

Garcia-Closas, M., Rothman, N., Figueroa, J. D., Prokunina-Olsson, L., Han, S. S., Baris, D., et al. (2013). Common genetic polymorphisms modify the effect of smoking on absolute risk of bladder cancer. *Cancer Res.* 73, 2211–2220.

Gaziano, J. M., Concato, J., Brophy, M., Fiore, L., Pyarajan, S., Breeling, J., et al. (2016). Million Veteran Program: a mega-biobank to study genetic influences on health and disease. *J. Clin. Epidemiol.* 70, 214–223. doi: 10.1016/j.jclinepi.2015.09.016

Giambartolomei, C., Vukcevic, D., Schadt, E. E., Franke, L., Hingorani, A. D., Wallace, C., et al. (2014). Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genet.* 10:e1004383. doi: 10.1371/journal.pgen.1004383

Giambartolomei, C., Zhenli Liu, J., Zhang, W., Hauberg, M., Shi, H., Boocock, J., et al. (2018). A Bayesian framework for multiple trait colocalization from summary association statistics. *Bioinformatics* 34, 2538–2545. doi: 10.1093/bioinformatics/bty147

Global Lipids Genetics Consortium, Willer, C. J., Schmidt, E. M., Sengupta, S., Peloso, G. M., Gustafsson, S., et al. (2013). Discovery and refinement of loci associated with lipid levels. *Nat. Genet.* 45.

Gosselin, D., Skola, D., Coufal, N. G., Holtman, I. R., Schlachetzki, J. C. M., Sajti, E., et al. (2017). An environment-dependent transcriptional network specifies human microglia identity. *Science* 356:eaal3222. doi: 10.1126/science.aal3222

Grosselin, K., Durand, A., Marsolier, J., Poitou, A., Marangoni, E., Nemati, F., et al. (2019). High-throughput single-cell ChIP-seq identifies heterogeneity of chromatin states in breast cancer. *Nat. Genet.* 51, 1060–1066. doi: 10.1038/s41588-019-0424-9

GTEx Consortium (2013). The genotype-tissue expression (GTEx) project. *Nat. Genet.* 45, 580–585.

Gusev, A., Ko, A., Shi, H., Bhatia, G., Chung, W., Penninx, B. W. J. H., et al. (2016). Integrative approaches for large-scale transcriptome-wide association studies. *Nat. Genet.* 48, 245–252.

Gusev, A., Mancuso, N., Won, H., Kousi, M., Finucane, H. K., Reshef, Y., et al. (2018). Transcriptome-wide association study of schizophrenia and chromatin activity yields mechanistic disease insights. *Nat. Genet.* 50, 538–548. doi: 10.1038/s41588-018-0092-1

Hakonarson, H., Qu, H.-Q., Bradfield, J. P., Marchand, L., Kim, C. E., Glessner, J. T., et al. (2008). A novel susceptibility locus for type 1 diabetes on Chr12q13 identified by a genome-wide association study. *Diabetes Metab. Res. Rev.* 57, 1143–1146. doi: 10.2337/db07-1305

Hannon, E., Spiers, H., Viana, J., Pidsley, R., Burrage, J., Murphy, T. M., et al. (2016). Methylation QTLs in the developing brain and their enrichment in schizophrenia risk loci. *Nat. Neurosci.* 19, 48–54. doi: 10.1038/nn.4182

Hari Dass, S. A., McCracken, K., Pokhvisneva, I., Chen, L. M., Garg, E., Nguyen, T. T. T., et al. (2019). A biologically-informed polygenic score identifies endophenotypes and clinical conditions associated with the insulin receptor function on specific brain regions. *EBioMedicine* 42, 188–202. doi: 10.1016/j.ebiom.2019.03.051

He, X., Fuller, C. K., Song, Y., Meng, Q., Zhang, B., Yang, X., et al. (2013). Sherlock: detecting gene-disease associations by matching patterns of expression QTL and GWAS. *Am. J. Hum. Genet.* 92, 667–680. doi: 10.1016/j.ajhg.2013.03.022

Henriksson, J., Chen, X., Gomes, T., Ullah, U., Meyer, K. B., Miragaia, R., et al. (2019). Genome-wide CRISPR Screens in T helper cells reveal pervasive crosstalk between activation and differentiation. *Cell* 176, 882–896.e18. doi: 10.1016/j.cell.2018.11.044

Hindorff, L. A., Gillanders, E. M., and Manolio, T. A. (2011). Genetic architecture of cancer and other complex diseases: lessons learned and future directions. *Carcinogenesis* 32, 945–954. doi: 10.1093/carcin/bgr056

Hormozdiari, F., Kostem, E., Kang, E. Y., Pasaniuc, B., and Eskin, E. (2014). Identifying causal variants at loci with multiple signals of association. *Genetics* 198, 497–508. doi: 10.1534/genetics.114.167908

Hormozdiari, F., van de Bunt, M., Segrè, A. V., Li, X., Joo, J. W. J., Bilow, M., et al. (2016). Colocalization of GWAS and eQTL Signals Detects Target Genes. *Am. J. Hum. Genet.* 99, 1245–1260. doi: 10.1016/j.ajhg.2016.10.003

Hormozdiari, F., Zhu, A., Kichaev, G.-T., Ju, C. J., Segre, A. V., Joo, J. W. J., et al. (2017). Widespread allelic heterogeneity in complex traits | Elsevier enhanced reader. *Am. J. Hum. Genet.* 100, 789–802. doi: 10.1016/j.ajhg.2017.04.005

Hu, X., Kim, H., Stahl, E., Plenge, R., Daly, M., and Raychaudhuri, S. (2011). Integrating autoimmune risk loci with gene-expression data identifies specific pathogenic immune cell subsets. *Am. J. Hum. Genet.* 89, 496–506. doi: 10.1016/j.ajhg.2011.09.002

Hu, Y., Lu, Q., Powles, R., Yao, X., Yang, C., Fang, F., et al. (2017). Leveraging functional annotations in genetic risk prediction for human complex diseases. *PLoS Comput. Biol.* 13:e1005589. doi: 10.1371/journal.pcbi.1005589

Huang, H., Fang, M., Jostins, L., Umićević Mirkov, M., Boucher, G., Anderson, C. A., et al. (2017). Fine-mapping inflammatory bowel disease loci to single-variant resolution. *Nature* 547, 173–178.

Ihry, R. J., Worringer, K. A., Salick, M. R., Frias, E., Ho, D., Theriault, K., et al. (2018). p53 inhibits CRISPR-Cas9 engineering in human pluripotent stem cells. *Nat. Med.* 24, 939–946.

Insull, W. (2009). The pathology of atherosclerosis: plaque development and plaque responses to medical treatment. *Am. J. Med.* 122, S3–S14.

International HapMap Consortium (2003). The international HapMap project. *Nature* 426, 789–796. doi: 10.1038/nature02168

International Genetics of Ankylosing Spondylitis Consortium, Cortes, A., Hadler, J., Pointon, J. P., Robinson, P. C., Karaderi, T., et al. (2013). Identification of multiple risk variants for ankylosing spondylitis through high-density genotyping of immune-related loci. *Nat. Genet.* 45, 730–738. doi: 10.1038/ng.2667

Iotchkova, V., Ritchie, G. R. S., Geihs, M., Morganella, S., Min, J. L., Walter, K., et al. (2019). GARFIELD classifies disease-relevant genomic features through integration of functional annotations with association signals. *Nat. Genet.* 51, 343–353. doi: 10.1038/s41588-018-0322-6

Jostins, L., Ripke, S., Weersma, R. K., Duerr, R. H., McGovern, D. P., Hui, K. Y., et al. (2012). Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature* 491, 119–124.

Khera, A. V., Chaffin, M., Aragam, K. G., Haas, M. E., Roselli, C., Choi, S. H., et al. (2018). Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nat. Genet.* 50, 1219–1224. doi: 10.1038/s41588-018-0183-z

Khera, A. V., Chaffin, M., Wade, K. H., Zahid, S., Brancale, J., Xia, R., et al. (2019). Polygenic prediction of weight and obesity trajectories from birth to adulthood. *Cell* 177, 587–596.e9. doi: 10.1016/j.cell.2019.03.028

Khera, A. V., Emdin, C. A., Drake, I., Natarajan, P., Bick, A. G., Cook, N. R., et al. (2016). Genetic risk, adherence to a healthy lifestyle, and coronary disease. *N. Engl. J. Med.* 375, 2349–2358.

Kimmerling, R. J., Lee Szeto, G., Li, J. W., Genshaft, A. S., Kazer, S. W., Payer, K. R., et al. (2016). A microfluidic platform enabling single-cell RNA-seq of multigenerational lineages. *Nat. Commun.* 7:10220.

Kircher, M., Xiong, C., Martin, B., Schubach, M., Inoue, F., Bell, R. J. A., et al. (2019). Saturation mutagenesis of twenty disease-associated regulatory elements at single base-pair resolution. *Nat. Commun.* 10:3583.

Koike-Yusa, H., Li, Y., Tan, E.-P., Velasco-Herrera, M. D. C., and Yusa, K. (2014). Genome-wide recessive genetic screening in mammalian cells with a lentiviral CRISPR-guide RNA library. *Nat. Biotechnol.* 32, 267–273. doi: 10.1038/nbt.2800

Kreins, A. Y., Ciancanelli, M. J., Okada, S., Kong, X.-F., Ramírez-Alejo, N., Kilic, S. S., et al. (2015). Human TYK2 deficiency: mycobacterial and viral infections without hyper-IgE syndrome. *J. Exp. Med.* 212, 1641–1662.

Kumasaka, N., Knights, A. J., and Gaffney, D. J. (2016). Fine-mapping cellular QTLs with RASQUAL and ATAC-seq. *Nat. Genet.* 48, 206–213. doi: 10.1038/ng.3467

Kunkel, S. D., Elmore, C. J., Bongers, K. S., Ebert, S. M., Fox, D. K., Dyle, M. C., et al. (2012). Ursolic acid increases skeletal muscle and brown fat and decreases diet-induced obesity, glucose intolerance and fatty liver disease. *PLoS ONE* 7:e39332. doi: 10.1371/journal.pone.0039332

Lähnemann, D., Köster, J., Szczurek, E., McCarthy, D. J., Hicks, S. C., Robinson, M. D., et al. (2020). Eleven grand challenges in single-cell data science. *Genome Biol.* 21:31.

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.

Li, Y., Oosting, M., Smeekens, S. P., Jaeger, M., Aguirre-Gamboa, R., Le, K. T. T., et al. (2016). A functional genomics approach to understand variation in cytokine production in humans. *Cell* 167, 1099–1110.e14. doi: 10.1016/j.cell.2016.10.017

Li, Y. I., Knowles, D. A., Humphrey, J., Barbeira, A. N., Dickinson, S. P., Im, H. K., et al. (2018). Annotation-free quantification of RNA splicing using LeafCutter. *Nat. Genet.* 50, 151–158. doi: 10.1038/s41588-017-0004-9

Liu, B., Gloudemans, M. J., Rao, A. S., Ingelsson, E., and Montgomery, S. B. (2019). Abundant associations with gene expression complicate GWAS follow-up. *Nat. Genet.* 51, 768–769. doi: 10.1038/s41588-019-0404-0

Liu, X., Li, Y. I., and Pritchard, J. K. (2019). Trans effects on gene expression can drive omnigenic inheritance. *Cell* 177, 1022–1034.e6. doi: 10.1016/j.cell.2019.04.014

Liu, B., Pjanic, M., Wang, T., Nguyen, T., Gloudemans, M., Rao, A., et al. (2018). Genetic regulatory mechanisms of smooth muscle cells map to coronary artery disease risk loci. *Am. J. Hum. Genet.* 103, 377–388. doi: 10.1016/j.ajhg.2018.08.001

Lloyd-Jones, L. R., Holloway, A., McRae, A., Yang, J., Small, K., Zhao, J., et al. (2017). The genetic architecture of gene expression in peripheral blood. *Am. J. Hum. Genet.* 100, 228–237.

Maas, P., Barrdahl, M., Joshi, A. D., Auer, P. L., Gaudet, M. M., Milne, R. L., et al. (2016). Breast cancer risk from modifiable and nonmodifiable risk factors among white women in the united states. *JAMA Oncol.* 2, 1295–1302.

MacArthur, J., Bowler, E., Cerezo, M., Gil, L., Hall, P., Hastings, E., et al. (2017). The new NHGRI-EBI catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Res.* 45, D896–D901.

Macosko, E. Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., et al. (2015). Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* 161, 1202–1214. doi: 10.1016/j.cell.2015.05.002

Manrai, A. K., Funke, B. H., Rehm, H. L., Olesen, M. S., Maron, B. A., Szolovits, P., et al. (2016). Genetic misdiagnoses and the potential for health disparities. *N. Engl. J. Med.* 375, 655–665. doi: 10.1056/nejmsa1507092

Marigorta, U. M., and Navarro, A. (2013). High trans-ethnic replicability of GWAS results implies common causal variants. *PLoS Genet.* 9:e1003566. doi: 10.1371/journal.pgen.1003566

Márquez-Luna, C., Gazal, S., Loh, P.-R., Furlotte, N., Auton, A., Me Research Team, et al. (2018). Modeling functional enrichment improves polygenic prediction accuracy in UK Biobank and 23andMe data sets. *bioRxiv [Preprint]*

Martin, A. R., Gignoux, C. R., Walters, R. K., Wojcik, G. L., Neale, B. M., Gravel, S., et al. (2017). Human demographic history impacts genetic risk prediction across diverse populations. *Am. J. Hum. Genet.* 100, 635–649. doi: 10.1016/j.ajhg.2017.03.004

Martin, A. R., Kanai, M., Kamatani, Y., Okada, Y., Neale, B. M., and Daly, M. J. (2019). Clinical use of current polygenic risk scores may exacerbate health disparities. *Nat. Genet.* 51, 584–591. doi: 10.1038/s41588-019-0379-x

Maurano, M. T., Humbert, R., Rynes, E., Thurman, R. E., Haugen, E., Wang, H., et al. (2012). Systematic localization of common disease-associated variation in regulatory DNA. *Science* 337, 1190–1195. doi: 10.1126/science.1222794

Mavaddat, N., Pharoah, P. D. P., Michailidou, K., Tyrer, J., Brook, M. N., Bolla, M. K., et al. (2015). Prediction of breast cancer risk based on profiling with common genetic variants. *J. Natl. Cancer Inst.* 107:djv036. doi: 10.1093/jnci/djv036

McCullum, E. O., Williams, B. A. R., Zhang, J., and Chaput, J. C. (2010). "Random mutagenesis by Error-Prone PCR," in *In Vitro Mutagenesis Protocols*, 3rd Edn, ed. J. Braman (Totowa, NJ: Humana Press), 103–109. doi: 10.1007/978-1-60761-652-8_7

Mega, J. L., Stitziel, N. O., Smith, J. G., Chasman, D. I., Caulfield, M., Devlin, J. J., et al. (2015). Genetic risk, coronary heart disease events, and the clinical benefit of statin therapy: an analysis of primary and secondary prevention trials. *Lancet* 385, 2264–2271. doi: 10.1016/s0140-6736(14)61730-x

Melé, M., Ferreira, P. G., Reverter, F., DeLuca, D. S., Monlong, J., Sammeth, M., et al. (2015). Human genomics. The human transcriptome across tissues and individuals. *Science* 348, 660–665.

Melzer, D., Perry, J. R. B., Hernandez, D., Corsi, A.-M., Stevens, K., Rafferty, I., et al. (2008). A genome-wide association study identifies protein quantitative trait loci (pQTLs). *PLoS Genet.* 4:e1000072. doi: 10.1371/journal.pgen.1000072

Monlong, J., Calvo, M., Ferreira, P. G., and Guigó, R. (2014). Identification of genetic variants associated with alternative splicing using sQTLseekeR. *Nat. Commun.* 5:4698.

Musunuru, K., Strong, A., Frank-Kamenetsky, M., Lee, N. E., Ahfeldt, T., Sachs, K. V., et al. (2010). From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol locus. *Nature* 466, 714–719. doi: 10.1038/nature09266

Myocardial Infarction Genetics Consortium, Kathiresan, S., Voight, B. F., Purcell, S., Musunuru, K., Ardissino, D., et al. (2009). Genome-wide association of early-onset myocardial infarction with single nucleotide polymorphisms and copy number variants. *Nat. Genet.* 41, 334–341. doi: 10.1038/ng.327

Nagai, A., Hirata, M., Kamatani, Y., Muto, K., Matsuda, K., Kiyohara, Y., et al. (2017). Overview of the BioBank Japan Project: study design and profile. *J. Epidemiol.* 27, S2–S8.

Nica, A. C., and Dermitzakis, E. T. (2013). Expression quantitative trait loci: present and future. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 368, 20120362. doi: 10.1098/rstb.2012.0362

Nica, A. C., Montgomery, S. B., Dimas, A. S., Stranger, B. E., Beazley, C., Barroso, I., et al. (2010). Candidate causal regulatory effects by integration of expression QTLs with complex trait genetic associations. *PLoS Genet.* 6:e1000895. doi: 10.1371/journal.pgen.1000895

Nicolae, D. L., Gamazon, E., Zhang, W., Duan, S., Dolan, M. E., and Cox, N. J. (2010). Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. *PLoS Genet.* 6:e1000888. doi: 10.1371/journal.pgen.1000888

Okada, Y., Wu, D., Trynka, G., Raj, T., Terao, C., Ikari, K., et al. (2014). Genetics of rheumatoid arthritis contributes to biology and drug discovery. *Nature* 506, 376–381.

Onengut-Gumuscu, S., Chen, W.-M., Burren, O., Cooper, N. J., Quinlan, A. R., Mychaleckyj, J. C., et al. (2015). Fine mapping of type 1 diabetes susceptibility loci and evidence for colocalization of causal variants with lymphoid gene enhancers. *Nat. Genet.* 47, 381–386. doi: 10.1038/ng.3245

Ongen, H., and Dermitzakis, E. T. (2015). Alternative splicing QTLs in european and african populations. *Am. J. Hum. Genet.* 97, 567–575. doi: 10.1016/j.ajhg.2015.09.004

Pasaniuc, B., Zaitlen, N., Shi, H., Bhatia, G., Gusev, A., Pickrell, J., et al. (2014). Fast and accurate imputation of summary statistics enhances evidence of functional enrichment. *Bioinformatics* 30:2906. doi: 10.1093/bioinformatics/btu416

Pasquali, L., Gaulton, K. J., Rodríguez-Seguí, S. A., Mularoni, L., Miguel-Escalada, I., Akerman, ı, et al. (2014). Pancreatic islet enhancer clusters enriched in type 2 diabetes risk-associated variants. *Nat. Genet.* 46, 136–143. doi: 10.1038/ng.2870

Pelikan, R. C., Kelly, J. A., Fu, Y., Lareau, C. A., Tessneer, K. L., Wiley, G. B., et al. (2018). Enhancer histone-QTLs are enriched on autoimmune risk haplotypes and influence gene expression within chromatin networks. *Nat. Commun.* 9:2905.

Picelli, S., Björklund, ÅK., Faridani, O. R., Sagasser, S., Winberg, G., and Sandberg, R. (2013). Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nat. Methods* 10, 1096–1098. doi: 10.1038/nmeth.2639

Pickrell, J. K. (2014). Joint analysis of functional genomic data and genome-wide association studies of 18 human traits. *Am. J. Hum. Genet.* 94, 559–573. doi: 10.1016/j.ajhg.2014.03.004

Pickrell, J. K., Berisa, T., Liu, J. Z., Ségurel, L., Tung, J. Y., and Hinds, D. A. (2016). Detection and interpretation of shared genetic influences on 42 human traits. *Nat. Genet.* 48, 709–717. doi: 10.1038/ng.3570

Plagnol, V., Smyth, D. J., Todd, J. A., and Clayton, D. G. (2009). Statistical independence of the colocalized association signals for type 1 diabetes and RPS26 gene expression on chromosome 12q13. *Biostatistics* 10, 327–334. doi: 10.1093/biostatistics/kxn039

ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature* 489, 57–74. doi: 10.1038/nature11247

Qi, L. S., Larson, M. H., Gilbert, L. A., Doudna, J. A., Weissman, J. S., Arkin, A. P., et al. (2013). Repurposing CRISPR as an RNA-guided platform for sequence-specific control of gene expression. *Cell* 152, 1173–1183. doi: 10.1016/j.cell.2013.02.022

Regev, A., Teichmann, S. A., Lander, E. S., Amit, I., Benoist, C., Birney, E., et al. (2017). The human cell atlas. *Elife* 6:e27041. doi: 10.7554/eLife.27041

Richardson, T. G., Harrison, S., Hemani, G., and Davey Smith, G. (2019). An atlas of polygenic risk score associations to highlight putative causal relationships across the human phenome. *Elife* 8:e43657. doi: 10.7554/eLife.43657

Ripatti, S., Tikkanen, E., Orho-Melander, M., Havulinna, A. S., Silander, K., Sharma, A., et al. (2010). A multilocus genetic risk score for coronary heart disease: case-control and prospective cohort analyses. *Lancet* 376, 1393–1400. doi: 10.1016/s0140-6736(10)61267-6

Roadmap Epigenomics Consortium, Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., et al. (2015). Integrative analysis of 111 reference human epigenomes. *Nature* 518, 317–330.

Rotem, A., Ram, O., Shoresh, N., Sperling, R. A., Goren, A., Weitz, D. A., et al. (2015). Single-cell ChIP-seq reveals cell subpopulations defined by chromatin state. *Nat. Biotechnol.* 33, 1165–1172. doi: 10.1038/nbt.3383

Saelens, W., Cannoodt, R., Todorov, H., and Saeys, Y. (2019). A comparison of single-cell trajectory inference methods. *Nat. Biotechnol.* 37, 547–554. doi: 10.1038/s41587-019-0071-9

Schadt, E. E., Molony, C., Chudin, E., Hao, K., Yang, X., Lum, P. Y., et al. (2008). Mapping the genetic architecture of gene expression in human liver. *PLoS Biol.* 6:e107. doi: 10.1371/journal.pbio.0060107

Schaid, D. J., Chen, W., and Larson, N. B. (2018). From genome-wide associations to candidate causal variants by statistical fine-mapping. *Nat. Rev. Genet.* 19, 491–504. doi: 10.1038/s41576-018-0016-z

Schmidt, E. M., Zhang, J., Zhou, W., Chen, J., Mohlke, K. L., Chen, Y. E., et al. (2015). GREGOR: evaluating global enrichment of trait-associated variants in epigenomic features using a systematic, data-driven approach. *Bioinformatics* 31, 2601–2606. doi: 10.1093/bioinformatics/btv201

Shifrut, E., Carnevale, J., Tobin, V., Roth, T. L., Woo, J. M., Bui, C. T., et al. (2018). Genome-wide CRISPR screens in primary human T cells reveal key regulators of immune function. *Cell* 175, 1958–1971.e15. doi: 10.1016/j.cell.2018.10.024

Slatkin, M. (2008). Linkage disequilibrium–understanding the evolutionary past and mapping the medical future. *Nat. Rev. Genet.* 9, 477–485. doi: 10.1038/nrg2361

Slowikowski, K., Hu, X., and Raychaudhuri, S. (2014). SNPsea: an algorithm to identify cell types, tissues and pathways affected by risk loci. *Bioinformatics* 30, 2496–2497. doi: 10.1093/bioinformatics/btu326

Smith, G. D., Ebrahim, S., Lewis, S., Hansell, A. L., Palmer, L. J., and Burton, P. R. (2005). Genetic epidemiology and public health: hope, hype, and future prospects. *Lancet* 366, 1484–1498. doi: 10.1016/s0140-6736(05)67601-5

Soskic, B., Cano-Gamez, E., Smyth, D. J., Rowan, W. C., Nakic, N., Esparza-Gordillo, J., et al. (2019). Chromatin activity at GWAS loci identifies T cell states driving complex immune diseases. *Nat. Genet.* 51, 1486–1493. doi: 10.1038/s41588-019-0493-9

Strober, B. J., Elorbany, R., Rhodes, K., Krishnan, N., Tayeb, K., Battle, A., et al. (2019). Dynamic genetic regulation of gene expression during cellular differentiation. *Science* 364, 1287–1290. doi: 10.1126/science.aaw0040

Sun, W., Poschmann, J., Cruz-Herrera Del, Rosario, R., Parikshak, N. N., Hajan, H. S., et al. (2016). Histone acetylome-wide association study of autism spectrum disorder. *Cell* 167, 1385–1397.e11. doi: 10.1016/j.cell.2016.10.031

Tehranchi, A., Hie, B., Dacre, M., Kaplow, I., Pettie, K., Combs, P., et al. (2019). Fine-mapping cis-regulatory variants in diverse human populations. *Elife* 8:e39595. doi: 10.7554/eLife.39595

Teslovich, T. M., Musunuru, K., Smith, A. V., Edmondson, A. C., Stylianou, I. M., Koseki, M., et al. (2010). Biological, clinical and population relevance of 95 loci for blood lipids. *Nature* 466, 707–713.

The PsychENCODE Consortium, Akbarian, S., Liu, C., Knowles, J. A., Vaccarino, F. M., Farnham, P. J., et al. (2015). The PsychENCODE project. *Nat. Neurosci.* 18, 1707–1712

Torkamani, A., Wineinger, N. E., and Topol, E. J. (2018). The personal and clinical utility of polygenic risk scores. *Nat. Rev. Genet.* 19, 581–590. doi: 10.1038/s41576-018-0018-x

Trapnell, C. (2015). Defining cell types and states with single-cell genomics. *Genome Res.* 25, 1491–1498. doi: 10.1101/gr.190595.115

Trynka, G., and Raychaudhuri, S. (2013). Using chromatin marks to interpret and localize genetic associations to complex human traits and diseases. *Curr. Opin. Genet. Dev.* 23, 635–641. doi: 10.1016/j.gde.2013.10.009

Trynka, G., Sandor, C., Han, B., Xu, H., Stranger, B. E., Liu, X. S., et al. (2013). Chromatin marks identify critical cell types for fine mapping complex trait variants. *Nat. Genet.* 45, 124–130. doi: 10.1038/ng.2504

Trynka, G., Westra, H.-J., Slowikowski, K., Hu, X., Xu, H., Stranger, B. E., et al. (2015). Disentangling the effects of colocalizing genomic annotations to functionally prioritize non-coding variants within complex-trait loci. *Am. J. Hum. Genet.* 97, 139–152. doi: 10.1016/j.ajhg.2015.05.016

Udalova, I. A., Mantovani, A., and Feldmann, M. (2016). Macrophage heterogeneity in the context of rheumatoid arthritis. *Nat. Rev. Rheumatol.* 12:472. doi: 10.1038/nrrheum.2016.91

Ulirsch, J. C., Lareau, C. A., Bao, E. L., Ludwig, L. S., Guo, M. H., Benner, C., et al. (2019). Interrogation of human hematopoiesis at single-cell and single-variant resolution. *Nat. Genet.* 51, 683–693. doi: 10.1038/s41588-019-0362-6

Ulirsch, J. C., Nandakumar, S. K., Wang, L., Giani, F. C., Zhang, X., Rogov, P., et al. (2016). Systematic functional dissection of common genetic variation affecting red blood cell traits. *Cell* 165, 1530–1545. doi: 10.1016/j.cell.2016.04.048

van der Wijst, M. G. P., Brugge, H., de Vries, D. H., Deelen, P., Swertz, M. A., LifeLines Cohort, et al. (2018). Single-cell RNA sequencing identifies celltype-specific cis-eQTLs and co-expression QTLs. *Nat. Genet.* 50, 493–497. doi: 10.1038/s41588-018-0089-9

Vilhjálmsson, B. J., Yang, J., Finucane, H. K., Gusev, A., Lindström, S., Ripke, S., et al. (2015). Modeling linkage disequilibrium increases accuracy of polygenic risk scores. *Am. J. Hum. Genet.* 97, 576–592.

Visscher, P. M., and Goddard, M. E. (2019). From R.A. fisher's 1918 Paper to GWAS a century later. *Genetics* 211, 1125–1130. doi: 10.1534/genetics.118.301594

Visscher, P. M., Wray, N. R., Zhang, Q., Sklar, P., McCarthy, M. I., Brown, M. A., et al. (2017). 10 Years of GWAS Discovery: Biology, Function, and Translation. *Am. J. Hum. Genet.* 101, 5–22. doi: 10.1016/j.ajhg.2017.06.005

Viswanath, B., Jose, S. P., Squassina, A., Thirthalli, J., Purushottam, M., Mukherjee, O., et al. (2015). Cellular models to study bipolar disorder: a systematic review. *J. Affect. Disord.* 184, 36–50.

Võsa, U., Claringbould, A., Westra, H.-J., Bonder, M. J., Deelen, P., Zeng, B., et al. (2018). Unraveling the polygenic architecture of complex traits using blood eQTL metaanalysis. *bioRxiv [Preprint]*

Wainberg, M., Sinnott-Armstrong, N., Mancuso, N., Barbeira, A. N., Knowles, D. A., Golan, D., et al. (2019). Opportunities and challenges for transcriptome-wide association studies. *Nat. Genet.* 51, 592–599. doi: 10.1038/s41588-019-0385-z

Wallace, C. (2013). Statistical testing of shared genetic control for potentially related traits. *Genet. Epidemiol.* 37, 802–813. doi: 10.1002/gepi.21765

Wallace, C., Rotival, M., Cooper, J. D., Rice, C. M., Yang, J. H. M., McNeill, M., et al. (2012). Statistical colocalization of monocyte gene expression and genetic risk variants for type 1 diabetes. *Hum. Mol. Genet.* 21, 2815–2824. doi: 10.1093/hmg/dds098

Wang, H., La Russa, M., and Qi, L. S. (2016). CRISPR/Cas9 in genome editing and beyond. *Annu. Rev. Biochem.* 85, 227–264.

Wellcome Trust Case Control Consortium (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* 447, 661–678. doi: 10.1038/nature05911

Wen, X., Pique-Regi, R., and Luca, F. (2017). Integrating molecular QTL data into genome-wide genetic association analysis: probabilistic assessment of enrichment and colocalization. *PLoS Genet.* 13:e1006646. doi: 10.1371/journal.pgen.1006646

Westra, H.-J., Peters, M. J., Esko, T., Yaghootkar, H., Schurmann, C., Kettunen, J., et al. (2013). Systematic identification of trans eQTLs as putative drivers of known disease associations. *Nat. Genet.* 45, 1238–1243.

Wright, F. A., Sullivan, P. F., Brooks, A. I., Zou, F., Sun, W., Xia, K., et al. (2014). Heritability and genomics of gene expression in peripheral blood. *Nat. Genet.* 46, 430–437.

Xu, Y., Vuckovic, D., Ritchie, S. C., Akbari, P., Jiang, T., Grealey, J., et al. (2020). Learning polygenic scores for human blood cell traits. *bioRxiv [Preprint]*

Yang, J., Benyamin, B., McEvoy, B. P., Gordon, S., Henders, A. K., Nyholt, D. R., et al. (2010). Common SNPs explain a large proportion of the heritability for human height. *Nat. Genet.* 42, 565–569. doi: 10.1038/ng.608

Yang, J., Zeng, J., Goddard, M. E., Wray, N. R., and Visscher, P. M. (2017). Concepts, estimation and interpretation of SNP-based heritability. *Nat. Genet.* 49, 1304–1310. doi: 10.1038/ng.3941

Yao, C., Chen, G., Song, C., Keefe, J., Mendelson, M., Huan, T., et al. (2018). Genome-wide mapping of plasma protein QTLs identifies putatively causal genes and pathways for cardiovascular disease. *Nat. Commun.* 9:3268.

Zhang, W., Voloudakis, G., Rajagopal, V. M., Readhead, B., Dudley, J. T., Schadt, E. E., et al. (2019). Integrative transcriptome imputation reveals tissue-specific and shared biological mechanisms mediating susceptibility to complex traits. *Nat. Commun.* 10:3834.

Zheng, G. X. Y., Terry, J. M., Belgrader, P., Ryvkin, P., Bent, Z. W., Wilson, R., et al. (2017). Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.* 8:14049.

Zhu, Z., Zhang, F., Hu, H., Bakshi, A., Robinson, M. R., Powell, J. E., et al. (2016). Integration of summary data from GWAS. 48, 481–487. doi: 10.1038/ng.3538

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.

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