RECENT ADVANCES IN THE PREDICTION OF FUNCTIONAL EFFECTS OF NON-CODING VARIANTS AND THEIR CLINICAL APPLICATIONS

EDITED BY: Le Li, Dapeng Xiong, Lixin Cheng and Kevin Yip

PUBLISHED IN: Frontiers in Genetics

frontiers Research Topics

frontiers

ISSN 1664-8714
ISBN 978-2-83250-274-7
DOI 10.3389/978-2-83250-274-7

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

RECENT ADVANCES IN THE PREDICTION OF FUNCTIONAL EFFECTS OF NON-CODING VARIANTS AND THEIR CLINICAL APPLICATIONS

Topic Editors:
Le Li, Cornell University, United States
Dapeng Xiong, Cornell University, United States
Lixin Cheng, Jinan University, China
Kevin Yip, The Chinese University of Hong Kong, China

Citation: Li, L., Xiong, D., Cheng, L., Yip, K., eds. (2022). Recent Advances in the Prediction of Functional Effects of Non-Coding Variants and their Clinical Applications. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-83250-274-7

Table of Contents

frontiers in Genetics

ORIGINAL RESEARCH
published: 11 June 2019
doi: 10.3389/fgene.2019.00550

The Effect of Genetic Variation on the Placental Transcriptome in Humans

*Triin Kikas[1], Kristiina Rull[1,2,3], Robin N. Beaumont[4], Rachel M. Freathy[4] and Maris Laan[1]**

[1] Human Genetics Research Group, Institute of Biomedicine and Translational Medicine, University of Tartu, Tartu, Estonia, [2] Women's Clinic, Tartu University Hospital, Tartu, Estonia, [3] Department of Obstetrics and Gynecology, University of Tartu, Tartu, Estonia, [4] Institute of Biomedical and Clinical Science, University of Exeter Medical School, University of Exeter, Exeter, United Kingdom

The knowledge of genetic variants shaping human placental transcriptome is limited and they are not cataloged in the Genotype-Tissue Expression project. So far, only one whole genome analysis of placental expression quantitative trait loci (eQTLs) has been published by Peng et al. (2017) with no external independent validation. We report the second study on the landscape of placental eQTLs. The study aimed to generate a high-confidence list of placental *cis*-eQTLs and to investigate their potential functional implications. Analysis of *cis*-eQTLs (±100 kbp from the gene) utilized 40 placental RNA sequencing and respective whole genome genotyping datasets. The identified 199 placental *cis*-eSNPs represented 88 independent eQTL signals (FDR < 5%). The most significant placental eQTLs (FDR < 10^{-5}) modulated the expression of ribosomal protein RPL9, transcription factor ZSCAN9 and aminopeptidase ERAP2. The analysis confirmed 50 eSNP-eGenes pairs reported by Peng et al. (2017) and thus, can be claimed as robust placental eQTL signals. The study identified also 13 novel placental eGenes. Among these, *ZSCAN9* is modulated by several eSNPs (experimentally validated: rs1150707) that have been also shown to affect the methylation level of genes variably escaping X-chromosomal inactivation. The identified 63 placental eGenes exhibited mostly mixed or ubiquitous expression. Functional enrichment analysis highlighted 35 Gene Ontology categories with the top ranking pathways "ruffle membrane" (FDR = 1.81×10^{-2}) contributing to the formation of motile cell surface and "ATPase activity, coupled" (FDR = 2.88×10^{-2}), critical for the membrane transport. Placental eGenes were also significantly enriched in pathways implicated in development, signaling and immune function. However, this study was not able to confirm a significant overrepresentation of genome-wide association studies top hits among the placental eSNP and eGenes, reported by Peng et al. (2017). The identified eSNPs were further analyzed in association with newborn and pregnancy traits. In the discovery step, a suggestive association was detected between the eQTL of *ALPG* (rs11678251) and reduced placental, newborn's and infant's weight. Meta-analysis across REPROMETA, HAPPY PREGNANCY, ALSPAC cohorts (n = 6830) did not replicate these findings. In summary, the study emphasizes the role of genetic variation in driving the transcriptome profile of the human placenta and the importance to explore further its functional implications.

Keywords: human placenta, *cis*-eQTL, complex traits, REPROMETA, ALSPAC, HAPPY PREGNANCY

OPEN ACCESS

Edited by:
Serena Sanna, University Medical Center Groningen, Netherlands

Reviewed by:
Alexander Teumer, University of Greifswald, Germany
Wouter Van Rheenen, University Medical Center Utrecht, Netherlands

***Correspondence:**
Maris Laan
maris.laan@ut.ee

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

Received: *20 January 2019*
Accepted: *24 May 2019*
Published: *11 June 2019*

Citation:
Kikas T, Rull K, Beaumont RN, Freathy RM and Laan M (2019) The Effect of Genetic Variation on the Placental Transcriptome in Humans. Front. Genet. 10:550. doi: 10.3389/fgene.2019.00550

INTRODUCTION

The placenta is a unique mammal-specific organ. It serves the developing fetus during its short intrauterine period not only as a mediator of maternal resources, but also as a contributor to its developmental programming, growth, and maturation for the postnatal life. Throughout its restricted lifetime, coordinated dynamics of placental gene expression across the entire gestation reflects the growing needs of the fetus in order to guarantee the normal course of pregnancy (Winn et al., 2007; Uusküla et al., 2012). Disturbances in placental gene expression at any gestational time point may lead to or reflect placental malfunction and various pregnancy-related complications of the mother and/or the fetus, such as recurrent pregnancy loss, preeclampsia, intra-uterine growth restriction (Sõber et al., 2015, 2016).

Several modulators of placental gene expression have been investigated in order to uncover their role in placental function and in predisposition to develop pregnancy complications. Amongst the most studied effectors are disturbances in intrauterine conditions [e.g., hypoxia (Soares et al., 2017)], communication at the maternal-fetal interface (Pavličev et al., 2017), placental microRNAs (Deshpande and Balasinor, 2018) and epigenetic alterations modulating the transcriptional activity of critical placental genes (Deshpande and Balasinor, 2018). Additional intrinsic genomic factors determining the expression level of one or many genes are expression quantitative trait loci (eQTLs), alternatively referred as eSNPs. Based on the analysis of 48 different human tissues in the framework of the Genotype-Tissue Expression (GTEx) project – the largest coordinated activity aiming to uncover the relationship between genetic variation and gene expression (Lonsdale et al., 2013; Ardlie et al., 2015), the number of eSNPs across the human genome has been estimated to exceed 150,000 (Aguet et al., 2017). The contribution of eSNPs to human pathological conditions has been shown directly in numerous targeted studies [e.g., breast (Li et al., 2013) and pancreatic cancer (Zhang et al., 2018), nephrotic syndrome (Gillies et al., 2018)] and indirectly via enrichment of eSNPs among top associated loci in GWA studies of various diseases and traits, such as psoriasis (Ding et al., 2010) and height (Yengo et al., 2018). As the GTEx project is based on the analysis of donated post-mortem tissues, it does not contain any placental samples. However, it is well acknowledged that eSNPs vary across different tissues (Grundberg et al., 2012; McKenzie et al., 2014) and even as much as 12–40% of them have been estimated to be tissue-specific (Gillies et al., 2018; Zhang et al., 2018). Therefore, it is critical to uncover placental eSNPs in order to understand their role in modulating placental function and risk to gestational disturbances.

Until recently, studies of human placental eSNPs have been restricted to reports focusing on variants modulating transcript levels of single placental genes and their effect on pregnancy outcomes. Of the few studies on placental eQTLs, one has linked variants in *FKBP5* with infant neurobehavioral phenotypes (Paquette et al., 2014), and another reported an association between *STC1* eSNPs and preeclampsia (Juhanson et al., 2016). The only published large-scale placental eQTL study demonstrated that the majority of placental eSNPs are located in the vicinity of the modulated genes (eGenes) (Peng et al., 2017). Analysis of 159 placental transcriptomes identified 3218 (98.9%) eSNPs with *cis*-, but only 35 (1.1%) variants with significant *trans*-effects on the expression of affected genes. The study also demonstrated that >5% of all unique loci associated with any disease in genome-wide association studies (GWAS) represented placental eSNPs. The potential role of placental transcriptome profile in "(mis)programming" the offspring's risks to human late-onset common diseases was previously discussed in the context of a substantial overlap between GWAS top loci and genes exhibiting high mid-gestational placental expression (Uusküla et al., 2012).

The current study aimed to generate a high-confidence list of genetic variants modulating placental gene expression. The analysis confirmed 50 eSNP-eGenes pairs reported recently by Peng et al. (2017) and identified 13 novel placental eGenes. The variants were further explored for their potential link to complex diseases and newborn growth parameters.

MATERIALS AND METHODS

Ethics Statement

The REPROgrammed fetal and/or maternal METAbolism (REPROMETA) and HAPPY PREGNANCY (full study name "Development of novel non-invasive biomarkers for fertility and healthy pregnancy") studies were approved by the Ethics Review Committee of Human Research of the University of Tartu, Estonia (Permissions No 146/18, 27.02.2006; 150/33, 18.06.2006; 158/80, 26.03.2007; 221/T-6, 17.12.2012; 286/M-18, 15.10.2018). All study participants were recruited, and the study material was collected at the Women's Clinic of Tartu University Hospital, Estonia in 2006–2011 (REPROMETA) and in 2013–2015 (HAPPY PREGNANCY). All participants in the REPROMETA and HAPPY PREGNANCY studies were of white European ancestry and living in Estonia. A written informed consent to participate in the study was obtained from each individual prior to recruitment.

Ethical approval for the Avon Longitudinal Study of Parents and Children (ALSPAC) was obtained from the ALSPAC Ethics and Law Committee and the Local Research Ethics Committees. The study recruited pregnant women in Bristol area, United Kingdom in 1991–1992.[1] Consent for biological samples was collected in accordance with the Human Tissue Act (2004). Informed consent for the use of data collected via questionnaires and clinics was obtained from participants following the recommendations of the ALSPAC Ethics and Law Committee.

All procedures and methods in the three studies have been carried out in compliance with the guidelines of the Declaration of Helsinki.

[1] http://www.alspac.bris.ac.uk

REPROMETA Dataset for the Discovery Placental eQTL Analysis

The REPROMETA study represents family trios (mother, father, placenta) recruited before or shortly after delivery of a singleton newborn at the Women's Clinic of Tartu University Hospital, Estonia. The study was designed to include well-defined, clinically diagnosed diverse scenarios of pregnancy outcomes at term (gestational age 36–42 weeks). The full REPROMETA placental sample set analyzed in the current study (n = 336) is comprised of five clinical subgroups: delivery of a small-for-gestational-*age* [SGA, birth weight <10th centile (Sildver et al., 2015); n = 65] or large-for-gestational-age newborn (LGA, >90th centile; n = 83), cases of maternal gestational diabetes (GD; n = 41) or severe late-onset preeclampsia (PE; n = 43), and healthy term pregnancies (birth weight 10th–90th centile; n = 104). Criteria for the clinical subgrouping are detailed in **Supplementary Methods**. Clinical and epidemiologic data of the mother and the newborn were collected from medical records.

The current eQTL discovery study utilized previously published RNA sequencing (RNA-Seq) (Sõber et al., 2015; Reiman et al., 2017), and corresponding genome-wide genotyping datasets of 40 term placental samples (Kasak et al., 2015), where each REPROMETA subgroup (NORM, PE, GD, SGA, LGA) was represented by eight placental transcriptomes maximally matched for the gestational age, delivery mode and proportions of male/female newborns (**Table 1**). Placental sampling, RNA extraction, library sequencing and basic informatics are detailed in the **Supplementary Methods** and in Sõber et al. (2015) and Reiman et al. (2017).

Genome-wide genotyping of the same placental samples was carried out on Illumina HumanOmniExpress-12-v1 BeadChip (>733,000 SNPs, median spacing 2.1 kb; Kasak et al., 2015). Samples were genotyped with an average overall call rate >99% per individual per genotype.

We excluded SNPs if they deviated from Hardy-Weinberg Equilibrium (HWE; $P < 1 \times 10^{-6}$) or had no minor allele carriers in our dataset. In total of 661,354 genotyped SNPs were taken forward to the next step.

TABLE 1 | Maternal and newborn data of the term placentas utilized *cis*-eQTL discovery analysis (n = 40).

Parameter	Unit	eQTL analysis samples
Maternal age	year	28.5 [18-39]
Gestational age at delivery	day	274.0 [260-284]
Newborn birth weight	gram	3587 [2004-4986]
Newborn length	cm	50.2 [45-55]
Placental weight	gram	584.0 [200-1060]
Newborn sex: female/male	*n*	21/19
Delivery mode: EmCS/ECS/Vag	*n* [%]	9/12/19 [22.5%/30.0%/47.5%]
Labor activity: no/yes/NA	*n* [%]	19/20/1 [47.5%/50.0%/2.5%]
Pregnancy complications	*n*	32 [80%] PE, GDM, SGA, LGA: 8 each

All cases were recruited, clinical data and placental tissue was collected in the framework of the REPROMETA study at the Women's Clinic of Tartu University Hospital, Estonia. Detailed information on this sample set and respective the placental RNA-Seq data is provided in ***Supplementary Methods*** *and original references (Sõber et al., 2015; Reiman et al., 2017). Data is given as mean [minimum–maximum] unless indicated differently. ECS, elective cesarean section; EmCS, emergency cesarean section; GD, gestational diabetes; LGA, large-for-gestational age newborn; NA, not available; PE, preeclampsia; SGA, small-for-gestational age newborn; Vag, vaginal delivery.*

Bioinformatics for eQTL Detection

The initial unfiltered RNA-Seq dataset of 40 placental samples (Sõber et al., 2015; Reiman et al., 2017) included gene expression data for 53,893 genes (Ensembl v67[2]). Gene expression was quantified using htseq-count (as raw read counts) and then normalized for read depth of the sample. Genes with low expression (applied empirical cutoff: median expression <100 normalized read counts) were excluded. After further filtering out mitochondrial genes, 11,733 genes were retained for the eQTL analysis (**Supplementary Figure S1**).

As a very low number of placental *trans*-eQTLs was expected (Peng et al., 2017) and the chance to detect spurious *trans*-eQTL associations is high (Yao et al., 2017), our study design focused on *cis*-eQTL discovery to maximize true-positive, functionally relevant findings. In the current study, we defined *cis*-SNPs to locate within ±100 kbp from the gene start/end coordinates as these regions have high probability of containing significant and functionally relevant eQTL hits (Veyrieras et al., 2008). SNP and gene coordinates were extracted from the Ensembl database using BioMart (Ensembl v54) and *cis*-eQTL testing was implemented in Matrix eQTL package for R (Shabalin, 2012). The final number of SNP included into eQTL testing was 353,599 and the total number of performed tests was 659,826 (**Supplementary Data S1**). Association between gene expression level (quantified as normalized read counts) and *cis*-SNPs was carried out using linear regression adjusted by the pregnancy outcome (normal term or PE, GD, SGA, LGA pregnancy), labor activity and newborn sex. Nominal P-values were corrected for multiple testing using a built-in Benjamini and Hochberg FDR method in Matrix eQTL. A statistically significant eSNP-eGene association was defined as FDR < 0.05. For each identified eQTL, the proportion of gene expression variability explained by the SNP was calculated (R^2).

eSNPs were clustered into high LD-groups ($r^2 > 0.8$) and only the empirically selected lead SNP with the lowest P-value was taken forward to represent the LD-group in subsequent analyses. For determining the top eGenes, the lowest P-value of any associated eSNP was considered as the P-value for the gene.

eGenes were analyzed for the functional enrichment among Gene Ontology pathways (biological processes, cellular components, molecular functions). The statistical significance testing of the enrichment was implemented in FUMA[3] (Watanabe et al., 2017) and FDR was estimated using Benjamini–Hochberg correction for multiple testing. GWAS data was downloaded from GWAS Catalog.[4] SNPnexus (Dayem Ullah et al., 2018) was used to determine the closest gene for each

[2]http://www.ensembl.org/index.html

[3]http://fuma.ctglab.nl/

[4]https://www.ebi.ac.uk/gwas/docs/file-downloads, all associations v1.0_e91

GWAS hit. χ^2 test was used to determine enrichment of eGenes and eSNPs among GWAS results.

Taqman RT-qPCR Validation of *cis*-eQTLs in REPROMETA Placental Samples

Experimental validation of the discovery study targeted protein coding genes with multiple identified *cis*-SNPs that exhibited low FDR and at least two-fold gene expression difference between the placentas with heterozygote and major homozygote genotypes. The *cis*-eSNPs selected for the validation were rs1150707 (*ZSCAN9* c.568+1990 C > T); rs10044354 (*ERAP2* g.96984791 C > T) and rs11678251 (*ALPG* c.-318 G > A). For these SNPs, an extended REPROMETA placental sample set (n = 336; **Table 2**) was genotyped using Sequenom iPLEX Gold genotyping system (Sequenom, Agena, United States). Primers are provided in **Supplementary Table S1**. Based on the genotyping outcome, 24 placentas per each eSNP were selected for the Taqman RT-qPCR gene expression quantification (**Supplementary Table S2**). Taqman RT-qPCR and testing of eQTL effects targeted also three neighboring genes (*ALPP*, *ERAP1*, *LNPEP*).

eQTL testing was conducted as previously, except for using gestational age as an additional cofactor to account for the wider range of gestational age in validation samples. The expression levels for all samples were transformed to represent the fold-change from the median expression of major homozygotes used as the reference. Details for Taqman RT-qPCR validation are provided in **Supplementary Methods** and **Supplementary Table S3**.

HAPPY PREGNANCY and ALSPAC Pregnancy Cohorts and Data

The HAPPY PREGNANCY study recruited prospectively 2334 pregnant women during their first antenatal visit to Women's Clinic of Tartu University Hospital, Estonia. Longitudinal clinical and epidemiological data includes reproductive history, parental lifestyle, the course and outcome of pregnancy. The current study analyzed 408 placental samples with a specific focus on SGA newborns (n = 158; **Table 2**).

The ALSPAC initially recruited 14,541 pregnant women resident in Avon, United Kingdom, with expected dates of delivery April 1, 1991 to December 31, 1992 (Boyd et al., 2013; Fraser et al., 2013). For all recruited cases medical data from obstetric and perinatal records were documented. From the initial pregnancies, 14,062 resulted in live births. Gestational age at the delivery was recorded the nearest gestational week. The current study analyzed 7669 newborns with available genotype data (**Table 2**). Please note that the ALSPAC study website contains details of all the data that are available through a fully searchable data dictionary and variable search tool: http://www.bristol.ac.uk/alspac/researchers/our-data/.

The HAPPY PREGNANCY placental samples were genotyped for the rs11678251 (*ALPG* c.-318 G > A) using a pre-made Taqman assay (ID C__27838320_10, Applied Biosystems, Foster City, CA, United States). For the ALSPAC cohort, genotypes of the proxy SNP rs744873 (r^2 = 1.0 with rs11678251) were obtained from the genome-wide array dataset (Boyd et al., 2013).

Genetic Association Testing

Genetic association analysis was implemented in PLINK, ver. 1.09 (Purcell et al., 2007).[5] Differences in allelic distributions between term cases of normal pregnancy and gestational complications (PE, GD, SGA, LGA) in the discovery REPROMETA sample set was assessed using logistic regression (additive genetic model).

Genetic association tests with newborn parameters applied linear regression analysis. The REPROMETA placental genotypes (n = 336) for three *cis*-eQTLs rs1150707 (*ZSCAN9* c.568+1990 C > T), rs10044354 (*ERAP2* g.96984791 C > T) and rs11678251 (*ALPG*, c.-318 G > A) were tested in association with birth weight and length, placental weight, newborns' head and chest circumference and postnatal growth (**Supplementary Table S4**). The tests with the birth parameters were implemented under additive and recessive models, adjusted to newborn sex and gestational age (in days). Based on the outcome of the association testing with birth parameters, only the sex-adjusted recessive model was applied in further association analysis with infant postnatal growth. Infant height and weight had been documented at 6 (n = 233) and 12 (n = 216) months of age (**Supplementary Table S5**). For prematurely born children (<259 g days) postnatal growth was adjusted to the gestational age at birth (**Supplementary Methods**).

Suggestive association of the *ALPG* c.-318 G > A (rs11678251) with the newborn growth parameters in the REPROMETA study (under recessive model) was further assessed in independent cohorts HAPPY PREGNANCY (n = 408; Estonia) and ALSPAC (n = 7669; United Kingdom). For ALSPAC a proxy SNP rs744873 was analyzed. The test results of individual REPROMETA, HAPPY PREGNANCY and ALSPAC studies were combined in a meta-analysis under fixed effects model.

The obtained nominal P-values < 0.05 were considered as suggestive association. The Bonferroni corrected statistical significance level was estimated $P < 1.2 \times 10^{-3}$ for the 42 tests in the initial REPROMETA study and $P < 3.6 \times 10^{-3}$ for the 14 tests in the meta-analysis, respectively.

RESULTS

Discovery Analysis of Placental Expression Quantitative Trait Loci

Discovery analysis of the placental eQTLs was implemented using published RNA sequencing (Sõber et al., 2015) and whole genome genotyping (Kasak et al., 2015) datasets representing 40 unrelated term placentas collected during the REPROMETA study and representing a diverse range of pregnancy outcomes (**Table 1**). The analysis design targeted *cis*-eQTLs (±100 kbp from gene) and included genes with sufficient transcript levels (>100 reads; n = 11,733 genes after filtering) in order to avoid spurious associations (**Supplementary Figure S1**). Linear regression testing with 353,599 SNPs (Illumina HumanOmniExpress array) identified 199 placental *cis*-eSNPs corresponding to the applied stringent statistical significance threshold (FDR < 5%) (**Figure 1A**,

[5]http://pngu.mgh.harvard.edu/$\sim$purcell/plink/

TABLE 2 | REPROMETA, HAPPY PREGNANCY, and ALSPAC datasets utilized in the genetic association testing with newborn traits.

Parameter	REPROMETA (*n* = 336)	HAPPY PREGNANCY (*n* = 408)	ALSPAC (*n* = 7669)
Gestational age at delivery	273.4 [196 – 299] days	271.9 [167 – 294] days	39.5 [25 – 47] weeks
Newborn sex (F/M)	170/166	204/204	3732/3937
Placental weight (g)	585.1 [190 – 1122]	539.4 [190 – 1132]	657.1 [240 – 1260][a]
Birth weight (g)	3545 [990 – 5740]	3127.1 [560 – 4418]	3448.8 [645 – 5640]
Birth length (cm)	49.9 [35 – 57]	49.0 [29.5 – 56]	50.8 [34 – 62]
Head circumference (cm)	34.8 [26 – 40]	34.2 [22 – 38]	34.9 [20.3 – 54]
Chest circumference (cm)	34 [21 – 43]	32.8 [22.5 – 37]	NA
Preeclampsia (*n* [%])	43 [12.8%]	8 [2.0%]	150 [2.0%]
Gestational diabetes (*n* [%])	41 [12.2%]	10 [2.5%]	37 [0.5%]
SGA newborn (*n* [%])[b]	65 [19.3%]	158 [38.7%]	771 [10.1%]
LGA newborn (*n* [%])	83 [24.7%]	4 [1.0%]	764 [10.0%]
Maternal age (years)	28.7 [16 – 43]	28.9 [15.8 – 47.2]	28.6 [15 – 44]
Maternal height (cm)	167.1 [150 – 185]	166.4 [150 – 186]	164.2 [143.5 – 188.3][a]
Maternal pre-pregnancy weight (kg)	67.9 [43 – 142]	63.3 [40 – 106]	60.9 [31.6 – 141.6]
Gestational weight gain (kg)	15.5 [(−3) – 40]	12.1 [1.2 – 28]	12.6 [– 6.7 to 37.8]
Maternal smoking: no/yes (*n* [%])	306/21 [91.1%/6.3%]	385/23 [94.4%/5.6%]	5845/1595 [78.6%/21.4%]

Data is given as mean [minimum–maximum] unless indicated differently. REPROMETA clinical cases were recruited in 2006–2011 and HAPPY PREGNANCY cohort cases in 2013–2015 at the Women's Clinic of Tartu University Hospital, Estonia. ALSPAC cohort cases were recruited in 1991–1992 in Bristol area, United Kingdom. Detailed information on each study is provided in Materials and Methods. LGA, large-for-gestational age newborn; NA, not available; SGA, small-for-gestational age newborn. [a]Data were available for fewer than half of the samples. [b]SGA includes preeclampsia and preterm cases.

Supplementary Table S6A, and **Supplementary Data S1**). The proportion of gene expression explained by the identified eQTLs (R^2) varied from 0.45 to 0.77 (mean R^2 = 0.52) (**Figure 1B**). When taking into account linkage disequilibrium (LD; $r^2 > 0.8$) between the eSNPs, there were in total 88 independent *cis*-eQTL association signals representing 47 singleton eSNPs and 41 eSNP-clusters localized across the genome, except for some smaller chromosomes (#18, 20, 21, Y; **Supplementary Figure S2**). The densest coverage of independent eSNPs-eGene pairs was detected on chromosome 19 (n = 10 SNPs, 0.16 SNP/Mbp; n = 6 genes, 0.09 gene/Mbp), the most gene-rich chromosome in the human genome (Venter et al., 2001) (**Supplementary Table S7**). The highest number of eSNPs (23, including seven independent signals) was mapped for the *ZSCAN9* gene that has a potential function in the X-chromosome inactivation process (Luijk et al., 2018) (**Supplementary Table S6A**).

The associated eSNPs modulated the expression level of 63 eGenes expressed moderate to high in the human placenta (**Supplementary Tables S6A,B** and **Figure 1C**). The most significant placental eQTLs (FDR $< 10^{-5}$) with a positive minor allele effect were modulating the expression of genes encoding the ribosomal protein RPL9 [fold change, fc(het) = 1.64, fc(min hom) = 2.32], the transcription factor ZSCAN9 [fc(het) = 3.62, fc(min hom) = 4.54] and the aminopeptidase ERAP2 [fc(het) = 3.22, fc(min hom) = 4.65] (**Table 3**). In the literature, genetic variants around *ERAP2* have been associated with autoimmunity related disorders, whereas the *ZSCAN9* gene is surrounded by the risk variants for neuropsychiatric diseases and cancer (**Figures 2A,B**).

The largest eSNP effect was detected for the phospholipid transporter *ATP8A1* transcript levels [FDR = 7.44×10^{-4}; fc(het) = 4.64, fc(min hom) = 7.92]. Overall, the number of eQTLs with a positive effect on gene expression was twice as many compared to those with a negative effect. Top eQTLs exhibiting negative minor allele effect were associated with the expression of *DDX11*, *PLEKHG1*, and *THNSL2* implicated in fetal development, signaling and cellular proliferation, respectively (**Table 3**). Nearly 80% of the identified eGenes (n = 50) overlapped with the gene list reported by Peng et al. (**Figure 1D**). Despite the differences in the study design, ten of the 63 identified placental eGenes in our dataset fall within 1% and 45 genes within 25% of the loci with the lowest P-value in the previous independent dataset (Peng et al., 2017; **Supplementary Tables S6B,C**). As an added value we identified 13 novel placental eGenes (10 protein coding, 3 pseudogenes), including nine associations that were supported by the GTEx data in other tissues (**Table 4**). Amongst these, the identified eQTLs for the genes *ZSCAN9* and *TLDC1* were ranked within the top 10 gene expression associated variants in our analysis (**Table 3**). The expression level of *RBPJ*, *TCIM*, *TPRN*, and *THUMPD2* was associated with potential placenta specific eQTLs that have not been detected in the GTEx project.

Placental eQTLs and Risk to Term Pregnancy Complications

In the discovery sample, none of the placental eQTLs were significantly associated with adverse pregnancy outcomes at term after applying multiple testing correction (**Supplementary Table S6D**). Among nominal suggestive associations a substantial effect of *ZSCAN9* eSNP (rs1150707) was detected to the risk of GD (logistic regression; OR = 21, P = 0.02) or the birth of a SGA newborn (OR = 15.4, P = 0.04). This was not confirmed in the

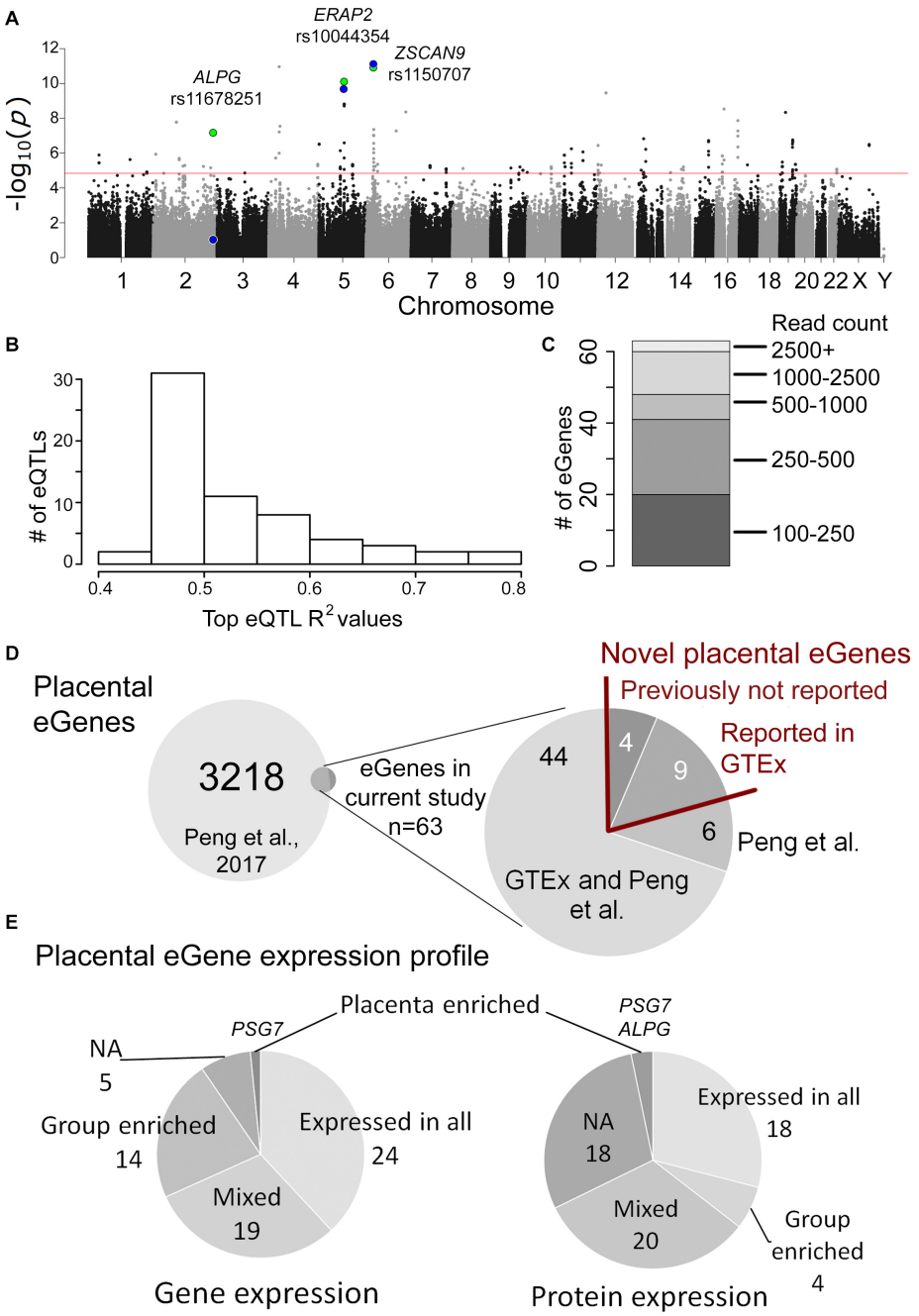


FIGURE 1 | Overview of the placental *cis*-eQTL analysis outcome. **(A)** Manhattan plot representing the landscape of *P*-values from the discovery analysis. The association *P*-values for the three eSNPs-eGene pairs selected for validation experiments are shown for the discovery (green) and validation (blue) analysis. The red line indicates the statistical significance threshold (FDR 5%). **(B)** The proportion of eGene expression (R^2) explained by the eSNPs detected in this study. **(C)** The expression levels of identified placental eGenes, shown in read counts from the RNA-Seq dataset (Söber et al., 2015). **(D)** The overlap of identified placental eGenes between the current study and Peng et al. (2017). **(E)** Profile of mRNA and protein expression of the placental eGenes according to Human Protein Atlas (www.proteinatlas.org).

analysis of the full REPROMETA sample set (GD: OR = 1.53, $P = 0.14$; SGA: OR = 1.18, $P = 0.45$).

Additional suggestive association with GD was detected for the eSNP of *PLEKHG1* (rs7738394 OR = 0.05, $P = 0.02$), previously linked to the risk of preeclampsia (Gray et al., 2018). An eSNP modulating the expression of *DNAJC15* was nominally associated with LGA (rs17553284; OR = 0.05, $P = 0.02$). Interestingly, this variant has been reported as a top signal driving differentiated allele-specific expression between African and European populations (Tian et al., 2018).

TABLE 3 | Top 10 protein coding eGenes from the discovery placental *cis*-eQTL analysis.

Lead eQTL chr:position	Alleles	MAF	eGene (*n*, eQTLs)	*P*-value *P* (FDR)[a]	Read count[b] (AA)	Read count (Aa)	Read count (aa)	fc[c] (Aa)	fc (aa)	R^2 for eQTL and gene expression[d]	eQTL in GTEx[e]	mRNA/ protein[f]	Peng et al.
rs2249563 4:39453758	G > A	21%	*RPL9* (3)	1.09×10^{-11} 9.68×10^{-7}	239 (180, 317)	392 (304, 454)	555 (*singleton*)	1.64	2.32	0.768	All tissues	All/all	Y
rs1150707 6:28229827[g]	C > T	30%	*ZSCAN9*[g] (23)	1.17×10^{-11} 9.68×10^{-7}	131 (92, 230)	475 (267, 686)	596 (493, 644)	3.62	4.54	0.767	Skin	All/mixed	N
rs10044354 5:96984791[g]	C > T	41%	*ERAP2*[g] (11)	8.63×10^{-11} 5.69×10^{-6}	470 (339, 1182)	1512 (813, 2354)	2189 (1252, 2602)	3.22	4.65	0.737	All tissues	Mixed/mixed	Y
rs10743750 12:31087400	C > T	28%	*DDX11* (2)	3.49×10^{-10} 1.92×10^{-5}	503 (326, 700)	355 (299, 502)	182 (145, 252)	0.71	0.36	0.713	All tissues	All/mixed	Y
rs7738394 6:150685788	T > C	30%	*PLEKHG1* (1)	4.34×10^{-9} 1.43×10^{-4}	2079 (1602, 2393)	1617 (1039, 1784)	825 (755, 895)	0.78	0.40	0.665	testis, esophagus	Enhanced (testis)/all	Y
rs7252798 19:21734730	C > T	46%	*ZNF100* (1)	4.57×10^{-9} 1.44×10^{-4}	187 (119, 280)	310 (201, 480)	427 (304, 588)	1.66	2.28	0.664	Most tissues	Mixed/NA	Y
rs567637 16:84500760	G > A	48%	*TLDC1* (5)	1.36×10^{-8} 4.07×10^{-4}	1241 (1085, 1837)	1886 (1157, 2722)	2465 (1870, 2956)	1.52	1.99	0.640	esophagus	Mixed/ mixed	N
rs10865489 2:88147373	T > G	31%	*THNSL2* (2)	1.66×10^{-8} 4.57×10^{-4}	577 (351, 724)	319 (217, 622)	171 (141, 191)	0.55	0.30	0.636	All tissues	Enhanced (parathyroid)/ mixed	Y
rs2612520 4:42645785	G > T	38%	*ATP8A1* (1)	2.86×10^{-8} 7.55×10^{-4}	60 (25, 263)	281 (42, 816)	479 (280, 821)	4.64	7.92	0.623	10 tissues adipose, breast etc.	Enhanced (parathyroid)/ mixed	Y
rs11678251 2:232406577[g]	G > A	11%	*ALPG*[g] (4)	7.62×10^{-8} 1.52×10^{-3}	372 (33, 834)	1336 (579, 1708)	NA	3.60	NA	0.600	colon, lung	Enhanced (fallopian tube)/placenta	Y

eQTL association testing was implemented in Matrix eQTL (Shabalin, 2012) using linear regression adjusted by the pregnancy outcome (normal term or PE, GD, SGA, LGA pregnancy), labor activity and newborn sex. [a]False discovery rate was calculated according to Benjamini and Hochberg method. [b]Normalized read count, median (min, max). [c]Fold change from median normalized read count of major homozygotes. [d]Fraction of gene expression variation explained by the eQTL genotypes (range 0–1). [e]GTEx reports eQTL data for 48 tissues, excluding placenta (https://gtexportal.org). [f]The Human Protein Atlas (https://www.proteinatlas.org/). [g]Selected for validation. ALPG, alkaline phosphatase, germ cell; ATP8A1, ATPase phospholipid transporting 8A1; chr, chromosome; DDX11, DEAD/H-box helicase 11; DNAJC15, DnaJ heat shock protein family (Hsp40) member C15; ERAP2, endoplasmic reticulum aminopeptidase 2; fc, fold change; MAF, minor allele frequency; NA, not available; PLEKHG1, pleckstrin homology and RhoGEF domain containing G1; ptg, parathyroid gland; RPL9, ribosomal protein L9; THNSL2, threonine synthase like 2; TLDC1, TBC/LysM-associated domain containing 1; ZNF100, zinc finger protein 100; ZSCAN9, zinc finger and SCAN domain containing 9.

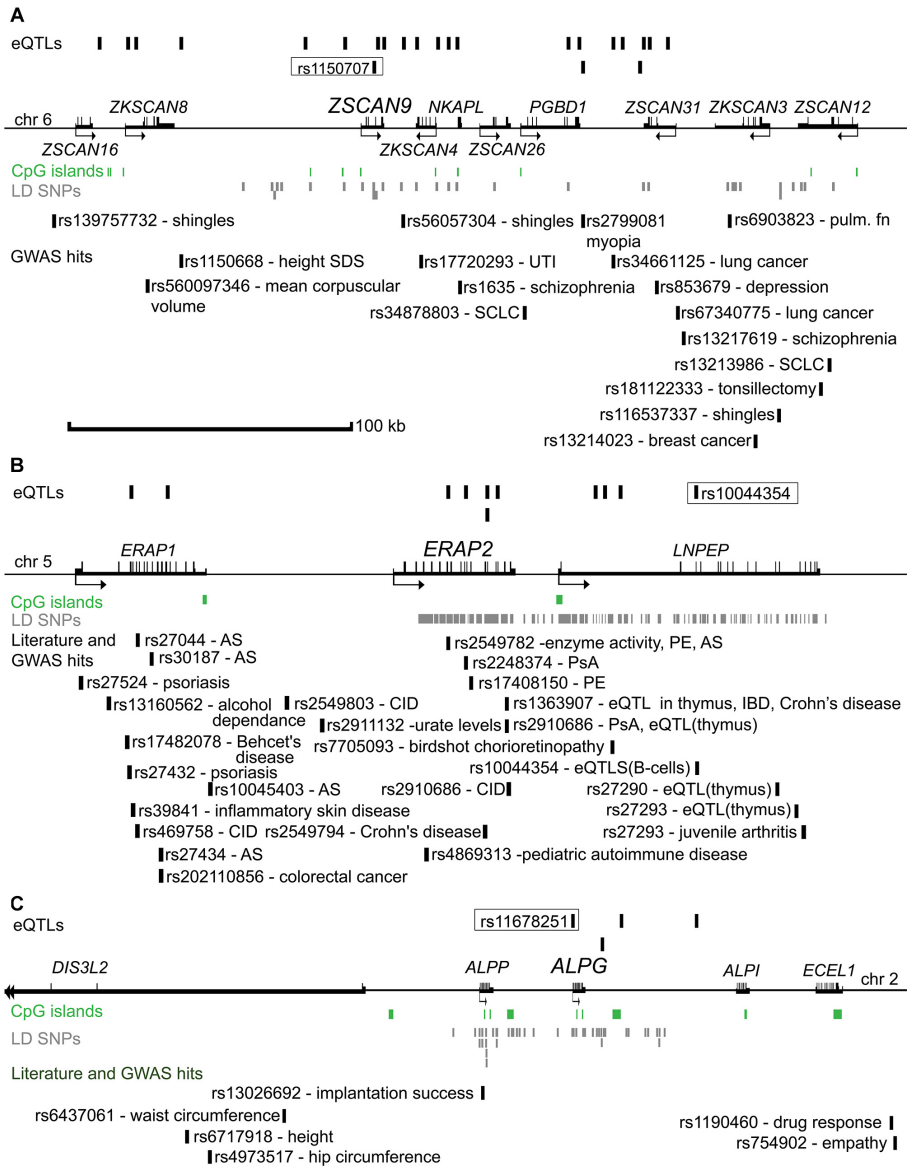


FIGURE 2 | Genomic regions surrounding the eSNP-eGene pairs selected for validation **(A)** *ZSCAN9*, **(B)** *ERAP2*, and **(C)** *ALPG*. Distances between the genes are drawn in approximate scale and the eGene of the region is highlighted with an increased font. eSNPs identified in the discovery study are shown above the genes. The eSNP chosen for experimental validation is boxed and its proxy SNPs in LD ($r^2 > 0.8$) are shown in gray below the genes. The landscape of reported genetic associations with common traits and diseases demonstrated at the bottom part of each subfigure was derived from the GWAS catalog and literature reports. The respective references are provided in **Supplementary Table S9**. *ALPG*, alkaline phosphatase, germ cell; *ALPI*, alkaline phosphatase, intestinal; *ALPP*, alkaline phosphatase, placental; AS, ankylosing spondylitis; chr, chromosome; CID, chronic inflammatory disease; *DIS3L2*, DIS3 like 3′-5′ exoribonuclease 2; *ECEL1*, endothelin converting enzyme like 1; eQTL, expression quantitative trait loci; *ERAP1*, Endoplasmic Reticulum Aminopeptidase 1; *ERAP2*, endoplasmic reticulum aminopeptidase 2; GWAS, genome-wide association study; IBD, inflammatory bowel disease; kb, kilobasepairs; LD, linkage disequilibrium; *LNPEP*, leucyl and cystinyl aminopeptidase; *NKAPL*, NFKB activating protein like; PE, preeclampsia; *PGBD1*, piggyBac transposable element derived 1; PsA, psoriatic arthritis; pulm. fn, pulmonary function; SCLC, Squamous cell lung carcinoma; SDS, standard deviation score; UTI, urinary tract infection; *ZKSCAN4*, zinc finger with KRAB and SCAN domains 4; *ZKSCAN8*, zinc finger with KRAB and SCAN domains 8; *ZSCAN9*, zinc finger and SCAN domain containing 9; *SCAN12*, zinc finger and SCAN domain containing 12; *ZSCAN16*, zinc finger and SCAN domain containing 16; *ZSCAN26*, zinc finger and SCAN domain containing 26; *ZSCAN31*, zinc finger and SCAN domain containing 31.

Placental eGenes Exhibit Broad Functional Portfolio

According to the Human Protein Atlas, the identified 63 eGenes exhibited mostly mixed or ubiquitous gene and protein expression (**Figure 1E**). Only few genes are specifically expressed in the placenta, e.g., placental alkaline phosphatase *ALPG* and *PSG7* (**Supplementary Tables S6A,B**). Based on the available literature evidence, the placental eGenes represent a broad

TABLE 4 | Characteristics of 13 novel placental eGenes, identified in the current study.

Lead *cis*-eSNP Chr: Position	Alleles MAF	eGene (number of *cis*-eSNPs)	*P*-value *P* (FDR)[a]	Read count[b]			Fold change[c]		R^2 for eQTL and gene expression[d]	Biological function	mRNA/protein[e]	GTEx[f]
				Maj hom	Het	Min hom	Het	Min hom				
rs1150707 chr6:28229827	T < C 0.30	*ZSCAN9* (23)	1.17×10^{-11} 9.68×10^{-7}	131	475	596	3.62	4.54	0.77	Transcription factor[g]	All/mixed	Y
rs1132812 chr16:30186830	A < G 0.45	*SMG1P5* (2)	2.93×10^{-9} 1.02×10^{-4}	145	440	610	3.03	4.21	0.67	Pseudogene	NA/NA	Y
rs567637 chr16:84500760	A < G 0.48	*TLDC1* (5)	1.36×10^{-8} 4.07×10^{-4}	1242	1886	2465	1.52	1.99	0.64	Cell proliferation (Nguyen et al., 2018)	Mixed/all	Y
rs9320475 chr6:113922342	T < C 0.21	*FO393415.1* (1)	5.34×10^{-8} 1.26×10^{-3}	91	241	255	2.65	2.80	0.61	Pseudogene	NA/NA	Y
rs3810756 chrX:119417694	T < C 0.05	*SLC25A43* (2)	3.67×10^{-7} 4.48×10^{-3}	192	131	123	0.68	0.64	0.56	Mitochondrial transporter (Gabrielson et al., 2016)	Mixed/enhanced (liver)	Y
rs10767971 chr11:32874118	T < C 0.35	*PRRG4* (4)	5.64×10^{-7} 6.45×10^{-3}	1023	1400	1923	1.37	1.89	0.55	Neuronal regulation (Justice et al., 2017)	Mixed/mixed	Y
rs1053846 chr1:39757468	G < T 0.21	*PPIE* (3)	1.29×10^{-6} 1.24×10^{-2}	199	306	394	1.53	1.98	0.52	Splicing (Chanarat and Sträßer, 2013)	All/all	Y
rs2871198 chr4:26290933	C < T 0.33	*RBPJ* (1)	1.91×10^{-6} 1.68×10^{-2}	2663	3105	3703	1.17	1.39	0.51	Transcriptional regulation (Han et al., 2002)	All/NA	N
rs7046565 chr9:109539940	C < T 0.43	*YBX1P6* (1)	6.26×10^{-6} 3.22×10^{-2}	32	156	164	4.87	5.15	0.48	Pseudogene	NA/NA	Y
rs10869496 chr9:75085168	A < C 0.36	*NMRK1* (1)	7.20×10^{-6} 3.52×10^{-2}	433	353	261	0.82	0.60	0.47	NAD metabolism (Bieganowski and Brenner, 2004)	All/mixed	Y
rs4370521 chr8:40111873	G < A 0.33	*TCIM* (1)	7.63×10^{-6} 3.62×10^{-2}	300	473	631	1.58	2.10	0.47	Cell cycle regulator (Jung et al., 2006)	Mixed/mixed	N
rs7850758 chr9:137293773	G < A 0.06	*TPRN* (1)	1.29×10^{-5} 4.69×10^{-2}	218	304	NA	1.40	NA	0.45	Hearing (Li et al., 2010)	All/all	N
rs11692913 chr2:39853203	A < G 0.35	*THUMPD2* (1)	1.44×10^{-5} 4.77×10^{-2}	217	230	281	1.06	1.29	0.45	tRNA metabolism[g]	Mixed/mixed	N

eQTLs were tested using linear regression adjusted by the pregnancy outcome (normal term or PE, GD, SGA, LGA pregnancy), labor activity and newborn sex. [a]False discovery rate was calculated according to Benjamini and Hochberg method. [b]Normalized read count, median (min, max). [c]Fold change compared to median normalized read count of major homozygotes. [d]Fraction of gene expression variation explained by the cis-eSNP genotypes (scale 0–1). [e]Data from the Human Protein Atlas (https://www.proteinatlas.org/). [f]eQTL identified in GTEx for at least one tissue. [g]Predicted annotation from UniprotKB. Maj hom, major homozygote; Het, heterozygote; Min hom, Minor homozygote; NMRK1, nicotinamide riboside kinase 1; PPIE, peptidylprolyl isomerase E; PRRG4, proline rich and Gla domain 4; RBPJ, Recombination signal binding protein for immunoglobulin kappa J region; SLC25A43, solute carrier family 25 member 43; SMG1P5, SMG1 pseudogene 5; TCIM, transcriptional and immune response regulator; THUMPD2, THUMP domain containing 2; TLDC1, TBC/LysM-associated domain containing 1; TPRN, taperin; YBX1P6, Y-box binding protein 1 pseudogene 6; ZSCAN9, Zinc finger and SCAN domain containing 9.

portfolio of functional categories (**Table 5**). Almost 1/3 of eGenes were associated with general cellular functions [structure (n = 10, e.g., *DCTN5*, *TTLL4*) and transport (n = 10, e.g., *AQP11*, *SLC27A6*)]. About 10% of genes are implicated in the immune function and immune-defense mechanisms that are critical in maintaining healthy pregnancy (n = 6, e.g., *ERAP2*, *TRIM5*). A functional enrichment analysis of eGenes highlighted 35 Gene Ontology (GO) categories (**Supplementary Table S6E**). Interestingly, the most significantly enriched pathways were "ruffle membrane" (*IFIT5*, *ADAM17*, *EPB41L5*, FDR = 1.81×10^{-2}) contributing to the formation of motile cell surface and "ATPase activity, coupled" (*ATP1A4*, *DDX11*, *RFC3*, *ATP8A1*, *PEX6*; FDR = 2.88×10^{-2}), critical for the active transport of molecules across cell membrane. Among the significantly (FDR < 0.05) enriched GO categories there were also several pathways implicated in development, signaling and immune function.

To obtain more potential functional insight, the identified eSNPs and eGenes were investigated for the overlap with association signals from published GWAS. Six placental eSNPs have been directly associated with either blood metabolite levels [eSNP-eGenes pairs: rs2576452-*TMC4*, rs2041073-*HEATR4* (Shin et al., 2014), rs6743376-*IL36RN* (Matteini et al., 2014)], risk to neurodegenerative diseases [rs10767971-*PRRG4* (Latourelle et al., 2009), rs1129187-*PEX6* (Jun et al., 2016)] or cancer [rs12309274-*RAD52* (Zhang et al., 2014)] (**Supplementary Table S6F**). These eSNP-eGene pairs have also been reported in GTEx with an effect in at least one other tissue. However, there was no statistically significant enrichment of GWAS hits among the identified eSNPs (6 of 199) compared to the overall proportion of GWAS SNPs (11,475 of 353,599) among the tested variants (χ^2 test: OR = 0.93 [0.41 – 2.09], P = 0.85). In total 58 of 63 of the eGenes were located within 100 kb of various reported GWAS loci and 45 of them were assigned as the closest gene (**Supplementary Table S6G**). However, due to the abundance of the GWAS associated genes (GWAS Catalog: 7433 of 11,733 genes tested in this study), this does not represent a significant enrichment (OR = 1.44 [0.84 – 2.50], P = 0.19).

Validation of the eSNPs of *ALPG*, *ERAP2*, and *ZSCAN9* Using Taqman RT-qPCR

From amongst the most significant eQTL signals, three (*ZSCAN9* c.568+1990 C > T, *ERAP2* g.96984791 C > T, *ALPG* c.-318 G > A) were selected for validation experiments in an extended set of genotyped REPROMETA placental samples (n = 24/gene; see section "Materials and Methods"; **Supplementary Table S2**). In order to clarify potential eSNP effects on the neighboring genes, the analysis included also the *ALPP* gene that represents a duplicate locus to *ALPG*, and the *ERAP1* and *LNPEP* genes flanking *ERAP2* (**Figure 2**). The eQTL effect on expression levels of *ZSCAN9* (linear regression, additive model FDR = 3.6×10^{-11}) and *ERAP2* (FDR = 1.2×10^{-9}) were robustly validated (**Figure 3A** and **Supplementary Table S8**). The expression of *ERAP1* and *LNPEP* was not modulated by *ERAP2* g.96984791 C > T. Although in the validation dataset the effect of *ALPG* c.-318 G > A on *ALPG* expression level did not reach statistical significance (FDR = 0.18), the placentas carrying the minor A-allele exhibited a trend for increased *ALPG* and decreased *ALPP* expression (*ALPP/ALPG* ratio FDR = 1.2×10^{-2}; **Supplementary Table S8**). This observation requires further confirmation as there is a substantial difference between the expression levels of *ALPP* and *ALPG* that may have led to a statistical artifact.

Suggestive Association of the *ALPG* -318 G > A Variant With Newborn Growth Parameters in the REPROMETA Study

Association testing targeting the highlighted eSNPs near *ZSCAN9*, *ERAP2*, and *ALPG* was carried out with newborn parameters in the full REPROMETA sample set (n = 336) applying linear regression (**Table 2**). None of the tested associations resulted in a statistically significant outcome after correction for multiple testing. However, the *ALPG* c.-318 G > A exhibited a suggestive association under recessive model with the placental weight (AA-genotype effect −202 g, nominal $P = 9.3 \times 10^{-3}$), as well as newborns' chest circumference (−2.1 cm, $P = 3.2 \times 10^{-3}$) and weight (−856 g, $P = 1.0 \times 10^{-2}$)

TABLE 5 | Identified eGenes classified by their main functional category.

Function	*n*	Genes
Cellular transport	10	*AQP11*, *ATP1A4*, *ATP8A1*, *CNIH4*, *HEATR5A*, *SLC25A43*, *SLC27A6*, *SLC44A1*, *SNX25*, *TMC4*, *DNAJC15*
Cell structure	10	*CEP128*, *CEP72*, *DCTN5*, *EPB41L5*, *FAM118A*, *HEATR4*, *LYPD5*, *NEO1*, *PEX6*, *TPRN*, *TSGA10*, *TTLL4*
Transcription	7	*TCIM*, *TRIM66*, *RBPJ*, *ZNF100*, *ZNF266*, *ZNF749*, *ZSCAN9*
Immunity/defense	6	*CBLB*, *ERAP2*, *IFIT5*, *IL36RN*, *PSG7*, *TRIM5*
Enzymatic activity	5	*ALPG*, *ATG10*, *IP6K3*, *NMRK1*, *THNSL2*
DNA replication/repair	4	*DDX11*, *RAD52*, *RFC3*, *CYREN*
Translational regulation and protein modification	4	*SPSB2*, *RPL9*, *PPIE*, *THUMPD2*
Signaling	4	*ADAM17*, *GLS*, *PLEKHG1*, *PRRG4*
Cellular proliferation and differentiation	2	*TLDC1*, *SLFN5*
Other	4	*ABHD11*, *PSMD5*, *THNSL2*, *WDR91*
Unclassified	1	*HEATR5A*
Pseudogene	5	*FO393415*, *GUCY1B2*, *YBX1P6*, *SMG1P5*, *HTR7P1*

Genes were grouped according to data retrieved from UniProtKB/Swiss-Prot database, Entrez gene database and/or literature.

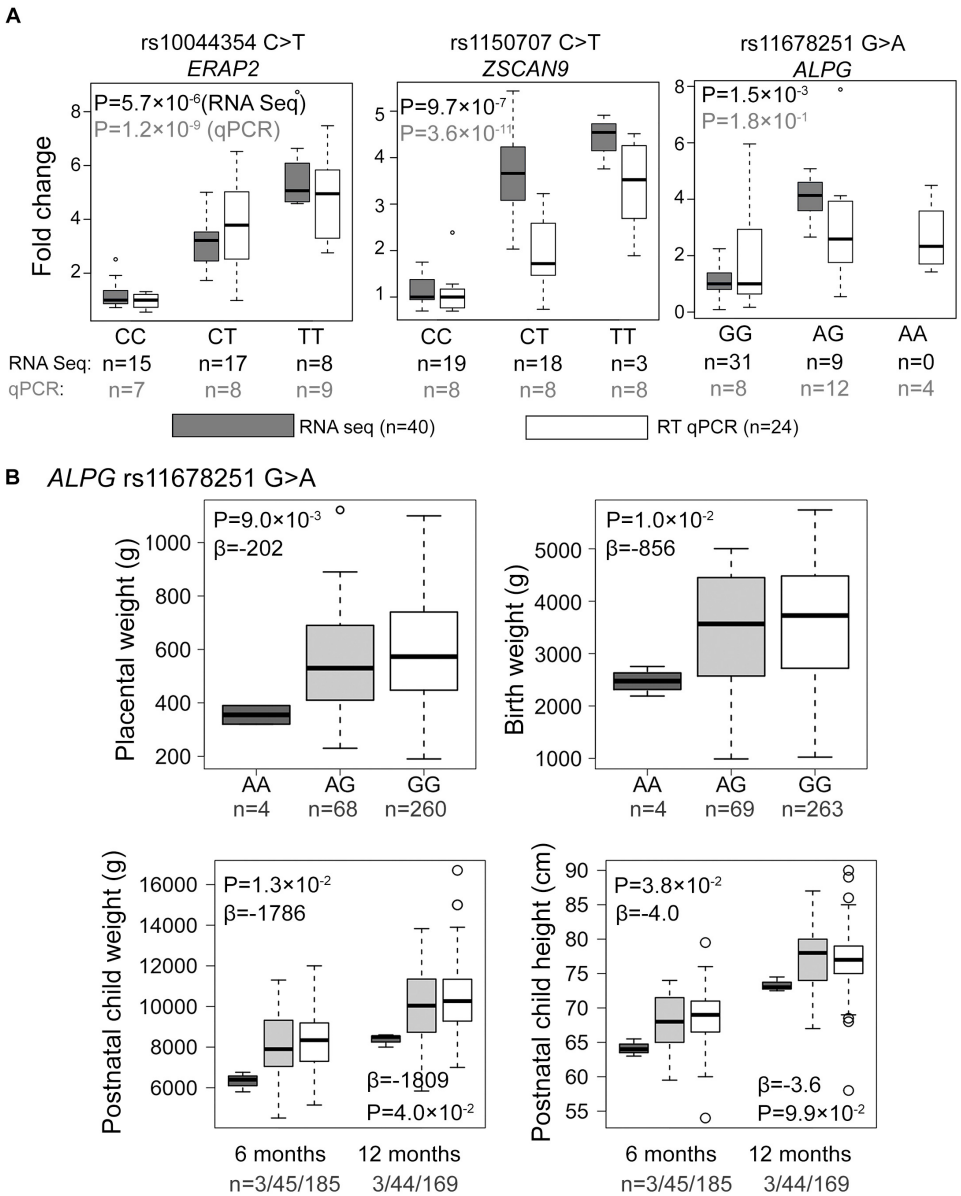


FIGURE 3 | Experimental validation and genetic association testing with newborn parameters. **(A)** Comparative effect sizes from the RNA-Seq discovery and Taqman RT qPCR experiments for the eSNP-eGene associations selected for the experimental validation. For the estimation of fold changes, median expression level in the placentas with the major homozygote genotype was considered as the reference. Statistical analysis for association testing was performed under linear regression using additive model adjusted by the newborn sex, pregnancy complication group and labor activity. As the validation dataset was not matched for gestational age at delivery, this parameter was additionally incorporated as a covariate in the validation step. The shown *P*-values have been corrected multiple testing using FDR method. *n*, number of samples. **(B)** Effect of the *ALPG* c.-318 G > A (rs11678251) eSNP on the offspring growth parameters at birth and during infancy in the REPROMETA dataset. The data was available for 336 newborns at delivery, and their follow-up data at the of 6 (n = 233) and 12 (n = 216) months of age. Genetic association testing was performed using linear regression under recessive model adjusted by fetal sex. In testing the newborn parameters, gestational age at delivery was used as an additional covariate. The obtained nominal *P*-values < 0.05 were considered as supportive for the trend of a tested association. Beta values reflect the estimated effect of the AA-homozygosity on the tested parameter.

(**Figure 3B** and **Table 6**). Supportingly, previous studies have associated SNPs near *ALPG* with other anthropometric traits (height, waist and hip circumference) and implantation success (**Figure 2C**). Also, *ZSCAN9* c.568+1990 C > T showed a nominal association with placental weight (TT-homozygotes: −54 g, $P = 4.4 \times 10^{-2}$).

The eSNPs were further tested for the association with the growth of the REPROMETA children during their first year of life. Based on the outcome of the association testing with birth parameters, only the effect of homozygosity (recessive model) was applied. Six and twelve months old infants who were homozygous for the *ALPG* c.-318 A-allele had maintained

TABLE 6 | Association testing of validated eQTL with newborn traits in the REPROMETA study (n = 335).

Tested parameter and genetic model	*ALPG*		*ERAP2*		*ZSCAN9*	
	G > A rs11678251		C > T rs10044354		C > T rs1150707	
	Effect (SE)	*P*-value	Effect (SE)	*P*-value	Effect (SE)	*P*-value
Placental weight (g)						
Additive	−48 (19)	**1.0×10^{-2}**	6 (12)	6.1×10^{-1}	−12 (13)	3.4×10^{-1}
Recessive	−202 (77)	**9.0×10^{-3}**	−3 (21)	8.9×10^{-1}	−54 (27)	**4.4×10^{-2}**
Birth weight (g)						
Additive	−161 (81)	**4.7×10^{-2}**	42 (50)	4.1×10^{-1}	−59 (55)	2.8×10^{-1}
Recessive	−856 (332)	**1.0×10^{-2}**	64 (92)	4.9×10^{-1}	−145 (116)	2.1×10^{-1}
Birth length (cm)						
Additive	−0.4 (0.3)	1.4×10^{-1}	0.2 (0.2)	3.0×10^{-1}	−0.2 (0.2)	4.3×10^{-1}
Recessive	−1.9 (1.2)	1.2×10^{-1}	0.4 (0.3)	2.7×10^{-1}	−0.4 (0.4)	4.1×10^{-1}
Head circumference at birth (cm)						
Additive	−0.3 (0.2)	2.0×10^{-1}	0.1 (0.1)	3.4×10^{-1}	−0.3 (0.1)	7.2×10^{-2}
Recessive	−1.1 (0.9)	2.0×10^{-1}	0.2 (0.2)	3.4×10^{-1}	−0.5 (0.3)	1.2×10^{-1}
Chest circumference at birth (cm)						
Additive	−0.3 (0.2)	1.6×10^{-1}	0.3 (0.2)	1.0×10^{-1}	−0.1 (0.2)	5.1×10^{-1}
Recessive	−2.1 (0.7)	**3.2×10^{-3}**	0.3 (0.4)	3.7×10^{-1}	−0.4 (0.4)	3.2×10^{-1}
At 6 months of age (only recessive model)[a]						
Weight	−1786 (711)	**1.3×10^{-2}**	120 (206)	5.6×10^{-1}	169 (263)	5.2×10^{-1}
Height	−4.0 (1.9)	**3.8×10^{-2}**	0.4 (0.6)	5.0×10^{-1}	0.2 (0.7)	8.4×10^{-1}
At 12 months of age (only recessive model)[a]						
Weight	−1809 (876)	**4.0×10^{-2}**	−85 (262)	7.5×10^{-1}	45.2 (330.3)	8.9×10^{-1}
Height	−3.6 (2.2)	9.9×10^{-2}	0.5 (0.7)	4.2×10^{-1}	0.27 (0.82)	7.4×10^{-1}

Association testing with birth parameters was implemented using linear regression analysis under additive and recessive model, adjusted by newborn sex and gestational as covariates. The obtained nominal P-values < 0.05 (in bold) were considered as supportive for the trend of a tested association. The Bonferroni corrected statistical significance level was estimated $P < 1.2 \times 10^{-3}$. [a]Postnatal growth data was available for 231 children (121 girls and 110 boys). Based on the outcome of the primary analysis with birth parameters, only recessive model (sex-adjusted) was applied for association testing with postnatal growth. ALPG, alkaline phosphatase, germ cell; ERAP2, endoplasmic reticulum aminopeptidase 2; ZSCAN9, zinc finger and SCAN domain containing 9.

lower weight compared to alternative genotype carriers (effect at 6 m: −1786 g, nominal $P = 1.3 \times 10^{-2}$; 12 m: 1809 g, $P = 4.0 \times 10^{-2}$; **Figure 3B**). Also, their height remained lower during 6 postnatal months (effect −4.0 cm, $P = 3.8 \times 10^{-2}$), but it was caught up by the 1st birthday ($P > 0.05$). eSNPs of *ERAP2* and *ZSCAN9* showed no association with infant growth parameters.

Meta-Analysis of *ALPG* -318 G > A in the HAPPY PREGNANCY and ALSPAC Cohort Samples

Association testing of the *ALPG* c.-318 G > A variant with the newborn anthropometric parameters was extended to two population-based cohorts HAPPY PREGNANCY (Estonia, n = 408) and ALSPAC (United Kingdom, n = 7669). The analysis in the individual cohorts and the meta-analysis combining the results from three studies did not replicate the initial associations with newborns' and placental weight detected in the REPROMETA study (**Table 7**). However, only in the meta-analysis the association of *ALPG* c.-318 G > A with the newborn's smaller head circumference reached nominal statistical significance (P = 0.042; effect −0.4 cm). The effect of the AA-genotype across the studies varied from −0.3 to −1.1 cm.

DISCUSSION

Genetic variants modulating human placental gene expression have been understudied and so far only one genome-wide analysis of placental eQTLs has been published with no external independent validation (Peng et al., 2017). The current report represents the second study addressing specifically the landscape of placental *cis*-eQTLs, including experimental validation of selected top eSNP-eGene pairs and further exploration of their potential effect on human traits. The study identified 88 independent eSNP signals modulating the expression of 63 placental genes (termed as eGenes). Importantly, 50 loci overlapped with the placental eGene list from Peng et al. (2017) and thus, can be claimed as robust placental eQTL signals to be explored further for their effect on placental function, fetal development, neonatal outcomes and postnatal disease risks (**Supplementary Table S6B**). Unfortunately, the direction of these eSNP effects on the expression of the eGenes could not be compared among the studies as the reference allele was not equivocally stated in Peng et al. (2017).

The current study identified additional 13 novel placental eGenes (**Table 4**). Among these, *ZSCAN9* represented an interesting novel locus modulated by several linked genetic variants and their highly significant eQTL effect was confirmed in

experimental validation (**Figure 3A**, **Table 3**, and **Supplementary Table S6A**). A recent study reported that *ZSCAN9 cis*-eQTLs exhibit female specific effects on human X-chromosomal methylation (Luijk et al., 2018). The ancestral A-allele of the variant rs1736891 was associated *in cis* with hypomethylated CpGs and high *ZSCAN9* expression, and *in trans* with hypomethylation of the CpG islands near X-chromosomal genes variably escaping X-chromosomal inactivation. Also, in the current study, rs1736891 was detected among the top placental eQTLs associated with the *ZSCAN9* transcript levels (**Supplementary Table S6A**). The placenta is well known for its unusual hypomethylated epigenome and thus, the *ZSCAN9* eSNPs may further contribute to the organ-specific enhanced expression of X-chromosomal genes that escape inactivation. As an additional interesting observation from the current study, worth to be followed-up in a larger cohort, was a potential effect of *ZSCAN9* eQTLs to the risk of GD. Other interesting placental eSNP-eGene pairs are rs2871198-*RBPJ* and rs4370521-*TCIM* that have not been reported in GTEx. TCIM is a positive regulator of the Wnt/beta-catenin pathway (Jung et al., 2006) and RBPJ regulates the transcription of the Notch signaling pathway genes via recruitment of chromatin remodeling complexes to its targets (Han et al., 2002). Both pathways are critical in early human development.

When comparing the current placental eSNP analysis outcome with the previous study (Peng et al., 2017), the number of identified eGenes appears not as extensive (n = 63 vs. 3218; **Supplementary Table S6C**). However, in order to minimize false-positive predictions and to avoid spurious associations that are common in eQTL analysis (Huang et al., 2018), several stringent criteria for the eQTL detection were applied. One core difference between the two studies was in defining *cis*-eQTLs (±100 kbp vs. ±500 kbp from gene), resulting in a five-fold smaller proportion of the genome targeted to the *cis*-eQTL testing compared to the previous study (Peng et al., 2017). Another critical aspect was in applying a more conservative statistical significance threshold for claiming eQTL effect (5% vs. 10% FDR) that further reduced the number of robust claims. Thirdly, while this study relied only on directly genotyped genetic variants (n = 661,354), the previous placental eQTL analysis included also imputed SNPs (n = 5,748,854) that possibly increased the number of identified associations. Taken together, the careful study design facilitated the detection of high-confidence placental eSNPs, the majority of those were confirmed either by the independent dataset (Peng et al., 2017) or by the experimental approach. However, it has to be acknowledged that as the number of analyzed samples was modest [n = 40 vs. 199 (Peng et al., 2017)], this possibly affected the study power.

A critical role of placental function in fetal programming and its further impact on the health across the life span has been proposed already several years ago (Bonnin et al., 2011; Longtine and Nelson, 2011; Kwon and Kim, 2017). As an example, the two haplotype variants of the placenta-specific *GH2* gene encoding placental growth hormone have been associated with the programming of adult height (Timasheva et al., 2013). In the current study, six eSNPs represented directly GWAS hits (6.8% of independent eQTL signals) and 45/63 eGenes were assigned

TABLE 7 | Meta-analysis of the genetic association testing of the *ALPG* c.-318 G > A (rs11678251) with newborn traits in the REPROMETA, HAPPY PREGNANCY, and ALSPAC studies.

Newborn parameter	REPROMETA			HAPPY PREGNANCY			ALSPAC			Meta-analysis		
	Effect (SE)	*P*-value	*n*	Effect (SE)	*P*-value	*n*	Effect (SE)	*P*-value	*n*	Effect (SE)	*P*-value	*n*
All newborns												
Birth weight (g)	−856 (332)	**1.0×10^{-2}**	335	−199 (164)	2.3×10^{-1}	408	−61 (57)	2.9×10^{-1}	7669	−96 (53)	7.2×10^{-2}	8412
Placental weight (g)	−202 (77)	**9.0×10^{-3}**	331	−19 (46)	6.8×10^{-1}	406	29 (27)	2.9×10^{-1}	3123	−2 (22)	9.2×10^{-1}	3860
Birth length (cm)	−1.9 (1.2)	1.2×10^{-1}	330	−0.7 (0.7)	3.3×10^{-1}	390	−0.3 (0.3)	3.6×10^{-1}	6077	−0.4 (0.3)	1.2×10^{-1}	6797
Head circumference (cm)	−1.1 (0.9)	2.0×10^{-1}	328	−0.8 (0.5)	1.1×10^{-1}	343	−0.3 (0.2)	1.8×10^{-1}	6159	−0.4 (0.2)	**4.2×10^{-2}**	6830
Chest circumference (cm)	−3.53 (1.25)	**5.1×10^{-3}**	325	−0.86 (0.7)	2.2×10^{-1}	337	*NA*	NA	NA	−1.49 (0.61)	**1.5×10^{-2}**	662

Linear regression analysis under recessive model adjusted by covariates newborn sex and gestational age (in days for REPROMETA and HAPPY PREGNANCY, in weeks for ALSPAC) was applied. The obtained nominal P-values < 0.05 (in bold) were considered as a suggestive association. Bonferroni significance level corrected for the number of individual tests was calculated 0.05/14 = 3.6×10^{-3}. ALPG, alkaline phosphatase, germ cell; NA, not available; SE, standard error.

as the closest gene to various GWAS SNPs for biomedical traits or diseases (**Supplementary Table S6G** and **Figure 2**). However, we were not able to confirm a significant enrichment of GWAS top hits among the detected placental eSNP and eGenes that was reported by the previous placental eQTL study (Peng et al., 2017).

Altered placental gene expression and the malfunctioning placenta are well-acknowledged risk factors for pregnancy complications and impaired fetal development (Sõber et al., 2015, 2016; Ashar-Patel et al., 2017; Sundrani et al., 2017; Sifakis et al., 2018). Our pilot analysis in the REPROMETA study detected a suggestive association of the AA-genotype of the eSNP *ALPG* c.-318 (rs11678251) with the reduced weight of the placenta, newborn, and infant until 1 year of age (**Figure 3B** and **Table 6**). *ALPG* represents one of the core trophectoderm (TE)-specific genes that drives the first cellular differentiation to inner cell mass and TE in early human development (Bai et al., 2012). However, the identified associations were not replicated in two independent and larger pregnancy cohorts of the HAPPY PREGNANCY and ALSPAC studies (**Table 7**). As the number of common genetic (and non-genetic) factors affecting anthropometric parameters is considered high and the effect sizes of individual variants are expected to be small (Wood et al., 2014; Tachmazidou et al., 2017), even a larger dataset than utilized in the current meta-analysis (n = 8412) would be required to confirm or reject the preliminary findings.

In summary, the study robustly demonstrated the role of genetic variation in driving the transcriptome profile of the human placenta and emphasized the importance to explore further the link between placental eQTLs, prenatal developmental programming and susceptibility to complex diseases.

DATA AVAILABILITY

The raw data supporting the conclusions of this manuscript will not be made publicly available to protect participant confidentiality. Clinical RNA-seq data will be made available by the authors upon request to any qualified researcher, pending ethical approval.

ETHICS STATEMENT

The REPROgrammed fetal and/or maternal METAbolism (REPROMETA) and HAPPY PREGNANCY (full study name "Development of novel non-invasive biomarkers for fertility and healthy pregnancy") studies were approved by the Ethics Review Committee of Human Research of the University of Tartu, Estonia (Permissions Nos. 146/18, 27.02.2006; 150/33, 18.06.2006; 158/80, 26.03.2007; 221/T-6, 17.12.2012; and 286/M-18, 15.10.2018). All study participants were recruited, and the study material was collected at the Women's Clinic of Tartu University Hospital, Estonia in 2006–2011 (REPROMETA) and in 2013–2015 (HAPPY PREGNANCY). All participants in the REPROMETA and HAPPY PREGNANCY studies were of white European ancestry and living in Estonia. A written informed consent to participate in the study was obtained from each individual prior to recruitment. Ethical approval for the ALSPAC was obtained from the ALSPAC Ethics and Law Committee and the Local Research Ethics Committees. The study recruited pregnant women in Bristol area, United Kingdom in 1991–1992 (http://www.alspac.bris.ac.uk). Consent for biological samples was collected in accordance with the Human Tissue Act (2004). Informed consent for the use of data collected via questionnaires and clinics was obtained from participants following the recommendations of the ALSPAC Ethics and Law Committee. All procedures and methods in the three studies have been carried out in compliance with the guidelines of the Declaration of Helsinki.

AUTHOR CONTRIBUTIONS

ML conceived the study. TK, KR, and ML designed the analyses and experiments. RF and ML provided the study materials. KR coordinated the clinical phenotyping and sampling in Estonia. TK conducted the experiments. TK and RB analyzed the data. TK, KR, RB, RF, and ML interpreted the data. TK and ML wrote the manuscript. All authors contributed to critical reading and commenting of the manuscript, and final approval of manuscript.

FUNDING

This study was supported by the Estonian Research Council (Grant: IUT34-12 for ML) and the European Union through the European Regional Development Fund (project HAPPY PREGNANCY, 3.2.0701.12-0047; for ML and KR). The UK Medical Research Council and Wellcome Trust (Grant ref: 102215/2/13/2) and the University of Bristol provide core support for ALSPAC. This publication is the work of the authors and RF will serve as guarantor for the contents of this manuscript. ALSPAC GWAS data was generated by Sample Logistics and Genotyping Facilities at Wellcome Sanger Institute and LabCorp (Laboratory Corporation of America) using support from 23andMe. RF and RB are supported by Sir Henry Dale Fellowship (Wellcome Trust and Royal Society Grant: WT104150).

ACKNOWLEDGMENTS

We are grateful to all the participants of the REPROMETA, HAPPY PREGNANCY, and ALSPAC studies. Clinical personnel at the Women's Clinic, Tartu University Hospital for the assistance in recruitment and phenotyping in the REPROMETA and HAPPY PREGNANCY studies. Siim Sõber and Mario Reiman are thanked for the advice in data analysis, Evelin Gross for the assistance in collecting postnatal data of the REPROMETA children and Eve Laasik for the contribution in extracting placental DNAs of the HAPPY PREGNANCY cohort.

We are extremely grateful to all the families who took part in the ALSPAC study, the midwives for their help in recruiting them, and the whole ALSPAC team, which includes interviewers, computer and laboratory technicians, clerical workers, research scientists, volunteers, managers, receptionists, and nurses.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.00550/full#supplementary-material

REFERENCES

Aguet, F., Brown, A. A., Castel, S. E., Davis, J. R., He, Y., Jo, B., et al. (2017). Genetic effects on gene expression across human tissues. *Nature* 550, 204–213. doi: 10.1038/nature24277

Ardlie, K. G., Deluca, D. S., Segre, A. V., Sullivan, T. J., Young, T. R., Gelfand, E. T., et al. (2015). The genotype-tissue expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* 348, 648–660. doi: 10.1126/science.1262110

Ashar-Patel, A., Kaymaz, Y., Rajakumar, A., Bailey, J. A., Karumanchi, S. A., and Moore, M. J. (2017). FLT1 and transcriptome-wide polyadenylation site (PAS) analysis in preeclampsia. *Sci. Rep.* 7:12139. doi: 10.1038/s41598-017-11639-6

Bai, Q., Assou, S., Haouzi, D., Ramirez, J.-M., Monzo, C., Becker, F., et al. (2012). Dissecting the first transcriptional divergence during human embryonic development. *Stem Cell Rev.* 8, 150–162. doi: 10.1007/s12015-011-9301-3

Bieganowski, P., and Brenner, C. (2004). Discoveries of nicotinamide riboside as a nutrient and conserved NRK genes establish a Preiss-Handler independent route to NAD+ in fungi and humans. *Cell* 117, 495–502. doi: 10.1016/s0092-8674(04)00416-7

Bonnin, A., Goeden, N., Chen, K., Wilson, M. L., King, J., Shih, J. C., et al. (2011). A transient placental source of serotonin for the fetal forebrain. *Nature* 472, 347–350. doi: 10.1038/nature09972

Boyd, A., Golding, J., Macleod, J., Lawlor, D. A., Fraser, A., Henderson, J., et al. (2013). Cohort profile: the 'Children of the 90s'—the index offspring of the avon longitudinal study of parents and children. *Int. J. Epidemiol.* 42, 111–127. doi: 10.1093/ije/dys064

Chanarat, S., and Sträßer, K. (2013). Splicing and beyond: the many faces of the Prp19 complex. *Biochim. Biophys.* 1833, 2126–2134. doi: 10.1016/j.bbamcr.2013.05.023

Dayem Ullah, A. Z., Oscanoa, J., Wang, J., Nagano, A., Lemoine, N. R., and Chelala, C. (2018). SNPnexus: assessing the functional relevance of genetic variation to facilitate the promise of precision medicine. *Nucleic Acids Res.* 46, W109–W113. doi: 10.1093/nar/gky399

Deshpande, S. S., and Balasinor, N. H. (2018). Placental defects: an epigenetic perspective. *Reprod. Sci.* 25, 1143–1160. doi: 10.1177/1933719118766265

Ding, J., Gudjonsson, J. E., Liang, L., Stuart, P. E., Li, Y., Chen, W., et al. (2010). Gene expression in skin and lymphoblastoid cells: refined statistical method reveals extensive overlap in cis-eQTL signals. *Am. J. Hum. Genet.* 87, 779–789. doi: 10.1016/J.AJHG.2010.10.024

Fraser, A., Macdonald-Wallis, C., Tilling, K., Boyd, A., Golding, J., Davey Smith, G., et al. (2013). Cohort profile: the avon longitudinal study of parents and children: ALSPAC mothers cohort. *Int. J. Epidemiol.* 42, 97–110. doi: 10.1093/ije/dys066

Gabrielson, M., Reizer, E., Stål, O., and Tina, E. (2016). Mitochondrial regulation of cell cycle progression through SLC25A43. *Biochem. Biophys. Res. Commun.* 469, 1090–1096. doi: 10.1016/j.bbrc.2015.12.088

Gillies, C. E., Putler, R., Menon, R., Otto, E., Yasutake, K., Nair, V., et al. (2018). An eQTL landscape of kidney tissue in human nephrotic syndrome. *Am. J. Hum. Genet.* 103, 232–244. doi: 10.1016/j.ajhg.2018.07.004

Gray, K. J., Kovacheva, V. P., Mirzakhani, H., Bjonnes, A. C., Almoguera, B., DeWan, A. T., et al. (2018). Gene-centric analysis of preeclampsia identifies maternal association at PLEKHG1. *Hypertension* 72, 408–416. doi: 10.1161/HYPERTENSIONAHA.117.10688

Grundberg, E., Small, K. S., Hedman, Å. K., Nica, A. C., Buil, A., Keildson, S., et al. (2012). Mapping cis- and trans-regulatory effects across multiple tissues in twins. *Nat. Genet.* 44, 1084–1089. doi: 10.1038/ng.2394

Han, H., Tanigaki, K., Yamamoto, N., Kuroda, K., Yoshimoto, M., Nakahata, T., et al. (2002). Inducible gene knockout of transcription factor recombination signal binding protein-J reveals its essential role in T versus B lineage decision. *Int. Immunol.* 14, 637–645. doi: 10.1093/intimm/dxf030

Huang, Q. Q., Ritchie, S. C., Brozynska, M., and Inouye, M. (2018). Power, false discovery rate and Winner's Curse in eQTL studies. *Nucleic Acids Res.* 46:e133. doi: 10.1093/nar/gky780

Juhanson, P., Rull, K., Kikas, T., Laivuori, H., Vaas, P., Kajantie, E., et al. (2016). Stanniocalcin-1 hormone in nonpreeclamptic and preeclamptic pregnancy: clinical, life-style, and genetic modulators. *J. Clin. Endocrinol. Metab.* 101, 4799–4807. doi: 10.1210/jc.2016-1873

Jun, G., Ibrahim-Verbaas, C. A., Vronskaya, M., Lambert, J.-C., Chung, J., Naj, A. C., et al. (2016). A novel Alzheimer disease locus located near the gene encoding tau protein. *Mol. Psychiatry* 21, 108–117. doi: 10.1038/mp.2015.23

Jung, Y., Bang, S., Choi, K., Kim, E., Kim, Y., Kim, J., et al. (2006). TC1 (C8orf4) Enhances the Wnt/β-catenin pathway by relieving antagonistic activity of chibby. *Cancer Res.* 66, 723–728. doi: 10.1158/0008-5472.CAN-05-3124

Justice, E. D., Barnum, S. J., and Kidd, T. (2017). The WAGR syndrome gene PRRG4 is a functional homologue of the commissureless axon guidance gene. *PLoS Genet.* 13:e1006865. doi: 10.1371/journal.pgen.1006865

Kasak, L., Rull, K., Vaas, P., Teesalu, P., and Laan, M. (2015). Extensive load of somatic CNVs in the human placenta. *Sci. Rep.* 5:8342. doi: 10.1038/srep08342

Kwon, E. J., and Kim, Y. J. (2017). What is fetal programming?: a lifetime health is under the control of in utero health. *Obstet. Gynecol. Sci.* 60:506. doi: 10.5468/ogs.2017.60.6.506

Latourelle, J. C., Pankratz, N., Dumitriu, A., Wilk, J. B., Goldwurm, S., Pezzoli, G., et al. (2009). Genomewide association study for onset age in Parkinson disease. *BMC Med. Genet.* 10:98. doi: 10.1186/1471-2350-10-98

Li, Q., Seo, J.-H., Stranger, B., McKenna, A., Pe'er, I., LaFramboise, T., et al. (2013). Integrative eQTL-based analyses reveal the biology of breast cancer risk loci. *Cell* 152, 633–641. doi: 10.1016/J.CELL.2012.12.034

Li, Y., Pohl, E., Boulouiz, R., Schraders, M., Nürnberg, G., Charif, M., et al. (2010). Mutations in TPRN cause a progressive form of autosomal-recessive nonsyndromic hearing loss. *Am. J. Hum. Genet.* 86, 479–484. doi: 10.1016/j.ajhg.2010.02.003

Longtine, M., and Nelson, D. M. (2011). Placental dysfunction and fetal programming: the importance of placental size, shape, histopathology, and molecular composition. *Semin. Reprod. Med.* 29, 187–196. doi: 10.1055/s-0031-1275515

Lonsdale, J., Thomas, J., Salvatore, M., Phillips, R., Lo, E., Shad, S., et al. (2013). The genotype-tissue expression (GTEx) project. *Nat. Genet.* 45, 580–585. doi: 10.1038/ng.2653

Luijk, R., Wu, H., Ward-Caviness, C. K., Hannon, E., Carnero-Montoro, E., Min, J. L., et al. (2018). Autosomal genetic variation is associated with DNA methylation in regions variably escaping X-chromosome inactivation. *Nat. Commun.* 9:3738. doi: 10.1038/s41467-018-05714-3

Matteini, A. M., Li, J., Lange, E. M., Tanaka, T., Lange, L. A., Tracy, R. P., et al. (2014). Novel gene variants predict serum levels of the cytokines IL-18 and IL-1ra in older adults. *Cytokine* 65, 10–16. doi: 10.1016/j.cyto.2013.10.002

McKenzie, M., Henders, A. K., Caracella, A., Wray, N. R., and Powell, J. E. (2014). Overlap of expression quantitative trait loci (eQTL) in human brain and blood. *BMC Med. Genomics* 7:31. doi: 10.1186/1755-8794-7-31

Nguyen, J. T., Ray, C., Fox, A. L., Mendonça, D. B., Kim, J. K., and Krebsbach, P. H. (2018). Mammalian EAK-7 activates alternative mTOR signaling to regulate cell proliferation and migration. *Sci. Adv.* 4:eaao5838. doi: 10.1126/sciadv.aao5838

Paquette, A. G., Lester, B. M., Koestler, D. C., Lesseur, C., Armstrong, D. A., and Marsit, C. J. (2014). Placental FKBP5 genetic and epigenetic variation is associated with infant neurobehavioral outcomes in the RICHS cohort. *PLoS One* 9:e104913. doi: 10.1371/journal.pone.0104913

Pavličev, M., Wagner, G. P., Chavan, A. R., Owens, K., Maziarz, J., Dunn-Fletcher, C., et al. (2017). Single-cell transcriptomics of the human placenta: inferring the cell communication network of the maternal-fetal interface. *Genome Res.* 27, 349–361. doi: 10.1101/gr.207597.116

Peng, S., Deyssenroth, M. A., Di Narzo, A. F., Lambertini, L., Marsit, C. J., Chen, J., et al. (2017). Expression quantitative trait loci (eQTLs) in human placentas suggest developmental origins of complex diseases. *Hum. Mol. Genet.* 26, 3432–3441. doi: 10.1093/hmg/ddx265

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., et al. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* 81, 559–575. doi: 10.1086/519795

Reiman, M., Laan, M., Rull, K., and Sõber, S. (2017). Effects of RNA integrity on transcript quantification by total RNA sequencing of clinically collected human placental samples. *FASEB J.* 31, 3298–3308. doi: 10.1096/fj.201601031RR

Shabalin, A. A. (2012). Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinformatics* 28, 1353–1358. doi: 10.1093/bioinformatics/bts163

Shin, S.-Y., Fauman, E. B., Petersen, A.-K., Krumsiek, J., Santos, R., Huang, J., et al. (2014). An atlas of genetic influences on human blood metabolites. *Nat. Genet.* 46, 543–550. doi: 10.1038/ng.2982

Sifakis, S., Androutsopoulos, V., Pontikaki, A., Velegrakis, A., Papaioannou, G., Koukoura, O., et al. (2018). Placental expression of PAPPA, PAPPA-2 and PLAC-1 in pregnacies is associated with FGR. *Mol. Med. Rep.* 17, 6435–6440. doi: 10.3892/mmr.2018.8721

Sildver, K., Veerus, P., and Lang, K. (2015). Sünnikaalukõverad Eestis ja sünnikaalu mõjutavad tegurid: registripõhine uuring - Eesti Arst - Eesti Arstide Liidu ajakiri. *Eesti Arst* 94, 465–470.

Soares, M. J., Iqbal, K., and Kozai, K. (2017). Hypoxia and placental development. *Birth Defects Res.* 109, 1309–1329. doi: 10.1002/bdr2.1135

Sõber, S., Reiman, M., Kikas, T., Rull, K., Inno, R., Vaas, P., et al. (2015). Extensive shift in placental transcriptome profile in preeclampsia and placental origin of adverse pregnancy outcomes. *Sci. Rep.* 5:13336. doi: 10.1038/srep13336

Sõber, S., Rull, K., Reiman, M., Ilisson, P., Mattila, P., and Laan, M. (2016). RNA sequencing of chorionic villi from recurrent pregnancy loss patients reveals impaired function of basic nuclear and cellular machinery. *Sci. Rep.* 6:38439. doi: 10.1038/srep38439

Sundrani, D., Narang, A., Mehendale, S., Joshi, S., and Chavan-Gautam, P. (2017). Investigating the expression of *MMPs* and *TIMPs* in preterm placenta and role of CpG methylation in regulating *MMP-9* expression. *IUBMB Life* 69, 985–993. doi: 10.1002/iub.1687

Tachmazidou, I., Süveges, D., Min, J. L., Ritchie, G. R. S., Steinberg, J., Walter, K., et al. (2017). Whole-genome sequencing coupled to imputation discovers genetic signals for anthropometric traits. *Am. J. Hum. Genet.* 100, 865–884. doi: 10.1016/j.ajhg.2017.04.014

Tian, L., Khan, A., Ning, Z., Yuan, K., Zhang, C., Lou, H., et al. (2018). Genome-wide comparison of allele-specific gene expression between African and European populations. *Hum. Mol. Genet.* 27, 1067–1077. doi: 10.1093/hmg/ddy027

Timasheva, Y., Putku, M., Kivi, R., Kožich, V., Männik, J., and Laan, M. (2013). Developmental programming of growth: genetic variant in GH2 gene encoding placental growth hormone contributes to adult height determination. *Placenta* 34, 995–1001. doi: 10.1016/j.placenta.2013.08.012

Uusküla, L., Männik, J., Rull, K., Minajeva, A., Kõks, S., Vaas, P., et al. (2012). Mid-gestational gene expression profile in placenta and link to pregnancy complications. *PLoS One* 7:e49248. doi: 10.1371/journal.pone.0049248

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., et al. (2001). The sequence of the human genome. *Science* 291, 1304–1351. doi: 10.1126/science.1058040

Veyrieras, J.-B., Kudaravalli, S., Kim, S. Y., Dermitzakis, E. T., Gilad, Y., Stephens, M., et al. (2008). High-resolution mapping of expression-QTLs yields insight into human gene regulation. *PLoS Genet.* 4:e1000214. doi: 10.1371/journal.pgen.1000214

Watanabe, K., Taskesen, E., van Bochoven, A., and Posthuma, D. (2017). Functional mapping and annotation of genetic associations with FUMA. *Nat. Commun.* 8:1826. doi: 10.1038/s41467-017-01261-5

Winn, V. D., Haimov-Kochman, R., Paquet, A. C., Yang, Y. J., Madhusudhan, M. S., Gormley, M., et al. (2007). Gene expression profiling of the human maternal-fetal interface reveals dramatic changes between midgestation and term. *Endocrinology* 148, 1059–1079. doi: 10.1210/en.2006-0683

Wood, A. R., Esko, T., Yang, J., Vedantam, S., Pers, T. H., Gustafsson, S., et al. (2014). Defining the role of common variation in the genomic and biological architecture of adult human height. *Nat. Genet.* 46, 1173–1186. doi: 10.1038/ng.3097

Yao, C., Joehanes, R., Johnson, A. D., Huan, T., Liu, C., Freedman, J. E., et al. (2017). Dynamic role of trans regulation of gene expression in relation to complex traits. *Am. J. Hum. Genet.* 100, 571–580. doi: 10.1016/j.ajhg.2017.02.003

Yengo, L., Sidorenko, J., Kemper, K. E., Zheng, Z., Wood, A. R., Weedon, M. N., et al. (2018). Meta-analysis of genome-wide association studies for height and body mass index in ∼700000 individuals of European ancestry. *Hum. Mol. Genet.* 27, 3641–3649. doi: 10.1093/hmg/ddy271

Zhang, B., Jia, W.-H., Matsuda, K., Kweon, S.-S., Matsuo, K., Xiang, Y.-B., et al. (2014). Large-scale genetic study in East Asians identifies six new loci associated with colorectal cancer risk. *Nat. Genet.* 46, 533–542. doi: 10.1038/ng.2985

Zhang, M., Lykke-Andersen, S., Zhu, B., Xiao, W., Hoskins, J. W., Zhang, X., et al. (2018). Characterising cis-regulatory variation in the transcriptome of histologically normal and tumour-derived pancreatic tissues. *Gut* 67, 521–533. doi: 10.1136/gutjnl-2016-313146

Conflict of Interest Statement: 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: 30 April 2020
doi: 10.3389/fgene.2020.00337

Exome Sequencing Analysis Identifies Rare Variants in *ATM* and *RPL8* That Are Associated With Shorter Telomere Length

*Ashley van der Spek[1,2], Sophie C. Warner[3], Linda Broer[4], Christopher P. Nelson[3,5], Dina Vojinovic[1], Shahzad Ahmad[1], Pascal P. Arp[4], Rutger W. W. Brouwer[6], Matthew Denniff[3], Mirjam C. G. N. van den Hout[6], Jeroen G. J. van Rooij[4,7], Robert Kraaij[4], Wilfred F. J. van IJcken[6], Nilesh J. Samani[3,5], M. Arfan Ikram[1], André G. Uitterlinden[1,4], Veryan Codd[3,5], Najaf Amin[1,8] and Cornelia M. van Duijn[1,8]**

[1] Department of Epidemiology, Erasmus MC University Medical Center Rotterdam, Rotterdam, Netherlands, [2] SkylineDx B.V., Rotterdam, Netherlands, [3] Department of Cardiovascular Sciences, University of Leicester, Leicester, United Kingdom, [4] Department of Internal Medicine, Erasmus MC University Medical Center Rotterdam, Rotterdam, Netherlands, [5] NIHR Leicester Biomedical Research Centre, Glenfield Hospital, Leicester, United Kingdom, [6] Center for Biomics, Erasmus MC University Medical Center Rotterdam, Rotterdam, Netherlands, [7] Department of Neurology, Erasmus MC University Medical Center Rotterdam, Rotterdam, Netherlands, [8] Nuffield Department of Population Health, University of Oxford, Oxford, United Kingdom

OPEN ACCESS

Edited by:
Serena Sanna,
Institute for Genetic and Biomedical Research (IRGB), Italy

Reviewed by:
Eleonora Porcu,
Université de Lausanne, Switzerland
Nabila Bouatia-Naji,
Institut National de la Santé et de la Recherche Médicale (INSERM), France

***Correspondence:**
Cornelia M. van Duijn
Cornelia.vanDuijn@ndph.ox.ac.uk

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

Received: 17 September 2019
Accepted: 20 March 2020
Published: 30 April 2020

Citation:
van der Spek A, Warner SC, Broer L, Nelson CP, Vojinovic D, Ahmad S, Arp PP, Brouwer RWW, Denniff M, van den Hout MCGN, van Rooij JGJ, Kraaij R, van IJcken WFJ, Samani NJ, Ikram MA, Uitterlinden AG, Codd V, Amin N and van Duijn CM (2020) Exome Sequencing Analysis Identifies Rare Variants in ATM and RPL8 That Are Associated With Shorter Telomere Length. Front. Genet. 11:337. doi: 10.3389/fgene.2020.00337

Telomeres are important for maintaining genomic stability. Telomere length has been associated with aging, disease, and mortality and is highly heritable (∼82%). In this study, we aimed to identify rare genetic variants associated with telomere length using whole-exome sequence data. We studied 1,303 participants of the Erasmus Rucphen Family (ERF) study, 1,259 of the Rotterdam Study (RS), and 674 of the British Heart Foundation Family Heart Study (BHF-FHS). We conducted two analyses, first we analyzed the family-based ERF study and used the RS and BHF-FHS for replication. Second, we combined the summary data of the three studies in a meta-analysis. Telomere length was measured by quantitative polymerase chain reaction in blood. We identified nine rare variants significantly associated with telomere length (p-value $< 1.42 \times 10^{-7}$, minor allele frequency of 0.2–0.5%) in the ERF study. Eight of these variants (in *C11orf65*, *ACAT1*, *NPAT*, *ATM*, *KDELC2*, and *EXPH5*) were located on chromosome 11q22.3 that contains *ATM*, a gene involved in telomere maintenance. Although we were unable to replicate the variants in the RS and BHF-FHS (p-value ≥ 0.21), segregation analysis showed that all variants segregate with shorter telomere length in a family. In the meta-analysis of all studies, a nominally significant association with LTL was observed with a rare variant in *RPL8* (p-value $= 1.48 \times 10^{-6}$), which has previously been associated with age. Additionally, a novel rare variant in the known *RTEL1* locus showed suggestive evidence for association (p-value $= 1.18 \times 10^{-4}$) with LTL. To conclude, we identified novel rare variants associated with telomere length. Larger samples size are needed to confirm these findings and to identify additional variants.

Keywords: telomere, aging, whole exome sequencing, meta-analysis, ATM, RPL8

INTRODUCTION

Telomeres are DNA structures located at the ends of chromosomes and consist of tandem hexanucleotide sequence repeats (TTAGGG) (Blackburn and Gall, 1978). They are important for maintaining genomic stability by preventing DNA degradation and chromosomal fusions (Blackburn, 2000). Telomeres are shortened with each cell division due to the inability of DNA polymerase to fully extend the 3′ end of the DNA strand during replication. When the telomeres reach a critical length, this leads to cellular senescence and ultimately to cell death, making them regulators of the replicative capacity of a cell and markers of biological age (Lindsey et al., 1991; Allsopp et al., 1992).

Shorter leukocyte telomere length (LTL) has been associated with several age-related diseases including cardiovascular diseases (Brouilette et al., 2003, 2007; Benetos et al., 2004; Fitzpatrick et al., 2007; Willeit et al., 2010a; Haycock et al., 2014), cancer (Hastie et al., 1990; Willeit et al., 2010b, 2011; Bojesen et al., 2013) and dementia (Martin-Ruiz et al., 2006; Grodstein et al., 2008; Honig et al., 2012). LTL has also been associated with mortality (Cawthon et al., 2003; Kimura et al., 2008; Fitzpatrick et al., 2011; Honig et al., 2012; Deelen et al., 2014; Rode et al., 2015; Marioni et al., 2016). However, this association has been inconsistent (Martin-Ruiz et al., 2005; Harris et al., 2006; Njajou et al., 2009; Houben et al., 2011; Strandberg et al., 2011; Needham et al., 2015). LTL is highly heritable with heritability estimates ranging from 34 to 82% (Slagboom et al., 1994; Bischoff et al., 2005; Vasa-Nicotera et al., 2005; Andrew et al., 2006; Broer et al., 2013). Previously, genome-wide association studies (GWASs) in European ancestry studies have identified common variants associated with LTL located in multiple genes, including: *TERC* (Codd et al., 2010, 2013; Mangino et al., 2012; Pooley et al., 2013), *TERT* (Codd et al., 2013; Pooley et al., 2013), *NAF1* (Codd et al., 2013), *OBFC1* (Levy et al., 2010; Mangino et al., 2012; Codd et al., 2013; Pooley et al., 2013), *RTEL1* (Codd et al., 2013), *CTC1* (Mangino et al., 2012), *ZNF676* (Mangino et al., 2012), *ZNF208* (Codd et al., 2013), *ACYP2* (Codd et al., 2013), *DCAF4* (Mangino et al., 2015), and *PXK* (Pooley et al., 2013). However, these variants explain < 5% of the heritability.

Up until now, no systemic whole exome or genome screen for rare variants has been published, despite the fact that these may explain part of the heritability (Manolio et al., 2009). Rare variants are not well captured by microarrays used for GWAS and remain difficult to impute, despite the recent improvements in imputation panels (McCarthy et al., 2016). Next generation sequencing technologies, such as whole-exome sequencing (WES), are better suited to study rare variants. In this study, we present a dual analysis. First, we conducted a genome-wide WES analysis of LTL in 1,303 individuals of the Dutch Erasmus Rucphen Family (ERF) study to search for rare genetic variants associated with LTL. The advantage of a family-based study is that the segregation of rare variants can be studied. We performed a replication analysis in the Rotterdam Study (RS) and the British Heart Foundation Family Heart Study (BHF-FHS). Next, we pooled the data together and conducted a meta-analysis of the association results of all three cohorts.

RESULTS

Descriptive statistics of the family-based and population-based studies are provided in **Table 1**. Mean age at LTL measurement was 49 years (SD = 15.0) in the ERF study and 61% of the study participants were female. The RS participants were older ($\bar{x}_{age}$ = 75 years, SD = 7.7) and 57% of the participants were female, while mean age in the BHF-FHS was 58 years (SD = 8.2) and most study participants were male (26% female). Mean LTL values, measured in each participant using a quantitative polymerase chain reaction (qPCR) based technique as the ratio of telomere repeat length to copy number of the single copy gene *36B4*, were higher in the ERF study ($\bar{x}_{LTL}$ = 1.85, SD = 0.35) than in the RS ($\bar{x}_{LTL}$ = 0.94, SD = 0.18) and BHF-FHS ($\bar{x}_{LTL}$ = 1.37, SD = 0.22). After adjusting LTL values for age and sex, mean LTL was comparable between studies (**Table 1**).

The Manhattan plot and the distribution of the test statistic [quantile-quantile (QQ) plot, λ = 1.04] of the WES analysis in the ERF study are presented in **Figures 1**, **2**, respectively. We observed significant association of nine rare variants [Minor Allele Frequency (MAF) between 0.2% and 0.5%] with LTL as shown in **Table 2**. The significance threshold (p-value < 1.42×10^{-7}) was adjusted for multiple testing using Bonferroni correction based on the number of variants in the analysis (0.05/353,075 variants). Each variant was negatively associated with LTL and the estimated effects of the minor allele of these variants were large (-2.18 < standardized β < -1.34), suggesting a significant decrease in LTL for each minor allele.

The top eight variants are located in a dense region on chromosome 11q22.3 (position range: 108004687 – 108384666, **Figure 3**) and appear to be a part of a haplotype that spans the *C11orf65*, *ACAT1*, *NPAT*, *ATM*, *KDELC2*, and *EXPH5* genes. A haplotype can describe a pair of genes inherited together from one parent on one chromosome, or it can describe all of the genes on a chromosome that were inherited together from a single parent. This haplotype segregates with shorter LTL in a family (**Supplementary Figure 1**), where it is carried by 14 individuals, 11 of whom were related within 4 generations according to the

TABLE 1 | Descriptive statistics of the study populations.

	ERF	RS	BHF-FHS
N	1303	1257	674
Mean age (SD)	48.9 (15.0)	74.5 (7.7)	58.0 (8.2)
Age range	18.2–95.7	55.0–105.8	36.0–81.0
% female	60.5	57.0	25.8
Mean LTL (SD)	1.85 (0.35)	0.94 (0.18)	1.37 (0.22)
LTL range	0.77–3.17	0.31–1.79	0.69–2.14
Adjusted mean LTL (SD)*	8.85×10^{-18} (0.32)	1.37×10^{-17} (0.18)	1.11×10^{-10} (0.21)
Adjusted LTL range*	−1.15–1.08	−0.71–0.89	−0.63–0.71

*ERF = Erasmus Rucphen Family study; RS = Rotterdam Study; BHF-FHS = British Heart Foundation Family Heart Study; N = Number of variants; SD = Standard Deviation, LTL = Leukocyte Telomere Length. *LTL values were adjusted for age and sex and information on residuals is shown.*

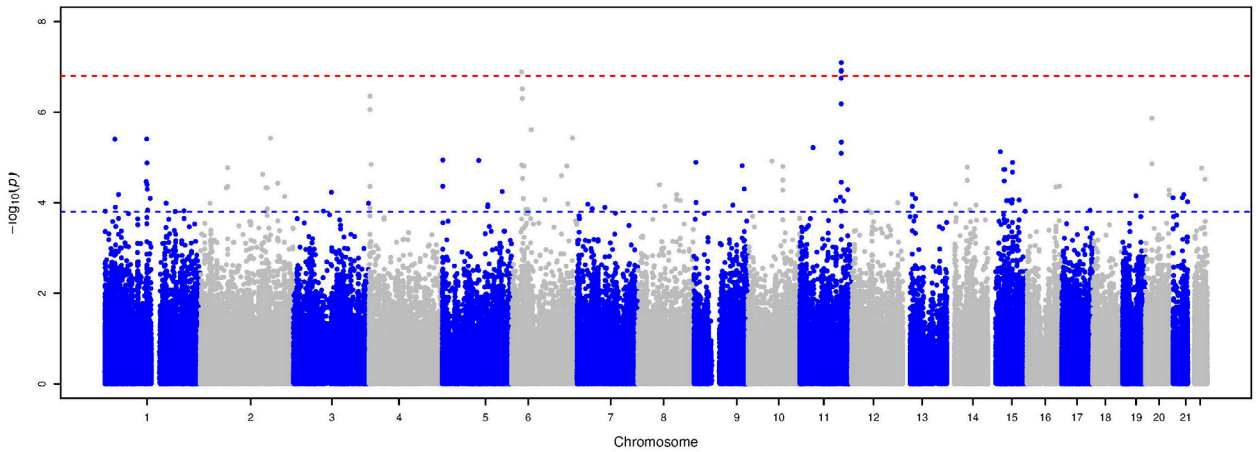


FIGURE 1 | Manhattan plot of the association analysis with LTL in the ERF study. This plot shows –log10 transformed *p*-values (*y*-axis) for all variants present in the association analysis according to their position on each chromosome (*x*-axis). The red dashed line represents the Bonferroni corrected *p*-value threshold for significance (p-value $< 1.42 \times 10^{-7}$), while the blue dashed line represents the *p*-value threshold for suggestive significance (p-value $\leq 1.42 \times 10^{-4}$).

pedigree data (**Figure 4**). Further, the genetic kinship estimates show that the other three individuals are also related within 3–4 generations. These 8 variants are in strong linkage disequilibrium (pairwise LD: r^2 between 0.93 and 1.00, D′ = 1) and show very similar *p*-values. The top variant rs185270276 is located in an intron of the *C11orf65* gene (MAF = 0.5%, β = −1.34, SE = 0.25, p-value = 7.99×10^{-8}). The next six variants significantly associated with LTL (MAF = 0.5%, β = −1.38, SE = 0.26, p-value = 1.21×10^{-7}) are located in the *ACAT1*, *NPAT*, *ATM*/*C11orf65*, *KDELC2*, and *EXPH5* genes. Two of these six variants are missense variants: rs79119325 (*NPAT*, PolyPhen = 1, CADD score = 18.1) and rs12146512 (*EXPH5*, PolyPhen = 0.624, CADD score = 4.5). The eighth significant variant, rs2234993, is located within an intron of *ATM* (MAF = 0.5%, β = −1.37, SE = 0.26, p-value = 1.25×10^{-7}). The ninth significant variant, rs144114619 (MAF = 0.2%, p-value = 1.29×10^{-7}), is a missense variant located on chromosome six in the *BTN3A1* gene, which is predicted to be damaging (PolyPhen = 1, CADD score = 12.2) and has the largest effect size (β = −2.18, SE = 0.41). There were six carriers of this variant in the ERF population. Interestingly, four of these carriers are also carriers of the rare variants in the chromosome 11q22.3 region (**Supplementary Figure 2**).

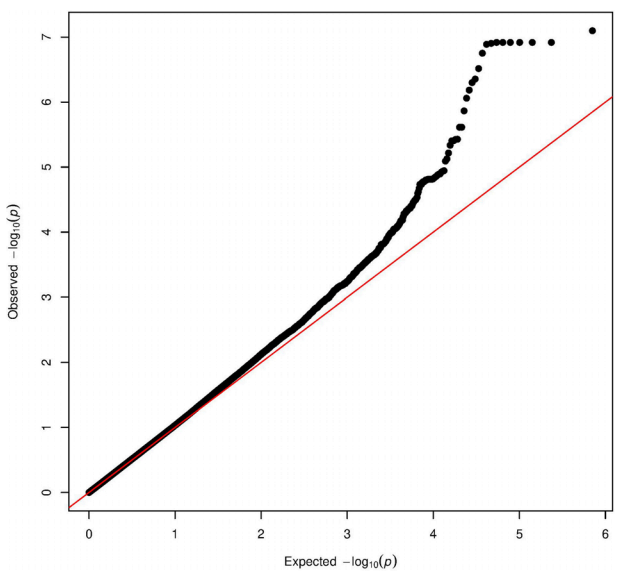


FIGURE 2 | Quantile-quantile plot of the association analysis with LTL in the ERF study. The QQ-plot shows the observed test statistics (*y*-axis) plotted against the expected values of the test statistics (*x*-axis) (X^2-distribution). The red line shows the distribution under the null hypothesis.

For replication analysis, we used WES data from two independent cohorts of European ancestry, the RS and the BHF-FHS. Results of the replication analysis are shown in **Table 3**, together with the results of the meta-analysis of summary statistics from all three cohorts for these variants. Six out of nine rare variants significantly associated with LTL in the ERF study (located in *ACAT1*, *NPAT*, *ATM/C11orf65*, *EXPH5*, and *BTN3A1*) were present in the RS and/or the BHF-FHS but were not significantly (p-value ≥ 0.21) associated with LTL (p-value < 0.025, 0.05/2 independent tests). The direction of effect of most variants was similar in the ERF study and the RS, while the BHF-FHS showed opposite direction of effect for most variants. Three variants, located in *C11orf65*, *KDELC2*, and *ATM*, were not present in the RS and BHF-FHS data. As these are unique to an isolated population, we were unable to confirm or reject their association with LTL in the replication cohorts.

Finally, to increase the statistical power, we performed an inverse-variance weighted meta-analysis of the association results from all three cohorts using METAL software. Variants were included in the meta-analysis if they were present in at least two out of three cohorts and had a minimum minor allele count of five in one or more cohorts, resulting in a multiple testing corrected significance threshold of 3.02×10^{-7} (0.05/165,311 variants). The top results of the meta-analysis (p-value $< 3.02 \times 10^{-4}$) are

TABLE 2 | Significant variants from the association analysis in the ERF study.

rsID	Gene	Chr	Position*	MAF	REF/ ALT	GVS function**	PolyPhen score**	PhastCons score**	CADD score**	β	SE	*p*-value
rs185270276	*C11orf65*	11	108263828	0.005	T/C	Intron	Unknown	0	6.51	−1.34	0.25	7.99×10^{-8}
rs12365364	*ACAT1*	11	108004687	0.005	G/A	Intron	Unknown	0	3.28	−1.38	0.26	1.21×10^{-7}
rs79119325	*NPAT*	11	108032614	0.005	C/T	Missense	1	0.998	18.13	−1.38	0.26	1.21×10^{-7}
rs3092910	*ATM,C11orf65*	11	108180917	0.005	T/C	Intron, synonymous	Unknown	0.997	9.87	−1.38	0.26	1.21×10^{-7}
rs3218711	*ATM,C11orf65*	11	108236264	0.005	C/G	3-prime-UTR, intron	Unknown	0.002	5.56	−1.38	0.26	1.21×10^{-7}
rs11212668	*KDELC2*	11	108352576	0.005	T/C	Intron	Unknown	0	1.44	−1.38	0.26	1.21×10^{-7}
rs12146512	*EXPH5*	11	108384666	0.005	T/C	Missense	0.624	0.011	4.47	−1.38	0.26	1.21×10^{-7}
rs2234993	*ATM*	11	108129599	0.005	C/G	Intron	Unknown	0	4.25	−1.37	0.26	1.25×10^{-7}
rs144114619	*BTN3A1*	6	26408145	0.002	T/A	Missense	1	0.002	12.24	−2.18	0.41	1.29×10^{-7}

*Chr = Chromosome; MAF = Minor Allele Frequency; REF = Reference allele; ALT = Alternative allele; CADD = Combined Annotation Dependent Depletion; β = effect of the minor allele; SE = Standard Error. *Position according to Hg19. **Seattleseq Annotation Database 138.*

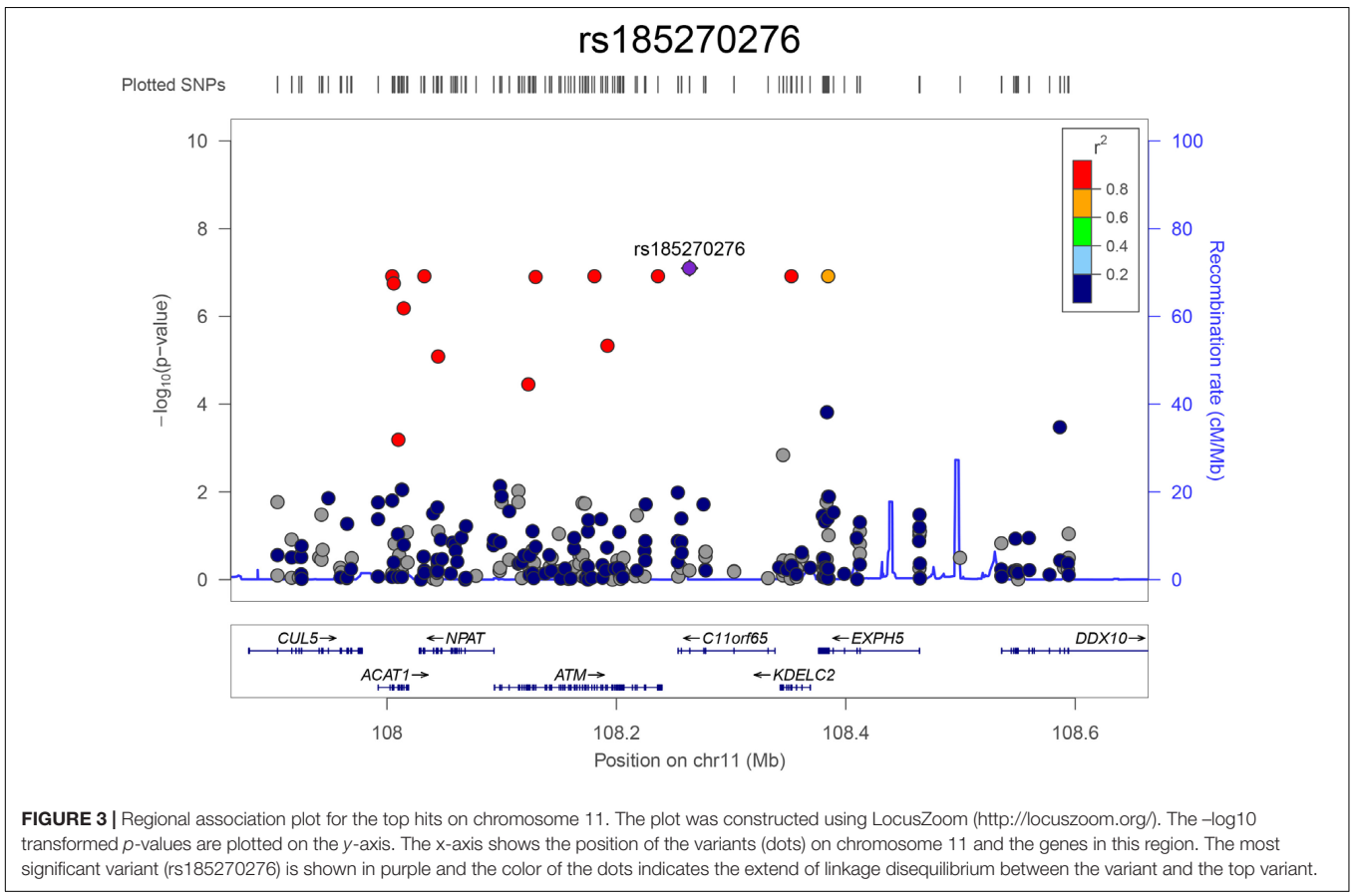


FIGURE 3 | Regional association plot for the top hits on chromosome 11. The plot was constructed using LocusZoom (http://locuszoom.org/). The –log10 transformed *p*-values are plotted on the *y*-axis. The x-axis shows the position of the variants (dots) on chromosome 11 and the genes in this region. The most significant variant (rs185270276) is shown in purple and the color of the dots indicates the extend of linkage disequilibrium between the variant and the top variant.

available in **Table 4** and in **Supplementary Table 1**, which also contains cohort specific information. The Manhattan and QQ-plots are provided in **Supplementary Figures 3, 4**, respectively. The λ of 0.97 suggests the power has been low. Although there were no variants genome-wide significantly associated with LTL in the meta-analysis after adjusting for multiple testing, many of the top findings show a consistent effect across cohorts. The variant most significantly associated with LTL was a highly conserved synonymous variant (PhastCons score = 0.999) located in the *RPL8* gene on chromosome 8 (p-value = 1.48×10^{-6}), which is predicted to be deleterious (CADD score = 15.11). Additionally, we used the meta-analysis results to perform a look-up of variants in loci identified by previous European ancestry GWASs (**Supplementary Table 2**). There were 153 variants present in these loci and we found suggestive evidence ($p\text{-value}_{meta} < 3.02 \times 10^{-4}$) for a positive association of a rare variant with LTL (rs181080831, β = 0.74, SE = 0.19, p-value = 1.18×10^{-4}) in the known *RTEL1* locus.

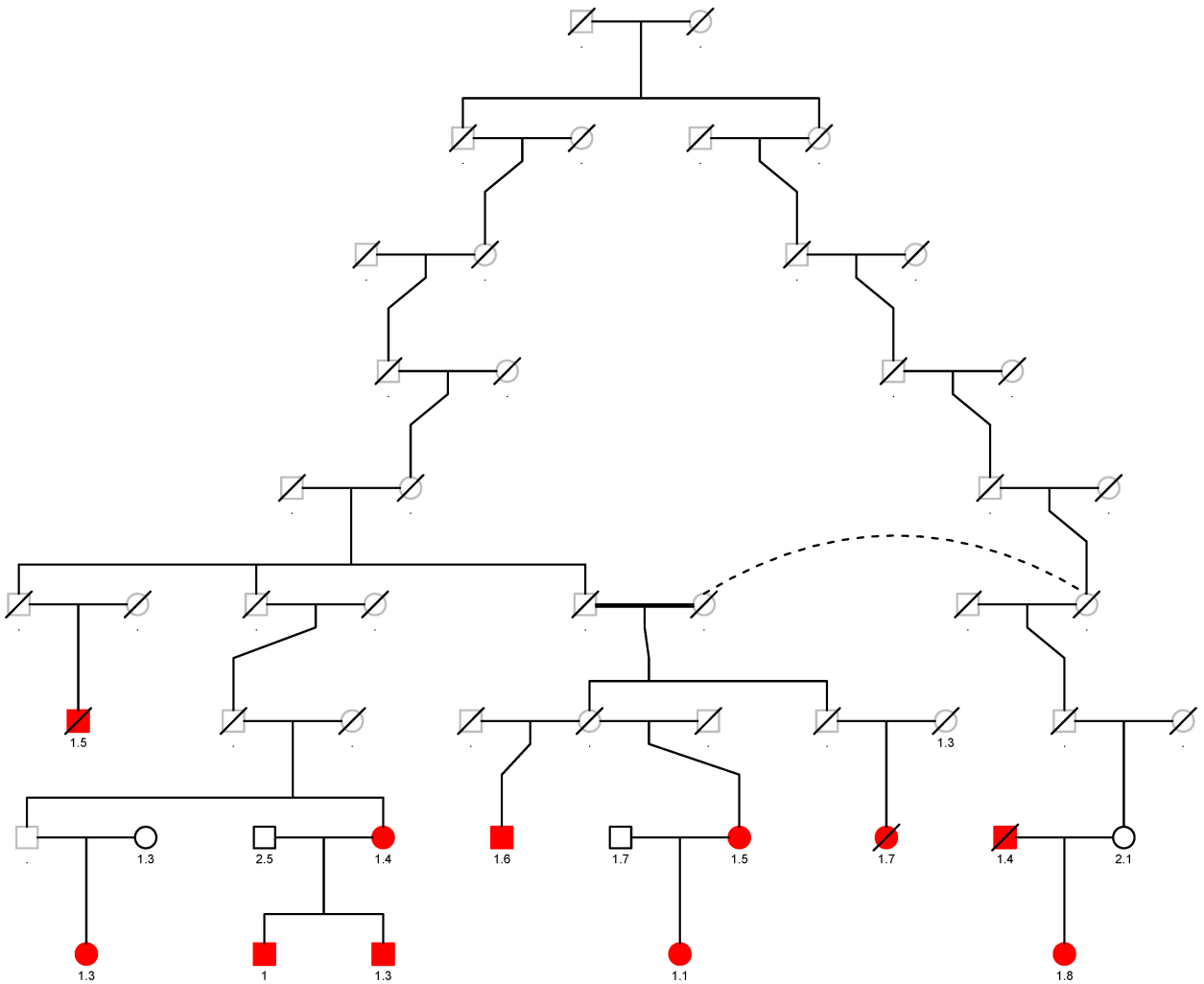


FIGURE 4 | Segregation plot of the rare variants on chromosome 11q22.3 in the ERF study. The carriers of the rare variants located on chromosome 11q22.3 are depicted in red. The T/S ratio is added below the individuals of whom whole-exome sequencing data was available. Squares are males and rounds are females. Deceased individuals are denoted by a line through that specific individual. One person is twice in the pedigree, this is shown with the dotted line.

DISCUSSION

In the family-based ERF study, we identified nine rare variants (MAF between 0.2% and 0.5%) associated with LTL by performing a WES association analysis. The eight most significantly associated variants are located in a region on chromosome 11q22.3 and segregate together with shorter LTL in one family. This region contains the *ATM* locus that has previously been shown to be involved in telomere maintenance and genomic stability and is thus an obvious candidate gene. In the meta-analysis of discovery and replication cohorts, we identified another rare missense variant in the *RPL8* gene strongly associated with LTL. Although we were not able to replicate either of the associations, both *ATM* and *RPL8* have been previously found to be strong predictors of telomere length (*ATM*) and chronological age (*ATM* and *RPL8*).

Interestingly, we identified three unique rare variants in the *ATM* (Ataxia Telangiectasia Mutated) gene associated with LTL in the ERF study. *ATM* is the homolog of the *Tel1* gene in yeast (Greenwell et al., 1995) and has been implicated in important telomere maintenance processes (Pandita, 2002; Lee et al., 2015; Tong et al., 2015). The ATM protein kinase is a master controller of cell cycle checkpoint signaling pathways required for cell response to DNA damage and for genomic stability (https://www.ncbi.nlm.nih.gov/gene/?term=472). Additionally, ATM kinase is necessary for telomere elongation (Lee et al., 2015; Tong et al., 2015). *ATM* is involved in the genetic disorder ataxia telangiectasia (AT), which is characterized by cerebellar ataxia, oculocutaneous telangiectasia, immunodeficiency, and a predisposition to cancer (Savitsky et al., 1995). Cells of AT patients also show telomeric fusions and have accelerated telomere shortening with increasing age (Metcalfe et al., 1996). Furthermore, *ATM* was significantly associated with

TABLE 3 | Replication results of the association analysis.

rsID	Gene	ERF (*N* = 1,303)				RS (*N* = 1,257)				BHF-FHS (*N* = 674)				Meta-analysis (*N* = 3,234)			
		MAF	β	SE	*p*-value	MAF	β	SE	*p*-value	MAF	β	SE	*p*-value	Direction*	β	SE	*p*-value
rs185270276	*C11orf65*	0.005	−1.34	0.25	7.99×10^{-8}	–	–	–	–	–	–	–	–	–	–	–	–
rs12365364	*ACAT1*	0.005	−1.38	0.26	1.21×10^{-7}	0.006	−0.20	0.25	0.43	0.010	0.30	0.26	0.25	--+	−0.40	0.15	7.84×10^{-3}
rs79119325	*NPAT*	0.005	−1.38	0.26	1.21×10^{-7}	0.006	−0.24	0.28	0.39	0.010	0.30	0.26	0.25	--+	−0.43	0.16	5.91×10^{-3}
rs3092910	*ATM,C11orf65*	0.005	−1.38	0.26	1.21×10^{-7}	0.007	−0.18	0.27	0.50	0.010	0.30	0.26	0.25	--+	−0.41	0.16	8.67×10^{-3}
rs3218711	*ATM,C11orf65*	0.005	−1.38	0.26	1.21×10^{-7}	0.005	−0.24	0.28	0.39	0.010	0.30	0.26	0.25	--+	−0.43	0.16	5.93×10^{-3}
rs11212668	*KDELC2*	0.005	−1.38	0.26	1.21×10^{-7}	–	–	–	–	–	–	–	–	–	–	–	–
rs12146512	*EXPH5*	0.005	−1.38	0.26	1.21×10^{-7}	–	–	–	–	0.016	0.27	0.21	0.21	-?+	−0.38	0.17	2.61×10^{-2}
rs2234993	*ATM*	0.005	−1.37	0.26	1.25×10^{-7}	–	–	–	–	–	–	–	–	–	–	–	–
rs144114619	*BTN3A1*	0.002	−2.18	0.41	1.29×10^{-7}	0.004	0.008	0.29	0.98	0.003	−0.43	0.48	0.38	-+-	−0.63	0.22	3.73×10^{-3}

*ERF = Erasmus Rucphen Family study; RS = Rotterdam Study; BHF-FHS = British Heart Foundation Family Heart Study; N = sample size; MAF = Minor Allele Frequency; β = Effect of the minor allele; SE = Standard Error. *Order of cohorts in direction column: ERF – RS – BHF-FHS; Direction of effect represented by – (negative association), +(positive association), or ? (not available).*

chronological age in a meta-analysis of gene expression profiles, showing lower transcript abundance in older individuals (Peters et al., 2015). Although genetic variants in *ATM* have been associated with various cancers (Barrett et al., 2011; Helgason et al., 2015; Ransohoff et al., 2017; Scelo et al., 2017), only one genetic variant in *ATM*, rs227080, was genome-wide significantly associated with LTL in a Singaporean Chinese population (Dorajoo et al., 2019). This variant was not significantly associated with LTL in the ERF study, the Rotterdam Study or the BHF-FHS, implicating it was not driving the association observed in the current study.

Segregation analysis of the variants within the chromosome 11q22.3 region in the ERF study showed that the variants segregate with shorter LTL in one family. We did not detect a specific disease that segregates in this family, which may be explained by the relatively young age of most of the carriers. In the replication analysis, we were not able to confirm the association of the nine variants associated with LTL in the RS and BHF-FHS, due to lack of association (six variants) or absence of the variant in the replication cohorts (three variants). However, it is possible this signal comes from variants that are family-specific and thus may not be transferable to the general population. Likewise, it may also be that the whole haplotype has an effect on telomere length, an haplotype that is most likely unique to the isolated ERF population. At this locus, the *ATM* gene is currently the most likely causative gene. Lastly, we performed a meta-analysis of the three cohorts. The top variant of the meta-analysis is located in *RPL8* (Ribosomal Protein L8). This is interesting as *RPL8*, together with six other *RPL* genes, was negatively associated with chronological age (Peters et al., 2015). In this transcriptomic study, Peters et al. identified 1,497 genes whose expression level changes associated with age; this gene list includes *RPL8* and *ATM*. We additionally found suggestive evidence for an association of a novel rare variant in the known *RTEL1* locus. It would be interesting to perform a WES or whole-genome sequencing meta-analysis with larger sample size to increase the statistical power.

The advantage of our study is that we used a family-based setting that allowed us to show segregation of the variants located on chromosome 11q22.3 in a family. The ERF study population has shown a low immigration rate and a high level of inbreeding, which has increased the frequency of many rare alleles (Pardo et al., 2005). Another advantage is that all three studies used the standardized qPCR method to quantify LTL. However, there are also several limitations to our study. The first limitation is that our findings in the ERF study are not easy to generalize to the general population as these findings may be family-specific. However, we identified genes that are known to be related to telomere processes and thus are plausible candidate genes. The second limitation of our study may be that we used blood measurements of telomere length instead of tissue specific measurements, although previous studies have shown that mean telomere length in blood and other tissues are highly correlated (Okuda et al., 2002; Wilson et al., 2008) and, therefore, we expect this did not influence our findings. The third limitation is the difference in LTL distribution between the three studies, which may be explained by the different age distributions in the studies.

TABLE 4 | Suggestive findings of the meta-analysis (p-value $\leq 3.02 \times 10^{-4}$).

rsID	Gene	Chr	Position*	REF/ ALT	GVS function**	PolyPhen score**	PhastCons score**	CADD score**	β	SE	p-value	Direction†
8:146017422	*RPL8*	8	146017422	G/A	Synonymous	Unknown	0.999	15.11	1.93	0.40	1.48×10^{-6}	+++
rs77919685	*LATS2*	13	21563311	G/A	Missense	0.002	0	7.001	0.44	0.09	2.23×10^{-6}	+++
rs56041036	*ZFPM1*	16	88599023	A/G	Synonymous-near-splice	Unknown	0.998	10.77	−0.20	0.04	7.44×10^{-6}	---
rs199779997	*MGA*	15	42058958	A/C	Missense	0.001	0.866	9.45	1.77	0.40	8.60×10^{-6}	+++
rs4895944	*VNN2*	6	133070995	G/T	Missense, non-coding-exon	0.038	0.003	16.69	−1.29	0.29	8.99×10^{-6}	--?
rs7735563	*RAPGEF6*	5	130764936	T/C	Intron	Unknown	1	19.08	−0.77	0.18	1.53×10^{-5}	---
rs189691392	*SF3B5*	6	144416667	G/A	5-prime-UTR	Unknown	0	7.119	−1.82	0.42	1.77×10^{-5}	-?-
rs138765444	*ZKSCAN4*	6	28219377	A/G	Synonymous, intron	Unknown	0.036	9.324	−1.27	0.30	2.59×10^{-5}	---
rs1783091	*none*	21	33964605	T/C	Intergenic	Unknown	0.004	0.154	0.22	0.05	2.67×10^{-5}	+?+
rs2170177	*DOCK3*	3	51349887	C/A	Intron	Unknown	0	0.044	−0.36	0.09	5.06×10^{-5}	---
rs5930	*LDLR*	19	11224265	A/G	Synonymous	Unknown	0.001	0.579	−0.10	0.03	5.73×10^{-5}	---
rs55648406	*TUB*	11	8060566	G/A	Missense, intron	0.086	1	15.76	−0.62	0.16	5.85×10^{-5}	---
rs140456008	*SLC35G2*	3	136574420	A/G	Missense	0.941	1	10.84	1.89	0.47	6.67×10^{-5}	++?
rs139380413	*COL8A1*	3	99513830	G/A	Missense	0.071	0.966	11.23	−1.26	0.32	7.88×10^{-5}	---
rs11656725	*LRRC48*	17	17900726	C/T	Intron	Unknown	0	0.471	0.38	0.10	8.25×10^{-5}	+?+
rs7193541	*RFWD3*	16	74664743	T/C	Missense	0.008	0.485	9.1	−0.10	0.03	8.28×10^{-5}	---
rs17222435	*SLC28A1*	15	85488335	C/T	Intron	Unknown	0	3.598	0.69	0.18	9.07×10^{-5}	+++
rs117223521	*none*	8	38964715	T/C	Upstream-gene	Unknown	0	1.194	−1.36	0.35	1.01×10^{-4}	-?-
rs11700220	*MTG2*	20	60770931	A/G	Missense	1	1	21.6	0.47	0.12	1.03×10^{-4}	+++
rs187466877	*GPN1*	2	27862872	A/G	Intron	Unknown	0	1.47	0.85	0.22	1.05×10^{-4}	+?+
rs56188826	*MARK1*	1	220791870	C/T	Synonymous	Unknown	0.123	6.414	0.41	0.11	1.14×10^{-4}	+?+
rs1872592	*PIF1*	15	65113493	G/A	Intron	Unknown	0	0.005	−0.10	0.02	1.15×10^{-4}	---
rs73056605	*CLEC4C*	12	7894056	G/A	Missense	0.037	0	0.005	0.11	0.03	1.18×10^{-4}	+++
rs181080831	*RTEL1*, *RTEL1-TNFRSF6B*	20	62326874	G/A	Intron, non-coding-exon, synonymous	Unknown	0	4.04	0.74	0.19	1.18×10^{-4}	+++
rs13014800	*CENPA*	2	27015118	G/A	Intron	Unknown	0.163	11.67	−0.13	0.03	1.23×10^{-4}	-?-
rs143463783	*TRIM27*	6	28889741	G/A	Synonymous	Unknown	1	9.216	−1.22	0.32	1.24×10^{-4}	-?-
rs374215951	*PIEZO1*	16	88788318	G/A	Synonymous	Unknown	0.21	0.893	−2.57	0.67	1.31×10^{-4}	-?-
rs181215404	*EPPK1*	8	144941659	G/A	Synonymous	Unknown	0.011	4.835	0.94	0.25	1.32×10^{-4}	+?+
rs377359525	*FTCD*	21	47572869	A/G	Missense	1	1	16.44	1.68	0.44	1.33×10^{-4}	+++
rs10936599	*MYNN*	3	169492101	C/T	Synonymous, non-coding-exon, 5-prime-UTR	Unknown	1	10.1	−0.11	0.03	1.38×10^{-4}	---
rs74730846	*STXBP5L*	3	120924764	C/T	Intron-near-splice	Unknown	0.629	5.818	−0.19	0.05	1.39×10^{-4}	-?-
rs137853096	*HSD17B4*	5	118788316	G/A	Missense, 5-prime-UTR	1	1	24	−0.73	0.19	1.67×10^{-4}	--?
rs41284136	*IFIT3*	10	91087805	G/C	5-prime-UTR	Unknown	0	7.796	0.40	0.11	1.76×10^{-4}	+++
rs58106741	*SLC4A1AP*	2	27886820	G/T	Synonymous	Unknown	0	5.939	0.70	0.19	1.80×10^{-4}	+?+
rs58068845	*UTP6*	17	30200363	G/A	Intron	Unknown	0	4.223	0.35	0.09	1.90×10^{-4}	+++
rs200602887	*GREB1*	2	11751072	G/C	Synonymous	Unknown	0.986	10.5	−1.02	0.27	2.01×10^{-4}	---
rs141180155	*LRP2*	2	170127559	G/A	Synonymous, intron	Unknown	0	13.06	−0.36	0.10	2.08×10^{-4}	---
rs7837242	*LONRF1*	8	12600622	C/T	Intron	Unknown	0.001	4.941	−0.18	0.05	2.08×10^{-4}	-?-
rs115018606	*C2orf16*	2	27799773	A/C	Missense	0.972	0.002	5.869	0.41	0.11	2.12×10^{-4}	+?+
rs151309008	*REXO2*	11	114310345	C/T	Missense	0.437	1	17.76	−0.64	0.17	2.12×10^{-4}	---
rs143759519	*PYGL*	14	51382637	G/A	Missense	1	0.975	34	0.55	0.15	2.13×10^{-4}	+++
rs10936600	*LRRC34*	3	169514585	A/T	Missense	1	0.001	12.07	−0.11	0.03	2.14×10^{-4}	---
rs117178504	*DYNC2H1*	11	103153788	C/A	Synonymous	Unknown	0.996	8.615	0.38	0.10	2.18×10^{-4}	+++
rs367644268	*COA5*	2	99224742	C/T	Intron	Unknown	0.015	8.626	−1.42	0.39	2.28×10^{-4}	--?
rs146979490	*GPN1*	2	27864089	A/G	Intron	Unknown	0.025	11.52	0.51	0.14	2.43×10^{-4}	+++
rs146033252	*MTA3*, *OXER1*	2	42991127	G/A	Missense, intron	0.084	0.023	8.404	0.71	0.19	2.48×10^{-4}	+++

(Continued)

TABLE 4 | Continued

rsID	Gene	Chr	Position*	REF/ ALT	GVS function**	PolyPhen score**	PhastCons score**	CADD score**	β	SE	*p*-value	Direction†
rs141280036	*PADI4*	1	17634718	A/G	Missense	0.992	0.881	13.72	0.79	0.22	2.51×10^{-4}	+++
rs79400176	*DZANK1*	20	18414309	C/T	Missense	0.129	0.994	2.801	−0.36	0.10	2.55×10^{-4}	---
rs116604207	*RBM5*	3	50147061	G/A	Synonymous, non-coding-exon	Unknown	0.453	11.12	0.79	0.22	2.62×10^{-4}	+++
rs41307740	*LBR*	1	225601614	C/A	Intron	Unknown	0.004	5.956	−0.65	0.18	2.76×10^{-4}	-?-
rs369623673	*ELAC2*	17	12899160	C/T	Intron	Unknown	0	6	0.93	0.26	2.77×10^{-4}	+?+
rs9997727	*C4orf50*	4	5969113	G/A	Intron	Unknown	0	0.613	−0.10	0.03	2.85×10^{-4}	---
rs150538926	*PDZD2*	5	32037369	C/T	Synonymous	Unknown	0	0.482	0.40	0.11	2.87×10^{-4}	+++
rs55868421	*TRIM5*	11	5688948	A/G	Intron-near-splice, intron	Unknown	0	5.512	0.44	0.12	2.94×10^{-4}	+++
rs323895	*ACY1*, *ABHD14A-ACY1*	3	52021316	A/G	Intron	Unknown	0	1.818	−0.36	0.10	3.00×10^{-4}	---
rs7188880	*RFWD3*	16	74664810	A/T	Synonymous	Unknown	1	10.78	−0.09	0.02	3.01×10^{-4}	---

*Chr = Chromosome; REF = Reference allele; ALT = Alternative allele; CADD = Combined Annotation Dependent Depletion; β = effect of the minor allele; SE = Standard Error. *Position according to Hg19. **Seattleseq Annotation Database 138. †Order of cohorts in direction column: ERF, RS, BHF-FHS; Direction of effect represented by − (negative association) + (positive association) or ? (not available).*

Mean age is lowest in the ERF study (49 years), highest in the RS (75 years) and mean age in the BHF-FHS is in between (58 years). As LTL decreases significantly with age (Lindsey et al., 1991; Slagboom et al., 1994; Blackburn, 2001) and is associated with mortality (Cawthon et al., 2003; Kimura et al., 2008; Fitzpatrick et al., 2011; Honig et al., 2012; Deelen et al., 2014; Rode et al., 2015; Marioni et al., 2016), the variation in LTL becomes less with the aging populations. Mean LTL values were comparable between studies after adjusting LTL values for age and sex. To further standardize across the three cohorts, a z-transformation of the LTL values was performed resulting in comparable distributions with mean of zero and standard deviation of one. Nevertheless, these differences between studies, together with the small sample size of the BHF-FHS cohort, could potentially explain the lack of replication. The fourth limitation is that we were unable to calculate the effect of the variants on telomere length in base pairs because of the quantification method of LTL in our study together with the z-transformation that was applied to the LTL values to standardize across the cohorts.

To conclude, this first study using WES data to search for rare genetic variants associated with LTL has identified interesting variants and genes associated with shorter LTL. Eight out of nine rare variants associated with LTL are located on chromosome 11q22.3 and all variants segregate within an ERF family. As we were not able to replicate findings, future studies should further investigate this region and the other genes identified in this study to confirm their involvement in telomere length regulation.

MATERIALS AND METHODS

Study Populations

Our discovery population consisted of participants from the family-based Erasmus Rucphen Family (ERF) study. The ERF study comprises approximately 3,000 inhabitants of a recent genetically isolated community in the Southwest of the Netherlands, studied as part of the Genetic Research in Isolated Population program (Pardo et al., 2005). All ERF participants are descendants of 22 founder couples who had at least six children baptized in the community church in the 18th century, or their spouses. Baseline data collection, including blood withdrawal, took place between 2002 and 2005. As the ERF population shows a low rate of immigration and a high level of inbreeding, the frequency of several rare alleles is increased in this population (Pardo et al., 2005). The ERF study was approved by the Medical Ethics Commitee of the Erasmus Medical Center (MC), Rotterdam, Netherlands. All participants provided written informed consents and all investigations were carried out in accordance with the Declaration of Helsinki.

The replication cohorts included participants from the Rotterdam Study (RS) and the British Heart Foundation Family Heart Study (BHF-FHS). The RS is a prospective cohort study ongoing since 1990 in the well-defined Ommoord district in Rotterdam, The Netherlands. The original RS cohort (RS-I) included 7,983 individuals of 55 years of age or over. At baseline, participants were interviewed at home and had an extensive set of examinations, which were repeated every 3–4 years (Ikram et al., 2017). The Rotterdam Study has been approved by the Medical Ethics Committee of the Erasmus MC and by the Dutch Ministry of Health, Welfare and Sport. The Rotterdam Study has been entered into the Netherlands National Trial Register (NTR; www.trialregister.nl) and into the WHO International Clinical Trials Registry Platform (ICTRP; www.who.int/ictrp/network/primary/en/) under shared catalogue number NTR6831. All participants provided written informed consent to participate in the study and to have their information obtained from treating physicians.

The British Heart Foundation Family Heart study recruited families with at least two siblings diagnosed with premature (<66 years) coronary artery disease (CAD) within the

United Kingdom between 1998 and 2003. Full details are provided elsewhere (Samani et al., 2005, 2007).

Telomere Length

Mean LTL values in the ERF study and the BHF-FHS study were measured using a qPCR method in all samples (Cawthon, 2002). The measurements were performed in Leicester, United Kingdom, and details of the measurements were previously described (Codd et al., 2010, 2013). In summary, mean LTL was measured in leukocytes and expressed as the ratio (T/S ratio) of telomere repeat length (T) to the copy number of a single-copy gene, *36B4* (S). Samples were quantified relative to a calibrator sample used on each run (DNA from the K562 cell line) (Codd et al., 2010).

In the RS, mean LTL values were also measured using a qPCR assay based on the method described by Cawthon (2002) with minor modifications. For each sample the telomere and *36B4* assay were run in separate wells but in the same 384 wells PCR plate. Each reaction contained 5 ng DNA, 1 uM of each of the telomere primers (tel1b-forward: GGTT TGTTTGGGTTTGGGTTTGGGTTTGGGTTTGGGTT, tel2b-reverse: GGCTTGCCTTACCCTTACCCTTACCCTTACCCTTA CCCT) or 250 nM of the *36B4* primers (36B4u-forward: CAG CAAGTGGGAAGGTGTAATCC, 36B4d-reverse: CCCATTCT ATCATCAACGGGTACAA) and 1x Quantifast SYBR green PCR Mastermix (Qiagen). The reactions for both assays were performed in duplicate for each sample in a 7900HT machine (Applied Biosystems). Ct values and PCR efficiencies were calculated per plate using the MINER algorithm (Zhao and Fernald, 2005). Duplicate Ct values that had a Coefficient of Variance (CV) of more than 1% were excluded from further analysis. Using the average Ct value per sample and the average PCR efficiency per plate the samples were quantified using the formula Q = 1/(1 + PCR eff)^Ct. The relative telomere length was calculated by dividing the Q of the telomere assay by the Q of the *36B4* assay. To validate the assay 96 random samples were run twice and the CV of that experiment was 4.5%.

Exome Sequencing

The exomes of 1,336 ERF participants were sequenced at the Erasmus Center for Biomics of the department of Cell Biology, Erasmus MC, The Netherlands. The exomes of a randomly selected subset of 2,628 individuals from the RS-I population were sequenced at the Human Genotyping facility of the Department of Internal Medicine, Erasmus MC, The Netherlands. Details of the methods and quality control for ERF and RS are described elsewhere (Amin et al., 2016; van Rooij et al., 2017). In total, 1,303 ERF participants and 1,257 RS participants had both exome sequence and LTL data available and were included in this analysis. A subset of the BHF-FHS, comprising of 674 unrelated individuals of Caucasian ancestry who had previously undergone exome sequencing as part of the Leicester Myocardial Infarction Study (Khera et al., 2017) and had LTL data available (Codd et al., 2010) were included in this analysis.

Statistical Analyses

For each individual cohort quantitative trait association analysis was performed using Rare Variant tests (RVtests) software, which supports the analysis of related individuals (Zhan et al., 2016). Association analysis was performed using a score test, assuming an additive model, suitable for analysis with related and unrelated individuals. We applied a z-transformation of LTL values for the three cohorts separately to standardize values across cohorts. All analyses were adjusted for age, sex and batch effects (if needed). Furthermore, we adjusted for familial relationships in ERF using the kinship matrix estimated from the genotyped data, while in the RS we corrected for the first four principal components as the fourth principal component was significantly associated with LTL. Only variants with a minor allele count ≥ 5 were included.

In the ERF study, we calculated the pairwise LD (r^2 and D′) between the top eight variants on chromosome 11 that were significantly associated with LTL using the –ld command of PLINK 1.9 software (Chang et al., 2015) (www.cog-genomics.org/plink/1.9/). Additionally, we performed an inverse-variance weighted meta-analysis using METAL software (Willer et al., 2010). In the analysis using data of the family-based ERF study, we corrected the significance thresholds for multiple testing using Bonferroni correction, resulting in a significance threshold of 1.42×10^{-7} (0.05/353,075). In the replication analysis, the multiple testing corrected *p*-value thresholds was 0.025 (0.05/2 independent tests). In the meta-analysis, we adjusted for the number of variants tested, resulting in a significance threshold of 3.02×10^{-7} (0.05/165,311).

DATA AVAILABILITY STATEMENT

The datasets analyzed for each individual cohort can be requested by contacting the responsible Principal Investigator. Because of restrictions based on privacy regulations and informed consent of the participants, data cannot be made freely available in a public repository. For the Rotterdam Study data, requests should be directed toward the management team of the Rotterdam Study (secretariat.epi@erasmusmc.nl), which has a protocol for approving data requests.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by their respective Medical Ethics Boards and all investigations were carried out in accordance with the Declaration of Helsinki. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

AS, NA, and CD designed the study and wrote the manuscript. AS, SW, CN, and SA performed the analyses. LB, PA, RB, MD, MH, JR, RK, and NA were involved in the data collection and

provided the data. WI, NS, MI, AU, and CD were involved in the supervision of individual cohorts. AS, SW, CN, DV, NS, VC, NA, and CD interpreted the results. All authors critically reviewed and approved the manuscript.

FUNDING

The ERF study as a part of EUROSPAN (European Special Populations Research Network) was supported by European Commission FP6 STRP grant number 018947 (LSHG-CT-2006-01947) and also received funding from the European Community's Seventh Framework Programme (FP7/2007-2013)/grant agreement HEALTH-F4-2007-201413 by the European Commission under the programme "Quality of Life and Management of the Living Resources" of 5th Framework Programme (No. QLG2-CT-2002-01254). High-throughput analysis of the ERF data was supported by joint grant from Netherlands Organization for Scientific Research and the Russian Foundation for Basic Research (NWO-RFBR 047.017.043). Exome sequencing analysis in ERF was supported by the ZonMw grant (project 91111025). AS and CD have used exchange grants from Personalized pREvention of Chronic DIseases consortium (PRECeDI) (H2020-MSCA-RISE-2014). The generation and management of the exome sequencing data for the Rotterdam Study was executed by the Human Genotyping Facility of the Genetic Laboratory of the Department of Internal Medicine, Erasmus MC, the Netherlands. The Exome Sequencing data set was funded by the Netherlands Genomics Initiative (NGI)/Netherlands Organisation for Scientific Research (NWO) sponsored Netherlands Consortium for Healthy Aging (NCHA; project nr. 050-060-810), by the Genetic Laboratory of the Department of Internal Medicine, Erasmus MC, and by the and by a Complementation Project of the Biobanking and Biomolecular Research Infrastructure Netherlands (BBMRI-NL; www.bbmri.nl; project number CP2010-41). The Rotterdam Study was funded by Erasmus Medical Center and Erasmus University, Rotterdam, Netherlands Organization for the Health Research and Development (ZonMw), the Research Institute for Diseases in the Elderly (RIDE), the Ministry of Education, Culture and Science, the Ministry for Health, Welfare and Sports, the European Commission (DG XII), and the Municipality of Rotterdam. The BHF-FHS study was funded by the British Heart Foundation (BHF). Exome sequencing of the Leicester Myocardial Infarction Study was supported by grant 5U54HG003067 from the National Institutes of Health. CN and NS were supported by the BHF. NS is a NIHR Senior Investigator.

ACKNOWLEDGMENTS

ERF study: we are grateful to all study participants and their relatives, general practitioners, and neurologists for their contributions and to P. Veraart for her help in genealogy, J. Vergeer for the supervision of the laboratory work and P. Snijders for his help in data collection. Rotterdam Study: we thank Mr. Pascal Arp, Ms. Mila Jhamai, and Mr. Marijn Verkerk for their help in creating the RS-Exome Sequencing database. We are grateful to the study participants, the staff from the Rotterdam Study and the participating general practitioners and pharmacists.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.00337/full#supplementary-material

REFERENCES

Allsopp, R. C., Vaziri, H., Patterson, C., Goldstein, S., Younglai, E. V., Futcher, A. B., et al. (1992). Telomere length predicts replicative capacity of human fibroblasts. *Proc. Natl. Acad. Sci. U.S.A.* 89, 10114–10118. doi: 10.1073/pnas.89.21.10114

Amin, N., Jovanova, O., Adams, H. H., Dehghan, A., Kavousi, M., Vernooij, M. W., et al. (2016). Exome-sequencing in a large population-based study reveals a rare Asn396Ser variant in the LIPG gene associated with depressive symptoms. *Mol. Psychiatry* 22, 537–543. doi: 10.1038/mp.2016.101

Andrew, T., Aviv, A., Falchi, M., Surdulescu, G. L., Gardner, J. P., Lu, X., et al. (2006). Mapping genetic loci that determine leukocyte telomere length in a large sample of unselected female sibling pairs. *Am. J. Hum. Genet.* 78, 480–486. doi: 10.1086/500052

Barrett, J. H., Iles, M. M., Harland, M., Taylor, J. C., Aitken, J. F., Andresen, P. A., et al. (2011). Genome-wide association study identifies three new melanoma susceptibility loci. *Nat. Genet.* 43, 1108–1113.

Benetos, A., Gardner, J. P., Zureik, M., Labat, C., Xiaobin, L., Adamopoulos, C., et al. (2004). Short telomeres are associated with increased carotid atherosclerosis in hypertensive subjects. *Hypertension* 43, 182–185. doi: 10.1161/01.hyp.0000113081.42868.f4

Bischoff, C., Graakjaer, J., Petersen, H. C., Hjelmborg, J., Vaupel, J. W., Bohr, V., et al. (2005). The heritability of telomere length among the elderly and oldest-old. *Twin Res. Hum. Genet.* 8, 433–439. doi: 10.1375/twin.8.5.433

Blackburn, E. H. (2000). Telomere states and cell fates. *Nature* 408, 53–56. doi: 10.1038/35040500

Blackburn, E. H. (2001). Switching and signaling at the telomere. *Cell* 106, 661–673. doi: 10.1016/s0092-8674(01)00492-5

Blackburn, E. H., and Gall, J. G. (1978). A tandemly repeated sequence at the termini of the extrachromosomal ribosomal RNA genes in Tetrahymena. *J. Mol. Biol.* 120, 33–53. doi: 10.1016/0022-2836(78)90294-2

Bojesen, S. E., Pooley, K. A., Johnatty, S. E., Beesley, J., Michailidou, K., Tyrer, J. P., et al. (2013). Multiple independent variants at the TERT locus are associated with telomere length and risks of breast and ovarian cancer. *Nat. Genet.* 45, 371–384.

Broer, L., Codd, V., Nyholt, D. R., Deelen, J., Mangino, M., Willemsen, G., et al. (2013). Meta-analysis of telomere length in 19,713 subjects reveals high heritability, stronger maternal inheritance and a paternal age effect. *Eur. J. Hum. Genet.* 21, 1163–1168. doi: 10.1038/ejhg.2012.303

Brouilette, S., Singh, R. K., Thompson, J. R., Goodall, A. H., and Samani, N. J. (2003). White cell telomere length and risk of premature myocardial infarction. *Arterioscler. Thromb. Vasc. Biol.* 23, 842–846. doi: 10.1161/01.atv.0000067426.96344.32

Brouilette, S. W., Moore, J. S., Mcmahon, A. D., Thompson, J. R., Ford, I., Shepherd, J., et al. (2007). Telomere length, risk of coronary heart disease, and statin treatment in the West of Scotland Primary Prevention Study: a nested case-control study. *Lancet* 369, 107–114. doi: 10.1016/s0140-6736(07)60071-3

Cawthon, R. M. (2002). Telomere measurement by quantitative PCR. *Nucleic Acids Res.* 30:e47.

Cawthon, R. M., Smith, K. R., O'brien, E., Sivatchenko, A., and Kerber, R. A. (2003). Association between telomere length in blood and mortality in people aged 60 years or older. *Lancet* 361, 393–395. doi: 10.1016/s0140-6736(03)12384-7

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.

Codd, V., Mangino, M., Van Der Harst, P., Braund, P. S., Kaiser, M., Beveridge, A. J., et al. (2010). Common variants near TERC are associated with mean telomere length. *Nat. Genet.* 42, 197–199. doi: 10.1038/ng.532

Codd, V., Nelson, C. P., Albrecht, E., Mangino, M., Deelen, J., Buxton, J. L., et al. (2013). Identification of seven loci affecting mean telomere length and their association with disease. *Nat. Genet.* 45, 422–427.

Deelen, J., Beekman, M., Codd, V., Trompet, S., Broer, L., Hagg, S., et al. (2014). Leukocyte telomere length associates with prospective mortality independent of immune-related parameters and known genetic markers. *Int. J. Epidemiol.* 43, 878–886. doi: 10.1093/ije/dyt267

Dorajoo, R., Chang, X., Gurung, R. L., Li, Z., Wang, L., Wang, R., et al. (2019). Loci for human leukocyte telomere length in the Singaporean Chinese population and trans-ethnic genetic studies. *Nat. Commun.* 10:2491.

Fitzpatrick, A. L., Kronmal, R. A., Gardner, J. P., Psaty, B. M., Jenny, N. S., Tracy, R. P., et al. (2007). Leukocyte telomere length and cardiovascular disease in the cardiovascular health study. *Am. J. Epidemiol.* 165, 14–21. doi: 10.1093/aje/kwj346

Fitzpatrick, A. L., Kronmal, R. A., Kimura, M., Gardner, J. P., Psaty, B. M., Jenny, N. S., et al. (2011). Leukocyte telomere length and mortality in the cardiovascular health study. *J. Gerontol. A Biol. Sci. Med. Sci.* 66, 421–429.

Greenwell, P. W., Kronmal, S. L., Porter, S. E., Gassenhuber, J., Obermaier, B., and Petes, T. D. (1995). TEL1, a gene involved in controlling telomere length in *S. cerevisiae*, is homologous to the human ataxia telangiectasia gene. *Cell* 82, 823–829. doi: 10.1016/0092-8674(95)90479-4

Grodstein, F., Van Oijen, M., Irizarry, M. C., Rosas, H. D., Hyman, B. T., Growdon, J. H., et al. (2008). Shorter telomeres may mark early risk of dementia: preliminary analysis of 62 participants from the nurses' health study. *PLoS One* 3:e1590. doi: 10.1371/journal.pone.0001590

Harris, S. E., Deary, I. J., Macintyre, A., Lamb, K. J., Radhakrishnan, K., Starr, J. M., et al. (2006). The association between telomere length, physical health, cognitive ageing, and mortality in non-demented older people. *Neurosci. Lett.* 406, 260–264. doi: 10.1016/j.neulet.2006.07.055

Hastie, N. D., Dempster, M., Dunlop, M. G., Thompson, A. M., Green, D. K., and Allshire, R. C. (1990). Telomere reduction in human colorectal carcinoma and with ageing. *Nature* 346, 866–868. doi: 10.1038/346866a0

Haycock, P. C., Heydon, E. E., Kaptoge, S., Butterworth, A. S., Thompson, A., and Willeit, P. (2014). Leucocyte telomere length and risk of cardiovascular disease: systematic review and meta-analysis. *BMJ* 349:g4227. doi: 10.1136/bmj.g4227

Helgason, H., Rafnar, T., Olafsdottir, H. S., Jonasson, J. G., Sigurdsson, A., Stacey, S. N., et al. (2015). Loss-of-function variants in ATM confer risk of gastric cancer. *Nat. Genet.* 47, 906–910. doi: 10.1038/ng.3342

Honig, L. S., Kang, M. S., Schupf, N., Lee, J. H., and Mayeux, R. (2012). Association of shorter leukocyte telomere repeat length with dementia and mortality. *Arch. Neurol.* 69, 1332–1339.

Houben, J. M., Giltay, E. J., Rius-Ottenheim, N., Hageman, G. J., and Kromhout, D. (2011). Telomere length and mortality in elderly men: the zutphen elderly study. *J. Gerontol. A Biol. Sci. Med. Sci.* 66, 38–44. doi: 10.1093/gerona/glq164

Ikram, M. A., Brusselle, G. G. O., Murad, S. D., Van Duijn, C. M., Franco, O. H., Goedegebure, A., et al. (2017). The Rotterdam Study: 2018 update on objectives, design and main results. *Eur. J. Epidemiol.* 32, 807–850. doi: 10.1007/s10654-017-0321-4

Khera, A. V., Won, H. H., Peloso, G. M., O'dushlaine, C., Liu, D., Stitziel, N. O., et al. (2017). Association of rare and common variation in the lipoprotein lipase gene with coronary artery disease. *JAMA* 317, 937–946.

Kimura, M., Hjelmborg, J. V., Gardner, J. P., Bathum, L., Brimacombe, M., Lu, X., et al. (2008). Telomere length and mortality: a study of leukocytes in elderly Danish twins. *Am. J. Epidemiol.* 167, 799–806. doi: 10.1093/aje/kwm380

Lee, S. S., Bohrson, C., Pike, A. M., Wheelan, S. J., and Greider, C. W. (2015). ATM kinase is required for telomere elongation in mouse and human cells. *Cell Rep.* 13, 1623–1632. doi: 10.1016/j.celrep.2015.10.035

Levy, D., Neuhausen, S. L., Hunt, S. C., Kimura, M., Hwang, S. J., Chen, W., et al. (2010). Genome-wide association identifies OBFC1 as a locus involved in human leukocyte telomere biology. *Proc. Natl. Acad. Sci. U.S.A.* 107, 9293–9298. doi: 10.1073/pnas.0911494107

Lindsey, J., Mcgill, N. I., Lindsey, L. A., Green, D. K., and Cooke, H. J. (1991). In vivo loss of telomeric repeats with age in humans. *Mutat. Res.* 256, 45–48. doi: 10.1016/0921-8734(91)90032-7

Mangino, M., Christiansen, L., Stone, R., Hunt, S. C., Horvath, K., Eisenberg, D. T., et al. (2015). DCAF4, a novel gene associated with leucocyte telomere length. *J. Med. Genet.* 52, 157–162. doi: 10.1136/jmedgenet-2014-102681

Mangino, M., Hwang, S. J., Spector, T. D., Hunt, S. C., Kimura, M., Fitzpatrick, A. L., et al. (2012). Genome-wide meta-analysis points to CTC1 and ZNF676 as genes regulating telomere homeostasis in humans. *Hum. Mol. Genet.* 21, 5385–5394. doi: 10.1093/hmg/dds382

Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., et al. (2009). Finding the missing heritability of complex diseases. *Nature* 461, 747–753.

Marioni, R. E., Harris, S. E., Shah, S., Mcrae, A. F., Von Zglinicki, T., Martin-Ruiz, C., et al. (2016). The epigenetic clock and telomere length are independently associated with chronological age and mortality. *Int. J. Epidemiol.* 45, 424–432. doi: 10.1093/ije/dyw041

Martin-Ruiz, C., Dickinson, H. O., Keys, B., Rowan, E., Kenny, R. A., and Von Zglinicki, T. (2006). Telomere length predicts poststroke mortality, dementia, and cognitive decline. *Ann. Neurol.* 60, 174–180. doi: 10.1002/ana.20869

Martin-Ruiz, C. M., Gussekloo, J., Van Heemst, D., Von Zglinicki, T., and Westendorp, R. G. (2005). Telomere length in white blood cells is not associated with morbidity or mortality in the oldest old: a population-based study. *Aging Cell* 4, 287–290. doi: 10.1111/j.1474-9726.2005.00171.x

McCarthy, S., Das, S., Kretzschmar, W., Delaneau, O., Wood, A. R., Teumer, A., et al. (2016). A reference panel of 64,976 haplotypes for genotype imputation. *Nat. Genet.* 48, 1279–1283. doi: 10.1038/ng.3643

Metcalfe, J. A., Parkhill, J., Campbell, L., Stacey, M., Biggs, P., Byrd, P. J., et al. (1996). Accelerated telomere shortening in ataxia telangiectasia. *Nat. Genet.* 13, 350–353. doi: 10.1038/ng0796-350

Needham, B. L., Rehkopf, D., Adler, N., Gregorich, S., Lin, J., Blackburn, E. H., et al. (2015). Leukocyte telomere length and mortality in the national health and nutrition examination survey, 1999-2002. *Epidemiology* 26, 528–535. doi: 10.1097/ede.0000000000000299

Njajou, O. T., Hsueh, W. C., Blackburn, E. H., Newman, A. B., Wu, S. H., Li, R., et al. (2009). Association between telomere length, specific causes of death, and years of healthy life in health, aging, and body composition, a population-based cohort study. *J. Gerontol. A Biol. Sci. Med. Sci.* 64, 860–864. doi: 10.1093/gerona/glp061

Okuda, K., Bardeguez, A., Gardner, J. P., Rodriguez, P., Ganesh, V., Kimura, M., et al. (2002). Telomere length in the newborn. *Pediatr. Res.* 52, 377–381.

Pandita, T. K. (2002). ATM function and telomere stability. *Oncogene* 21, 611–618. doi: 10.1038/sj.onc.1205060

Pardo, L. M., Mackay, I., Oostra, B., Van Duijn, C. M., and Aulchenko, Y. S. (2005). The effect of genetic drift in a young genetically isolated population. *Ann. Hum. Genet.* 69, 288–295. doi: 10.1046/j.1469-1809.2005.00162.x

Peters, M. J., Joehanes, R., Pilling, L. C., Schurmann, C., Conneely, K. N., Powell, J., et al. (2015). The transcriptional landscape of age in human peripheral blood. *Nat. Commun.* 6:8570.

Pooley, K. A., Bojesen, S. E., Weischer, M., Nielsen, S. F., Thompson, D., Amin Al Olama, A., et al. (2013). A genome-wide association scan (GWAS) for mean telomere length within the COGS project: identified loci show little association with hormone-related cancer risk. *Hum. Mol. Genet.* 22, 5056–5064. doi: 10.1093/hmg/ddt355

Ransohoff, K. J., Wu, W., Cho, H. G., Chahal, H. C., Lin, Y., Dai, H. J., et al. (2017). Two-stage genome-wide association study identifies a novel susceptibility locus associated with melanoma. *Oncotarget* 8, 17586–17592.

Rode, L., Nordestgaard, B. G., and Bojesen, S. E. (2015). Peripheral blood leukocyte telomere length and mortality among 64,637 individuals from the general population. *J. Natl. Cancer Inst.* 107:djv074.

Samani, N. J., Burton, P., Mangino, M., Ball, S. G., Balmforth, A. J., Barrett, J., et al. (2005). A genomewide linkage study of 1,933 families affected

by premature coronary artery disease: the british heart foundation (BHF) family heart study. *Am. J. Hum. Genet.* 77, 1011–1020. doi: 10.1086/498653

Samani, N. J., Erdmann, J., Hall, A. S., Hengstenberg, C., Mangino, M., Mayer, B., et al. (2007). Genomewide association analysis of coronary artery disease. *N. Engl. J. Med.* 357, 443–453.

Savitsky, K., Bar-Shira, A., Gilad, S., Rotman, G., Ziv, Y., Vanagaite, L., et al. (1995). A single ataxia telangiectasia gene with a product similar to PI-3 kinase. *Science* 268, 1749–1753. doi: 10.1126/science.7792600

Scelo, G., Purdue, M. P., Brown, K. M., Johansson, M., Wang, Z., Eckel-Passow, J. E., et al. (2017). Genome-wide association study identifies multiple risk loci for renal cell carcinoma. *Nat. Commun.* 8:15724.

Slagboom, P. E., Droog, S., and Boomsma, D. I. (1994). Genetic determination of telomere size in humans: a twin study of three age groups. *Am. J. Hum. Genet.* 55, 876–882.

Strandberg, T. E., Saijonmaa, O., Tilvis, R. S., Pitkala, K. H., Strandberg, A. Y., Miettinen, T. A., et al. (2011). Association of telomere length in older men with mortality and midlife body mass index and smoking. *J. Gerontol. A Biol. Sci. Med. Sci.* 66, 815–820. doi: 10.1093/gerona/glr064

Tong, A. S., Stern, J. L., Sfeir, A., Kartawinata, M., De Lange, T., Zhu, X. D., et al. (2015). ATM and ATR signaling regulate the recruitment of human telomerase to telomeres. *Cell Rep.* 13, 1633–1646. doi: 10.1016/j.celrep.2015.10.041

van Rooij, J. G. J., Jhamai, M., Arp, P. P., Nouwens, S. C. A., Verkerk, M., Hofman, A., et al. (2017). Population-specific genetic variation in large sequencing data sets: why more data is still better. *Eur. J. Hum. Genet.* 25, 1173–1175. doi: 10.1038/ejhg.2017.110

Vasa-Nicotera, M., Brouilette, S., Mangino, M., Thompson, J. R., Braund, P., Clemitson, J. R., et al. (2005). Mapping of a major locus that determines telomere length in humans. *Am. J. Hum. Genet.* 76, 147–151. doi: 10.1086/426734

Willeit, P., Willeit, J., Brandstatter, A., Ehrlenbach, S., Mayr, A., Gasperi, A., et al. (2010a). Cellular aging reflected by leukocyte telomere length predicts advanced atherosclerosis and cardiovascular disease risk. *Arterioscler. Thromb. Vasc. Biol.* 30, 1649–1656. doi: 10.1161/atvbaha.110.205492

Willeit, P., Willeit, J., Kloss-Brandstatter, A., Kronenberg, F., and Kiechl, S. (2011). Fifteen-year follow-up of association between telomere length and incident cancer and cancer mortality. *JAMA* 306, 42–44.

Willeit, P., Willeit, J., Mayr, A., Weger, S., Oberhollenzer, F., Brandstatter, A., et al. (2010b). Telomere length and risk of incident cancer and cancer mortality. *JAMA* 304, 69–75.

Willer, C. J., Li, Y., and Abecasis, G. R. (2010). METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* 26, 2190–2191. doi: 10.1093/bioinformatics/btq340

Wilson, W. R., Herbert, K. E., Mistry, Y., Stevens, S. E., Patel, H. R., Hastings, R. A., et al. (2008). Blood leucocyte telomere DNA content predicts vascular telomere DNA content in humans with and without vascular disease. *Eur. Heart J.* 29, 2689–2694. doi: 10.1093/eurheartj/ehn386

Zhan, X., Hu, Y., Li, B., Abecasis, G. R., and Liu, D. J. (2016). RVTESTS: an efficient and comprehensive tool for rare variant association analysis using sequence data. *Bioinformatics* 32, 1423–1426. doi: 10.1093/bioinformatics/btw079

Zhao, S., and Fernald, R. D. (2005). Comprehensive algorithm for quantitative real-time polymerase chain reaction. *J. Comput. Biol.* 12, 1047–1064. doi: 10.1089/cmb.2005.12.1047

Conflict of Interest: AS is an employee of the company SkylineDx.

The remaining 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

BRIEF RESEARCH REPORT
published: 25 August 2020
doi: 10.3389/fgene.2020.00874

A Pilot Study on the Whole Exome Sequencing of Prostate Cancer in the Indian Phenotype Reveals Distinct Polymorphisms

*Ayam Gupta[1,2†], Nidhi Shukla[1,3†], Mamta Nehra[1], Sonal Gupta[1], Babita Malik[3], Ashwani Kumar Mishra[4], Maneesh Vijay[5], Jyotsna Batra[6], Nirmal Kumar Lohiya[7], Devendra Sharma[5] and Prashanth Suravajhala[1]**

[1] Department of Biotechnology and Bioinformatics, Birla Institute of Scientific Research, Jaipur, India, [2] Vignan's Foundation for Science, Technology & Research (Deemed to be University), Guntur, India, [3] Department of Chemistry, School of Basic Sciences, Manipal University Jaipur, Jaipur, India, [4] DNA Xperts Private Limited, Noida, India, [5] Rukmani Birla Hospitals, Jaipur, India, [6] Australian Prostate Cancer Research Centre, Queensland Institute of Health and Biomedical Innovation and School of Biomedical Science, Queensland University of Technology, Brisbane, QLD, Australia, [7] Department of Zoology, University of Rajasthan, Jaipur, India

OPEN ACCESS

Edited by:
Marika Kaakinen, University of Surrey, United Kingdom

Reviewed by:
Jyoti Sharma, Institute of Bioinformatics (IOB), India
Pritish Kumar Varadwaj, Indian Institute of Information Technology Allahabad, India

***Correspondence:**
Prashanth Suravajhala
prash@bisr.res.in

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

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

Received: *20 March 2020*
Accepted: *16 July 2020*
Published: *25 August 2020*

Citation:
Gupta A, Shukla N, Nehra M, Gupta S, Malik B, Mishra AK, Vijay M, Batra J, Lohiya NK, Sharma D and Suravajhala P (2020) A Pilot Study on the Whole Exome Sequencing of Prostate Cancer in the Indian Phenotype Reveals Distinct Polymorphisms. Front. Genet. 11:874. doi: 10.3389/fgene.2020.00874

Prostate cancer (PCa) is the third most common cancer among men in India, and no next-generation sequencing (NGS) studies have been attempted earlier. Recent advances in NGS have heralded the discovery of biomarkers from Caucasian/European and Chinese ancestry, but not much is known about the Indian phenotype/variant of PCa. In a pilot study using the whole exome sequencing of benign/PCa patients, we identified characteristic mutations specific to the Indian sub-population. We observed a large number of mutations in DNA repair genes, *viz.* helicases, TP53, and BRCA besides the variants of unknown significance with a possibly damaging rare variant (rs730881069/chr19:55154172C/TR136Q) in the TNNI3 gene that has been previously reported as a semi-conservative amino acid substitution. Our pilot study attempts to bring an understanding of PCa prognosis and recurrence for the Indian phenotype.

Keywords: prostate cancer, genomics, exome sequencing, prognosis, biomarkers

INTRODUCTION

Prostate cancer (PCa) is the second most prevalent cancer worldwide and the third most prevalent cancer in India (Jain et al., 2014). Over 1 million PCa cases are diagnosed per year globally, and the mortality rate has grown to more than 300,000 deaths per year (Farhat et al., 2000). The Population Based Cancer Registries (PBCRs) of different cities for the time period shows that PCa has ranked among the top leading sites of cancer in many cities in India (NCRP, 2013). Prostate cancer leads in large Indian cities like Delhi, Kolkata, Pune, and Thiruvananthapuram after oral cancer, and is the third leading site of cancer in cities like Bangalore and Mumbai besides being among the top 10 leading sites of cancers in the rest of the PBCRs of India (Jain et al., 2014). The data show that almost all regions of India are equally affected, with the incidences of PCa relatively low in some states like Gujarat (Ahmedabad and Wardha PBCRs) and Madhya Pradesh (Bhopal PBCR), the lowest being the northeast regions of India (Jain et al., 2014). This could be because of lack of PBCRs in addition to PCa not actively reported in states where awareness about it is lacking (Kimura, 2012).

Although studies on PCa have dealt with the genetics, genomics, and environmental influence in the causality of PCa, no association of genotype and phenotype employing the whole exome sequencing (WES) of PCa has been done to date for the Indian population. As many Indians are vegetarian, a polygenic risk with other diseases such as diabetes could be associated with PCa, which would be interesting to deliberate (Agrawal et al., 2014). On the other hand, studies at the sequence level in India have been limited where the genetic variants have been shown to be associated with PCa but not linked to pathogenesis (Abate-Shen and Shen, 2000; Shen and Abate-Shen, 2010). In the recent past, the next-generation sequencing (NGS) approaches have provided an efficient, rapid, economical, and global biological means to understand millions of sequenced DNA reads besides recognizing ample variants from the samples (Mardis, 2008). It has an increasing potential to carry a wide spectrum of purposes including many research fields, for example, in the molecular diagnostics of genetic diseases, infectious diseases, cancer, and pharmacogenomics (Ten Bosch and Grody, 2008; Guchelaar et al., 2014). On the other hand, the integration of systems biology with genomic data can lead to the profound knowledge of the disease pathway besides allowing us to identify prognostic biomarkers for drug target identification.

In this pilot study, while we report mutations from northwestern Indian patients, we aimed to understand the marked impact of the risk of PCa specific genes on affected individuals versus controls. We claim that this is perhaps the first study of its kind, where WES was employed to screen mutations and infer the genetic evaluation of PCa in India. With the new Genome India Project initiative (GenomeIndia Initiative, 2017), the genetic makeup for such diseases could allow us to understand the pathways associated with PCa native to India. We also deliberate on mutations associated with polygenic diseases and vitamin D deficiency, besides those that are unique to human diet, socioeconomic imbalances, and PCa manifestation when compared to the Caucasian population (Layne et al., 2019).

MATERIALS AND METHODS

Subjects

The samples were collected from Rukmani Birla Hospital (RBH), Jaipur, and all patients were identified as native to north or northwest India. All methods were performed in accordance with relevant guidelines and regulations. Archival pathological FFPE blocks containing specimens were obtained retrospectively after clearance from the institutional ethics committee of RBH, and informed consent wherever applicable was duly taken. Annotated FFPE blocks of five benign and five tumors (malignant) were taken and sent to Xcelris Labs, Ahmedabad, for WES (**Table 1**). DNA from sample tissues was extracted using the FFPE tissue gDNA isolation kit (Qiagen/56404) for DNA isolation, and quality control (QC) for contamination was done. The baits for exon capture with a mean target coverage depth of 110× per sample were achieved with as many as several targets covered for more than 20×.

TABLE 1 | Samples used for the WES along with their Gleason scores which constituted one high-grade tumor sample, with three intermediate samples and one just on par above grade 6, while five others are with less than six from benign cases.

Sample ID	Condition	Gleason Score (primary + secondary)
Z785	Adenocarcinoma	3 + 4 = 7
Z786	Acinar Adenocarcinoma	3 + 4 = 7
Z789	Acinar Adenocarcinoma	3 + 3 = 6
Z791	Acinar Adenocarcinoma	3 + 4 = 7
Z794	Cribriform Adenocarcinoma	4 + 4 = 8
Z787	Benign Nodular Prostatic Hyperplasia	<6
Z788	Benign Prostatic Hyperplasia	<6
Z790	Benign Nodular Prostatic Hyperplasia	<6
Z792	Benign Nodular Prostatic Hyperplasia	<6
Z793	Benign Nodular Prostatic Hyperplasia	<6

Sample Preparation, Exome Capture, and Sequencing

DNA was isolated from samples using the Qiagen FFPE DNA extraction kit. The quality of genomic DNA was checked on 0.8% agarose gel (loaded 3 μl) for the single intact band. The gel was run at 110 V for 30 min; 1 μl of each sample was used for determining the concentration using the Qubit 2.0 Fluorometer. All the libraries were prepared using the Agilent SureSelect XT target enrichment system by following the manufacturer's instruction. Briefly, 200 ng of each DNA sample was used for fragmentation using the Covaris S2 system. While the fragmented DNA was subjected to end repair, A-tailing, it was followed by adapter ligation which was hybridized to RNA baits that are designed to physically capture specific DNA sequences. The captured DNA is then eluted from the baits, purified using a biotin-based precipitation, and then amplified by PCR to yield an exome library for sequencing. All prepared libraries were checked on the Agilent high-sensitivity (HS) chip on Bioanalyzer 2100 and quantified on fluorometer by the Qubit dsDNA HS Assay kit (Life Technologies). The average size and concentration of each library was calculated from the HS chip and Qubit, respectively (see **Table 2**).

The amplified library was analyzed on Bioanalyzer 2100 (Agilent Technologies) using a HS DNA chip as per the manufacturer's instructions. After obtaining the Qubit concentration for the library and mean peak size from Bioanalyzer, the profile library was loaded onto the Illumina platform for cluster generation and sequencing. Paired-end (PE) sequencing allowing the template fragments to be sequenced in both forward and reverse directions was ensured with the library molecules bound to complementary adapter oligos on PE flow cells. The adapters were designed to allow the selective cleavage of the forward strands after the re-synthesis of the reverse strand during sequencing. Finally, the copied reverse strand was used to sequence from the opposite end of the fragment, and all the libraries were sequenced with 150 × 2 PE reads, which resulted in 8–10 GB of data per sample.

TABLE 2 | Data statistics of the sequenced samples.

File Name	Total Reads	Total Bases		Size in GB
Z785	34,589,627	4,524,820,258	9,035,044,752	9.03
Z785	34,589,627	4,510,224,494		
Z786	36,192,378	4,782,261,247	9,552,530,147	9.55
Z786	36,192,378	4,770,268,900		
Z787	29,342,454	3,504,276,014	7,023,683,778	7.02
Z787	29,342,454	3,519,407,764		
Z788	32,007,925	4,590,702,434	9,142,339,304	9.14
Z788	32,007,925	4,551,636,870		
Z789	32,912,966	4,688,597,419	9,332,868,751	9.33
Z789	32,912,966	4,644,271,332		
Z790	44,212,484	6,398,843,741	12,730,582,043	12.73
Z790	44,212,484	6,331,738,302		
Z791	33,311,709	4,822,305,019	9,600,461,768	9.60
Z791	33,311,709	4,778,156,749		
Z792	36,245,502	5,367,958,348	10,658,814,541	10.65
Z792	36,245,502	5,290,856,193		
Z793	37,261,051	5,344,729,573	10,640,702,373	10.64
Z793	37,261,051	5,295,972,800		
Z794	33,200,697	4,584,192,242	9,148,425,756	9.14
Z794	33,200,697	4,564,233,514		

Quality Control and Variant Calling

The datasets were run through our in-house pipeline described earlier (Meena et al., 2018).The quality assessment was done for all the samples using FastQC (Andrews, 2010) with raw reads checked for quality, GC bias, K-mer quality, and duplication levels. Calling and filtering of variants and indels were performed using Vt (Tan et al., 2015) and Annovar (Wang et al., 2010) before we established the sensitivity and specificity of called variants. Mutations were counted as heterozygous ("het") using awk/bash one-liners and the two sets of average depth ($\geq$5 DP $\leq$ 20 and DP $\geq$ 20) were parsed for further downstream annotation (**Supplementary Information**). Population stratification at a gross-level was deemed not necessary given the small size of the population on a small scale; however, contamination crosscheck both at sequence (fastq and bam levels) were ensured to assess the heterogeneity of each given site of pooled dataset. Variant analysis using Vt predicted the variant types and were filtered by setting criteria for a false discovery rate (FDR) that yielded the genomic variants in affected individuals versus controls. The samples were analyzed for mutations predicted as causal, and the corresponding genes were filtered against GeneMania (Warde-Farley et al., 2010) and checked for the significant association of pathogenesis with the risk of genitourinary/PCa-specific genes. A brief overview of the methodology is described in **Figure 1**.

Downstream Analyses

Downstream annotation was done using various open source tools. The reported PCa variants were confirmed from NCBIs' ClinVar (NCBI ClinVar Database, 2019) using the search term "prostate" with filters "pathogenic" and "single nucleotide" and checked against the generated VCF files. The bcftools was employed to retrieve the heterozygous mutations with the varied depth as mentioned earlier. A final list of variants was validated using Sanger sequencing. For the latter, we have maintained an average of 400–600 bp of product size with the 200–250 bp flanking from both sides of the location of the SNP mutation. Primers were designed in such a way that secondary structures in the genome were avoided. We adjusted the GC content as 50% to maintain a balanced distribution of GC-rich in CDS regions. The PCR was set with initial denaturation at 95°C for 5 min, followed by 35 cycles of denaturation at 95°C for 30 s, annealing at 60°C for 30 s and an extension at 72°C for 45 s and final extension of 5 min at 72°C.

RESULTS AND DISCUSSION

Characteristics of PCa Relevant to the Indian Phenotype

Ten subjects, including five cases, those including one high Gleason grade sample (8), and four between 6 and 7 grades were subjected to WES. The sequences were aligned to human genome reference (build hg38) using bowtie2 to produce the sequence alignment file. We observed 944,509 annotated variants and upon further normalization, we discovered 897,926 SNPs, indels and a few copy number variants (CNVs) among them. Post downstream analyses, the generated VCF files were filtered for the variants which were heterozygous and 42 SNPs were confirmed that exhibited significant association with PCa in affected subjects (**Supplementary Table 1**). Among them, 17 SNPs were subjected further to Sanger validation in an independent platform wherein primers were designed with a length of 18–25 nucleotides (**Supplementary Table 2**). We found that all the SNPs/samples except one unaffected (rs73598374) passed the sequencing quality thresholds from quality checks with a mean read depth of 14.4. The false positives were carefully checked between benign prostate hyperplasia (BPH) and PCa samples even as a number of exome sites, low-coverage sites, and large deletions could not be validated. In total, we observed 671,609 heterozygous variants confirming the above thresholds that were filtered across all samples.

Association of Variants With PCa

The variants segregating from cases were retained, while those from controls were excluded to rule out the false positives. The minor allele frequency cut-off 0.05 intersecting the allele frequency were judiciously checked for low penetrated variants that are associated with a large number of disease susceptibility genes, primarily cancers. Therefore, we setup the variant identification to as low as DP = 5 and further, segregated the variants those that fall in $\geq$ 5 DP $\leq$ 20 and DP $\geq$ 20 (**Figure 2A**).

High-Impact Variants in High-Grade Tumors

The identification of tumor-specific somatic mutations was possible with our pipeline. Among the inferred mutations, the prominent were BRCA2 mutations (rs276174889, rs80358600,

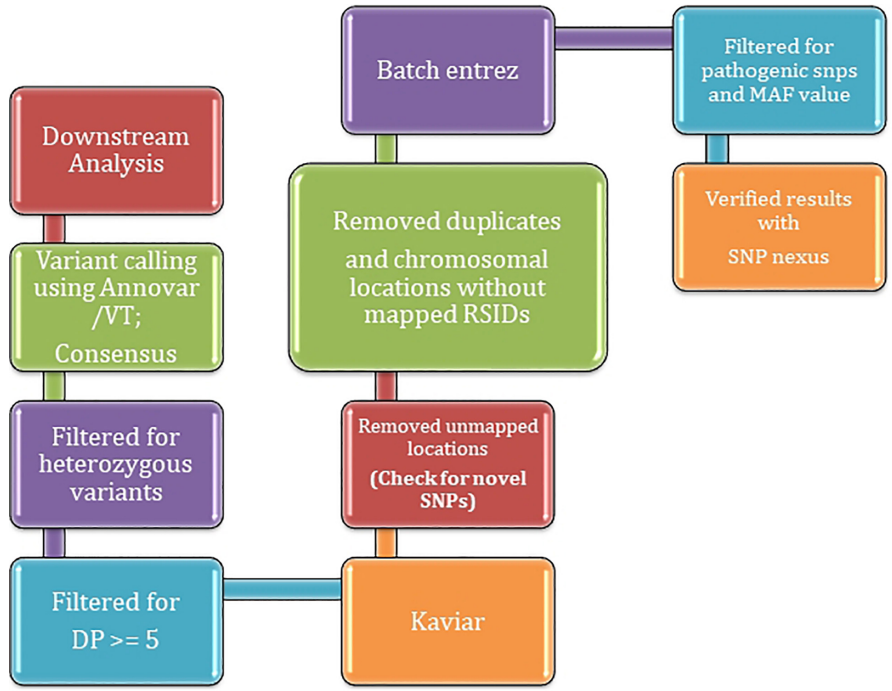


FIGURE 1 | A brief overview of methodology employed for inferring variants using WES.

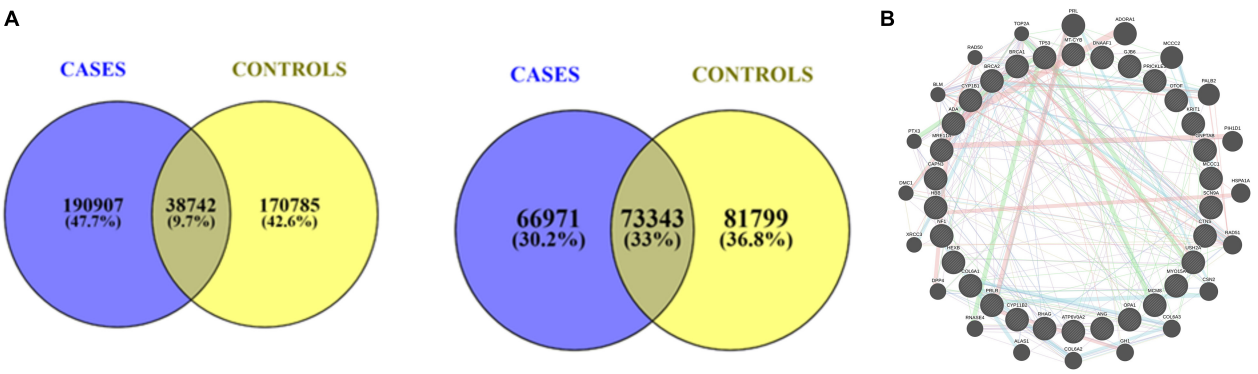


FIGURE 2 | (A) A union of intersection of variants between cases and controls with respect to depth ≥ 5 DP ≤ 20 respectively. **(B)** Protein-Protein Interaction (PPI) map of genes associated with DNA repair genes.

rs80359171, rs771203198, and rs145988146) associated with high-grade and intermediate tumors (Z794, Z785, and Z786) (Couch et al., 2007). While, BRCA gene mutations lead to development of breast and ovarian cancer, they have been linked to PCa in addition to pancreatic cancer and other myelomas (Paul and Paul, 2014; Cavanagh and Rogers, 2015). The BRCA2 mutations in particular have been associated with an 8.6-fold increased risk for the PCa manifestation (Castro and Eeles, 2012) even as BRCA1 was also seen characteristically in one sample (Z785): rs28897696. Common genetic alterations are usually associated with heterozygous BRCA1 or BRCA2 mutations, and these include loss of the wild-type BRCA1 or BRCA2 allele (LOH), loss of TP53 (which encodes p53), and ATM or CHK2 function (Roy et al., 2012). These additional alterations may allow cells to bypass checkpoint controls and evade apoptosis, and thereby initiate tumorigenesis. Additionally, tumor complexity makes the detection of cancer-specific CNVs even more difficult (Zare et al., 2017), and the variants observed are frameshift, loss, or gain of stop codons which have significant insertions/deletions at a number of bases in different reading frames with altered protein sequences. While we found them to be significantly associated, we observed that mutations in TP53 and COL6A1 among the other 102 frameshift variants were shown to be affected with spliced regions. Sanger validation (orthogonal) was used to disprove a false-positive variant wherein we performed a

systematic validation of variants of these genes and found no discrepancies among them.

The majority of mutations that are detected in the current study (*viz., GJB6, KRIT1, GNPTAB, ANG, MCM8*, and *NF1*) are only significant to high-grade PCa. The SNPs have been widely studied in context to cancer genomes and its predisposition, and so, we believe these SNPs cannot be associated with the familial aspect of the cancer. On the other hand, we observed that a few pathogenic SNPs were absent in high- and median-grade PCa samples and present in the low-grade sample, *viz., BRAF, RAD51C, RHAG, CYP11B2, PRICKLE1, CAPN3.* With the CNVs considered one of the most important somatic aberrations in cancer since oncogene activation, it is often attributed to chromosomal copy number amplification and tumor suppressor gene inactivation leading to largely heterozygous deletion. With the identification of somatic CNVs known to have an important role in cancer prognosis and treatment, we asked if any of them were significantly associated. On the other hand, we also found 76 CNVs associated with the gain/loss of function that we mapped using SNP nexus (Dayem Ullah et al., 2018) (**Supplementary Table 3**).

As expected, a large number of DNA repair genes (e.g., helicase, TP53, BRCA2) harbor characteristic mutations, which is in agreement with the WES studies. We also found possibly damaging rare variants of unknown significance (VUS) (rs730881069/chr19:55154172: C/T: R136Q) in the *TNNI3* gene that has been previously reported as a semi-conservative amino acid substitution. The mutation, howev2wer, is known to impact the PCa cases and seen across multiple hypertrophic cardiomyopathy case reports and, hence, classified as likely pathogenic. From the pathway and interaction network analyses, we inferred that *BRCA2* and *BRCA1* have a majority of interacting partners that are largely associated with either DNA repair genes or helicases, *viz., KRIT1, CTNS, USH2A, HEXB, MCCC1, MYO15A, CAPN3, GNPTAB, RHAG, PRLR, OTOF, GJB6, MCM8, CYP1B1, PRICKLE1, TP53, COL6A1, DNAAF1, SCN9A, CYP11B2, ATP6V0A2, NF1, ADA, MT-CYB, OPA1, ANG, HBB, MRE11* (**Figure 2B**). Notable among them is RecQ/BLM helicase, which has been recently reported to regulate cancer cell proliferation (Qian et al., 2017). In the recent past, recommendations demonstrated the role of prostate serum antigen (PSA) levels correlating the patient's race, age, and prostate volume (Wu et al., 2017). Since the use of PSA is limited and controversial, the search for novel PCa-specific biomarkers, especially from non-invasive bio-fluids, is an important task (Wu et al., 2017). Furthermore, while it is assumed that the impacts of the discovered SNPs on tumor initiation and progression cannot be established due to small sample size, we found that certain genes have a vivid association with the expression or tumor type in general agreement. For example, CYP1B1 is known for increases in high-grade PCa and correlates with Gleason grades (Chang et al., 2017), P53 and PRLR are shown to correlate with Gleason grades (Jacobson et al., 2011; Martinez-Gonzalez et al., 2018), MRE11 is known to have an elevated expression associated with progression and poor outcome in PCa (Wang et al., 2019), and ANG is up-regulated in the prostate and involved in prostate intraepithelial neoplasia (PIN) formation (Li et al., 2011; Vanli and Guo-Fu, 2015). As no single biomarker is available for diagnostic and prognostic use (Velonas et al., 2013), the protein–protein interactions complementing our analyses show characteristic associations with a large number of pathways enriched between them (**Figure 2B**). Although only 5.42% of the protein interaction map density constitutes the pathways, COL6A1, which is known to be upregulated in castration-resistant prostate cancer (CRPC), remains to be the most critical oncogene regulating the androgen signaling pathways (Zhu et al., 2015). This is also in agreement with the pathway association of its sibling genes COL6A2 and COL6A3 (as indicated in light blue edges). Taken together, our results suggest the importance of these pathways and support the potential use of COL6A family members in PCa.

In our PCa cohort, TP53 is the highest prevalent mutation which is in agreement with the previous studies reported (Ecke et al., 2010). Apart from well-known and promising cancer genes, our study uncovered several genes with poorly characterized functional roles in cancer which needs further experimental investigation. For example, *KRIT1* is an intracellular protein with ankyrin repeats and a FERM domain that interacts with the Ras-family GTPase Krev1/Rap1a inferring a role in GTPase signaling cascades. Most of the studies linked were its involvement in cardiovascular development but it is known to be ubiquitously expressed in many cells and tissues (Laberge-le Couteulx et al., 1999). *USH2A*, a gene responsible for Usher syndrome, has been found to be mutated in our cohort demonstrating its role as tumor suppressor (Van Wijk et al., 2004). Although, some other genes such as MCC1, COL6A1, and CYP11B2 are associated with different cancers such as pancreatic, colorectal, and hepatocellular carcinoma, their involvement in PCa is not yet explored (Fukuyama et al., 2008; Fan et al., 2016; Owusu-Ansah et al., 2019). Previous studies, however, have reported the role of diet in PCa (Labbé et al., 2015), so it will be enticing to explore if the genes mentioned above have any role in diet induced PCa. One of the genes, *viz., OPA1*, *is* mutated in our cohort that is important for maintaining normal mitochondrial morphology and function, with deficiency of *OPA1* checking diet-induced obesity and insulin resistance (Pereira et al., 2018). Likewise, *CYP1B1* prevents diet-induced obesity and glucose intolerance through AMPK activation (Liu et al., 2015). We also observed that CTNS (17p13), a gene that causes cystinosis, is associated with multisystem genetic disorder, wherein cysteine is accumulated in different tissues including the adipose and prostate. Strikingly, it is known that vitamin D repletion could be associated with CTNS knockout or mutated mice indicating vitamin D as a key-player in the pathogenesis of PCa (Cheung et al., 2019).

CONCLUSION

Prostate cancer is burgeoning in India. Our comprehensive bioinformatics analysis confirmed some characteristic known or unknown mutations from a WES study native to India. Although our study shows characteristic mutations in certain genes, an assay comprising multiple biomarkers that are differentially expressed could be attempted in the future. If this is successful, the number of biomarkers developed will depend on their

validation in a large cohort of patients and the translation of these findings to clinical practice. This pilot study, we believe, is decisive to understand the inherent genes and mutations responsible for PCa in India. Furthermore, an attempt was made to develop a conceptual framework for research particularly in propagating information on the causal genes and mutations responsible for PCa. Although the work was limited to a small number of samples studied, we deem this pilot work would have an impending role in understanding mutations that are of particular interest to Indian genealogy.

DATA AVAILABILITY STATEMENT

The raw reads of WES are deposited in NCBI's sequence read archive (SRA) with accession PRJNA616165.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by CK Birla/Rukmani Birla Hospital. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

AG, MN, and PS analyzed the PCA samples using our NGS pipeline. NS, SG, and BM cross-checked the downstream analysis. MV and DS provided the clinical samples. AM, JB, and NL helped in the Sequenom analysis. AG wrote the first draft with PS. PS proofread the manuscript before uploading and all others agreed to the changes in the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

The authors declare that this study received seed funding from Brainpan Innovations. The funder was not involved in the study design, collection, analysis, and interpretation of data, the writing of this article or the decision to submit it for publication.

ACKNOWLEDGMENTS

The authors gratefully acknowledge the critical comments of Jayendra Nath Shukla, Praveen Mathur, and Krishna Mohan Medicherla for their kind inputs and suggestions.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.00874/full#supplementary-material

REFERENCES

Abate-Shen, C., and Shen, M. M. (2000). Molecular genetics of prostate cancer. *Genes Dev.* 14, 2410–2434. doi: 10.1101/gad.819500

Agrawal, S., Millett, C. J., Dhillon, P. K., Subramanian, S., and Ebrahim, S. (2014). Type of vegetarian diet, obesity and diabetes in the adult Indian population. *Nutr. J.* 13:89. doi: 10.1186/1475-2891-13-89

Andrews, S. (2010). FastQC: a quality control tool for high throughput sequence data. *Babraham Bioinform.*

Castro, E., and Eeles, R. (2012). The role of BRCA1 and BRCA2 in prostate cancer. *Asian J. Androl.* 14, 409–414. doi: 10.1038/aja.2011.150

Cavanagh, H., and Rogers, K. M. A. (2015). The role of BRCA1 and BRCA2 mutations in prostate, pancreatic and stomach cancers. *Hered. Cancer Clin. Pract.* 13:16. doi: 10.1186/s13053-015-0038-x

Chang, I., Mitsui, Y., Kim, S. K., Sun, J. S., Jeon, H. S., Kang, J. Y., et al. (2017). Cytochrome P450 1B1 inhibition suppresses tumorigenicity of prostate cancer via caspase-1 activation. *Oncotarget* 8:34190. doi: 10.18632/oncotarget.16598

Cheung, W. W., Hao, S., and Wang, Z. (2019). Vitamin D repletion ameliorates adipose tissue browning and muscle wasting in infantile nephropathic cystinosis-associated cachexia. *J. Cachexia Sarcopenia Muscle* 11, 120–134. doi: 10.1002/jcsm.12497

Couch, F. J., Johnson, M. R., Rabe, K. G., Brune, K., De Andrade, M., Goggins, M., et al. (2007). The prevalence of BRCA2 mutations in familial pancreatic cancer. *Cancer Epidemiol. Biomark. Prev.* 16, 342–346. doi: 10.1158/1055-9965.EPI-06-0783

Dayem Ullah, A. Z., Oscanoa, J., Wang, J., Nagano, A., Lemoine, N. R., and Chelala, C. (2018). SNPnexus: assessing the functional relevance of genetic variation to facilitate the promise of precision medicine. *Nucleic Acids Res.* 46, W109–W113. doi: 10.1093/nar/gky399

Ecke, T. H., Schlechte, H. H., Schiemenz, K., Sachs, M. D., Lenk, S. V., Rudolph, B. D., et al. (2010). TP53 gene mutations in prostate cancer progression. *Anticancer Res.* 30, 1579–1586.

Fan, Z., Wang, Z., Chen, W., Cao, Z., and Li, Y. (2016). Association between the CYP11 family and six cancer types. *Oncol. Lett.* 12, 35–40. doi: 10.3892/ol.2016.4567

Farhat, W. A., Habbal, A. A., and Khauli, R. B. (2000). A guideline to clinical utility of prostate specific antigen. *Saudi Med. J.* 21, 223–227.

Fukuyama, R., Niculaita, R., Ng, K. P., Obusez, E., Sanchez, J., Kalady, M., et al. (2008). Mutated in colorectal cancer, a putative tumor suppressor for serrated colorectal cancer, selectively represses β-catenin-dependent transcription. *Oncogene* 27, 6044–6055. doi: 10.1038/onc.2008.204

Guchelaar, H. J., Gelderblom, H., Van Der Straaten, T., Schellens, J., and Swen, J. (2014). Pharmacogenetics in the cancer clinic: from candidate gene studies to next-generation sequencing. *Clin. Pharmacol. Ther.* 95, 383–385. doi: 10.1038/clpt.2014.13

GenomeIndia Initiative (2017). *GenomeIndia Initiative Launched to Unlock India's DNA. 1.* Available online at: https://www.iisc.ac.in/events/genomeindia-initiative-launched-to-unlock-indias-dna/ (accessed August 27, 2019).

Jacobson, E. M., Hugo, E. R., Borcherding, D. C., and Ben-Jonathan, N. (2011). Prolactin in breast and prostate cancer: molecular and genetic perspectives. *Discov. Med.* 11, 315–324.

Jain, S., Saxena, S., and Kumar, A. (2014). Epidemiology of prostate cancer in India. *Meta Gene* 2, 596–605. doi: 10.1016/j.mgene.2014.07.007

Kimura, T. (2012). East meets west: ethnic differences in prostate cancer epidemiology between East Asians and Caucasians. *Chin. J. Cancer* 31, 421–429. doi: 10.5732/cjc.011.10324

Labbé, D. P., Zadra, G., Ebot, E. M., Mucci, L. A., Kantoff, P. W., Loda, M., et al. (2015). Role of diet in prostate cancer: the epigenetic link. *Oncogene* 34, 4683–4691. doi: 10.1038/onc.2014.422

Laberge-le Couteulx, S., Jung, H. H., Labauge, P., Houtteville, J. P., Lescoat, C., Cecillon, M., et al. (1999). Truncating mutations in CCM1, encoding KRIT1, cause hereditary cavernous angiomas. *Nat. Genet.* 23, 189–193. doi: 10.1038/13815

Layne, T. M., Graubard, B. I., Ma, X., Mayne, S. T., and Albanes, D. (2019). Prostate cancer risk factors in black and white men in the NIH-AARP diet and health study. *Prostate Cancer Prostatic Dis.* 22, 91–100. doi: 10.1038/s41391-018-0070-79

Li, S., Ibaragi, S., and Hu, G. (2011). Angiogenin as a molecular target for the treatment of prostate cancer. *Curr. Cancer Ther. Rev.* 7, 83–90. doi: 10.2174/157339471107020083

Liu, X., Huang, T., Li, L., Tang, Y., Tian, Y., Wang, S., et al. (2015). CYP1B1 deficiency ameliorates obesity and glucose intolerance induced by high fat diet in adult C57BL/6J mice. *Am. J. Transl. Res.* 7, 761–771.

Mardis, E. R. (2008). The impact of next-generation sequencing technology on genetics. *Trends Genet.* 24, 133–141. doi: 10.1016/j.tig.2007.12.007

Martinez-Gonzalez, L. J., Pascual Geler, M., Robles Fernandez, I., Cozar, J. M., Lorente, J. A., and Alvarez Cubero, M. J. (2018). Improving the genetic signature of prostate cancer, the somatic mutations. *Urol. Oncol. Semin. Orig. Investig.* 36, 312.e17–312.e23. doi: 10.1016/j.urolonc.2018.03.012

Meena, N., Mathur, P., Medicherla, K., and Suravajhala, P. (2018). A bioinformatics pipeline for WHOLE exome sequencing: overview of the processing and steps from raw data to downstream analysis. *Biol. Protoc.* 8:2805. doi: 10.21769/BioProtoc.2805

NCBI ClinVar Database (2019). *NCBI NCBI ClinVar Database*. Available online at: https://www.ncbi.nlm.nih.gov/clinvar (accessed October 25, 2019).

NCRP (2013). *Three-Year Report of the Population Based Cancer Registries- 2009-2011*. Bethesda, MD: NCRP.

Owusu-Ansah, K. G., Song, G., Chen, R., Edoo, M. I. A., Li, J., Chen, B., et al. (2019). COL6A1 promotes metastasis and predicts poor prognosis in patients with pancreatic cancer. *Int. J. Oncol.* 55, 391–404. doi: 10.3892/ijo.2019.4825

Paul, A., and Paul, S. (2014). The breast cancer susceptibility genes (BRCA) in breast and ovarian cancers. *Front. Biosci. Landmark* 19, 605–618. doi: 10.2741/4230

Pereira, R., Olvera, A. C., Marti, A. A., Hewezi, R., Bui Tran, W. A., Potthoff, M. J., et al. (2018). Resistance to diet-induced obesity in mice lacking OPA1 in adipose tissue occurs independently of fat-derived FGF-21 and BAT function. *Diabetes* 67(Supplement 1) 276-LB. doi: 10.2337/db18-276-lb

Qian, X., Feng, S., Xie, D., Feng, D., Jiang, Y., and Zhang, X. (2017). RecQ helicase BLM regulates prostate cancer cell proliferation and apoptosis. *Oncol. Lett.* 14, 4206–4212. doi: 10.3892/ol.2017.6704

Roy, R., Chun, J., and Powell, S. N. (2012). BRCA1 and BRCA2: different roles in a common pathway of genome protection. *Nat. Rev. Cancer* 12, 68–78. doi: 10.1038/nrc3181

Shen, M., and Abate-Shen, C. (2010). Molecular genetics of prostate cancer: new prospects for old challenges. *Genes Dev.* 24, 1967–2000. doi: 10.1101/gad.1965810.GENES

Tan, A., Abecasis, G. R., and Kang, H. M. (2015). Unified representation of genetic variants. *Bioinformatics* 31, 2202–2204. doi: 10.1093/bioinformatics/btv112

Ten Bosch, J. R., and Grody, W. W. (2008). Keeping up with the next generation: massively parallel sequencing in clinical diagnostics. *J. Mol. Diagnost.* 10, 484–492. doi: 10.2353/jmoldx.2008.080027

Van Wijk, E., Pennings, R. J. E., Te Brinke, H., Claassen, A., Yntema, H. G., Hoefsloot, L. H., et al. (2004). Identification of 51 novel exons of the usher syndrome Type 2A (USH2A) gene that encode multiple conserved functional domains and that are mutated in patients with usher syndrome Type II. *Am. J. Hum. Genet.* 74, 738–744. doi: 10.1086/383096

Vanli, N., and Guo-Fu, H. U. (2015). Mechanism and function of angiogenin in prostate cancer. *Chin. J. Biochem. Mol. Biol.* 31, 1261–1266. doi: 10.13865/j.cnki.cjbmb.2015.12.06

Velonas, V., Woo, H., Remedios, C., and Assinder, S. (2013). Current status of biomarkers for prostate cancer. *Int. J. Mol. Sci.* 14, 11034–11060. doi: 10.3390/ijms140611034

Wang, J., Xu, W.-H., Wei, Y., Zhu, Y., Qin, X.-J., Zhang, H.-L., et al. (2019). Elevated MRE11 expression associated with progression and poor outcome in prostate cancer. *J. Cancer* 10, 4333–4340. doi: 10.7150/jca.31454

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

Warde-Farley, D., Donaldson, S. L., Comes, O., Zuberi, K., Badrawi, R., Chao, P., et al. (2010). The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. *Nucleic Acids Res.* 38, W214–W220. doi: 10.1093/nar/gkq537

Wu, D., Ni, J., Beretov, J., Cozzi, P., Willcox, M., Wasinger, V., et al. (2017). Urinary biomarkers in prostate cancer detection and monitoring progression. *Crit. Rev. Oncol. Hematol.* 118, 15–26. doi: 10.1016/j.critrevonc.2017.08.002

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 Bioinform.* 18:286. doi: 10.1186/s12859-017-1705-x

Zhu, Y. P., Wan, F. N., Shen, Y. J., Wang, H. K., Zhang, G. M., and Ye, D. W. (2015). Reactive stroma component COL6A1 is upregulated in castration-resistant prostate cancer and promotes tumor growth. *Oncotarget* 6, 14488–14496. doi: 10.18632/oncotarget.3697

Conflict of Interest: AM was employed by company DNA Xperts Pvt Ltd.

The remaining 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: 27 November 2020
doi: 10.3389/fgene.2020.539862

A Variant in the Nicotinic Acetylcholine Receptor Alpha 3 Subunit Gene Is Associated With Hypertension Risks in Hypogonadic Patients

Tao Wu[1†], Yujia Wang[2,3†], Wei Shi[2†], Bi-Qi Zhang[1], John Raelson[2], Yu-Mei Yao[4], Huan-Dong Wu[1], Zao-Xian Xu[1], Francois-Christophe Marois-Blanchet[2], Jonathan Ledoux[5], Rikard Blunck[6], Jian-Zhong Sheng[7], Shen-Jiang Hu[1]*, Hongyu Luo[2]* and Jiangping Wu[2,8]**

[1] Institute of Cardiology, The First Affiliated Hospital, College of Medicine, Zhejiang University, Hangzhou, China, [2] Research Centre, Centre Hospitalier de l'Université de Montréal (CHUM), Montreal, QC, Canada, [3] Children's Hospital, Zhejiang University School of Medicine, Hangzhou, China, [4] Department of Cardiology, The Third Affiliated Hospital of Zhejiang Chinese Medical University, Hangzhou, China, [5] Montreal Heart Institute, Université de Montréal, Montreal, QC, Canada, [6] Department of Physics, University of Montreal, Montreal, QC, Canada, [7] Department of Pathology and Physiopathology, College of Medicine, Zhejiang University, Hangzhou, China, [8] Nephrology Service, Centre Hospitalier de l'Université de Montréal (CHUM), Montreal, QC, Canada

Ephb6 gene knockout causes hypertension in castrated mice. EPHB6 controls catecholamine secretion by adrenal gland chromaffin cells (AGCCs) in a testosterone-dependent way. Nicotinic acetylcholine receptor (nAChR) is a ligand-gated Ca^{2+}/Na^{+} channel, and its opening is the first signaling event leading to catecholamine secretion by AGCCs. There is a possibility that nAChR might be involved in EPHB6 signaling, and thus sequence variants of its subunit genes are associated with hypertension risks. CHRNA3 is the major subunit of nAChR used in human and mouse AGCCs. We conducted a human genetic study to assess the association of CHRNA3 variants with hypertension risks in hypogonadic males. The study cohort included 1,500 hypogonadic Chinese males with (750 patients) or without (750 patients) hypertension. The result revealed that SNV *rs3743076* in the fourth intron of *CHRNA3* was significantly associated with hypertension risks in the hypogonadic males. We further showed that EPHB6 physically interacted with CHRNA3 in AGCCs, providing a molecular basis for nAChR being in the EPHB6 signaling pathway.

Keywords: CHRNA3, acetylcholine receptor, adrenal gland chromaffin cells, single nucleotide variant, hypogonadism, hypertension

OPEN ACCESS

Edited by:
Jingyun Yang,
Rush University Medical Center,
United States

Reviewed by:
Chuntao Zhao,
Cincinnati Children's Hospital Medical Center, United States
Paulina Carmona-Mora,
University of California, Davis,
United States

***Correspondence:**
Jian-Zhong Sheng
shengjz@zju.edu.cn
Shen-Jiang Hu
s0hu0001@hotmail.com
Hongyu Luo
hongyu.luo@umontreal.ca
Jiangping Wu
jianping.wu@umontreal.ca

† These authors have contributed equally to this work

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

Received: *02 March 2020*
Accepted: *10 November 2020*
Published: *27 November 2020*

Citation:
Wu T, Wang Y, Shi W, Zhang B-Q, Raelson J, Yao Y-M, Wu H-D, Xu Z-X, Marois-Blanchet F-C, Ledoux J, Blunck R, Sheng J-Z, Hu S-J, Luo H and Wu J (2020) A Variant in the Nicotinic Acetylcholine Receptor Alpha 3 Subunit Gene Is Associated With Hypertension Risks in Hypogonadic Patients. Front. Genet. 11:539862. doi: 10.3389/fgene.2020.539862

Abbreviations: ACh, acetylcholine; AChR, acetylcholine receptors; AGCCs, adrenal gland chromaffin cells; BK channel, large-conductance calcium-activated potassium channel; BMI, body mass index; C.H.B., Han Chinese in Beijing; CHRNA3, nicotinic acetylcholine receptor alpha 3 subunit; EFNs, ephrins; EPHs, erythropoietin-producing hepatocellular kinases; KO, knockout; LD, linkage-disequilibrium; mAChR, muscarinic AChR; PSA, prostate-specific antigen; SNVs, single-nucleotide variants; VSMC: vascular smooth muscle cell; WT, wild type.

INTRODUCTION

Erythropoietin-producing hepatocellular kinases (EPHs) comprise the largest family of receptor tyrosine kinases (Eph Nomenclature Committee, 1997). Their ligands are called ephrins (EFNs), which are also cell surface molecules. When EPH and EFN on neighboring cells interact, both molecules are capable of transducing signals into their respective cells. Interactions between EPHs and EFNs are promiscuous. One EPH can interact with multiple EFNs and vice versa. In general, EPHA members bind preferentially to EFNA members, and EPHB members, to EFNB members (Pasquale, 2008).

Erythropoietin-producing hepatocellular kinases and EFNs have functions in many biological systems and processes, such as the nervous (Flanagan and Vanderhaeghen, 1998; Wilkinson, 2000), the vascular (Kuijper et al., 2007), the digestive (Batlle et al., 2002), the immune (Wu and Luo, 2005), and the endocrine systems (Konstantinova et al., 2007). Our lab is the first to report the function of EPHs and EFNs in the immune system (Luo et al., 2001, 2002). We have recently discovered the previously unknown functions of EPHs and EFNs in regulating blood pressure (Luo et al., 2012; Wu et al., 2012; Wang et al., 2015). Our human genetics studies have provided corroborating evidence that five single-nucleotide variants (SNVs) in the EFNB2 (Wang et al., 2016) and two SNVs in the EFNB3 (Tremblay et al., 2017) are significantly associated with human hypertension in a sex-specific manner.

The deletion of EPHB6 results in increased vascular smooth muscle cell (VSMC) contractility but does not affect blood pressure in male *Ephb6* knockout (KO) mice unless the mice are castrated: *Ephb6* KO plus castration leads to increased blood pressure (Luo et al., 2012). We have shown that, in addition to its effects on VSMCs, EPHB6 also regulates catecholamine biosynthesis (Shi et al., 2019) and secretion by adrenal gland chromaffin cells (AGCCs) (Luo et al., 2012). Male *Ephb6* KO mice have reduced 24-hour urine catecholamine levels, counteracting the outcome of increased VSMC contractility, and the sum of these two effects results in normal blood pressure. The castration of male *Ephb6* KO mice leads to a return to normal levels of catecholamine secretion (Luo et al., 2012). This change, concomitantly with enhanced VSMC contractility, results in blood pressure elevation in the castrated *Ephb6* KO mice. These results suggest that EPHB6 and male sex hormones act in concert to regulate catecholamine secretion and blood pressure.

Our further investigation revealed that AGCCs from male *Ephb6* KO mice presented reduced acetylcholine (ACh)-triggered Ca^{2+} influx, depending on the non-genomic effects of testosterone (Wang et al., 2018). We demonstrated that the decreased Ca^{2+} influx resulted from enhanced large-conductance calcium-activated potassium (BK) currents in these cells (Wang et al., 2018).

In addition to the abnormality of BK channels in the *Ephb6*-KO AGCCs, EPHB6 might regulate other ion channels such as the ACh receptor (AChR), whose opening is the first event in triggering catecholamine release in AGCC. There are two types of AChRs, nicotinic AChR (nAChR) and muscarinic AChR, based on their agonists (nicotine and muscarine). nAChR is a ligand-gated ion channel mediating fast responses of ACh (Itier and Bertrand, 2001), while muscarinic AChR (mAChR) is a G-protein-coupled metabotropic receptor with a slower signaling pace (Eglen, 2006).

nAChR is a pentamer with five homomeric or heteromeric subunits, depending on species and cell types. The subunits consist of α, β, δ. ε and γ types. The α type has ten members (α1–10), β type five members (β1–5). In the mouse AGCCs, α1, α3, α4, α7, α9, β1, β2, and β4 are detectable, with α3 and β4 being most abundant (Mousavi and Nordberg, 2006; Wu et al., 2010), similar to that in rat and bovine AGCCs (Campos-Caro et al., 1997; Di Angelantonio et al., 2003). According to subunit-specific blocker/agonists and the measurement of mRNA, α3b4 nAChR is also present and functional in human AGCCs (Mousavi et al., 2001; Perez-Alvarez et al., 2012).

Multiple proteins can interact with nAChR and positively or negatively modulate its function (Jones et al., 2010). EPHB6 might be one such regulator, and nAChR might lie in the EPHB6 signaling pathway explaining its role in blood pressure control. If so, variants in nAChR subunits should be associated with hypertension risks in a testosterone-dependent way, similar to those of EPHB6. In this study, we conducted a human genetic study to assess the association of variants in CHRNA3, the major subunit of nAChR in AGCCs, with hypertension risks in hypogonadic, hypertensive patients. We also investigated the possible physical interaction of EPHB6 with nAChR.

MATERIALS AND METHODS

Patient Population

The details of the patient population were previously described (Wu et al., 2018) but are presented here again for the convenience of readers. A total of 4,480 male patients ≥40 years old from the Cardiology Ward, Endocrinology Ward, and Physical Examination Center of First Affiliated Hospital, College of Medicine, Zhejiang University in Hangzhou, China, were recruited for this study. They were tested for total plasma testosterone levels. Those with hypogonadism [plasma total testosterone levels <346 ng/dL, the International Society's cut-off level recommended for the Study of the Aging Male (Traish et al., 2011)] were retained. Among those hypogonadic individuals, 982 were diagnosed with primary hypertension and were considered as cases. The primary hypertension phenotype was defined as having recorded systolic pressure >140 mm Hg or diastolic pressure >90 mmHg or having been actively treated for hypertension, excluding known conditions or medication use that could cause BP to increase. All other medical conditions were allowed as long as they were not likely to cause hypertension. All types of medication in the last 3 months were allowed except those that are known to affect testosterone levels, such as testosterone replacement therapy. Seven hundred and eighty-eight normotensive hypogonadic patients among the cohort were retained as controls. Seven hundred and fifty cases and 750 controls were selected with an attempt to match their ages as closely as possible. The systolic and diastolic BP and the values of potential covariate parameters (i.e., age, plasma testosterone

level, heart rates, body mass index (BMI), serum uric acid levels, and smoking status), which could be implicated hypertension risks, of the case and control groups are shown in **Supplementary Tables 1**, **2**.

Blood Sample Collection and Plasma Total Testosterone Measurements

Venous blood samples were drawn from all subjects after an overnight fast of at least 8 h. Five mL of blood was collected into vacuum tubes with the anticoagulant EDTA-K^+ and centrifuged at the collection site within 1 h. Cell pellets were frozen until DNA extraction.

Plasma total testosterone levels were measured with Siemens Immulite 2000 Total Testosterone Kits on Siemens Immulite 2000 Immunoassay Analyzer according to the manufacturer's protocols.

DNA Extraction and Purification

Sample DNA was extracted using DNeasy Blood and Tissue Kit (Cat. 69506, QIAGEN, Hilden, Germany) according to the manufacturer's instructions. Purified DNA quantity and quality were assessed by Qubit®2.0 Fluorometer (Q32866, Invitrogen, Carlsbad, CA, United States) and 1% agarose gel electrophoresis. Samples with DNA quantity $\geq$ 2 μg and optical density (OD) 260 nm/280 nm = 1.8–2.0 were submitted to SNV assay.

Candidate Tag Single Nucleotide Variants (SNVs)

Tag SNVs were chosen from the chromosomal regions at 15q25.1 between positions 78,837,801 and 78,959,037 (Build 37 /hg 19), containing the *CHRNA3* gene and 50 kb on either side of the gene, using the Tagger Program (De Bakker et al., 2005) with Han Chinese in Beijing (CHB) linkage-disequilibrium (LD) data. Tag SNVs were chosen with minimum $r^2 > 0.80$ and minor allele frequency >0.05. Candidate tag SNVs were then submitted for analysis by Illumina Software for compatibility with the GoldenGate multiplexing process. Alternative Tag SNVs were chosen for those Tag SNVs determined to be incompatible with the multiplexing technology, and the new Tag SNVs were then re-submitted for a new Illumina software analysis to determine the compatibility of the new set of SNVs including previously compatible SNVs and the new alternative Tag SNVs. This process continued iteratively until all Tag SNVs were found to be compatible for use by the multiplexing technology. Ultimately, 14 tag SNVs were chosen for this region. The *Bonferroni*-corrected critical *p*-values (*Pcrit* = 0.05/14 = 0.0037) were calculated for SNVs with the hypertension association test, assuming tag SNVs represent independent statistical tests, when performing association analysis across the region.

Tag SNV Genotyping

The tag SNVs were genotyped by the Shanghai Biotechnology Corporation using the Illumina GoldenGate genotyping platform according to the manufacturer's instructions. Those SNVs with a call rate of less than 90% were filtered out and not analyzed.

Association Analysis

The genotyped SNVs were tested for Hardy-Weinberg equilibrium and were analyzed for association with hypertensive versus normotensive status using the PLINK program (Purcell et al., 2007) and a logistic regression model both with and without covariates, the choice of which is described in the Results section. Each SNV was assigned a reference allele and an alternative allele, the SNV of the reference allele being given a numeric code of 1, and the alternative allele, 0. The sum of the numeric codes of the whole reference allele in each individual was used as the genotype value. The sums of the genotype values of all cases and controls were calculated, respectively, and were entered into the logistic regression equation as the genotype terms for each group. This approach is analogous to an additive genetic model.

Reverse Transcription-Quantitative Polymerase Chain Reaction (RT-qPCR)

mRNA levels of *Chrna3* and *Chrnb4* in the mouse adrenal glands were measured by RT-qPCR. Total RNA from adrenal glands of male *Ephb6* KO and WT mice was extracted with TRIzol® (Invitrogen, Burlington, ON, Canada) and reverse-transcribed with iScript™ cDNA Synthesis Kit (Bio-Rad Laboratories (Canada) Ltd., Mississauga, ON, Canada). The sequences of qPCR primers used were listed in **Table 1**. The qPCR condition was as follows: 2 min at 50°C, 2 min at 95°C, followed by 40 cycles of 10 s at 94°C, 20 s at 58°C, and 20 s at 72°C. β-actin mRNA levels were used as internal controls. Data from 20 to 30 cycles of amplification were used. Samples were assayed in duplicate. Data were expressed as signal ratios of target RNA/β-actin mRNA.

Primary AGCC Culture

The adrenal glands from 8- to 10-week old mice were isolated, and the fat and the cortex were removed from these glands. Papain (P4762, Sigma-Aldrich, Oakville, ON, Canada) was activated with 5 mM L-cysteine. Adrenal gland medullae were digested with papain in Hank's buffer (2 medullae/100 ml Hank's buffer containing 4 units of activated papain) at 37°C for 25 min. They were washed twice with Hank's buffer and then triturated by pipetting in 300 ml Hank's buffer until they became feather-like. Cells were pelleted at 3,700 *g* for 3 min and resuspended in DMEM containing 15% FCS for culture.

Immunoprecipitation and Immunoblotting

HEK293 cells were co-transfected with plasmid pCEP4-HA-m*Ephb6* (expressing the HA-tagged mouse EPHB6 intracellular domain (aa 2,268 to 3,537)) and plasmid MR226689 (OriGene, Rockville, MD, United States; expressing Myc-tagged mouse CHRNA3), or control empty pCMV6-Entry Vector (PS100001; OriGene). After 48 h, the cells were lysed using radioimmunoprecipitation assay buffer (RIPA), which contained a cocktail of protease inhibitors (Roche Applied Science, Meylan, France). The lysates were pre-cleared with 30 μl protein G-agarose beads (17061801; GE Healthcare Life Sciences, Mississauga, ON, Canada) and then precipitated with mouse anti-Myc mAb (sc-40; Santa Cruz Biotechnology, Dallas,

TABLE 1 | RT-qPCR primer sequences.

Gene	Sense sequences	Antisense sequences
b-actin	5′-TCGTACCACAGGCATTGTGATGGA-3′	5′-TGATGTCACGCACGATTTCCCTCT-3′
Chrna3	5′-CTGGTGAAGGTGGATGAAGTAA-3′	5′-GGTAGTCAGAGGGTTTCCATTT-3′
Chrnb4	5′-CTGGGTTGTAGTGGGATGATATG-3′	5′-GGCTGACTGCCAATAGTCTTAG-3′

United States) or normal mouse IgG (sc-2025; Santa Cruz Biotechnology) plus protein G-agarose beads at 4°C with gentle rotation overnight. The protein-bead complexes were mixed with SDS-PAGE-loading buffer and boiled for 5 min. Samples were resolved in 10% SDS-PAGE and transferred to PVDF membranes, which were blotted with rabbit anti-HA mAb (3724; Cell Signaling Technology, Danvers, MA, United States), mouse anti-Myc Ab (sc-40; Santa Cruz Biotechnology), and rabbit anti-b-actin Ab (4967; Cell Signaling Technology). Blots were washed and incubated with fluorophore-conjugated secondary Abs (IRDye®680RD-conjugated donkey anti-mouse IgG Ab, P/N 925-68072, LI-COR Biosciences, Lincoln, NE, United States; or IRDye®800CW-conjugated donkey anti-rabbit IgG Ab, P/N 925-32213, LI-COR Biosciences). All the Abs were used at the manufacturer's recommended dilutions. Signals were visualized by the Odyssey infrared imaging system (LI-COR Biosciences).

Immunofluorescence Microscopy

Adrenal gland chromaffin cells were cultured in 6-well plates with cover glass placed at the bottom of the wells. After one day, the cells were washed once with PBS and fixed with paraformaldehyde (4%) for 20 min. The cells were blocked with 10% FBS in PBS for 20 min and then incubated with goat anti-mouse EPHB6 Ab (AF611; R&D Systems, Minneapolis, MN, United States) and rabbit anti-mouse CHRNA3 Ab (ABN281; Millipore, Etobicoke, ON, Canada) overnight at 4°C. Cells were then reacted with Alexa Fluor-488-conjugated donkey anti-goat IgG Ab (A-11055; Thermo Fisher Scientific, Waltham, MA, United States) for EPHB6 and with rhodamine-conjugated goat anti-rabbit IgG Ab (31670; Thermo Fisher Scientific) for CHRNA3 at room temperature for 2 h. The cells were embedded with ProLong® Gold anti-fade reagent (Invitrogen). Cell staining was examined with a Zeiss microscope.

Fluorescence Resonance Energy Transfer Microscopy (FRET)

Adrenal gland chromaffin cells were cultured in 6-well plates with cover glass placed at the bottom of the wells. After one day, the cells were washed once with PBS and fixed with paraformaldehyde (4%) for 20 min. The cells were blocked with 10% FBS in PBS for 20 min and then incubated with goat anti-mouse EPHB6 Ab (AF611; R&D Systems, Minneapolis, MN, United States) and rabbit anti-mouse CHRNA3 Ab (ABN281; Millipore, Etobicoke, ON, Canada) overnight at 4°C. The cells were then reacted with Alexa Fluor-488-conjugated donkey anti-goat IgG Ab (A-11055; Thermo Fisher Scientific, Waltham, MA, United States) for EPHB6 and with rhodamine-conjugated goat anti-rabbit IgG Ab (31670; Thermo Fisher Scientific) for CHRNA3 at room temperature for 2 h. The cells were embedded with ProLong® Gold anti-fade reagent (Invitrogen). The FRET signal was examined under a Leica TCS SP5 laser-scanning confocal microscope (Leica Microsystems Inc., Concord, ON, Canada). Rhodamine was the donor fluorophore, and Alexa Fluor-488, the acceptor fluorophore.

Fluorescence resonance energy transfer microscopy was measured with acceptor photobleaching (AB) by FRET AB Wizard software (Leica Microsystems Inc.). All necessary controls for AB, such as cells with relevant single fluorescence staining, were performed to satisfy background deductions in calculating FRET efficiency, as required by the software.

Acceptor photobleaching FRET efficiency was calculated by fluorescence intensity of the donor before (D^{pre}) and after (D^{post}) acceptor-selective photobleaching, according to the following formula:

$$\text{AB FRET efficiency} = (D^{post} - D^{pre}) / D^{post}$$

RESULTS

We used 750 hypogonadic men with hypertension as cases and 750 age-matched normotensive hypogonadic men as controls to assess the association of *CHRNA3* variants with hypertension. The mean measured blood pressure (systolic as well as diastolic) was significantly higher in the cases (125.42 ± 0.54 /76.45 ± 0.36 mmHg, mean $\pm$ SE) than the controls ($116.6 + 0.43$ /$72.06 + 0.32$ mmHg) ($p = 2.2 \times 10^{-16}$) for both systolic and diastolic pressures (**Supplementary Table 1**), but the case blood pressure did not reach the hypertension diagnostic criteria, i.e., systolic pressure >140 mmHg and/or diastolic pressure >90 mmHg. This was due to the fact that the majority of the cases (93.7%) were previously diagnosed and were under anti-hypertension medication, which controlled their measured blood pressure within the normal range.

We performed a covariate analysis to test the effects of non-genetic factors in the determination of case and control status (hypertension) in our sample (**Supplementary Table 2**). We conducted simple logistic regression analysis on several phenotype candidate covariates that we suspected might differ between cases and controls (variable 1 for cases, 0 for controls). These were age, plasma testosterone levels, heart rates, body mass index (BMI), serum uric acid levels, and smoking status. Each regression was done for a single phenotype covariate value only without any genotype factor included.

Age, heart rate, BMI, and serum uric acid levels were significantly different between cases and controls in the simple regression analyses and were retained for further analysis. In addition, the testosterone level was nearly significant at $p = 0.05$, so it is also included since the assumption of hypogonadism in the whole cohort is central to this analysis.

Realizing that some of these covariates could be correlated among themselves, we attempted to consider their combined effect on case and control status. In order to do so, we performed a multiple logistic regression with all of the retained candidate covariates but without any genotype factor. All of the retained covariates from the simple regression analyses, when combined in a multiple regression model without a genotype factor, remained significantly different between cases and controls, again with the exception of testosterone levels, which were nearly significant ($p = 0.0506$). (**Supplementary Table 3**). Therefore, we retained age, heart rate, BMI, serum uric acid levels, and blood testosterone levels in the genetic regression models subsequently performed for each SNV separately. All covariate factors were included in each of the genetic multiple regression analysis. There may have been slight differences in covariate measures for each SNV due only to differences in missing values, but this effect will have been slight as the data genotype completeness was controlled and variation kept within strict limits. Essentially, the covariate effects remained constant for all SNV analyses, and the significance of each genotype term was tested once the consistent covariate effects were removed.

The genetic association was tested for all SNVs using logistic regression models, both with and without covariates. SNV *rs3743706* (chromosome 15, position 78,909,227; build 37/hg 19), which is intronic to *CHRNA3* lying 138 bp 5′ of exon 5, was significant at *p*-values below the critical multiple testing *p*-value *(p-crit)* of 0.0037 for models both with and without covariates ($p = 0.00190$ and $p = 0.00143$, respectively). **Table 2** presents the genotyped tag SNVs along with their positions, reference alleles, minor alleles, minor allele frequencies, odds ratios with respect to the reference alleles, standard errors of those odds ratios, and the *p*-values for the significance of the odds ratios for association in the models with or without covariates. The SNV with a *p*-value below *p*-crit along with its associated parameters is in bold. **Figure 1** shows LocusZoom plots illustrating association *p*-values for all the 14 SNVs calculated without or with covariates, with *rs3743706* showing the highest $-\log_{10}$ *p*-value. The plots also illustrate the location of this significant SNV in the *CHRNA3* and the recombination rate in the analyzed region. Analysis with covariates did not reduce the *p*-value of *rs3743706*, compared to the analysis without covariates, suggesting that this SNV's association with hypertension is independent of the phenotypes used in the logistic multiple regression model.

The results of the human genetic study were compatible with our hypothesis that AChR in AGCCs lies within the EPHB6 signaling pathway. If so, EPHB6 might also regulate AChR expression or interact with AChR directly to regulate its function. We first investigated whether *Ephb6* KO resulted in altered expression of the major nAChR subunits α3 and β4 in AGCCs. As shown in **Figure 2**, the mRNA expression of these two major subunits in KO AGCCs from male mice remained unchanged, compared to their wild type (WT) counterparts. This excludes the possibility that EPHB6 regulates AChR function by modulating the expression of its major components α3 and β4 at the gene level and favors the possibility that the putative regulation might occur through EPHB6 and AChR interaction at the protein level.

TABLE 2 | Logistic regression association tests for SNVs in the CHRNA3 gene region with and without covariates.

SNV	Marker					Model with Covariates			Model without Covariates		
	Position (Build 7/hg 19)	Reference Allele	Alternate Allele	Minor Allele	Minor Allele Frequency	Odds Ratio	Standard Error of Odds Ratio	*p*-value	Odds Ratio	Standard Error of Odds Ratio	*p*-value
rs4887062	78837801	G	A	G	0.193	1.052	0.091	0.57860	1.045	0.090	0.62210
rs588765	78865425	T	C	T	0.469	1.082	0.115	0.49470	1.076	0.113	0.51810
rs680244	78871288	T	C	T	0.492	1.025	0.115	0.82760	1.018	0.113	0.87200
rs3743078	78894759	C	G	C	0.229	0.980	0.088	0.81740	0.967	0.087	0.69600
rs1317286	78896129	A	G	G	0.105	1.246	0.121	0.06814	1.271	0.118	0.04167
rs11637630	78899719	G	A	G	0.448	0.891	0.073	0.11500	0.897	0.072	0.13090
rs2869546	78907345	C	T	C	0.195	1.196	0.118	0.13100	1.229	0.117	0.07733
rs3743076	**78909227**	**T**	**A**	**T**	**0.257**	**1.307**	**0.086**	**0.00190**	**1.311**	**0.085**	**0.00143**
rs1948	78917399	A	G	A	0.471	1.125	0.074	0.11090	1.132	0.073	0.08789
rs950776	78926018	T	C	C	0.266	0.942	0.103	0.56150	0.908	0.102	0.34360
rs12440014	78926726	C	G	G	0.454	1.039	0.073	0.60040	1.043	0.072	0.56370
rs1021070	78946863	C	G	C	0.148	0.939	0.103	0.54010	0.956	0.101	0.65100
rs7166158	78948753	T	A	T	0.228	1.155	0.088	0.10160	1.153	0.087	0.10100
rs12594550	78959037	C	G	C	0.213	0.922	0.091	0.36610	0.926	0.088	0.37930

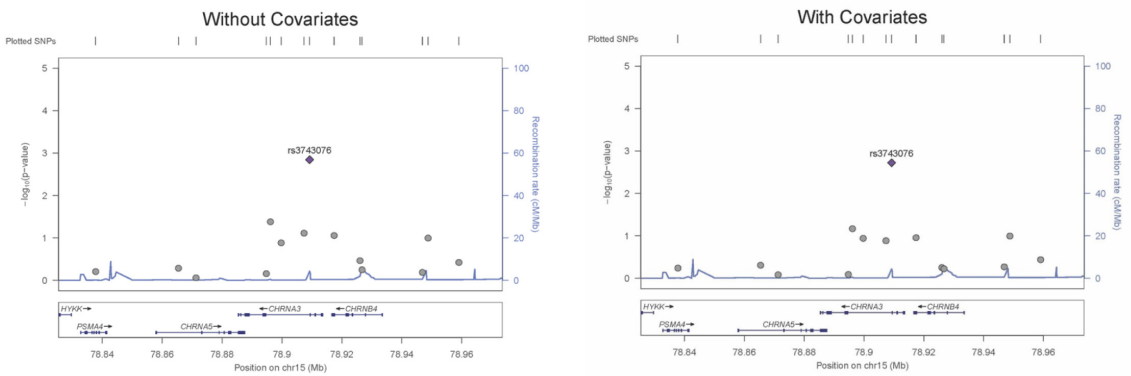


FIGURE 1 | LocusZoom plots for an association of all SNVs assayed across the CHRNA3 gene region for regression models with and without covariates. Left *Y*-axis: $-\text{Log}_{10}$ (*p*-value) of the SNV association with hypertension. Right *Y*-axis: recombination rate (cM/Mb). *X*-axis: the position of SNVs and the *CHRNA3* gene on chromosome 15.

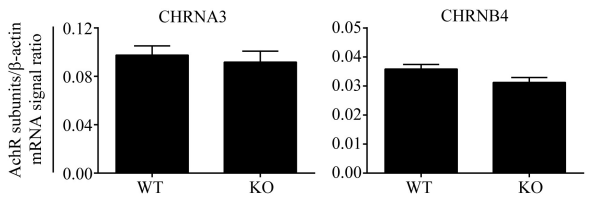


FIGURE 2 | Similar Chrna3 and Chrnb4 mRNA expression levels in AGCCs from WT and EPHB6 KO mice. Mouse *Chrna3* and *Chrnb4* mRNA levels were measured by RT-qPCR, using total RNA from adrenal glands of male *Ephb6* KO and WT mice. β-actin mRNA levels were employed as internal controls. Each sample was analyzed in duplicate. Three independent experiments were conducted, and the pooled data of these experiments were expressed as signal ratios (means ± SE) of target R.N.A./β-actin mRNA. There was no significant difference between WT and *Ephb6* KO AGCCs in their *Chrna3* and *Chrnb4* mRNA levels (two-way paired Student's *t*-test).

We then examined the possible interaction between EPHB6 and AChR. HEK293 cells were transfected with the HA-tagged mouse EPHB6 intra-cellular domain and Myc-tagged mouse CHRNA3. As shown in **Figure 3, EPHB6-HA** and **CHRNA3**-Myc both presented in the lysates of HEK293 cells that were transfected with EPHB6-HA or CHRNA3-Myc, but not in cells transfected with the control plasmid. EPHB6-HA was found in the anti-Myc (for CHRNA3-Myc) precipitates (first lane) of cells transfected with both EPHB6-HA and CHRNA3-Myc expressing plasmids, but not in the control precipitates from cells transfected with Myc-vector (second lane). CHRNA3 (the two major bands indicated by arrows) only appeared in lysates and precipitates from cells transfected with CHRNA3-Myc plasmid (lanes 1 and 5). There were some minor non-specific bands observed in anti-My Ab blotting. The abundance of the non-specific proteins was different in the lysates and immunoprecipitates, so the patterns of the non-specific protein bands were different in lanes 1 and 5. Nevertheless, the two CHRNA3 bands were present in both the lysates and immunoprecipitates. Non-specific Ab did not precipitate CHRNA3-Myc or EPHB6-HA (middle two lanes). Beta-actin immunoblotting was used to show similar input of lysates. For immunoprecipitates, the similar intensity of b-actin bands were present in all the lanes, including those controls of normal IgG and empty vectors. The b-actin in the immunoprecipitates was due to non-specific carryover by the beads, but such non-specific carryover indicated that the absence of EPHB6 signals in the control lanes was not due to lack of or unequal protein inputs. This result indicates that EPHB6 physically interacts with AChR via its major subunit CHRNA3. Since the EPHB6 intracellular domain-expressing plasmid was used in this experiment, the results also suggest that the EPHB6 intracellular domain is responsible and sufficient to bind to CHRNA3.

To assess the EPHB6 and CHRNA3 interaction in a more physiological setting, we examined EPHB6 and CHRNA3 expression in mouse AGCCs using confocal microscopy and found that these two molecules were colocalized (**Figure 4**).

The possible physical interaction between EPHB6 and CHRNA3 in WT mouse AGCCs was further examined using a different approach: acceptor photobleaching (AB) FRET. During AB, fluorescence from the acceptor EPHB6 was selectively depleted without influencing donor CHRNA3 signals. The donor and acceptor fluorescence intensity before and after photobleaching in the bleached region (rectangles in **Figure 5A**) and the unbleached region of AGCCs were quantified. Data from a representative cell are listed in the inset table (**Figure 5B**). After AB, the signal intensity of donor CHRNA3 was significantly increased in the bleached but not in the unbleached area, compared to that observed before photobleaching. This indicates that donor fluorophore energy could not be transferred to the acceptor due to the acceptor fluorophore bleaching, demonstrating that the donor and acceptor existed in proximity. The acceptor fluorophore intensity was reduced in the photobleached but not unbleached area, again as expected. FRET efficiency from CHRNA3 to EPHB6 of areas inside and outside the bleached regions for more than 7–10 randomly selected cells from 3 independent experiments is shown in **Figure 5C**. The results show that CHRNA3 and EPHB6 in mouse AGCCs are in very close

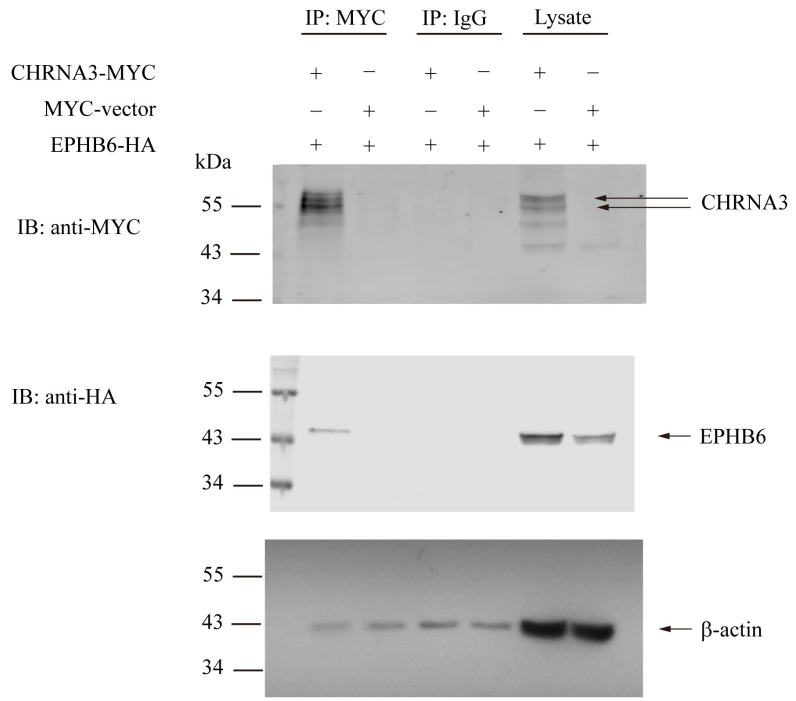


FIGURE 3 | Mouse EPHB6 physically interacts with CHRNA3 according to immunoprecipitation. HEK293 cells were co-transfected with plasmids expressing HA-tagged EPHB6 intracellular domain (aa 2268 to 3537) (EPHB6-HA) and Myc-tagged CHRNA3 (CHRNA-MYC), or an empty control vector (MYC-vector), as indicated. The cell lysates were precipitated with anti-Myc mAb, or normal mouse IgG, as shown. The precipitates were resolved in 10% SDS-PAGE and immunoblotted with anti-Myc (first row) and anti-HA (second row) Abs followed by fluorophore-conjugated secondary Abs. The membranes were also blotted with anti-b-actin Ab (row 3) to show equal loading. Signals were visualized by the Odyssey infrared imaging system (LI-COR Biosciences). The experiments were conducted three times, and representative results are presented. Arrows point to CHRNA3 and EPHB6 bands.

proximity of <10 nm, which is conventionally considered as a direct association.

DISCUSSION

We previously demonstrated that EPHB6 and testosterone jointly regulated adrenal gland catecholamine secretion and blood pressure in mice (Luo et al., 2012). The major event triggering catecholamine secretion in AGCCs from the male EPHB6 KO, Ca^{2+} influx, was compromised, and this was accompanied by decreased catecholamine release (Wang et al., 2018). We found that one of the causes for the compromised Ca^{2+} influx was the earlier closure of the BK channel in AGCCs. This premature closure led to quicker AGCC repolarization, hence, to earlier termination of the Ca^{2+} influx and reduced total Ca^{2+} influx (Wang et al., 2018). We wondered whether EPHB6 also affected the Ca^{2+} influx by interacting with other ion channels. nAChR is a ligand-gated ion channel. The opening of nAChR after ligand Ach binding is the first event in AGCC depolarization, which results in the subsequent large Ca^{2+} influx followed by catecholamine secretion. Thus, nAChR in AGCCs is possibly involved in EPHB6 signaling pathways related to blood pressure regulation. This prompted us to conduct a human genetic study to assess the association of variants of a major nAChR subunit CHRNA3 with hypertension risks in hypogonadic male hypertensive patients. These patients had augmented chances of having detrimental variants in the EPHB6 signaling pathways (including AChR subunits) due to their hypogonadic condition because hypertension was only revealed in castrated *Ephb6* KO mice (Luo et al., 2012). Indeed, SNV *rs3743706* within *CHRNA3* was found to be significantly associated with hypertension risks in this cohort.

The significant SNV, rs3743706, is located in the fourth intron of *CHRNA3*. This SNV might and might not be a functional variant. This SNV itself or nearby functionally active variants in linkage disequilibrium (LD) with it in this intron could be part of an intronic enhancer/repressor regulating CHRNA3 expression levels. Such intronic regulators are not uncommon. For example, HCN4 has a potent enhancer in its first intron controlling its expression (Kuratomi et al., 2009). The genomic sequence from 100-bp upstream to 100-bp downstream of this SNV was surveyed for the presence of possible regulatory elements using software available in the UCSC Genome Browser (EPDnew, EPDnewNC, ENCODE Candidate *Cis*-Regulatory Elements (cCREs), GeneHancer, CpG islands, H3K27Ac Mark, ORegAnno, NCBI RefSeq Functional Elements, Double Elite, Hi-C and Micro-C). No conserved enhancer/repressor motifs or transcription factor-binding sites were found in this region (**Supplementary Figure 1**). We cannot exclude the existence of cryptic and non-canonical regulator elements for which *rs3743706* is a part.

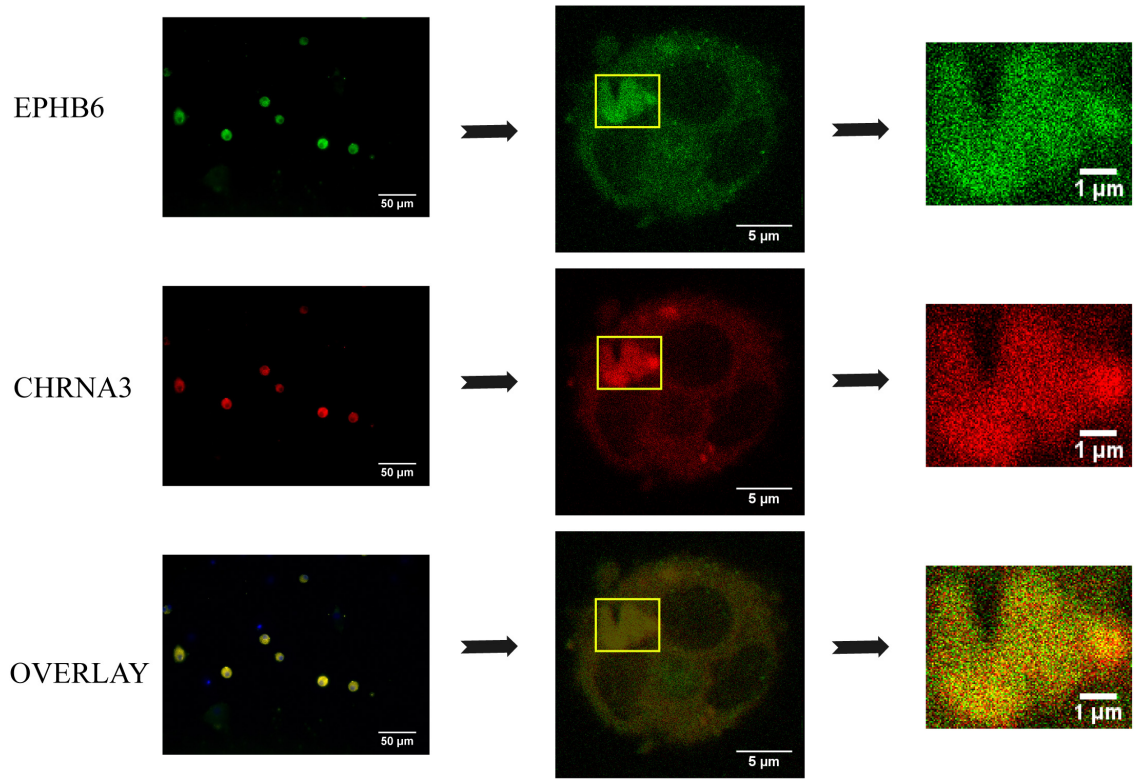


FIGURE 4 | EPHB6 co-localized with CHRNA3 in mouse AGCCs according to confocal microscopy. AGCCs from male WT mice were fixed and stained with anti-EPHB6 and anti-CHRNA3 Abs followed by Alexa Fluor-488-conjugated donkey anti-goat IgG Ab and rhodamine-conjugated goat anti-rabbit IgG Ab, respectively. Three independent experiments were conducted, and micrographs from a representative experiment are shown. The rectangles in the left panels were enlarged at the right. EPHB6 was in pseudo green, and CHRNA3, pseudo red.

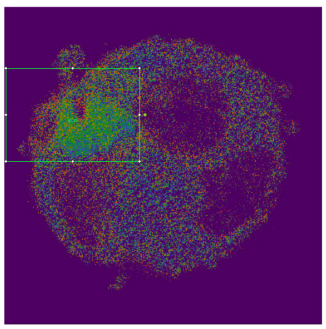
Alternatively, the functional variants could be in LD with this SNV and located in the nearby exons or introns. If so, sequence variants in CHRNA3 protein or enhancer/repressor of its gene might be responsible for the alteration of its function or expression, respectively.

CHRNA3 SNVs have not been found to be significantly associated with hypertension risks in many genome-wide association studies (Wellcome Trust Case Control Consortium, 2007; Adeyemo et al., 2009; Levy et al., 2009; Org et al., 2009; Hiura et al., 2010; Padmanabhan et al., 2010; International Consortium for Blood Pressure Genome-Wide Association Studies et al., 2011; Slavin et al., 2011; Guo et al., 2012). However, testosterone levels are not stratified in these studies. Since the majority of the males are testosterone sufficient in the general population, *CHRNA3* variants would not be associated with blood pressure in the majority of individuals in these studies, while the smaller subpopulation of patients with both hypogonadism and *CHRNA3* variants would be obscured statistically and so could not reach significance in a GWAS even if they were hypertensive, unless samples were stratified according to testosterone levels.

Our genetic study was performed using an ethnic Han cohort in China. Additional studies using a different Han cohort and a cohort of different genetic backgrounds will be needed to confirm and generalize our conclusion.

There are different views regarding whether catecholamines released from the adrenal glands play a critical role in primary hypertension. The adrenal glands produce all three major species of catecholamines: epinephrine, norepinephrine, and dopamine. However, while plasma epinephrine is mainly derived from the adrenal gland, plasma norepinephrine predominantly comes from sympathetic nerve endings. Dopamine levels in plasma are very low. Some early studies showed that epinephrine levels are associated with hypertension in humans and animal models (Axelrod, 1976; Buhler et al., 1982; Dominiak and Grobecker, 1982; Goldstein, 1983; Borkowski and Quinn, 1984; Jablonskis and Howe, 1994). In recent decades, chronic stress (Sparrenberger et al., 2009; Liu et al., 2017) and systemic low-level sterile inflammation (De Miguel et al., 2015) are positively correlated to primary hypertension. Such conditions are normally associated with elevation of plasma norepinephrine, implicating the involvement of the nervous system, although in many cases, augmentation of epinephrine levels is also observed (; Rivier et al., 1989; Pende et al., 1990; Corssmit et al., 1996; Kannan et al., 1996; Esler et al., 2008a,b; Byrne et al., 2018). In a mouse model with genetic manipulation, TRPM4 deletion resulted in increased catecholamine secretion by AGCCs and hypertension

A



B

		Bleached Area	Unbleached Area
Donor (CHRNA3)	Pre-bleaching	45.1	46.22
	Post-bleaching	79.97	50.1
Acceptor (EPHB6)	Pre-bleaching	29.51	34.95
	Post-bleaching	13.64	35.19
	Efficiency	0.436	0.0774

C

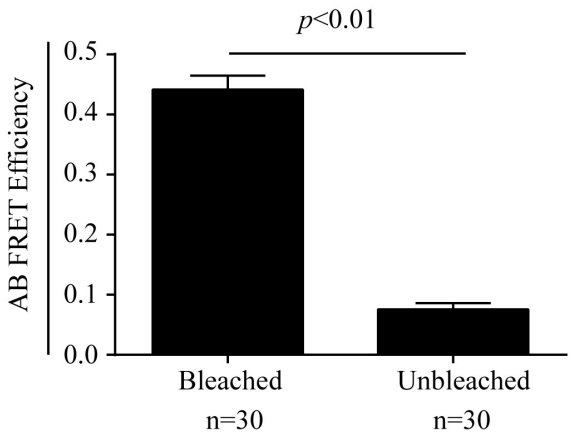


FIGURE 5 | Mouse EPHB6 interacts with CHRNA3 in WT AGCCs according to FRET. FRET was measured with acceptor photobleaching (AB). All the necessary controls for AB, such as cells with relevant single fluorescence staining, were performed to satisfy background deductions in calculating FRET efficiency. AB FRET efficiency was calculated by fluorescence intensity of the donor before (D^{pre}) and after (D^{post}) acceptor-selective photobleaching, according to the following formula: AB FRET efficiency = (D^{post} – D^{pre}) / D^{post}. **(A)** Merged image of an EPHB6-(pseudo-green) and CHRNA3-(pseudo-red) stained AGCC from a WT male mouse. The rectangle indicates the bleached area. **(B)** Donor and acceptor signal measurements and AB FRET efficiency of a representative AGCC from a WT male mouse. **(C)** AB FRET efficiency of bleached and unbleached areas. Data of 7–10 AGCCs from 3 independent experiments were pooled and presented with means ± SE. The differences between the AB FRET efficiency values of bleached areas and unbleached areas were highly significant ($p < 0.01$; Two-way Student's *t*-test).

(Mathar et al., 2010). These data collectively suggest that the catecholamine release from the adrenal glands contributes to primary hypertension.

The role of testosterone in blood pressure is controversial. Testosterone increases blood pressure in many animal studies (Masubuchi et al., 1982; Chen and Meng, 1991; Reckelhoff et al., 1998). However, hypogonadism in humans also increases the risk of hypertension (Hughes et al., 1989; Phillips et al., 1993; Jaffe et al., 1996; Liu et al., 2003; Zitzmann, 2009; Garcia-Cruz et al., 2012). In several human studies, testosterone replacement therapy increases cardiovascular disease risks (Miner et al., 2014; Anaissie et al., 2017), but the effect sizes are too small to be convincing. More importantly, such studies cannot exclude the possibility that testosterone can reduce blood pressure for a subpopulation of males with disease-associated variants in molecules of the EPHB6 signaling pathway (including EPHB6 and CHRNA3) but that such a subpopulation, being a minor one, is obscured by others in statistical analysis in studies without stratification based on testosterone levels.

Our study with stratification of testosterone levels has enhanced the association of variants of molecules in the EPHB6 signaling pathway with hypertension risks and revealed the association of a CHRNA3 variant with hypertension risks. Our results further suggest that for the subpopulation carrying the *rs3743706* SNV, testosterone replacement therapy could be a personalized treatment for the cause of their hypertension. Although testosterone treatment might raise the concern of increased prostate cancer risks, this could be mitigated by routine examination of the patients for elevated serum prostate-specific antigen (PSA) levels. Moreover, a meta-analysis of testosterone replacement therapy outcome has shown no evidence of increased prostate cancer risks and no significant increase in PSA levels (Kang and Li, 2015).

AChR function, which is affected by its assembly, cycling, and clustering in addition to its opening, is known to be modulated by its associated molecules (Christianson and Green, 2004; Dobbins et al., 2008; Jones et al., 2010; Baier et al., 2011). We used three different experimental approaches (i.e., immunoprecipitation, colocalization, and FRET) to show that

EPHB6 physically interacted with CHRNA3 in AGCCs. This provides a molecular basis for EPHB6 to function as a modulator of nAChR. AChR opening is the very first triggering event leading to a subsequent large Ca^{2+} influx, which is required for proper adrenaline secretion. Therefore, the physical interaction between EPHB6 and a dominant AChR subunit CHRNA3 implies the possibility that EPHB6 modulates AChR opening and hence the subsequent Ca^{2+} influx that controls adrenaline secretion.

To demonstrate that such interaction has functional relevance, we need to prove experimentally that KO AGCCs have reduced initial Ca^{2+}/N^{+} influx through nAChR upon acetylcholine stimulation. We spent significant efforts to demonstrate this but were not able to succeed due to the technical difficulties of patching primary AGCCs. We also need to test in human patients with adequate testosterone levels but with loss-of-function variants of EPHB6 and/or its downstream signaling molecules whether less AGCC-derived catecholamines are produced according to 24-hour urine tests.

We previously demonstrated that EPHB6 KO in castrated mice led to hypertension (Luo et al., 2012). This prior knowledge plus our current findings from the human genetic study and mouse AGCC experiments allow us to propose the following model for the implication of the regulation of AGCC function by CHRNA3 and EPHB6 in human hypertension pathogenesis.

EPHB6 and molecules in its signaling pathways (e.g., CHRNA3) normally have a positive effect on catecholamine secretion by AGCCs. The results of our current study suggest that such a positive effect could be exerted by two different mechanisms. (1) A proper expression or function of molecules (e.g., CHRNA3) in the EPHB6 signaling pathway; or (2) an appropriate direct interaction at the protein level between EPHB6 and its signaling molecules (e.g., the physical interaction between EPHB6 and CHRNA3). Compromises in either of these two mechanisms will reduce catecholamine secretion by AGCCs, and such reduction has a protective effect against hypertension. The significant variant found in the CHRNA3 gene is related to mechanism 1. The physical interaction between EPHB6 and CHRNA3 in AGCCs implicates mechanism 2. Such a protective effect depends on the presence of sufficient testosterone, and hence only occurs in testosterone-sufficient males (Wang et al., 2018). At the same time, EPHB6 has other targets (e.g., vascular smooth muscle cells), and EPHB6 KO results in increased blood vessel contractility (Luo et al., 2012). Although we can only speculate at this time, the deleterious CHRNA3 variant might also have other targets leading to increased blood pressure. As a result, the protective effect due to the reduced catecholamine secretion is neutralized by the above-described blood increasing effects. Hence, males with deleterious variants in EPHB6 or its signaling molecules (e.g., CHRNA3) have normal BP. However, when their testosterone level is decreased due to either old ages or some pathological conditions, the protective effect due to reduced catecholamine secretion is lost. This is evidenced in our previous report that the castrated EPHB6 KO mice presented hypertension (Luo et al., 2012). Consistent with this hypothesis, we demonstrated that hypogonadic patients with deleterious CHRNA3 variants had increased hypertension risks according to our human genetic study.

Molecules in the EPHB6 signaling pathway (including EPHB6, EFNBs, AChR, BK, etc.) are also expressed in catecholamine-producing neuronal cells in the nervous system. Whether their sequence variants alone or in concert with sex hormones compromise catecholamine release and consequently modify the pathophysiology of diseases caused by dysfunctional catecholamine secretion (e.g., Parkinson's disease) is an area worth investigating.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusion of this article will be made available by the authors, without undue reservation.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Ethics Committee of the First Affiliated Hospital of Zhejiang University (No. 2013-145). The patients/participants provided their written informed consent to participate in this study. The animal study was reviewed and approved by Animal Protection Committee of CHUM Research Center. Written informed consent was obtained from the owners for the participation of their animals in this study.

AUTHOR CONTRIBUTIONS

JW, HL, S-JH, and J-ZS generated the concept and initiated this project. TW, YW, WS, B-QZ, Y-MY, H-DW, Z-XX, JW, J-ZS, S-JH, JL, and RB conducted the experiments. JR and FM-B performed the genetic data analysis. YW and WS analyzed the animal data. TW, JR, YW, WS, HL, and JW drafted the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by the Fonds de recherche du Québec – Santé, the Natural Sciences and Engineering Research Council of Canada (RGPIN-2017-04790), the Canadian Institutes of Health Research (MOP 123389), and the J.-Louis Lévesque Foundation to JW. This study was also made possible by a group grant from the National Sciences Foundation of China (#81361120264) to J-ZS, S-JH, TW, and JW, and a grant from the National Natural Science Foundation of China (#81700363) to YW.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.539862/full#supplementary-material

REFERENCES

Adeyemo, A., Gerry, N., Chen, G., Herbert, A., Doumatey, A., Huang, H., et al. (2009). A genome-wide association study of hypertension and blood pressure in African Americans. *PLoS Genet.* 5:e1000564. doi: 10.1155/2018/7259704

Anaissie, J., Roberts, N. H., Wang, P., and Yafi, F. A. (2017). Testosterone replacement therapy and components of the metabolic syndrome. *Sex Med. Rev.* 5, 200–210. doi: 10.1016/j.sxmr.2017.01.003

Axelrod, J. (1976). Catecholamines and hypertension. *Clin. Sci. Mol. Med. Suppl.* 3, 415s–421s.

Baier, C. J., Fantini, J., and Barrantes, F. J. (2011). Disclosure of cholesterol recognition motifs in transmembrane domains of the human nicotinic acetylcholine receptor. *Sci. Rep.* 1:69.

Batlle, E., Henderson, J. T., Beghtel, H., Van Den Born, M. M., Sancho, E., Huls, G., et al. (2002). Beta-catenin and TCF mediate cell positioning in the intestinal epithelium by controlling the expression of EphB/ephrinB. *Cell* 111, 251–263. doi: 10.1016/s0092-8674(02)01015-2

Borkowski, K. R., and Quinn, P. (1984). Adrenaline and the development of genetic hypertension. *J. Hypertens. Suppl.* 2, S81–S83.

Buhler, F. R., Amann, F. W., Bolli, P., Hulthen, L., Kiowski, W., Landmann, R., et al. (1982). Elevated adrenaline and increased alpha-adrenoceptor-mediated vasoconstriction in essential hypertension. *J. Cardiovasc. Pharmacol.* 4(Suppl. 1), S134–S138.

Byrne, C. J., Khurana, S., Kumar, A., and Tai, T. C. (2018). Inflammatory signaling in hypertension: regulation of adrenal catecholamine biosynthesis. *Front. Endocrinol.* 9:343. doi: 10.3389/fendo.2018.00343

Campos-Caro, A., Smillie, F. I., Dominguez Del Toro, E., Rovira, J. C., Vicente-Agullo, F., Chapuli, J., et al. (1997). Neuronal nicotinic acetylcholine receptors on bovine chromaffin cells: cloning, expression, and genomic organization of receptor subunits. *J. Neurochem.* 68, 488–497. doi: 10.1046/j.1471-4159.1997.68020488.x

Chen, Y. F., and Meng, Q. C. (1991). Sexual dimorphism of blood pressure in spontaneously hypertensive rats is androgen dependent. *Life Sci.* 48, 85–96. doi: 10.1016/0024-3205(91)90428-e

Christianson, J. C., and Green, W. N. (2004). Regulation of nicotinic receptor expression by the ubiquitin-proteasome system. *EMBO J.* 23, 4156–4165. doi: 10.1038/sj.emboj.7600436

Corssmit, E. P., Heijligenberg, R., Endert, E., Ackermans, M. T., Sauerwein, H. P., and Romijn, J. A. (1996). Endocrine and metabolic effects of interferon-alpha in humans. *J. Clin. Endocrinol. Metab.* 81, 3265–3269. doi: 10.1210/jc.81.9.3265

De Bakker, P. I., Yelensky, R., Pe'er, I., Gabriel, S. B., Daly, M. J., and Altshuler, D. (2005). Efficiency and power in genetic association studies. *Nat. Genet.* 37, 1217–1223. doi: 10.1038/ng1669

De Miguel, C., Rudemiller, N. P., Abais, J. M., and Mattson, D. L. (2015). Inflammation and hypertension: new understandings and potential therapeutic targets. *Curr. Hypertens. Rep.* 17:507.

Di Angelantonio, S., Matteoni, C., Fabbretti, E., and Nistri, A. (2003). Molecular biology and electrophysiology of neuronal nicotinic receptors of rat chromaffin cells. *Eur. J. Neurosci.* 17, 2313–2322. doi: 10.1046/j.1460-9568.2003.02669.x

Dobbins, G. C., Luo, S., Yang, Z., Xiong, W. C., and Mei, L. (2008). alpha-Actinin interacts with rapsyn in agrin-stimulated AChR clustering. *Mol Brain* 1:18. doi: 10.1186/1756-6606-1-18

Dominiak, P., and Grobecker, H. (1982). Elevated plasma catecholamines in young hypertensive and hyperkinetic patients: effect of pindolol. *Br. J. Clin. Pharmacol.* 13, 381S–390S.

Eglen, R. M. (2006). Muscarinic receptor subtypes in neuronal and non-neuronal cholinergic function. *Auton. Autacoid. Pharmacol.* 26, 219–233. doi: 10.1111/j.1474-8673.2006.00368.x

Eph Nomenclature Committee (1997). Unified nomenclature for Eph family receptors and their ligands, the ephrins. *Cell* 90, 403–404. doi: 10.1016/s0092-8674(00)80500-0

Esler, M., Eikelis, N., Schlaich, M., Lambert, G., Alvarenga, M., Dawood, T., et al. (2008a). Chronic mental stress is a cause of essential hypertension: presence of biological markers of stress. *Clin. Exp. Pharmacol. Physiol.* 35, 498–502. doi: 10.1111/j.1440-1681.2008.04904.x

Esler, M., Schwarz, R., and Alvarenga, M. (2008b). Mental stress is a cause of cardiovascular diseases: from scepticism to certainty. *Stress Health* 24, 175–180. doi: 10.1002/smi.1198

Flanagan, J. G., and Vanderhaeghen, P. (1998). The ephrins and Eph receptors in neural development. *Annu. Rev. Neurosci.* 21, 309–345. doi: 10.1146/annurev.neuro.21.1.309

Garcia-Cruz, E., Piqueras, M., Huguet, J., Perez-Marquez, M., Gosalbez, D., Peri, L., et al. (2012). Hypertension, dyslipidemia and overweight are related to lower testosterone levels in a cohort of men undergoing prostate biopsy. *Int. J. Impot. Res.* 24, 110–113. doi: 10.1038/ijir.2011.55

Goldstein, D. S. (1983). Plasma catecholamines and essential hypertension. An analytical review. *Hypertension* 5, 86–99. doi: 10.1161/01.hyp.5.1.86

Guo, Y., Tomlinson, B., Chu, T., Fang, Y. J., Gui, H., Tang, C. S., et al. (2012). A genome-wide linkage and association scan reveals novel loci for hypertension and blood pressure traits. *PLoS One* 7:e31489. doi: 10.1371/journal.pone.0031489

Hiura, Y., Tabara, Y., Kokubo, Y., Okamura, T., Miki, T., Tomoike, H., et al. (2010). A genome-wide association study of hypertension-related phenotypes in a Japanese population. *Circ. J.* 74, 2353–2359. doi: 10.1253/circj.cj-10-0353

Hughes, G. S., Mathur, R. S., and Margolius, H. S. (1989). Sex steroid hormones are altered in essential hypertension. *J. Hypertens.* 7, 181–187.

International Consortium for Blood Pressure Genome-Wide Association Studies, Ehret, G. B., Munroe, P. B., Rice, K. M., Bochud, M., Johnson, A. D., et al. (2011). Genetic variants in novel pathways influence blood pressure and cardiovascular disease risk. *Nature* 478, 103–109. doi: 10.1038/nature10405

Itier, V., and Bertrand, D. (2001). Neuronal nicotinic receptors: from protein structure to function. *FEBS Lett.* 504, 118–125. doi: 10.1016/s0014-5793(01)02702-8

Jablonskis, L. T., and Howe, P. R. (1994). Elevated plasma adrenaline in spontaneously hypertensive rats. *Blood Press.* 3, 106–111. doi: 10.3109/08037059409101529

Jaffe, A., Chen, Y., Kisch, E. S., Fischel, B., Alon, M., and Stern, N. (1996). Erectile dysfunction in hypertensive subjects. Assessment of potential determinants. *Hypertension* 28, 859–862. doi: 10.1161/01.hyp.28.5.859

Jones, A. K., Buckingham, S. D., and Sattelle, D. B. (2010). Proteins interacting with nicotinic acetylcholine receptors: expanding functional and therapeutic horizons. *Trends Pharmacol. Sci.* 31, 455–462. doi: 10.1016/j.tips.2010.07.001

Kang, D. Y., and Li, H. J. (2015). The effect of testosterone replacement therapy on prostate-specific antigen (PSA) levels in men being treated for hypogonadism: a systematic review and meta-analysis. *Medicine* 94:e410. doi: 10.1097/md.0000000000000410

Kannan, H., Tanaka, Y., Kunitake, T., Ueta, Y., Hayashida, Y., and Yamashita, H. (1996). Activation of sympathetic outflow by recombinant human interleukin-1 beta in conscious rats. *Am. J. Physiol.* 270, R479–R485.

Konstantinova, I., Nikolova, G., Ohara-Imaizumi, M., Meda, P., Kucera, T., Zarbalis, K., et al. (2007). EphA-Ephrin-A-mediated beta cell communication regulates insulin secretion from pancreatic islets. *Cell* 129, 359–370. doi: 10.1016/j.cell.2007.02.044

Kuijper, S., Turner, C. J., and Adams, R. H. (2007). Regulation of angiogenesis by Eph-ephrin interactions. *Trends Cardiovasc. Med.* 17, 145–151. doi: 10.1016/j.tcm.2007.03.003

Kuratomi, S., Ohmori, Y., Ito, M., Shimazaki, K., Muramatsu, S., Mizukami, H., et al. (2009). The cardiac pacemaker-specific channel Hcn4 is a direct transcriptional target of MEF2. *Cardiovasc. Res.* 83, 682–687. doi: 10.1093/cvr/cvp171

Levy, D., Ehret, G. B., Rice, K., Verwoert, G. C., Launer, L. J., Dehghan, A., et al. (2009). Genome-wide association study of blood pressure and hypertension. *Nat. Genet.* 41, 677–687.

Liu, M. Y., Li, N., Li, W. A., and Khan, H. (2017). Association between psychosocial stress and hypertension: a systematic review and meta-analysis. *Neurol. Res.* 39, 573–580. doi: 10.1080/01616412.2017.1317904

Liu, P. Y., Death, A. K., and Handelsman, D. J. (2003). Androgens and cardiovascular disease. *Endocr. Rev.* 24, 313–340.

Luo, H., Wan, X., Wu, Y., and Wu, J. (2001). Cross-linking of EphB6 resulting in signal transduction and apoptosis in Jurkat cells. *J. Immunol.* 167, 1362–1370. doi: 10.4049/jimmunol.167.3.1362

Luo, H., Wu, Z., Tremblay, J., Thorin, E., Peng, J., Lavoie, J. L., et al. (2012). Receptor tyrosine kinase Ephb6 regulates vascular smooth muscle contractility and modulates blood pressure in concert with sex hormones. *J. Biol. Chem.* 287, 6819–6829. doi: 10.1074/jbc.m111.293365

Luo, H., Yu, G., Wu, Y., and Wu, J. (2002). EphB6 crosslinking results in costimulation of T cells. *J. Clin. Invest.* 110, 1141–1150. doi: 10.1172/jci0215883

Masubuchi, Y., Kumai, T., Uematsu, A., Komoriyama, K., and Hirai, M. (1982). Gonadectomy-induced reduction of blood pressure in adult spontaneously hypertensive rats. *Acta Endocrinol.* 101, 154–160. doi: 10.1530/acta.0.1010154

Mathar, I., Vennekens, R., Meissner, M., Kees, F., Van Der Mieren, G., Camacho Londono, J. E., et al. (2010). Increased catecholamine secretion contributes to hypertension in TRPM4-deficient mice. *J. Clin. Invest.* 120, 3267–3279. doi: 10.1172/jci41348

Miner, M., Barkin, J., and Rosenberg, M. T. (2014). Testosterone deficiency: myth, facts, and controversy. *Can. J. Urol.* 21(Suppl. 2), 39–54.

Mousavi, M., Hellstrom-Lindahl, E., Guan, Z. Z., Bednar, I., and Nordberg, A. (2001). Expression of nicotinic acetylcholine receptors in human and rat adrenal medulla. *Life Sci.* 70, 577–590. doi: 10.1016/s0024-3205(01)01427-8

Mousavi, M., and Nordberg, A. (2006). Expression of the alpha7, alpha4 and alpha3 nicotinic receptor subtype in the brain and adrenal medulla of transgenic mice carrying genes coding for human AChE and beta-amyloid. *Int. J. Dev. Neurosci.* 24, 269–273. doi: 10.1016/j.ijdevneu.2006.02.002

Org, E., Eyheramendy, S., Juhanson, P., Gieger, C., Lichtner, P., Klopp, N., et al. (2009). Genome-wide scan identifies CDH13 as a novel susceptibility locus contributing to blood pressure determination in two European populations. *Hum. Mol. Genet.* 18, 2288–2296. doi: 10.1093/hmg/ddp135

Padmanabhan, S., Melander, O., Johnson, T., Di Blasio, A. M., Lee, W. K., Gentilini, D., et al. (2010). Genome-wide association study of blood pressure extremes identifies variant near UMOD associated with hypertension. *PLoS Genet* 6:e1001177. doi: 10.1371/journal.pgen.1001177

Pasquale, E. B. (2008). Eph-ephrin bidirectional signaling in physiology and disease. *Cell* 133, 38–52. doi: 10.1016/j.cell.2008.03.011

Pende, A., Musso, N. R., Vergassola, C., Puppo, F., Ioverno, A., Criscuolo, D., et al. (1990). Neuroendocrine effects of interferon alpha 2-a in healthy human subjects. *J. Biol. Regul. Homeost. Agents* 4, 67–72.

Perez-Alvarez, A., Hernandez-Vivanco, A., Mcintosh, J. M., and Albillos, A. (2012). Native alpha6beta4* nicotinic receptors control exocytosis in human chromaffin cells of the adrenal gland. *FASEB J.* 26, 346–354. doi: 10.1096/fj.11-190223

Phillips, G. B., Jing, T. Y., Resnick, L. M., Barbagallo, M., Laragh, J. H., and Sealey, J. E. (1993). Sex hormones and hemostatic risk factors for coronary heart disease in men with hypertension. *J. Hypertens.* 11, 699–702. doi: 10.1097/00004872-199307000-00003

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., et al. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* 81, 559–575. doi: 10.1086/519795

Reckelhoff, J. F., Zhang, H., and Granger, J. P. (1998). Testosterone exacerbates hypertension and reduces pressure-natriuresis in male spontaneously hypertensive rats. *Hypertension* 31, 435–439. doi: 10.1161/01.hyp.31.1.435

Rivier, C., Vale, W., and Brown, M. (1989). In the rat, interleukin-1 alpha and -beta stimulate adrenocorticotropin and catecholamine release. *Endocrinology* 125, 3096–3102. doi: 10.1210/endo-125-6-3096

Shi, W., Wang, Y., Peng, J., Qi, S., Vitale, N., Kaneda, N., et al. (2019). EPHB6 controls catecholamine biosynthesis by up-regulating tyrosine hydroxylase transcription in adrenal gland chromaffin cells. *J. Biol. Chem.* 294, 6871–6887. doi: 10.1074/jbc.ra118.005767

Slavin, T. P., Feng, T., Schnell, A., Zhu, X., and Elston, R. C. (2011). Two-marker association tests yield new disease associations for coronary artery disease and hypertension. *Hum. Genet.* 130, 725–733. doi: 10.1007/s00439-011-1009-6

Sparrenberger, F., Cichelero, F. T., Ascoli, A. M., Fonseca, F. P., Weiss, G., Berwanger, O., et al. (2009). Does psychosocial stress cause hypertension? A systematic review of observational studies. *J. Hum. Hypertens.* 23, 12–19. doi: 10.1038/jhh.2008.74

Traish, A. M., Miner, M. M., Morgentaler, A., and Zitzmann, M. (2011). Testosterone deficiency. *Am. J. Med.* 124, 578–587.

Tremblay, J., Wang, Y., Raelson, J., Marois-Blanchet, F. C., Wu, Z., Luo, H., et al. (2017). Evidence from single nucleotide polymorphism analyses of ADVANCE study demonstrates EFNB3 as a hypertension risk gene. *Sci. Rep.* 7:44114.

Wang, Y., Hamet, P., Thorin, E., Tremblay, J., Raelson, J., Wu, Z., et al. (2016). Reduced blood pressure after smooth muscle EFNB2 deletion and the potential association of EFNB2 mutation with human hypertension risk. *Eur. J. Hum. Genet.* 24, 1817–1825.

Wang, Y., Shi, W., Blanchette, A., Peng, J., Qi, S., Luo, H., et al. (2018). EPHB6 and testosterone in concert regulate epinephrine release by adrenal gland chromaffin cells. *Sci. Rep.* 8:842.

Wang, Y., Thorin, E., Luo, H., Tremblay, J., Lavoie, J. L., Wu, Z., et al. (2015). EPHB4 protein expression in vascular smooth muscle cells regulates their contractility, and EPHB4 deletion leads to hypotension in mice. *J. Biol. Chem.* 290, 14235–14244. doi: 10.1074/jbc.m114.621615

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

Wilkinson, D. G. (2000). Eph receptors and ephrins: regulators of guidance and assembly. *Int. Rev. Cytol.* 196, 177–244. doi: 10.1016/s0074-7696(00)96005-4

Wu, J., and Luo, H. (2005). Recent advances on T-cell regulation by receptor tyrosine kinases. *Curr. Opin. Hematol.* 12, 292–297. doi: 10.1097/01.moh.0000166497.26397.9f

Wu, P. C., Fann, M. J., and Kao, L. S. (2010). Characterization of Ca2+ signaling pathways in mouse adrenal medullary chromaffin cells. *J. Neurochem.* 112, 1210–1222. doi: 10.1111/j.1471-4159.2009.06533.x

Wu, T., Zhang, B. Q., Raelson, J., Yao, Y. M., Wu, H. D., Xu, Z. X., et al. (2018). Analysis of the association of EPHB6, EFNB1 and EFNB3 variants with hypertension risks in males with hypogonadism. *Sci. Rep.* 8:14497.

Wu, Z., Luo, H., Thorin, E., Tremblay, J., Peng, J., Lavoie, J. L., et al. (2012). Possible role of Efnb1 protein, a ligand of Eph receptor tyrosine kinases, in modulating blood pressure. *J. Biol. Chem.* 287, 15557–15569. doi: 10.1074/jbc.m112.340869

Zitzmann, M. (2009). Testosterone deficiency, insulin resistance and the metabolic syndrome. *Nat. Rev. Endocrinol.* 5, 673–681. doi: 10.1038/nrendo.2009.212

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: 28 January 2021
doi: 10.3389/fgene.2021.637780

A Review of Genetic and Physiological Disease Mechanisms Associated With Cav1 Channels: Implications for Incomplete Congenital Stationary Night Blindness Treatment

Tal T. Sadeh[1], Graeme C. Black[1,2] and Forbes Manson[1]*

[1] Division of Evolution and Genomic Sciences, Faculty of Biology, Medicine and Health, University of Manchester, Manchester, United Kingdom, [2] Manchester Centre for Genomic Medicine, Manchester Academic Health Sciences Centre, Manchester University NHS Foundation Trust, St Mary's Hospital, Manchester, United Kingdom

OPEN ACCESS

Edited by:
Gabrielle Wheway, University of Southampton, United Kingdom

Reviewed by:
Katarzyna Szymanska, University of Leeds, United Kingdom
Zhipeng Liu, Purdue University, United States

****Correspondence:***
Graeme C. Black
graeme.black@manchester.ac.uk

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

Received: *04 December 2020*
Accepted: *12 January 2021*
Published: *28 January 2021*

Citation:
Sadeh TT, Black GC and Manson F (2021) A Review of Genetic and Physiological Disease Mechanisms Associated With Cav1 Channels: Implications for Incomplete Congenital Stationary Night Blindness Treatment. Front. Genet. 12:637780. doi: 10.3389/fgene.2021.637780

Calcium channels are crucial to a number of cellular functions. The high voltage-gated calcium channel family comprise four heteromeric channels (Cav1.1-1.4) that function in a similar manner, but that have distinct expression profiles. Three of the pore-forming α_1 subunits are located on autosomes and the forth on the X chromosome, which has consequences for the type of pathogenic mutation and the disease mechanism associated with each gene. Mutations in this family of channels are associated with malignant hyperthermia (Cav1.1), various QT syndromes (Cav1.2), deafness (Cav1.3), and incomplete congenital stationary night blindness (iCSNB; Cav1.4). In this study we performed a bioinformatic analysis on reported mutations in all four Cav α_1 subunits and correlated these with variant frequency in the general population, phenotype and the effect on channel conductance to produce a comprehensive composite Cav1 mutation analysis. We describe regions of mutation clustering, identify conserved residues that are mutated in multiple family members and regions likely to cause a loss- or gain-of-function in Cav1.4. Our research highlights that therapeutic treatments for each of the Cav1 channels will have to consider channel-specific mechanisms, especially for the treatment of X-linked iCSNB.

Keywords: L-type calcium channels, mutation analysis, incomplete congenital stationary night blindness, treatment, CaV1.4 calcium channel

INTRODUCTION

Voltage-gated calcium channels perform multiple functions including signaling, hormone and neurotransmitter secretion, muscle contraction, and gene expression (Catterall et al., 2005). The family is grouped on the voltage sensitivities of the α_1 subunits, with Cav1.1-1.4 comprising the high voltage L-type channels (Catterall et al., 2005). Cav1.1-1.3 channels are encoded by autosomal genes; the Cav1.4 channel is encoded by the X-linked *CACNA1F* gene. Cav1 channels diverged from a common evolutionary ancestor and have 60–75% overall polypeptide sequence identity.

Within the voltage-gated calcium channels, the α subunit forms the conductive pore. It associates with a β subunit, an $\alpha_2\delta$ subunit, and, for some channels, a γ unit, in an equal ratio (Catterall et al., 2005; **Supplementary Figure 1**). The α_1 protein structure consists of four homologous transmembrane repeats (I–IV), each comprising of six helical segments (S1–S6) connected by extracellular or cytoplasmic loops. The fourth segment of each repeat carries positively charged amino acids at every third position that function as voltage-sensor domains (VSDs) sensitive to membrane depolarization impulses (Catterall et al., 2005). The S5 and S6 segments form the pore, which is lined by the S5–S6 loops. Each of these S5–S6 loops contains a conserved glutamate residue that forms a negatively charged ring that functions as the ion-selectivity filter (Yang et al., 1993). Thus, their conserved properties enable the four Cav1 channels to function in a similar fashion, albeit in different temporo-spatial circumstances, for calcium homeostasis (Zamponi et al., 2015): The α_1 subunit of Cav1.1 is encoded by the *CACNA1S* gene located on chromosome 1q32.1 and is exclusively expressed in skeletal muscle (Catterall et al., 2005). The Cav1.4 α_1 subunit is located at Xp11.23 (encoded by *CACNA1F*) and is expressed in retinal interneurons (Bech-Hansen et al., 1998). Cav1.2 (*CACNA1C*, Chr12p13.33) and Cav1.3 (*CACNA1D*, Chr3p21.1) are expressed in many electrically excitable cells and are often expressed in the same cell (e.g., adrenal chromaffin cells, sinoatrial node, neurons, and atrial cardiomyocytes) (Zamponi et al., 2015). The channels' unique tissue expression profiles are reflected in their different voltage-dependencies i.e., they activate and inactivate at different voltages.

Missense mutations in the autosomal Cav1.1-1.3 α genes, although variously described as gain-of-function (GoF) and loss-of-function (LoF), mostly cause a gain of activity that is associated with autosomal dominant phenotypes. For instance, Timothy syndrome (TS) is an autosomal dominant, multiorgan condition, that is predominantly caused by GoF Cav1.2 mutations [e.g., p.Gly406Arg (Splawski et al., 2005)]. A smaller number of LoF variants have been described and are generally associated with autosomal recessive traits, such as Cav1.3 insertion mutations that result in autosomal recessive congenital deafness (Baig et al., 2011).

X-linked congenital stationary night blindness (X-linked iCSNB) is a static monogenic disorder that results in visual disability predominantly in males. Pathogenic *CACNA1F* mutations disrupt Cav1.4 function and impair normal retinal synaptic transmission (Strom et al., 1998). There is a small number of reports of affected females, presumably as a result of skewed X-inactivation (Rigaudière et al., 2003; Hemara-Wahanui et al., 2005). In the majority of cases, recessive X-linked mutations in *CACNA1F* abolish or decrease Cav1.4 calcium current density. Missense mutations can increase or decrease current density. Null (amorphic) alleles abolish calcium currents, and hypomorphs reduce currents by either dysregulating the current window (e.g., a shift in the voltage dependence of activation or inactivation) or the quantity of calcium influx. These changes may be through the production of no active protein products (null), or by reducing transcription or producing a protein lacking full functionality (hypomorph). By contrast, hypermorphs result in hyperactive channels by either increasing the current window or calcium influx. These consequences are seen in electrophysiological recordings by a shift in the voltage dependence of activation or inactivation (hyperpolarized leftward shift or depolarized rightward shift), or by changing the amount of calcium influx.

In this study, we collate and analyze reported mutations in the α_1 subunit-encoding genes of L-type channels to identify the similarities and differences between the autosomal Cav1.1-1.3 and X-linked Cav1.4 channels with the aim to inform the pathophysiology of Cav1.4 variants, which is an important prerequisite for future therapeutic intervention.

MATERIALS AND METHODS

Reported Mutations

Mutations in the α_1 subunits of Cav1.1-1.4 genes were retrieved from the Human Gene Mutation Database (HGMD, licensed version accessed on June 2020; Stenson et al., 2017) and functional studies were collated from PubMed. These were classed as missense, nonsense, splicing, deletion, insertion/duplications, or other (complex rearrangements or regulatory substitutions). The associated phenotypes and mode of inheritance were recorded from HGMD and Online Mendelian Inheritance in Man (OMIM; Amberger et al., 2009).

Population Database Search (gnomAD)

The tolerance and constraint scores of mutation types in Cav1 genes and their minor allele frequencies (MAFs) in the general population were derived from The Genome Aggregation Database (gnomAD; Karczewski et al., 2019). The most common version of a gene in a population is referred to as the wildtype allele and variations are annotated relative to it; the MAF is the number of times a variant allele occurs in a population for any data set. A high Z score indicates more constraints and intolerance to synonymous and missense variations, and pLI score close to 1 implies that the gene is intolerant to protein-truncating variants (i.e., nonsense, frameshift, splice sites variants), which likely cause LoF (predicted LoF; pLoF). In addition, the observed/expected (oe) ratio compares the observed pLoF to the expected frequency of the variation in the general population, supporting the Z and pLI probabilities (90% CI).

Physicochemical and Pathogenicity Prediction

The physicochemical properties of each mutation were manually analyzed using NCBI's Amino Acid Explorer tools "Structure and Chemistry" and "Common Substitutions" (Bulka et al., 2006).[1] These tools compare specific physicochemical constraints of the amino acid pair such as a change in amino acid size, charge, and hydrophobicity. The latter tool relies on BLOSUM62 matrix to sort the frequency of the substitution (Henikoff and Henikoff, 1992). Pathogenicity was predicted using

[1] https://www.ncbi.nlm.nih.gov/Class/Structure/aa/aa_explorer.cgi

polymorphism phenotyping v2 (PolyPhen-2; Adzhubei et al., 2010), which considers protein structural properties such as amino acid surface area accessibility and generates a score where 1 is damaging and 0 is benign.

UniProt was used to annotate channel-specific domains and amino acids (The UniProt Consortium, 2018). Functionally characterized mutations were assigned as null, hypomorphic, or hypermorphic, dependent on the published electrophysiological properties and protein expression.

Conservation Analysis

Evolutionary constraints were analyzed by comparing protein conservation between Cav1 paralogs (protein accession numbers: Cav1.1 NP_000060.2; Cav1.2 NP_955630.3; Cav1.3 NP_000711.1; and Cav1.4 NP_005174.2) and 10 orthologs (*Mus* musculus, *Rattus* norvegicus, *Canis* lupus familiaris, *Felis* catus, *Macaca* mulatta, *Sus* scrofa, *Danio* rerio, *Halichoerus* grypus, *Zootoca* vivipara, and *Xenopus* tropicalis), using Clustal Omega (Sievers et al., 2011). *Xenopus* was not included for Cav1.3 as only low-quality sequences are available.

RESULTS

Cav1 Proteins: Incidence of Loss-of-Function (LoF) Mutations

For the four Cav1 encoding genes the population frequencies of LoF variants were examined and respective scores were ascertained from gnomAD. Cav1.1 has a total of 76 different LoF variants recorded on gnomAD with pLI = 0; oe = 0.39; the total number of Cav1.1 LoF alleles was 686. The low pLI score predicts that Cav1.1 is tolerant to LoF variants and therefore tolerant of haploinsufficiency (i.e., there will be sufficient protein function from the remaining wildtype allele). By contrast, Cav1.2 and Cav1.3 have a lower incidence of LoF variations in the population than Cav1.1. There are 31 LoF variants for Cav1.2, comprising 81 alleles (pLI = 1; oe = 0.1). For Cav1.3 there are 36 different LoF allele that collectively occur 296 times (pLI = 1; oe = 0.21). The low incidence of LoF variants in these channels indicates they are less tolerant of haploinsufficiency. It should be stressed that despite these statistical predictions only functional analyses will confirm if a LoF variant is associated with a loss of protein function, i.e., reduced current density and changes in protein expression. In Cav1.4 there are 35 different LoF variants collated in gnomAD, with an allele count of 54. Most of these occur only once in a single individual with 17 hemizygote males and 37 heterozygote females. A LoF pLI = 0 and oe = 0.45 indicates that Cav1.4 is not under selection against such variants.

Cav1 Mutation Spectrum in Disease

Overall, for monogenic disorders caused by pathogenic variation in Cav 1 channels, the proportion of missense vs LoF variants is very different (**Figure 1**). Cav1.1-1.3 have 0–6% nonsense mutations, whereas, Cav1.4 has 18% nonsense mutations. Cav1.1-1.3 have 81–85% missense mutations and Cav1.4 has 35% missense mutations.

Cav 1.1

There are 66 Cav1.1 mutations reported in HGMD, of which, 56 are missense mutations (85%). These missense alterations are the sole group of pathogenic variants associated with common Mendelian Cav1.1 conditions: malignant hyperthermia (MH; Monnier et al., 1997), muscular dystrophy (Zenagui et al., 2018), primary and hypokalemia periodic paralysis (hypoPP; Ptácek et al., 1994), exertional heat illness (Fiszer et al., 2015), and rhabdomyolysis (Vivante et al., 2017). These all follow an autosomal dominant inheritance, apart from muscular dystrophy and rhabdomyolysis, which are also inherited recessively. Overall, 60% of the reported mutations result in MH (16 mutations), myopathy (12 mutations), and hypoPP (11 mutations), with hypoPP being a monogenic condition; caused by Cav1.1 mutations only.

The 10 reported LoF variants (five deletions, four nonsense, one splicing) have been associated with complex and multifactorial diseases such as autism spectrum disorder (ASD) (Torrico et al., 2019), schizophrenia (Purcell et al., 2014), exertional heat illness, and myopathy (Hunter et al., 2015). While such Cav1.1 variants can be considered risk factors for these disorders, it is not possible to make definitive correlations due to variable expression and incomplete penetrance.

Six missense mutations in Cav1.1 have previously been functionally characterized *in vitro* – these are associated with hypoPP and MH (**Supplementary Figure 2A**). Interestingly, these result in both hypermorphic (one mutation) and hypomorphic (five mutations) proteins, although the latter are the most common (**Table 1**). Mutations reducing the current density are due to the loss of a positive charge in S4 VSDs, and three of these also delay the time course of activation [p.Arg528His (Morrill et al., 1998), p.Arg1239His and p.Arg1239Gly (Morrill and Cannon, 1999)].

Two of the functionally analyzed mutations (p.Arg175Trp, p.Arg1086His) are reported in the heterozygous state in gnomAD and have very low MAFs (0.00003 and 0.000004, respectively); no other analyzed mutation is present in gnomAD.

Cav1.2

There are 85 Cav1.2 mutations reported in HGMD, with 72 missense mutations (85%). These result in three monogenic cardiovascular disorders including Long QT syndrome 8 (LQT8; Boczek et al., 2013), Brugada syndrome 3 (BRGDA3; Antzelevitch et al., 2007), and TS (Splawski et al., 2004), which are autosomal dominant traits. These account for 60% of the reported mutations: LQT8 (25 mutations), BRGDA3 (12 mutations), and TS (11 mutations). In addition, Cav1.2 mutations cause a range of common disorders including cardiomyopathy (D'Argenio et al., 2014), atrial or ventricular fibrillation (Maltese et al., 2019), bradycardia (Zhu et al., 2018), and cerebellar ataxia (Chen et al., 2019), which are both autosomal dominant and recessively inherited. Cav1.2 has been reported in association with four complex diseases: ASD (Jiang et al., 2013), intellectual disability (Hu et al., 2019), schizophrenia (Roussos et al., 2014), and epileptic encephalopathies (Bozarth et al., 2018).

Missense mutations are the most frequent type associated with Cav1.2 disease. However, a small number of LoF variants (two

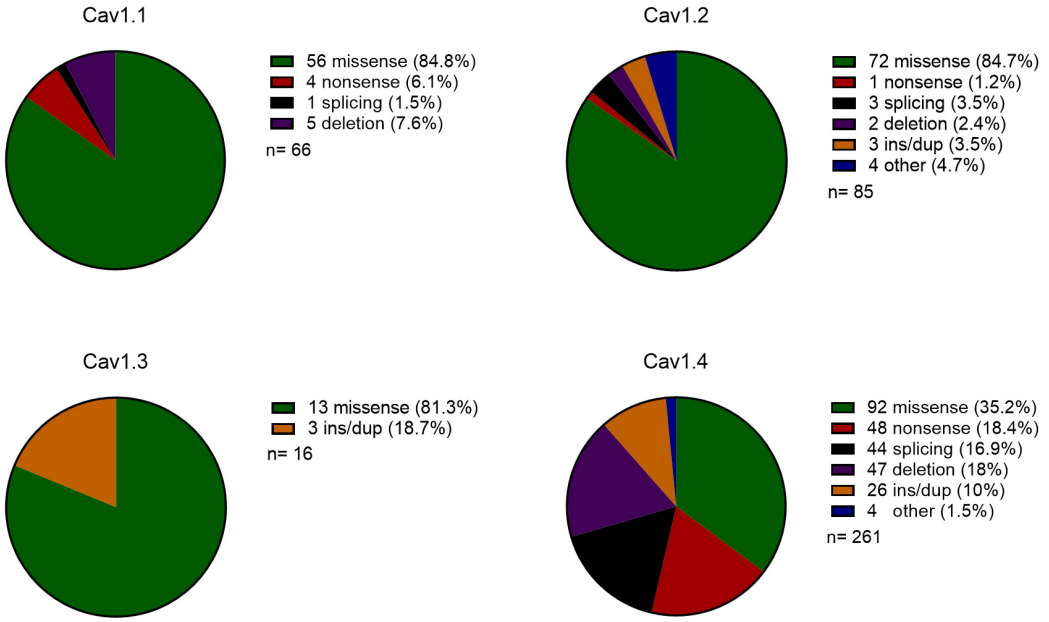


FIGURE 1 | Comparison of the type of mutations in Cav1 genes reported on HGMD. Pie chart of the type of mutations cataloged for Cav1.1-1.4 genes. Mutation key and numbers are shown on the right: ins/dup: insertion/duplication, other: complex rearrangements or regulatory substitutions, n: number of reported mutations.

TABLE 1 | Summary of functionally analyzed Cav1 missense mutations.

Channel (gene)	Number of tested mutations	Effect on current	Details
Cav1.1 (*CACNA1S*)	6	5 decreased currents. 1 increased currents.	1 null and 4 hypomorphs are all Arg substitution that reduce the current density due to the loss of charge in the VSDs. 1 hypermorph in S5-S6 loop lines the pore and results in faster activation of currents upon depolarisation.
Cav1.2 (*CACNA1C*)	17	5 decreased currents. 12 increased currents.	2 hypermorph in S6 affect the pore charge or hydrophobicity. 14 mutations are in cytoplasmic loops: 10 hyper- and 4 hypo-morphs. 4/5 hypomorphs affect trafficking and reduce membrane expression. All hypermorphs causing LQT8 increase the current window, which is consistent with a long QT interval. p.Ile1186Thr increase the current window and causes both LQT8 and TS but substitution to Val only seen in LQT8.
Cav1.3 (*CACNA1D*)	14	2 decreased currents. 12 increased currents.	6 hypermorph in S6 affect the pore charge or hydrophobicity and cause 3 phenotypes. 2 hypomorph mutations: 1 Arg substitution reduces the current density due to the loss of charge in the VSDs. 1 His substitution unbalances the charge on an extracellular loop and reduces expression on the membrane. 6 hypermorphs are in cytoplasmic loops that mostly cause a hyperpolarised shift.
Cav1.4 (*CACNA1F*)	10	7 decreased currents. 3 increased currents.	3 hypermorph in S6 affect the pore charge or hydrophobicity. 3 hypomorphs mutations: 2 affect trafficking and reduce protein expression. 1 reduce the current density with normal protein expression. 4 null mutations have no currents with normal protein expression. (p.Gly1007Arg introduces an extra charge to the VSDs).

VSDs, voltage sensor domains, LQT8, Long QT syndrome 8, TS, Timothy syndrome. Additional details in ***Supplementary Table 1****.*

splicing and one insertion mutations) are described in patients with BRGDA3 and cerebellar ataxia. Ten other mutation types have been described associated with two complex disorders; schizophrenia (including four regulatory substitutions, two deletions, one splicing, and one nonsense mutation), and ASD (two insertions). As for Cav1.1, the genetic heterogeneity of polygenic disorders makes it difficult to be conclusive regarding their correlation.

Seventeen Cav1.2 missense mutations have previously been functionally characterized by cell electrophysiology

(**Table 1**). These are associated with BRGDA3, LQT8, and TS (**Supplementary Figure 2B**). Hypermorphs are the most common and mutations in transmembrane domains increase the current window by different mechanisms including a shift in the activation and/or inactivation or by increasing the maximum ion conductance (detailed in **Supplementary Table 1**). There are five hypomorphs that reduce the current density, most of which express poorly at the plasma membrane [p.Ala39Val (Antzelevitch et al., 2007), p.Asn300Asp (Béziau et al., 2014), p.Arg518Cys/His (Boczek et al., 2015)].

Three functionally analyzed mutations (two hypermorphs: p.Ala28Thr, p.Gly406Arg, and one hypomorph: p.Arg518His) are present in the general population with very low MAFs (0.00006, 0.0004, and 0.00003, respectively), and only in the heterozygous state. The other 13 functionally analyzed mutations are absent in gnomAD.

Cav1.3

Cav1.3 has 16 reported mutations in HGMD, of which, 13 are missense mutations (81%). These are associated with four Mendelian diseases including the autosomal dominant conditions primary aldosteronism with seizures and neurologic abnormalities (PASNA; Semenova et al., 2018), aldosterone-producing adenomas (APAs; Scholl et al., 2013), and epilepsy (Tumiene et al., 2018), and the autosomal recessive condition sinoatrial node dysfunction and deafness (SANDD; Liaqat et al., 2019). Five other complex conditions [ASD (O'Roak et al., 2012), bipolar disorder (Ross et al., 2016), developmental delays (Di Gregorio et al., 2017), and hearing impairment and intellectual disability (Garza-Lopez et al., 2018)] have been associated with Cav1.3 variants, but the exact causality remains unclear.

Missense mutations are the most frequent mutation type, however, three other mutation types have been associated with recessive (SANDD; one insertion) or complex (developmental delay; one insertion and ASD; one duplication) conditions.

The channel function of 14 missense mutations have previously been analyzed and are associated with APAs, ASD, hearing impairment and intellectual disability, and PASNA (**Supplementary Figure 2C**). Hypermorphs that increase the current density is the most frequent mechanism causing these conditions (**Table 1**). The hypomorphs have charge differences that reduce the current density.

None of the functionally analyzed mutations are present in gnomAD.

Cav1.4

Cav1.4 is the only X-linked Cav1 channel. 261 mutations have been reported in HGMD, of which, 92 are missense mutations (35%). 206 of all mutations are associated with iCSNB (Strom et al., 1998), the remainder causing cone-rod dystrophy (Jalkanen et al., 2006) and Aland island eye disease (Jalkanen et al., 2007). Retinitis pigmentosa (Xu et al., 2015), high myopia (Sun et al., 2015), and Usher syndrome (Song et al., 2011) have all been described although these remain unconfirmed.

About 169 truncating variants are predominantly associated with recessive iCSNB (48 nonsense, 42 deletions, 41 splicing, 25 insertion/duplications, and 4 complex rearrangements), and a small number are reportedly associated with cone-rod dystrophy [two deletions (Hauke et al., 2013; Huang et al., 2016) and one insertion (Jalkanen et al., 2006)] and Aland island eye disease [one deletion (Jalkanen et al., 2007)], or high myopia [one deletion and one splicing (Sun et al., 2015)] and retinitis pigmentosa (two splicing (Xu et al., 2015; Jespersgaard et al., 2019) and one deletion (Martin-Merida et al., 2019)].

Reduction or loss of activity from the single X chromosome allele is presumed to be the pathogenic mechanism for most truncating alleles in males since most will result in a lack of protein due to nonsense mediated decay (NMD; Sharma et al., 2020). The 48 nonsense alleles (out of 261 reported mutations, 18%) in Cav1.4 are predicted to be degraded through NMD. Nonsense mutations in the final exon often escape NMD, which may have different mechanistic consequences.

Ten iCSNB missense mutations have previously been functionally characterized. As expected for a gene associated with a large number of LoF mutations, hypomorphic alleles are the most common: seven hypomorphic or null alleles that reduce or abolish channel conductance. However, three hypermorphs increase channel conductance (**Supplementary Figure 2D**); these are in the S6 repeats. All mutations in loops are hypomorphic or null alleles (**Table 1**). The four null alleles express at normal global protein levels, which suggests that the proteins are unstable and escaped degradation and are not trafficked to the membrane correctly.

The null p.Gly1007Arg mutant has a very low MAF (0.00003) in the general population and there are no homozygous individuals. No other functionally analyzed mutations are present in gnomAD.

Composite Cav1 Mutation Analysis

In order to understand the mechanistic consequences of missense alleles we next analyzed the missense mutations reported in the four Cav1 α_1 genes. However, since causality is so hard in complex diseases we only looked at the mutations associated with verified monogenic phenotypes.

In total 429 mutations have been described in the Cav1 α_1 genes and are associated with 35 phenotypes. 234 are missense variants, of which, 47 have previously been functionally characterized by *in vitro* assays and are pathogenic in 10 phenotypes (**Figure 2**). Of the functionally analyzed mutations, hyperactive channels are the most common and most are associated with autosomal dominant conditions, although this predominance is likely a result of bias by researchers selecting what mutants to study. Twenty-eight hypermorphic alleles increase the current density, and 19 hypomorphic or null alleles reduce it (**Table 2**). **Supplementary Table 1** summarizes the functionally characterized mutations reported for Cav1 channels.

By plotting all mutations on the composite Cav1 channel (**Figure 2**) it is possible to identify mutated residues that overlap between the channels and have specific consequences that are either shared or that are distinct between the channels. There are

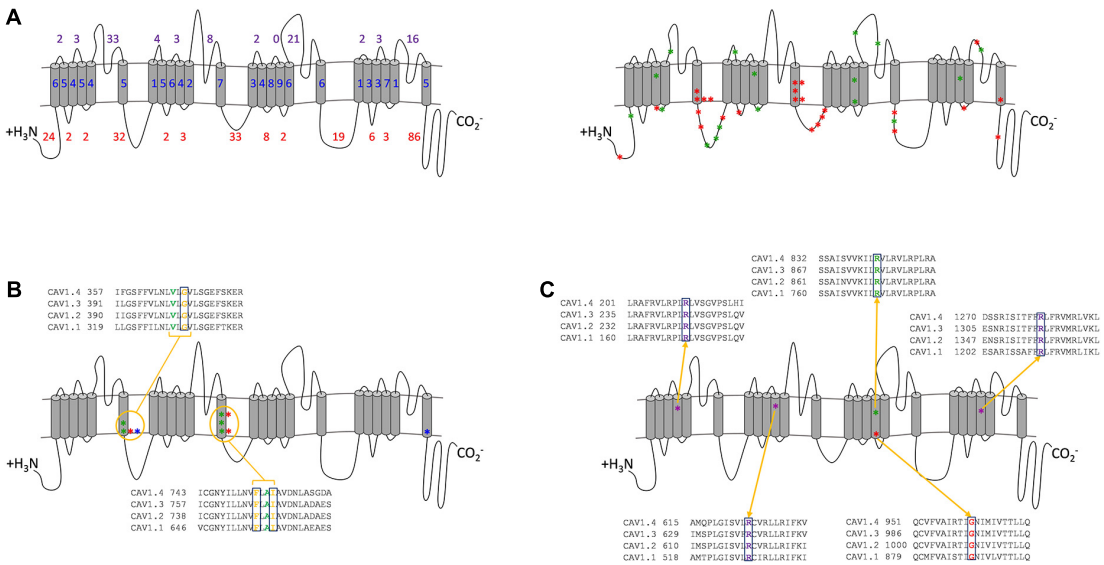


FIGURE 2 | Combined analysis of Cav1 mutation positions. **(A)** Location of Cav1 α_1 subunit mutations. **(Left)** The number of mutations in transmembrane (blue), extracellular (purple) and cytoplasmic (red) regions. **(Right)** The 47 functionally tested missense mutations are shown as red (GoF) and green (LoF) stars. Stars represent a mutated residue and may include multiple amino acid substitution. The cytoplasmic loops are the most frequently mutated regions (222 mutations), of which, 86 are in the C-terminal tail, whereas the extracellular loops have 97 mutations, and the transmembrane domains have 110 mutations. 114 mutations are in the pore-forming domains, and 25 are in the S4 VSDs. **(B)** Three functionally characterized hypermorphs in S6 repeats I and II of Cav1.4 [p.Gly369Asp (Hoda et al., 2005), p.Phe753Cys (Hope et al., 2005), p.Ile756Thr (Peloquin et al., 2007)] (red) are present at the equivalent position in Cav1.2: p.Gly402Ser (Splawski et al., 2005) (equivalent to p.Gly369Asp) (blue) and Cav1.3: p.Gly403Asp/Arg (Scholl et al., 2013) (equivalent to p.Gly369Asp), p.Ile750Met (Scholl et al., 2013) (equivalent to p.Ile756Thr), p.Phe747Leu (Pinggera et al., 2018) (equivalent to p.Phe753Cys) (green). **(C)** Five mutations to the VSDs (four hypomorphs and one null) are functionally analyzed mutants that damage the sensor's voltage sensitivity and reduce the channel's conductance in Cav1.1 [p.Arg174Trp (Bannister and Beam, 2013), p.Arg528His (Morrill et al., 1998), p.Arg1239His/Gly (Morrill and Cannon, 1999)] (purple), Cav1.3 [p.Arg990His (Monteleone et al., 2017)] (green), and Cav1.4 [p.Gly1007Arg (Peloquin et al., 2007)] (red). The conservation in paralogs is shown for each residue.

46 missense mutations in the C-terminal tail, which may interfere with the calcium-dependent inactivation (CDI) machinery. The unique regulatory functions of this region in Cav1.4 suggest that mutations will have different effects in Cav1.1-1.3 than in Cav1.4.

There are 114 pathogenic mutations in the pore-forming domains (S5, S6, and S5–S6 loops), 10 of which are transmembrane and have been functionally analyzed *in vitro* (**Figure 2B**). Most of these amino acid substitutions result in a charge change (e.g., uncharged non-polar glycine to positively charged arginine) or the hydrophobicity (e.g., hydrophobic isoleucine to hydrophilic threonine). All increase the time course of inactivation, resulting in a net current increase. Taken together the amino acid substitutions at these three positions are likely to have a similar hypermorphic effect across all channels. Similarly, mutations to the S4 VSDs result in similar functional consequences across the channels (**Figure 2C**). The functionally analyzed hypomorphs are due to the loss of the positively charged arginine residue, which delays the channel's opening and reduces the net current. These charged residues are highly conserved and sensitive to change that even the arginine substitution to another positively charged amino acid (histidine) has a deleterious effect in Cav1.1 and Cav1.3. This is possibly because histidine has a shorter R-chain than arginine. Interestingly, in Cav1.4 the addition of a positive amino acid in this already charged region abolishes the channel's currents (p.Gly1007Arg). These accumulative consequences demonstrate that the smallest charge difference effects the fine-tuning of the VSDs and have a similar hypomorphic effect for all Cav1 channels.

DISCUSSION

The different chromosomal location of the four Cav1 genes suggests that there are more similarities in the type of mutations in autosomal Cav1.1–1.3 genes than there are when comparing all Cav1 genes; Cav1.1–1.3 are predominantly associated with missense mutations (81–85%), whereas missense mutations in Cav1.4 account for 35% of all mutations (**Figure 1**). Missense mutations in the autosomal Cav1 channels mostly result in hyperactive channels and manifest dominant Mendelian diseases. This implies that in dominant disorders the main disease mechanism is likely to be via GoF rather than LoF/haploinsufficiency. As Cav1 channels are heteromers, to decrease the function the mutant α-subunit protein must aggregate with the wildtype protein or preferentially bind with the interacting protein subunits to reduce their binding to the wildtype α-subunit protein.

Conversely, for Cav1.4, hypomorphic and null alleles are associated with recessive X-linked disease. This explains the high incidence of LoF mutations in Cav1.4 where there is no compensating wildtype allele in hemizygous males, underlining the observation that these are 'hypomorphic phenotypes',

TABLE 2 | Collated summary of Cav1 mutations.

Number of tested mutations	Effect on current	Details
47	19 decreased currents. 28 increased currents.	Cav1.1-1.3 are predominantly missense mutations (85%) and more tolerant of truncating mutations. Cav1.4 is intolerant to missense and truncating mutations. Cav1.1-1.3 mutations mostly cause hyperactive channels (n = 25 hyperactive and n = 12 hypoactive). Cav1.4 predominantly cause hypoactive channels (n = 7 hypoactive and n = 3 hyperactive). Mutations in S4 are hypomorphs – VSDs change of charge. n = 5 (3 in Cav1.1, 1 in Cav1.3, 1 in Cav1.4). Apart from S4, all mutations to helix are hypermorphs. n = 9 (2 in Cav1.2, 4 in Cav1.3, 3 in Cav1.4). Two hotspots in S6 of repeat I and II are all hypomorphs due to dysregulating the pore. n = 10 (1 in Cav1.2, 6 in Cav1.3, 3 in Cav1.4). Introducing a Pro constrains the loop resulting in less flexibility and hypomorphic alleles. Ser229Pro, Leu860Pro, Leu1068Pro (Cav1.4). Gly is important in a helix. Small size enables helix flexibility and substituting it results in hypermorphs. Gly402Ser (Cav1.2), Gly403Asp/Arg (Cav1.3), and Gly369Asp (Cav1.4).

VSDs, voltage sensor domains. Detailed analysis on the effect on current is presented in ***Supplementary Table 1****.*

which suggests that LoF is likely the predominant disease mechanism in Cav1.4 channelopathies. Nevertheless, a small number of *CACNA1F* mutations result in increased overall calcium currents by increasing the current density or by causing larger currents at more hyperpolarized potentials (Hemara-Wahanui et al., 2005; Hoda et al., 2005; Peloquin et al., 2007). For example, the p.Il756Thr mutation is electrophysiologically a GoF mutation as it shifts the voltage dependence of activation to more negative potentials and reduced the inactivation, however, in the retina it causes a LoF with respect to the retinal function by damaging photoreceptors signaling to bipolar cells (second-order neurons) (Hemara-Wahanui et al., 2005).

The large number (85) of mutations in the C-terminal tail indicate that mutations in this important regulatory region disturb proper channel inactivation, and so, tissue function (**Figure 2A**). The morphological changes required for the inactivation of the channels are controlled by the voltage-dependent inactivation (VDI) and CDI. Transient calcium influx is controlled by CDI via the proximal C-terminal regulatory domains (PCRD) that require the coupling of the IQ and pre-IQ calmodulin-binding domains to an EF hand domain motif when calmodulin (CaM) is bound by four calcium ions (Hulme et al., 2006; Singh et al., 2006). As neurotransmitting photoreceptors require a sustained calcium influx the CDI in Cav1.4 is inhibited by distal C-terminal regulatory domain (DCRD), although, the binding site and precise mechanism is currently unknown. This negative inhibitory CDI (ICDI) feedback mechanism ensures that Cav1.4 is non-inactivating. The absence of CDI is unique to Cav1.4 and changes the voltage dependencies (electrophysiologically recorded as a leftward shift of the IV curve which changes the steepness, relative to Cav1.1–1.3) (Singh et al., 2006). The domain differences of this region between the channels suggest that mutations in Cav1.1–1.3 PCRD are likely to dysregulate the inactivation properties, which prevents transient influxes. In Cav1.4 the functional consequence is dependent on whether the mutation is in the PCRD or DCRD tail regions; mutations in the CDI may be tolerated to a degree as it is usually inhibited by ICDI, whereas mutations in the ICDI are likely to have a hypomorphic effect by preventing the inhibition of CDI and consequently allowing the channel to close prematurely.

From this analysis, we predict that there are similar functional consequences across Cav1 channels. For example, mutations in specific residues (Cav1.4: p.Gly369, p.Phe753, p.Ile756 in S6, **Figure 2B**) will have similar hypermorphic effects, and mutations in the VSDs will have similar hypomorphic consequences (**Figure 2C**). The extent of these effects is mutation specific.

There are some conflicting studies of *CACNA1F* mutations. One study reported no electrophysiological changes in a missense (p.Gly369Asp) and nonsense (p.Trp1459*) mutation (McRory et al., 2004), whereas a second study reported both to be deleterious (Hoda et al., 2005). Our analyses suggest that both are likely to be pathogenic. The p.Gly369Asp mutation is consistent with the equivalent glycine residue being pathogenic in Cav1.2 and Cav1.3 (**Figure 2B**) and is predicted to be damaging by PolyPhen-2 (**Supplementary Table 1**). No protein expression was found for p.Trp1459* channels (Hoda et al., 2005), probably due to NMD. Functional investigations into these mutations and variants within close proximity will enhance these predictions and validate their pathogenicity.

The diverse functional consequences of Cav1 mutations suggests that certain therapies will be more pertinent to different channel dysfunctions. Over 60% of Cav1 pathogenic mutations are missense mutations, and 40% of these are associated with hypomorphic or null alleles. The differences between autosomal and X-linked channels suggest that therapeutic approaches for Cav1.4 X-linked iCSNB will differ from Cav1.1–1.3 channelopathies since the predominant mechanism for X-linked iCSNB is LoF. Treating GoF variants will require a different approach such as the use of inhibitors that act locally and specifically to inhibit or reduce the excess function. Although a reduced function as a consequence of protein dysregulation could potentially be corrected, a complete LoF can be embryonically lethal for all Cav1 channels as reported in Cav1.2 α_1 knockout mice which died in utero (Zamponi et al., 2015).

For X-linked iCSNB caused through null alleles, gene replacement therapy would compensate for the lack of channel expression by providing a functional product sufficient to restore neurotransmission to normal levels. However, when Cav1.4 channels with reduced function are present then augmentation therapy may be more appropriate. Specific modulation drugs target different protein defects. Hypomorphs causing trafficking abnormalities and protein instability may be amenable to modulation drugs that increase or decrease protein expression, stabilize the protein, and open or close the channel gate. Drug "correctors" can stabilize protein folding and trafficking to the cell surface [e.g., elexacaftor or tezacaftor used to modulate the cystic fibrosis transmembrane conductance regulator protein (CFTR; Hoy, 2019)], or drug 'potentiators' that can bind proteins that are correctly localised but inactive and help them open to restore ion influx [e.g., ivacaftor used to activate mutant CFTR (Van Goor et al., 2009)]. As for cystic fibrosis, a combination therapy of both correctors and potentiators may be effective to facilitate protein stability, trafficking, and channel activation.

CONCLUSION

Although Cav1 channels function in a similar way, genetic mutations have distinct consequences depending on whether the inheritance pattern is autosomal or X-linked. Functional *in vitro* studies have shown that a change in normal function that either increases or decreases conductance is pathogenic. Understanding the different mechanistic effect of Cav1 mutations reveals that effective therapeutic approaches for X-linked Cav1.4 diseases will be different from autosomal Cav1.1–1.3 channelopathies.

The Cav1.4 α_1 gene has been known for over 20 years and a number of studies have contributed to our understanding of how this channel functions. Since its discovery, nearly 300 pathogenic mutations (92 missense) in *CACNA1F* have been cataloged, with functionally analyzed mutations affecting the channel's ability to conduct calcium ions. Our analyses clarify the genetic mechanisms driving Cav1.4 channelopathies and differentiate the types of mutation and effect they have on channel function. This has highlighted that some mutated residues or domains are likely to have similar hypomorphic or hypermorphic effects across all Cav1 channels.

Our analyses have shown that LoF is the predominant mechanism causing Cav1.4 channelopathies through missense and truncating variations. However, as the number of variants that have been functionally analyzed in Cav1.4 is small, further studies are warranted to better understand the potentially treatable mutation types, as well as facilitating the prediction of variants of unknown significance identified during molecular diagnosis.

DATA AVAILABILITY STATEMENT

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

AUTHOR CONTRIBUTIONS

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

ACKNOWLEDGMENTS

We thank professor Alexandra Koschak for her constructive comments during the preparation of this manuscript. We thank the Manchester Academic Health Science Centre and the Manchester National Institute for Health Research Biomedical Research Centre for support. Supported by Fight for Sight (5079/5080) (TTS).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.637780/full#supplementary-material

REFERENCES

Adzhubei, I. A., Schmidt, S., Peshkin, L., Ramensky, V. E., Gerasimova, A., Bork, P., et al. (2010). A method and server for predicting damaging missense mutations. *Nat. Methods* 7, 248–249. doi: 10.1038/nmeth0410-248

Amberger, J., Bocchini, C. A., Scott, A. F., and Hamosh, A. (2009). McKusick's online mendelian inheritance in man (OMIM®). *Nucleic Acids Res.* 37, D793–D796.

Antzelevitch, C., Pollevick, G. D., Cordeiro, J. M., Casis, O., Sanguinetti, M. C., Aizawa, Y., et al. (2007). Loss-of-function mutations in the cardiac calcium channel underlie a new clinical entity characterized by ST-segment elevation, short QT intervals, and sudden cardiac death. *Circulation* 115, 442–449. doi: 10.1161/circulationaha.106.668392

Baig, S. M., Koschak, A., Lieb, A., Gebhart, M., Dafinger, C., Nürnberg, G., et al. (2011). Loss of Ca v 1.3 (CACNA1D) function in a human channelopathy with bradycardia and congenital deafness. *Nat. Neurosci.* 14, 77–84. doi: 10.1038/nn.2694

Bannister, R. A., and Beam, K. G. (2013). Impaired Gating of an L-Type Ca2+ channel carrying a mutation linked to malignant hyperthermia. *Biophys. J.* 104, 1917–1922. doi: 10.1016/j.bpj.2013.03.035

Bech-Hansen, N. T., Naylor, M. J., Maybaum, T. A., Pearce, W. G., Koop, B., Fishman, G. A., et al. (1998). Loss-of-function mutations in a calcium-channel α1-subunit gene in Xp11.23 cause incomplete X-linked congenital stationary night blindness. *Nat. Genet.* 19, 264–267. doi: 10.1038/947

Béziau, D. M., Barc, J., O'Hara, T., Le Gloan, L., Amarouch, M. Y., Solnon, A., et al. (2014). Complex Brugada syndrome inheritance in a family harbouring compound SCN5A and CACNA1C mutations. *Basic Res. Cardiol.* 109:446.

Boczek, N. J., Best, J. M., Tester, D. J., Giudicessi, J. R., Middha, S., Evans, J. M., et al. (2013). Exome sequencing and systems biology converge to identify novel mutations in the L-type calcium channel, CACNA1C, linked to autosomal dominant long QT syndrome. *Circ. Cardiovasc. Genet.* 6, 279–289. doi: 10.1161/circgenetics.113.000138

Boczek, N. J., Miller, E. M., Ye, D., Nesterenko, V. V., Tester, D. J., Antzelevitch, C., et al. (2015). Novel Timothy syndrome mutation leading to increase in

CACNA1C window current. *Heart Rhythm* 12, 211–219. doi: 10.1016/j.hrthm.2014.09.051

Bozarth, X., Dines, J. N., Cong, Q., Mirzaa, G. M., Foss, K., Lawrence Merritt, J., et al. (2018). Expanding clinical phenotype in CACNA1C related disorders: from neonatal onset severe epileptic encephalopathy to late-onset epilepsy. *Am. J. Med. Genet. A* 176, 2733–2739. doi: 10.1002/ajmg.a.40657

Bulka, B., desJardins, M., and Freeland, S. J. (2006). An interactive visualization tool to explore the biophysical properties of amino acids and their contribution to substitution matrices. *BMC Bioinformatics* 7:329. doi: 10.1186/1471-2105-7-329

Catterall, W. A., Perez-Reyes, E., Snutch, T. P., and Striessnig, J. (2005). International union of pharmacology. XLVIII. Nomenclature and structure-function relationships of voltage-gated calcium channels. *Pharmacol. Rev.* 57, 411–425. doi: 10.1124/pr.57.4.5

Chen, J., Sun, Y., Liu, X., and Li, J. (2019). Identification of a novel mutation in the CACNA1C gene in a Chinese family with autosomal dominant cerebellar ataxia. *BMC Neurol.* 19:157. doi: 10.1186/s12883-019-1381-8

D'Argenio, V., Frisso, G., Precone, V., Boccia, A., Fienga, A., Pacileo, G., et al. (2014). DNA sequence capture and next-generation sequencing for the molecular diagnosis of genetic cardiomyopathies. *J. Mol. Diagn. JMD* 16, 32–44. doi: 10.1016/j.jmoldx.2013.07.008

Di Gregorio, E., Riberi, E., Belligni, E. F., Biamino, E., Spielmann, M., Ala, U., et al. (2017). Copy number variants analysis in a cohort of isolated and syndromic developmental delay/intellectual disability reveals novel genomic disorders, position effects and candidate disease genes. *Clin. Genet.* 92, 415–422. doi: 10.1111/cge.13009

Fiszer, D., Shaw, M.-A., Fisher, N. A., Carr, I. M., Gupta, P. K., Watkins, E. J., et al. (2015). Next-generation Sequencing of RYR1 and CACNA1S in malignant hyperthermia and exertional heat illness. *Anesthesiology* 122, 1033–1046. doi: 10.1097/aln.0000000000000610

Garza-Lopez, E., Lopez, J. A., Hagen, J., Sheffer, R., Meiner, V., and Lee, A. (2018). Role of a conserved glutamine in the function of voltage-gated Ca2+ channels revealed by a mutation in human CACNA1D. *J. Biol. Chem.* 293, 14444–14454. doi: 10.1074/jbc.ra118.003681

Hauke, J., Schild, A., Neugebauer, A., Lappa, A., Fricke, J., Fauser, S., et al. (2013). A novel large in-frame deletion within the CACNA1F gene associates with a cone-rod dystrophy 3-like phenotype. *PLoS One* 8:e76414. doi: 10.1371/journal.pone.0076414

Hemara-Wahanui, A., Berjukow, S., Hope, C. I., Dearden, P. K., Wu, S.-B., Wilson-Wheeler, J., et al. (2005). CACNA1F mutation identified in an X-linked retinal disorder shifts the voltage dependence of Cav1.4 channel activation. *Proc. Natl. Acad. Sci.U.S.A.* 102, 7553–7558. doi: 10.1073/pnas.0501907102

Henikoff, S., and Henikoff, J. G. (1992). Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci.U.S.A.* 89, 10915–10919. doi: 10.1073/pnas.89.22.10915

Hoda, J.-C., Zaghetto, F., Koschak, A., and Striessnig, J. (2005). Congenital stationary night blindness type 2 mutations S229P, G369D, L1068P, and W1440X alter channel gating or functional expression of Ca(v)1.4 L-type Ca2+ channels. *J. Neurosci. Off. J. Soc. Neurosci.* 25, 252–259. doi: 10.1523/jneurosci.3054-04.2005

Hope, C. I., Sharp, D. M., Hemara-Wahanui, A., Sissingh, J. I., Lundon, P., Mitchell, E. A., et al. (2005). Clinical manifestations of a unique X-linked retinal disorder in a large New Zealand family with a novel mutation in CACNA1F, the gene responsible for CSNB2. *Clin. Exp. Ophthalmol.* 33, 129–136. doi: 10.1111/j.1442-9071.2005.00987.x

Hoy, S. M. (2019). Elexacaftor/ivacaftor/tezacaftor: first approval. *Drugs* 79, 2001–2007. doi: 10.1007/s40265-019-01233-7

Hu, H., Kahrizi, K., Musante, L., Fattahi, Z., Herwig, R., Hosseini, M., et al. (2019). Genetics of intellectual disability in consanguineous families. *Mol. Psychiatry* 24, 1027–1039.

Huang, L., Xiao, X., Li, S., Jia, X., Wang, P., Sun, W., et al. (2016). Molecular genetics of cone-rod dystrophy in Chinese patients: new data from 61 probands and mutation overview of 163 probands. *Exp. Eye Res.* 146, 252–258. doi: 10.1016/j.exer.2016.03.015

Hulme, J. T., Yarov-Yarovoy, V., Lin, T. W.-C., Scheuer, T., and Catterall, W. A. (2006). Autoinhibitory control of the CaV1.2 channel by its proteolytically processed distal C-terminal domain. *J. Physiol.* 576, 87–102. doi: 10.1113/jphysiol.2006.111799

Hunter, J. M., Ahearn, M. E., Balak, C. D., Liang, W. S., Kurdoglu, A., Corneveaux, J. J., et al. (2015). Novel pathogenic variants and genes for myopathies identified by whole exome sequencing. *Mol. Genet. Genomic Med.* 3, 283–301.

Jalkanen, R., Bech-Hansen, N. T., Tobias, R., Sankila, E.-M., Mäntyjärvi, M., Forsius, H., et al. (2007). A novel CACNA1F gene mutation causes Aland Island eye disease. *Invest. Ophthalmol. Vis. Sci.* 48, 2498–2502.

Jalkanen, R., Mäntyjärvi, M., Tobias, R., Isosomppi, J., Sankila, E.-M., Alitalo, T., et al. (2006). X linked cone-rod dystrophy, CORDX3, is caused by a mutation in the CACNA1F gene. *J. Med. Genet.* 43, 699–704. doi: 10.1136/jmg.2006.040741

Jespersgaard, C., Fang, M., Bertelsen, M., Dang, X., Jensen, H., Chen, Y., et al. (2019). Molecular genetic analysis using targeted NGS analysis of 677 individuals with retinal dystrophy. *Sci. Rep.* 9:1219.

Jiang, Y., Yuen, R. K. C., Jin, X., Wang, M., Chen, N., Wu, X., et al. (2013). Detection of clinically relevant genetic variants in autism spectrum disorder by whole-genome sequencing. *Am. J. Hum. Genet.* 93, 249–263.

Karczewski, K. J., Francioli, L. C., Tiao, G., Cummings, B. B., Alföldi, J., Wang, Q., et al. (2019). Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes. *BioRxiv* [Preprint] 531210.

Liaqat, K., Schrauwen, I., Raza, S. I., Lee, K., Hussain, S., Chakchouk, I., et al. (2019). Identification of CACNA1D variants associated with sinoatrial node dysfunction and deafness in additional Pakistani families reveals a clinical significance. *J. Hum. Genet.* 64, 153–160. doi: 10.1038/s10038-018-0542-8

Maltese, P. E., Aldanova, E., Kriuchkova, N., Averianov, A., Manara, E., Paolacci, S., et al. (2019). Putative role of Brugada syndrome genes in familial atrial fibrillation. *Eur. Rev. Med. Pharmacol. Sci.* 23, 7582–7598.

Martin-Merida, I., Avila-Fernandez, A., Del Pozo-Valero, M., Blanco-Kelly, F., Zurita, O., Perez-Carro, R., et al. (2019). Genomic landscape of sporadic retinitis pigmentosa: findings from 877 Spanish cases. *Ophthalmology* 126, 1181–1188. doi: 10.1016/j.ophtha.2019.03.018

McRory, J. E., Hamid, J., Doering, C. J., Garcia, E., Parker, R., Hamming, K., et al. (2004). The CACNA1F gene encodes an L-type calcium channel with unique biophysical properties and tissue distribution. *J. Neurosci. Off. J. Soc. Neurosci.* 24, 1707–1718. doi: 10.1523/jneurosci.4846-03.2004

Monnier, N., Procaccio, V., Stieglitz, P., and Lunardi, J. (1997). Malignant-hyperthermia susceptibility is associated with a mutation of the alpha 1-subunit of the human dihydropyridine-sensitive L-type voltage-dependent calcium-channel receptor in skeletal muscle. *Am. J. Hum. Genet.* 60, 1316–1325. doi: 10.1086/515454

Monteleone, S., Lieb, A., Pinggera, A., Negro, G., Fuchs, J. E., Hofer, F., et al. (2017). Mechanisms responsible for ω-pore currents in cav calcium channel voltage-sensing domains. *Biophys. J.* 113, 1485–1495. doi: 10.1016/j.bpj.2017.08.010

Morrill, J. A., Brown, R. H., and Cannon, S. C. (1998). Gating of the L-type Ca channel in human skeletal myotubes: an activation defect caused by the hypokalemic periodic paralysis mutation R528H. *J. Neurosci.* 18, 10320–10334. doi: 10.1523/jneurosci.18-24-10320.1998

Morrill, J. A., and Cannon, S. C. (1999). Effects of mutations causing hypokalaemic periodic paralysis on the skeletal muscle L-type Ca2+ channel expressed in *Xenopus laevis* oocytes. *J. Physiol.* 520, 321–336. doi: 10.1111/j.1469-7793.1999.00321.x

O'Roak, B. J., Vives, L., Girirajan, S., Karakoc, E., Krumm, N., Coe, B. P., et al. (2012). Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations. *Nature* 485, 246–250. doi: 10.1038/nature10989

Peloquin, J. B., Rehak, R., Doering, C. J., and McRory, J. E. (2007). Functional analysis of congenital stationary night blindness type-2 CACNA1F mutations F742C, G1007R, and R1049W. *Neuroscience* 150, 335–345. doi: 10.1016/j.neuroscience.2007.09.021

Pinggera, A., Negro, G., Tuluc, P., Brown, M. J., Lieb, A., and Striessnig, J. (2018). Gating defects of disease-causing de novo mutations in Cav1.3 Ca2+ channels. *Channels* 12, 388–402. doi: 10.1080/19336950.2018.1546518

Ptácek, L. J., Tawil, R., Griggs, R. C., Engel, A. G., Layzer, R. B., Kwiecinski, H., et al. (1994). Dihydropyridine receptor mutations cause hypokalemic periodic paralysis. *Cell* 77, 863–868. doi: 10.1016/0092-8674(94)90135-x

Purcell, S. M., Moran, J. L., Fromer, M., Ruderfer, D., Solovieff, N., Roussos, P., et al. (2014). A polygenic burden of rare disruptive mutations in schizophrenia. *Nature* 506, 185–190. doi: 10.1038/nature12975

Rigaudière, F., Roux, C., Lachapelle, P., Rosolen, S. G., Bitoun, P., Gay-Duval, A., et al. (2003). ERGs in female carriers of incomplete Congenital Stationary Night Blindness (I-CSNB) a family report. *Doc. Ophthalmol.* 107, 203–212.

Ross, J., Gedvilaite, E., Badner, J. A., Erdman, C., Baird, L., Matsunami, N., et al. (2016). Rare variant in CACNA1D segregates with 7 bipolar I disorder cases in a large pedigree. *Mol. Neuropsychiatry* 2, 145–150. doi: 10.1159/000448041

Roussos, P., Mitchell, A. C., Voloudakis, G., Fullard, J. F., Pothula, V. M., Tsang, J., et al. (2014). A role for noncoding variation in schizophrenia. *Cell Rep.* 9, 1417–1429.

Scholl, U. I., Goh, G., Stölting, G., de Oliveira, R. C., Choi, M., Overton, J. D., et al. (2013). Somatic and germline CACNA1D calcium channel mutations in aldosterone-producing adenomas and primary aldosteronism. *Nat. Genet.* 45, 1050–1054. doi: 10.1038/ng.2695

Semenova, N. A., Ryzhkova, O. R., Strokova, T. V., and Taran, N. N. (2018). [The third case report a patient with primary aldosteronism, seizures, and neurologic abnormalities (PASNA) syndrome de novo variant mutations in the CACNA1D gene]. *Zh. Nevrol. Psikhiatr. Im. S S Korsakova* 118, 49–52. doi: 10.17116/jnevro201811812149

Sharma, J., Keeling, K. M., and Rowe, S. M. (2020). Pharmacological approaches for targeting cystic fibrosis nonsense mutations. *Eur. J. Med. Chem.* 200, 112436. doi: 10.1016/j.ejmech.2020.112436

Sievers, F., Wilm, A., Dineen, D., Gibson, T. J., Karplus, K., Li, W., et al. (2011). Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.* 7:539. doi: 10.1038/msb.2011.75

Singh, A., Hamedinger, D., Hoda, J.-C., Gebhart, M., Koschak, A., Romanin, C., et al. (2006). C-terminal modulator controls Ca^{2+}-dependent gating of $Ca_v1.4$ L-type Ca^{2+} channels. *Nat. Neurosci.* 9, 1108–1116.

Song, J., Smaoui, N., Ayyagari, R., Stiles, D., Benhamed, S., MacDonald, I. M., et al. (2011). High-throughput retina-array for screening 93 genes involved in inherited retinal dystrophy. *Invest. Ophthalmol. Vis. Sci.* 52, 9053–9060. doi: 10.1167/iovs.11-7978

Splawski, I., Timothy, K. W., Decher, N., Kumar, P., Sachse, F. B., Beggs, A. H., et al. (2005). Severe arrhythmia disorder caused by cardiac L-type calcium channel mutations. *Proc. Natl. Acad. Sci. U.S.A.* 102, 8089–8096. doi: 10.1073/pnas.0502506102

Splawski, I., Timothy, K. W., Sharpe, L. M., Decher, N., Kumar, P., Bloise, R., et al. (2004). Ca(V)1.2 calcium channel dysfunction causes a multisystem disorder including arrhythmia and autism. *Cell* 119, 19–31. doi: 10.1016/j.cell.2004.09.011

Stenson, P. D., Mort, M., Ball, E. V., Evans, K., Hayden, M., Heywood, S., et al. (2017). The human gene mutation database: towards a comprehensive repository of inherited mutation data for medical research, genetic diagnosis and next-generation sequencing studies. *Hum. Genet.* 136, 665–677. doi: 10.1007/s00439-017-1779-6

Strom, T. M., Nyakatura, G., Apfelstedt-Sylla, E., Hellebrand, H., Lorenz, B., Weber, B. H. F., et al. (1998). An L-type calcium-channel gene mutated in incomplete X-linked congenital stationary night blindness. *Nat. Genet.* 19, 260–263. doi: 10.1038/940

Sun, W., Huang, L., Xu, Y., Xiao, X., Li, S., Jia, X., et al. (2015). Exome sequencing on 298 probands with early-onset high myopia: approximately one-fourth show potential pathogenic mutations in retnet genes. *Invest. Ophthalmol. Vis. Sci.* 56, 8365–8372. doi: 10.1167/iovs.15-17555

The UniProt Consortium (2018). UniProt: the universal protein knowledgebase. *Nucleic Acids Res.* 46, 2699–2699. doi: 10.1093/nar/gky092

Torrico, B., Shaw, A. D., Mosca, R., Vivó-Luque, N., Hervás, A., Fernàndez-Castillo, N., et al. (2019). Truncating variant burden in high-functioning autism and pleiotropic effects of LRP1 across psychiatric phenotypes. *J. Psychiatry Neurosci. JPN* 44, 350–359. doi: 10.1503/jpn.180184

Tumienė, B., Maver, A., Writzl, K., Hodžić, A., Čuturilo, G., Kuzmanić-Šamija, R., et al. (2018). Diagnostic exome sequencing of syndromic epilepsy patients in clinical practice. *Clin. Genet.* 93, 1057–1062. doi: 10.1111/cge.13203

Van Goor, F., Hadida, S., Grootenhuis, P. D. J., Burton, B., Cao, D., Neuberger, T., et al. (2009). Rescue of CF airway epithelial cell function in vitro by a CFTR potentiator, VX-770. *Proc. Natl. Acad. Sci. U.S.A.* 106, 18825–18830. doi: 10.1073/pnas.0904709106

Vivante, A., Ityel, H., Pode-Shakked, B., Chen, J., Shril, S., van der Ven, A. T., et al. (2017). Exome sequencing in Jewish and Arab patients with rhabdomyolysis reveals single-gene etiology in 43% of cases. *Pediatr. Nephrol. Berl. Ger.* 32, 2273–2282. doi: 10.1007/s00467-017-3755-8

Xu, Y., Guan, L., Xiao, X., Zhang, J., Li, S., Jiang, H., et al. (2015). Mutation analysis in 129 genes associated with other forms of retinal dystrophy in 157 families with retinitis pigmentosa based on exome sequencing. *Mol. Vis.* 21, 477–486.

Yang, J., Elllnor, P. T., Sather, W. A., Zhang, J.-F., and Tsien, R. W. (1993). Molecular determinants of Ca 2+ selectivity and ion permeation in L-type Ca 2+ channels. *Nature* 366, 158–161. doi: 10.1038/366158a0

Zamponi, G. W., Striessnig, J., Koschak, A., and Dolphin, A. C. (2015). The physiology, pathology, and pharmacology of voltage-gated calcium channels and their future therapeutic potential. *Pharmacol. Rev.* 67, 821–870. doi: 10.1124/pr.114.009654

Zenagui, R., Lacourt, D., Pegeot, H., Yauy, K., Juntas Morales, R., Theze, C., et al. (2018). reliable targeted next-generation sequencing strategy for diagnosis of myopathies and muscular dystrophies, especially for the giant titin and nebulin genes. *J. Mol. Diagn. JMD* 20, 533–549. doi: 10.1016/j.jmoldx.2018.04.001

Zhu, Y.-B., Luo, J.-W., Jiang, F., and Liu, G. (2018). Genetic analysis of sick sinus syndrome in a family harboring compound CACNA1C and TTN mutations. *Mol. Med. Rep.* 17, 7073–7080.

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.

The handling editor declared a past co-authorship with one of the authors GB.

frontiers in Genetics

ORIGINAL RESEARCH
published: 05 August 2021
doi: 10.3389/fgene.2021.701652

Splicing Outcomes of 5′ Splice Site GT>GC Variants That Generate Wild-Type Transcripts Differ Significantly Between Full-Length and Minigene Splicing Assays

Jin-Huan Lin[1,2†], Hao Wu[1,2†], Wen-Bin Zou[1,2†], Emmanuelle Masson[3,4], Yann Fichou[3,5], Gerald Le Gac[3,4,5], David N. Cooper[6], Claude Férec[3,4], Zhuan Liao[1,2] and Jian-Min Chen[3*]*

[1] Department of Gastroenterology, Changhai Hospital, Second Military Medical University, Shanghai, China, [2] Shanghai Institute of Pancreatic Diseases, Shanghai, China, [3] Univ Brest, Inserm, EFS, UMR 1078, GGB, Brest, France, [4] Service de Génétique Médicale et de Biologie de la Reproduction, CHRU Brest, Brest, France, [5] Laboratory of Excellence GR-Ex, Paris, France, [6] Institute of Medical Genetics, School of Medicine, Cardiff University, Cardiff, United Kingdom

OPEN ACCESS

Edited by:
Stephen J. Bush,
University of Oxford, United Kingdom

Reviewed by:
Alex Star Nord,
University of California, Davis, United States
Ludovica Celli,
Institute of Molecular Genetics, Italian National Research Council, Italy

***Correspondence:**
Zhuan Liao
liaozhuan@smmu.edu.cn
Jian-Min Chen
jian-min.chen@univ-brest.fr

†These authors share first authorship

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

Received: 28 April 2021
Accepted: 13 July 2021
Published: 05 August 2021

Citation:
Lin J-H, Wu H, Zou W-B, Masson E, Fichou Y, Le Gac G, Cooper DN, Férec C, Liao Z and Chen J-M (2021) Splicing Outcomes of 5′ Splice Site GT>GC Variants That Generate Wild-Type Transcripts Differ Significantly Between Full-Length and Minigene Splicing Assays. Front. Genet. 12:701652. doi: 10.3389/fgene.2021.701652

Combining data derived from a meta-analysis of human disease-associated 5′ splice site GT>GC (i.e., +2T>C) variants and a cell culture-based full-length gene splicing assay (FLGSA) of forward engineered +2T>C substitutions, we recently estimated that ∼15–18% of +2T>C variants can generate up to 84% wild-type transcripts relative to their wild-type counterparts. Herein, we analyzed the splicing outcomes of 20 +2T>C variants that generate some wild-type transcripts in two minigene assays. We found a high discordance rate in terms of the generation of wild-type transcripts, not only between FLGSA and the minigene assays but also between the different minigene assays. In the pET01 context, all 20 wild-type minigene constructs generated the expected wild-type transcripts; of the 20 corresponding variant minigene constructs, 14 (70%) generated wild-type transcripts. In the pSPL3 context, only 18 of the 20 wild-type minigene constructs generated the expected wild-type transcripts whereas 8 of the 18 (44%) corresponding variant minigene constructs generated wild-type transcripts. Thus, in the context of a particular type of variant, we raise awareness of the limitations of minigene splicing assays and emphasize the importance of sequence context in regulating splicing. Whether or not our findings apply to other types of splice-altering variant remains to be investigated.

Keywords: aberrant transcript, full-length gene splicing assay, genetic variant, minigene splicing assay, splice site, SpliceAI

INTRODUCTION

In principle, both coding and intronic variants within a gene have the potential to affect splicing (Cooper et al., 2009; Scotti and Swanson, 2016; Anna and Monika, 2018; Truty et al., 2021). Genetic variants occurring within the 5′ splice site GT dinucleotide, whenever found in disease-causing or disease-predisposing genes, have generally been classified as pathogenic (Mount et al., 2019;

Stenson et al., 2020). However, a problem is posed by 5′ splice site GT>GC variants (henceforth simply termed +2T>C variants) due to the fact that in the human genome, a small but nevertheless significant minority (∼1%) of introns contain the 5′ splice site GC dinucleotide (Burset et al., 2000, 2001; Abril et al., 2005; Sheth et al., 2006; Parada et al., 2014). Recently, combining data derived from a meta-analysis of human inherited disease-associated +2T>C variants and a cell culture-based Full-Length Gene Splicing Assay (FLGSA) of forward engineered +2T>C substitutions, we estimated that ∼15–18% of +2T>C variants can generate up to 84% wild-type transcripts relative to their wild-type counterparts (Lin et al., 2019). This finding was corroborated by a re-analysis (Chen et al., 2020) of the saturation genome editing data on 12 *BRCA1* +2T>C substitutions (Findlay et al., 2018).

Our aforementioned findings have two direct clinical implications. Firstly, many +2T>C variants in human disease genes that have been capable of generating some wild-type transcripts are likely to have gone largely unreported; this represents a significant deficiency in terms of our understanding of genotype-phenotype relationships and tailored treatment options given that even the minor retention of wild-type transcripts derived from a variant allele might significantly impact disease expression and severity (Ramalho et al., 2002; Den Uijl et al., 2011; Raraigh et al., 2018; Lin et al., 2019; Scalet et al., 2019; Joynt et al., 2020). In this regard, it is pertinent to mention that *CFTR* c.3873+2T>C and c.4242+2T>C transitions (Joynt et al., 2020) and *SRP68* c.184+2T>C (Schmaltz-Panneau et al., 2021) are among the most recently reported examples of disease-causing +2T>C variants that generated some wild-type transcripts. Secondly, +2T>C variants in human disease genes may not invariably be pathogenic, a notion that has received support from at least two recent publications, which reclassified *BRCA2* c.8331+2T>C (Nix et al., 2021) and *BAP1* c.783+2T>C (Goldberg et al., 2021) as variants of unknown significance.

Another important finding arising from our study was that none of the widely used splicing prediction tools were capable of reliably distinguishing those +2T>C variants that generated wild-type transcripts from those that did not (Lin et al., 2019). The root of this problem is twofold: apart from the use of GC instead of GT as the 5′ splice site dinucleotide in ∼1% of introns, these prediction tools only take into consideration short local DNA sequence motifs (Chen et al., 2020). The recently developed deep learning-based tool, SpliceAI (Jaganathan et al., 2019), performed somewhat better in this regard but was still far from perfect (Chen et al., 2020). These observations underscored the importance of experimentally determining the splicing outcomes of +2T>C variants in a clinical as well as a basic research setting. Whilst RNA analysis, using pathophysiologically relevant tissues, provides the most accurate and reliable mRNA phenotyping information on human splicing variants, this is often not possible if appropriate tissue samples are not available (Aicher et al., 2020). RNA analysis using either patient blood cells or immortalized lymphoblastoid cells represents an alternative option, providing that the gene of interest is normally expressed in these cells (Wai et al., 2020). In case of the non-feasibility of both approaches, a cell culture-based minigene splicing assay has often been devised (for some most recent examples, see Damasio et al., 2021; Hao et al., 2021; Kim et al., 2021; Kortum et al., 2021; Le Tertre et al., 2021; Morbidoni et al., 2021; Qian et al., 2021; Saint-Martin et al., 2021; Torrado et al., 2021).

Our FLGSA assay (focused on genes whose genomic sizes were < 8 kb) (Lin et al., 2019, 2020) cannot be readily used for large genes for various practical and/or technical reasons. Genome editing (Findlay et al., 2018) is a promising trend but its wide application is still some way from becoming reality. Thus, the minigene splicing assay will for the time being remain the mainstream approach for functionally characterizing potential splice-altering variants. However, an inherent drawback of the minigene splicing assay is the lack of the wider genomic sequence context of the gene under study (Zou et al., 2016; Lin et al., 2019, 2020; Tang et al., 2019). This could lead to inaccurate results and incorrect conclusions being drawn owing to the complexity of the splicing code (Fu and Ares, 2014; Drexler et al., 2020), as exemplified by the contrasting findings from the study of the *SPINK1* c.194G>A variant in a minigene assay (Beer and Sahin-Toth, 2014) and our own FLGSA assay (Wu et al., 2017). Herein, we explored whether the splicing outcomes of 20 +2T>C variants that have been previously shown to generate some wild-type transcripts by means of FLGSA and/or patient RNA analysis (Lin et al., 2019) could be replicated in two minigene assays.

MATERIALS AND METHODS

+2T>C Variants Included for Minigene Splicing Assay and Variant Nomenclature

A total of 26 +2T>C variants were previously shown to generate some wild-type transcripts by means of FLGSA and/or patient RNA analyses (Lin et al., 2019). Of these, six variants that occurred within the first intron of their respective genes could not be readily analyzed by the minigene assay and hence were excluded from further consideration.

All the remaining 20 variants were included in the current analysis (**Table 1**). Of these, six had been originally reported to be both naturally occurring and disease causing. These six pathogenic variants included the five variants that had previously been demonstrated to generate some wild-type transcripts by means of patient RNA analysis (i.e., *CD3E* IVS7+2T>C, *CD40LG* IVS3+2T>C, *DMD* IVS54+2T>C, *PLP1* IVS5+2T>C and *SPINK1* IVS3+2T>C) plus *HBB* IVS2+2T>C. Although the latter *HBB* IVS2+2T>C variant had no accompanying patient RNA data, it was suggested to have had a limited impact on splicing due to its associated hematological phenotype that was milder than would have been expected from a null allele (Frischknecht et al., 2009); its orthologous counterpart in the rabbit *Hbb* gene has been experimentally shown to generate wild-type transcripts (Aebi et al., 1986, 1987); and the human variant was also shown to generate wild-type transcripts in a FLGSA assay (Lin et al., 2019). Of the five pathogenic variants subjected to patient RNA analysis, only *SPINK1* IVS3+2T>C was also analyzed by FLGSA; the findings from the patient RNA analysis

(Kume et al., 2006) and FLGSA assay (Zou et al., 2016; Lin et al., 2019) were remarkably similar.

The remaining 14 variants included in this study were not known to be disease causing at the time (Lin et al., 2019). They represent forward engineered +2T>C substitutions, all being found to generate wild-type transcripts by means of FLGSA (Lin et al., 2019).

For ease of description and to be consistent with our previous publications (Lin et al., 2019, 2020; Chen et al., 2020), all included +2T>C variants were described in accordance with the traditional IVS (intervening sequence or intron) nomenclature (**Table 1**). Their respective chromosome locations, hg38 coordinates, reference alleles in hg38 and HGVS nomenclature (den Dunnen et al., 2016) are, however, provided in **Supplementary Table 1**.

Construction of pET01 and pSPL3 Wild-Type Minigene Expression Vectors by Means of In-Fusion Cloning

For a given +2T>C variant, the corresponding wild-type genomic sequences cloned into the pET01 and pSPL3 exon trapping vectors were always identical. Of the 20 wild-type inserts, 18 comprised 63-330 bp sequence from the 3′ end of N-1 intron, the entire exon N and 65–328 bp sequence from the 5′ end of intron N (N is the number of the variant-affected intron) (see upper panel in **Figure 1A**). The other two wild-type inserts (for *PSMC5* IVS10+2T>C and *SELENOS* IVS5+2>C, respectively) instead comprised 50–116 bp sequence from intron N-2, entire exon N-1, entire intron N-1, entire exon N and 96–294 bp sequence from intron N (lower panel in **Figure 1A**); this was done primarily due to the small size (<100 bp) of the respective intron N-1 in these two cases. See **Supplementary Table 2** for the sequences of all inserts.

Two pairs of 5′ *Xho*I-harboring and 3′ *BamH*I-harboring primers, one for in-fusion cloning into the pET01 trapping vector and the other for in-fusion cloning into the pSPL3 exon trapping vector, were designed to amplify each insert (**Figure 1B**). Primer sequences are provided in **Supplementary Table 3**. PCR was performed in a 25 μL reaction mixture containing 0.5 U KAPA HiFi HotStart DNA Polymerase (Kapa Biosystems), 0.75 μL KAPA dNTP Mix (300 μM final), 5 μL 5× KAPA HiFi Buffer, 50 ng DNA (from a healthy Chinese subject), and 0.3 μM forward and reverse primers. The PCR program comprised an initial denaturation at 95°C for 5 min, followed by 30 cycles of denaturation at 98°C for 20 s, annealing at 66°C for 15 s, extension at 72°C for 1 min, and a final extension at 72°C for 5 min.

PCR products of the expected size were purified with the Cloning Enhancer kit (TaKaRa). The purified products were

TABLE 1 | Results from minigene splicing analyses of 20 +2T>C variants that were previously reported to generate wild-type transcripts.

Gene	mRNA reference	Variant[a]	Summary of previous data		Current results[c]	
			Analytical method employed	Expression level of the wild-type transcript[b]	pET01	pSPL3
CD3E	NM_000733.3	IVS7+2T>C	Patient RNA analysis	1-5%	–	–
CD40LG	NM_000074.2	IVS3+2T>C	Patient RNA analysis	15%	+	+
DBI	NM_001079862.2	IVS2+2T>C	FLGSA[d]	Not quantified	+	–
DMD	NM_004006.2	IVS54+2T>C	Patient RNA analysis	10%	+	+
DNAJC19	NM_145261.3	IVS5+2T>C	FLGSA	42%	+	×
FOLR3	NM_000804.3	IVS4+2T>C	FLGSA	Not quantified	+	+
HBB	NM_000518.5	IVS2+2T>C	FLGSA	Not quantified	+	–
IFNL2	NM_172138.1	IVS5+2T>C	FLGSA	5%	+	+
IL10	NM_000572.3	IVS3+2T>C	FLGSA	Not quantified	+	+
MGP	NM_000900.4	IVS2+2T>C	FLGSA	80%	–	–
PLP1	NM_000533.4	IVS5+2T>C	Patient RNA analysis	8%	–	–
PSMC5	NM_001199163.1	IVS6+2T>C	FLGSA	56%	–	–
		IVS8+2T>C	FLGSA	56%	–	–
		IVS10+2T>C	FLGSA	46%	+	–
RPL11	NM_000975.5	IVS2+2T>C	FLGSA	Not quantified	+	+
		IVS3+2T>C	FLGSA	Not quantified	+	+
RPS27	NM_001030.4	IVS2+2T>C	FLGSA	63%	–	–
		IVS3+2T>C	FLGSA	Not quantified	+	×
SELENOS	NM_203472.2	IVS5+2T>C	FLGSA	14%	+	+
SPINK1	NM_003122.3	IVS3+2T>C	Patient RNA analysis/FLGSA	10%	+	–

[a] *In accordance with the traditional IVS (intervening sequence or intron) nomenclature as previously described (Lin et al., 2019, 2020). See* ***Supplementary Table 1*** *for additional information including specifically the HGVS nomenclature (den Dunnen et al., 2016).*

[b] *Expression level of the wild-type transcripts generated from the variant allele relative to that (set as 1) generated from the wild-type allele (see Lin et al., 2019; Chen et al., 2020).*

[c] *+, the variant generated wild-type transcripts; –, the variant did not generate wild-type transcripts; ×, the variant was not analyzed by the minigene assay owing to the non-expression of the wild-type transcripts from the corresponding wild-type minigene vectors.*

[d] *FLGSA, full-length gene splicing assay.*

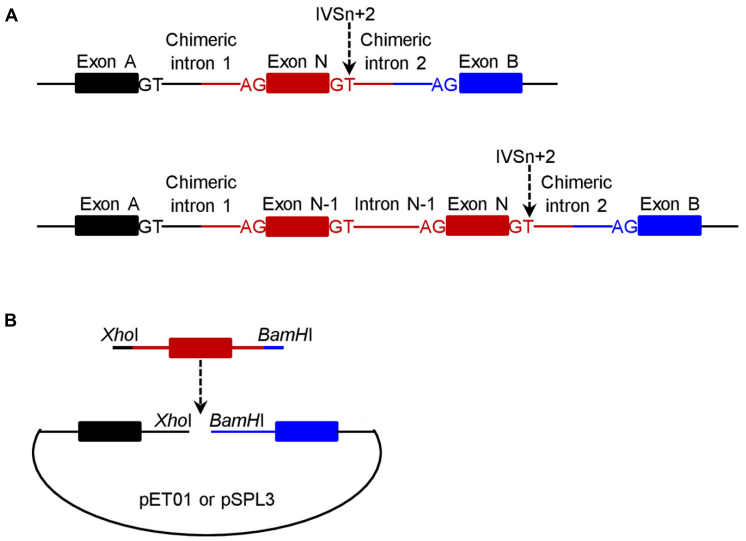


FIGURE 1 | Schematic illustrations of the minigene expression constructs in the context of wild-type inserts. **(A)** Illustration of the target gene sequence (in red) inserted into the pET01 and pSPL3 exon trapping vectors. The number of the +2T>C variant-affected intron is N and the IVSn+2T site is indicated by a downward pointing arrow. For most variants (n = 18), the insert comprised a single exon (i.e., exon N) plus flanking intronic sequences on both sides (upper panel). For two variants, the insert comprised two exons (exon N-1 and exon N) plus flanking and intervening intronic sequences (lower panel). The two exons (exon A and exon B) located within the vector and the canonical splice donor GT and acceptor AG sites defining the two chimeric introns (upper panel) or the two chimeric introns and intron N-1 (lower panel) are also denoted in the figure. **(B)** Illustration of how a wild-type minigene expression vector was constructed. The insert (in red) was PCR amplified with 5′ *Xho*I-harboring and 3′ *BamH*I-harboring in-fusion primers with respect to the pET01 or pSPL3 vector. The resulting PCR products were inserted into their respective linearized vectors by means of in-fusion cloning.

then cloned into the *Xho*I and *BamH*I restriction sites of the linearized pET01 or pSPL3 vector with the In-Fusion HD Cloning kit (TaKaRa) according to the manufacturer's instructions. Transformation was performed using Stellar Competent Cells (TaKaRa). Transformed cells were spread onto LB agar plates with 50 μg/mL ampicillin and incubated at 37°C overnight. Plasmid constructs containing inserts were confirmed by Sanger sequencing.

Primers were designed by our laboratory in the Changhai Hospital. Primer synthesis, insert amplification, in-fusion cloning and verification of the inserted fragments were all performed by GENEWIZ, Beijing, China.

Generation of pET01 and pSPL3 +2T>C Variant Minigene Expression Vectors by Means of Site-Directed Mutagenesis

+2T>C variants were introduced into their respective wild-type minigene expression constructs by means of the QuikChange II XL Site-Directed Mutagenesis Kit (Agilent Technologies). Mutagenesis, transformation, plasmid preparation and validation of the introduced variants were performed as previously described (Lin et al., 2019). Sequences of the mutagenesis primers are provided in **Supplementary Table 4**.

Cell Culture, Transfection, RNA Extraction, and Reverse Transcription

These were performed as previously described (Lin et al., 2019).

Reverse Transcription-Polymerase Chain Reaction (RT-PCR) Analysis

RT-PCR was performed in a 25-μL reaction mixture containing 12.5 μL HotStarTaq Master Mix (Qiagen), 1 μL cDNA, and 0.4 μM each primer (5′-GAGGGATCCGCTTCCTGGCCC-3′ (forward) and 5′-CTCCCGGGCCACCTCCAGTGCC-3′ (reverse) for pET01 expression vectors (both primers are located within the pET01 vector sequence); 5′-TCTGAGTCACCTGGACAACC-3′ (forward) and 5′-ATCTCAGTGGTATTTGTGAGC-3′ (reverse) for pSPL3 expression vectors (both primers are located within the pSPL3 vector sequence)). The PCR program had an initial denaturation step at 95°C for 15 min, followed by 30 cycles of denaturation at 94°C for 45 s, annealing at 58°C for 45 s, extension at 72°C for 1 min/kb, and a final extension step at 72°C for 10 min. RT-PCR products of a single band were cleaned by ExoSAP-IT (Affymetrix). In the case of multiple bands, the bands were excised from the agarose gel and then purified by QIAquick Gel Extraction Kit (Qiagen). Sequencing primers were those used for the RT-PCR analyses and sequencing was performed using the BigDye Terminator v1.1 Cycle Sequencing Kit (Applied Biosystems).

RESULTS

Rationale of Experimental Protocol

The experimental procedures adopted in this study are summarized in **Figure 2**. Before presenting the results obtained

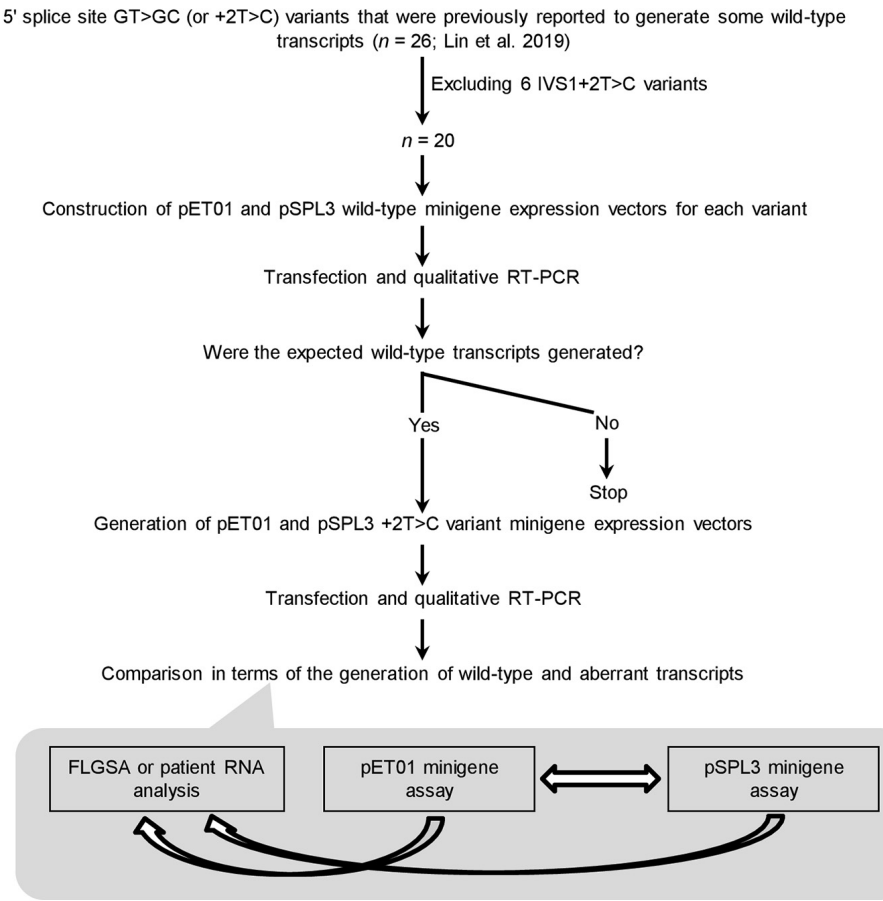


FIGURE 2 | Outline of the experimental procedures. IVS, intervening sequence or intron; RT-PCR, reverse transcription-polymerase chain reaction. FLGSA, full-length gene splicing assay.

in some detail, we would like to summarize our protocol in terms of its four components.

First, the 20 +2T>C variants included for minigene assay analysis represent the totality of the variants that were previously found to generate wild-type transcripts as assessed by FLGSA and/or patient RNA analyses excluding all IVS1+2T>C ones (Lin et al., 2019). (NB. Intronic variants located near the first or last exon of the gene cannot be readily evaluated by a minigene assay without special adaptation (Chen et al., 2018; Raud et al., 2019; Tang et al., 2019).) The accuracy and reliability of the FLGSA-obtained functional assessment of the +2T>C variants have been extensively addressed in our previous publications (Lin et al., 2019, 2020; Chen et al., 2020).

Second, for each +2T>C variant under study, the corresponding wild-type genomic sequences inserted into the two minigene vectors, pET01 and pSPL3, were always identical.

Third, the generation (or not) of wild-type transcripts—determined by qualitative RT-PCR analyses as previously described (Lin et al., 2019)—was used as the basis for comparison.

Fourth, a wild-type transcript refers to the product containing precisely Exon A, Exon N and Exon B or Exon A, Exon N-1, Exon N and Exon B as depicted in **Figure 1A**. The authenticity of all wild-type transcripts was confirmed by Sanger sequencing. Most aberrant transcripts were also Sanger sequenced.

Generation (or Not) of Wild-Type Transcripts

Two Exceptions to the Rule That Wild-Type Minigene Constructs Invariably Express the Expected Wild-Type Transcripts

Of the 40 wild-type minigene constructs (20 in the pET01 context and 20 in the pSPL3 context), only two did not express the expected wild-type transcripts (**Supplementary Figures 1–20**), both of them in the pSPL3 context. Specifically, the pSPL3 *DNAJC19* IVS5+2T minigene construct expressed a transcript lacking *DNAJC19* exon 5 but containing instead a 118-bp pseudoexon (**Figure 3**). The pSPL3 *RPS27* IVS3+2T minigene construct expressed a transcript with *RPS27* exon 3 being skipped (**Supplementary Figure 18**). These two wild-type minigene constructs were thus not mutated to their corresponding variant versions. In other words, *DNAJC19* IVS5+2T>C and pSPL3 *RPS27* IVS3+2T>C were not analyzed by the minigene assay in the pSPL3 context.

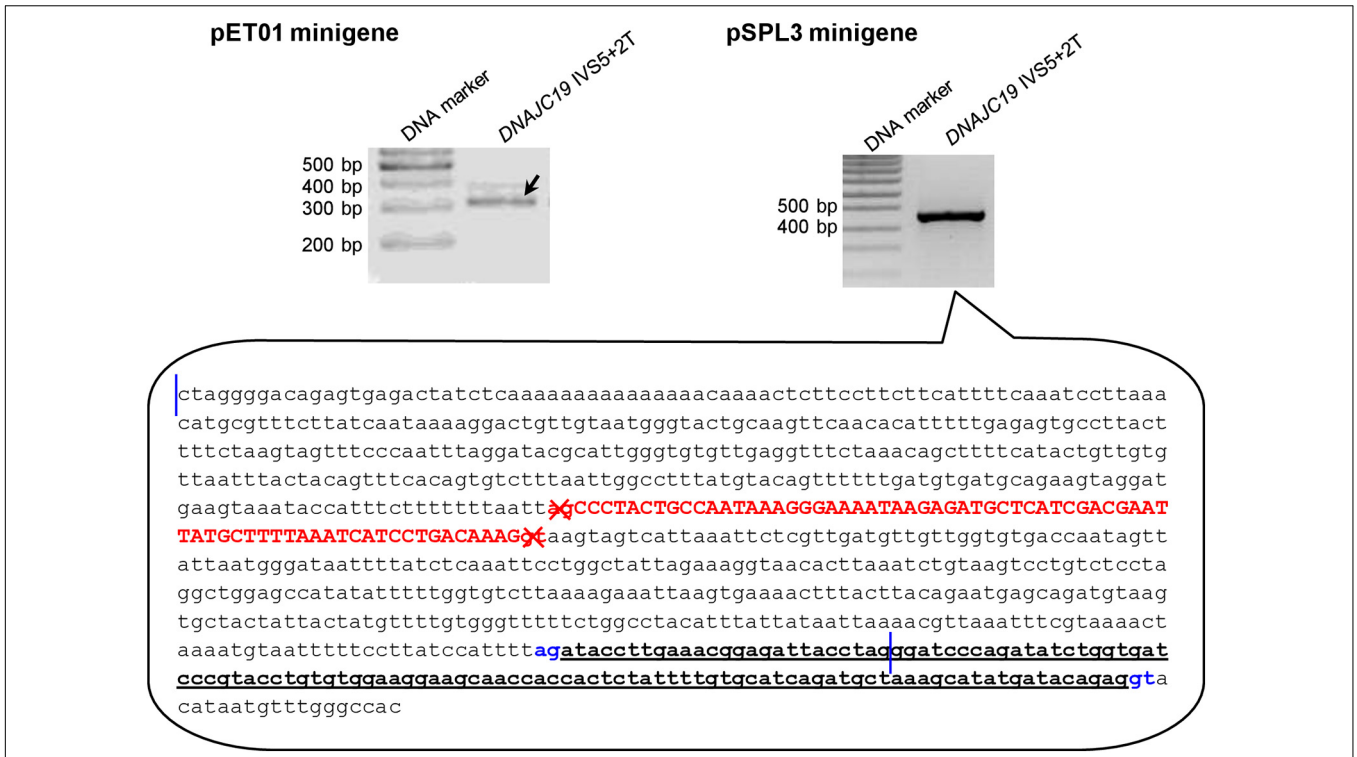


FIGURE 3 | Reverse transcription-polymerase chain reaction results from minigene assays with respect to the *DNAJC19* IVS5+2T (i.e., wild-type) construct. In the pET01 minigene assay, the *DNAJC19* IVS5+2T construct yielded the expected wild-type transcripts (indicated by the left oblique downward pointing arrow). In the pSPL3 minigene assay, the *DNAJC19* IVS5+2T construct yielded an aberrant transcript, whose nature was illustrated in the call-out box. In the call-out box, the sequence delimited by the two vertical blue lines refers to the entire wild-type *DNAJC19* DNA insert, which comprised exon 5 (in upper case and in red) and partial intronic sequences on both sides. The sequence downstream of the second vertical line refers to partial downstream pSPL3 vector sequence. The aberrant transcript did not contain *DNAJC19* exon 5 but instead contained a 118-bp pseudoexon (in bold and underlined) that spanned the chimeric region of the chimeric intron 2 (see **Figure 1A** for term definition). The aberrantly inactivated AG-GT splice sites flanking *DNAJC19* exon 5 are highlighted in red and denoted by crosses. The aberrantly activated cryptic AG-GT splice sites are highlighted in blue. See **Supplementary Figure 5** for the full gel photographs.

Four +2T>C Variants Exhibited Discordance Between the Two Minigene Assays

Four variants, *DBI* IVS2+2T>C (**Figure 4**), *HBB* IVS2+2>C (**Supplementary Figure 7**), *PSMC5* IVS10+2T>C (**Supplementary Figure 14**) and *SPINK1* IVS3+2T>C (**Supplementary Figure 20**), generated wild-type transcripts in the pET01 minigene assay but not in the pSPL3 minigene assay.

Six +2T>C Variants Failed to Generate Wild-Type Transcripts in Both Minigene Assays

The six variants are *CD3E* IVS7+2T>C (**Supplementary Figure 1**), *MGP* IVS2+2T>C (**Supplementary Figure 10**), *PLP1* IVS5+2T>C (**Supplementary Figure 11**), *PSMC5* IVS6+2T>C (**Supplementary Figure 12**), *PSMC5* IVS8+2T>C (**Supplementary Figure 13**) and *RPS27* IVS2+2T>C (**Figure 5**).

Eight Variants Generated Wild-Type Transcripts in Both Minigene Assays

These eight variants are *CD40LG* IVS3+2T>C (**Supplementary Figure 2**), *DMD* IVS54+2T>C (**Supplementary Figure 4**), *FOLR3* IVS4+2T>C (**Figure 6**), *IFNL2* IVS5+2T>C (**Supplementary Figure 8**), *IL10* IVS3+2T>C (**Supplementary Figure 9**), *RPL11* IVS2+2T>C (**Supplementary Figure 15**), *RPL11* IVS3+2T>C (**Supplementary Figure 16**) and *SELENOS* IVS5+2T>C (**Supplementary Figure 19**).

Synthesis

The above findings are summarized in **Table 1**. In short, in the pET01 context, all 20 wild-type minigene constructs generated the expected wild-type transcripts; of the 20 corresponding variant minigene constructs, 14 (70%) generated wild-type transcripts. In the pSPL3 context, only 18 of the 20 wild-type minigene constructs generated the expected wild-type transcripts; of the 18 corresponding variant minigene constructs, only 8 (44%) generated wild-type transcripts.

DISCUSSION

In this study, we set out to systematically analyze the splicing outcomes of 20 +2T>C variants that had been previously shown to generate varying levels of wild-type transcripts by means of FLGSA and/or patient RNA analyses (Lin et al., 2019), in two minigene systems. We found a fairly high level of discordance between the different systems in terms of the generation of wild-type transcripts (**Table 1**). First and foremost,

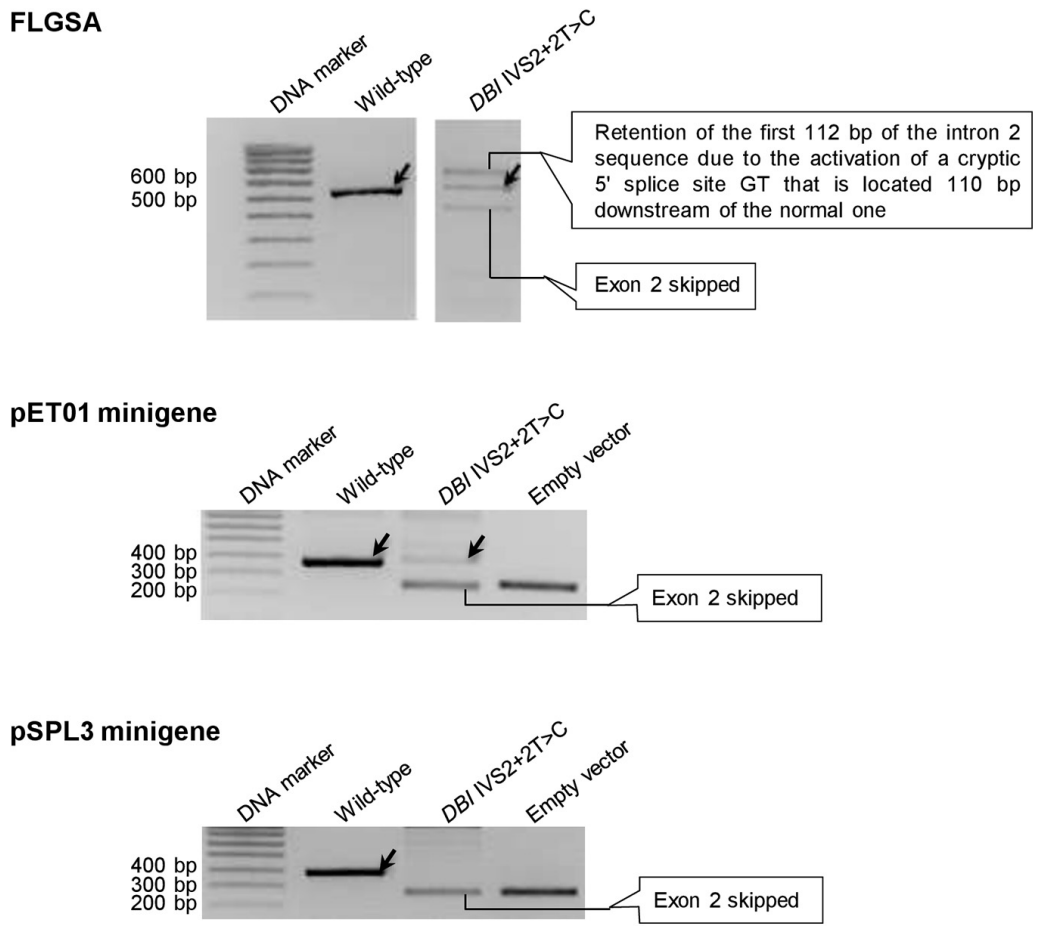


FIGURE 4 | Reverse transcription-polymerase chain reaction (RT-PCR) results from the pET01 and pSPL3 minigene assays with respect to the *DBI* IVS2+2T>C variant. Results from the previously performed full-length gene splicing assay (FLGSA) (Lin et al., 2019) are included for the sake of comparison (NB. The two aberrant transcripts were newly sequenced in this study). In all panels, wild-type transcripts are indicated by oblique downward pointing arrows. See **Supplementary Figure 3** for the full gel photographs with respect to the minigene assays. The FLGSA data were adapted from Lin et al. (2019) with permission (Copyright 2020 Wiley Periodicals LLC).

30% (n = 6) of the 20 +2T>C variants analyzed in the pET01 minigene assay and 56% (n = 10) of the 18 +2T>C variants analyzed in the pSPL3 minigene assay failed to generate wild-type transcripts. It would thus appear that the minigene assays have a tendency to exaggerate the negative effect of the +2T>C variants on splicing. Whether this is bound up with the artificiality of the minigene structure or simply represents a chance finding, remains to be established. In line with our own findings, the aforementioned reclassified *BRCA2* c.8331+2T>C variant (Nix et al., 2021) had been previously found to generate no wild-type transcripts at all by means of a minigene assay (Fraile-Bethencourt et al., 2017). However, using exonic tag-SNP analysis of transcripts expressed in Epstein-Barr virus-immortalized lymphoblastoid cells from a heterozygous *BRCA2* c.8331+2T>C carrier, Gelli et al. (2019) demonstrated that wild-type transcripts were derived from both the wild-type and c.8331+2T>C alleles, although they did not specify the relative levels of wild-type transcript emanating from the wild-type and variant alleles. More recently, using exonic tag-SNP analysis of transcripts expressed in blood cells from a *BRCA2* c.8331+2T>C heterozygote, Nix et al. (2021) demonstrated that 62 and 38% of the wild-type transcripts were derived from the wild-type and variant alleles, respectively. For the purpose of comparison, *BRCA2* c.68-7T>A, which causes an ∼20% functional loss of the variant allele, has been firmly established to be nonpathogenic (Colombo et al., 2018) whilst analysis of a neutral leaky variant (c.231T>G) has served to demonstrate that a reduction of ∼60% of full-length *BRCA2* transcripts from the mutant allele does not give rise to any measurable increase in cancer risk (Tubeuf et al., 2020).

Significant discordance was also apparent between the two minigene systems, into which identical inserts were cloned for each variant under study, in terms of the results obtained (**Table 1**). Moreover, even in the cases that showed concordance in terms of the generation (or not) of wild-type transcripts, the splicing outcomes may have differed in terms of the nature of the aberrant transcripts and/or relative levels of the wild-type transcripts. Take, for example, the *FOLR3* IVS4+2T>C variant

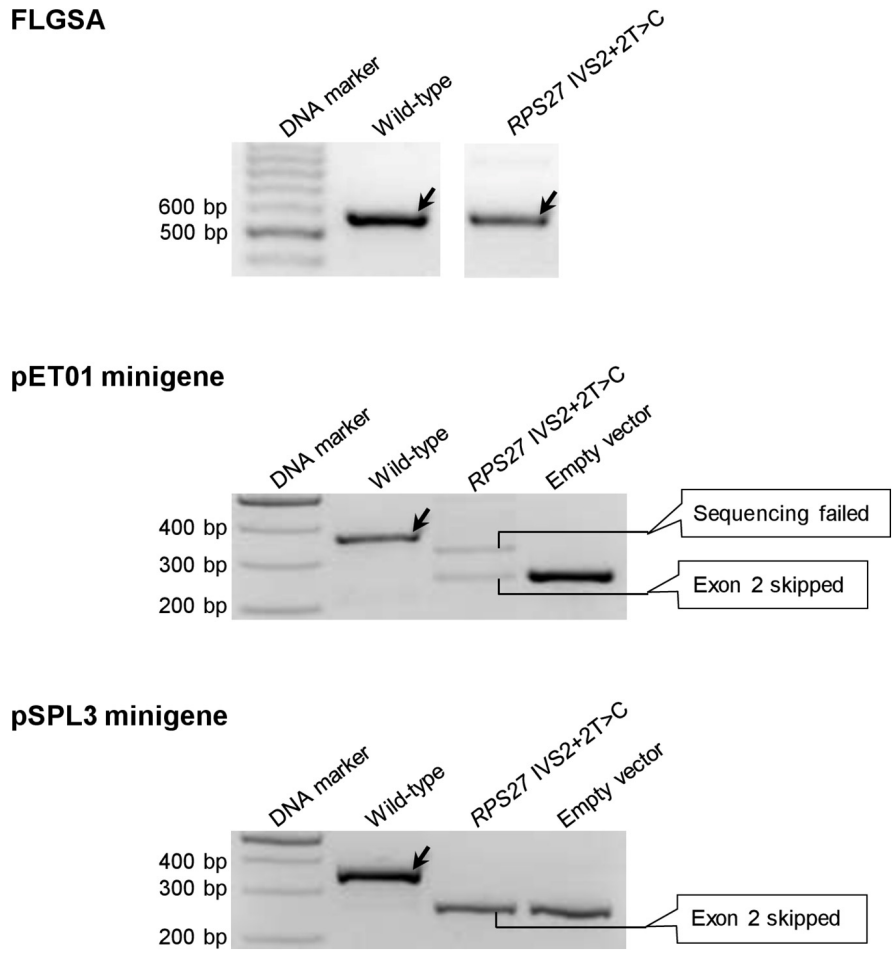


FIGURE 5 | Reverse transcription-polymerase chain reaction results from the pET01 and pSPL3 minigene assays with respect to the *RPS27* IVS2+2T>C variant. Results from the previously performed full-length gene splicing assay (FLGSA) (Lin et al., 2019) are included for the sake of comparison. In all panels, wild-type transcripts are indicated by oblique downward pointing arrows. See **Supplementary Figure 17** for the full gel photographs with respect to the minigene assays. The FLGSA data were adapted from Lin et al. (2019) with permission (Copyright 2020 Wiley Periodicals LLC).

that generated wild-type transcripts in all three systems: the aberrant transcripts generated from FLGSA were different from those generated from the two minigene assays; moreover, the relative levels of the wild-type transcript were markedly different between the two minigene assays. Specifically, the FLGSA-derived aberrant transcript had retained intron 4 whereas the minigene-derived aberrant transcript had skipped exon 4; further, the level of the pET01-derived wild-type transcripts was much higher than that of the pSPL3-derived wild-type transcripts, as indicated by the relative intensities of the wild-type and aberrant transcript bands (**Figure 6**).

All the above mentioned discordant findings could be attributed primarily (if not solely) to differences in the underlying sequence contexts because our previous FLGSA and the current minigene assays were all performed under the same experimental conditions and employing the same procedures. As such, the high level of discordant findings between the different systems used should not be regarded as surprising given that (i) the sequence determinants for the 5′ splice site go beyond the best studied 9-bp consensus sequence motif (see Lin et al., 2019 and references therein) and (ii) splicing is a complicated as well as a coordinated process across different introns (Fu and Ares, 2014; Drexler et al., 2020). In this context, it is pertinent to cite a previous study, in which two splicing reporter minigenes were found to exhibit very different sensitivities in relation to the effects of 13 *MLH1* variants on exon 10 skipping; it was the one that most closely approximated the pattern of exon 10 skipping *in vivo* (in the context of the wild-type *MLH1* exon 10 minigene construct) that was used for the final analysis (Soukarieh et al., 2016). Taken together with our current findings, this indicates that it is most unlikely that a universal splicing reporter minigene could ever be developed that would be suitable for the analysis of all splicing variants. In other words, different exon trapping vectors carrying a particular wild-type target gene insert might need to be tested in advance with a view to selecting one empirically that most closely resembled the normal expression pattern of the gene designated for functional analysis. Alternatively, a midigene splicing assay (Sangermano et al., 2018) might be considered

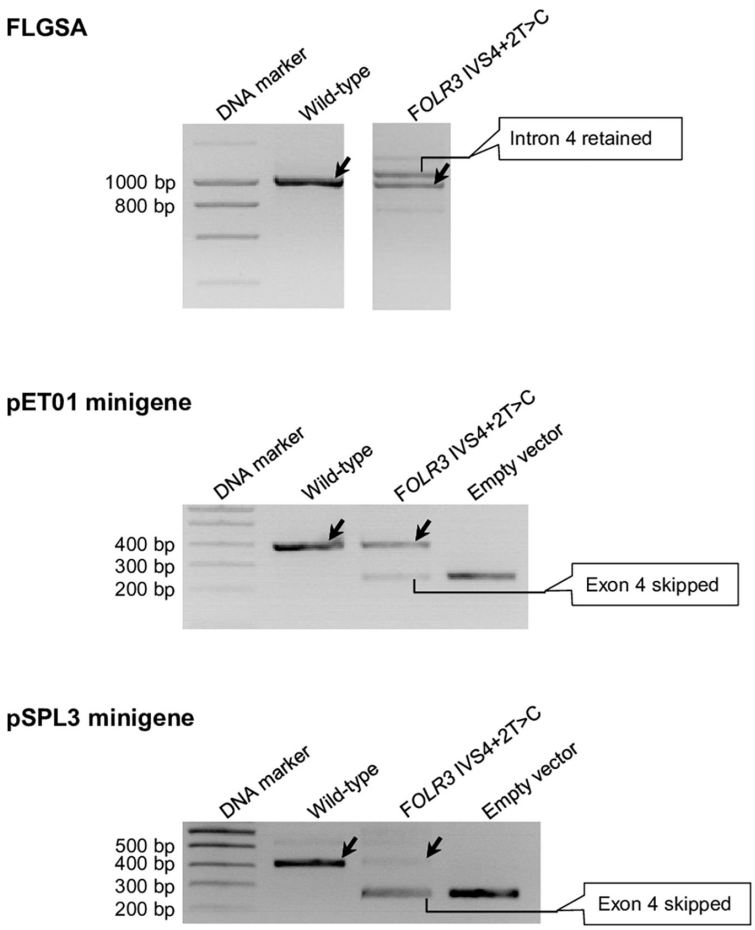


FIGURE 6 | Reverse transcription-polymerase chain reaction results from the pET01 and pSPL3 minigene assays with respect to the *FOLR3* IVS4+2T>C variant. Results from the previously performed full-length gene splicing assay (FLGSA) (Lin et al., 2019) are included for the sake of comparison (NB. the aberrant transcript with retained intron 4 was newly sequenced in this study). In all panels, wild-type transcripts are indicated by oblique downward pointing arrows. See **Supplementary Figure 6** for the full gel photographs with respect to the minigene assays. The FLGSA data were adapted from Lin et al. (2019) with permission (Copyright 2020 Wiley Periodicals LLC).

with a view to increasing the natural sequence context of the variant under study.

Although we have provided experimental evidence that genomic sequence context has influenced the splicing outcome of +2T>C variants capable of generating wild-type transcripts, it is beyond our current ability to discern precisely how and why these differences originated. For illustrative purposes, let us take the two pSPL3 wild-type minigene constructs that did not generate the expected wild-type transcripts. First, the pSPL3 *DNAJC19* IVS5+2T minigene construct expressed a transcript lacking *DNAJC19* exon 5 and containing instead a 118-bp pseudoexon. As shown in **Figure 3**, this was due to the inactivation of the physiological GT-AG splice sites defining *DNAJC19* exon 5 and the concurrent activation of cryptic splice sites located within the chimeric intron 2. However, based upon comparisons with the 3′ splice site consensus sequence (CAG|G) and 5′ splice site consensus sequence (MAG|GTRAGT where M is A or C and R is A or G)[1], we could not draw any meaningful conclusions about the alternative use of the *DNAJC19* exon 5-defining GT-AG splice sites and the aberrantly activated cryptic GT-AG splice sites observed in the minigene construct. Second, the pSPL3 *RPS27* IVS3+2T minigene construct expressed an *RPS27* transcript skipping exon 3 (**Supplementary Figure 18**). One might argue that this could somehow be associated with alternative splicing. However, although *RPS27* has three alternative transcripts,[2] exon 3 (in the context of NM_001030.6) is common to all three. Even if *RPS27* exon 3 was differentially used by the three transcripts, the fact of its being skipped only in the *pSPL3* context points to differences in sequence that extend beyond the gene inserts.

There is one final point to make. Although we have provided experimental evidence that points to limitations in the minigene-based analysis of splicing, our findings should not be interpreted as a challenge to the current preeminence of the minigene splicing

[1]https://science.umd.edu/labs/mount/RNAinfo/consensus.html

[2]https://www.ncbi.nlm.nih.gov/gene/6232

assay which is one of the most widely used analytical tools employed for the interpretation of potentially pathogenic variants. Indeed, all methods for assessing splicing have their advantages and inconveniences.

Our study has its limitations. For example, we used only one cell line for transfection and subsequent RT-PCR analysis. It would be interesting to see whether the same results were obtained employing another cell line. Here it may nonetheless be pertinent to mention that in our previous study, we have analyzed 10 +2T>C substitutions that generated wild-type transcripts and 10 +2T>C substitutions that did not generate wild-type transcripts in HEK293T cells for FLGSA in HeLa cells; we observed entirely consistent findings in the two cell lines in terms of the generation of wild-type transcripts or not (Lin et al., 2019). Moreover, in common with our previous studies (Lin et al., 2019, 2020), our findings were based on qualitative RT-PCR/gel analysis in terms of the absence or presence of the wild-type transcripts. Repeating the experiments using another more precise method would strengthen our findings. However, we believe that since an aberrant transcript band was always generated by the variant minigene expression vectors (or in other words, the aberrant transcripts always served as an internal control for gene expression), our findings should be highly reliable.

CONCLUSION

Our study provides experimental evidence that +2T>C variants capable of generating some wild-type transcripts exhibit remarkable differences not only between minigene and full-length gene splicing assays but also between different minigene assays. Our results therefore bring fresh glimpses of the limitations that are inherent to minigene splicing assays and emphasize the role of sequence context in regulating splicing of a particular variant type. Whether our findings also apply to other types of splice-altering variant remains to be investigated.

DATA AVAILABILITY STATEMENT

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

AUTHOR CONTRIBUTIONS

J-HL, HW, and W-BZ designed the study, performed the experiments, and assisted in writing the manuscript. EM, YF, GLG, DNC, and CF analyzed the data and revised the manuscript with important intellectual input. ZL contributed to study design, obtained funding, supervised the experiments, and revised the manuscript with important intellectual input. J-MC conceived and coordinated the study and drafted the manuscript. All authors approved the final manuscript.

FUNDING

This research was funded by the National Natural Science Foundation of China (82000611 to J-HL and 81800569 to HW), the Shanghai Pujiang Program (2020PJD061 to J-HL and 18PJD057 to HW), the Scientific Innovation Program of Shanghai Municipal Education Committee, China (201901070007E00052 to ZL), and the Institut National de la Santé et de la Recherche Médicale (INSERM), France. The funding bodies did not play any role in the study design, collection, analysis and interpretation of data or in the writing of the article and the decision to submit it for publication.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.701652/full#supplementary-material

REFERENCES

Abril, J. F., Castelo, R., and Guigo, R. (2005). Comparison of splice sites in mammals and chicken. *Genome Res.* 15, 111–119. doi: 10.1101/gr.3108805

Aebi, M., Hornig, H., and Weissmann, C. (1987). 5′ cleavage site in eukaryotic pre-mRNA splicing is determined by the overall 5′ splice region, not by the conserved 5′ GU. *Cell* 50, 237–246. doi: 10.1016/0092-8674(87)90219-4

Aebi, M., Hornig, H., Padgett, R. A., Reiser, J., and Weissmann, C. (1986). Sequence requirements for splicing of higher eukaryotic nuclear pre-mRNA. *Cell* 47, 555–565. doi: 10.1016/0092-8674(86)90620-3

Aicher, J. K., Jewell, P., Vaquero-Garcia, J., Barash, Y., and Bhoj, E. J. (2020). Mapping RNA splicing variations in clinically accessible and nonaccessible tissues to facilitate Mendelian disease diagnosis using RNA-seq. *Genet. Med.* 22, 1181–1190. doi: 10.1038/s41436-020-0780-y

Anna, A., and Monika, G. (2018). Splicing mutations in human genetic disorders: examples, detection, and confirmation. *J. Appl. Genet.* 59, 253–268. doi: 10.1007/s13353-018-0444-7

Beer, S., and Sahin-Toth, M. (2014). Exonic variants affecting pre-mRNA splicing add to genetic burden in chronic pancreatitis. *Gut* 63, 860–861. doi: 10.1136/gutjnl-2013-305981

Burset, M., Seledtsov, I. A., and Solovyev, V. V. (2000). Analysis of canonical and non-canonical splice sites in mammalian genomes. *Nucleic Acids Res.* 28, 4364–4375. doi: 10.1093/nar/28.21.4364

Burset, M., Seledtsov, I. A., and Solovyev, V. V. (2001). SpliceDB: database of canonical and non-canonical mammalian splice sites. *Nucleic Acids Res.* 29, 255–259. doi: 10.1093/nar/29.1.255

Chen, J. M., Lin, J. H., Masson, E., Liao, Z., Férec, C., Cooper, D. N., et al. (2020). The experimentally obtained functional impact assessments of 5′ splice site GT>GC variants differ markedly from those predicted. *Curr. Genomics* 21, 56–66. doi: 10.2174/1389202921666200210141701

Chen, Y., Huang, L., Jiao, X., Riazuddin, S., Riazuddin, S. A., and Fielding Hetmancik, J. (2018). A novel *LRAT* mutation affecting splicing in a family with early onset retinitis pigmentosa. *Hum. Genomics* 12:35.

Colombo, M., Lopez-Perolio, I., Meeks, H. D., Caleca, L., Parsons, M. T., Li, H., et al. (2018). The *BRCA2* c.68-7T >A variant is not pathogenic: a model for clinical calibration of spliceogenicity. *Hum. Mutat.* 39, 729–741.

Cooper, T. A., Wan, L., and Dreyfuss, G. (2009). RNA and disease. *Cell* 136, 777–793.

Damasio, J., Santos, D., Sara Morais, S., Bras, J., Guerreiro, R., Sardoeira, A., et al. (2021). Congenital ataxia due to novel variant in *ATP8A2*. *Clin. Genet.* 100, 79–83. doi: 10.1111/cge.13954

den Dunnen, J. T., Dalgleish, R., Maglott, D. R., Hart, R. K., Greenblatt, M. S., McGowan-Jordan, J., et al. (2016). HGVS recommendations for the description of sequence variants: 2016 update. *Hum. Mutat.* 37, 564–569. doi: 10.1002/humu.22981

Den Uijl, I. E., Mauser Bunschoten, E. P., Roosendaal, G., Schutgens, R. E., Biesma, D. H., Grobbee, D. E., et al. (2011). Clinical severity of haemophilia A: does the classification of the 1950s still stand? *Haemophilia* 17, 849–853. doi: 10.1111/j.1365-2516.2011.02539.x

Drexler, H. L., Choquet, K., and Churchman, L. S. (2020). Splicing kinetics and coordination revealed by direct nascent RNA sequencing through nanopores. *Mol. Cell* 77, 985–998. doi: 10.1016/j.molcel.2019.11.017

Findlay, G. M., Daza, R. M., Martin, B., Zhang, M. D., Leith, A. P., Gasperini, M., et al. (2018). Accurate classification of *BRCA1* variants with saturation genome editing. *Nature* 562, 217–222. doi: 10.1038/s41586-018-0461-z

Fraile-Bethencourt, E., Diez-Gomez, B., Velasquez-Zapata, V., Acedo, A., Sanz, D. J., and Velasco, E. A. (2017). Functional classification of DNA variants by hybrid minigenes: identification of 30 spliceogenic variants of *BRCA2* exons 17 and 18. *PLoS Genet.* 13:e1006691. doi: 10.1371/journal.pgen.1006691

Frischknecht, H., Dutly, F., Walker, L., Nakamura-Garrett, L. M., Eng, B., and Waye, J. S. (2009). Three new beta-thalassemia mutations with varying degrees of severity. *Hemoglobin* 33, 220–225. doi: 10.1080/03630260903089060

Fu, X. D., and Ares, M. Jr. (2014). Context-dependent control of alternative splicing by RNA-binding proteins. *Nat. Rev. Genet.* 15, 689–701. doi: 10.1038/nrg3778

Gelli, E., Colombo, M., Pinto, A. M., De Vecchi, G., Foglia, C., Amitrano, S., et al. (2019). Usefulness and limitations of comprehensive characterization of mRNA splicing profiles in the definition of the clinical relevance of *BRCA1/2* variants of uncertain significance. *Cancers (Basel)* 11:295. doi: 10.3390/cancers11030295

Goldberg, Y., Laitman, Y., Ben David, M., Bazak, L., Lidzbarsky, G., Salmon, L. B., et al. (2021). Re-evaluating the pathogenicity of the c.783+2T>C *BAP1* germline variant. *Hum. Mutat.* 42, 592–599. doi: 10.1002/humu.24189

Hao, Z., Jin, D. Y., Chen, X., Schurgers, L. J., Stafford, D. W., and Tie, J. K. (2021). Gamma-Glutamyl carboxylase mutations differentially affect the biological function of vitamin K-dependent proteins. *Blood* 137, 533–543. doi: 10.1182/blood.2020006329

Jaganathan, K., Kyriazopoulou Panagiotopoulou, S., McRae, J. F., Darbandi, S. F., Knowles, D., Li, Y. I., et al. (2019). Predicting splicing from primary sequence with deep learning. *Cell* 176, 535–548.e24.

Joynt, A. T., Evans, T. A., Pellicore, M. J., Davis-Marcisak, E. F., Aksit, M. A., Eastman, A. C., et al. (2020). Evaluation of both exonic and intronic variants for effects on RNA splicing allows for accurate assessment of the effectiveness of precision therapies. *PLoS Genet.* 16:e1009100. doi: 10.1371/journal.pgen.1009100

Kim, B. J., Jeon, H. W., Jeon, W., Han, J. H., Oh, J., Yi, N., et al. (2021). Rising of *LOXHD1* as a signature causative gene of down-sloping hearing loss in people in their teens and 20s. *J. Med. Genet.* doi: 10.1136/jmedgenet-2020-107594 [Epub ahead of print],

Kortum, F., Kieninger, S., Mazzola, P., Kohl, S., Wissinger, B., Prokisch, H., et al. (2021). X-linked retinitis pigmentosa caused by non-canonical splice site variants in *RPGR*. *Int. J. Mol. Sci.* 22:850. doi: 10.3390/ijms22020850

Kume, K., Masamune, A., Kikuta, K., and Shimosegawa, T. (2006). [-215G>A; IVS3+2T>C] mutation in the *SPINK1* gene causes exon 3 skipping and loss of the trypsin binding site. *Gut* 55:1214. doi: 10.1136/gut.2006.095752

Le Tertre, M., Ka, C., Raud, L., Berlivet, I., Gourlaouen, I., Richard, G., et al. (2021). Splicing analysis of *SLC40A1* missense variations and contribution to hemochromatosis type 4 phenotypes. *Blood Cells Mol. Dis.* 87:102527. doi: 10.1016/j.bcmd.2020.102527

Lin, J. H., Masson, E., Boulling, A., Hayden, M., Cooper, D. N., Férec, C., et al. (2020). 5' splice site GC>GT and GT>GC variants differ markedly in terms of their functionality and pathogenicity. *Hum. Mutat.* 41, 1358–1364. doi: 10.1002/humu.24029

Lin, J. H., Tang, X. Y., Boulling, A., Zou, W. B., Masson, E., Fichou, Y., et al. (2019). First estimate of the scale of canonical 5' splice site GT>GC variants capable of generating wild-type transcripts. *Hum. Mutat.* 40, 1856–1873. doi: 10.1002/humu.23821

Morbidoni, V., Baschiera, E., Forzan, M., Fumini, V., Ali, D. S., Giorgi, G., et al. (2021). Hybrid minigene assay: an efficient tool to characterize mRNA splicing profiles of *NF1* variants. *Cancers (Basel)* 13:999. doi: 10.3390/cancers13050999

Mount, S. M., Avsec, Z., Carmel, L., Casadio, R., Celik, M. H., Chen, K., et al. (2019). Assessing predictions of the impact of variants on splicing in CAGI5. *Hum. Mutat.* 40, 1215–1224. doi: 10.1002/humu.23869

Nix, P., Mundt, E., Coffee, B., Goossen, E., Warf, B. M., Brown, K., et al. (2021). Interpretation of *BRCA2* splicing variants: a case series of challenging variant interpretations and the importance of functional RNA analysis. *Fam. Cancer* doi: 10.1007/s10689-020-00224-y [Epub ahead of print],

Parada, G. E., Munita, R., Cerda, C. A., and Gysling, K. (2014). A comprehensive survey of non-canonical splice sites in the human transcriptome. *Nucleic Acids Res.* 42, 10564–10578. doi: 10.1093/nar/gku744

Qian, X., Wang, J., Wang, M., Igelman, A. D., Jones, K. D., Li, Y., et al. (2021). Identification of deep-intronic splice mutations in a large cohort of patients with inherited retinal diseases. *Front. Genet.* 12:647400. doi: 10.3389/fgene.2021.647400

Ramalho, A. S., Beck, S., Meyer, M., Penque, D., Cutting, G. R., and Amaral, M. D. (2002). Five percent of normal cystic fibrosis transmembrane conductance regulator mRNA ameliorates the severity of pulmonary disease in cystic fibrosis. *Am. J. Respir Cell Mol. Biol.* 27, 619–627. doi: 10.1165/rcmb.2001-0004oc

Raraigh, K. S., Han, S. T., Davis, E., Evans, T. A., Pellicore, M. J., McCague, A. F., et al. (2018). Functional assays are essential for interpretation of missense variants associated with variable expressivity. *Am. J. Hum. Genet.* 102, 1062–1077. doi: 10.1016/j.ajhg.2018.04.003

Raud, L., Ka, C., Gourlaouen, I., Callebaut, I., Ferec, C., Le Gac, G., et al. (2019). Functional analysis of novel *RHD* variants: splicing disruption is likely to be a common mechanism of variant D phenotype. *Transfusion* 59, 1367–1375. doi: 10.1111/trf.15210

Saint-Martin, C., Cauchois-Le Miere, M., Rex, E., Soukarieh, O., Arnoux, J. B., Buratti, J., et al. (2021). Functional characterization of *ABCC8* variants of unknown significance based on bioinformatics predictions, splicing assays, and protein analyses: benefits for the accurate diagnosis of congenital hyperinsulinism. *Hum. Mutat.* 42, 408–420. doi: 10.1002/humu.24164

Sangermano, R., Khan, M., Cornelis, S. S., Richelle, V., Albert, S., Garanto, A., et al. (2018). *ABCA4* midigenes reveal the full splice spectrum of all reported noncanonical splice site variants in stargardt disease. *Genome Res* 28, 100–110. doi: 10.1101/gr.226621.117

Scalet, D., Maestri, I., Branchini, A., Bernardi, F., Pinotti, M., and Balestra, D. (2019). Disease-causing variants of the conserved +2T of 5′ splice sites can be rescued by engineered U1snRNAs. *Hum. Mutat.* 40, 48–52. doi: 10.1002/humu.23680

Schmaltz-Panneau, B., Pagnier, A., Clauin, S., Buratti, J., Marty, C., Fenneteau, O., et al. (2021). Identification of biallelic germline variants of *SRP68* in a sporadic case with severe congenital neutropenia. *Haematologica* 106, 1216–1219. doi: 10.3324/haematol.2020.247825

Scotti, M. M., and Swanson, M. S. (2016). RNA mis-splicing in disease. *Nat. Rev. Genet.* 17, 19–32. doi: 10.1038/nrg.2015.3

Sheth, N., Roca, X., Hastings, M. L., Roeder, T., Krainer, A. R., and Sachidanandam, R. (2006). Comprehensive splice-site analysis using comparative genomics. *Nucleic Acids Res.* 34, 3955–3967. doi: 10.1093/nar/gkl556

Soukarieh, O., Gaildrat, P., Hamieh, M., Drouet, A., Baert-Desurmont, S., Frebourg, T., et al. (2016). Exonic splicing mutations are more prevalent than currently estimated and can be predicted by using in silico tools. *PLoS Genet.* 12:e1005756. doi: 10.1371/journal.pgen.1005756

Stenson, P. D., Mort, M., Ball, E. V., Chapman, M., Evans, K., Azevedo, L., et al. (2020). The human gene mutation database (HGMD®): optimizing its use in a clinical diagnostic or research setting. *Hum. Genet.* 139, 1197–1207. doi: 10.1007/s00439-020-02199-3

Tang, X. Y., Lin, J. H., Zou, W. B., Masson, E., Boulling, A., Deng, S. J., et al. (2019). Toward a clinical diagnostic pipeline for *SPINK1* intronic variants. *Hum. Genomics* 13:8.

Torrado, M., Fernandez, G., Ganoza, C. A., Maneiro, E., Garcia, D., Sonicheva-Paterson, N., et al. (2021). A cryptic splice-altering *KCNQ1* variant in trans with R259L leading to Jervell and Lange-Nielsen syndrome. *NPJ Genom. Med.* 6:21.

Truty, R., Ouyang, K., Rojahn, S., Garcia, S., Colavin, A., Hamlington, B., et al. (2021). Spectrum of splicing variants in disease genes and the ability of RNA

analysis to reduce uncertainty in clinical interpretation. *Am. J. Hum. Genet.* 108, 696–708. doi: 10.1016/j.ajhg.2021.03.006

Tubeuf, H., Caputo, S. M., Sullivan, T., Rondeaux, J., Krieger, S., Caux-Moncoutier, V., et al. (2020). Calibration of pathogenicity due to variant-induced leaky splicing defects by using *BRCA2* exon 3 as a model system. *Cancer Res.* 80, 3593–3605. doi: 10.1158/0008-5472.can-20-0895

Wai, H. A., Lord, J., Lyon, M., Gunning, A., Kelly, H., Cibin, P., et al. (2020). Blood RNA analysis can increase clinical diagnostic rate and resolve variants of uncertain significance. *Genet. Med.* 22, 1005–1014. doi: 10.1038/s41436-020-0766-9

Wu, H., Boulling, A., Cooper, D. N., Li, Z. S., Liao, Z., Chen, J. M., et al. (2017). *In vitro* and *in silico* evidence against a significant effect of the *SPINK1* c.194G>A variant on pre-mRNA splicing. *Gut* 66, 2195–2196. doi: 10.1136/gutjnl-2017-313948

Zou, W. B., Boulling, A., Masson, E., Cooper, D. N., Liao, Z., Li, Z. S., et al. (2016). Clarifying the clinical relevance of *SPINK1* intronic variants in chronic pancreatitis. *Gut* 65, 884–886. doi: 10.1136/gutjnl-2015-311168

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: 19 November 2021
doi: 10.3389/fgene.2021.780874

TMPRSS3 Gene Variants With Implications for Auditory Treatment and Counseling

In Seok Moon[1,2†], *Andrew R. Grant*[3,4†], *Varun Sagi*[5,6], *Heidi L. Rehm*[3,7] *and Konstantina M. Stankovic*[1,5*]

[1]Department of Otolaryngology—Head and Neck Surgery, Massachusetts Eye and Ear and Harvard Medical School, Boston, MA, United States , [2]Department of Otorhinolaryngology, Yonsei University College of Medicine, Seoul, Korea, [3]Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA, United States, [4]New York Medical College, Valhalla, NY, United States, [5]Department of Otolaryngology—Head and Neck Surgery, Stanford University School of Medicine, Stanford, CA, United States, [6]University of Minnesota Medical School, Minneapolis, MN, United States, [7]Center for Genomic Medicine and Departments of Pathology and Medicine, Massachusetts General Hospital and Harvard Medical School, Boston, MA, United States

OPEN ACCESS

Edited by:
Gavin R. Oliver,
Mayo Clinic, United States

Reviewed by:
Katarina Trebušak Podkrajšek,
University of Ljubljana, Slovenia
Sedigheh Delmaghani,
Institut Pasteur, France
Jourdan Holder,
Vanderbilt University Medical Center,
United States

***Correspondence:**
Konstantina M. Stankovic
kstankovic@stanford.edu

[†]These authors have contributed equally to this work

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

Received: 21 September 2021
Accepted: 18 October 2021
Published: 19 November 2021

Citation:
Moon IS, Grant AR, Sagi V, Rehm HL and Stankovic KM (2021) TMPRSS3 Gene Variants With Implications for Auditory Treatment and Counseling. Front. Genet. 12:780874. doi: 10.3389/fgene.2021.780874

Objective: To identify and report novel variants in the *TMPRSS3* gene and their clinical manifestations related to hearing loss as well as intervention outcomes. This information will be helpful for genetic counseling and treatment planning for these patients.

Methods: Literature review of previously reported *TMPRSS3* variants was conducted. Reported variants and associated clinical information was compiled. Additionally, cohort data from 18 patients, and their families, with a positive result for *TMPRSS3*-associated hearing loss were analyzed. Genetic testing included sequencing and copy number variation (CNV) analysis of *TMPRSS3* and the Laboratory for Molecular Medicine's OtoGenome-v1, -v2, or -v3 panels. Clinical data regarding patient hearing rehabilitation was interpreted along with their genetic testing results and in the context of previously reported cochlear implant outcomes in individuals with *TMPRSS3* variants.

Results: There have been 87 previously reported *TMPRSS3* variants associated with non-syndromic hearing loss in more than 20 ancestral groups worldwide. Here we report occurrences of known variants as well as one novel variant: deletion of Exons 1–5 and 13 identified from our cohort of 18 patients. The hearing impairment in many of these families was consistent with that of previously reported patients with *TMPRSS3* variants (i.e., typical down-sloping audiogram). Four patients from our cohort underwent cochlear implantation.

Conclusion: Bi-allelic variants of *TMPRSS3* are associated with down-sloping hearing loss regardless of ancestry. The outcome following cochlear implantation in patients with variants of *TMPRSS3* is excellent. Therefore, cochlear implantation is strongly recommended for hearing rehabilitation in these patients.

Keywords: TMPRSS3, cochlear implantation, sensorineural hearing loss, genetic counseling, hereditary hearing loss

1 INTRODUCTION

Autosomal recessive non-syndromic hearing loss (ARNSHL) is the most common form of hereditary hearing loss. It accounts for about 70–80% of congenital hereditary hearing loss. ARNSHL is an extremely heterogenous condition as more than 98 loci have been mapped and 77 causative genes have been identified to date (http://hereditaryhearingloss.org/).

The *TMPRSS3* gene encodes a type III transmembrane serine protease that is structurally defined by four functional domains: a transmembrane domain, low density lipoprotein receptor A domain, scavenger receptor cysteine rich domain, and a carboxyl terminal serine protease domain (Südhof et al., 1985; van Driel et al., 1987; Sarrias et al., 2004; Rawlings et al., 2010). The *TMPRSS3* gene is expressed in inner hair cells, spiral ganglion neurons (SGNs), the stria vascularis, and cochlear aqueducts of fetal cochlea (Guipponi et al., 2002). Four alternatively spliced transcripts have been described (DiStefano et al., 2018). The transmembrane serine protease 3 protein is thought to be involved in the development and maintenance of the inner ear, perilymph, endolymph and SGNs (Guipponi et al., 2002). While the function of the *TMPRSS3* gene in the auditory system is not fully understood, its alteration has been linked with non-syndromic genetic hearing loss (DiStefano et al., 2018).

The incidence of *TMPRSS3*-associated ARNSHL is variable amongst different ancestral backgrounds but *TMPRSS3* is a significant contributor in some populations. Pathogenic *TMPRSS3* variants account for 0.7% of Japanese (Miyagawa et al., 2015), 3% of Pakistani (Ben-Yosef et al., 2001), 4.6% of Chinese (Gao et al., 2017), 5–6% of Tunisian (Masmoudi et al., 2001), 5.9% of Korean (Chung et al., 2014), and 11% of Turkish (Wattenhofer et al., 2005) ARNSHL cases. However, this gene has been reported in less than 1% of non-syndromic genetic deafness in White individuals (Wattenhofer et al., 2002). In contrast, pathogenic variants in the *GJB2* gene are found in up to 50% of patients with ARNSHL. Despite the relatively low proportion of ARNSHL cases attributed to *TMPRSS3*, the gene remains a prime candidate for post lingual progressive ARNHSL in North European populations once *GJB2* variants are ruled out (Seligman et al., 2021).

Patients with pathogenic variants in the *TMPRSS3* gene have been described as having one of two discrete hearing phenotypes: severe, prelingual or progressive, post-lingual hearing loss. Weegerink et al. (2011) proposed that the phenotypic outcome of hearing loss is dependent on the combination and severity of *TMPRSS3* variants (i.e., mild or severe). They assert that having two "severe" pathogenic variants leads to profound deafness with prelingual onset (DFNB10), whereas a single 'severe' pathogenic variant *in trans* with a milder *TMPRSS3* pathogenic variant yields an initially less severe, but progressive and post-lingual onset hearing loss (DFNB8) (Weegerink et al., 2011). The *TMPRSS3* gene encodes for a transmembrane serine protease which is expressed in SGNs (Guipponi et al., 2002). Therefore, the differential hearing phenotype may reflect the extent of loss of protease activity from a given variant.

In this study, we compile previously reported *TMPRSS3* variants and present a novel variant along with their associated hearing phenotypes. We also aggregate reported outcomes and present new findings regarding the therapeutic effects of cochlear implantation (CI) in patients with pathogenic *TMPRSS3* variants. Together, this information may assist with genetic counseling and treatment planning for patients with *TMPRSS3* variants.

2 METHODS

2.1 Review of the Literature

Literature databases were searched using different combinations of keywords such as "transmembrane serine protease 3," "TMPRSS3," "ear," "hearing loss," "non-syndromic hearing loss," and "cochlear implantation." The databases searched were PubMed, Google Scholar, and two selected gene database websites (https://hereditaryhearingloss.org; https://www.ncbi.nlm.nih.gov/clinvar/). The titles and abstracts were screened using following inclusion criteria: 1) written in English, 2) dealing with non-syndromic hearing loss, and 3) reporting human data.

Based on the search strategy, 39 TMPRSS3-associated papers published from May 2000 to Aug 2021 were reviewed and summarized (**Figure 1**; **Table 1**). Among those 39 studies, eleven studies described patients who underwent cochlear implantation (**Table 2**).

Previously reported variants and their associated hearing phenotypes and clinical outcomes following CI, when available, were compiled. Additionally, our own cohort of patients was genetically screened as described below.

2.2 Cohort Description

Our study included genetic and phenotypic data from 18 patients and their family members (when available), who were largely White, though Family A was a consanguineous White Egyptian family, Family B was "mixed," and Families M and I were of Hispanic or Latino ethnicity. Of the patients with characterized hearing loss, the severity ranged from moderate to profound with some individuals experiencing congenital onset and others experiencing a childhood onset or an onset in the second decade of life. Patients were referred to the Laboratory for Molecular Medicine (LMM) at Mass General Brigham Personalized Medicine (Cambridge, MA, United States) from 2009 to 2017. Patients were referred from various clinics and hospitals across the United States. The LMM collected information pertinent to the nature of the hearing loss in the patients (if available) including family history of hearing loss and/or disease, audiological testing, temporal bone CT/MRI results, and CI status. Further information was requested through physicians via the Mass General Brigham Human Research Committee's IRB protocol for the study of the genetics of hearing loss. Patients were selected based on whether they received a positive result for *TMPRSS3*-associated hearing loss with the intent of follow up of the outcome of CI, if received.

2.3 *TMPRSS3* Screening and OtoGenome Next-Generation Sequencing Testing

Patient DNA was extracted from whole blood from patients who were referred to the LMM for hearing-loss genetic testing. Our

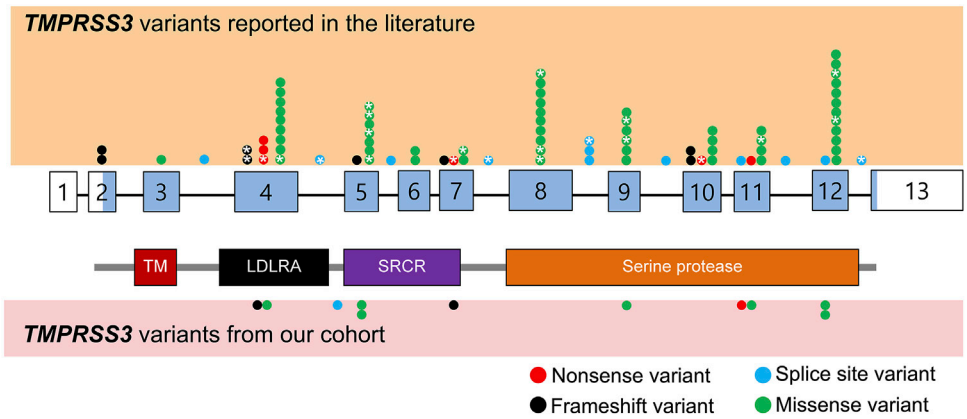


FIGURE 1 | Schematic representation of *TMPRSS3* variants identified from literature review as well as our cohort. Benign variants are not displayed. The exons are numbered with coding sequence shaded blue and untranslated regions unshaded. *: variants with reported cochlear implant outcomes. TM, transmembrane domain; LDLRA, LDL receptor-like domain; SRCR, scavenger receptor cysteine-rich domain.

cohort contains patients from 2009 to 2017. The genetic testing varied for each patient based on the judgment of the ordering physician and the nature of the patient's hearing loss. Testing was performed by single gene sequencing that included *TMPRSS3*, or LMM's OtoGenome-v1,-v2, or -v3 panels.

The LMM's bioinformatics pipeline for targeted next generation sequencing (NGS) panels has been described previously (Pugh et al., 2016). Patients with hearing loss who underwent genetic testing between 2010 and 2014 were tested with the Otogenome-v1 which included the following 71 genes: *ACTG1*, *ATP6V1B1*, *BSND*, *CCDC50*, *CDH23*, *CLDN14*, *CLRN1*, *COCH*, *COL11A2*, *CRYM*, *DFNA5*, *DFNB31*, *DFNB59*, *DIAPH1*, *ESPN*, *ESRRB*, *EYA1*, *EYA4*, *GIPC3*, *GJB2*, *GJB3*, *GJB6*, *GPR98*, *GPSM2*, *GRHL2*, *GRXCR1*, *HGF*, *ILDR1*, *KCNE1*, *KCNQ1*, *KCNQ4*, *LHFPL5*, *LOXHD1*, *LRTOMT*, *MARVELD2*, *MIR183*, *MIR96*, *MSRB3*, *MTRNR1 (12S rRNA)*, *MTTS1 (tRNAser(UCN))*, *MYH14*, *MYH9*, *MYO15A*, *MYO1A*, *MYO3A*, *MYO6*, *MYO7A*, *OTOA*, *OTOF*, *PCDH15*, *PDZD7*, *POU3F4*, *POU4F3*, *PRPS1*, *RDX*, *SERPINB6*, *SLC17A8*, *SLC26A4 (PDS)*, *SLC26A5*, *TECTA*, *TIMM8A*, *TJP2*, *TMC1*, *TMIE*, *TMPRSS3*, *TPRN*, *TRIOBP*, *USH1C*, *USH1G*, *USH2A*, and *WFS1*.

OtoGenome-v2 was used in patients who underwent testing at the LMM from 2014 to 2015. For this iteration, *PDZD7* and *SLC26A5* genes were removed and the *STRC* gene was added. In addition, copy number variant (CNV) detection was added using VisCap as previously described (Pugh et al., 2016; Tayoun et al., 2016).

OtoGenome-v3, used from 2015 to 2017, included 87 genes but did not include the following genes included in v2: *CRYM*, *GJB3*, *MIR182*, *MYO1A*, *SLC17A8*, and *TJP2*. The following 23 genes were added *CACNA1D*, *CATSPER2*, *CEACAM16*, *CIB2*, *CLPP*, *DIABLO*, *EDN3*, *EDNRB*, *HARS2*, *HSD17B4*, *KARS*, *LARS2*, *MITF*, *OTOG*, *OTOGL*, *P2RX2*, *PAX3*, *SIX1*, *SMPX*, *SOX10*, *SYNE4*, *TBC1D24*, and *TSPEAR*. Parents and other unaffected/affected family members, when available, were tested for detected variants. Variants were confirmed *via* Sanger sequencing for single-nucleotide variants (SNVs), or droplet digital PCR for CNVs called by VisCap (Pugh et al., 2016; Tayoun et al., 2016).

2.4 LMM Variant Classification

The LMM's early variant classification methods are as previously described (Duzkale et al., 2013) and were subsequently updated to conform to more recent professional guidelines (Richards et al., 2015). Data used to classify variants included that from population databases (e.g., Exome Aggregation Consortium (ExAC); gnomAD), internal or external disease databases (e.g., ClinVar, LOVD, HGMD), the literature, functional studies, segregation, allelic observations and *in silico* missense and splicing prediction tools. Variants were classified as pathogenic (P), likely pathogenic (LP), of uncertain significance (VUS), likely benign, or benign. The VUS category was further subdivided into VUS-5, -4, and -3 where VUS-5 indicated leaning towards pathogenic, and VUS-3 indicated leaning towards benign. Likely benign and benign variants are not reported in this article but were submitted to ClinVar (www.ncbi.nlm.nih.gov/clinvar/) along with all other variants observed at the LMM.

3 RESULTS

We reviewed the type, position, origin, and variant classification of 87 previously reported *TMPRSS3* variants and present one novel variant identified from our cohort (**Figure 1**; **Table 1**). Compiled variants are associated with non-syndromic hearing loss in more than 20 ancestral groups worldwide. Fourteen of the identified variants were predicted loss-of-function (pLOF) (frameshift, stop-codon, or splice-site variants) with either prematurely terminated protein products or nonsense-mediated

TABLE 1 | Overview of TMPRSS3 variants resulting in non-syndromic hearing loss, including those identified in the present study.

DNA change	Protein change	Exon	Domain	Variant classification	Origin	Phenotype severity at testing	References
Deletion of E1-5 and 13	—	**E1-5 and E13**	—	Pathogenic	**United States**	**Severe**	**This study**
c.36delC	p.Pro12fs	E2			Chinese	Severe	Gao et al. (2017)
c.36dupC	p.Phe13fs	E2			Turkish	—	Diaz-Horta et al. (2012)
c.157G>A	p.Val53Ile	E3	TM		Palestinian	—	Scott et al. (2001)
					Pakistani	—	Ben-Yosef et al. (2001)
					United States		
					Korean	—	Lee et al. (2013)
					Taiwanese	—	Wong et al. (2020)
c.205+38C>T	—	Intron3	—		Taiwanese	—	Wong et al. (2020)
c.207delC	p.Thr70fs	E4			Dutch	—	Ahmed et al. (2004)
					Newfoundlander	—	
					Dutch	—	Weegerink et al. (2011)
c.208delC	p.Thr70fs*19	E4		Pathogenic	Slovenian	Severe	Battelino et al. (2016)
					Polish	—	Lechowicz et al. (2017)
					United States	—	Shearer et al. (2018)
					Slovenian	—	Likar et al. (2018)
					Czech	—	Safka Brozkova et al. (2020)
					United States	**Severe**	**This study**
c.212T>C	p.Phe71Ser	E4	LDLRA		Korean	—	Lee et al. (2013)
					Japanese	—	Miyagawa et al. (2015)
c.218G>A	p.Cys73Tyr	E4	LDRLA		Polish	—	Lechowicz et al. (2017)
c.226C>T	p.Gln76X	E4			Japanese	—	Miyagawa et al. (2013)
							Miyagawa et al. (2015)
c.238C>T	p.Arg80Cys	E4	LDRLA	Likely pathogenic	**Europe**	**—**	Capalbo et al., (2019)
					United States	**Mild**	**This study**
c.239G>A	p.Arg80His	E4	LDRLA		Taiwanese	—	Wong et al. (2020)
c.268G>A	p.Ala90Thr	E4	LDLRA		UK Caucasian	—	Charif et al. (2012)
					Moroccan		
c.280G>A	p.Gly94Arg	E4	LDLRA		Japanese	—	Miyagawa et al. (2015)
c.296C>A	p.Ser99X	E4			Chinese	Severe	Gu et al. (2015)
c.308A>G	p.Asp103Gly	E4	LDLRA		Greek	—	Wattenhofer et al. (2005)
c.310G>A	p.Glu104Lys	E4	LDLRA		Pakistani	—	Lee et al. (2012)
c.310G>T	p.Glu104X	E4			Pakistani	—	Lee et al. (2012)
c.316C>T	p.Arg106Cys	E4	LDLRA		Japanese	Mild	Miyagawa et al. (2013)
					Chinese		Gao et al. (2017)
c.323-6G>A	—	Intron4	—	Pathogenic	Pakistani	—	Scott et al. (2001)
					Korean	—	Ahmed et al. (2004)
					Dutch	—	Weegerink et al. (2011)
					Chinese	Mild	Gao et al. (2017)
					Pakistani	—	Singh et al. (2020)
					United States	**Severe**	**This study**
c.325C>T	p.Arg109Trp	E5	SRCR	Pathogenic	Pakistani	—	Ben-Yosef et al. (2001)
					Pakistani	—	Ahmed et al. (2004)
					Korean	—	Chung et al. (2014)
					Czech	—	Safka Brozkova et al. (2020)
					United States	**Mild**	**This study**
c.326G>A	p.Arg109Gln	E5	SRCR		Chinese	—	Gu et al. (2015)
					Polish	Mild	Lechowicz et al. (2017)
c.331G>A	p.Gly111Ser	E5	SRCR		United States	—	Ben-Yosef et al. (2001)
c.346G>A	p.Val116Met	E5	SRCR		Indian	—	Ganapathy et al. (2014)
					Korean	—	Kim et al. (2017)
					Czech	—	Safka Brozkova et al. (2020)
c.371C>T	p.Ser124Leu	E5	SRCR		Polish	—	Lechowicz et al. (2017)
c.390C>G	p.His130Arg	E5	SRCR		Japanese	—	Miyagawa et al. (2015)

(Continued on following page)

TABLE 1 | (*Continued*) Overview of TMPRSS3 variants resulting in non-syndromic hearing loss, including those identified in the present study.

DNA change	Protein change	Exon	Domain	Variant classification	Origin	Phenotype severity at testing	References
c.413C>G	p.Ala138Glu	E5	SRCR	Pathogenic	British	Mild	Weegerink et al. (2011)
					Korean		
					United States	—	Eppsteiner et al. (2012)
					Polish	—	Lechowicz et al. (2017)
					United States	—	Shearer et al. (2018)
					Pakistani	—	Singh et al. (2020)
					United States	**Mild**	**This study**
c.432delA	p.Gln144fs	E5			Chinese	—	Sang et al. (2019)
c.447-13A>G	—	Intron 5	—		Pakistani	—	Ben-Yosef et al. (2001)
					United States		
					Taiwanese	—	Wong et al. (2020)
c.453G>A	p.Val151Val	E6	SRCR		Palestinian	—	Scott et al. (2001)
					Pakistani	—	Ben-Yosef et al. (2001)
					United States		
					Korean	—	Lee et al. (2013)
					Taiwanese	—	Wong et al. (2020)
c.551T>C	P.Leu184Ser	E6	SRCR		Chinese	—	Sang et al. (2019)
					Chinese	—	Li et al. (2019)
					Taiwanese	—	Wong et al. (2020)
c.581G>T	p.Cys194Phe	E7	SRCR		Pakistani	—	Ben-Yosef et al. (2001)
						Severe	Ahmed et al. (2004)
c.579dupA	p.Cys194Mfs*17	E7		Pathogenic	Polish	—	Lechowicz et al. (2017)
					United States	**Severe**	**This study**
c.595G>A	p.Val199Met	E7	SRCR		Dutch	Severe	Weegerink et al. (2011)
					Korean		
c.607C>T	p.Gln203X	E7			Japanese	Severe	Miyagawa et al. (2013)
c.617-4_-3dupAT	—	Intron7	—		Japanese	—	Miyagawa et al. (2015)
					Taiwanese	—	Wong et al. (2020)
c.621T>C	p.Cys207Cys	E8	Serine protease		Taiwanese	—	Wong et al. (2020)
c.636C>T	p.Gly212Gly	E8	Serine protease		Korean	—	Lee et al. (2013)
c.646C>T	p.Arg216Cys	E8	Serine protease		German	Mild	Elbracht et al. (2007)
					United States (Caucasian)	—	Eppsteiner et al. (2012)
c.647G>T	p.Arg216Leu	E8	Serine protease		Turkish	Severe	Wattenhofer et al. (2005)
					Japanese	—	Miyagawa et al. (2015)
c.726C>G	p.Cys242Trp	E8	Serine protease		Pakistani	Severe	Shafique et al. (2014)
					Czech	—	Safka Brozkova et al. (2020)
c.727G>A	p.Gly243Arg	E8	Serine protease		Indian	—	Ganapathy et al. (2014)
					Pakistani	—	Khan et al. (2019)
c.734C>T	p.Ser245Phe	E8	Serine protease		Czech	—	Safka Brozkova et al. (2020)
c.743C>T	p.Thr248Met	E8	Serine protease		Korean	Mild	Chung et al. (2014)
c.753G>C	p.Trp251Cys	E8	Serine protease		Tunisian	Severe	Masmoudi et al. (2001)
c.757A>G	p.Ile253Val	E8	Serine protease		Pakistani	—	Ben-Yosef et al. (2001)
					United States		
					Korean	—	Lee et al. (2003)
					Taiwanese	—	Wong et al. (2020)
c.767C>T	p.Arg256Val	E8	Serine protease		Pakistani	—	Lee et al. (2012)
c.778G>A	p.Ala260Thr	E8	Serine protease		Japanese	—	Miyagawa et al. (2015)
c.782+8insT	—	Intron8	—		Pakistani	Severe	Ahmed et al. (2004)
c.782+2T>A	—	Intron8	—		Polish	—	Lechowicz et al. (2017)
c.783-1G>A	—	Intron8	—		Korean	—	Kim et al. (2017)

(Continued on following page)

TABLE 1 | (*Continued*) Overview of TMPRSS3 variants resulting in non-syndromic hearing loss, including those identified in the present study.

DNA change	Protein change	Exon	Domain	Variant classification	Origin	Phenotype severity at testing	References
c.809T>A	p.Ile270Asn	E9	Serine protease		Chinese	Severe	Gao et al. (2017)
c.830C>T	p.Pro277Leu	E9	Serine protease		Turkish	—	Masmoudi et al. (2001)
c.871G>C	p.Val291Leu	E9	Serine protease		Korean	—	Lee et al. (2013) Kim et al. (2017)
c.916G>A	p.Ala306Thr	E9	Serine protease	Likely pathogenic	German	Severe	Elbracht et al. (2007)
					Dutch	—	Weegerink et al. (2011)
					United States (Caucasian)	—	Eppsteiner et al. (2012)
					Korean	—	Lee et al. (2013)
						—	Chung et al. (2014)
					Tibetan	—	Fan et al. (2014)
					Chinese	—	Gao et al. (2017)
					Korean	—	Song et al. (2020)
					United States	**Mild**	**This study**
c.933C>T	p.Ala311Ala	E9	Serine protease		Taiwanese	—	Wong et al. (2020)
c.941T>C	p.Leu314Pro	E9	Serine protease		Pakistani	—	Zhou et al. (2020)
c.953-5A>G	—	Intron 9	—		Polish	—	Lechowicz et al. (2017)
c.974T>A	p.Leu325Gln	E10	Serine protease		Polish	—	Lechowicz et al. (2017)
c.988delA	p.Glu330fs	E10			Pakistani	Severe	Walsh et al. (2006)
c.999delC	p.Asp334Mfs*24	E10			Polish	—	Lechowicz et al. (2017)
c.1019C>G	p.Thr340Arg	E10	Serine protease		Italian	Severe	Vozzi et al. (2014)
c.1025G>A	p.Gly342Glu	E10	Serine protease		Turkish	—	Duman et al. (2011)
c.1028G>C	p.Trp343Ser	E10	Serine protease		Czech	—	Safka Brozkova et al. (2020)
c.1039G>T	p.Glu347X	E10			Korean	—	Song et al. (2020)
c.1128C>T	p.Tyr376Tyr	E11	Serine protease		United States	—	Ben-Yosef et al. (2001)
c.1151T>G	p.Met384Arg	E11	Serine protease		Chinese	Severe	Gao et al. (2017)
c.1156T>C	p.Cys386Arg	E11	Serine protease		Indian	—	Ganapathy et al. (2014)
c.1159G>A	p.Ala387Thr	E11	Serine protease		Japanese	Mild	Miyagawa et al. (2013)
c.1180_1187del8ins68	—	E11	Serine protease		Palestinian	Severe	Scott et al. (2001)
c.1183G>C	p.Asp395His	E11	Serine protease	Unknown	**United States**	**Severe**	**This study**
c.1192C>T	p.Gln398X	E11		Pathogenic	Turkish	Severe	Wattenhofer et al. (2005)
					United States	**Severe**	**This study**
c.1194+15C>A	—	Intron 11	—		Taiwanese	—	Wong et al. (2020)
c.1204G>A	p.Gly402Arg	E12	Serine protease		Chinese	Severe	Gao et al. (2017)
					Pakistani	—	Noman et al. (2019)
					United States	—	Bowles et al. (2021)
c.1211C>T	p.Pro404Leu	E12	Serine protease		Tunisian	Severe	Masmoudi et al. (2001) Wattenhofer et al. (2005)
					United States	—	Bowles et al. (2021)
c.1219T>C	p.Cys407Arg	E12	Serine protease		Pakistani	Severe	Ben-Yosef et al. (2001) Ahmed et al. (2004) Lee et al. (2012)
						—	Khan et al. (2019)
						—	Zafar et al. (2020)
c.1244T>C	p.Leu415Ser	E12	Serine protease		Chinese	Severe	Gao et al. (2017)

(Continued on following page)

TABLE 1 | (*Continued*) Overview of TMPRSS3 variants resulting in non-syndromic hearing loss, including those identified in the present study.

DNA change	Protein change	Exon	Domain	Variant classification	Origin	Phenotype severity at testing	References
c.1250G>A	p.Gly417Glu	E12	Serine protease		Chinese	Severe	Gao et al. (2017)
c.1253C>T	p.Ala418Val	E12	Serine protease		Taiwanese	—	Wong et al. (2020)
c.1269C>T	p.Ile423Ile	E12	Serine protease		Taiwanese	—	Wong et al. (2020)
c.1273T>C	p.Cys425Arg	E12	Serine protease		Pakistani	—	Lee et al. (2012)
c.1276G>A	p.Ala426Thr	E12	Serine protease	Likely pathogenic	Dutch	Mild	Weegerink et al. (2011)
					Italian		Leone et al. (2017)
					Polish	—	Lechowicz et al. (2017)
					United States	—	Shearer et al. (2018)
						Mild	**This study**
c.1291C>T	p.Pro431Ser	E12	Serine protease		Italian	Severe	Vozzi et al. (2014)
c.1306C>G	p.Arg436Gly	E12	Serine protease	Likely pathogenic	Polish	—	Lechowicz et al. (2017)
					Czech	—	Safka Brozkova et al. (2020)
					United States	**Severe**	**This study**
c.1343T>C	p.Met448Thr	E12	Serine protease	Likely pathogenic	Polish	—	Lechowicz et al. (2017)
					Czech	—	Safka Brozkova et al. (2020)
					United States	**Mild**	**This study**
c.1345-2A>G	—	E12			United States	—	Shearer et al. (2018)

TM, transmembrane domain; LDLRA, LDL receptor-like domain; SRCR, scavenger receptor cysteine-rich domain; serine protease, trypsin-like serine protease domain. Naming of variants and labeling of domains and exons are based on the NM_001256317.3 transcript. Variant classification based on LMM variant classification. Only predicted loss-of-function and coding variants were included in the table. Bolded text refers to variants identified in this study. Of note, the phenotype severity is provided at the time of testing. While some patients may initially have milder phenotypes, the hearing loss can progress and become more severe.

decay of mRNA. Fifty-eight of the identified variants were missense variants. Nearly all variants were predicted to disrupt the proteolytic activity of the protein. Both prelingual and post lingual hearing impairment was reported, with most patients showing a typical ski-slope audiogram configuration. CI outcomes were reported for 32 patients with bi-allelic variants in *TMPRSS3* across 11 different studies (**Table 2**) (Weegerink et al., 2011; Eppsteiner et al., 2012; Miyagawa et al., 2013; Chung et al., 2014; Miyagawa et al., 2015; Battelino et al., 2016; Gao et al., 2017; Kim et al., 2017; Shearer et al., 2018; Song et al., 2020; Holder et al., 2021). While degree of hearing improvement varied between patients, the majority of those who underwent CI had positive outcomes.

Our cohort included 18 patients—7 females and 11 males—with ages ranging from 3 months to 36 years (**Figure 2**). 15 patients were White with the remaining 3 identifying as Hispanic/Latino or mixed. We identified 12 different *TMPRSS3* variants of which 1 has not been previously reported: deletion of Exons 1–5 and 13 (**Table 3**). This novel variant was classified as *pathogenic* as it met the criteria outlined by previous professional guidelines (Richards et al., 2015) with specifications provided by ClinGen (https://clinicalgenome.org/working-groups/sequence-variant-interpretation), specifically the combination of PVS1 (predicted loss of function), PM2 (absence in gnomAD), and PM3 (homozygous observation in an individual with phenotype matching the gene). The most commonly identified variants were p.Thr70fs*19 and p.Ala138Glu. Eight patients had congenital hearing loss, four of whom had biallelic pLOF variants.

Four patients in our cohort underwent CI, and outcome information was available for two patients. The first patient, from family B, was found to have congenital profound hearing loss and was homozygous for p.Thr70fs*19. It is unclear when the patient underwent CI. However, at a follow up at 4 years of age, the patient had functional speech. Clinical records indicated that the patient had ongoing articulation errors and required speech therapy but was able to maintain adequate hearing. The second patient, from family K, was compound heterozygous for p.Glu104Lys and p.Ala306Thr. Clinical records have suggested positive CI outcome for her moderate-profound hearing loss. The remaining two patients who underwent CI were the siblings from family A who both had profound congenital hearing loss and were homozygous for a deletion of Exons 1–5 and 13. Their current hearing status is unknown.

4 DISCUSSION

The genotype-phenotype correlations of *TMPRSS3* variants have not been well characterized. It has been previously shown that the frequency of *TMPRSS3*-induced ARSNHL was low in White

TABLE 2 | Overview of clinical characteristics and genotypes of patients with *TMPRSS3* variants who have received cochlear implantation.

Study (country)	DNA change	Protein change	Exon	Domain	Hearing loss severity	Age at CI (gender)	Age at severe-profound HL	Pre-operative hearing	CI type	CI outcomes
Weegerink et al., 2011 (Netherlands)	c.207delC c.916G>A	p.Thr70fs p.Ala306Thr	E4 E9	 Serine protease	—	4.5 years	—	Sloping HL 40–60–100–110–110 dB (0.25, 0.5, 1, 2, 4 kHz)	Nucleus Freedom (Cochlear)	91% Phoneme (76% WRS)
	c.595G>A c.916G>A	p.Val199Met p.Ala306Thr	E7 E9	SRCR Serine protease	—	6 years	—	Sloping HL 40–50–110–110–110 dB (0.25, 0.5, 1, 2, 4 kHz)	Nucleus Freedom (Cochlear)	80% Phoneme (65% WRS)
	c.413C>G c.916G>A	p.Ala138Glu p.Ala306Thr	E5 E9	SRCR Serine protease	—	29 years	—	Decreasing HL 80–90–100–110–110 dB (0.25, 0.5, 1, 2, 4 kHz); 5% Phoneme	Nucleus CI24M (Cochlear)	—
	c.207delC c.1276G>A	p.Thr70fs p.Ala426Thr	E4 E12	 Serine protease	—	49 years	—	Decreasing HL 70–95–110–110–110 dB (0.25, 0.5, 1, 2, 4 kHz); 20% Phoneme	Nucleus Contour CI24R (Cochlear)	89% Phoneme (75% WRS)
	c.207delC c.1276G>A	p.Thr70fs p.Ala426Thr	E4 E12	 Serine protease	—	45 years	—	Decreasing HL 80–90–100–110–120 dB (0.25, 0.5, 1, 2, 4 kHz); 5% Phoneme	Clarion AB-5100H (Advanced Bionics)	76% Phoneme (60% WRS)
	c.207delC c.1276G>A	p.Thr70fs p.Ala426Thr	E4 E12	 Serine protease	—	46 years	—	Flat 100–100–110–120–120 dB (0.25, 0.5, 1, 2, 4 kHz); 0% Phoneme	Clarion AB-5100H (Advanced Bionics)	82% Phoneme (58% WRS)
	c.207delC c.1276G>A	p.Thr70fs p.Ala426Thr	E4 E12	 Serine protease	—	43 years	—	Flat 100–90–110–120–120 dB (0.25, 0.5, 1, 2, 4 kHz); 0% Phoneme	Clarion AB-5100H (Advanced Bionics)	83% Phoneme (62% WRS)
	c.413C>G c.595G>A	p.Ala138Glu p.Val199Met	E5 E7	SRCR SRCR	—	51 years	—	Decreasing HL 80–90–100–110–120 dB (0.25, 0.5, 1, 2, 4 kHz); 2.5% Phoneme	Nucleus Contour CI24R (Cochlear)	88% Phoneme (68% WRS)
	c.413C>G c.323-6G>A	p.Ala138Glu —	E5 In4	SRCR SRCR	—	30 years	—	Sloping HL 50–90–110–110–110 dB (0.25, 0.5, 1, 2, 4 kHz); 10% Phoneme	Nucleus Freedom (Cochlear)	—
Eppsteiner et al., 2012 (United States)	c.413C>G c.646C>T	p.Ala138Glu p.Arg216Cys	E5 E8	SRCR Serine protease	Mild Mild	45 years (male)	45 years	93 dB (PTA at 0.5, 1, 2, and 4 kHz)	Advanced Bionics CII	Poor performance (Combined CNC & HINT Score: 37)
	c.413C>G c.916G>A	p.Ala138Glu p.Ala306Thr	E5 E9	SRCR Serine protease	Mild Severe	32 years (female)	17 years	98 dB (PTA at 0.5, 1, 2, and 4 kHz)	Advanced Bionics CII	Poor performance (Combined CNC & HINT Score: 23)
Miyagawa et al., 2013 (Japan)	c.607C>T c.1159G>A	p.Gln203X p.Ala387Thr	E7 E11	 Serine protease	Severe Mild	40 years (female)	—	Sloping HL 25–30–65–100–110–110–100 dB (0.125, 0.25, 0.5, 1, 2, 4, 8 kHz)	MED-EL Pulsar FLEXeas	40–35–30–35–40–40–45 dB (0.125, 0.25, 0.5, 1, 2, 4, 8 kHz)

(Continued on following page)

TABLE 2 | (*Continued*) Overview of clinical characteristics and genotypes of patients with *TMPRSS3* variants who have received cochlear implantation.

Study (country)	DNA change	Protein change	Exon	Domain	Hearing loss severity	Age at CI (gender)	Age at severe-profound HL	Pre-operative hearing	CI type	CI outcomes
Chung et al., 2014 (Korea)	c.325C>T c.916G>A	p.Arg109Trp p.Ala306Thr	E5 E9	SRCR Serine protease	— —	12 years (female)	—	Flat (<sloping) 100–110–110–110–110–110 dB (0.25, 0.5, 1, 2, 4, 8 kHz)	—	Mean open set sentence score at 6 months following CI was 88.5%
	c.325C>T c.916G>A	p.Arg109Trp p.Ala306Thr	E5 E9	SRCR Serine protease	— Profound	6 years (male)	—	Decreasing HL 70–80–90–100–110–100 dB (0.25, 0.5, 1, 2, 4, 8 kHz)	—	Mean open set sentence score at 6 months following CI was 88.5%
Miyagawa et al., 2015 (Japan)	c.390C>G c.647G>T	p.His130Arg p.Arg216Leu	E5 E8	SRCR Serine protease	— —	45 years (male)	—	Sloping HL 25–30–65–100–110–110–100 dB (0.125, 0.25, 0.5, 1, 2, 4, 8 kHz); 30% WRS w/HA	MED-EL PULSAR FLEX24	90% discrimination score on Japanese monosyllable test at 24 months
	c.226C>T c.778G>A	p.Gln76X p.Ala260Thr	E4 E8	 Serine protease	— —	39 years (female)	—	Flat (<-Sloping) 70–90–100–100–110–110–100 dB (0.125, 0.25, 0.5, 1, 2, 4, 8 kHz); 24% WRS w/HA	MED-EL PULSAR FLEX24	70% discrimination score on Japanese monosyllable test at 12 months
	c.212T>C c.617-4_-3dupAT	p.Phe71Ser p.Thr205fs	E4 In7	LDLRA —	— —	51 years (female)	—	Sloping HL 30–40–40–40–100–110–100 dB (0.125, 0.25, 0.5, 1, 2, 4, 8 kHz); 40% WRS w/HA	MED-EL PULSAR FLEX24	80% discrimination score on Japanese monosyllable test at 12 months
Battelino et al., 2016 (Slovenia)	c.208delC[a]	p.Thr70fs[a]19	E4		—	11 months (male)	—	80–110 dB (unclear methodology)	—	25 dB (unclear methodology)
	c.208delC[a]	p.Thr70fs[a]19	E4		—	30 months (male)	—	95–110 dB (unclear methodology)	—	45 dB (unclear methodology)
	c.208delC[a]	p.Thr70fs[a]19	E4		—	13 months (male)	—	80–100 dB (unclear methodology)	—	25 dB (unclear methodology)
	c.208delC[a]	p.Thr70fs[a]19	E4		—	11 months (male)	—	70–85 dB (unclear methodology)	—	25 dB (unclear methodology)
Gao et al., 2017 (China)	c.916G>A	p.Ala306Thr	E9	Serine protease	Severe	3 years (female)	—	Decreasing HL	—	Described as “improved”
	c.1250G>A	p.Gly417Glu	E12	Serine protease	Severe			60–80–80–100–100 dB (0.25, 0.5, 1, 2, 4 kHz)		
	c.916G>A	p.Ala306Thr	E9	Serine protease	Severe	14 years (female)	—	Sloping HL	—	Described as “improved”
	c.323-6G>A	—	In4	—	Severe			20–20–60–100–100–100 dB (0.25, 0.5, 1, 2, 4 kHz)		
Kim et al., 2017 (Korea)	c.346G>A c.783-1G>A	p.Val116Met —	E5 In8	SRCR —	— Uncertain	4 years (female)	—	Decreasing HL 90–100–100–1,100–110 dB (0.25, 0.5, 1, 2, 4 kHz)	—	Not described, unofficially good
	c.346G>A c.871G>C	p.Val116Met p.Val291Leu	E5 E9	SRCR Serine protease	Profound Uncertain	10 years (female)	—	Sloping HL 45–90–100–100–110 dB (0.25, 0.5, 1, 2, 4 kHz)	—	Not described, unofficially good

(Continued on following page)

TABLE 2 | (*Continued*) Overview of clinical characteristics and genotypes of patients with *TMPRSS3* variants who have received cochlear implantation.

Study (country)	DNA change	Protein change	Exon	Domain	Hearing loss severity	Age at CI (gender)	Age at severe-profound HL	Pre-operative hearing	CI type	CI outcomes
Shearer et al., 2018 (United States)	c.208delC c.1276G>A	p.Thr70fs[a]19 p.Ala426Thr	E4 E12	 Serine protease	—	64 years	—	—	Nucelus Hybrid CI L24 Array	80–90–110–110–110 dB (0.125, 0.25, 0.5, 1, 2 kHz)
	c.413C>G c.1276G>A	p.Ala138Glu p.Ala426Thr	E5 E12	SRCR Serine protease	— —	53 years	—	—	Nucleus Hybrid CI S8 Array	50–60–90–110–110 dB (0.125, 0.25, 0.5, 1, 2 kHz)
	c.1345–2A>G[a]	—	In12	—	—	38 years	—	—	Nucelus Hybrid CI L24 Array	35–30–55–110–110 dB (0.125, 0.25, 0.5, 1, 2 kHz)
Song et al., 2020 (Korea)	c.916G>A c.1039G>T	p.Ala306Thr p.Glu347Ter	E9 E10	Serine protease Serine protease	— —	17 years (female)	3–5 years	Sloping HL 40–90–100–100–110–110 dB (0.25, 0.5, 1, 2, 4, 8 kHz)	—	86% WRS at 12 months following implantation
Holder et al., 2021 (United States)	c.208delC c.916G>A	p.Thr70fs[a]19 p.Ala306Thr	E4 E9	 Serine protease	—	54 months (female)	—	Sloping HL 20–25–95–110–100 dB (0.25, 0.5, 1, 2, 4 kHz)	Cochlear Nucleus 522/ 532 (left/right)	CNC 84%; BabyBio Quiet 94%/ 92% (left/right)
	c.208delC c.916G>A	p.Thr70fs[a]19 p.Ala306Thr	E4 E9	 Serine protease	—	47 months (female)	—	Sloping HL 20–20–75–115–115 dB (0.25, 0.5, 1, 2, 4 kHz)	Cochlear Nucleus 522/ 522 (left/right)	CNC 72%; BabyBio Quiet 55%
	c.208delC c.916G>A	p.Thr70fs[a]19 p.Ala306Thr	E4 E9	 Serine protease	—	43 months (female)	—	Sloping HL 20–25–15–95–110 (0.25, 0.5, 1, 2, 4 kHz)	Cochlear Nucleus 532/ 532 (left/right)	LNT 92%/82% (left/right); HINT 62%

HL, hearing loss; CI, cochlear implant; LDLRA, LDL receptor-like domain; dB, decibel; WRS, word-recognition score; SRCR, scavenger receptor cysteine-rich domain; serine protease, trypsin-like serine protease domain; PTA, pure tone average; CNC, consonant-nucleus-consonant; HINT, hearing in noise test; HA, hearing aid; LNT, lexical neighborhood test. Naming of variants and labeling of domains and exons are based on the NM_001256317.3 transcript. Of note, the phenotype severity is provided at the time of testing. While some patients may initially have milder phenotypes, the hearing loss can progress and become more severe.

[a]*Patient is homozygous for the specified variant.*

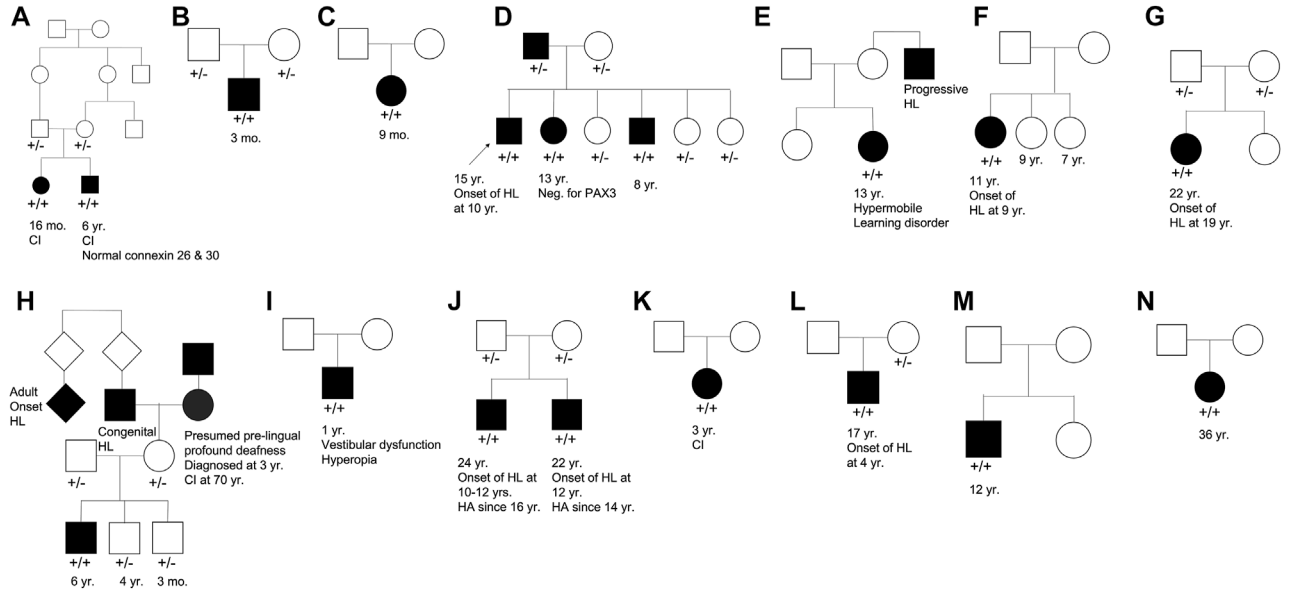


FIGURE 2 | Pedigree chart for enrolled patient families **(A–N)**. Age at genetic testing, age at onset of hearing loss, and other relevant clinical information is provided, when available, for patients and family members. CI, cochlear implant; HL, hearing loss; HA, hearing aid.

TABLE 3 | Genotype and phenotype overview of our patient cohort.

Family	Age	Gender	DNA change	Protein change	Configuration	HL onset	HL severity
A	16 months	F	**Deletion of Exons 1–5 and13**[a]	—	—	Congenital	—
	6 years	M	**Deletion of Exons 1–5 and13**[a]	—	—	Congenital	—
B	3 months	M	c.208delC[a]	p.Thr70fs[a]19	—	Congenital	Profound
C	9 months	F	c.208delC; c.1192C>T	p.Thr70fs[a]19; p.Gln398X	—	Congenital	Profound
D	8 years	M	c.208delC; c.1276G>A	p.Thr70fs[a]19; p.Ala426Thr	—	—	Sloping hearing loss
	13 years	F			—	—	Sloping sensorineural hearing loss
	15 years	M			Trans	10 years old	Progressive sloping, moderate left, severe right
E	13 years	F	c.208delC; c.413C>G	p.Thr70fs[a]19; p.Ala138Glu	—	—	Progressive, sloping, severe
F	11 years	F	c.208delC; c.413C>G	p.Thr70fs[a]19; p.Ala138Glu	—	9 years old	Sloping, profound
G	22 years	F	c.208delC; c.413C>G	p.Thr70fs[a]19; p.Ala138Glu	—	19 years old	—
H	6 years	M	c.323-6G>A; c.325C>T	-; p.Arg109Trp	—	Congenital	Moderately severe to profound
I	1 year	M	c.579dupA; c.1183G>C	p.Cys194MetfsX17; p.Asp395His	Trans	Congenital	Severe to profound
J	22 years	M	c.238C>T; c.1343T>C	p.Arg80Cys; p.Met448Thr	—	12 years old	Progressive, moderate-severe left, severe right
	24 years	M	c.238C>T; c.1343T>C	p.Arg80Cys; p.Met448Thr	—	10–12 years old	—
K	3 years	F	c.310G>A; c.916G>A	p.Glu104Lys; p.Ala306Thr	—	—	Moderately severe at low frequencies, profound at high frequencies
L	17 years	M	c.325C>T; c.413C>G	p.Arg109Trp; p.Ala138Glu	—	4 years old	Moderate-severe
M	12 years	M	c.413C>G; c.916G>A	p.Ala138Glu; p.Ala306Thr	—	Congenital	Progressive, high frequency, moderate
N	36 years	M	c.208delC; c.1306C>T	p.Thr70fs[a]19; p.Arg436Gly	—	Congenital	Progressive, profound

HL, hearing loss. Novel variant is bolded. Naming of variants is based on the NM_001256317.3 transcript.
[a]*patient is homozygous for the specified variant.*

individuals (Wattenhofer et al., 2002). However, a recent epidemiological study of patients undergoing CI revealed that 10% (13) of patients with positive genetic testing had *TMPRSS3* gene variants (Seligman et al., 2021). As adoption of genetic testing in clinical practice continues to grow, it is important to be aware of common *TMPRSS3* variants and associated phenotypes to best counsel patients.

In our cohort of 18 patients, 15 of whom were White, the most frequently observed variants were p.Thr70fs*19 and p.Ala138Glu implying that those were either hot spots or founder variants. The combination of the p.Thr70fs*19 frameshift variant with a missense variant appeared to cause sloping hearing loss that varied in severity. Biallelic pLOF variants appeared to cause congenital profound hearing loss. This phenotype information is valuable when trying to understand potential patient prognosis based on genetic testing results.

Previous studies on the role of CI in patients with *TMPRSS3* variants have reported variable results. In one study, poor outcomes following CI in patients with *TMPRSS3* variants were attributed to the expression of the *TMPRSS3* gene in SGNs as opposed to other locations in the cochlea such as the membranous labyrinth (Eppsteiner et al., 2012). These authors also suggested that patients with pathogenic *TMPRSS3* variants may have continued loss of SGNs over time which could contribute to ongoing hearing deterioration even after CI. However, recent studies have shown predominantly positive outcomes following CI in patients with *TMPRSS3* variants (Weegerink et al., 2011; Miyagawa et al., 2013; Chung et al., 2014; Miyagawa et al., 2015; Battelino et al., 2016; Gao et al., 2017; Shearer et al., 2018; Song et al., 2020; Holder et al., 2021). This discrepancy might be related to the large duration of deafness and older age of the two patients in Eppsteiner et al. (2012) and Holder et al. (2021). In addition, a study of CI outcomes in pediatric patients with *TMPRSS3* variants reported positive outcomes with no evidence of SGN degeneration leading to decreased performance over time (Holder et al., 2021). Furthermore, it was suggested that even if SGN degeneration does contribute to a longitudinal decline in performance, early CI may help slow or reverse this process (Holder et al., 2021). Even so, many clinics do not implant patients with precipitously sloping hearing loss as they do not meet *labeled* indications for CI. However, *off-label* implantation has been shown to be beneficial and is being employed much more frequently at major academic medical centers (Carlson et al., 2015; Leigh et al., 2016; Carlson et al., 2018).

Taken together with the positive clinical outcomes following CI in two patients from our cohort, it is evident that CI is a promising treatment strategy for patients with *TMPRSS3* variants. Active intervention with CI is likely to be beneficial, particularly in patients in whom residual hearing is preserved. It is imperative that the benefits of CI are made clear when counseling patients on their potential treatment options.

DATA AVAILABILITY STATEMENT

The evidence for all variants classified by the authors is included in submissions to ClinVar by the Laboratory for Molecular Medicine (Organization ID: 21766). All other data supporting the conclusions of this article, if not directly included in the paper, will be made available by the authors, without undue reservation.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Mass General Brigham Human Research Committee's IRB. Written informed consent to participate in this study was provided by the participants' legal guardian/next of kin.

AUTHOR CONTRIBUTIONS

IM and AG co-wrote the manuscript and prepared the tables and figures, VS edited the manuscript and prepared the tables and figures for submission, HR edited the manuscript and provided technical feedback, KS conceived, designed, and supervised the manuscript writing and editing.

FUNDING

We gratefully acknowledge support from the National Institutes of Health grant R01 DC015824 (KMS) and Jennifer and Louis Hernandez (KMS).

REFERENCES

Ahmed, Z. M., Cindy Li, X., Powell, S. D., Riazuddin, S., Young, T.-L., Ramzan, K., et al. (2004). Characterization of a New Full Length TMPRSS3 Isoform and Identification of Mutant Alleles Responsible for Nonsyndromic Recessive Deafness in Newfoundland and Pakistan. *BMC Med. Genet.* 5, 24. doi:10.1186/1471-2350-5-24

Battelino, S., Klancar, G., Kovac, J., Battelino, T., and Trebusak Podkrajsek, K. (2016). TMPRSS3 Mutations in Autosomal Recessive Nonsyndromic Hearing Loss. *Eur. Arch. Otorhinolaryngol.* 273 (5), 1151–1154. doi:10.1007/s00405-015-3671-0

Ben-Yosef, T., Wattenhofer, M., Riazuddin, S., Ahmed, Z. M., Scott, H. S., Kudoh, J., et al. (2001). Novel Mutations of TMPRSS3 in Four DFNB8/B10 Families Segregating Congenital Autosomal Recessive Deafness. *J. Med. Genet.* 38 (6), 396–400. doi:10.1136/jmg.38.6.396

Bowles, B., Ferrer, A., Nishimura, C. J., Pinto E Vairo, F., Rey, T., Leheup, B., et al. (2021). TSPEAR Variants Are Primarily Associated with Ectodermal Dysplasia and Tooth Agenesis but Not Hearing Loss: A Novel Cohort Study. *Am. J. Med. Genet.* 185 (8), 2417–2433. doi:10.1002/ajmg.a.62347

Capalbo, A., Valero, R. A., Jimenez-Almazan, J., Pardo, P. M., Fabiani, M., Jiménez, D., et al. (2019). Optimizing Clinical Exome Design and Parallel Gene-Testing for Recessive Genetic Conditions in Preconception Carrier Screening: Translational Research Genomic Data from 14,125 Exomes. *Plos Genet.* 15 (10), e1008409. doi:10.1371/journal.pgen.1008409

Carlson, M. L., Sladen, D. P., Gurgel, R. K., Tombers, N. M., Lohse, C. M., and Driscoll, C. L. (2018). Survey of the American Neurotology Society on Cochlear Implantation: Part 1, Candidacy Assessment and Expanding Indications. *Otol. Neurotol.* 39 (1), e12–e19. doi:10.1097/MAO.0000000000001632

Carlson, M. L., Sladen, D. P., Haynes, D. S., Driscoll, C. L., DeJong, M. D., Erickson, H. C., et al. (2015). Evidence for the Expansion of Pediatric

Cochlear Implant Candidacy. *Otol. Neurotol.* 36 (1), 43–50. doi:10.1097/MAO.0000000000000607

Charif, M., Abidi, O., Boulouiz, R., Nahili, H., Rouba, H., Kandil, M., et al. (2012). Molecular Analysis of the TMPRSS3 Gene in Moroccan Families with Non-syndromic Hearing Loss. *Biochem. Biophysical Res. Commun.* 419 (4), 643–647. doi:10.1016/j.bbrc.2012.02.066

Chung, J., Park, S. M., Chang, S. O., Chung, T., Lee, K. Y., Kim, A. R., et al. (2014). A Novel Mutation of TMPRSS3 Related to Milder Auditory Phenotype in Korean Postlingual Deafness: a Possible Future Implication for a Personalized Auditory Rehabilitation. *J. Mol. Med.* 92 (6), 651–663. doi:10.1007/s00109-014-1128-3

Diaz-Horta, O., Duman, D., Foster, J., Sırmacı, A., Gonzalez, M., Mahdieh, N., et al. (2012). Whole-exome Sequencing Efficiently Detects Rare Mutations in Autosomal Recessive Nonsyndromic Hearing Loss. *PLoS One* 7 (11), e50628. doi:10.1371/journal.pone.0050628

DiStefano, M. T., Hemphill, S. E., Cushman, B. J., Bowser, M. J., Hynes, E., Grant, A. R., et al. (2018). Curating Clinically Relevant Transcripts for the Interpretation of Sequence Variants. *J. Mol. Diagn.* 20 (6), 789–801. doi:10.1016/j.jmoldx.2018.06.005

Duman, D., Sirmaci, A., Cengiz, F. B., Ozdag, H., and Tekin, M. (2011). Screening of 38 Genes Identifies Mutations in 62% of Families with Nonsyndromic Deafness in Turkey. *Genet. Test. Mol. Biomarkers* 15 (1-2), 29–33. doi:10.1089/gtmb.2010.0120

Duzkale, H., Shen, J., McLaughlin, H., Alfares, A., Kelly, M., Pugh, T., et al. (2013). A Systematic Approach to Assessing the Clinical Significance of Genetic Variants. *Clin. Genet.* 84 (5), 453–463. doi:10.1111/cge.12257

Elbracht, M., Senderek, J., Eggermann, T., Thurmer, C., Park, J., Westhofen, M., et al. (2007). Autosomal Recessive Postlingual Hearing Loss (DFNB8): Compound Heterozygosity for Two Novel TMPRSS3 Mutations in German Siblings. *J. Med. Genet.* 44 (6), e81. doi:10.1136/jmg.2007.049122

Eppsteiner, R. W., Shearer, A. E., Hildebrand, M. S., Deluca, A. P., Ji, H., Dunn, C. C., et al. (2012). Prediction of Cochlear Implant Performance by Genetic Mutation: the Spiral Ganglion Hypothesis. *Hearing Res.* 292 (1-2), 51–58. doi:10.1016/j.heares.2012.08.007

Fan, D., Zhu, W., Li, D., Ji, D., and Wang, P. (2014). Identification of a Novel Homozygous Mutation, TMPRSS3: c.535G>A, in a Tibetan Family with Autosomal Recessive Non-syndromic Hearing Loss. *PLoS One* 9 (12), e114136. doi:10.1371/journal.pone.0114136

Ganapathy, A., Pandey, N., Srisailapathy, C. R. S., Jalvi, R., Malhotra, V., Venkatappa, M., et al. (2014). Non-syndromic Hearing Impairment in India: High Allelic Heterogeneity Among Mutations in TMPRSS3, TMC1, USHIC, CDH23 and TMIE. *PLoS One* 9 (1), e84773. doi:10.1371/journal.pone.0084773

Gao, X., Huang, S.-S., Yuan, Y.-Y., Xu, J.-C., Gu, P., Bai, D., et al. (2017). Identification ofTMPRSS3as a Significant Contributor to Autosomal Recessive Hearing Loss in the Chinese Population. *Neural Plasticity* 2017, 1–8. doi:10.1155/2017/3192090

Gu, X., Guo, L., Ji, H., Sun, S., Chai, R., Wang, L., et al. (2015). Genetic Testing for Sporadic Hearing Loss Using Targeted Massively Parallel Sequencing Identifies 10 Novel Mutations. *Clin. Genet.* 87 (6), 588–593. doi:10.1111/cge.12431

Guipponi, M., Vuagniaux, G., Wattenhofer, M., Shibuya, K., Vazquez, M., Dougherty, L., et al. (2002). The Transmembrane Serine Protease (TMPRSS3) Mutated in Deafness DFNB8/10 Activates the Epithelial Sodium Channel (ENaC) *In Vitro*. *Hum. Mol. Genet.* 11 (23), 2829–2836. doi:10.1093/hmg/11.23.2829

Holder, J. T., Morrel, W., Rivas, A., Labadie, R. F., and Gifford, R. H. (2021). Cochlear Implantation and Electric Acoustic Stimulation in Children with TMPRSS3 Genetic Mutation. *Otol. Neurotol.* 42 (3), 396–401. doi:10.1097/MAO.0000000000002943

Khan, A., Han, S., Wang, R., Ansar, M., Ahmad, W., and Zhang, X. (2019). Sequence Variants in Genes Causing Nonsyndromic Hearing Loss in a Pakistani Cohort. *Mol. Genet. Genomic Med.* 7 (9), e917. doi:10.1002/mgg3.917

Kim, A., Chung, J., Kim, N., Lee, C., Park, W.-Y., Oh, D.-Y., et al. (2017). The Analysis of A Frequent TMPRSS3 Allele Containing P.V116M and P.V291L in A Cis Configuration Among Deaf Koreans. *Ijms* 18 (11), 2246. doi:10.3390/ijms18112246

Lechowicz, U., Gambin, T., Pollak, A., Podgorska, A., Stawinski, P., Franke, A., et al. (2017). Iterative Sequencing and Variant Screening (ISVS) as a Novel Pathogenic Mutations Search Strategy - Application for TMPRSS3 Mutations Screen. *Sci. Rep.* 7 (1), 2543. doi:10.1038/s41598-017-02315-w

Lee, J., Baek, J.-I., Choi, J. Y., Kim, U.-K., Lee, S.-H., and Lee, K.-Y. (2013). Genetic Analysis of TMPRSS3 Gene in the Korean Population with Autosomal Recessive Nonsyndromic Hearing Loss. *Gene* 532 (2), 276–280. doi:10.1016/j.gene.2013.07.108

Lee, K., Khan, S., Islam, A., Ansar, M., Andrade, P. B., Kim, S., et al. (2012). Novel TMPRSS3 Variants in Pakistani Families with Autosomal Recessive Non-syndromic Hearing Impairment. *Clin. Genet.* 82 (1), 56–63. doi:10.1111/j.1399-0004.2011.01695.x

Lee, Y. J., Park, D., Kim, S. Y., and Park, W. J. (2003). Pathogenic Mutations but Not Polymorphisms in Congenital and Childhood Onset Autosomal Recessive Deafness Disrupt the Proteolytic Activity of TMPRSS3. *J. Med. Genet.* 40 (8), 629–631. doi:10.1136/jmg.40.8.629

Leigh, J. R., Moran, M., Hollow, R., and Dowell, R. C. (2016). Evidence-based Guidelines for Recommending Cochlear Implantation for Postlingually Deafened Adults. *Int. J. Audiol.* 55 (2), S3–S8. doi:10.3109/14992027.2016.1146415

Leone, M. P., Palumbo, P., Ortore, R., Castellana, S., Palumbo, O., Melchionda, S., et al. (2017). Putative TMPRSS3/GJB2 Digenic Inheritance of Hearing Loss Detected by Targeted Resequencing. *Mol. Cell Probes* 33, 24–27. doi:10.1016/j.mcp.2017.03.001

Li, X., Tan, B., Wang, X., Xu, X., Wang, C., Zhong, M., et al. (2019). Identification of a Complex Genomic Rearrangement in TMPRSS3 by Massively Parallel Sequencing in Chinese Cases with Prelingual Hearing Loss. *Mol. Genet. Genomic Med.* 7 (6), e685. doi:10.1002/mgg3.685

Likar, T., Hasanhodžić, M., Teran, N., Maver, A., Peterlin, B., and Writzl, K. (2018). Diagnostic Outcomes of Exome Sequencing in Patients with Syndromic or Non-syndromic Hearing Loss. *PLoS One* 13 (1), e0188578. doi:10.1371/journal.pone.0188578

Masmoudi, S., Antonarakis, S. E., Schwede, T., Ghorbel, A. M., Gratri, M. h., Pappasavas, M.-P., et al. (2001). Novel Missense Mutations of TMPRSS3 in Two Consanguineous Tunisian Families with Non-syndromic Autosomal Recessive Deafness. *Hum. Mutat.* 18 (2), 101–108. doi:10.1002/humu.1159

Miyagawa, M., Nishio, S.-y., Ikeda, T., Fukushima, K., and Usami, S.-i. (2013). Massively Parallel DNA Sequencing Successfully Identifies New Causative Mutations in Deafness Genes in Patients with Cochlear Implantation and EAS. *PLoS One* 8 (10), e75793. doi:10.1371/journal.pone.0075793

Miyagawa, M., Nishio, S.-y., Sakurai, Y., Hattori, M., Tsukada, K., Moteki, H., et al. (2015). The Patients Associated with TMPRSS3 Mutations Are Good Candidates for Electric Acoustic Stimulation. *Ann. Otol. Rhinol. Laryngol.* 124 (1), 193S–204S. doi:10.1177/0003489415575056

Noman, M., Ishaq, R., Bukhari, S. A., Ahmed, Z. M., and Riazuddin, S. (2019). Delineation of Homozygous Variants Associated with Prelingual Sensorineural Hearing Loss in Pakistani Families. *Genes* 10 (12), 1031. doi:10.3390/genes10121031

Pugh, T. J., Amr, S. S., Bowser, M. J., Gowrisankar, S., Hynes, E., Mahanta, L. M., et al. (2016). VisCap: Inference and Visualization of Germ-Line Copy-Number Variants from Targeted Clinical Sequencing Data. *Genet. Med.* 18 (7), 712–719. doi:10.1038/gim.2015.156

Rawlings, N. D., Barrett, A. J., and Bateman, A. (2010). MEROPS: the Peptidase Database. *Nucleic Acids Res.* 38, D227–D233. doi:10.1093/nar/gkp971

Richards, S., Aziz, N., Aziz, N., Bale, S., Bick, D., Das, S., et al. (2015). Standards and Guidelines for the Interpretation of Sequence Variants: a Joint Consensus Recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet. Med.* 17 (5), 405–423. doi:10.1038/gim.2015.30

Safka Brozkova, D., Poisson Marková, S., Mészárosová, A. U., Jenčík, J., Čejnová, V., Čada, Z., et al. (2020). Spectrum and Frequencies of Non GJB2 Gene Mutations in Czech Patients with Early Non-syndromic Hearing Loss Detected by Gene Panel NGS and Whole-exome Sequencing. *Clin. Genet.* 98 (6), 548–554. doi:10.1111/cge.13839

Sang, S., Ling, J., Liu, X., Mei, L., Cai, X., Li, T., et al. (2019). Proband Whole-Exome Sequencing Identified Genes Responsible for Autosomal Recessive Non-syndromic Hearing Loss in 33 Chinese Nuclear Families. *Front. Genet.* 10, 639. doi:10.3389/fgene.2019.00639

Sarrias, M. R., Gronlund, J., Padilla, O., Madsen, J., Holmskov, U., and Lozano, F. (2004). The Scavenger Receptor Cysteine-Rich (SRCR) Domain: an Ancient and Highly Conserved Protein Module of the Innate Immune System. *Crit. Rev. Immunol.* 24 (1), 1–38. doi:10.1615/critrevimmunol.v24.i1.10

Scott, H. S., Kudoh, J., Wattenhofer, M., Shibuya, K., Berry, A., Chrast, R., et al. (2001). Insertion of β-satellite Repeats Identifies a Transmembrane Protease Causing Both Congenital and Childhood Onset Autosomal Recessive Deafness. *Nat. Genet.* 27 (1), 59–63. doi:10.1038/83768

Seligman, K. L., Shearer, A. E., Frees, K., Nishimura, C., Kolbe, D., Dunn, C., et al. (2021). Genetic Causes of Hearing Loss in a Large Cohort of Cochlear Implant Recipients. *Otolaryngol. Head Neck Surg.*, 019459982110213. doi:10.1177/01945998211021308

Shafique, S., Siddiqi, S., Schraders, M., Oostrik, J., Ayub, H., Bilal, A., et al. (2014). Genetic Spectrum of Autosomal Recessive Non-syndromic Hearing Loss in Pakistani Families. *PLoS One* 9 (6), e100146. doi:10.1371/journal.pone.0100146

Shearer, A. E., Tejani, V. D., Brown, C. J., Abbas, P. J., Hansen, M. R., Gantz, B. J., et al. (2018). *In Vivo* Electrocochleography in Hybrid Cochlear Implant Users Implicates TMPRSS3 in Spiral Ganglion Function. *Sci. Rep.* 8 (1), 14165. doi:10.1038/s41598-018-32630-9

Singh, K., Bijarnia-Mahay, S., Ramprasad, V. L., Puri, R. D., Nair, S., Sharda, S., et al. (2020). NGS-based Expanded Carrier Screening for Genetic Disorders in North Indian Population Reveals Unexpected Results - a Pilot Study. *BMC Med. Genet.* 21 (1), 216. doi:10.1186/s12881-020-01153-4

Song, M. H., Jung, J., Rim, J. H., Choi, H. J., Lee, H. J., Noh, B., et al. (2020). Genetic Inheritance of Late-Onset, Down-Sloping Hearing Loss and its Implications for Auditory Rehabilitation. *Ear Hear* 41 (1), 114–124. doi:10.1097/aud.0000000000000734

Südhof, T. C., Goldstein, J. L., Brown, M. S., and Russell, D. W. (1985). The LDL Receptor Gene: a Mosaic of Exons Shared with Different Proteins. *Science* 228 (4701), 815–822. doi:10.1126/science.2988123

Tayoun, A. N. A., Mason-Suares, H., Frisella, A. L., Bowser, M., Duffy, E., Mahanta, L., et al. (2016). Targeted Droplet-Digital PCR as a Tool for Novel Deletion Discovery at the DFNB1 Locus. *Hum. Mutat.* 37 (1), 119–126. doi:10.1002/humu.22912

van Driel, I. R., Goldstein, J. L., Südhof, T. C., and Brown, M. S. (1987). First Cysteine-Rich Repeat in Ligand-Binding Domain of Low Density Lipoprotein Receptor Binds Ca2+ and Monoclonal Antibodies, but Not Lipoproteins. *J. Biol. Chem.* 262 (36), 17443–17449. doi:10.1016/s0021-9258(18)45399-9

Vozzi, D., Morgan, A., Vuckovic, D., D'Eustacchio, A., Abdulhadi, K., Rubinato, E., et al. (2014). Hereditary Hearing Loss: a 96 Gene Targeted Sequencing Protocol Reveals Novel Alleles in a Series of Italian and Qatari Patients. *Gene* 542 (2), 209–216. doi:10.1016/j.gene.2014.03.033

Walsh, T., Rayan, A. A., Sa'ed, J. A., Shahin, H., Shepshelovich, J., Lee, M. K., et al. (2006). Genomic Analysis of a Heterogeneous Mendelian Phenotype: Multiple Novel Alleles for Inherited Hearing Loss in the Palestinian Population. *Hum. Genomics* 2 (4), 203–211. doi:10.1186/1479-7364-2-4-203

Wattenhofer, M., Di Iorio, M., Rabionet, R., Dougherty, L., Pampanos, A., Schwede, T., et al. (2002). Mutations in the TMPRSS3 Gene Are a Rare Cause of Childhood Nonsyndromic Deafness in Caucasian Patients. *J. Mol. Med.* 80 (2), 124–131. doi:10.1007/s00109-001-0310-6

Wattenhofer, M., Sahin-Calapoglu, N., Andreasen, D., Kalay, E., Caylan, R., Braillard, B., et al. (2005). A Novel TMPRSS3 Missense Mutation in a DFNB8/10 Family Prevents Proteolytic Activation of the Protein. *Hum. Genet.* 117 (6), 528–535. doi:10.1007/s00439-005-1332-x

Weegerink, N. J. D., Schraders, M., Oostrik, J., Huygen, P. L. M., Strom, T. M., Granneman, S., et al. (2011). Genotype-phenotype Correlation in DFNB8/10 Families with TMPRSS3 Mutations. *Jaro* 12 (6), 753–766. doi:10.1007/s10162-011-0282-3

Wong, S.-H., Yen, Y.-C., Li, S.-Y., and Yang, J.-J. (2020). Novel Mutations in the TMPRSS3 Gene May Contribute to Taiwanese Patients with Nonsyndromic Hearing Loss. *Ijms* 21 (7), 2382. doi:10.3390/ijms21072382

Zafar, S., Shahzad, M., Ishaq, R., Yousaf, A., Shaikh, R. S., Akram, J., et al. (2020). Novel Mutations in *CLPP, LARS2, CDH23,* and *COL4A5* Identified in Familial Cases of Prelingual Hearing Loss. *Genes* 11 (9), 978. doi:10.3390/genes11090978

Zhou, Y., Tariq, M., He, S., Abdullah, U., Zhang, J., and Baig, S. M. (2020). Whole Exome Sequencing Identified Mutations Causing Hearing Loss in Five Consanguineous Pakistani Families. *BMC Med. Genet.* 21 (1), 151. doi:10.1186/s12881-020-01087-x

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

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

frontiers in Genetics

ORIGINAL RESEARCH
published: 01 December 2021
doi: 10.3389/fgene.2021.747090

A New HPV score System Predicts the Survival of Patients With Cervical Cancers

Qunchao Hu[1], Yani Wang[2], Yuchen Zhang[2], Yanjun Ge[2], Yihua Yin[2] and Haiyan Zhu[2]**

[1]Department of Radiation Oncology, Shanghai First Maternity and Infant Hospital, Tongji University School of Medicine, Shanghai, China, [2]Department of Gynecology, Shanghai First Maternity and Infant Hospital, Tongji University School of Medicine, Shanghai, China

Persistent high-risk human papillomavirus (hrHPV) infection is confirmed as the major cause of cervical cancer. According to the HPV infection status, cervical cancer could be generalized as following three subgroups: HPV-negative, pure HPV-infection, and HPV-integration. Currently, the impact of HPV status on cervical cancer prognosis remains under dispute. Therefore, we explored the potential correlation between HPV status and the clinical outcome of cervical cancer, by establishing a robust prognostic predicting model based on a cervical cancer cohort using The Cancer Genome Atlas (TCGA) database. We performed an iCluster algorithm incorporating DNA copy number variation, SNP, DNA methylation, mRNA expression, and miRNA expression profile together and classified the cohort into three clusters. According to defined clusters, we established an HPV score system by weighing resultant gene alterations through random forest and COX models. This prediction tool could help to identify cervical cancer prognosis through evaluating individual HPV infection status and subsequent genetic modification, which might provide insights into HPV-related gene driven cervical cancer treatment strategies, yet its predictive power and robustness need to be further verified with independent cohorts.

Keywords: HPV status, multi-omics, cervical cancer, HPV score, prognosis

OPEN ACCESS

Edited by:
Xiangqian Guo,
Henan University, China

Reviewed by:
Jlayi Ji,
Icahn School of Medicine at Mount Sinai, United States
Musalula Sinkala,
University of Cape Town, South Africa

****Correspondence:***
Yihua Yin
YINyihua123@outlook.com
Haiyan Zhu
zhuhaiyandoc@sina.com

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

Received: *25 July 2021*
Accepted: *15 November 2021*
Published: *01 December 2021*

Citation:
Hu Q, Wang Y, Zhang Y, Ge Y, Yin Y and Zhu H (2021) A New HPV score System Predicts the Survival of Patients With Cervical Cancers. Front. Genet. 12:747090. doi: 10.3389/fgene.2021.747090

INTRODUCTION

Cervical cancer (CC) is one of the most perplexing women's health problems and the commonest gynecological cancers, ranking fifth in the incidence rate and mortality rate of women (Bray et al., 2018). Persistent high-risk HPV (hrHPV) infection has been identified as the main event leading to cervical cancer (Cohen et al., 2019). For women, the cumulative lifetime infection rate for HPV can range from 60 to 70%, however, only a few infections persist and eventually lead to cancer (Qingqing et al., 2020). It is generally believed that the integration of the HPV genome into the host chromosome is the key genetic step in the pathogenesis of cervical carcinomas (Pett and Coleman, 2007). It is reported that the majority of HPV18 and HPV16 related cervical cancer samples were found to have structural aberrations and increased target-gene(s) expression (Cancer Genome Atlas Research Network, 2017). HPV integration usually involves the inactivation of viral E1 and E2 regions, resulting in the upregulation of oncogenes E6 and E7. Then, E6 oncoprotein degrades p53, inhibits cancer cell apoptosis and viral DNA replication. E7 is known for suppressing RB1, which abrogates cell cycle arrest and stimulates proliferation (Hu and Ma, 2018).

Previous studies have classified cervical cancer into HPV-negative cervical cancer (HPV-) and HPV-positive cervical cancer (HPV+) and reported women with hrHPV-positive cervical tumors had a substantially better prognosis than those with hrHPV-negative tumors (Lei et al., 2018). Moreover, the HPV genotype showed an independent prognostic value in early-stage cervical cancer with contradictory findings (Lai et al., 2007; Cuschieri et al., 2014). Recently, HPV-positive cervical cancer was further classified into pure HPV-infection and HPV-integration (HPV-int), according to the characteristics of HPV invasion (Wei et al., 2015; Cancer Genome Atlas Research Network et al., 2017). Whether there are differential molecular characteristics between HPV-positive cancers and HPV-negative cervical cancers remains unclear. Furthermore, whether HPV status, including HPV-negative, HPV-infection, and HPV-int, can be stably utilized as a prognostic factor remains an issue.

The current study first explored different genomic characterizations according to HPV status, including HPV-negative, pure HPV-infection, and HPV-int, using TCGA multi-omics data, then established molecular subgroups based on HPV status. Finally, we developed an HPV score model that can effectively predict the survival of cervical cancer.

MATERIALS AND METHODS

Data Collection

Genomic, transcriptomic, and epigenomic data for cervical squamous cell carcinoma and endocervical adenocarcinoma (CESC) were downloaded from the TCGA database. We downloaded MAF files for cervical cancer (the reference genome of the MAF file is HG19) using the R-package of TCGAbiolinks, which contains the mutation detection results from 297 samples, while SNP Copy Number segment data for 287 cervical cancer samples and the methylation microarray data for 299 cervical cancer samples were drawn from Fire Browse (http://firebrowse.org/). The mRNA expression profiles of 307 cervical cancer samples were downloaded with UCSC Xena (https://xenabrowser.net/datapages/). There were 284 samples with multi-omics data, including DNA mutation, copy number alteration, methylation, mRNA, and miRNA expression profiles. Clinical data were downloaded from the TCGA data portal, including primary tumor histological type and grade, HPV infectious status, and FIGO stage. The follow-up analysis was based on 283 samples with the full information collected.

Data Analysis

The Differentially Expressed Genes and Differentially Methylation Genes in Cervical Squamous Cell Carcinoma and Endocervical Adenocarcinoma From The Cancer Genome Atlas

The differentially expressed genes (DEGs) and methylation genes (DMGs) between the HPV + group and the HPV- group were identified by R programming, using the limma package and the ChAMP package, respectively. As for DEGs of mRNA and miRNA, fold change values larger than 1.5 and p-value less than 0.05 were set as the cut-off criteria. DMGs were predicted based on the β-value file of CpG site probes, Δβ, and p-value. Only regions with $|\Delta\beta|$ larger than 0.1 and p-value less than 0.05 were identified as DMRs.

Gene Set Variation Analysis

Initially, we divided cervical cancer samples into two groups as HPV infected (HPV-positive) and non-HPV infected (HPV-negative), then downloaded hallmark reference gene set (c2.cp.kegg.v7.1) from the MSigDB database (https://www.gsea-msigdb.org/gsea/index.jsp) and set the p-value less than 0.05 and t-test value over 2 as cut-off criteria. We performed the GSVA enrichment analysis between two defined groups using R-package of GSVA (version 1.20.0), and also assessed the relative enrichment of gene sets across defined groups using a non-parametric approach, estimating changes in the activity of pathways and biological processes. Patients from the CESC cohort were divided into two groups according to the HPV infectious status. Genomic events like somatic mutations and somatic copy number alterations (CNAs) of samples from different groups were analyzed with the MutSigCV and GISTIC analysis modules in GenePattern (https://cloud.genepattern.org) accordingly. The mutation landscape was visualized by using the R-package maptools.

Subgroup Identification and Analysis of Different Genes Expression

As for subgroup identification, we integrated various molecular platforms by using iCluster, including DNA copy number, SNP, DNA methylation, mRNA expression, and miRNA expression profile (Keck et al., 2015). Then, the differentially expressed genes between subgroups of cervical cancer patients were screened by using the limma package in R. The screening criteria were $|\log FC| > \log 2\ (1.5)$ and p-value < 0.05.

Assessment of Infiltrated the Immune Cell

The "Cell type Identification By Estimating Relative Subsets Of RNA Transcripts (CIBERSORT)" algorithm (https://cibersort.stanford.edu/) was performed to complete the calculation, and each sample was assigned with a p-value. Samples with CIBERSORT output p-value less than 0.05 were gathered for subsequent analysis. Then it was used to indicate the composition of the tumor infiltrated immune cells with LM22 signature as a reference, which is based on the gene expression profiling of complex tissue (Newman et al., 2015). The proportion of 22 subtypes of immune cells was obtained, including myeloid subgroup, NK cell, plasma cell, juvenile, memory type B cells, and seven subgroups of T cells. Pearson's Chi-squared test was launched to estimate different immune cell infiltration statuses and the visualization was presented in the R program.

HPV Score Model Construction

According to the different expression gene patterns of the three clustered subgroups (as described previously in the section on "Subgroup identification and analysis of different genes expression"), the random forest method was applied to filtrate the redundant genes. Then, multivariate Cox regression combing patients' survival data was used to correlate candidate genes and

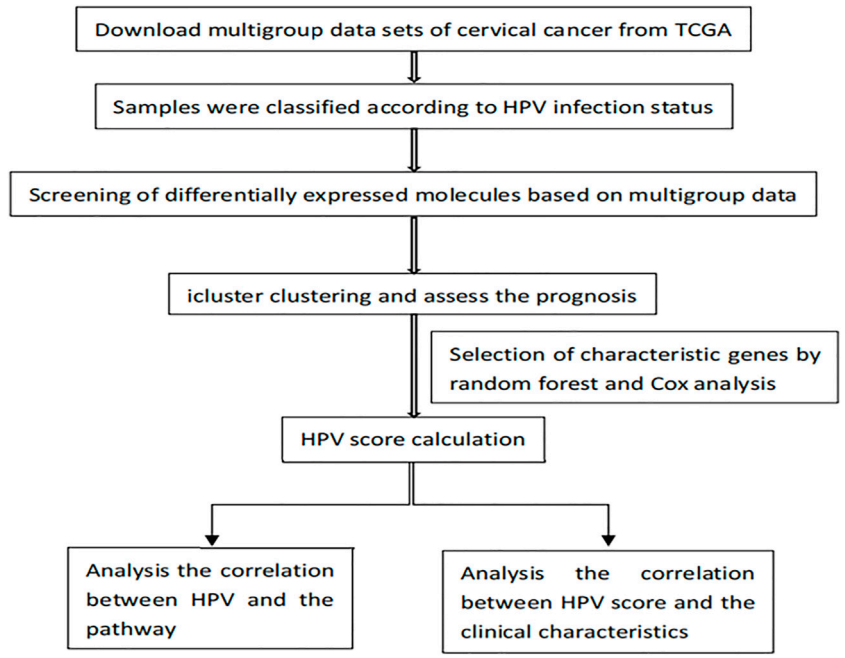


FIGURE1 | Flow chart.

divide them into positive or negative regulation groups. We took a *p*-value less than 0.05 as a threshold. Ultimately, the score of GCI was used to estimate the HPV score model as follows (Sotiriou et al., 2006; Zeng et al., 2019; Zhang et al., 2020).

$$\text{HPVscore} = \text{scale}\left(\sum X - \sum Y\right)$$

In this formulation, X indicates the expression value of the gene set with positive co-efficient(s), while Y refers to the negative correlation in Cox regression, and scale represents the standardization process.

The optimal threshold of predictive HPV score was determined by surv_cutpoint algorithm of survival R package. This function is used to explore the optimal cutoff value for one or multiple continuous variables by using the maximally selected rank statistics from the "maxstat" R package. Thus, the cervical cancer patients were divided into HPV high score and HPV low score groups according to individual calculation with the formula above. The survival difference between the HPV high score and low score was determined by the log-rank test with a *p*-value less than 0.05 as statistically significant.

Statistical Analysis

R software (version 3.6.2; https://www.r-project.org/) was used in data mining and statistical analyses. Wilcoxon test was used to evaluate the differences between the defined groups. The Student t-test, Pearson's Chi-squared test, or χ^2 test was applied to distinguish different expression profiles regards to defined groups. The survival curves were determined by the Kaplan-Meier method, while the log-rank test compared the clinical outcome differences. Based on specified genes of identified subgroups, the predictive HPV score model was constructed. Receiver operating characteristic (ROC) and area under the curve (AUC) were drawn with ROC and timeROC package in the R program to evaluate the model predictive ability and robustness. A *p*-value of <0.05 was considered statistically significant. The flow chart is shown in **Figure 1**.

RESULTS

Multi-Omics Data Analysis According to HPV Infection Status

We identified 263 differentially expressed mRNA, 28 differentially expressed miRNA, and 73 differentially DNA methylation sites between HPV-negative and HPV-positive groups. As shown in **Figure 2**, Gene Ontology clustering (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway analysis indicated that non-HPV related cervical cancers were associated with enrichment of biological function such as glycan biosynthesis, selenoamino acid metabolism, glyoxylate, and dicarboxylate metabolism, as well as aminoacyl Trina biosynthesis, when compared to those patients with HPV positive.

As for genomic aberrations between cervical cancer with pure HPV-infection and HPV-int, we found 98 genes SNP variants of significant differences, such as CNTNAP5, MYO7B, LMOD3, etc. Compared to pure HPV-infection, the HPV-int group presented a higher frequency of CNV events, including copy number gain in 1933 genes and copy number loss in 187 genes. As shown by the gene mutation in Figure 3AB, the most common mutation sites in pure HPV-infection and HPV-int samples were located in the

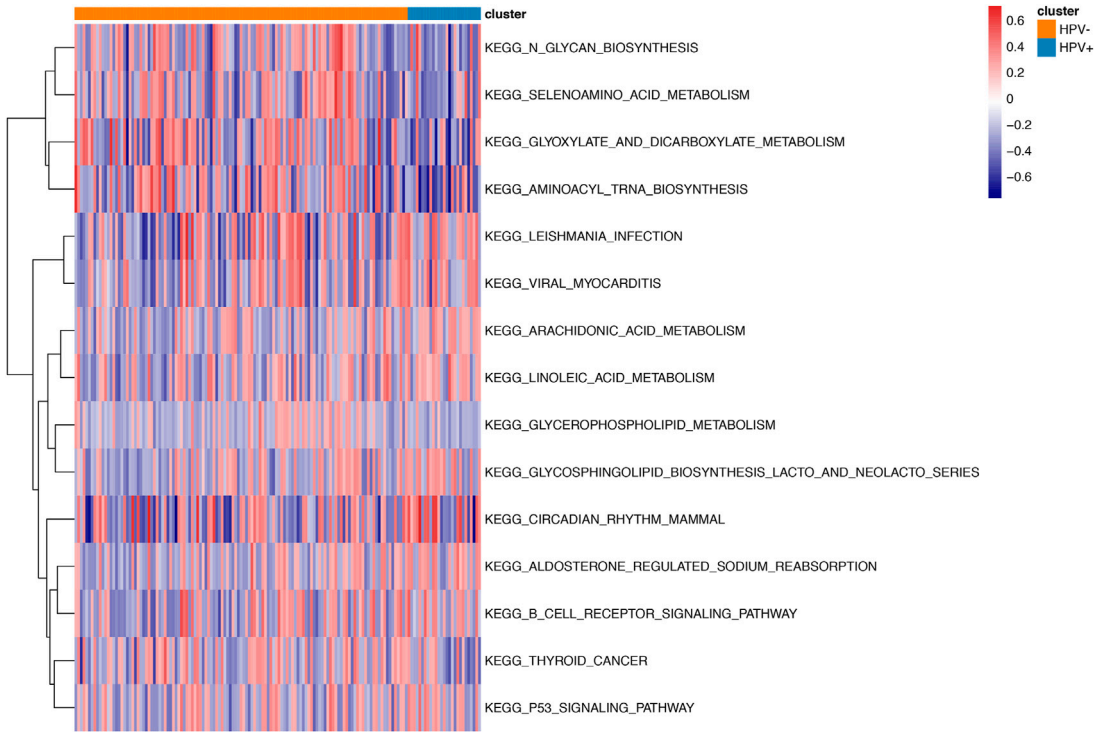


FIGURE2 | GSVA enrichment analysis between HPV positive and negative samples. Different colors stand for different status of biological pathway: red means activation and blue means inhibition.

genes TTN, MUC4, PIK3CA, and MUC16. HPV-int group demonstrated a higher frequency of gene alteration than the non-integrated group (96.3 vs 87.2%). PIK3CA gene missense mutation was found in 12 cases (44%, 12/27, **Figure 3B**). Figure 3CD shows the copy number variation enrichment in different genomic regions between pure HPV-infection and the HPV-int samples set. Copy number amplification is mainly located in chromosome 3q28 in the HPV-infection group and chromosome 3q26.2 in HPV-int set, while the copy number loss was mostly located in 2q37.1 in the HPV-infection samples and 2q37.3 in the HPV-int set.

HPV Related Subgroup Classification and Signature Analysis

We then performed an iCluster algorithm, incorporating five different omics of DNA copy, SNP, DNA methylation, mRNA, and miRNA expression profiles together and sorted them into three HPV-related clusters. As shown in **Figure 4A**, while cluster 1 and cluster 2 are mainly pure HPV-infection and HPV-int, cluster 3 is of a great proportion of non-HPV-related cervical cancer patients. Patients of cluster 1 presented superior overall survival to that of cluster 2 and cluster 3, though there is no statistical difference between cluster 2 and cluster 3, especially in the squamous cell cancer patients (**Figures 4B,C**). In addition, cluster 1 showed an improved OS compared to cluster 2/3 among patients in stage I/II cervical cancer (p = 0.0061). Among those three clusters, only patients of cluster 3 reached the median overall survival of 94.8 months (95%CI: 45.3–144.4). However, there was no significant difference in OS among the three clusters in patients with stage III/IV (**Figures 4D,E**).

CIBERSORT method combined with LM22 characteristic matrix was used to further estimate the differences of infiltrating immune cell subsets among those three clusters. The proportion of immune cells in each sample varied within and between groups (**Figure 5A**). There were significant differences in subsets of resting mast cells, macrophage M2, Neutrophils, and CD4 naïve T cells among the three groups (**Figure 5B**).

We analyzed the differential genes among these three clusters using the R-package limma. Finally, 460 differential expressed genes were identified (**Figure 6**).

HPV Score Model Construction

Since it is known that HPV status alone (HPV-negative, pure HPV-infection, HPV-int) cannot effectively predict the prognosis of cervical cancer, this study aimed to develop a multi-factor model, called the "HPV-score," according to virus infection status and resultant gene alterations according to defined clusters. We selected the characteristic genes that are most relevant to classification among those 460 differential genes using a random forest algorithm. The optimal threshold point (cutoff = 1.15482) for the classification of HPV-score was determined according to the surv_cutpoint function in the R-package survminer, and the samples were divided into two categories: HPV-score high and HPV-score

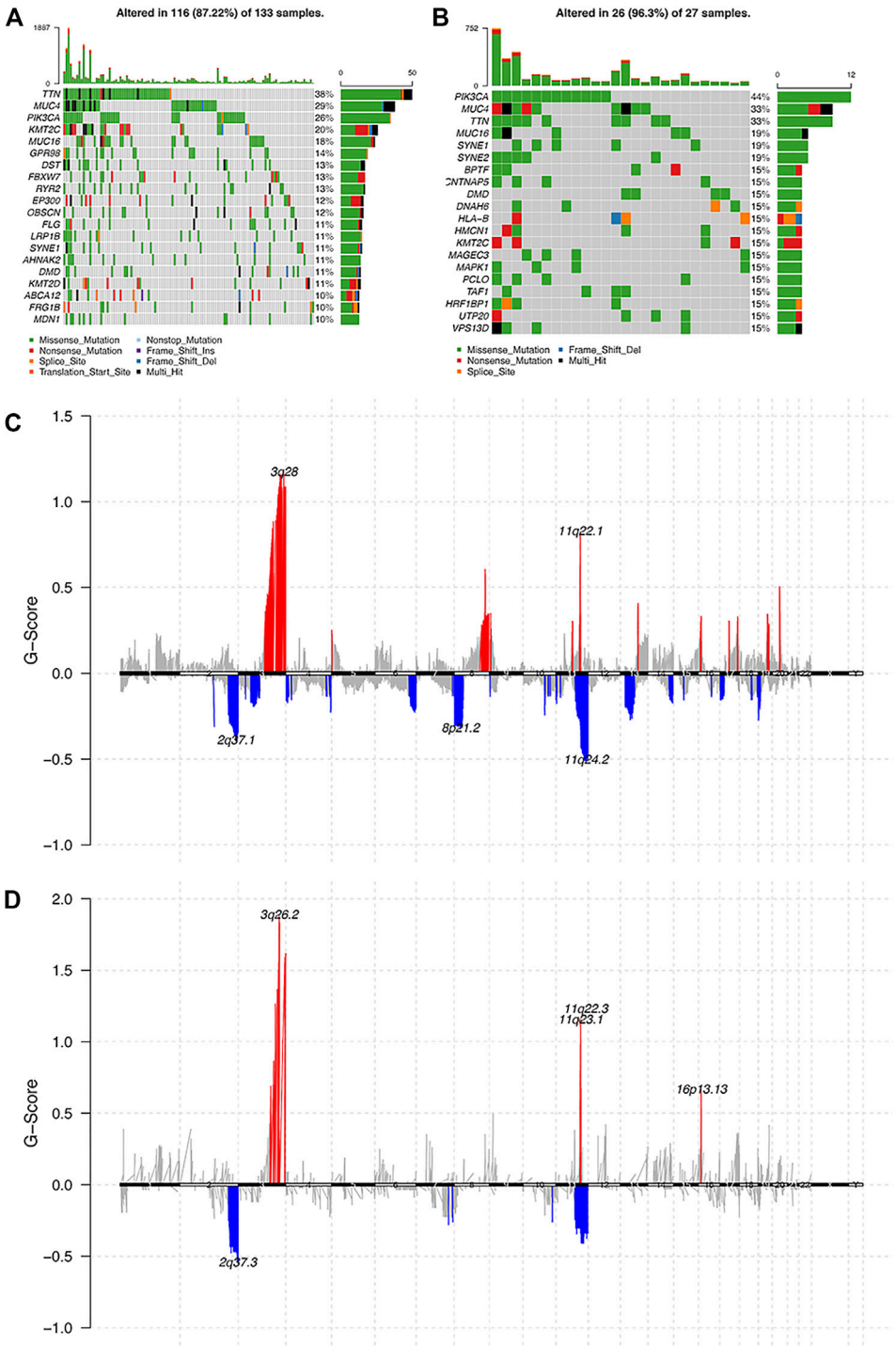


FIGURE 3 | Analysis of the molecular characters between pure HPV-infection and HPV-int samples. **(A)**: Gene mutation in pure HPV-infection samples, **(B)**: Gene mutation in HPV-int samples, **(C)**: Distribution of CNV regions among pure HPV-infection samples, **(D)**: Distribution of CNV regions among HPV-int samples.

low. As shown in **Figure 7A**, the group with HPV-score high had a relatively poorer prognosis compared to those with low HPV-score. The HPV-score-related genes identified are shown in **Figure 7B**. The correlation between HPV status, HPV-related clusters, and HPV-score are was depicted in **Figure 7C**. HPV-score was considered negatively correlated with genes from the FGFR3 pathway and cell cycle regulation and positively correlated with epithelial-mesenchymal transition (EMT) biological functions, which might promote cancer cell proliferation, invasion, and migration (**Figure 7D**).

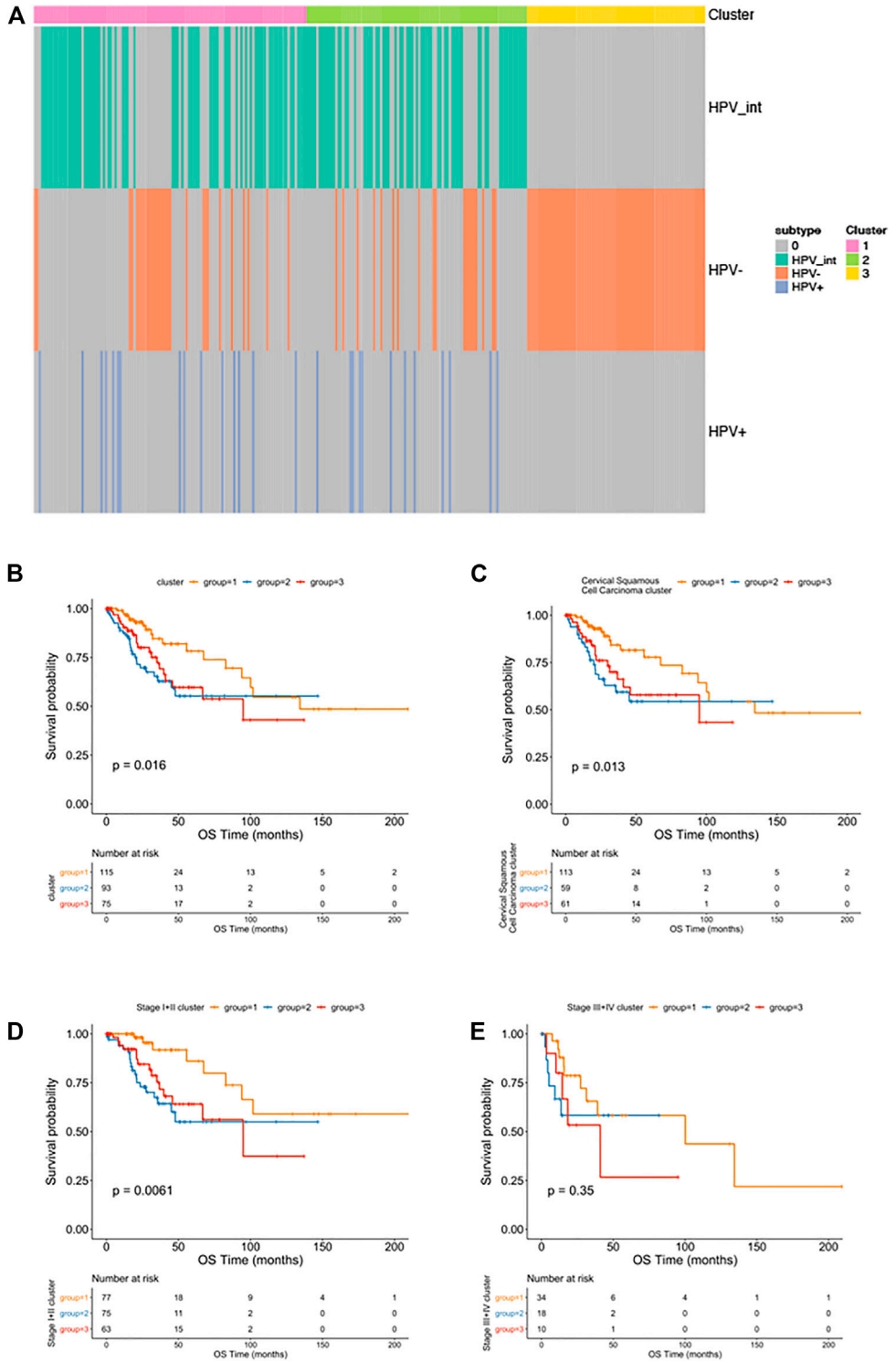


FIGURE 4 | Multi-omics clustering analysis. **(A)**: The cluster-of-clusters analysis separated cervical cancers into three subgroups; **(B–D)**: Kaplan-Meier analysis of overall survival stratified by three subgroups among all cervical cancer patients **(B)**; among squamous cervical cancer patients **(C)**; among patients with stage I-II cervical cancers **(D)** and stage III-IV cervical cancers **(E)**.

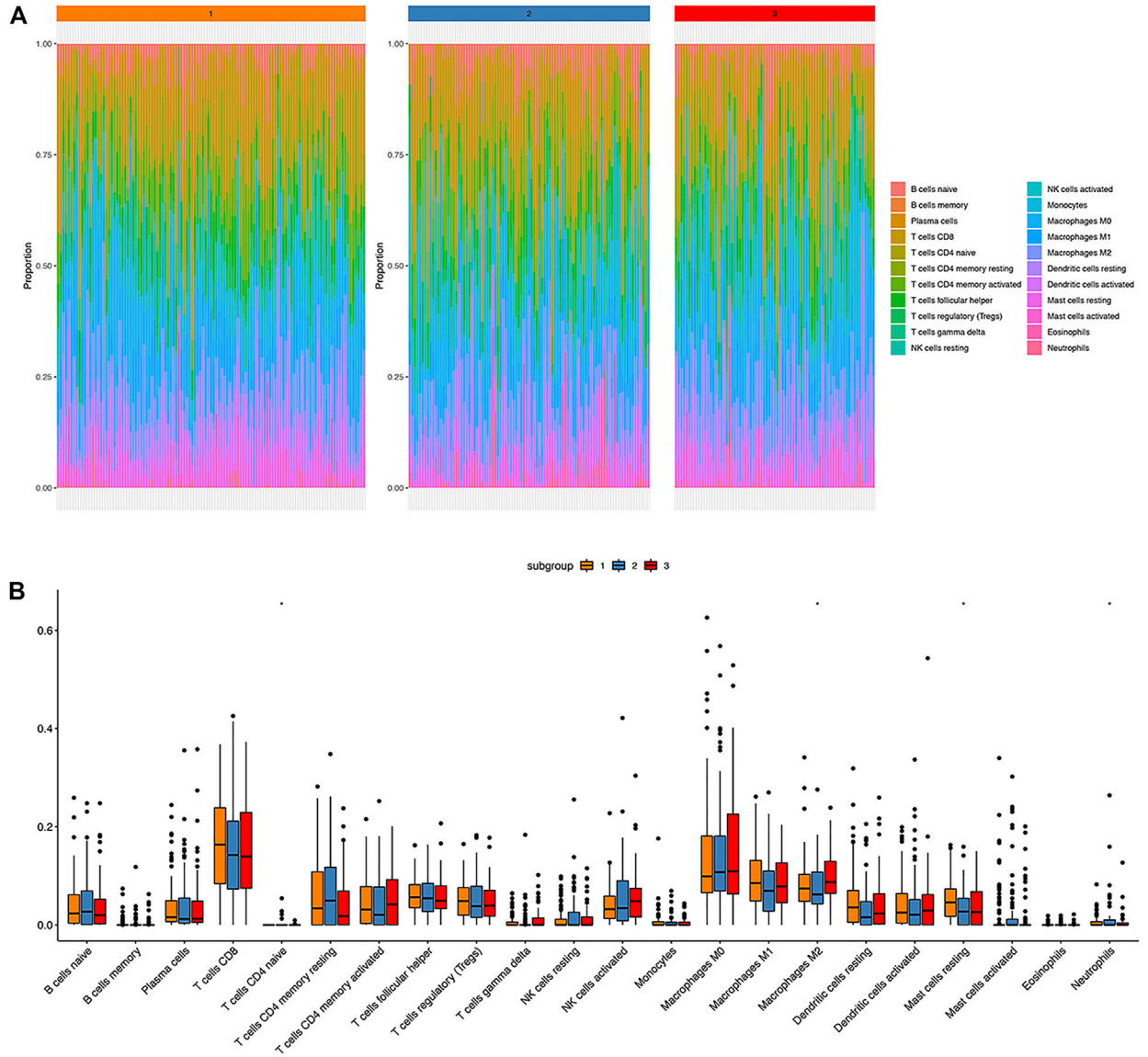


FIGURE 5 | Immune landscape atlas in three HPV clusters. **(A)**: The distribution of 22 types of immune cells in three classified clusters; **(B)**: Percentage distribution of immune cells in three classified clusters; ns: $p > 0.05$, *: $p \leq 0.05$, **: $p \leq 0.01$, ***: $p \leq 0.001$, ****: $p \leq 0.0001$.

Furthermore, we explored the relationship between HPV-scores and histological differentiation and found poor differentiation (grade3-4) showed higher HPV-scores than well differentiation (grade1-2) (**Figure 8A**). The HPV-scores showed no change before and after treatment (**Figure 8B**). Finally, we used the HPV-score to predict the 1,3,5-years survival rate among the TCGA CESC dataset, and the AUC values presented as 0.664, 0.623, 0.588 accordingly (**Figure 9**).

DISCUSSION

Cervical cancer is one of the most common gynecological cancers (Bray et al., 2018). Although it is generally believed that persistent infection of high-risk HPV is the main etiology of cervical cancer (Cohen et al., 2019), a variable proportion of tumors are reported to be negative for hrHPV (Lei et al., 2018). A previous study reported that women with hrHPV-positive cervical tumors had a substantially better prognosis than women with hrHPV-negative tumors (Lei et al., 2018). However, whether the pathogenesis of these two types of cervical cancer is different is unclear. In the current study, we firstly explored the different molecular characteristics between HPV-positive and HPV-negative cervical cancer and revealed that HPV-positive tumors were depleted in glycan biosynthesis and metabolism. Interestingly, in a recent study, Yang et al. explored the compositional and functional alterations on vaginal samples between HPV16-positive and age-matched HPV-negative controls, and

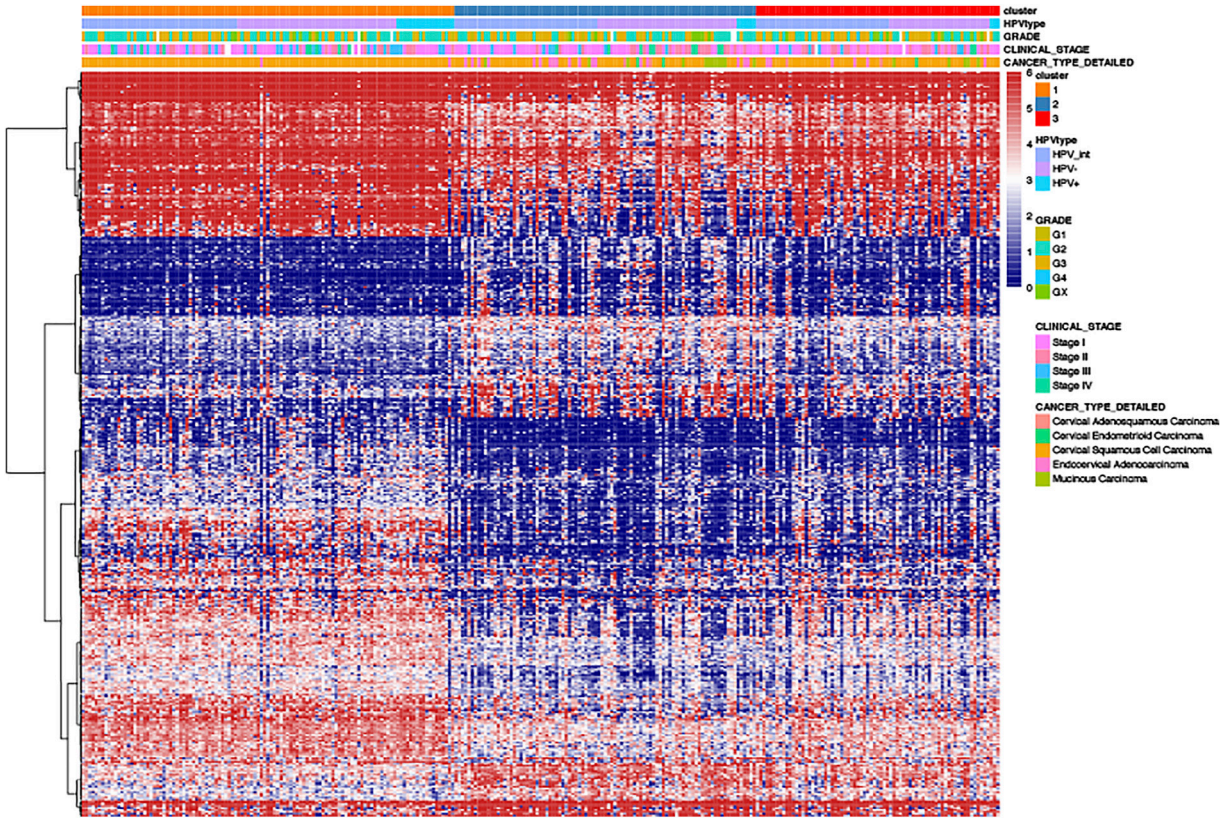


FIGURE 6 | The heat-map of differential gene distribution in cervical cancer stratified by HPV related clusters, HPV status, grades, stages, pathological types.

demonstrated that HPV16-positive candidates were enriched in metabolism and membrane transport, and depleted by glycan biosynthesis and metabolism, and replication and repair, when comparing to the HPV-negative samples (Yang et al., 2020). These results were consistent with our results, suggesting metabolic reprogramming involved in HPV-induced cervical cancer.

It is generally believed that the integration of the HPV genome into the host chromosome is the key genetic step in the pathogenesis of cervical carcinomas (Pett and Coleman, 2007). We then explored the genetic differences between pure HPV-infection and HPV-int tumors and verified that PIK3CA, TTN, MUC4, KMT2C, and SYNE1 were common somatic mutations both in pure HPV-infection and HPV-int women, suggesting that these genes play an important role in HPV-associated cervical cancer. Along similar lines, PIK3CA is reported as the most frequently mutated oncogene in cervical cancer associated with HPV infection, and PIK3CA mutation is the only variable significantly associated with disease recurrence (Kannan et al., 2017; Beaty et al., 2020). In addition, the copy number amplification of pure HPV-infection and HPV-int samples was 3q28 and 3q26.2, and the deletion sites were 2q37.1 and 2q37.3, respectively, which was consistent with our previous report (Xu et al., 2021). Copy number alteration driven dysregulated genes, such as PI3KCA, PI3KCB, DVL3, WWTR1, and ERBB2, are believed to play an important role in regulating immune cell infiltration, which might help to predict the prognosis of cervical cancer (Wen et al., 2019).

We then explored molecular classification based on HPV status. Samples were classified by iCluster clustering, and three subgroups were obtained. Cluster 1 and Cluster 2 are characterized as HPV-positive, while Cluster 3 is characterized by being HPV-negative. Interestingly, when we compared the prognosis among the three subgroups, we detected both Cluster 2 and 3 showed worse overall survival as compared to Cluster 1. Previous studies have shown that women with hrHPV-positive cervical tumors had a substantially better prognosis than women with hrHPV-negative tumors (Lei et al., 2018), which is consistent with our results, the majority of pure HPV-infection, as well as HPV-int patients (Cluster 1), showed a better prognosis than HPV-negative women (Cluster 3). However, there were a proportion of HPV-positive tumors (Cluster 2) that showed a similarly poor prognosis as compared with HPV-negative women (Cluster 3). Therefore, simply based on the fact that HPV negative, pure HPV infection, and HPV integration cannot clearly predict the prognosis of cervical cancer.

To identify the genetic heterogeneity among the three subgroups and determine their relationship with the prognosis of the disease, an HPV-score system was established for all the samples. Patients with high HPV-scores showed a much worse prognosis than those with low HPV scores, suggesting that the

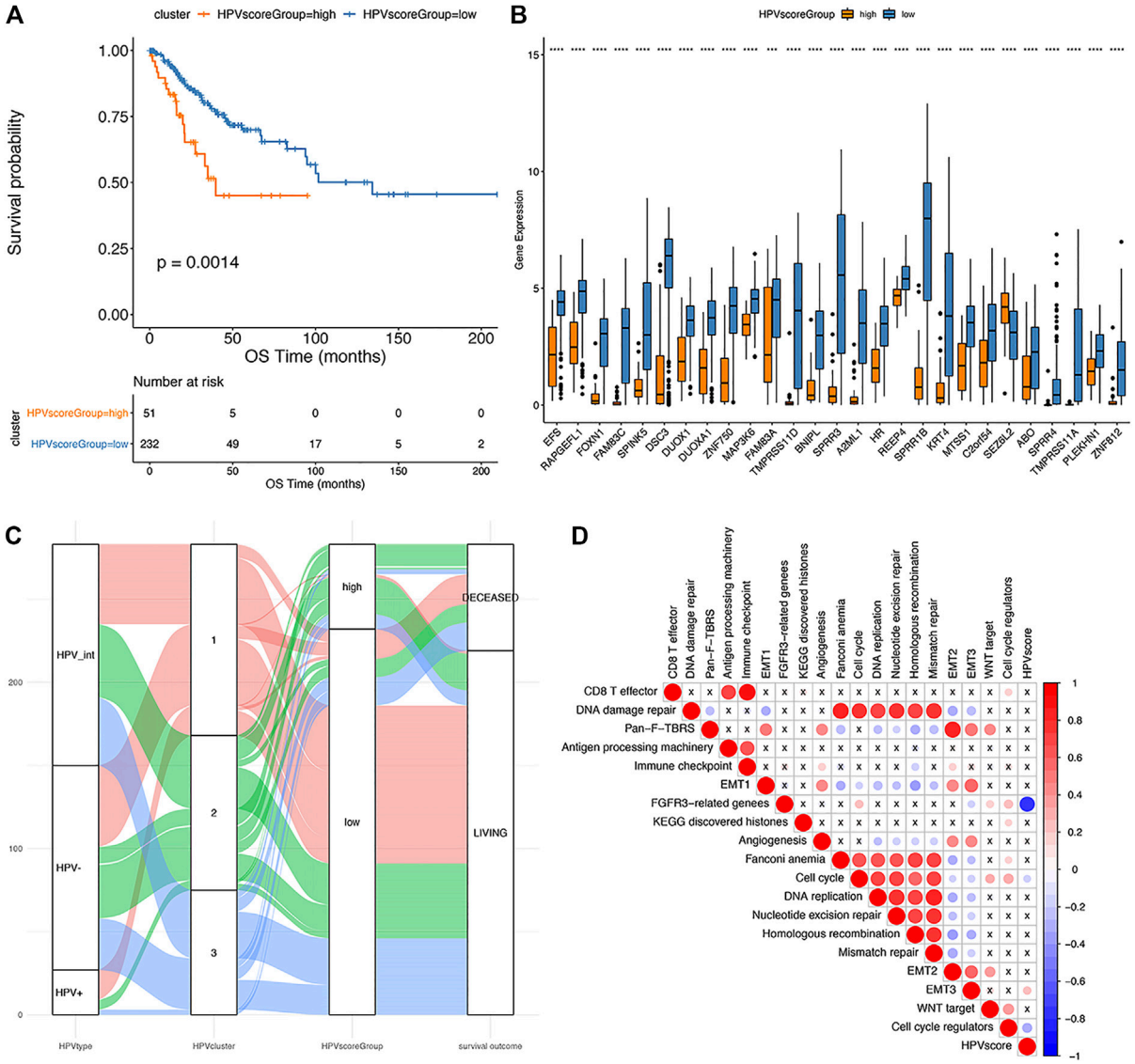


FIGURE 7 | The establishment of HPV score system. **(A)**: Kaplan-Meier analysis of overall survival in cervical cancer patients stratified by HPV score; **(B)**: The expression of HPV score associated genes in different HPV score groups (significance analysis by *t*-test); **(C)**: The alluvial plot showing the correlation among HPV status, iCluster clustering and HPV score; **(D)**: Pearson analysis of HPV score and its associated genes (The negative correlation is marked in blue and positively correlated with red. The X in the graph indicates that the correlation is not significant, and the larger the circle is, the more significant the correlation is).

calculation based on HPV score is a good prognostic predictor. Moreover, a high HPV-score was considered to be correlated with poor cancer cell differentiation, which might suggest that tumors with high HPV scores present more aggressive biological behavior. Additionally, this system was relatively stable, and not subject to treatment changes, as it showed no difference before and after treatment within a chosen cohort.

This HPV-score showed a significantly positive correlation with EMT3 biological functions, while negatively related to FGFR3-pathway genes and cell cycle regulation. Presumably, FGFR3 kinase plays a key role in regulating infected epithelium cells by limiting HPV replication, inhibiting tumorigenesis, and tumor growth (Bersani et al., 2017).

To further explore the immune basis of different HPV statuses, we investigated the correlation between HPV status and infiltrating immune cell subsets. The proportion of immune cells in each sample varied within and between groups. Notably, significant differences existed in the composition of resting mast cells, macrophage M2, Neutrophils, and CD4 naïve T cells among the three groups. A high fraction of activated mast cells was considered as independent factor associated with adverse outcomes in cervical cancer (Wang et al., 2019; Yang et al., 2019). Mast cells may contribute to an immunosuppressive environment that enables the persistence of HPV E7 protein induced pre-cancerous lesions (Bergot et al., 2014). In an esophageal squamous cell carcinoma model, HPV16 infection

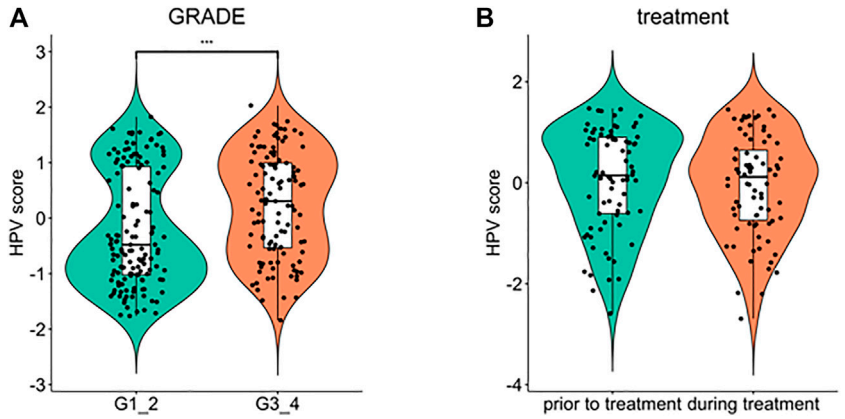


FIGURE 8 | Comparison of HPV score variation according to cancer cell differentiation **(A)** and treatment time schedule **(B)**.

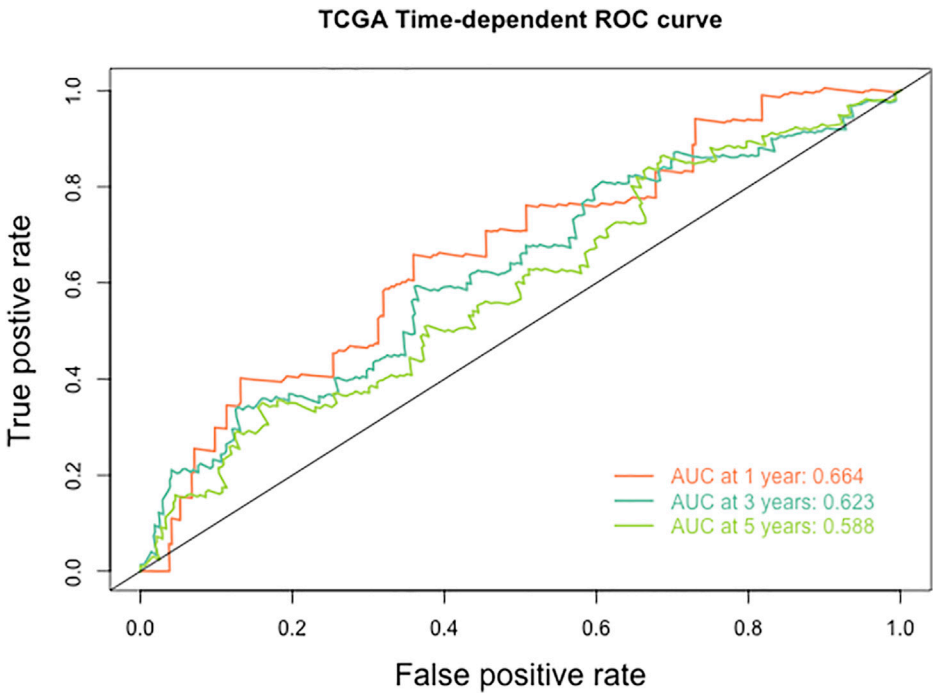


FIGURE 9 | The ROC curve of HPV score for predicting the survival rate of patients with cervical cancer in TCGA database.

promoted an M2 macrophage phenotype, thus, enhanced the invasion and metastasis (Yuan et al., 2021a). Tumor-associated neutrophils (TAN) can promote the formation of the tumor microenvironment. Besides, it is believed that a high proportion of neutrophil infiltration facilitates microenvironment formation for tumor progression and eventually results in a poorer prognosis in solid tumors. $CD4^+$ T cells play a crucial role in eliminating the virus. After activation, naive $CD4^+$ T could enhance cellular or humoral immune response through activating T-helper 1 (Th1) and Th2 subsets, respectively (Yuan et al., 2021b). Collectively, the immune medium constitutes a crucial role in regulating HPV-related cancers.

The main strength of our study is that we established an HPV scoring model that effectively predicts the prognosis of patients with cervical cancer. The main limitations of our study are that we only used the TCGA database. Another limitation is connected to the fact that the majority of the samples were pathological types of squamous cell carcinoma, meaning there was a lack of other pathological types and HPV infection status of the results. The status of HPV is mainly detected before treatment, and the lack of longitudinal data during treatment and follow-up may affect the extrapolation of the HPV-score during treatment and follow-up.

In conclusion, we explored the potential correlation between HPV status and clinical outcome of cervical cancer and then established a

HPV-score that can effectively predict the prognosis of patients with cervical cancer. This scoring system demonstrated good consistency with clinicopathologic characteristics and may provide guidance for HPV-related cervical cancer treatment decision-making in the future.

DATA AVAILABILITY STATEMENT

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

AUTHOR CONTRIBUTIONS

HZ and YY conceived and designed the experiments and revised the manuscript. YZ and YG performed the experiments. QH and YW analyzed the data and wrote the paper. HZ and QH contributed reagents/materials/analysis tools and revised the manuscript. All authors have read and approved the final manuscript.

FUNDING

This work was supported by a grant from National Natural Science Foundation of China (Grant No. 82102815).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.747090/full#supplementary-material

REFERENCES

Beaty, B. T., Moon, D. H., Shen, C. J., Amdur, R. J., Weiss, J., Grilley-Olson, J., et al. (2020). PIK3CA Mutation in HPV-Associated OPSCC Patients Receiving Deintensified Chemoradiation. *J. Natl. Cancer Inst.* 112, 855–858. doi:10.1093/jnci/djz224

Bergot, A.-S., Ford, N., Leggatt, G. R., Wells, J. W., Frazer, I. H., and Grimbaldeston, M. A. (2014). HPV16-E7 Expression in Squamous Epithelium Creates a Local Immune Suppressive Environment via CCL2- and CCL5- Mediated Recruitment of Mast Cells. *Plos Pathog.* 10, e1004466. doi:10.1371/journal.ppat.1004466

Bersani, C., Sivars, L., Haeggblom, L., DiLorenzo, S., Mints, M., Ährlund-Richter, A., et al. (2017). Targeted Sequencing of Tonsillar and Base of Tongue Cancer and Human Papillomavirus Positive Unknown Primary of the Head and Neck Reveals Prognostic Effects of Mutated FGFR3. *Oncotarget* 8, 35339–35350. doi:10.18632/oncotarget.15240

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: A Cancer J. Clinicians* 68, 394–424. doi:10.3322/caac.21492

Cancer Genome Atlas Research Network; Albert Einstein College of Medicine; Analytical Biological Services; Barretos Cancer Hospital; Baylor College of Medicine; Beckman Research Institute of City of Hope, et al. (2017). Integrated Genomic and Molecular Characterization of Cervical Cancer. *Nature* 543, 378–384. doi:10.1038/nature21386

Cancer Genome Atlas Research Network (2017). Integrated Genomic and Molecular Characterization of Cervical Cancer. *Nature* 543 (7645), 378–384. doi:10.1038/nature21386

Cohen, P. A., Jhingran, A., Oaknin, A., and Denny, L. (2019). Cervical Cancer. *Lancet* 393, 169–182. doi:10.1016/s0140-6736(18)32470-x

Cuschieri, K., Brewster, D. H., Graham, C., Nicoll, S., Williams, A. R. W., Murray, G. I., et al. (2014). Influence of HPV Type on Prognosis in Patients Diagnosed with Invasive Cervical Cancer. *Int. J. Cancer* 135, 2721–2726. doi:10.1002/ijc.28902

Hu, Z., and Ma, D. (2018). The Precision Prevention and Therapy of HPV-Related Cervical Cancer: New Concepts and Clinical Implications. *Cancer Med.* 7, 5217–5236. doi:10.1002/cam4.1501

Kannan, A., Hertweck, K. L., Philley, J. V., Wells, R. B., and Dasgupta, S. (2017). Genetic Mutation and Exosome Signature of Human Papilloma Virus Associated Oropharyngeal Cancer. *Sci. Rep.* 7, 46102. doi:10.1038/srep46102

Keck, M. K., Zuo, Z., Khattri, A., Stricker, T. P., Brown, C. D., Imanguli, M., et al. (2015). Integrative Analysis of Head and Neck Cancer Identifies Two Biologically Distinct HPV and Three Non-HPV Subtypes. *Clin. Cancer Res.* 21, 870–881. doi:10.1158/1078-0432.ccr-14-2481

Lai, C.-H., Chang, C.-J., Huang, H.-J., Hsueh, S., Chao, A., Yang, J.-E., et al. (2007). Role of Human Papillomavirus Genotype in Prognosis of Early-Stage Cervical Cancer Undergoing Primary Surgery. *J. Clin. Oncol.* 25, 3628–3634. doi:10.1200/jco.2007.11.2995

Lei, J., Ploner, A., Lagheden, C., Eklund, C., Nordqvist Kleppe, S., Andrae, B., et al. (2018). High-risk Human Papillomavirus Status and Prognosis in Invasive Cervical Cancer: A Nationwide Cohort Study. *Plos Med.* 15, e1002666. doi:10.1371/journal.pmed.1002666

Newman, A. M., Liu, C. L., Green, M. R., Gentles, A. J., Feng, W., Xu, Y., et al. (2015). Robust Enumeration of Cell Subsets from Tissue Expression Profiles. *Nat. Methods* 12, 453–457. doi:10.1038/nmeth.3337

Pett, M., and Coleman, N. (2007). Integration of High-Risk Human Papillomavirus: a Key Event in Cervical Carcinogenesis? *J. Pathol.* 212, 356–367. doi:10.1002/path.2192

Qingqing, B., Jie, Z., Songben, Q., Juan, C., Lei, Z., and Mu, X. (2020). Cervicovaginal Microbiota Dysbiosis Correlates with HPV Persistent Infection. *Microb. Pathog.* 152, 104617. doi:10.1016/j.micpath.2020.104617

Sotiriou, C., Wirapati, P., Loi, S., Harris, A., Fox, S., Smeds, J., et al. (2006). Gene Expression Profiling in Breast Cancer: Understanding the Molecular Basis of Histologic Grade to Improve Prognosis. *J. Natl. Cancer Inst.* 98, 262–272. doi:10.1093/jnci/djj052

Wang, J., Li, Z., Gao, A., Wen, Q., and Sun, Y. (2019). The Prognostic Landscape of Tumor-Infiltrating Immune Cells in Cervical Cancer. *Biomed. Pharmacother.* 120, 109444. doi:10.1016/j.biopha.2019.109444

Wei, W. F., Su, G. D., Wu, L. F., He, L. N., Lu, L., Zhou, J., et al. (2015). Study of Integrated State of HPV-16 Infection in Cervical Cancer and Precancerous Tissues. *Nan Fang Yi Ke Da Xue Xue Bao* 35, 47–50. doi:10.3969/j.issn.1673-4254.2015.01.09

Wen, Y., Zhang, S., Yang, J., and Guo, D. (2019). Identification of Driver Genes Regulating Immune Cell Infiltration in Cervical Cancer by Multiple Omics Integration. *Biomed. Pharmacother.* 120, 109546. doi:10.1016/j.biopha.2019.109546

Xu, Y., Luo, H., Hu, Q., and Zhu, H. (2021). Identification of Potential Driver Genes Based on Multi-Genomic Data in Cervical Cancer. *Front. Genet.* 12, 598304. doi:10.3389/fgene.2021.598304

Yang, S., Wu, Y., Deng, Y., Zhou, L., Yang, P., Zheng, Y., et al. (2019). Identification of a Prognostic Immune Signature for Cervical Cancer to Predict Survival and Response to Immune Checkpoint Inhibitors. *Oncoimmunology* 8 (12), e1659094. doi:10.1080/2162402x.2019.1659094

Yang, Q., Wang, Y., Wei, X., Zhu, J., Wang, X., Xie, X., et al. (2020). The Alterations of Vaginal Microbiome in HPV16 Infection as Identified by Shotgun Metagenomic Sequencing. *Front. Cel. Infect. Microbiol.* 10, 286. doi:10.3389/fcimb.2020.00286

Yuan, X., Liu, K., Li, Y., Zhang, A. Z., Wang, X. L., Jiang, C. H., et al. (2021). HPV16 Infection Promotes an M2 Macrophage Phenotype to Promote the Invasion and Metastasis of Esophageal Squamous Cell Carcinoma. *Clin. Transl Oncol.* 23 (11), 2382–2393. doi:10.1007/s12094-021-02642-5

Yuan, Y., Cai, X., Shen, F., and Ma, F. (2021). HPV post-infection Microenvironment and Cervical Cancer. *Cancer Lett.* 497, 243–254. doi:10.1016/j.canlet.2020.10.034

Zeng, D., Li, M., Zhou, R., Zhang, J., Sun, H., Shi, M., et al. (2019). Tumor Microenvironment Characterization in Gastric Cancer Identifies Prognostic and Immunotherapeutically Relevant Gene Signatures. *Cancer Immunol. Res.* 7, 737–750. doi:10.1158/2326-6066.cir-18-0436

Zhang, B., Wu, Q., Li, B., Wang, D., Wang, L., and Zhou, Y. L. (2020). m6A Regulator-Mediated Methylation Modification Patterns and Tumor Microenvironment Infiltration Characterization in Gastric cancerA Regulator-Mediated Methylation Modification Patterns and Tumor Microenvironment Infiltration Characterization in Gastric Cancer. *Mol. Cancer* 19, 53. doi:10.1186/s12943-020-01170-0

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

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

frontiers in Genetics

ORIGINAL RESEARCH
published: 01 December 2021
doi: 10.3389/fgene.2021.779155

Characterizing the Copy Number Variation of Non-Coding RNAs Reveals Potential Therapeutic Targets and Prognostic Markers of LUSC

*Jinfeng Ning[1], Fengjiao Wang[1], Kaibin Zhu[1], Binxi Li[2], Qing Shu[3] and Wei Liu[4]**

[1]Department of Thoracic Surgery, Harbin Medical University Cancer Hospital, Harbin, China, [2]Department of Management Science and Engineering, Harbin Engineering University, Harbin, China, [3]Department of Medical Imaging, Harbin Medical University Cancer Hospital, Harbin, China, [4]The Fourth Department of Medical Oncology, Harbin Medical University Cancer Hospital, Harbin, China

Lung squamous cell carcinoma (LUSC) has a poor clinical prognosis and a lack of available targeted therapies. Therefore, there is an urgent need to identify novel prognostic markers and therapeutic targets to assist in the diagnosis and treatment of LUSC. With the development of high-throughput sequencing technology, integrated analysis of multi-omics data will provide annotation of pathogenic non-coding variants and the role of non-coding sequence variants in cancers. Here, we integrated RNA-seq profiles and copy number variation (CNV) data to study the effects of non-coding variations on gene regulatory network. Furthermore, the 372 long non-coding RNAs (lncRNA) regulated by CNV were used as candidate genes, which could be used as biomarkers for clinical application. Nine lncRNAs including *LINC00896*, *MCM8-AS1*, *LINC01251*, *LNX1-AS1*, *GPRC5D-AS1*, *CTD-2350J17.1*, *LINC01133*, *LINC01121*, and *AC073130.1* were recognized as prognostic markers for LUSC. By exploring the association of the prognosis-related lncRNAs (pr-lncRNAs) with immune cell infiltration, *GPRC5D-AS1* and *LINC01133* were highlighted as markers of the immunosuppressive microenvironment. Additionally, the cascade response of pr-lncRNA-CNV-mRNA-physiological functions was revealed. Taken together, the identification of prognostic markers and carcinogenic regulatory mechanisms will contribute to the individualized treatment for LUSC and promote the development of precision medicine.

Keywords: copy number variation, non-coding RNAs, biomarkers, clinical prognosis, precision medicine

OPEN ACCESS

Edited by:
Lixin Cheng,
Jinan University, China

Reviewed by:
Hengzi Sun,
Peking Union Medical College Hospital (CAMS), China
Yan Zhang,
Harvard Medical School, United States

****Correspondence:***
Wei Liu
13945147542@163.com

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

Received: *18 September 2021*
Accepted: *01 November 2021*
Published: *01 December 2021*

Citation:
Ning J, Wang F, Zhu K, Li B, Shu Q and Liu W (2021) Characterizing the Copy Number Variation of Non-Coding RNAs Reveals Potential Therapeutic Targets and Prognostic Markers of LUSC. Front. Genet. 12:779155. doi: 10.3389/fgene.2021.779155

BACKGROUND

Lung cancer is the most common cause of cancer-related deaths worldwide (Ferlay et al., 2015), although our understanding of the pathogenesis and treatment options of lung cancer has improved. Lung squamous cell carcinoma (LUSC) is a subtype of non-small cell carcinoma, which accounts for about 40% of lung cancers (Goldstraw et al., 2011). Compared with lung adenocarcinoma (LUAD), another subtype of non-small cell lung cancer (NSCLC), LUSC has a poor prognosis and lacks effective clinical drugs (Hirsch et al., 2017). Immune checkpoint blockade (ICB) therapy is a hot spot in cancer treatment (Yu et al., 2016). However, this treatment only works in a subset of patients for LUSC and improves the prognosis of patients (Yuan et al., 2021). Currently, histopathological images

are the main method for clinical diagnosis of lung cancer (Cui et al., 2020); it is thus necessary to identify novel prognostic markers and therapeutic targets to assist the clinical treatment of LUSC.

Long non-coding RNAs (lncRNAs), which is defined as an ncRNA of at least 200 nucleotides (nt) in length (Wilusz et al., 2009; Ulitsky, 2016), have been found to play a critical role in the regulation of gene expression contributing to physiological function homeostasis, aging, and multiple cancers (Rinn and Chang, 2012; Schmitt and Chang, 2016). With the development of high-throughput technologies and comprehensive databases (Wang et al., 2019a; Wang et al., 2021), the integrated analysis of multiple omics data to reveal the roles of lncRNAs in tumor pathogenesis has become the norm. Genetic variation in the lncRNA region, such as single-nucleotide variation, somatic mutation, and copy number variation (CNV), may affect the expression level of the gene and its target genes, which may contribute to tumor occurrence and development (Wang et al., 2020). For example, lncRNAs with CNVs drive transcriptional perturbed functional pathways (Xu et al., 2020), and single-nucleotide variation in lncRNA regulates cancer-related pathways through ceRNA mechanism (Zhang et al., 2021).

Immune cells are an important part of the tumor microenvironment (TME) and have been proven to play an important role in tumor proliferation and metastasis. For example, increased abundance of specific T cell subtypes in cancer tissues is associated with better patient prognosis (Jiang et al., 2019; van der Leun et al., 2020). Macrophage polarization plays a critical role in subverting adaptive immunity and promoting tumor progression (Mantovani et al., 2002). Exploring the relationship between dysregulated lncRNA and immune cells is an important direction for deciphering the carcinogenic mechanism of lncRNA.

Here, we collected RNA-seq profiles and CNV data of 482 tumor samples for LUSC from The Cancer Genome Atlas (TCGA) (Tomczak et al., 2015) database to identify lncRNAs whose expression is driven by CNV. Then, Cox and least absolute shrinkage and selection operator (LASSO) regression analyses were performed. Prognosis-related lncRNAs (pr-lncRNAs) driven by CNV were recognized as prognostic markers of LUSC. Moreover, we comprehensively characterized the function and regulatory mechanism of pr-lncRNA by immuno-infiltration and functional enrichment analysis.

METHODS

Data Collection and Pre-Processing

The RNA-seq profiles (482 tumor samples), CNV data (570 tumor samples), and clinical information of LUSC collected by TCGA database were downloaded from UCSC Xena browser (https://xenabrowser.net/). The genome-wide annotation data of GRCH38 V37 including location of lncRNA was collected from the GENCODE (Frankish et al., 2021) database (https://www.gencodegenes.org/). The independent non-small cell lung cancer datasets including GSE37745 and GSE50081 were obtained from publicly available Gene Expression Omnibus (GEO, available at https://www.ncbi.nlm.nih.gov/geo/) (Okayama et al., 2012), which were used for verification of prognostic markers. The data of lncRNA-immune cell infiltration correlation was collected from the ImmLnc (Li et al., 2020) database (http://bio-bigdata.hrbmu.edu.cn/ImmLnc/). Additionally, the signature matrix of 22 types of immune cell was collected from previous studies (Newman et al., 2015) for the analysis of immune cell infiltration. For CNV data of LUSC, the software of GISTIC was used to analyze the CNV of the entire lncRNA genome, taking the number of copies greater than 1 as the threshold of copy amplification and less than −1 as the threshold of copy loss. The R package maftools (v2.4.12) (Mayakonda et al., 2018) was used to observe the distribution of the *G*-score of copy amplification and loss in the genome.

Identification of LncRNA Driven by CNV

Based on the copy number profiles for lncRNA, lncRNAs that have no overlapped segments in more than 50% of patients were deleted. We grouped patients according to the degree of copy number variation in a particular lncRNA region, and patients with copy numbers greater than 1 or less than −1 were assigned to the group of patients with CNVs, while other patients were assigned to the control group. Then, the Mann–Whitney *U* test was used to test whether the lncRNA expression was differentially expressed in the group of patients with CNVs (H0: there was no difference between the group of patients with CNVs and the control group). We also calculated the fold change (FC) of each lncRNA between the group of patients with CNVs and the control group, i.e., $FC = m_{cnv}/m_c$, where m_{cnv} and m_c represent the mean expression of an lncRNA in the group of patients with CNVs and the control group, respectively. The lncRNAs with $p < 0.01$ and $|FC| > 0$ were considered to be lncRNAs driven by CNV.

Identification of Prognosis-Related LncRNA Signature

Based on the lncRNAs driven by CNV, we generated the lncRNA expression matrix. The univariate Cox regression using R package survival (Guo et al., 2019) was employed to screen for lncRNAs with expression level significantly associated with patient's overall survival (OS) of LUSC patients (the cutoff of *p*-value was 0.05). Furthermore, the LASSO regression using R package glmnet (v4.0-2) (Alhamzawi and Ali, 2018) was used to further screen for pr-lncRNA driven by CNV (pr-lncRNA-CNV) based on a more rigorous algorithm. Next, we randomly selected 70% of all tumor samples as the training set and the remaining as the test set. The features selected by the LASSO regression were used to fit for a multivariate Cox risk regression model based on training sets. The reliability of the Cox risk regression model was validated by the receiver operating characteristic (ROC) curve, and the area under curve (AUC) also was calculated. Finally, we defined the lncRNAs with a *p*-value <0.05 as the pr-lncRNA signatures, which significantly contributed to the LUSC patient survival outcomes. Besides, the nomogram method was used to build a more intuitive prediction model, and the calibration curve was used to evaluate the predictive ability of nomograph on the patient's 1- and 3-year survival risk.

Construction of Risk Scoring Model

We used the linear combination of expression values weighted by the coefficient from the multivariate Cox regression analysis to calculate the risk score for each patient:

$$\text{Riskscore}(i) = \sum_{k=1}^{n} \beta_k e_{ki},$$

where n denotes the number of the pr-lncRNAs ($n = 9$), β was the coefficient of multivariate Cox regression analysis, and e_{ki} was the expression level of the kth pr-lncRNA expression of patient i. The median risk score was used as the cut-off to divide patients into high- and low-risk groups. Then, the samples of training set and test set were respectively divided into high-risk and low-risk categories. Additionally, Kaplan–Meier survival curve (Ranstam and Cook, 2017) was used to prove the difference of OS between the high-risk group and low-risk categories, and the bilateral logarithmic rank test (Guyot et al., 2012) was used to calculate the statistical significance. The whole pipeline above was also performed in the independent dataset [GSE37745 (Botling et al., 2013) and GSE50081 (Der et al., 2014), **Supplementary Table S1**] to confirm the robustness and stability of pr-lncRNAs.

Tumor Immune Microenvironment Analysis

First, the online tool CIBERSORTx (https://cibersortx.stanford.edu/), which is a method to characterize the cell composition of complex tissues from the gene expression profile based on signature matrix, was used to calculate the fraction of immune cell infiltration abundance. Then, the parameter perm, which is the number of permutations when calculating the p-value, was set to the max 1,000, and quantile normalization was disabled. The function ggboxplot from the ggpubr package visualized the difference in the abundance of 22 types of leukocytes between the high- and low-risk groups. Additionally, we identified immune cells associated with patient prognosis by linking their abundance to patient's OS. Besides, the lncRNA–immune cell type correlation was performed by the ImmLnc.

Construction of CNV-LncRNA-PCG Network

We first generated the expression matrix of protein-coding genes (PCGs) to calculate PCGs related to the expression of pr-lncRNA driven by CNV using Pearson's algorithm (Bishara and Hittner, 2012). Then, we have defined that gene pairs with a p-value <0.01 and $R > 0.3$ correlation coefficient have significant correlation in expression and are co-expressed genes with each other. For PCGs related to each pr-lncRNA, the Mann–Whitney U test was used to test whether the PCG expression was differentially expressed between the group of patients with CNVs and the control group. Finally, PCGs whose expressions were both differently expressed between the two groups and correlated with pr-lncRNA's expression were collected as signature PCGs. The CNV-lncRNA-PCG network was visualized using Cytoscape (v3.7.0) (Shannon et al., 2003).

Functional Enrichment Analysis

PCGs driven by the CNV of each pr-lncRNA were used to annotate the biological functions of pr-lncRNAs. The functional enrichment analysis was performed by the online tool Metascape (Zhou et al., 2019) (http://metascape.org/).

Statistical Analysis

All analyses were conducted using R (v3.6.3) software. The gene sets enrichment analysis was performed using the Fisher's exact test. Log-rank test was used to compare the difference of survival time between two groups.

RESULTS

Copy Number Amplitude Perturbs the Expression of LncRNA

Aneuploidy such as CNV is a hallmark of most solid tumors (including LUSC) (Ried et al., 2019). To explore the role of CNV on lncRNA in the carcinogenic mechanism of LUSC, we first analyzed the global characteristics of CNV. Maftool was used to visualize the copy number amplitude calculated by GISTIC on the whole genome. We found that there is an obvious copy number amplification on q26.33 of chromosome 3, p11.23/q24.21 of chromosome 8, and q.13.3 of chromosome 11 and copy number deletion on q37.1 of chromosome 2, p25.2 of chromosome 3, p23.2 of chromosome 8, and p21 of chromosome 9 (**Figure 1A**). Next, the copy number amplitude data was mapped to lncRNA to build copy number profiles, and lncRNAs with significant copy number amplification and deletion are defined as candidate lncRNAs (647 lncRNAs) that are used to link lncRNA expression profiles. We found the characteristics of the copy number amplitude of lncRNA in LUSC, that is, each lncRNA with CNV has global amplification or global deletion in patients (**Figure 1B**). Therefore, tumor patients were divided into a group with CNV in the lncRNA and a control group based on lncRNA copy number profiles. Moreover, lncRNAs tended to undergo overall copy number amplification in LUSC (**Figure 1B**), which is contrary to a previous study suggesting that copy number deletion pattern of the lncRNAs was widely observed in various cancer types (Xu et al., 2020). Furthermore, we identified 372 lncRNAs whose expression levels (FPKM) are driven by CNV (**Figure 1C**). Similar with the copy number amplitude results, most lncRNA expressions were up-regulated in the group with CNV in the lncRNA. Among them, the significantly up-regulated lncRNA SOX2-OT driven by CNV has been shown to regulate the proliferation and metastasis of multiple cancers through the ceRNA mechanism (Zhang and Li, 2019; Herrera-Solorio et al., 2021). By annotating the types of lncRNAs driven by CNV, we found that they were mainly long intergenic non-coding RNA (lincRNA) and antisense classes (**Figure 1D**). Taken together, the CNV feature and CNV-driven lncRNAs for LUSC were recognized.

The Nine Pr-LncRNAs for LUSC

LncRNA is an emerging biomarker for cancer development and patient's prognosis (Bolha et al., 2017). To identify CNV-driven

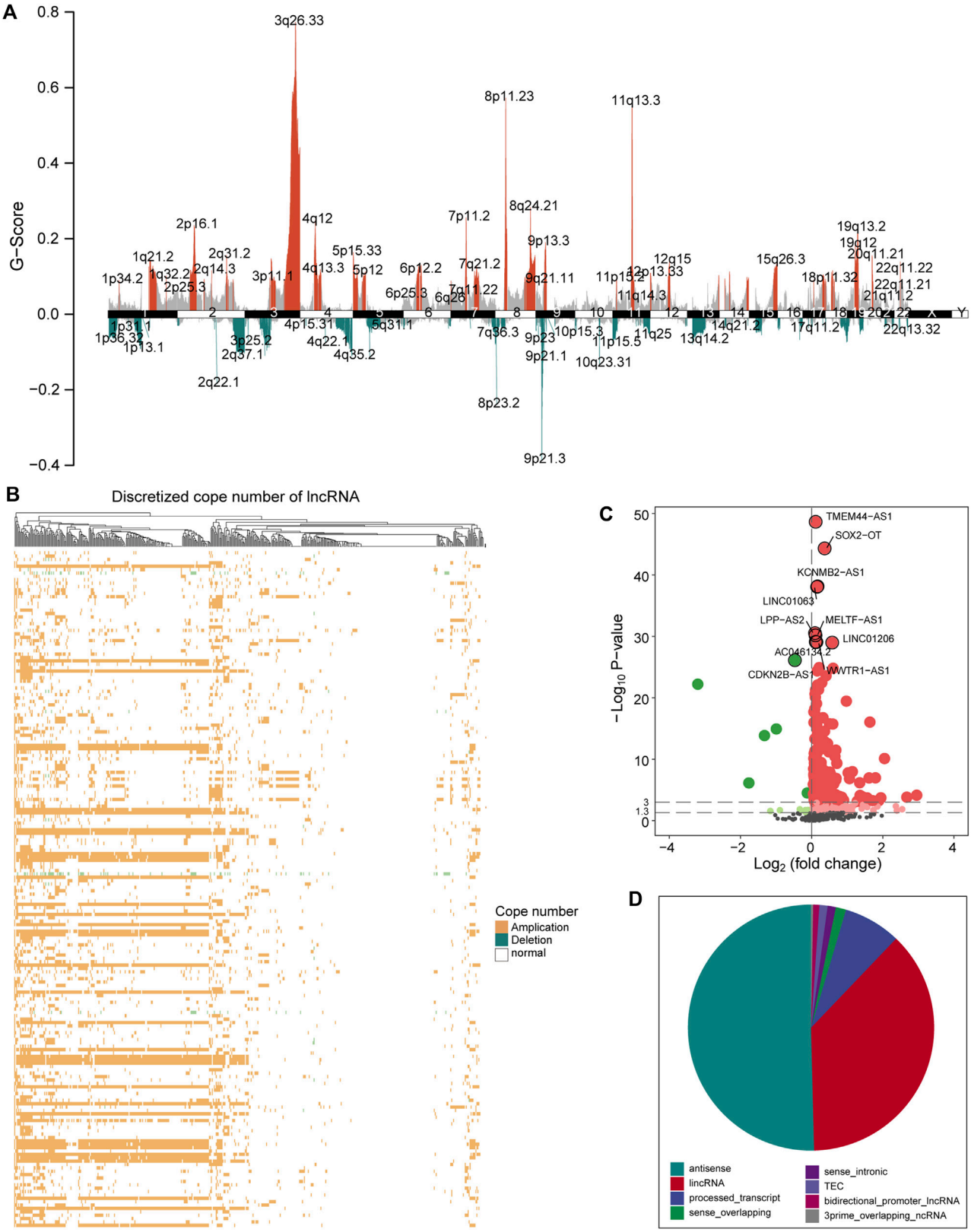


FIGURE 1 | A global view of CNV for LUSC. **(A)** The graph shows the distribution of the copy number amplitude on the genome for LUSC. The ordinate represents *G*-score, and the abscissa is the position of 22 homologous chromosomes and two sex chromosomes. **(B)** The amplification and deletion of lncRNA copy number in

(*Continued*)

FIGURE 1 | tumor samples are displayed by a heat map. Yellow indicates copy number amplification, blue indicates copy number deletion, and white indicates no CNV has occurred. **(C)** The expression difference of the candidate lncRNA between the group with CNV in the lncRNA and the control group is shown by a volcano graph. Red dots indicate up-regulation, and green dots indicate down-regulation. **(D)** The pie chart shows the proportion of various types of CNV-driven lncRNAs. Different colors represent specific lncRNA types. CNV, copy number variation; LUSC, lung squamous cell carcinoma; lncRNA, long non-coding RNA.

lncRNAs related to prognosis, we developed a computational model by combining LASSO regression and Cox regression models to identify pr-lncRNA. We performed the univariate Cox and LASSO algorithm to identify 16 lncRNAs that were significantly related to the patient's OS. After the construction of a multivariate Cox regression model, nine lncRNAs including *LINC00896*, *MCM8-AS1*, *LINC01251*, *LNX1-AS1*, *GPRC5D-AS1*, *CTD-2350J17.1*, *LINC01133*, *LINC01121*, and *AC073130.1* were finally identified as prognostic markers (**Figure 2A**, **Supplementary Figure S1**). For the nine pr-lncRNAs, *LNX1-AS1*, *AC073130.1*, *MCM8-AS1*, and *LINC01251* were treated as the risk factors for patient's survival of LUSC, while *LINC00896*, *GPRC5D-AS1*, *CTD-2350J17.1*, *LINC01133*, and *LINC01121* were treated as the protective factors. Furthermore, based on the expression of nine prognostic markers, the nomogram algorithm was used to construct 1- and 3-year survival probability prediction models (**Figure 2B**). Then, the calibration curve was used to evaluate the predictive performance of the nomogram, and the result showed that the nomogram algorithm has a good performance in predicting the survival risk for the patients of LUSC (**Figure 2C**). We divide the 1–3-year period into five time points, and the ROC curve was used to determine the best prediction time point for the risk prediction model. We found that the risk prediction result reached the maximum AUC value of 0.73 in 1,095 days (**Figure 2D**). All these suggest that the nine pr-lncRNAs can be used as prognostic markers to predict the survival risk for patients of LUSC.

Risk Score Model Supports the Diagnosis of Patient Prognosis

To accurately quantify the survival risk of patients, we constructed the risk score model based on the regression coefficients calculated by multivariate Cox regression model and patient's survival (Guo et al., 2020). The formula of the risk scoring model was as follows: Risk score = (−0.11 × *LINC00896* + 0.05 × *MCM8-AS1* + 0.04 × *LINC01251* + 0.07 × *LNX1-AS1* − 0.20 × *GPRC5D-AS1* − 0.08 × *CTD-2350J17.1* − 0.09 × *LINC01133* − 0.06 × *LINC01121* + 0.07 × *AC073130.1*). Furthermore, the samples of training and test sets were respectively divided into the high- and low-risk groups based on the median risk score. We found that there was an obvious expression difference of pr-lncRNAs between the high- and low-risk categories (**Figure 3A**), and the samples with high-risk score have poor prognosis (**Figure 3B**). Moreover, the high-risk samples in test sets also exhibited an association with poorer OS of LUSC (**Figure 3C,D**). All these suggest that the risk score can quantify the prognosis of patients and provide convenience for clinical diagnosis. Using the online analysis tool of GEPIA (Tang et al., 2019), we found that the high expression of single genes *GPRC5D-AS1* and *LINC01133* is associated with better patient prognosis (**Figure 3E,F**), which was consistent with the result in this study, suggesting *GPRC5D-AS1* and *LINC01133* as protective factors for patient's OS. Besides, *LNX1_AS1* has been reported as the poor prognosis marker for non-small cell lung cancer (Wang et al., 2019b) in previous studies. The evidence further emphasized that pr-lncRNAs may play an important role in the progression of LUSC.

Evaluation of the Robustness for Nine Pr-LncRNA Signatures

To further confirm that the pr-lncRNA signature is a robust biomarker in LUSC, we collected the expression profiles and clinical information of two sets (GSE37745 and GSE50081) including LUSC samples from public databases for prognostic marker testing. Furthermore, we divided samples from GSE37745 series into high-risk group (n = 33 LUSC samples) and low-risk group (n = 33 LUSC samples) based on median risk score. We found that only six pr-lncRNAs were detected in the microarray matrix, and the risk score distribution, survival status, and lncRNA expression of all patients were consistent with those observed in TCGA cohort (**Figure 4A**). There is a significant difference in the OS between the two groups, and the risk score was also identified as a poor prognosis marker (p = 0.018) (**Figure 4B**). In the set of GSE50081 series, 45 patients of LUSC were divided into high-risk group (n = 22) and low-risk group (n = 23). The results of log-rank test also showed that the risk score was significantly correlated with OS (**Figure 4C,D**). All these further supported that the pr-lncRNA signature was a robust prognosis indicator in LUSC.

Immune Cell Components Regulated by Pr-LncRNAs Support Tumor Progression

To investigate the effect of the nine pr-lncRNAs in the tumor immune microenvironment, we first identified the abundance of immune cell infiltration for TCGA cohort using the CIBERSORTx tool. The LUSC samples of TCGA cohort were divided into two groups based on the risk score of each sample, and the statistical difference in the infiltration abundance of 22 types of leukocytes between the two groups was calculated. We found that the infiltration abundance of multiple immune cells is significantly different in the high- and low-risk groups (**Figure 5A**). For example, the infiltration abundance of memory $CD4^+$ T cells was up-regulated in the high-risk group, the infiltration abundance of macrophage M0 was up-regulated in the high-risk group, and the infiltration abundance of resting dendritic cells was down-regulated in the high-risk group. By combining with the patient's survival information, we found that the infiltration abundance of the three cell types [resting

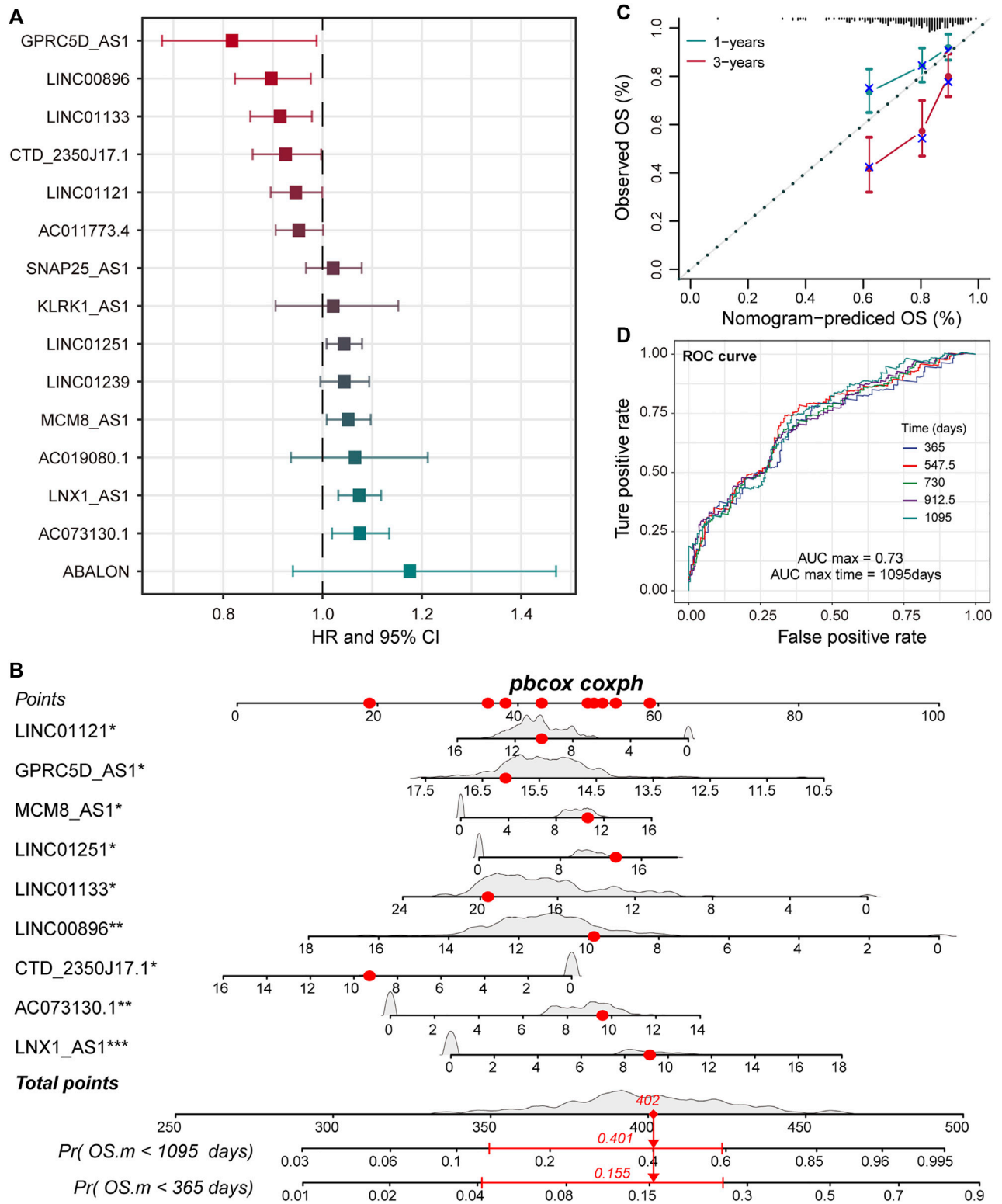


FIGURE 2 | Identification of prognostic markers for LUSC. **(A)** The error bar graphs are used to show the correlation between the expression level of lncRNA and the patient's overall survival (OS). The abscissa represents the hazard ratio (HR) value with the 95% confidence interval. **(B)** Nomogram for survival risk prediction of 1 and 3 years. The nine prognostic markers are marked by a star. Additionally, $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$. **(C)** The calibration curve of the nomogram. **(D)** The predictive performance of the survival predictive model at five time points from 1 to 3 years. The different colored curves represent specific time points.

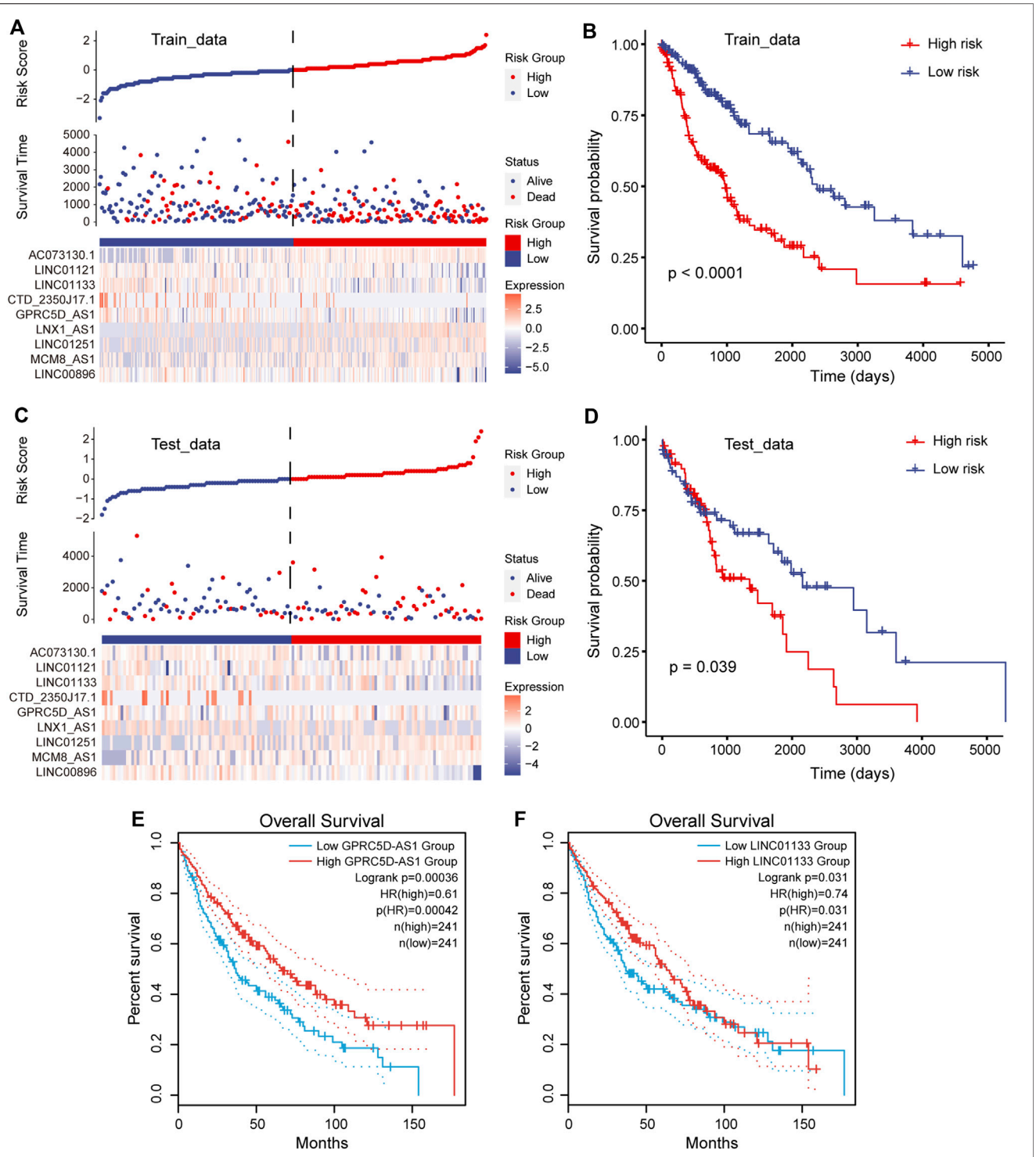


FIGURE 3 | Construction of risk scoring model. **(A)** Risk scores, survival status, and the expression of nine prognosis-related lncRNAs (pr-lncRNAs) for the training set. **(B)** The Kaplan–Meier curves for the two groups in the training set. Log-rank test is used to calculate statistical significance. The unit of time is days. **(C)** Same as in **(A)** but for the testing set. **(D)** Same as in **(B)** but for the two groups in the testing set. **(E–F)** The Kaplan–Meier curves for survival in groups with high and low expression (FPKM) of lncRNA GPRC5D-AS1 and LINC01133.

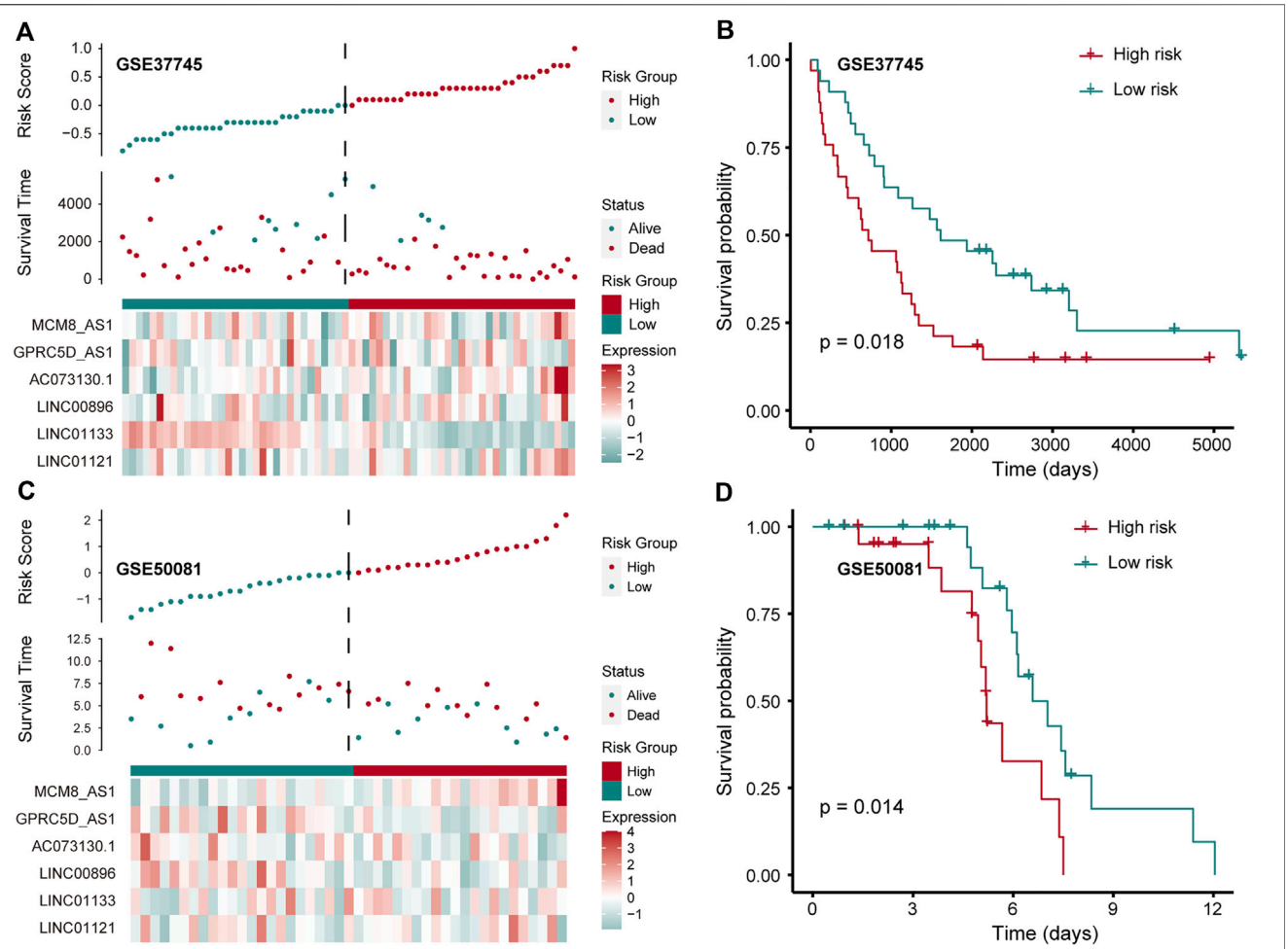


FIGURE 4 | Verification of prognostic markers in public data. **(A)** The expression pattern of lncRNA and survival status and risk scores of LUSC patients for GSE37745 series. **(B)** Kaplan–Meier curve of two groups (high-/low-risk score) for GSE37745 series. Log-rank test is used to calculate statistical significance. The unit of time is days. **(C)** The same as in **(A)** but for GSE50081 series. **(D)** Kaplan–Meier curve of two groups (high-/low-risk score) for GSE50081 series. The unit of time is days.

dendritic cells, resting natural killer (NK) cells, and macrophage M0] were significantly related to the patient's prognosis (**Figure 5B–D**). High infiltration abundance of resting dendritic cells was associated with better patient prognosis (**Figure 5B**), which is consistent with the phenomenon of low expression of resting dendritic cells in high-risk samples. We found that the expression abundance of resting NK cells can be used as a poor prognostic marker (**Figure 5C**), which may be due to the low solubility of resting NK cells to target cells (Bryceson et al., 2006). Moreover, the high fraction of macrophage M0 was significantly associated with poor patient prognosis (**Figure 5D**), which is consistent with a previous study suggesting that macrophage polarization plays a key role in promoting tumor progression (Mantovani et al., 2002). These suggest that the infiltration of these immune cell types could support tumor progression. Furthermore, we systemically analyzed the correlation between pr-lncRNA expression and immune cell infiltration abundance. For the pr-lncRNAs, *AC073130.1*, *GPRC5D-AS1*, and *LINC01133* exhibited significant associations with multiple immune cells, i.e., $CD4^+$ T cell, $CD8^+$ T cell, dendritic cell, macrophage, and neutrophil, suggesting that the three pr-lncRNAs may be immune-related lncRNAs for LUSC (**Figure 5E**). We also found that the pr-lncRNA *GPRC5D-AS1* showed negative correlations with the expression levels of major histocompatibility complex I (MHC I) and MHC II (**Figure 5F** and **Supplementary Figure S2A**) that assist in tumor cell recognition and antigen presentation (Lauss et al., 2017), indicating that a high expression of *GPRC5D-AS1* may reduce tumor immunogenicity.

The Cascade of Pr-LncRNAs Regulates the Carcinogenic Mechanisms for LUSC

To better understand the biological functions and regulatory mechanisms of pr-lncRNA-CNV, we explored the pr-lncRNA-mRNA regulatory mode. In determining the regulatory

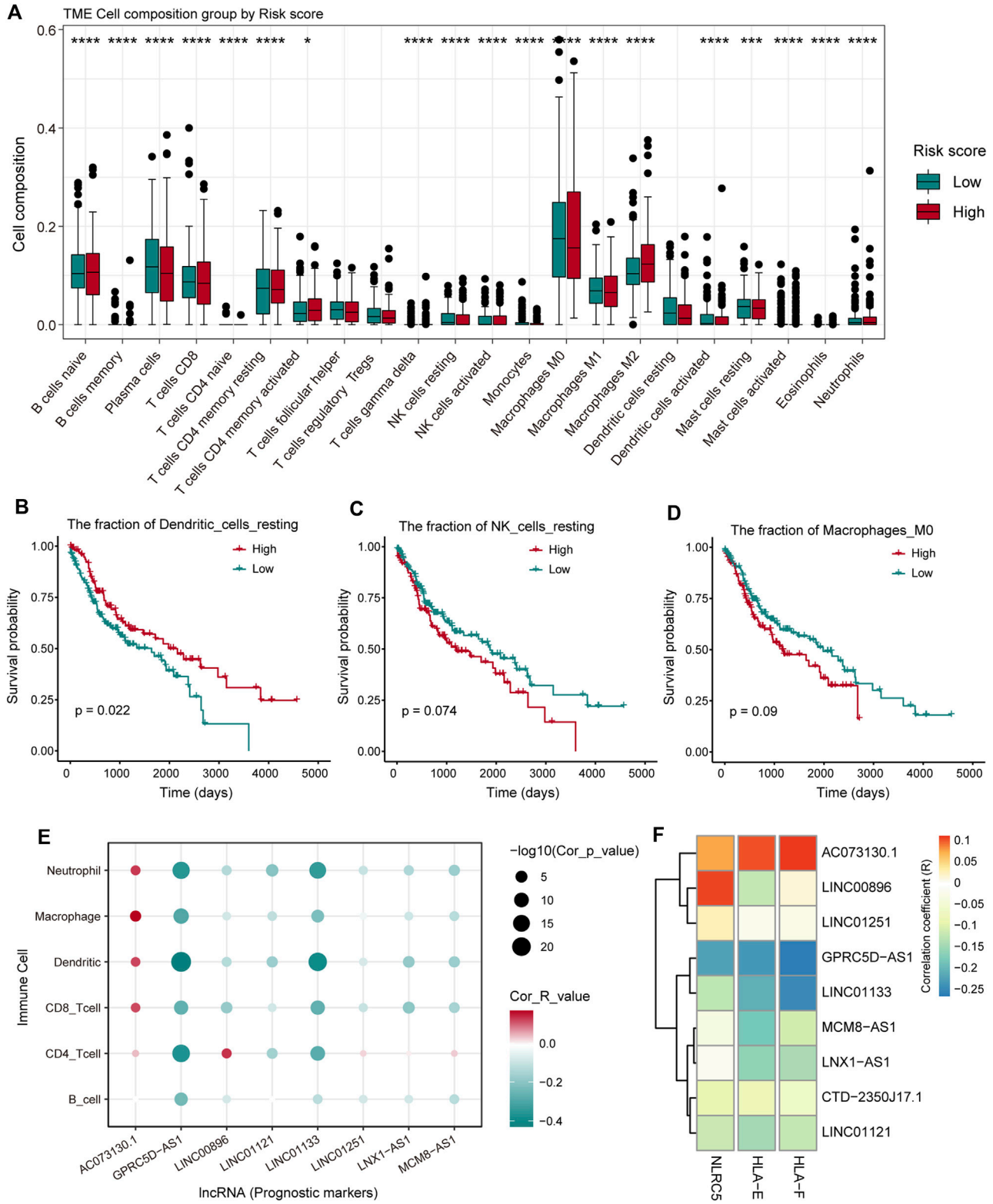


FIGURE 5 | The analysis of the relationship between pr-lncRNAs and tumor immune activity. **(A)** The infiltration abundance of 22 types of leukocytes between the high- and low-risk groups is displayed by a boxplot. The *p*-values were calculated by the Mann–Whitney *U* test. **(B–D)** Kaplan–Meier curve of the two groups (high-/low-infiltration abundance) for resting dendritic cells, resting natural killer (NK) cells, and macrophage M0. Log-rank test is used to calculate statistical significance. **(E)** The bubble chart shows the correlation between the infiltration abundance of the six immune cells and the expression of pr-lncRNA. **(F)** The relationship between the expression of genes encoding major histocompatibility complex I (MHC I) molecules and that of pr-lncRNAs is displayed by a heat map. The stronger the correlation, the darker the color.

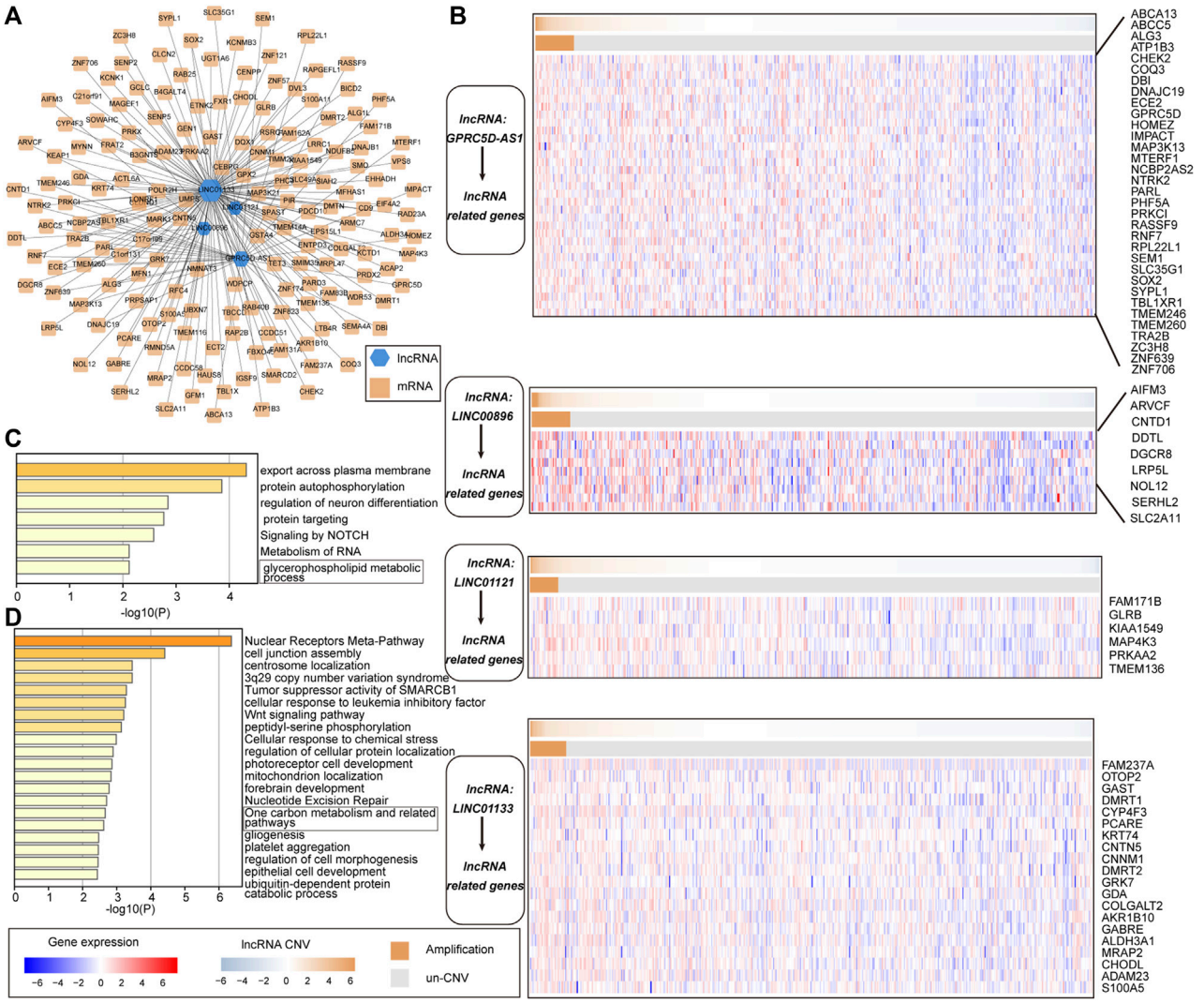


FIGURE 6 | The cascade response of pr-lncRNA-CNV-mRNA carcinogenic mechanism. **(A)** pr-lncRNA-mRNA regulatory network showing the dysregulation of functional genes driven by lncRNA-CNV. The green hexagon represents lncRNA, and the yellow square represents mRNA. **(B)** Four (lncRNA-CNV) gene cascade responses. The heat map combination includes the copy number and the expression of lncRNA-related genes for the four pr-lncRNAs. The first row of each combination represents the copy number amplitude of lncRNA, and the second row represents the copy number amplification and deletion of lncRNA. **(C–D)** The enrichment results for genes related to lncRNA GPRC5D-AS1 and LINC01133 are displayed by bar graphs, colored by *p*-values.

relationship of lncRNA-mRNA, we required the co-expression of lncRNA and mRNA, and mRNA was expressed significantly different between the CNV group and the control group as determined by the lncRNA. We identified 183 lncRNA-mRNA regulatory axes containing four pr-lncRNAs (*GPRC5D-AS1*, *LINC00896*, *LINC01121*, and *LINC01133*) and 167 mRNAs to construct the pr-lncRNA-mRNA network (**Figure 6A**). Furthermore, a series of lncRNA-CNV gene cascade reactions were proposed, and the physiological functions of its regulation were analyzed. The lncRNA *GPRC5D-AS1* with CNVs driving the dysregulation of downstream cancer genes such as *PRKCI*, *SEM1*, *TBL1XR1*, *IMPACT*, and *RPL22L1* further disturbed the activities of NOTCH signaling pathway that is usually inactivated in LUSC (Katoh and Katoh, 2020) (**Figure 6B,C**). Similarly, the lncRNA *LINC01133* may regulate the activity of tumor suppressor pathways by dysregulating *ACTL6A*, *SMARCD2*, *SMO*, *PHC3*, and *CENPP* (**Figure 6B,D**). It is worth noting that *LINC01133* has been shown to be closely related to the invasion and malignant proliferation of human non-small cell lung cancer (Zhang et al., 2020; Geng et al., 2021). The lncRNA *LINC00896* is associated with copy number variation syndrome by regulating the corresponding target gene (**Supplementary Figure S2B**). Taken together, all these results provide novel insights for understanding the function of lncRNA-CNV and the pathogenesis based on these lncRNA-CNV.

DISCUSSION

In this study, we integrated multiple omics data to identify the prognostic-related lncRNAs driven by CNV and define the cascade response of lncRNA-CNV-mRNA carcinogenic functions. Based on the principle of CNV dysregulated gene expression, we have identified 372 lncRNAs whose expression levels vary with copy number amplitude. Through the combined Cox regression algorithm and LASSO algorithm, nine CNV-driven lncRNAs were identified as prognostic markers. Immune cell infiltration analysis revealed the composition of the immune microenvironment for LUSC and the leukocytes associated with the patient's survival risk. Moreover, the correlation between the nine pr-lncRNAs and immune cell types has been revealed, emphasizing the important role of *AC073130.1*, *GPRC5D-AS1*, and *LINC01133* in immune disorders. We next identified mRNAs driven by lncRNA-CNV and defined the function of lncRNA with CNV, which highlight the novel mechanism of non-coding RNA with CNV driving oncogenic function in LUSC.

In the past decades, with the development of high-throughput sequencing dataset and technologies (Yu et al., 2018; Yu et al., 2021), the integrated analysis of multiple omics data to reveal the pathogenesis of cancer has become the norm. For example, Zheng et al. (2020) identified three lncRNA prognostic characteristics of ovarian cancer based on genome-wide CNV. Although this study is also based on genome-wide CNV to identify prognostic-related lncRNA, it did not reveal the functions and regulatory mechanisms of pr-lncRNA. The analysis of pr-lncRNA's involvement in tumor immune infiltration and carcinogenic function cascade is an important feature of this study.

The lncRNAs *GPRC5D-AS1* and *LINC01133* were emphasized in this study as prognostic markers that play an important role in immune cell infiltration and carcinogenic mechanisms. The expression levels of *GPRC5D-AS1* and *LINC01133* were negatively correlated with the infiltration of multiple immune cells ($CD4^+$/$CD8^+$ T cell, dendritic cell, and neutrophil), suggesting that a high expression of *GPRC5D-AS1* and *LINC01133* can inhibit cell infiltration. This provides novel therapeutic target for the development of immune checkpoint therapy.

In summary, we provided prognostic-related lncRNA driven by CNV for LUSC and revealed the regulatory mechanism of lncRNA-CNV-mRNA oncogenic function. By revealing the function of pr-lncRNA, potential targets that can be used for immunotherapy have been identified. Taken together, our research provides useful theoretical guidance for the clinical diagnosis and treatment for LUSC.

DATA AVAILABILITY STATEMENT

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

AUTHOR CONTRIBUTIONS

JN, FW, and WL conceived and designed the experiments. JN and KZ analyzed the data. FW and BL collected the data. JN, FW, and QS validated the method and data. JN and WL wrote the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by the "Chunhui Plan" project of the Ministry of Education of China (No. HLJ2019011), the Health Commission Foundation of Heilongjiang Province (2018241 and 2019061), the Top Young Talent Project of Harbin Medical University Cancer Hospital (BJQN2020-02), and The Fundamental Research Funds for the Provincial Universities (2020-KYYWF-1469 and 2019-KYYWF-0357).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.779155/full#supplementary-material

REFERENCES

Alhamzawi, R., and Ali, H. T. M. (2018). The Bayesian Adaptive Lasso Regression. *Math. Biosciences* 303, 75–82. doi:10.1016/j.mbs.2018.06.004

Bishara, A. J., and Hittner, J. B. (2012). Testing the Significance of a Correlation with Nonnormal Data: Comparison of Pearson, Spearman, Transformation, and Resampling Approaches. *Psychol. Methods* 17, 399–417. doi:10.1037/a0028087

Bolha, L., Ravnik-Glavač, M., and Glavač, D. (2017). Long Noncoding RNAs as Biomarkers in Cancer. *Dis. Markers* 2017, 7243968. doi:10.1155/2017/7243968

Botling, J., Edlund, K., Lohr, M., Hellwig, B., Holmberg, L., Lambe, M., et al. (2013). Biomarker Discovery in Non-small Cell Lung Cancer: Integrating Gene Expression Profiling, Meta-Analysis, and Tissue Microarray Validation. *Clin. Cancer Res.* 19, 194–204. doi:10.1158/1078-0432.ccr-12-1139

Bryceson, Y. T., March, M. E., Ljunggren, H.-G., and Long, E. O. (2006). Synergy Among Receptors on Resting NK Cells for the Activation of Natural Cytotoxicity and Cytokine Secretion. *Blood* 107, 159–166. doi:10.1182/blood-2005-04-1351

Cui, L., Li, H., Hui, W., Chen, S., Yang, L., Kang, Y., et al. (2020). A Deep Learning-Based Framework for Lung Cancer Survival Analysis with Biomarker Interpretation. *BMC Bioinformatics* 21, 112. doi:10.1186/s12859-020-3431-z

Der, S. D., Sykes, J., Pintilie, M., Zhu, C.-Q., Strumpf, D., Liu, N., et al. (2014). Validation of a Histology-independent Prognostic Gene Signature for Early-Stage, Non-small-cell Lung Cancer Including Stage IA Patients. *J. Thorac. Oncol.* 9, 59–64. doi:10.1097/jto.0000000000000042

Ferlay, J., Soerjomataram, I., Dikshit, R., Eser, S., Mathers, C., Rebelo, M., et al. (2015). Cancer Incidence and Mortality Worldwide: Sources, Methods and Major Patterns in GLOBOCAN 2012. *Int. J. Cancer* 136, E359–E386. doi:10.1002/ijc.29210

Frankish, A., Diekhans, M., Jungreis, I., Lagarde, J., Loveland, J. E., Mudge, J. M., et al. (2021). *Nucleic Acids Res.* 49 (2021), D916–D923. doi:10.1093/nar/gkaa1087

Geng, W., Lv, Z., Fan, J., Xu, J., Mao, K., Yin, Z., et al. (2021). Identification of the Prognostic Significance of Somatic Mutation-Derived LncRNA Signatures of Genomic Instability in Lung Adenocarcinoma. *Front. Cel Dev. Biol.* 9, 657667. doi:10.3389/fcell.2021.657667

Goldstraw, P., Ball, D., Jett, J. R., Le Chevalier, T., Lim, E., Nicholson, A. G., et al. (2011). Non-small-cell Lung Cancer. *The Lancet* 378, 1727–1740. doi:10.1016/s0140-6736(10)62101-0

Guo, Q., He, Y., Sun, L., Kong, C., Cheng, Y., Wang, P., et al. (2019). Identification of Potential Prognostic TF-associated lncRNAs for Predicting Survival in Ovarian Cancer. *J. Cel Mol Med* 23, 1840–1851. doi:10.1111/jcmm.14084

Guo, Q., Wang, J., Gao, Y., Li, X., Hao, Y., Ning, S., et al. (2020). Dynamic TF-lncRNA Regulatory Networks Revealed Prognostic Signatures in the Development of Ovarian Cancer. *Front. Bioeng. Biotechnol.* 8, 460. doi:10.3389/fbioe.2020.00460

Guyot, P., Ades, A., Ouwens, M. J., and Welton, N. J. (2012). Enhanced Secondary Analysis of Survival Data: Reconstructing the Data from Published Kaplan-Meier Survival Curves. *BMC Med. Res. Methodol.* 12, 9. doi:10.1186/1471-2288-12-9

Herrera-Solorio, A. M., Peralta-Arrieta, I., Armas López, L., Hernández-Cigala, N., Mendoza Milla, C., Ortiz Quintero, B., et al. (2021). LncRNA SOX2-OT Regulates AKT/ERK and SOX2/GLI-1 Expression, Hinders Therapy, and Worsens Clinical Prognosis in Malignant Lung Diseases. *Mol. Oncol.* 15, 1110–1129. doi:10.1002/1878-0261.12875

Hirsch, F. R., Scagliotti, G. V., Mulshine, J. L., Kwon, R., Curran, W. J., Jr., Wu, Y.-L., et al. (2017). Lung Cancer: Current Therapies and New Targeted Treatments. *The Lancet* 389, 299–311. doi:10.1016/s0140-6736(16)30958-8

Jiang, X., Xu, J., Liu, M., Xing, H., Wang, Z., Huang, L., et al. (2019). Adoptive CD8+ T Cell Therapy against cancer: Challenges and Opportunities. *Cancer Lett.* 462, 23–32. doi:10.1016/j.canlet.2019.07.017

Katoh, M., and Katoh, M. (2020). Precision Medicine for Human Cancers with Notch Signaling Dysregulation (Review). *Int. J. Mol. Med.* 45, 279–297. doi:10.3892/ijmm.2019.4418

Lauss, M., Donia, M., Harbst, K., Andersen, R., Mitra, S., Rosengren, F., et al. (2017). Mutational and Putative Neoantigen Load Predict Clinical Benefit of Adoptive T Cell Therapy in Melanoma. *Nat. Commun.* 8, 1738. doi:10.1038/s41467-017-01460-0

Li, Y., Jiang, T., Zhou, W., Li, J., Li, X., Wang, Q., et al. (2020). Pan-cancer Characterization of Immune-Related lncRNAs Identifies Potential Oncogenic Biomarkers. *Nat. Commun.* 11, 1000. doi:10.1038/s41467-020-14802-2

Mantovani, A., Sozzani, S., Locati, M., Allavena, P., and Sica, A. (2002). Macrophage Polarization: Tumor-Associated Macrophages as a Paradigm for Polarized M2 Mononuclear Phagocytes. *Trends Immunol.* 23, 549–555. doi:10.1016/s1471-4906(02)02302-5

Mayakonda, A., Lin, D.-C., Assenov, Y., Plass, C., and Koeffler, H. P. (2018). Maftools: Efficient and Comprehensive Analysis of Somatic Variants in Cancer. *Genome Res.* 28, 1747–1756. doi:10.1101/gr.239244.118

Newman, A. M., Liu, C. L., Green, M. R., Gentles, A. J., Feng, W., Xu, Y., et al. (2015). Robust Enumeration of Cell Subsets from Tissue Expression Profiles. *Nat. Methods* 12, 453–457. doi:10.1038/nmeth.3337

Okayama, H., Kohno, T., Ishii, Y., Shimada, Y., Shiraishi, K., Iwakawa, R., et al. (2012). Identification of Genes Upregulated in ALK-Positive and EGFR/KRAS/ALK-negative Lung Adenocarcinomas. *Cancer Res.* 72, 100–111. doi:10.1158/0008-5472.can-11-1403

Ranstam, J., and Cook, J. A. (2017). Kaplan-Meier Curve. *Br. J. Surg.* 104, 442. doi:10.1002/bjs.10238

Ried, T., Meijer, G. A., Harrison, D. J., Grech, G., Franch-Expósito, S., Briffa, R., et al. (2019). The Landscape of Genomic Copy Number Alterations in Colorectal Cancer and Their Consequences on Gene Expression Levels and Disease Outcome. *Mol. Aspects Med.* 69, 48–61. doi:10.1016/j.mam.2019.07.007

Rinn, J. L., and Chang, H. Y. (2012). Genome Regulation by Long Noncoding RNAs. *Annu. Rev. Biochem.* 81, 145–166. doi:10.1146/annurev-biochem-051410-092902

Schmitt, A. M., and Chang, H. Y. (2016). Long Noncoding RNAs in Cancer Pathways. *Cancer Cell* 29, 452–463. doi:10.1016/j.ccell.2016.03.010

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., et al. (2003). Cytoscape: a Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Res.* 13, 2498–2504. doi:10.1101/gr.1239303

Tang, Z., Kang, B., Li, C., Chen, T., and Zhang, Z. (2019). GEPIA2: an Enhanced Web Server for Large-Scale Expression Profiling and Interactive Analysis. *Nucleic Acids Res.* 47, W556–W560. doi:10.1093/nar/gkz430

Tomczak, K., Czerwińska, P., and Wiznerowicz, M. (2015). The Cancer Genome Atlas (TCGA): an Immeasurable Source of Knowledge. *Contemp. Oncol. (Pozn)* 19, A68–A77. doi:10.5114/wo.2014.47136

Ulitsky, I. (2016). Evolution to the rescue: Using Comparative Genomics to Understand Long Non-coding RNAs. *Nat. Rev. Genet.* 17, 601–614. doi:10.1038/nrg.2016.85

van der Leun, A. M., Thommen, D. S., and Schumacher, T. N. (2020). CD8+ T Cell States in Human Cancer: Insights from Single-Cell Analysis. *Nat. Rev. Cancer* 20, 218–232. doi:10.1038/s41568-019-0235-4

Wang, P., Li, X., Gao, Y., Guo, Q., Ning, S., Zhang, Y., et al. (2020). LnCeVar: a Comprehensive Database of Genomic Variations that Disturb ceRNA Network Regulation. *Nucleic Acids Res.* 48, D111–D117. doi:10.1093/nar/gkz887

Wang, P., Guo, Q., Hao, Y., Liu, Q., Gao, Y., Zhi, H., et al. (2021). LnCeCell: a Comprehensive Database of Predicted lncRNA-Associated ceRNA Networks at Single-Cell Resolution. *Nucleic Acids Res.* 49, D125–D133. doi:10.1093/nar/gkaa1017

Wang, P., Li, X., Gao, Y., Guo, Q., Wang, Y., Fang, Y., et al. (2019). LncACTdb 2.0: an Updated Database of Experimentally Supported ceRNA Interactions Curated from Low- and High-Throughput Experiments. *Nucleic Acids Res.* 47, D121–D127. doi:10.1093/nar/gky1144

Wang, X., Yin, H., Zhang, L., Zheng, D., Yang, Y., Zhang, J., et al. (2019). The Construction and Analysis of the Aberrant lncRNA-miRNA-mRNA Network in Non-small Cell Lung Cancer. *J. Thorac. Dis.* 11, 1772–1778. doi:10.21037/jtd.2019.05.69

Wilusz, J. E., Sunwoo, H., and Spector, D. L. (2009). Long Noncoding RNAs: Functional Surprises from the RNA World. *Genes Develop.* 23, 1494–1504. doi:10.1101/gad.1800909

Xu, Y., Wu, T., Li, F., Dong, Q., Wang, J., Shang, D., et al. (2020). Identification and Comprehensive Characterization of lncRNAs with Copy Number Variations and Their Driving Transcriptional Perturbed Subpathways Reveal Functional Significance for Cancer. *Brief Bioinform* 21, 2153–2166. doi:10.1093/bib/bbz113

Yu, F., Sankaran, V. G., and Yuan, G. C. (2021). CUT&RUNTools 2.0: A Pipeline for Single-Cell and Bulk-Level CUT&RUN and CUT&Tag Data Analysis. *Bioinformatics*. doi:10.1093/bioinformatics/btab507

Yu, F., Zhang, G., Shi, A., Hu, J., Li, F., Zhang, X., et al. (2018). *LnChrom: A Resource of Experimentally Validated lncRNA-Chromatin Interactions in Human and Mouse*. Oxford: Database, 2018.

Yu, H., Boyle, T. A., Zhou, C., Rimm, D. L., and Hirsch, F. R. (2016). PD-L1 Expression in Lung Cancer. *J. Thorac. Oncol.* 11, 964–975. doi:10.1016/j.jtho.2016.04.014

Yuan, H., Liu, J., and Zhang, J. (2021). The Current Landscape of Immune Checkpoint Blockade in Metastatic Lung Squamous Cell Carcinoma. *Molecules* 26. doi:10.3390/molecules26051392

Zhang, E., and Li, X. (2019). LncRNA SOX2-OT Regulates Proliferation and Metastasis of Nasopharyngeal Carcinoma Cells through miR-146b-5p/HNRNPA2B1 Pathway. *J. Cel Biochem* 120, 16575–16588. doi:10.1002/jcb.28917

Zhang, M., Han, Y., Zheng, Y., Zhang, Y., Zhao, X., Gao, Z., et al. (2020). ZEB1-activated LINC01123 Accelerates the Malignancy in Lung Adenocarcinoma through NOTCH Signaling Pathway. *Cell Death Dis* 11, 981. doi:10.1038/s41419-020-03166-6

Zhang, Y., Han, P., Guo, Q., Hao, Y., Qi, Y., Xin, M., et al. (2021). Oncogenic Landscape of Somatic Mutations Perturbing Pan-Cancer lncRNA-ceRNA Regulation. *Front. Cel Dev. Biol.* 9, 658346. doi:10.3389/fcell.2021.658346

Zheng, M., Hu, Y., Gou, R., Nie, X., Li, X., Liu, J., et al. (2020). Identification Three LncRNA Prognostic Signature of Ovarian Cancer Based on Genome-wide Copy Number Variation. *Biomed. Pharmacother.* 124, 109810. doi:10.1016/j.biopha.2019.109810

Zhou, Y., Zhou, B., Pache, L., Chang, M., Khodabakhshi, A. H., Tanaseichuk, O., et al. (2019). Metascape Provides a Biologist-Oriented Resource for the Analysis of Systems-Level Datasets. *Nat. Commun.* 10, 1523. doi:10.1038/s41467-019-09234-6

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

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

frontiers in Genetics

ORIGINAL RESEARCH
published: 04 January 2022
doi: 10.3389/fgene.2021.783970

Gene Co-Expression Network Analysis Identifies Vitamin D-Associated Gene Modules in Adult Normal Rectal Epithelium Following Supplementation

James P. Blackmur[1,2], Peter G. Vaughan-Shaw[1,2], Kevin Donnelly[1,2], Bradley T. Harris[2], Victoria Svinti[2], Anna-Maria Ochocka-Fox[1], Paz Freile[2], Marion Walker[1,2], Toby Gurran[1,2], Stuart Reid[1,2], Colin A. Semple[1], Farhat V. N. Din[2], Maria Timofeeva[2,3], Malcolm G. Dunlop[1,2] and Susan M. Farrington[2]**

[1]MRC Human Genetics Unit, Institute of Genetics and Cancer, University of Edinburgh, Edinburgh, United Kingdom, [2]Cancer Research UK Edinburgh Centre, Institute of Genetics and Cancer, University of Edinburgh, Edinburgh, United Kingdom, [3]Department of Public Health, Danish Institute for Advanced Study, University of Southern Denmark, Odense, Denmark

OPEN ACCESS

Edited by:
Stephen J. Bush,
University of Oxford, United Kingdom

Reviewed by:
Wei Liu,
Fujian Agriculture and Forestry University, China
Mohammad Farhadian,
University of Tabriz, Iran

***Correspondence:**
James P. Blackmur
jblackm2@ed.ac.uk
Susan M. Farrington
susan.farrington@ed.ac.uk

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

Received: 27 September 2021
Accepted: 02 December 2021
Published: 04 January 2022

Citation:
Blackmur JP, Vaughan-Shaw PG, Donnelly K, Harris BT, Svinti V, Ochocka-Fox A-M, Freile P, Walker M, Gurran T, Reid S, Semple CA, Din FVN, Timofeeva M, Dunlop MG and Farrington SM (2022) Gene Co-Expression Network Analysis Identifies Vitamin D-Associated Gene Modules in Adult Normal Rectal Epithelium Following Supplementation. Front. Genet. 12:783970. doi: 10.3389/fgene.2021.783970

Colorectal cancer (CRC) is a common, multifactorial disease. While observational studies have identified an association between lower vitamin D and higher CRC risk, supplementation trials have been inconclusive and the mechanisms by which vitamin D may modulate CRC risk are not well understood. We sought to perform a weighted gene co-expression network analysis (WGCNA) to identify modules present after vitamin D supplementation (when plasma vitamin D level was sufficient) which were absent before supplementation, and then to identify influential genes in those modules. The transcriptome from normal rectal mucosa biopsies of 49 individuals free from CRC were assessed before and after 12 weeks of 3200IU/day vitamin D (Fultium-D3) supplementation using paired-end total RNAseq. While the effects on expression patterns following vitamin D supplementation were subtle, *WGCNA* identified highly correlated genes forming gene modules. Four of the 17 modules identified in the post-vitamin D network were not preserved in the pre-vitamin D network, shedding new light on the biochemical impact of supplementation. These modules were enriched for GO terms related to the immune system, hormone metabolism, cell growth and RNA metabolism. Across the four treatment-associated modules, 51 hub genes were identified, with enrichment of 40 different transcription factor motifs in promoter regions of those genes, including VDR:RXR. Six of the hub genes were nominally differentially expressed in studies of vitamin D effects on adult normal mucosa organoids: *LCN2, HLA-C, AIF1L, PTPRU, PDE4B and IFI6.* By taking a gene-correlation network approach, we have described vitamin D induced changes to gene modules in normal human rectal epithelium *in vivo*, the target tissue from which CRC develops.

Keywords: colorectal cancer, vitamin D, vitamin D supplementation, gene correlation network, WGCNA (weighted gene co-expression network analysis)

INTRODUCTION

Colorectal cancer (CRC) is common, with over 40,000 incident cases and over 15,000 deaths associated with the disease per year in the United Kingdom (Cancer Research UK, 2019). In case-control and prospective cohort studies, higher plasma 25(OH)D level and higher dietary intake of vitamin D are associated with lower CRC risk (Jenab et al., 2010; Theodoratou et al., 2012; Theodoratou et al., 2014). Potential co-causality of CRC risk and vitamin D status (e.g. socioeconomic status, diet, physical exercise) and reverse causation (CRC or its treatment affecting serum vitamin D concentration) mean these observational studies may be confounded. Supplementation trials have been inconclusive, with randomised-controlled trials (RCTs) failing to show effect on CRC or adenoma (precursor lesion) incidence (Wactawski-Wende et al., 2006; Baron et al., 2015; Manson et al., 2019; Scragg, 2019). However, these supplementation studies are themselves confounded by short duration of follow up and vagaries in genetic, environmental, ethnic, dietary and ecological factors such as latitude, weather and sunlight exposure, with many participants in both control and intervention arms starting trials replete in vitamin D, and/or taking low-dose vitamin D supplementation. Whether vitamin D supplementation reduces risk of CRC therefore remains an open question. In addition, recent studies have suggested a beneficial effect for vitamin D supplementation on CRC mortality (Keum et al., 2019; Vaughan-Shaw et al., 2020).

The mechanisms by which vitamin D may modulate CRC risk and survival are not well understood (Alleyne et al., 2017). Potential mechanisms include induction of cell differentiation and apoptosis or inhibition of cell growth and proliferation (Lamprecht and Lipkin, 2003; Feldman et al., 2014). Understanding of vitamin D transcriptional responses in the colon and rectum has primarily come from studies in transformed cancer cell lines, which may not represent responses in normal healthy tissue (Mapes et al., 2014), hence it is advantageous and timely to prioritise human studies to explore potential mechanisms in the normal target tissue.

Gene co-expression networks are used to describe the pairwise relationships of a large number of gene expression variables (Zhang and Horvath, 2005). Genes related by high correlation coefficients are thought to be functionally related, members of the same pathway and/or co-regulated. Weighted gene co-expression network analysis (*WGCNA*) is one of the most widely used network methods, whereby gene correlations are raised by a soft-thresholding power to form a scale-free network. The advantage of this method is that all correlations are included for analysis. Methods that apply a hard-threshold (e.g. $r > 0.7$ being arbitrarily biologically relevant) lose connections below that threshold for further analysis (Zhang and Horvath, 2005). Highly connected genes within each module identified by *WGCNA* are defined as hub genes, with such genes thought to play key biological roles in that particular module or in regulation of a particular trait (Zhang and Horvath, 2005; Kogelman et al., 2014; Drag et al., 2017; Bakhtiarizadeh et al., 2018). By comparing differences in network structures (e.g. whether gene modules are preserved between conditions), it is possible to assess how groups of genes are perturbed by a certain condition (such as vitamin D supplementation) (Langfelder et al., 2011). This method has been successfully used to discover genes involved in endometriosis (Bakhtiarizadeh et al., 2018) and multiple cancer types (Sun et al., 2017; Zhang et al., 2018; Zhu et al., 2019).

In this study, we aimed to determine effects of vitamin D on normal rectal epithelium by assessing gene expression before and after 12 weeks of vitamin D supplementation in human subjects. We first assessed vitamin D effects on the expression of single genes, and then constructed a weighted correlation network to assess effects of vitamin D on groups of genes, and to identify hub genes in modules emerging after supplementation. We sought to functionally annotate genes and gene modules using Gene Ontology (GO) pathway analysis and to determine transcription factor binding sites common to hub genes in emergent modules. Finally, we sought to validate expression changes of hub genes in adult normal colorectal mucosa organoids treated with vitamin D. These organoid models are isolated from many of the vagaries in heterogeneity of the human population, yet maintain the genetic architecture and 3D-cell arrangement present in the parent tissue.

MATERIALS AND METHODS

Human Vitamin D Supplementation Study

In the Scottish Vitamin D (SCOVID) study, rectal normal mucosa biopsies (via rigid sigmoidoscopy) and blood were collected after informed consent from a cohort of human subjects free from colorectal cancer (n = 50, 49 whose samples passed QC). Demographic information is provided in **Table 1**, while the study protocol has been described elsewhere (Vaughan-Shaw et al., 2021). The study had approval from NHS Research Ethics Committee (REC No 13/SS/0248) and local Research and Development Committee (R&D Project ID 2014/0058). Participants received 3200IU daily oral vitamin D (Fultium-D3), with resampling at 12 weeks. Biopsy samples were stored immediately in RNA Later (Invitrogen) and kept for 48 h at 4°C before RNA extraction.

Plasma 25(OH)D was assayed from blood by mass spectrometry. Plasma extracted from blood taken in lithium heparin tubes was immediately frozen at −40°C and subsequently submitted to the Clinical Biochemistry department, Glasgow Royal Infirmary, United Kingdom for measurement of 25(OH)D.

To extract RNA, human biopsy samples transferred to 2 ml Eppendorf tubes and homogenised in Trizol. RNA was then extracted by Ribopure Kit (Invitrogen) according to the manufacturer's protocol. RNA samples from 49 subjects passed QC and were submitted to the Edinburgh Genomics facility, with sequencing on the Illumina HiSeq 2,500 in "rapid mode" with 150 bp paired-end reads as described in Supp Methods.

Analysis

Transcript quantification from RNAseq was conducted using *Salmon* v0.11 (Patro et al., 2017) using Ensembl version

TABLE 1 | Demographic and clinical information of SCOVID study participants.

Factor		
Age	Median years (IQR)	66 (58–72)
Gender	F/M	23/26
BMI	Median kg/m2 (IQR)	26.21 (23.66–31.64)
Current CRC	N/Y	49/0
Past History CRC	N/Y	31/18
Pre-supplementation		
Plasma 25(OH)D	Median nmol/l (IQR)	86 (23–54)
	n < 25 nmol/l	15
	n 25–50 nmol/l	20
	n > 50 nmol/l	14
Post-supplementation		
Plasma 25(OH)D	Median nmol/l (IQR)	89 (71–109)
	n < 25 nmol/l	0
	n 25–50 nmol/l	1
	n > 50 nmol/l	48

GRCh38, March 2017, Ensembl 88. Gene level counts were generated by R packages *txiimport* (Soneson et al., 2015) and annotated using *biomaRt* (Durinck et al., 2005; Durinck et al., 2009). Expression was normalised using a TMM algorithm based on gene expression thresholds of >0.1 Transcripts Per Million (TPM) and ≥6 reads in ≥20% of samples. Trimmed mean of M-values (TMM) between-sample normalisation was applied to the counts, as per the GTEx (v7) protocol. Effects of vitamin D on gene expression were assessed using *edgeR v3.32.1* (Robinson et al., 2010). Following dispersion estimation, a quasi-likelihood negative binomial generalized log-linear model was fitted to the count data. Differential expression was assessed by quasi-likelihood F-test, with the parent-tissue anonymous identifier included as a covariate in the model as per a paired-design. Significance was determined as FDR corrected p value < 0.05. GO pathway analysis was carried out by clusterProfiler (Yu et al., 2012).

Normalised counts were first log transformed before proceeding to correlation analysis by *WGCNA*. To minimise the biological noise from genes not functionally related to vitamin D and to limit the dataset for computational analysis, the top 25% most variable genes after vitamin D supplementation (determined by logFC) were taken forward for analysis. *WGCNA* analysis is sensitive to the presence of outliers (Horvath, 2011), therefore samples with a standardized connectivity score of less than −5 were removed. The goodSamplesGenes function was then used to remove samples and genes with missing entries (more than 50% missing entries) and genes with zero variance.

Vitamin D supplementation achieved a 25(OH)D concentration which might be termed "healthy" whereas the pre-supplementation samples were depleted in 25(OH)D, and could be termed the "disease" state. Based on the assumption that modules present after vitamin D supplementation but not present before would be of interest in discerning vitamin D activity, the post-vitamin D samples were considered as the reference set for module derivation. Module preservation analysis was then undertaken in the pre-vitamin D samples. Using *WGCNA*, we created a signed weighted gene co-expression network based on normal gene expression data. A weighted network was created from the pairwise biweight midcorrelation coefficients between genes using the *blockwiseModules* function, with module merge cut height of 0.25 and a minimum module size of 30 genes (Song et al., 2012; Bakhtiarizadeh et al., 2018). A weighted adjacency matrix was formed by raising correlations to the power of 7, which was chosen using the scale-free topology criterion (Zhang and Horvath, 2005; Horvath, 2011). The relationship between the power (β) and R^2 for a scale-free network is demonstrated along with sample dendrograms and their trait relationships in **Supplementary Figure S1**.

Age, gender and body mass index (BMI) have been reported to be associated with vitamin D concentration (Lagunova et al., 2009; Muscogiuri et al., 2019) and hence to assess association of those traits with identified gene modules, we assessed trait correlations to module eigenvectors (first principal component of each module). Each module eigenvector represents the expression profiles of all genes within that module.

To assess the preservation of post-vitamin D network modules in the pre-vitamin D dataset, the *modulePreservation* function in the *WGCNA* package was applied. We then applied Zsummary and medianRank (with 200 permutations) to detect module preservation. A module was considered as non-preserved if it had Zsummary<5 or medianRank≥8 (Langfelder et al., 2011).

To identify hub genes, intramodular connectivity (kIM) and module membership (kME) measures were used. Intramodular connectivity measures the degree of co-expression of a given gene with respect to the genes of a particular module. This was determined for both post- and pre-supplementation networks from the respective matrices of log transformed normalised counts using the *intramodularConnectivity.fromExpr* function (using pairwise biweight midcorrelation coefficients, power 7,

a signed network and using the module labels identified above). Module membership was determined similarly by the *signedKME* function, which determines the correlation between the expression profile of a gene and the module eigengene (first principal component of a particular module). Genes were kME ≥0.7 or kIM ≥0.7 were considered as hub genes to the respective module (Horvath, 2011).

Validation of Gene Modules in the STRING Dataset

We sought to validate gene modules identified by *WGCNA* in the STRING (Search Tool for the Retrieval of Interacting Genes/Proteins) curated database of protein-protein interactions (https://string-db.org/). The STRING database collects, scores and integrates publicly available sources of protein–protein interaction information, and can be used to assess if groups of genes identified by the user are enriched for protein-protein interactions in those sources (Szklarczyk et al., 2019). HGNC gene names within each module were uploaded to the STRING web-browser, and interactions assessed across all seven of the STRING "interaction sources".

Investigation of Common Transcription Factor Binding Sites

We were interested to assess if hub genes in non-preserved modules may be regulated by common transcription factors. As per Bakhtiarizadeh et al. (2018), the "TRANSFAC_and_JASPAR_PWMs" section of the Enrichr tool (Chen et al., 2013) (https://amp.pharm.mssm.edu/Enrichr/) was applied to determine common transcription factor binding sites in promoter regions of such genes.

Validation of Hub Gene Expression Changes in Adult Normal Colorectal Organoids

Fernandez-Barral et al. (2020) undertook differential expression analysis of adult normal mucosa organoids (FB-ANMO) derived from six individuals treated for 96 h with 100 nM calcitriol or 1% ethanol. Normalised counts and anonymised meta-data were downloaded from GEO (GSE100785; https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc = GSE100785). The same edgeR pipeline as described above was then used to determine differentially expressed genes in the FB-ANMO. Gene set testing was carried out using the *geneSetTest* function in *limma* (Ritchie et al., 2015).

RESULTS

No Genes Were Differentially Expressed at a Genome-Wide Level Following Vitamin D Supplementation

Fultium-D3 supplementation increased plasma 25(OH)D in all study participants [mean (SD) nmol/l baseline 39.4 (20.4), 12 weeks 92.2 (27.0), $p < 0.001$; mean increase 52.8 (28.3)]. 21,650 genes passed expression filters and were taken forward for differential expression analysis. 2,492 genes were nominally differentially expressed ($p < 0.05$), but none remained significant after genome-wide FDR correction (FDR $p < 0.05$, **Supplementary Table S1**). 150 GO Biological Process terms were enriched in nominally differentially expressed genes including "protein-containing complex localization", "DNA conformation change", and "RNA splicing" (**Supplementary Table S2**).

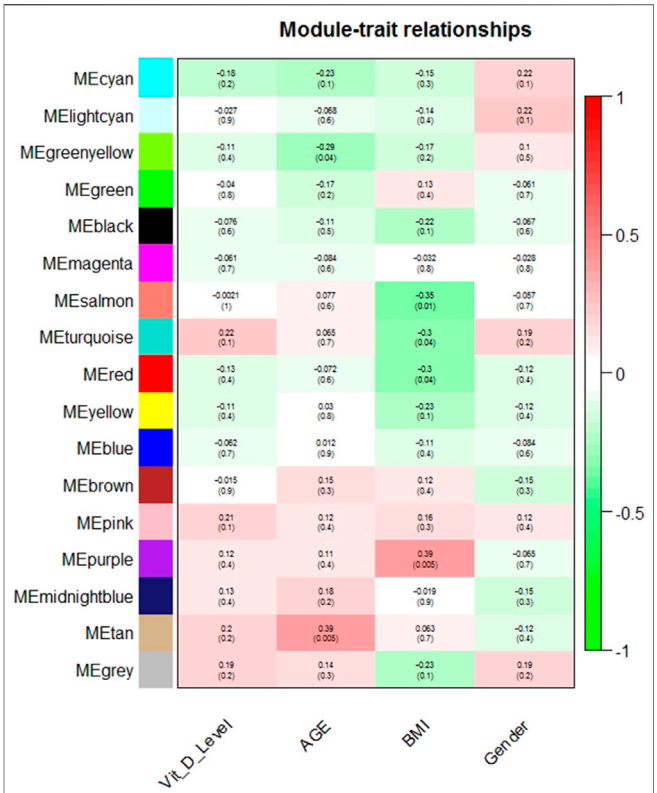


FIGURE 1 | Module-trait relationships in the post-vitamin D network. Pearson correlation of module eigenvector (first principal component) and trait (along with nominal *p*-value).

Gene Modules Identified by *WGCNA* After Vitamin D Supplementation

Having identified no genes differentially expressed at a genome-wide significance level, we were interested to examine if vitamin D changes occurred across groups of genes by network connectivity analysis. From the total pool of 21,650 genes, the top 25% (5,412) genes by logFC (irrespective of direction of effect) were taken forward for further analysis. This limited analysis to the genes potentially most responsive to vitamin D supplementation for computational reasons. No samples had outlying standardized connectivity score, and no genes were lost by the *goodSamplesGenes* function.

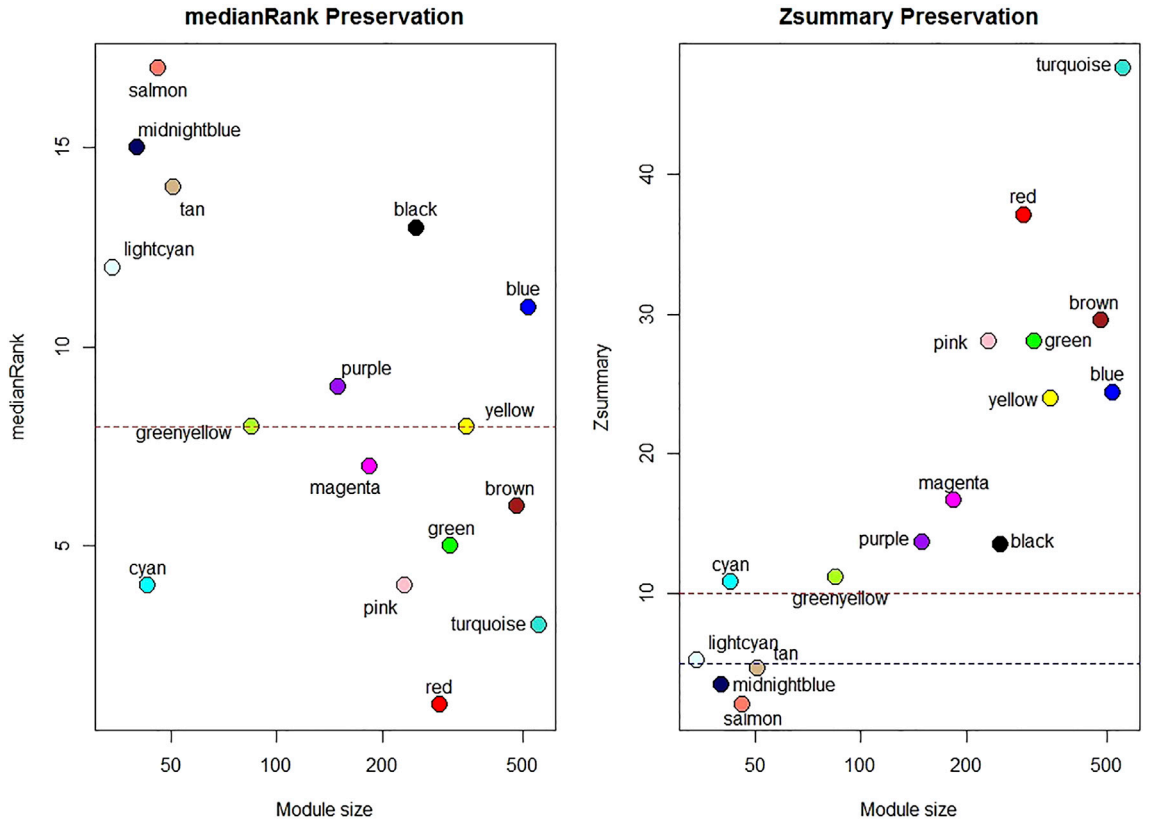


FIGURE 2 | Median rank and Z-summary statistics for preservation of modules from the post-vitamin D network in the pre-vitamin D network. Z-statistic >10 strong evidence of preservation, 5–10 moderate evidence of preservation, 2–5 weak evidence of preservation and <2 no evidence of preservation (Langfelder et al., 2011).

17 modules were identified in the post-vitamin D network including the unclassified module (grey). No module eigenvector was correlated with plasma 25(OH)D or gender (**Figure 1**, **Supplementary Table S3**). The tan and greenyellow module eigenvectors were nominally significantly correlated with age, though did not remain significantly correlated after correction for multiple testing (r = 0.39, p = 0.005, FDR p = 0.09; r = -0.29, p = 0.04, FDR p = 0.38 respectively). The purple, salmon, red and turquoise module eigenvectors were nominally correlated with BMI, however again were not significant after accounting for multiple testing (r = 0.39, p = 0.005, FDR p = 0.09; r = -0.35, p = 0.01, FDR p = 0.13; r = -0.30, p = 0.04, FDR p = 0.16; r = -0.30, p = 0.04, FDR p = 0.16 respectively).

Identification of Treatment-Associated Modules From the Pre-Vitamin D Transcriptome

One sample with outlying standardized connectivity score was removed from the pre-vitamin D network prior to module preservation analysis. 12 modules from the post-vitamin D network were strongly preserved in the pre-vitamin D network (defined as Zsummary statistic >10), i.e. were not associated with vitamin D treatment, while four modules from the post-vitamin D network were not preserved in the pre-vitamin D network, and hence were considered to be treatment-associated modules (Zsummary statistic <5 and median rank >8 for salmon, midnightblue and tan modules. Lightcyan module Zsummary statistic 5.2, Zdensity. pres 4.1 and Zconnectivity 6.3 with median rank 12). Module preservation statistics are described in **Supplementary Table S4** and their distribution in **Figure 2**. GO terms for each module are described in **Table 2** and **Supplementary Table S5**.

As a means of testing if *WGCNA* had identified gene modules which had also been found to be interacting in other datasets, we next sought to test if these modules were enriched in the STRING curated database of protein-protein interactions (PPI). Three of the four modules which were not preserved in the pre-vitamin D network were enriched for protein-protein interactions (**Table 3**). In addition 11 of the 12 preserved modules were enriched for protein-protein interactions (**Supplementary Table S6**).

Identification of Hub Genes in Treatment-Associated Modules

The genes with the highest degree of connectivity within a particular module are considered as hub genes. We were interested to identify hub genes in treatment-associated (non-preserved) modules, and in particular those genes which gained or lost hub-status. Hub genes were defined as those with kME ≥0.7 or with kIM≥0.7. By this

TABLE 2 | Select GO biological process terms in each of the modules present in the post-vitamin D network, with associated statistics for preservation in the pre-vitamin D network. The top three GO biological process terms for each module (determined by FDR *p*-value) are shown. Where highly similar terms related to overlapping genes existed (e.g., purine ribonucleoside binding and purine nucleoside binding) only one term is shown.

Module	Module size	Z summary	Median rank	N GO terms	GO description
Salmon	46	2.0	17	9	Defense response to virus; RNA helicase; GTP binding;
Midnightblue	40	3.47	15	158	viral mRNA export from host cell nucleus; RNA secondary structure unwinding; negative regulation of DNA damage checkpoint
Tan	51	4.60	14	5	positive regulation of hormone metabolic process; sphingolipid mediated signaling pathway; positive regulation of nuclear division
Lightcyan	34	5.21	12	49	adaptive immune response based on somatic recombination of immune receptors built from immunoglobulin superfamily domains; regulation of leukocyte mediated cytotoxicity
Cyan	43	10.83	4	141	ribosome biogenesis; ncRNA processing; nucleocytoplasmic transport
Greenyellow	85	11.19	8	202	homophilic cell adhesion via plasma membrane adhesion molecules; extracellular matrix organization; smoothened signaling pathway
Black	249	13.48	13	603	extracellular matrix organization; muscle contraction; axonogenesis
Purple	150	13.65	9	112	RNA splicing, via transesterification reactions; cilium organization; RNA transport
Magenta	184	16.64	7	166	positive regulation of viral release from host cell; vacuolar transport; endosome organization
Yellow	348	23.94	8	783	lymphocyte differentiation; regulation of T cell activation; positive regulation of leukocyte cell-cell adhesion
Blue	521	24.32	11	94	peptidyl-lysine modification; TORC1 signaling; histone modification
Pink	231	28.05	4	100	alcohol metabolic process; cellular response to extracellular stimulus; macroautophagy
Green	311	28.05	5	466	oxidative phosphorylation; cellular respiration; mitochondrial translation
Brown	481	29.54	6	5	apoptotic process involved in morphogenesis
Red	290	37.08	1	375	muscle tissue development; extracellular matrix organization; multicellular organismal signaling
Turquoise	556	47.57	3	445	leukocyte cell-cell adhesion; regulation of T cell activation; lymphocyte proliferation

TABLE 3 | Protein-protein interaction enrichment of genes in non-preserved modules identified in the STRING database. FDR correction for 16 modules tested (grey unclassified module excluded).

Module	N genes in string database	N genes in module	Number of connections	PPI enrichment *p* value	PPI enrichment FDR *p* value
Salmon	19	46	13	1.12e-05	1.28E-05
Midnightblue	22	40	12	5.14e-07	6.85E-07
Tan	38	47	6	0.093	0.099
Lightcyan	24	28	29	<1.0e-16	<2.67E-16

definition, eight hub genes were identified in the salmon module, 13 in the midnightblue, 21 in the tan and nine in the lightcyan module (**Supplementary Table S7**).

Enrichment of Transcription Factor Binding Sites in Hub Gene Promoter Regions

Given hub genes are thought to be functionally important within their respective modules, we next sought to identify transcription factors common to hub genes in the non-preserved modules. A total of 40 different transcription factor motifs were enriched in promoter regions of hub genes in treatment-associated modules (**Supplementary Table S8**). Each of the four treatment-associated modules included hub genes which contained either VDR:RXR or RXR binding motifs in their respective promoter regions.

Crossover With Differentially Expressed Genes in NM Organoids

Finally, we reviewed differential expression of hub genes in a study of adult normal mucosa organoids treated with vitamin D. 15 of the 51 hub genes in treatment-associated modules identified above were present in the FB-ANMO dataset after gene filters. Those 15 genes were significantly downregulated on gene set testing in FB-ANMO ($p = 0.02$). Six of the 15 hub genes were also nominally differentially expressed in FB-ANMO, with five remaining significant after correction for multiple testing (Genome-wide FDR <0.05, **Table 4**, **Supplementary Table S9**). *WGCNA* is not recommended on datasets with fewer than 15 samples (Langfelder and Horvath, 2017), and hence we did not proceed to network analysis in this organoid dataset (6 samples per condition).

TABLE 4 | Hub genes in non-preserved modules which are also differentially expressed in adult normal mucosa organoids from re-analysis of Fernandez-Barral et al. (FB-ANMO).

Gene	Module	logFC SCOVID	Pvalue	kME post	kME pre	kIM post	kIM pre	logFC organoid	Pvalue organoid	FDR organoid
LCN2	Tan	0.20	1.17E-01	0.47	0.57	0.75	0.85	-0.62	1.51E-04	2.27E-03
AIF1L	Tan	0.12	4.79E-02	0.44	0.40	0.58	0.71	-0.70	1.38E-03	1.03E-02
HLA-C	Lightcyan	0.14	4.30E-01	0.82	0.80	0.86	0.89	0.22	2.92E-03	1.31E-02
PTPRU	Tan	0.21	8.56E-02	0.64	0.61	0.94	0.92	-0.28	3.49E-03	1.31E-02
PDE4B	Tan	0.12	2.39E-01	0.50	0.63	0.72	0.92	0.88	8.90E-03	2.67E-02
IFI6	Salmon	0.17	1.60E-01	0.34	0.55	0.33	0.74	-0.33	3.38E-02	8.45E-02

DISCUSSION

In this study, we have demonstrated that while vitamin D supplementation does not affect the expression of any single gene at a genome-wide level of significance, it does induce changes in the inter-correlated expression patterns across genes, reflected in modules. 16 gene modules were identified after vitamin D supplementation (a state which reflects those individuals having a sufficient level of vitamin D for health). 14 of the 16 identified modules were also noted to be significantly enriched for protein-protein interactions in the STRING curated database, suggesting that *WGCNA* is able to identify modules which may have functional relevance.

Setting these results in context, other studies have similarly found no or few differentially expressed single genes following vitamin D supplementation (Hossein-nezhad et al., 2013; Munro, 2016; Pasing et al., 2017), however pathway analyses have identified effects on fatty acid metabolism and PPAR signaling (Munro, 2016), MAPK signaling, NF-kappa B signaling, T cell receptor signaling and prostate cancer (Hossein-nezhad et al., 2013). No study has previously investigated effects of vitamin D supplementation by the gene-correlation network approach. We were particularly interested to identify four treatment-associated modules that were not observed before vitamin D supplementation (a state of vitamin D insufficiency). Functional terms enriched in those modules included multiple GO terms related to the immune system, along with terms related to hormone metabolism, cell growth and RNA metabolism. Enrichment of GO terms associated with immunity is of particular note given the general interest in vitamin D and the immune system, in particular with risk of conditions such as multiple sclerosis and inflammatory bowel disease being associated with lower serum vitamin D (Munger et al., 2006; Limketkai et al., 2017; Fletcher et al., 2019). Vitamin D has also been postulated to favourably benefit the immune response in severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) (Martineau and Forouhi, 2020). Mechanistically, an excess of VDR binding variants identified by ChIP-exo has been reported to overlap with genomic variants associated with autoimmune disorders such as inflammatory bowel disease, Crohn's disease and rheumatoid arthritis (Gallone et al., 2017).

Hub genes represent the most connected genes within a module, and are thought to be functionally important. Interestingly each of the non-preserved modules contained hub genes which contained VDR:RXR or RXR motifs in their respective promoter regions. Vitamin D signaling occurs principally following the binding of the active form of vitamin D, $1,25(OH)_2D$, to the Vitamin D Receptor (VDR) (Kliewer et al., 1992). VDR forms a heterodimer complex with Retinoid X Receptor (RXR) to bind DNA via VDR-responsive elements (VDRE) largely characterized by the VDR-RXR motif (Kliewer et al., 1992; Zhang et al., 2011). While the presence of a motif does not indicate actual transcription factor binding *in-vivo*, it does add further support to the notion that these genes are central to the vitamin D response in normal rectal epithelium.

Finally, when the list of hub-genes in non-preserved modules was cross-referenced with a separate study of vitamin D response in adult normal mucosa organoids, a number of interesting candidate genes were identified including *LCN2*, *HLA-C* and *IFI6*. *LCN2,* found to have a RXR motif in its promoter region, is expressed by macrophages and epithelia in response to inflammation (Dahl et al., 2018) and inhibition of *LCN2*-modulated NF-kB pathway activation by vitamin D has been noted to promote cisplatin sensitivity of oral squamous cell carcinomas (Huang et al., 2019). *LCN2* acts in a bacteriostatic fashion (Dahl et al., 2018), which is noteworthy given the potential role of the gut microbiota in development of CRC (Saus et al., 2019). Human leukocyte antigens have been reported to be a major target of vitamin D physiological activity (Carlberg, 2019), with *HLA-C* being differentially expressed in peripheral blood mononuclear cells (PBMCs) following vitamin D supplementation of adult humans (Neme et al., 2019). Interferon alpha-inducible protein 6 (*IFI6*), also known as *G1P3* has been shown to contribute to hyperplasia, tamoxifen resistance and poor outcomes in breast cancer (Cheriyath et al., 2012). It was one of the top differentially expressed genes (log FC -3.04) following vitamin D treatment of airway smooth muscle cells derived from individuals following a fatal asthma episode (Himes et al., 2015).

It is worthy of note that no module eigenvector was correlated with plasma 25(OH)D. The relationships between vitamin D and gene expression may not be linear, and instead sigmoidal or U-shaped relationships may exist (Mizunashi et al., 1995; Ross et al., 2011; Macdonald et al., 2018). Assessing the relationships between plasma 25(OH)D and gene modules may therefore fail to show effects in individuals sufficient in vitamin D, as was the case following vitamin D supplementation in this

study. In addition, plasma vitamin D may not be an accurate measure of vitamin D in the target tissue of interest (in this case the rectal epithelium), with previous studies reporting a marked discrepancy between serum and tissue concentrations (Martinaityte et al., 2017).

This study represents a novel approach to assessing vitamin D effects. Unlike many of the large randomised-controlled trials of vitamin D effects (Wactawski-Wende et al., 2006; Baron et al., 2015; Manson et al., 2019; Scragg, 2019), the majority of participants were deficient in vitamin D at the start of the study period (n = 39 baseline plasma 25(OH)D < 50 nmol/l). The effects of vitamin D were also assessed directly in the target tissue of interest, the rectum, as opposed to assessing effects in blood which may have been technically easier to sample. This study is larger than many of the published studies assessing effects of vitamin D supplementation on gene expression (Hosseinnezhad and Holick, 2013; Gerke et al., 2014; Ryynänen et al., 2014; Protiva et al., 2016). Limitations of this study have been discussed elsewhere (Vaughan-Shaw et al., 2021). This study may have been too small to achieve sufficient power to assess individual gene significance. Individuals taking part in the study were not selected on the basis of initial plasma 25(OH)D; supplementation may have a sigmoidal or U-shaped relationship with gene expression (Mizunashi et al., 1995; Ross et al., 2011; Macdonald et al., 2018) and hence failing to select participants based on initial 25(OH)D could blunt the observed effect of supplementation. Finally sampling after 12 weeks of supplementation may not adequately capture early or later gene expression changes, however more frequent or delayed sampling would provide additional practical and ethical challenges.

SUMMARY

By taking a gene-correlation network approach, we have described vitamin D-induced changes to groups of genes in normal human rectal epithelium. By reviewing treatment-associated modules before and after vitamin D supplementation, we have identified hub genes which may play a key role in modulating vitamin D actions in normal rectal epithelium. This provides novel understanding of the mechanisms by which vitamin D may have beneficial effects on CRC risk and survival.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: https://www.ncbi.nlm.nih.gov/, GSE157982.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the NHS Research Ethics Committee (REC No 13/SS/0248). The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

Conceptualization, JPB, SMF, PVS, MGD; Methodology, JPB, PVS, KD, BH, VS, MT, TG, CS, FVND, SMF, MGD; Investigation, PVS, JPB, AMOF, PF, MT, MW, TG, SR, VS, KD; Writing - Original Draft JPB; Writing - Review & Editing, JPB, PVS, SMF, MGD, MT, KD, BH, FVND, CS, VS, AMOF, PF, MW, TG, SR; Funding Acquisition, SMF, MGD, PVS, JPB; Resources, SMF, MGD; Supervision, SMF, MGD.

FUNDING

This work was supported by funding for the infrastructure and staffing of the Edinburgh CRUK Cancer Research Centre; CRUK programme grant C348/A18927 (MGD/SMF). JB was supported by an ECAT-linked CRUK ECRC Clinical training award (C157/A23218). PV-S was supported by a NES SCREDS clinical lectureship, MRC Clinical Research Training Fellowship (MR/M004007/1), a Research Fellowship from the Harold Bridges bequest and by the Melville Trust for the Care and Cure of Cancer. The work received support from COST Action BM1206. FVND is supported by a CSO Senior Clinical Fellowship. This work was also funded by a grant to MGD as Programme Leader with the MRC Human Genetics Unit Centre Grant (U127527202 and U127527198 from 1/4/18).

ACKNOWLEDGMENTS

We are grateful to Donna Markie and Fiona McIntosh, and all those who continue to contribute to recruitment, data collection, and data curation for the Scottish Vitamin D Study. We acknowledge the expert support on sample preparation from the Genetics Core of the Edinburgh Wellcome Trust Clinical Research Facility in addition to the nursing and study facilities provided by the Clinical Research Facility.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.783970/full#supplementary-material

REFERENCES

A. C. Ross, C. L. Taylor, A. L. Yaktine, and H. B. Del Valle (Editors) (2011). *Dietary Reference Intakes for Calcium and Vitamin D* (Washington (DC): The National Academies Collection: Reports funded by National Institutes of Health).

Alleyne, D., Witonsky, D. B., Mapes, B., Nakagome, S., Sommars, M., Hong, E., et al. (2017). Colonic Transcriptional Response to 1α,25(OH) 2 Vitamin D 3 in African- and European-Americans. *J. Steroid Biochem. Mol. Biol.* 168, 49–59. doi:10.1016/j.jsbmb.2017.02.001

Bakhtiarizadeh, M. R., Hosseinpour, B., Shahhoseini, M., Korte, A., and Gifani, P. (2018). Weighted Gene Co-expression Network Analysis of Endometriosis and Identification of Functional Modules Associated with its Main Hallmarks. *Front. Genet.* 9, 453. doi:10.3389/fgene.2018.00453

Baron, J. A., Barry, E. L., Mott, L. A., Rees, J. R., Sandler, R. S., Snover, D. C., et al. (2015). A Trial of Calcium and Vitamin D for the Prevention of Colorectal Adenomas. *N. Engl. J. Med.* 373 (16), 1519–1530. doi:10.1056/nejmoa1500409

Cancer Research UK (2019). Colorectal Cancer Statistics 2019. Available from: https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/bowel-cancer#heading-Three.

Carlberg, C. (2019). Vitamin D Signaling in the Context of Innate Immunity: Focus on Human Monocytes. *Front. Immunol.* 10, 2211. doi:10.3389/fimmu.2019.02211

Chen, E. Y., Tan, C. M., Kou, Y., Duan, Q., Wang, Z., Meirelles, G. V., et al. (2013). Enrichr: Interactive and Collaborative HTML5 Gene List Enrichment Analysis Tool. *BMC Bioinformatics* 14, 128. doi:10.1186/1471-2105-14-128

Cheriyath, V., Kuhns, M. A., Jacobs, B. S., Evangelista, P., Elson, P., Downs-Kelly, E., et al. (2012). G1P3, an Interferon- and Estrogen-Induced Survival Protein Contributes to Hyperplasia, Tamoxifen Resistance and Poor Outcomes in Breast Cancer. *Oncogene* 31 (17), 2222–2236. doi:10.1038/onc.2011.393

Dahl, S. L., Woodworth, J. S., Lerche, C. J., Cramer, E. P., Nielsen, P. R., Moser, C., et al. (2018). Lipocalin-2 Functions as Inhibitor of Innate Resistance to Mycobacterium tuberculosis. *Front. Immunol.* 9, 2717. doi:10.3389/fimmu.2018.02717

Drag, M., Skinkyté-Juskiené, R., Do, D. N., Kogelman, L. J. A., and Kadarmideen, H. N. (2017). Differential Expression and Co-expression Gene Networks Reveal Candidate Biomarkers of Boar Taint in Non-castrated Pigs. *Sci. Rep.* 7 (1), 12205. doi:10.1038/s41598-017-11928-0

Durinck, S., Moreau, Y., Kasprzyk, A., Davis, S., De Moor, B., Brazma, A., et al. (2005). BioMart and Bioconductor: a Powerful Link between Biological Databases and Microarray Data Analysis. *Bioinformatics* 21 (16), 3439–3440. doi:10.1093/bioinformatics/bti525

Durinck, S., Spellman, P. T., Birney, E., and Huber, W. (2009). Mapping Identifiers for the Integration of Genomic Datasets with the R/Bioconductor Package biomaRt. *Nat. Protoc.* 4 (8), 1184–1191. doi:10.1038/nprot.2009.97

Feldman, D., Krishnan, A. V., Swami, S., Giovannucci, E., and Feldman, B. J. (2014). The Role of Vitamin D in Reducing Cancer Risk and Progression. *Nat. Rev. Cancer* 14 (5), 342–357. doi:10.1038/nrc3691

Fernández-Barral, A., Costales-Carrera, A., Buira, S. P., Jung, P., Ferrer-Mayorga, G., Larriba, M. J., et al. (2020). Vitamin D Differentially Regulates colon Stem Cells in Patient-Derived normal and Tumor Organoids. *FEBS J.* 287 (1), 53–72. doi:10.1111/febs.14998

Fletcher, J., Cooper, S. C., Ghosh, S., and Hewison, M. (2019). The Role of Vitamin D in Inflammatory Bowel Disease: Mechanism to Management. *Nutrients* 11 (5), 1019. doi:10.3390/nu11051019

Gallone, G., Haerty, W., Disanto, G., Ramagopalan, S. V., Ponting, C. P., and Berlanga-Taylor, A. J. (2017). Identification of Genetic Variants Affecting Vitamin D Receptor Binding and Associations with Autoimmune Disease. *Hum. Mol. Genet.* 26 (11), 2164–2176. doi:10.1093/hmg/ddx092

Gerke, A. K., Pezzulo, A. A., Tang, F., Cavanaugh, J. E., Bair, T. B., Phillips, E., et al. (2014). Effects of Vitamin D Supplementation on Alveolar Macrophage Gene Expression: Preliminary Results of a Randomized, Controlled Trial. *Multidiscip Respir. Med.* 9 (1), 18. doi:10.1186/2049-6958-9-18

Himes, B. E., Koziol-White, C., Johnson, M., Nikolos, C., Jester, W., Klanderman, B., et al. (2015). Vitamin D Modulates Expression of the Airway Smooth Muscle Transcriptome in Fatal Asthma. *PLoS One* 10 (7), e0134057. doi:10.1371/journal.pone.0134057

Horvath, S. (2011). *Weighted Network Analysis: Applications in Genomics and Systems Biology*. Berlin: Springer Science & Business Media.

Hossein-nezhad, A., and Holick, M. F. (2013). Vitamin D for Health: a Global Perspective. *Mayo Clinic Proc.* 88 (7), 720–755. doi:10.1016/j.mayocp.2013.05.011

Hossein-nezhad, A., Spira, A., and Holick, M. F. (2013). Influence of Vitamin D Status and Vitamin D3 Supplementation on Genome Wide Expression of white Blood Cells: a Randomized Double-Blind Clinical Trial. *PloS one* 8 (3), e58725. doi:10.1371/journal.pone.0058725

Huang, Z., Zhang, Y., Li, H., Zhou, Y., Zhang, Q., Chen, R., et al. (2019). Vitamin D Promotes the Cisplatin Sensitivity of Oral Squamous Cell Carcinoma by Inhibiting LCN2-Modulated NF-Kb Pathway Activation through RPS3. *Cell Death Dis* 10 (12), 936. doi:10.1038/s41419-019-2177-x

Jenab, M., Bueno-de-Mesquita, H. B., Ferrari, P., van Duijnhoven, F. J. B., Norat, T., Pischon, T., et al. (2010). Association between Pre-diagnostic Circulating Vitamin D Concentration and Risk of Colorectal Cancer in European Populations:a Nested Case-Control Study. *BMJ* 340, b5500. doi:10.1136/bmj.b5500

Keum, N., Lee, D. H., Greenwood, D. C., Manson, J. E., and Giovannucci, E. (2019). Vitamin D Supplementation and Total Cancer Incidence and Mortality: a Meta-Analysis of Randomized Controlled Trials. *Ann. Oncol.* 30 (5), 733–743. doi:10.1093/annonc/mdz059

Kliewer, S. A., Umesono, K., Mangelsdorf, D. J., and Evans, R. M. (1992). Retinoid X Receptor Interacts with Nuclear Receptors in Retinoic Acid, Thyroid Hormone and Vitamin D3 Signalling. *Nature* 355 (6359), 446–449. doi:10.1038/355446a0

Kogelman, L. J. A., Cirera, S., Zhernakova, D. V., Fredholm, M., Franke, L., and Kadarmideen, H. N. (2014). Identification of Co-expression Gene Networks, Regulatory Genes and Pathways for Obesity Based on Adipose Tissue RNA Sequencing in a Porcine Model. *BMC Med. Genomics* 7, 57. doi:10.1186/1755-8794-7-57

Lagunova, Z., Porojnicu, A. C., Lindberg, F., Hexeberg, S., and Moan, J. (2009). The Dependency of Vitamin D Status on Body Mass index, Gender, Age and Season. *Anticancer Res.* 29 (9), 3713–3720. doi:10.14341/2071-8713-4886

Lamprecht, S. A., and Lipkin, M. (2003). Chemoprevention of colon Cancer by Calcium, Vitamin D and Folate: Molecular Mechanisms. *Nat. Rev. Cancer* 3 (8), 601–614. doi:10.1038/nrc1144

Langfelder, P., and Horvath, S. (2017). WGCNA Package FAQ. Available from: https://horvath.genetics.ucla.edu/html/CoexpressionNetwork/Rpackages/WGCNA/faq.html.

Langfelder, P., Luo, R., Oldham, M. C., and Horvath, S. (2011). Is My Network Module Preserved and Reproducible. *Plos Comput. Biol.* 7 (1), e1001057. doi:10.1371/journal.pcbi.1001057

Limketkai, B. N., Mullin, G. E., Limsui, D., and Parian, A. M. (2017). Role of Vitamin D in Inflammatory Bowel Disease. *Nutr. Clin. Pract.* 32 (3), 337–345. doi:10.1177/0884533616674492

Macdonald, H. M., Reid, I. R., Gamble, G. D., Fraser, W. D., Tang, J. C., and Wood, A. D. (2018). 25-Hydroxyvitamin D Threshold for the Effects of Vitamin D Supplements on Bone Density: Secondary Analysis of a Randomized Controlled Trial. *J. Bone Miner Res.* 33 (8), 1464–1469. doi:10.1002/jbmr.3442

Manson, J. E., Cook, N. R., Lee, I.-M., Christen, W., Bassuk, S. S., Mora, S., et al. (2019). Vitamin D Supplements and Prevention of Cancer and Cardiovascular Disease. *N. Engl. J. Med.* 380 (1), 33–44. doi:10.1056/nejmoa1809944

Mapes, B., Chase, M., Hong, E., Ludvik, A., Ceryes, K., Huang, Y., et al. (2014). *Ex Vivo* culture of Primary Human Colonic Tissue for Studying Transcriptional Responses to 1α,25(OH)2 and 25(OH) Vitamin D. *Physiol. genomics* 46 (8), 302–308. doi:10.1152/physiolgenomics.00194.2013

Martinaityte, I., Kamycheva, E., Didriksen, A., Jakobsen, J., and Jorde, R. (2017). Vitamin D Stored in Fat Tissue during a 5-Year Intervention Affects Serum 25-Hydroxyvitamin D Levels the Following Year. *J. Clin. Endocrinol. Metab.* 102 (10), 3731–3738. doi:10.1210/jc.2017-01187

Martineau, A. R., and Forouhi, N. G. (2020). Vitamin D for COVID-19: a Case to Answer. *Lancet Diabetes Endocrinol.* 8 (9), 735–736. doi:10.1016/s2213-8587(20)30268-0

Mizunashi, K., Furukawa, Y., Abe, K., Takaya, K., and Yoshinaga, K. (1995). Resetting of Parathyroid Hormone Secretion after Vitamin D3 Treatment in Hypoparathyroidism and after Parathyroid Adenectomy in Primary Hyperparathyroidism. *Calcif Tissue Int.* 57 (1), 30–34. doi:10.1007/bf00298993

Munger, K. L., Levin, L. I., Hollis, B. W., Howard, N. S., and Ascherio, A. (2006). Serum 25-hydroxyvitamin D Levels and Risk of Multiple Sclerosis. *JAMA* 296 (23), 2832–2838. doi:10.1001/jama.296.23.2832

Munro, F. M. (2016). *The Effect of Vitamin D on Gene Expression in Colorectal Tumours and normal colon*. Otago: University of Otago. (Thesis, Master of Science) Retrieved from http://hdl.handle.net/10523/6170.

Muscogiuri, G., Barrea, L., Somma, C. D., Laudisio, D., Salzano, C., Pugliese, G., et al. (2019). Sex Differences of Vitamin D Status across BMI Classes: An Observational Prospective Cohort Study. *Nutrients* 11 (12), 3034. doi:10.3390/nu11123034

Neme, A., Seuter, S., Malinen, M., Nurmi, T., Tuomainen, T.-P., Virtanen, J. K., et al. (2019). *In Vivo* transcriptome Changes of Human white Blood Cells in Response to Vitamin D. *J. Steroid Biochem. Mol. Biol.* 188, 71–76. doi:10.1016/j.jsbmb.2018.11.019

Pasing, Y., Fenton, C. G., Jorde, R., and Paulssen, R. H. (2017). Changes in the Human Transcriptome upon Vitamin D Supplementation. *J. Steroid Biochem. Mol. Biol.* 173, 93–99. doi:10.1016/j.jsbmb.2017.03.016

Patro, R., Duggal, G., Love, M. I., Irizarry, R. A., and Kingsford, C. (2017). Salmon Provides Fast and Bias-Aware Quantification of Transcript Expression. *Nat. Methods* 14 (4), 417–419. doi:10.1038/nmeth.4197

Protiva, P., Pendyala, S., Nelson, C., Augenlicht, L. H., Lipkin, M., and Holt, P. R. (2016). Calcium and 1,25-dihydroxyvitamin D 3 Modulate Genes of Immune and Inflammatory Pathways in the Human colon: a Human Crossover Trial. *Am. J. Clin. Nutr.* 103 (5), 1224–1231. doi:10.3945/ajcn.114.105304

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., et al. (2015). Limma powers Differential Expression Analyses for RNA-Sequencing and Microarray Studies. *Nucleic Acids Res.* 43 (7), e47. doi:10.1093/nar/gkv007

Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2010). edgeR: a Bioconductor Package for Differential Expression Analysis of Digital Gene Expression Data. *Bioinformatics* 26 (1), 139–140. doi:10.1093/bioinformatics/btp616

Ryynänen, J., Neme, A., Tuomainen, T.-P., Virtanen, J. K., Voutilainen, S., Nurmi, T., et al. (2014). Changes in Vitamin D Target Gene Expression in Adipose Tissue Monitor the Vitamin D Response of Human Individuals. *Mol. Nutr. Food Res.* 58 (10), 2036–2045. doi:10.1002/mnfr.201400291

Saus, E., Iraola-Guzmán, S., Willis, J. R., Brunet-Vega, A., and Gabaldón, T. (2019). Microbiome and Colorectal Cancer: Roles in Carcinogenesis and Clinical Potential. *Mol. Aspects Med.* 69, 93–106. doi:10.1016/j.mam.2019.05.001

Scragg, R. K. R. (2019). Overview of Results from the Vitamin D Assessment (ViDA) Study. *J. Endocrinol. Invest.* 42 (12), 1391–1399. doi:10.1007/s40618-019-01056-z

Soneson, C., Love, M. I., and Robinson, M. D. (2015). Differential Analyses for RNA-Seq: Transcript-Level Estimates Improve Gene-Level Inferences. *F1000Res* 4, 1521. doi:10.12688/f1000research.7563.1

Song, L., Langfelder, P., and Horvath, S. (2012). Comparison of Co-expression Measures: Mutual Information, Correlation, and Model Based Indices. *BMC Bioinformatics* 13, 328. doi:10.1186/1471-2105-13-328

Sun, Q., Zhao, H., Zhang, C., Hu, T., Wu, J., Lin, X., et al. (2017). Gene Co-expression Network Reveals Shared Modules Predictive of Stage and Grade in Serous Ovarian Cancers. *Oncotarget* 8 (26), 42983–42996. doi:10.18632/oncotarget.17785

Szklarczyk, D., Gable, A. L., Lyon, D., Junge, A., Wyder, S., Huerta-Cepas, J., et al. (2019). STRING V11: Protein-Protein Association Networks with Increased Coverage, Supporting Functional Discovery in Genome-wide Experimental Datasets. *Nucleic Acids Res.* 47 (D1), D607–D613. doi:10.1093/nar/gky1131

Theodoratou, E., Palmer, T., Zgaga, L., Farrington, S. M., McKeigue, P., Din, F. V. N., et al. (2012). Instrumental Variable Estimation of the Causal Effect of Plasma 25-Hydroxy-Vitamin D on Colorectal Cancer Risk: a Mendelian Randomization Analysis. *PloS one* 7 (6), e37662. doi:10.1371/journal.pone.0037662

Theodoratou, E., Tzoulaki, I., Zgaga, L., and Ioannidis, J. P. A. (2014). Vitamin D and Multiple Health Outcomes: Umbrella Review of Systematic Reviews and Meta-Analyses of Observational Studies and Randomised Trials. *BMJ* 348, g2035. doi:10.1136/bmj.g2035

Vaughan-Shaw, P. G., Buijs, L. F., Blackmur, J. P., Theodoratou, E., Zgaga, L., Din, F. V. N., et al. (2020). The Effect of Vitamin D Supplementation on Survival in Patients with Colorectal Cancer: Systematic Review and Meta-Analysis of Randomised Controlled Trials. *Br. J. Cancer* 123 (11), 1705–1712. doi:10.1038/s41416-020-01060-8

Vaughan-Shaw, P. G., Grimes, G., Blackmur, J. P., Timofeeva, M., Walker, M., Ooi, L. Y., et al. (2021). Oral Vitamin D Supplementation Induces Transcriptomic Changes in Rectal Mucosa that Are Linked to Anti-tumour Effects. *BMC Med.* 19 (1), 174. doi:10.1186/s12916-021-02044-y

Wactawski-Wende, J., Kotchen, J. M., Anderson, G. L., Assaf, A. R., Brunner, R. L., O'Sullivan, M. J., et al. (2006). Calcium Plus Vitamin D Supplementation and the Risk of Colorectal Cancer. *N. Engl. J. Med.* 354 (7), 684–696. doi:10.1056/NEJMoa055222

Yu, G., Wang, L.-G., Han, Y., and He, Q.-Y. (2012). clusterProfiler: an R Package for Comparing Biological Themes Among Gene Clusters. *OMICS: A J. Integr. Biol.* 16 (5), 284–287. doi:10.1089/omi.2011.0118

Zhang, B., and Horvath, S. (2005). A General Framework for Weighted Gene Co-expression Network Analysis. *Stat. Appl. Genet. Mol. Biol.* 4, Article17. doi:10.2202/1544-6115.1128

Zhang, J., Chalmers, M. J., Stayrook, K. R., Burris, L. L., Wang, Y., Busby, S. A., et al. (2011). DNA Binding Alters Coactivator Interaction Surfaces of the Intact VDR-RXR Complex. *Nat. Struct. Mol. Biol.* 18 (5), 556–563. doi:10.1038/nsmb.2046

Zhang, X., Feng, H., Li, Z., Li, D., Liu, S., Huang, H., et al. (2018). Application of Weighted Gene Co-expression Network Analysis to Identify Key Modules and Hub Genes in Oral Squamous Cell Carcinoma Tumorigenesis. *Ott* 11, 6001–6021. doi:10.2147/ott.s171791

Zhu, Z., Jin, Z., Deng, Y., Wei, L., Yuan, X., Zhang, M., et al. (2019). Co-expression Network Analysis Identifies Four Hub Genes Associated with Prognosis in Soft Tissue Sarcoma. *Front. Genet.* 10, 37. doi:10.3389/fgene.2019.00037

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

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

frontiers in Genetics

ORIGINAL RESEARCH
published: 21 January 2022
doi: 10.3389/fgene.2021.795611

m^{6}A-mRNA Methylation Regulates Gene Expression and Programmable m^{6}A Modification of Cellular RNAs With CRISPR-Cas13b in Renal Cell Carcinoma

Ying Gan[1,2†], Aolin Li[1,2†], Jun Liu[3,4], Xiaofei Wang[1,2], Zhenan Zhang[1,2], Qinhan Li[1,2], Xiongjun Ye[3,4], Lin Yao[1,2]* and Qian Zhang[1,2]**

[1]Department of Urology, Peking University First Hospital, Beijing, China, [2]Beijing Key Laboratory of Urogenital Diseases (male) Molecular Diagnosis and Treatment Center, Beijing, China, [3]Urology and Lithotripsy Center, Peking University People's Hospital, Peking University, Beijing, China, [4]Peking University Applied Lithotripsy Institute, Peking University, Beijing, China

OPEN ACCESS

Edited by:
Dapeng Xiong,
Cornell University, United States

Reviewed by:
Dandan Yang,
Brigham and Women's Hospital and Harvard Medical School, United States
Heqiao Zhang,
ShanghaiTech University, China

***Correspondence:**
Lin Yao
poparies@163.com
Xiongjun Ye
yexiongjun@bjmu.edu.cn
Qian Zhang
zhangqianbjmu@126.com

†These authors have contributed equally to this work

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

Received: 15 October 2021
Accepted: 09 December 2021
Published: 21 January 2022

Citation:
Gan Y, Li A, Liu J, Wang X, Zhang Z, Li Q, Ye X, Yao L and Zhang Q (2022) m^{6}A-mRNA Methylation Regulates Gene Expression and Programmable m^{6}A Modification of Cellular RNAs With CRISPR-Cas13b in Renal Cell Carcinoma. Front. Genet. 12:795611. doi: 10.3389/fgene.2021.795611

Background: N^6-methyladenosine (m^{6}A) is the most extensive messenger RNA modification. Despite recent advances in the biological roles of m^{6}A, its role in the development and progression of renal cell carcinoma (RCC) remains unclear.

Methods: In this study, we gained the transcriptome-wide m^{6}A profile and gene expression pattern in RCC and paired adjacent peritumoral tissues by meRIP-seq and RNA-seq. m^{6}A modifications of mRNAs were validated by meRIP-qPCR in tissues, and targeted methylation or demethylation was validated by using a CRISPR-Cas13b-based tool in RCC cell lines.

Results: Our findings showed that there were 13,805 m^{6}A peaks among 5,568 coding gene transcripts (mRNAs) in adjacent tissues and 24,730 m^{6}A peaks among 6,866 mRNAs in tumor tissues. Furthermore, m^{6}A modification sites were usually located in the coding sequences (CDS), and some near the start and stop codons. Gene Ontology analysis revealed that coding genes had differential N^6-methyladenosine sites and were enriched in kidney development and cancer-related signaling pathways. We also found that different levels of m^{6}A modifications could regulate gene expression.

Conclusion: In summary, our results provided evidence for studying the potential function of RNA m^{6}A modification and m^{6}A-mediated gene expression regulation in human RCC.

Keywords: m6A, gene expression, Programmable Modification, CRISPR-Cas13b, Renal Cell Carcinoma

INTRODUCTION

Renal cell carcinoma (RCC) is one of the most common malignancies in the genitourinary system. Kidney cancer is the sixth most common cancer in men, with 73,820 estimated new cases and 14,770 estimated deaths a year in the United States, according to the latest cancer data (Siegel et al., 2019). Clear cell renal cell carcinoma (ccRCC) is the most common type of renal cell carcinoma, accounting for 80%. In clinical practice, 16% of ccRCCs were diagnosed with metastasis at the initial time, with a 5-year survival rate of 12% (Rydzanicz et al., 2013). Although the oncology research and surgical treatment of RCC have developed rapidly, the prognosis of RCC has not improved significantly. For RCC *in situ*, 20%–30% of patients relapse

after initial surgical treatment, and no treatment has been shown to reduce tumor recurrence and improve prognosis (Patel et al., 2016). In recent years, targeted agents have been shown to prolong survival and prognosis in patients with metastasis, but the median survival is still less than 3 years. Besides, in clinical practice, drug resistance and economic burden are still two prominent problems. Therefore, the study on the pathological mechanism and new therapeutic targets of RCC is still a challenging exploration.

The role of RNA in a variety of cellular processes has attracted a lot of attention and has become a rapidly developing field in the past decade. More than 100 chemically modified nucleotides have been found in different types of RNAs, such as mRNA, tRNA, rRNA, snRNA, etc. The modified RNA plays a crucial role in the post-transcriptional regulation of gene expression. In eukaryotes, m^6A is the most common form of mRNA modification, whose abundance accounts for 0.1%–0.4% of the total adenosine residues (Meyer et al., 2012; Zhu H et al., 2019). In general, m^6A is highly conserved between humans and mice, located in the 3′ terminal non-coding region, near the stop codon and long internal exons, and is closely associated with altered RNA stability, intracellular distribution, splicing, and translation (Csepany et al., 1990; Dominissini et al., 2012; Meyer et al., 2012). The cellular m^6A state is regulated by a set of genes called "writers" (METTL3, METTLL4, and WTAP), "erasers" (FTO and ALKBH5), and "readers" (YTHDF2, YTHDF2, YTHDF3, YTHDCL, and YTHDC2) (Fustin et al., 2013; Batista et al., 2014; Lin et al., 2016; Zhang et al., 2016; Bertero et al., 2018). The writer forms a multisubunit methyltransferase complex that could upregulate m^6A levels, while the eraser could downregulate m^6A levels, making this event reversible (Meyer et al., 2012; Dominissini et al., 2012).

In addition, it was demonstrated that METTL3, a major RNA N^6-adenosine methyltransferase, could promote hepatocellular carcinoma progression through YTHDF2-dependent post-transcriptional silencing of SOCS2. Although the m^6A modification of RNA has been reported to be associated with the occurrence of different types of cancer, the relationship between m^6A-related genes and RCC remains unclear (Aguilo et al., 2015; Shah et al., 2017; He et al., 2018; Panneerdoss et al., 2018; Zhu W et al., 2019; Fukumoto et al., 2019; Zhang et al., 2019). In this study, we used m^6A-RIP-seq and RNA-seq to research the m^6A modification profile and mRNA expression profile in RCC. In addition, we performed CRISPR-dCas13b fusion proteins to target methylation or demethylation of differentially methylated mRNAs (Wilson et al., 2020; Li et al., 2020a). This proved that abnormal RNA m^6A modifications could directly modulate gene expressions in RCC. Finally, we hope this study would facilitate further investigations of potential functions of m^6A modification in RCC pathogenesis.

MATERIALS AND METHODS

Patients and Samples

Five pairs of primary renal cell carcinoma (RCC) samples and adjacent peritumoral tissues with informed consent of the patients were gained in the Urology Department of Peking University First Hospital (PKUFH), Beijing, China. This study followed the Helsinki declaration and was approved by the Institutional Ethical Review Board of PKUFH. Samples were obtained immediately after surgical resection and stored in liquid nitrogen after rapid freezing for the subsequent RNA isolation.

RNA m^6A and mRNA Sequencing

MeRIP-seq and RNA-seq were performed by Cloudseq Biotech, Inc. (Shanghai, China) as described previously (Dominissini et al., 2013). Briefly, total RNAs were extracted from five pairs of tumor and adjacent peritumoral tissues using TRIzol (Invitrogen). Next, total RNAs were broken into almost 100 nt fragments and were incubated with anti-m^6A antibody (Manga) for 2 h at 4°C. The beads were prepared and incubated with the total RNAs for 2 h at 4°C. Then the mixture was washed, and the m^6A-bound RNA was purified with TE buffer. After purification, the library was constructed by Prep Kit (Illumina) on the Hiseq 3000.

Sequencing Data Analysis

After obtaining the sequencing data of control and IP samples, the read segment data should be preprocessed (such as filtering the read segment with poor sequencing quality), and then all the read segment sequence mapping of the two samples should be positioned on the reference genome, which is the basis of subsequent data processing and analysis. Then there were many read segments captured by methylation sites in the IP samples, which would be mapped to the reference genome to form a reading segment enrichment region or a "peak" near the methylation sites. Therefore, the methylation enrichment point detection algorithm is called the peak calling algorithm. The m^6A methylated peaks among the transcripts were identified by MACS (Zhang et al., 2008), and metagene m^6A distribution was researched by MetaPlotR (Olarerin-George and Jaffrey, 2017). The DMGs were identified by diffReps (Shen et al., 2013). To explore the DMGs and DEGs from MeRIP-seq and RNA-seq, the Gene Ontology (GO) analysis and KEGG pathway enrichment analysis were performed.

M^6A-IP-qPCR and Reverse Transcription-qPCR

Ten genes with differentially methylated sites were verified by reverse transcription (RT)-qPCR according to m^6A-seq. A small number of fragmented RNA was tested as the input control, while the rested RNA was incubated with anti-m^6A antibody-coupled beads. Then the immunoprecipitation complex was eluted from the beads. In the end, RT-qPCR was performed on the input control and m^6A -IP samples with gene-specific primers. Primers are listed in **Supplementary Table S1**.

Cell Line Culture and Plasmid Transfection

In this study, RCC cell lines, including 786-O, ACHN, and OSRC, and human kidney proximal tubular epithelial cells HK-2 were used. These cell lines were purchased from the American Type Culture Collection (ATCC, Manassas, VA, United States) and National Infrastructure of Cell Line Resource, China. Cell lines were cultured in RPMI 1640 or DMEM with 10% fetal bovine serum (Invitrogen, Carlsbad, CA, United States) and incubated in a 5% CO_2 environment at 37°C. The plasmids were transfected with Lipo3000 (Invitrogen) according to the protocol of the manufacturer, and the dosage of plasmids was 1 μg.

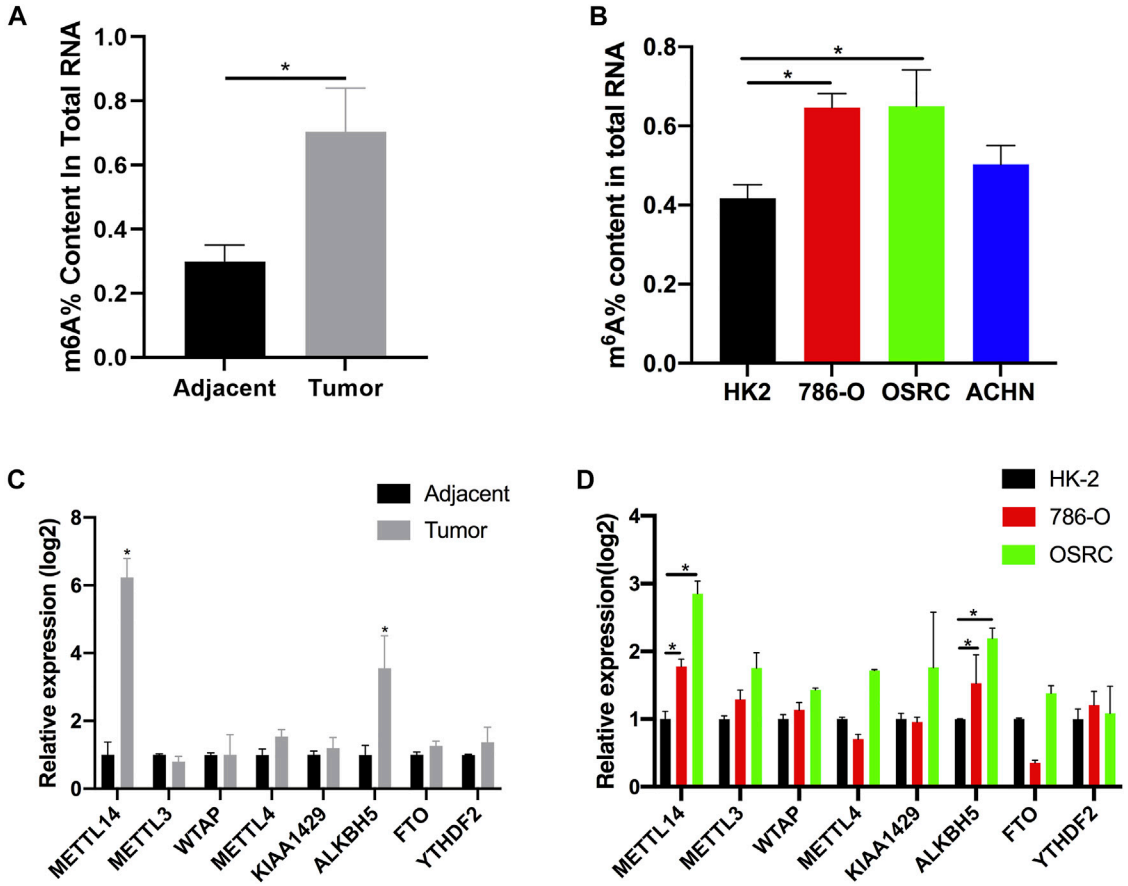


FIGURE 1 | The N^6-methyladenosine (m^{6}A) contents and modification enzymes expression in renal cell carcinoma (RCC) tissues and RCC cells. **(A)** The m^{6}A contents of total RNAs in tumor and adjacent tissues (n = 5). **(B)** The m^{6}A contents of total RNAs in normal renal cell line (HK-2) and RCC cell lines (786-O, OSRC, and ACHN). **(C)** The mRNA expression levels of m^{6}A modification enzymes in RCC and adjacent tissues (n = 5). **(D)** The mRNA expression levels of m^{6}A modification enzymes in normal renal cell line (HK-2) and RCC cell lines (786-O and OSRC). *p < 0.05.

Statistical Analysis

All statistical analyses were achieved and visualized using RStudio (version1.2.1335, Boston, MA, United States), GSEA (version4.0, UC San Diego and Broad Institute, United States) 23, Medcalc (version16.8, Ostend, Belgium), and GraphPad Prism (version 8.0, GraphPad, Inc., La Jolla, CA, United States). A two-tailed p < 0.05 was considered statistically significant.

RESULTS

The m^{6}A Content was Changed in Renal Cancer Cell

Recently, it has been found that RNA methylation could promote the progression of RCC (Dai et al., 2018). In order to explore the potential role of m^{6}A modification in RCC, we measured the m^{6}A content of total RNA in five pairs of tumor tissues and adjacent tissues using the m^{6}A quantitative kit. We found the content was higher in tumor tissues than those in their corresponding normal tissues (**Figure 1A**). At the same time, we also examined the m^{6}A contents of normal renal cell line HK-2 and renal cancer cell lines (786-O, OSRC, and ACHN), and found that m^{6}A contents were also increased in RCC cells compared with normal cells (**Figure 1B**). m^{6}A methyltransferase or demethylase could catalyze the modification of m^{6}A. Therefore, we speculated that the abnormal m^{6}A content might be caused by the abnormal expression of m^{6}A methyltransferase or demethylase in RCC. In addition, many proteins have been found to be associated with m^{6}A modification (Schwartz et al., 2014). To investigate our assumption, we examined the mRNA levels of eight genes associated with m^{6}A modification in tissues and cell lines by RT-qPCR. Finally, the results showed that the mRNA expression levels of METL14 and ALKBH5 were increased in tumor tissues and RCC cell lines, while the mRNA expression levels of other genes were not significantly changed (**Figures 1C,D**).

The mRNA m^{6}A Modification was Dynamic and Differed Between Paired Renal Tumor Tissues and Adjacent Tissues

In order to investigate whether m^{6}A modification could promote the progression of RCC, we used MeRIP-seq (Dominissini et al., 2013) to detect the differences of m^{6}A level in the paired renal

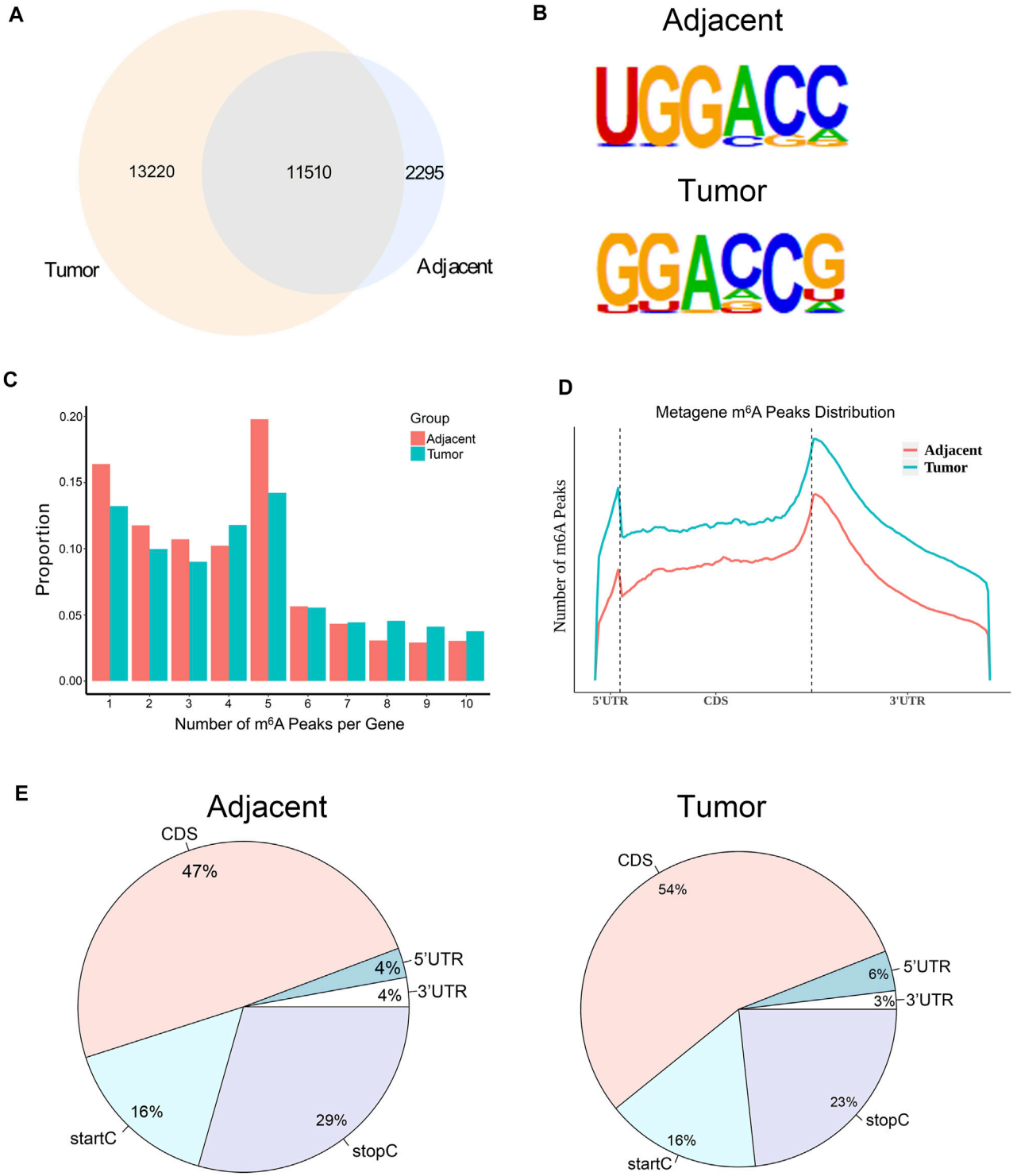


FIGURE 2 | The mRNA m^{6}A profile of five paired renal tumor tissues and adjacent tissues. **(A)** The Venn diagram showing the numbers of m^{6}A peaks in tumor and adjacent tissues. **(B)** The sequence motif of m^{6}A-containing peak regions in tumor and adjacent tissues, respectively. **(C)** The distribution of m^{6}A modifications in the human transcriptome. **(D)** The metagene profiles of transcripts peaks in RCC tissues and adjacent tissues. **(E)** The proportion of m^{6}A peaks located at transcripts. startC, start codon; stopC, stop codon.

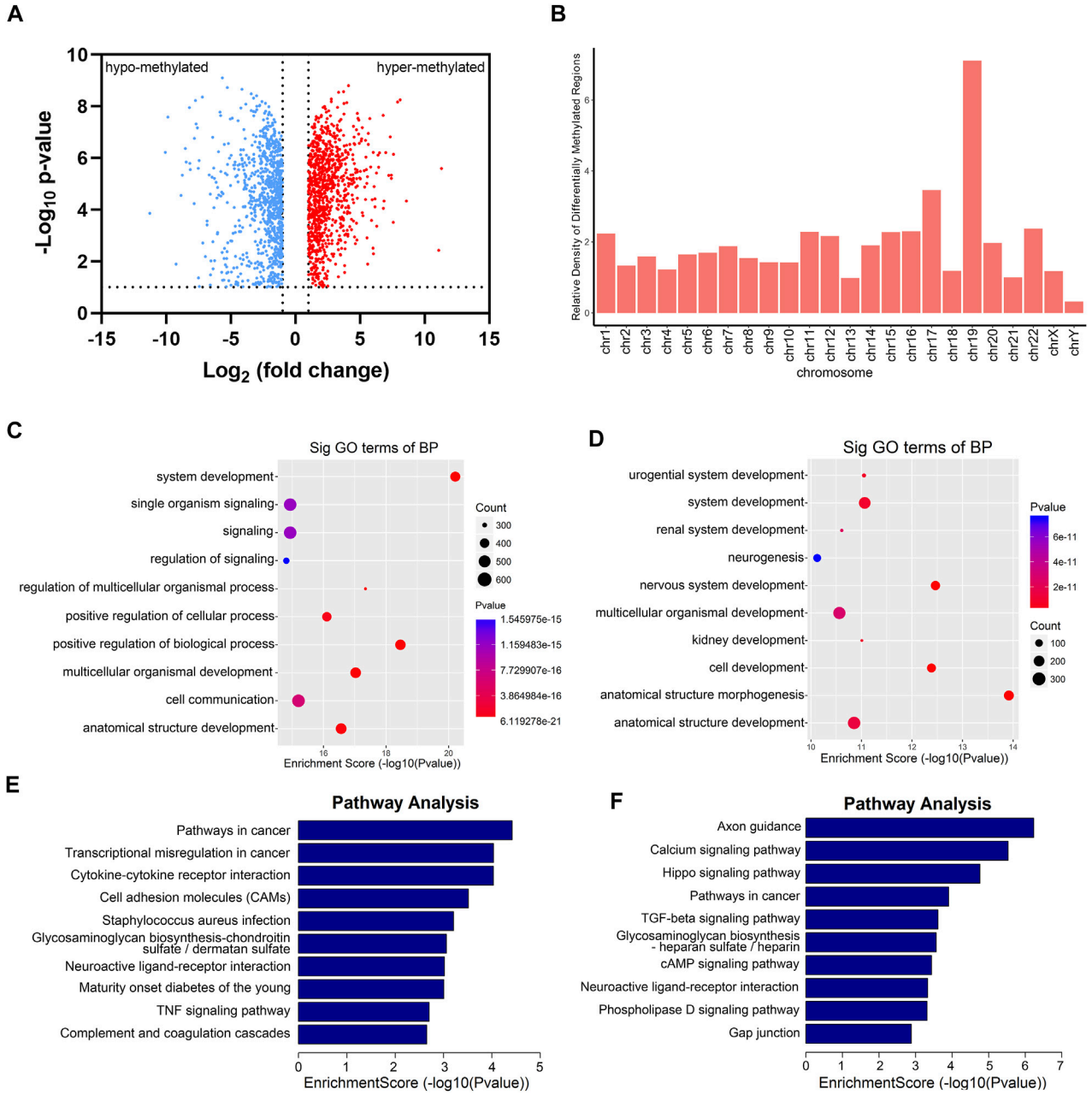


FIGURE 3 | Gene Ontology (GO) and KEGG analysis of DMGs between five paired renal tumor and adjacent tissues. **(A)** Volcano blot of differentially methylated mRNAs (DMGs) in tumor compared with adjacent tissues. **(B)** The container of differentially methylated m6A sites harbored by different chromosomes. **(C)** GO analysis of hypermethylated genes in RCC tumor tissues. **(D)** GO analysis of hypomethylated genes in RCC tumor tissues. **(E)** KEGG analysis of hypermethylated genes in RCC tumor tissues. **(F)** KEGG analysis of hypomethylated genes in RCC tumor tissues.

tumor tissues and adjacent tissues. MeRIP-seq analysis revealed that there were 13,805 m^6A peaks among 5,568 coding gene transcripts (mRNAs) in adjacent tissues, and 24,730 m^6A peaks among 6,866 mRNAs in tumor tissues. Of these, 11,510 peaks were overlapped between the adjacent tissues and tumor tissues (**Figure 2A**). The low overlapping m^6A peaks of mRNAs suggested that there were differences in the m^6A patterns between the two groups. To study whether the m^6A peaks had a conserved motif (Wei et al., 1976; Schibler et al., 1977), the m^6A peaks identified from the MeRIP-seq were analyzed by the HOMER motif software (Heinz et al., 2010). The results showed that there was a difference of m^6A motif between tumor tissues and adjacent tissues, while their motifs were similar to those previously reported (Fu et al., 2014; Chen et al., 2015) (**Figure 2B**). Besides, we examined the distribution of m^6A modifications in the human

transcriptome. We found that 70% of methylated sequences in the adjacent group (60% of methylated transcripts in the tumor group) contained fewer than five m^{6}A sites, while fewer transcripts contained six or more sites (**Figure 2C**).

In order to determine the priority position of m^{6}A in transcripts, we then studied the metagene profiles of the peaks in RCC tissues and adjacent normal tissues. We found that most m^{6}A peaks were located at the end of the 5′UTRs and start of the 3′UTRs (**Figure 2D**). At the same time, we found that the proportion of m^{6}A peaks in CDS was the highest and that in UTRs was the lowest in both tissues (**Figure 2E**). These results of the m^{6}A modification distributions were similar to those reported previously (Dominissini et al., 2012; Meyer et al., 2012).

mRNA Containing Differential m^{6}A Sites was Enriched in Kidney Development and Cancer-Related Signaling Pathways

Totally, we had identified 4,404 differential m^{6}A sites (DMMSs) within 1,877 nuclear mRNAs, of which 43% (1,887/4,404) were significantly downmethylated sites (tumor vs. adjacent). Compared with the adjacent group, we found 923 significantly hypomethylated coding genes and 954 significantly hypermethylated coding genes in the tumor group (**Figure 3A**). Besides, we analyzed the relative density of differential m^{6}A modification sites on chromosomes; the top three chromosomes were 19, 17, and 22 (**Figure 3B**). To explore the functions of m^{6}A in RCC, differentially methylated genes were performed for Gene Ontology enrichment analysis and Kyoto Encyclopedia of Genes and Genomes pathway analysis. The results indicated that differentially methylated genes in the tumor group were mostly enriched in many system development-associated processes (**Figures 3C,D**). Furthermore, differentially methylated genes were found to be significantly involved in cancer pathways, such as the transcriptional misregulation and TGF-β signaling pathway (**Figures 3E,F**).

Differentially Expressed RNAs were Involved in Important Biological Pathways

From the RNA-seq data, we found that 1,469 genes were upregulated in tumor tissues, and 1,402 were upregulated in adjacent tissues (**Figures 4A,B**). Next, we performed Gene Ontology enrichment analysis and Kyoto Encyclopedia of Genes and Genomes pathway analysis to investigate the function of the differentially expressed RNAs. We found that the differentially expressed RNAs were mainly related to single-organism process and urogenital system development (**Figures 4C,D**). Moreover, pathway analysis showed that systemic lupus erythematosus, tight junction, and MAPK signaling pathway were significantly changed in tumor samples (**Figures 4E,F**).

Conjoint Analysis of MeRIP-Seq and RNA-Seq

After conjoint analysis of meRIP-seq and RNA-seq, 369 hyper-up genes, 372 hypo-down genes, 14 hyper-down genes, and two hypo-up genes were found in RCC tissues compared with adjacent tissues (**Figure 5A**). This result promoted us to investigate the cancer-related genes in the RCC. For instance, cell division cycle-associated 2 (CDCA2) promotes the proliferation and development of colon cancer, within which m^{6}A was hypermethylated (tumor vs. adjacent) near the stop codon (**Figure 5B**) (Feng et al., 2019), and the m^{6}A peaks of choline-O-acetyltransferase (CHAT) (Yokomori et al., 1983; Pahud et al., 2001) were enriched around the 5′UTR of CHAT in tumors (**Figure 5C**), while mucin 15 (MUC15), a significantly hypermethylated peak enriched in coding sequence (CDS) was shown in the adjacent groups (**Figure 5D**), which had been reported to promote the progression of cancer (Zhang et al., 2020). It was reported that HRG could prevent the cancer development by inducing macrophage polarization (Rolny et al., 2011), and its m^{6}A peak was enriched around the beginning of the 3′UTR in the adjacent groups in our study (**Figure 5E**).

To further confirm the results of our m^{6}A-seq data, we analyzed gene-specific m^{6}A-IP qPCR assays for several hypermethylated genes (CDCA2, CHAT, TGFA, PLOD2, and EGLN1) and hypomethylated genes (MUC15, HRG, HIF1A, ZNF677, and PAK6), which might participate in RCC development. We verified the same m^{6}A level changes of 10 genes, confirming the validity of the meRIP-seq results (**Figure 5F**). Next, mRNA levels of the abovementioned 10 genes were measured in the five pairs of adjacent and tumor samples by RT-qPCR (**Figure 5G**). Results showed a similar tendency of m^{6}A-methylated levels and mRNA expressions in the two groups, which suggested a relationship between m^{6}A mRNA methylation and gene transcription.

dCas13b-ALKBH5 Fusion Protein Induced Demethylation of m^{6}A-Modified mRNAs *in vitro*

To study mRNA modifications *in vitro*, we performed dCas13b-ALKBH5 fusion protein for targeted mRNA demethylation (Li et al., 2020a). We designed two guide RNAs targeting the 3′UTR or CDS of CDCA2 mRNA, and another two gRNAs targeting 5′UTR or CDS of CHAT mRNA. These two mRNAs were hypermethylated in RCC, then all four gRNAs were transfected in 786-O RCC cell line with dCas13b-ALKBH5 fusion protein, respectively (**Figure 6A**). The results of m^{6}A-IP qPCR revealed that the m^{6}A levels of the targeted site were decreased after transfecting with dCas13b-ALKBH5 and gRNAs, compared with NT-gRNA (**Figures 6B,C**). Intriguingly, the mRNA expression levels of the targeted CDCA2 and CHAT were decreased after transfecting gRNAs (**Figures 6D,E**). YTHDF1, which is responsible for binding to the translational machinery and RNA translation activation, might be inhibited by dCas13b-ALKBH5 (Wang et al., 2015). Collectively, these results suggested that dCas13b-ALKBH5 could demethylate m^{6}A levels of targeted hypermethylated mRNAs.

dCas13b-M3M14 Fusion Protein Induced Methylation of m^{6}A-Modified mRNAs *in vitro*

In addition, we constructed the dCas13b-M3M14 fusion protein (Wilson et al., 2020) to promote m^{6}A modifications in RCC cells.

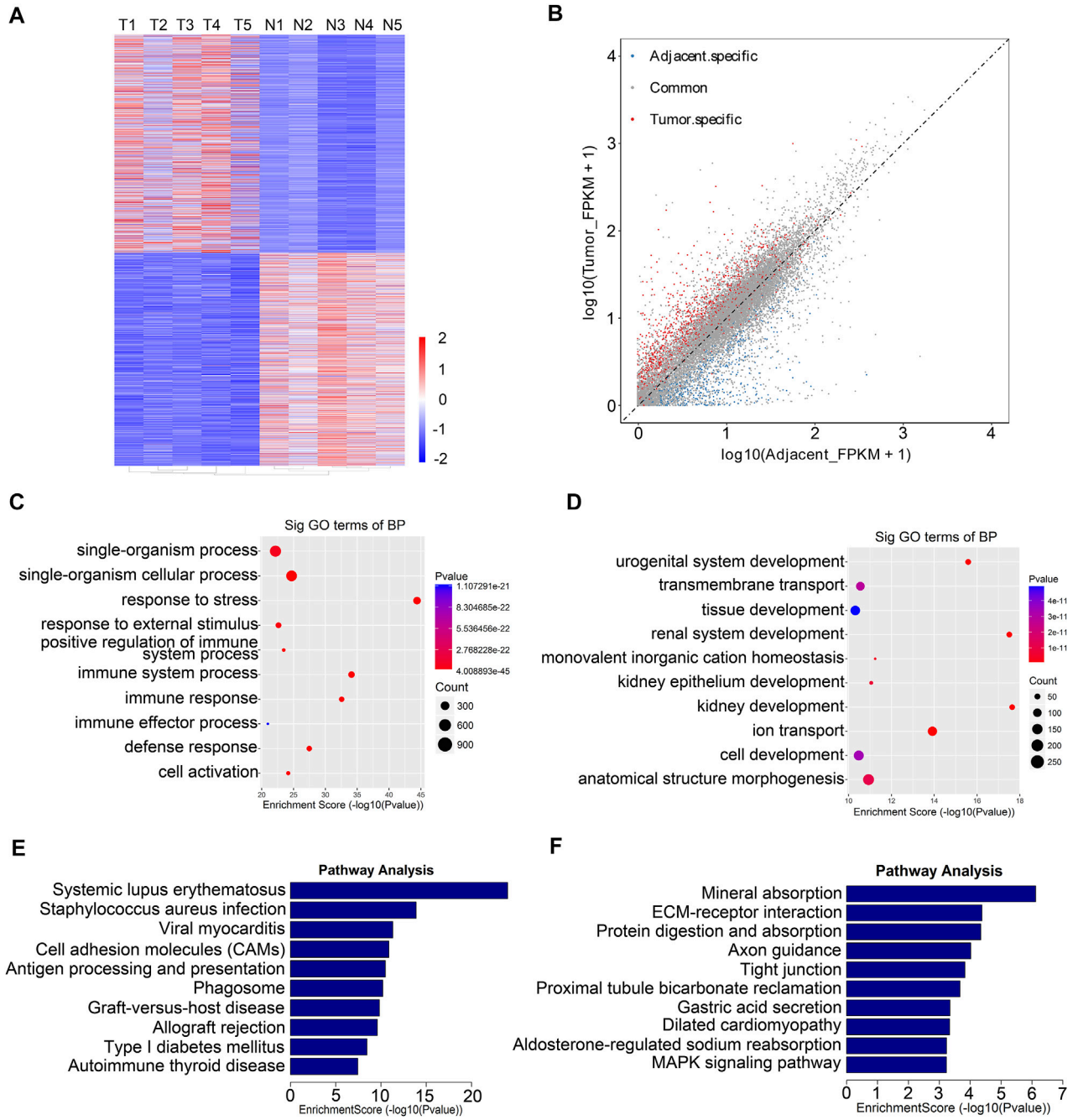


FIGURE 4 | Functional analysis of differentially expressed genes between five paired renal tumor and adjacent tissues. **(A)** Heatmap of differentially expressed mRNAs in renal tumor tissues and adjacent tissues (T, tumor group; N, adjacent normal group). **(B)** Scatter plots showing the tissue specific DEGs (fold changes ≥2 and $p < 0.05$). **(C)** GO analysis of upregulated genes in RCC tumor tissues. **(D)** GO analysis of downregulated genes in RCC tumor tissues. **(E)** KEGG analysis of upregulated genes in RCC tumor tissues. **(F)** KEGG analysis of downregulated genes in RCC tumor tissues.

The mRNAs of HRG and MUC15 were hypomethylated in RCC and were targeted by two gRNAs at different positions (**Figure 7A**). First, we measured the m^{6}A levels of HRG and MUC15 in 786-O RCC cell lines transfected with dCas13b-M3M14 and gRNAs or NT-gRNAs. The results showed that dCas13b-M3M14 significantly increased the m^{6}A levels of HRG and MUC15, suggesting that gRNAs could efficiently recognize their targeted mRNAs (**Figures 7B,C**). We then verified the effect of dCas13b-M3M14 fusion protein (gRNAs for HRG and MUC15) on the mRNA expressions of HRG and MUC15. The results of RT-qPCR showed that dCas13b-M3M14 fusion protein transfected with gRNAs

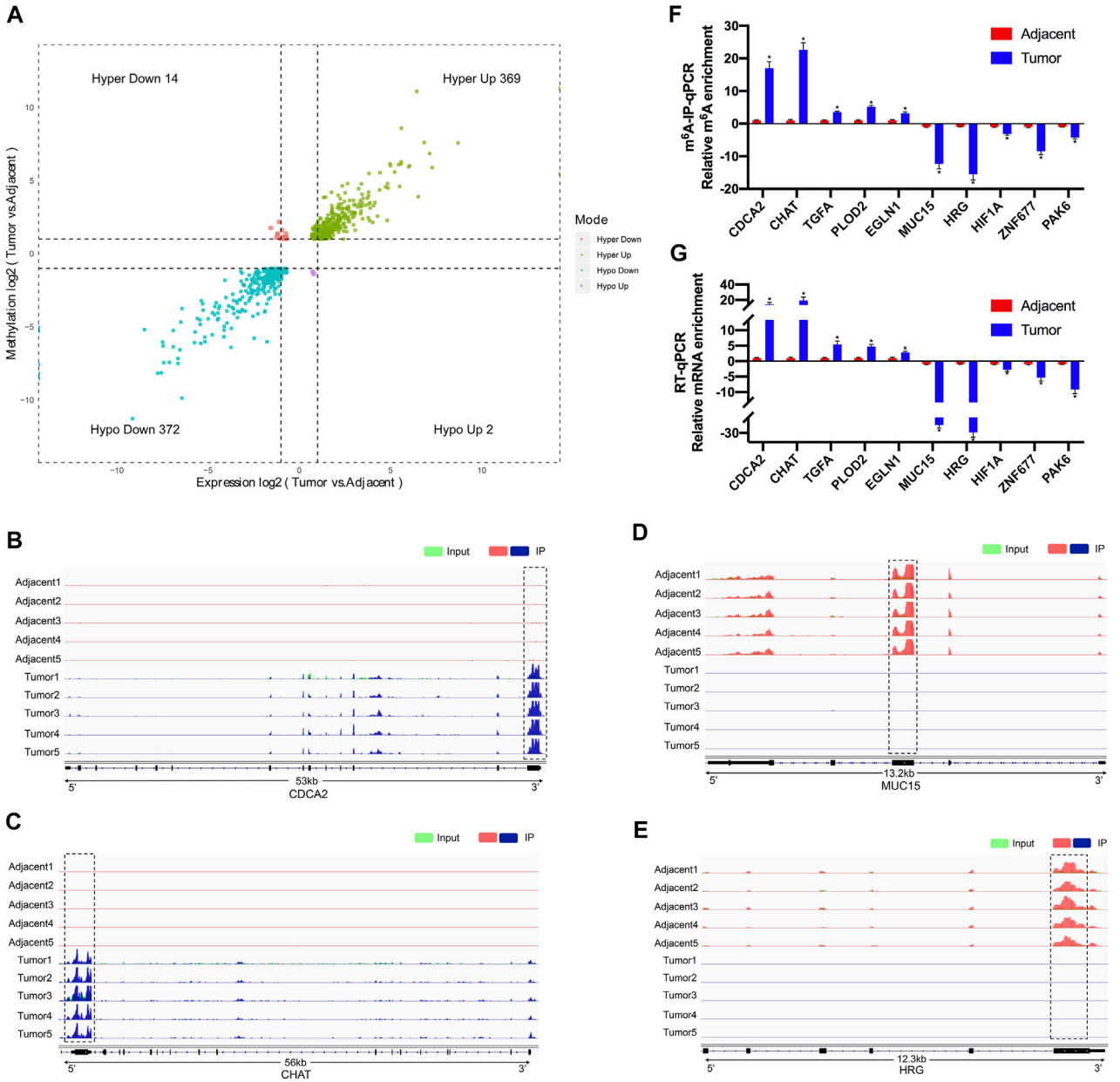


FIGURE 5 | Conjoint analysis of differentially methylated genes and differentially expressed genes between five paired renal tumor and adjacent tissues. **(A)** Four quadrant diagram showing the differentially methylated genes and differentially expressed genes. **(B,C)** Integrative Genome Viewer (IGV) software showing the m^{6}A peaks in hypermethylated and upregulated gene cell division cycle-associated 2 (CDCA2) **(B)** and choline-O-acetyltransferase (CHAT) **(C)**. **(D,E)** Integrative Genome Viewer (IGV) software showed the m^{6}A peaks in hypomethylated and downregulated genes mucin 15 (MUC15) **(D)** and histidine-rich glycoprotein (HRG) **(E)**. **(F)** Validations of the m^{6}A enrichment of five hypermethylated genes (CDCA2, CHAT, TGFA, PLOD2, and EGLN1) and five hypomethylated genes (MUC15, HRG, HIF1A, ZNF677, and PAK6) by m^{6}A-immunoprecipitation (IP)-qPCR (n = 5). **(G)** Validations of the mRNA expression level of five upregulated genes (CDCA2, CHAT, TGFA, PLOD2, and EGLN1) and five downregulated genes (MUC15, HRG, HIF1A, ZNF677, and PAK6) by RT-qPCR (n = 5). *p < 0.05.

significantly upregulated the mRNA levels of their targeted mRNAs (**Figures 7D,E**). This might be due to dCas13b-M3M14 fusion protein increasing the binding of targeted mRNA and YTHDF1. Finally, this dCas13b-M3M14 fusion protein could mediate efficient installation of m^{6}A in endogenous RNA transcripts.

DISCUSSION

m^{6}A is identified as a dynamic and reversible RNA modification in eukaryotes, due to the "writer" (methyltransferase) and "eraser" (demethylase) proteins. It has been reported that m^{6}A modification could take part in many cellular activities

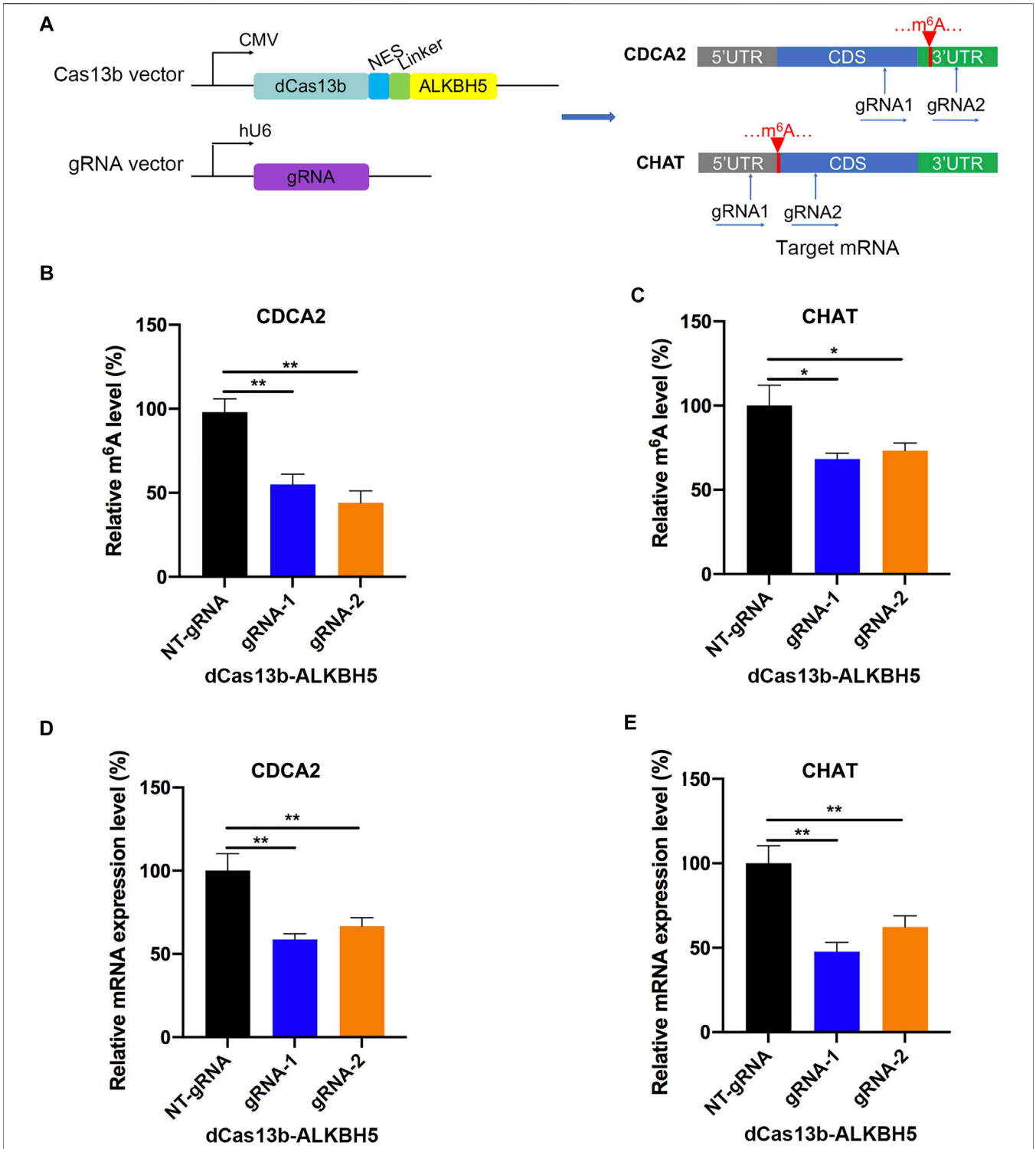


FIGURE 6 | dCas13b-ALKBH5 induces demethylation of m^{6}A hypermethylated mRNA in RCC cells. **(A)** The design of dCas13b-ALKBH5 fusion protein and sgRNAs targeted with CDCA2 and CHAT mRNA m^{6}A sites. **(B,C)** The m^{6}A level of CDCA2 **(B)** and CHAT **(C)** mRNA in RCC cell line 786-O after transfected with dCas13b-ALKBH5 and gRNAs. **(D,E)** The mRNA levels of CDCA2 **(D)** and CHAT **(E)** in 786-O cell after transfection with dCas13b-ALKBH5 and gRNAs. $^{**}p < 0.01$.

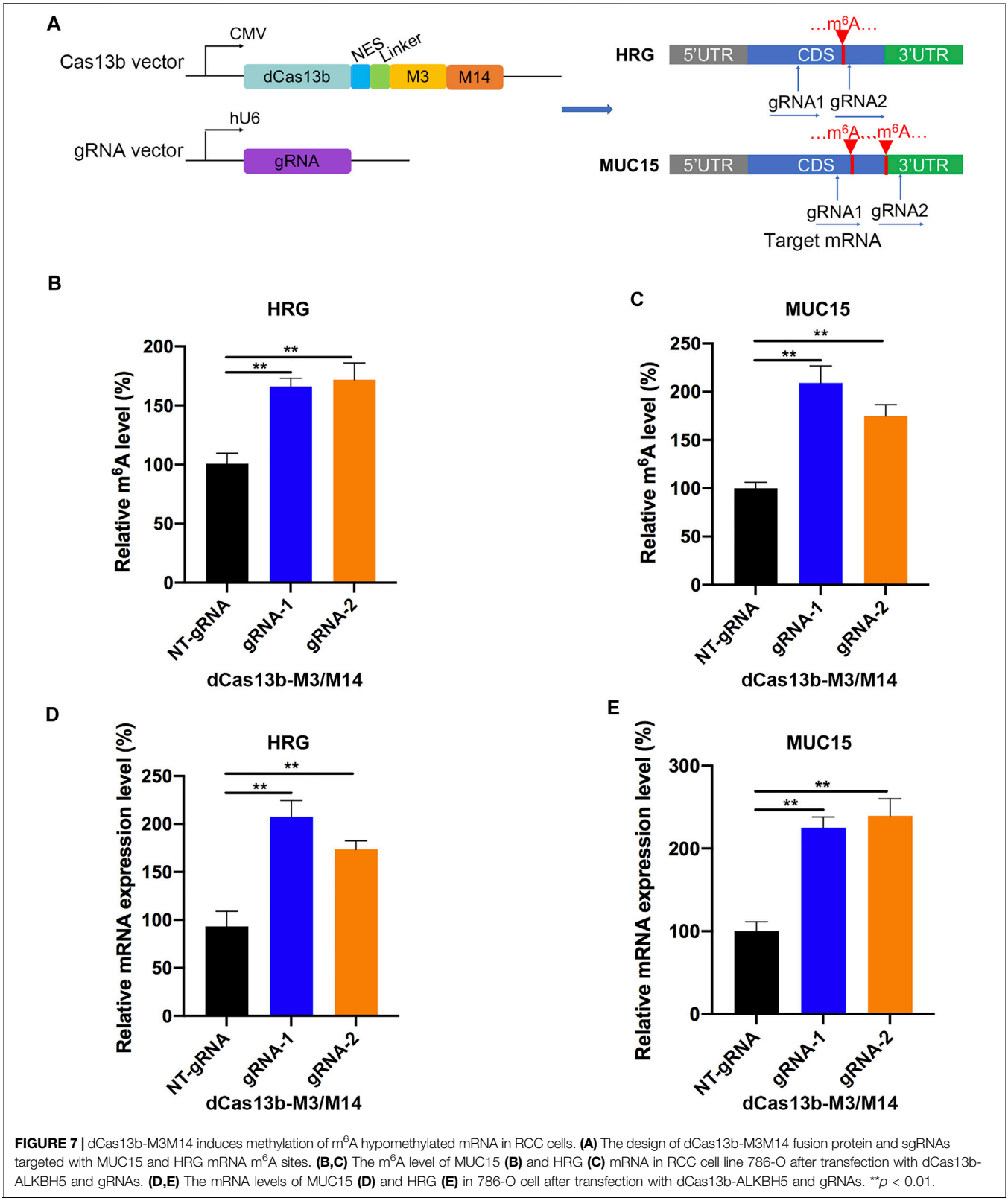


FIGURE 7 | dCas13b-M3M14 induces methylation of m^{6}A hypomethylated mRNA in RCC cells. **(A)** The design of dCas13b-M3M14 fusion protein and sgRNAs targeted with MUC15 and HRG mRNA m^{6}A sites. **(B,C)** The m^{6}A level of MUC15 **(B)** and HRG **(C)** mRNA in RCC cell line 786-O after transfection with dCas13b-ALKBH5 and gRNAs. **(D,E)** The mRNA levels of MUC15 **(D)** and HRG **(E)** in 786-O cell after transfection with dCas13b-ALKBH5 and gRNAs. $^{**}p < 0.01$.

and reaction, including heat shock (Zhou et al., 2015), ultraviolet light (Robinson et al., 2019), hypoxic stress (Fry et al., 2017), and oxidative stress (Anders et al., 2018). Moreover, there were numbers of evidences confirming that m^{6}A modification could promote the development of tumors. In this study, we identified many differentially methylated genes

in RCC samples and tumor-adjacent normal tissues based on meRIP-seq technology, analyzed and validated gene expression and cancer-related pathways modulated by abnormal m^{6}A RNA modifications.

We figured out that m^{6}A modification in tumor tissues and normal tissues mainly occurred in the GGACC motif, which was similar to the previous data, and the m^{6}A peaks of transcripts were mainly located at the CDS site. Almost 80% of the methylated genes had one to five m^{6}A sites, and others contained over eight m^{6}A sites in mRNAs. In the current study, differentially methylated mRNAs between tumor and normal tissues were shown to be involved in many important biological pathways. As observed, the gene function analyses of DMMSs showed that the hyper- and hypomethylated genes in the tumor group were significantly enriched in many phylogenetic processes, such as multicellular organismal development and kidney development, and they were also involved in the cancer pathways, such as the transcriptional misregulation in cancer and TGF-β signaling pathway, which supported the importance of m^{6}A in tumorigenesis. It was reported that m^{6}A could play an important role in carcinogenesis and the development of gastric cancer, and the genes with higher m6A levels were mainly enriched in transcriptional misregulation in carcinogenesis pathways, whereas the genes with decreased methylation mainly regulate digestion and absorption of protein (Sang et al., 2020). In addition, the dysregulated expression of m^{6}A was also involved in transcriptional misregulation in cancer and other malignancy-related pathways including primary immunodeficiency, regulation of autophagy, and response to oxidative stress (Zheng et al., 2020). It was also reported that the expression of m^{6}A methyltransferase METTL3 and m^{6}A modification were increased during TGF-β-induced epithelial–mesenchymal transition in A549 and LC2/ad lung cancer cells, and mechanistic investigations revealing that METTL3 could be indispensable for TGF-β-induced epithelial–mesenchymal transition of lung cancer cells through the regulation of JUNB (Wanna-Udom et al., 2020). Moreover, m^{6}A modification was involved in the epithelial–mesenchymal transition of cancer cells by regulating the expression and secretion of TGF-β1 (Li et al., 2020b). These studies revealed that differential m^{6}A modifications were involved in important biological pathways, which were consistent with our study.

Combined analysis of m^{6}A-seq and mRNA-seq data uncovered 369 hyper-up genes and 372 hypo-down genes in tumor tissues compared with adjacent normal tissues, which might play critical roles in the RCC development. Moreover, some of the genes have been reported to facilitate tumor growth and metastasis in different types of cancers. To further investigate mRNA m^{6}A modification of these specific genes, we used CRISPR-dCas13b fusion proteins to regulate the methylation levels of mRNA. For instance, the methylation levels of CDCA2 and CHAT were found to be approximately 20 times higher in the tumor group than that in the control group. Then, we applied dCas13b-ALKBH5 fusion protein to induce the demethylation of CDCA2 and CHAT mRNAs in RCC cell lines, and found that the m^{6}A levels and mRNA expression levels were significantly reduced after transfecting with their targeted gRNAs. Furthermore, another two hypomethylated genes, MUC15 and HRG, exhibited downregulated mRNA levels in the tumor group. Therefore, we performed dCas13b-M3M14 fusion protein combined with two gRNAs to methylate mRNA m^{6}A level of MUC15 and HRG in RCC cell line. The data showed hypermethylated m^{6}A levels and upregulated mRNA levels of MUC15 and HRG, which were in line with our expectations.

However, the study was still partially flawed. We cannot avoid the potential selection bias, since the m^{6}A-RIP-seq and RNA-seq were based on five paired ccRCC tumor tissues and adjacent tissues. Further functional experiments and mechanism explorations of m^{6}A modification in ccRCC should be performed. Despite the defects listed above, the presented findings still provided a link between mRNA m^{6}A modifications and renal tumorigenesis, which could be expected to be a new target for gene targeted therapy of renal cell carcinoma.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are publicly available. This data can be found here: https://www.ncbi.nlm.nih.gov/bioproject/PRJNA719065.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Institutional Ethical Review Board of PKUFH. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

QZ, XY, and LY conceptualized and designed the study. YG collected and analyzed the sequencing data. AL constructed plasmids and performed validation experiments. YG, JL, and XW processed tissue samples. ZZ and QL performed the data interpretation. All authors contributed to the writing of the manuscript and for the final approval of the manuscript.

FUNDING

This work was supported by the National Natural Science Foundation of China (Grant Nos. 82072826 and 81872088 to QZ; Nos. 62076007 to XY; Nos. 82103239 to YG).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.795611/full#supplementary-material

REFERENCES

Aguilo, F., Zhang, F., Sancho, A., Fidalgo, M., Di Cecilia, S., Vashisht, A., et al. (2015). Coordination of m(6)A mRNA Methylation and Gene Transcription by ZFP217 Regulates Pluripotency and Reprogramming. *Cell Stem Cell* 17, 689–704. doi:10.1016/j.stem.2015.09.005

Anders, M., Chelysheva, I., Goebel, I., Trenkner, T., Zhou, J., Mao, Y., et al. (2018). Dynamic m(6)A Methylation Facilitates mRNA Triaging to Stress Granules. *Life Sci. Alliance* 1, e201800113. doi:10.26508/lsa.201800113

Batista, P. J., Molinie, B., Wang, J., Qu, K., Zhang, J., Li, L., et al. (2014). m6A RNA Modification Controls Cell Fate Transition in Mammalian Embryonic Stem CellsA RNA Modification Controls Cell Fate Transition in Mammalian Embryonic Stem Cells. *Cell Stem Cell* 15, 707–719. doi:10.1016/j.stem.2014.09.019

Bertero, A., Brown, S., Madrigal, P., Osnato, A., Ortmann, D., Yiangou, L., et al. (2018). The SMAD2/3 Interactome Reveals that TGFβ Controls m(6)A mRNA Methylation in Pluripotency. *Nature* 555, 256–259. doi:10.1038/nature25784

Chen, T., Hao, Y.-J., Zhang, Y., Li, M.-M., Wang, M., Han, W., et al. (2015). m(6)A RNA Methylation Is Regulated by MicroRNAs and Promotes Reprogramming to PluripotencyA RNA Methylation Is Regulated by microRNAs and Promotes Reprogramming to Pluripotency. *Cell Stem Cell* 16, 289–301. doi:10.1016/j.stem.2015.01.016

Csepany, T., Lin, A., Baldick, C. J., and Beemon, K. (1990). Sequence Specificity of mRNA N6-Adenosine Methyltransferase. *J. Biol. Chem.* 265, 20117–20122. doi:10.1016/s0021-9258(17)30477-5

Dai, D., Wang, H., Zhu, L., Jin, H., and Wang, X. (2018). N6-methyladenosine Links RNA Metabolism to Cancer Progression. *Cell Death Dis.* 9, 124. doi:10.1038/s41419-017-0129-x

Dominissini, D., Moshitch-Moshkovitz, S., Salmon-Divon, M., Amariglio, N., and Rechavi, G. (2013). Transcriptome-wide Mapping of N(6)-Methyladenosine by m(6)A-Seq Based on Immunocapturing and Massively Parallel Sequencing. *Nat. Protoc.* 8, 176–189. doi:10.1038/nprot.2012.148

Dominissini, D., Moshitch-Moshkovitz, S., Schwartz, S., Salmon-Divon, M., Ungar, L., Osenberg, S., et al. (2012). Topology of the Human and Mouse m6A RNA Methylomes Revealed by m6A-Seq. *Nature* 485, 201–206. doi:10.1038/nature11112

Feng, Y., Qian, W., Zhang, Y., Peng, W., Li, J., Gu, Q., et al. (2019). CDCA2 Promotes the Proliferation of Colorectal Cancer Cells by Activating the AKT/CCND1 Pathway *In Vitro* and *In Vivo*. *BMC Cancer* 19, 576. doi:10.1186/s12885-019-5793-z

Fry, N. J., Law, B. A., Ilkayeva, O. R., Holley, C. L., and Mansfield, K. D. (2017). N(6)-methyladenosine Is Required for the Hypoxic Stabilization of Specific mRNAs. *RNA* 23, 1444–1455. doi:10.1261/rna.061044.117

Fu, Y., Dominissini, D., Rechavi, G., and He, C. (2014). Gene Expression Regulation Mediated through Reversible m(6)A RNA Methylation. *Nat. Rev. Genet.* 15, 293–306. doi:10.1038/nrg3724

Fukumoto, T., Zhu, H., Nacarelli, T., Karakashev, S., Fatkhutdinov, N., Wu, S., et al. (2019). N(6)-Methylation of Adenosine of FZD10 mRNA Contributes to PARP Inhibitor Resistance. *Cancer Res.* 79, 2812–2820. doi:10.1158/0008-5472.CAN-18-3592

Fustin, J.-M., Doi, M., Yamaguchi, Y., Hida, H., Nishimura, S., Yoshida, M., et al. (2013). RNA-methylation-dependent RNA Processing Controls the Speed of the Circadian Clock. *Cell* 155, 793–806. doi:10.1016/j.cell.2013.10.026

He, Y., Hu, H., Wang, Y., Yuan, H., Lu, Z., Wu, P., et al. (2018). ALKBH5 Inhibits Pancreatic Cancer Motility by Decreasing Long Non-coding RNA KCNK15-AS1 Methylation. *Cell Physiol. Biochem.* 48, 838–846. doi:10.1159/000491915

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. Cel.* 38, 576–589. doi:10.1016/j.molcel.2010.05.004

Li, J., Chen, F., Peng, Y., Lv, Z., Lin, X., Chen, Z., et al. (2020). N6-Methyladenosine Regulates the Expression and Secretion of TGFβ1 to Affect the Epithelial-Mesenchymal Transition of Cancer Cells. *Cells* 9, 296. doi:10.3390/cells9020296

Li, J., Chen, Z., Chen, F., Xie, G., Ling, Y., Peng, Y., et al. (2020). Targeted mRNA Demethylation Using an Engineered dCas13b-ALKBH5 Fusion Protein. *Nucleic Acids Res.* 48, 5684–5694. doi:10.1093/nar/gkaa269

Lin, S., Choe, J., Du, P., Triboulet, R., and Gregory, R. I. (2016). The m(6)A Methyltransferase METTL3 Promotes Translation in Human Cancer Cells. *Mol. Cel.* 62, 335–345. doi:10.1016/j.molcel.2016.03.021

Meyer, K. D., Saletore, Y., Zumbo, P., Elemento, O., Mason, C. E., and Jaffrey, S. R. (2012). Comprehensive Analysis of mRNA Methylation Reveals Enrichment in 3′ UTRs and Near Stop Codons. *Cell* 149, 1635–1646. doi:10.1016/j.cell.2012.05.003

Olarerin-George, A. O., and Jaffrey, S. R. (2017). MetaPlotR: a Perl/R Pipeline for Plotting Metagenes of Nucleotide Modifications and Other Transcriptomic Sites. *Bioinformatics* 33, 1563–1564. doi:10.1093/bioinformatics/btx002

Pahud, G., Bontron, S., and Eder-Colli, L. (2001). Modulation of Choline Acetyltransferase Synthesis by Okadaic Acid, a Phosphatase Inhibitor, and KN-62, a CaM Kinase Inhibitor, in NS-20Y Neuroblastoma. *Neurochem. Int.* 38, 75–82. doi:10.1016/s0197-0186(00)00064-4

Panneerdoss, S., Eedunuri, V. K., Yadav, P., Timilsina, S., Rajamanickam, S., Viswanadhapalli, S., et al. (2018). Cross-talk Among Writers, Readers, and Erasers of m(6)A Regulates Cancer Growth and Progression. *Sci. Adv.* 4, eaar8263. doi:10.1126/sciadv.aar8263

Patel, D. N., Figlin, R. A., and Kim, H. L. (2016). Adjuvant Treatment for Renal Cell Carcinoma: Do We Finally Have a Major Breakthrough? *Clin. Adv. Hematol. Oncol.* 14, 907–914.

Robinson, M., Shah, P., Cui, Y.-H., and He, Y.-Y. (2019). The Role of Dynamic m(6)A RNA Methylation in Photobiology. *Photochem. Photobiol.* 95, 95–104. doi:10.1111/php.12930

Rolny, C., Mazzone, M., Tugues, S., Laoui, D., Johansson, I., Coulon, C., et al. (2011). HRG Inhibits Tumor Growth and Metastasis by Inducing Macrophage Polarization and Vessel Normalization through Downregulation of PlGF. *Cancer Cell* 19, 31–44. doi:10.1016/j.ccr.2010.11.009

Rydzanicz, M., Wrzesiński, T., Bluyssen, H. A. R., and Wesoły, J. (2013). Genomics and Epigenomics of clear Cell Renal Cell Carcinoma: Recent Developments and Potential Applications. *Cancer Lett.* 341, 111–126. doi:10.1016/j.canlet.2013.08.006

Sang, L., Sun, L., Wang, A., Zhang, H., and Yuan, Y. (2020). The N6-Methyladenosine Features of mRNA and Aberrant Expression of m6A Modified Genes in Gastric Cancer and Their Potential Impact on the Risk and Prognosis. *Front. Genet.* 11, 561566. doi:10.3389/fgene.2020.561566

Schibler, U., Kelley, D. E., and Perry, R. P. (1977). Comparison of Methylated Sequences in Messenger RNA and Heterogeneous Nuclear RNA from Mouse L Cells. *J. Mol. Biol.* 115, 695–714. doi:10.1016/0022-2836(77)90110-3

Schwartz, S., Mumbach, M. R., Jovanovic, M., Wang, T., Maciag, K., Bushkin, G. G., et al. (2014). Perturbation of m6A Writers Reveals Two Distinct Classes of mRNA Methylation at Internal and 5′ Sites. *Cel. Rep.* 8, 284–296. doi:10.1016/j.celrep.2014.05.048

Shah, A., Rashid, F., Awan, H. M., Hu, S., Wang, X., Chen, L., et al. (2017). The DEAD-Box RNA Helicase DDX3 Interacts with m(6)A RNA Demethylase ALKBH5. *Stem Cell Int.* 2017, 8596135. doi:10.1155/2017/8596135

Shen, L., Shao, N.-Y., Liu, X., Maze, I., Feng, J., and Nestler, E. J. (2013). diffReps: Detecting Differential Chromatin Modification Sites from ChIP-Seq Data with Biological Replicates. *Plos One* 8, e65598. doi:10.1371/journal.pone.0065598

Siegel, R. L., Miller, K. D., and Jemal, A. (2019). Cancer Statistics, 2019. *CA A. Cancer J. Clin.* 69, 7–34. doi:10.3322/caac.21551

Wang, X., Zhao, B. S., Roundtree, I. A., Lu, Z., Han, D., Ma, H., et al. (2015). N(6)-methyladenosine Modulates Messenger RNA Translation Efficiency. *Cell* 161, 1388–1399. doi:10.1016/j.cell.2015.05.014

Wanna-Udom, S., Terashima, M., Lyu, H., Ishimura, A., Takino, T., Sakari, M., et al. (2020). The m6A Methyltransferase METTL3 Contributes to Transforming Growth Factor-Beta-Induced Epithelial-Mesenchymal Transition of Lung Cancer Cells through the Regulation of JUNB. *Biochem. Biophysical Res. Commun.* 524, 150–155. doi:10.1016/j.bbrc.2020.01.042

Wei, C. M., Gershowitz, A., and Moss, B. (1976). 5'-Terminal and Internal Methylated Nucleotide Sequences in HeLa Cell mRNA. *Biochemistry* 15, 397–401. doi:10.1021/bi00647a024

Wilson, C., Chen, P. J., Miao, Z., and Liu, D. R. (2020). Programmable M(6)A Modification of Cellular RNAs with a Cas13-Directed Methyltransferase. *Nat. Biotechnol.* 38, 1431. doi:10.1038/s41587-020-0572-6

Yokomori, K., Tsuchida, Y., and Saito, S. (1983). Tyrosine Hydroxylase and Choline Acetyltransferase Activity in Human Neuroblastoma Correlations

with Clinical Features. *Cancer* 52, 263–272. doi:10.1002/1097-0142(19830715)52:2<263:aid-cncr2820520213>3.0.co;2-o

Zhang, C., Samanta, D., Lu, H., Bullen, J. W., Zhang, H., Chen, I., et al. (2016). Hypoxia Induces the Breast Cancer Stem Cell Phenotype by HIF-dependent and ALKBH5-Mediated m(6)A-Demethylation of NANOG mRNA. *Proc. Natl. Acad. Sci. USA* 113, E2047–E2056. doi:10.1073/pnas.1602883113

Zhang, J., Guo, S., Piao, H.-y., Wang, Y., Wu, Y., Meng, X.-y., et al. (2019). ALKBH5 Promotes Invasion and Metastasis of Gastric Cancer by Decreasing Methylation of the lncRNA NEAT1. *J. Physiol. Biochem.* 75, 379–389. doi:10.1007/s13105-019-00690-8

Zhang, S., Zhang, W., Xiao, Y., Qin, T., Yue, Y., Qian, W., et al. (2020). Targeting MUC15 Protein in Cancer: Molecular Mechanisms and Therapeutic Perspectives. *Ccdt* 20, 647–653. doi:10.2174/1568009620666200601140639

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

Zheng, Z., Mao, S., Guo, Y., Zhang, W., Liu, J., Li, C., et al. (2020). N6-methyladenosine RNA M-ethylation R-egulators P-articipate in M-alignant P-rogression and H-ave P-rognostic V-alue in clear C-ell R-enal C-ell C-arcinoma. *Oncol. Rep.* 43, 1591–1605. doi:10.3892/or.2020.7524

Zhou, J., Wan, J., Gao, X., Zhang, X., Jaffrey, S. R., and Qian, S.-B. (2015). Dynamic m(6)A mRNA Methylation Directs Translational Control of Heat Shock Response. *Nature* 526, 591–594. doi:10.1038/nature15377

Zhu, H., Gan, X., Jiang, X., Diao, S., Wu, H., and Hu, J. (2019). ALKBH5 Inhibited Autophagy of Epithelial Ovarian Cancer through miR-7 and BCL-2. *J. Exp. Clin. Cancer Res.* 38, 163. doi:10.1186/s13046-019-1159-2

Zhu, W., Wang, J. Z., Xu, Z., Cao, M., Hu, Q., Pan, C., et al. (2019). Detection of N(6)-methyladenosine M-odification R-esidues (Review). *Int. J. Mol. Med.* 43, 2267–2278. doi:10.3892/ijmm.2019.4169

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

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

frontiers | Frontiers in Genetics

ORIGINAL RESEARCH
published: 19 July 2022
doi: 10.3389/fgene.2022.876325

Comprehensive Analysis of the Expression and Prognosis for Tripartite Motif-Containing Genes in Breast Cancer

Lvwen Ning[†], *Qin Huo*[†] *and Ni Xie*[*]

Biobank, Shenzhen Second People's Hospital, First Affiliated Hospital of Shenzhen University, Health Science Center, Shenzhen University, Shenzhen, China

Tripartite motif-containing genes (TRIMs), with a ubiquitin ligase's function, play critical roles in antitumor immunity by activating tumor-specific immune responses and stimulating tumor proliferation, thus affecting patient outcomes. However, the expression pattern and prognostic values of TRIMs in breast cancer (BC) are not well clarified. In this study, several datasets and software were integrated to perform a comprehensive analysis of the expression pattern in TRIMs and investigate their prognosis values in BC. We found that *TRIM59/46* were significantly upregulated and *TRIM66/52-AS1/68/7/2/9/29* were decreased in BC and validated them using an independent cohort. The expression of numerous TRIMs are significantly correlated with BC molecular subtypes, but not with tumor stages or patient age at diagnosis. Higher expression of *TRIM3/14/69/45* and lower expressions of *TRIM68/2* were associated with better overall survival in BC using the Kaplan–Meier analysis. The multivariate Cox proportional hazards model identified *TRIM45* as an independent prognostic marker. Further analysis of single-cell RNA-seq data revealed that most TRIMs are also expressed in nontumor cells. Higher expression of some TRIMs in the immune or stromal cells suggests an important role of TRIMs in the BC microenvironment. Functional enrichment of the co-expression genes indicates that they may be involved in muscle contraction and interferon-gamma signaling pathways. In brief, through the analysis, we provided several TRIMs that may contribute to the tumor progression and *TRIM45* as a potential new prognostic biomarker for BC.

Keywords: TRIM family, breast cancer, pronositic biomarker, bioinformatics, TRIM45

OPEN ACCESS

Edited by:
Le Li,
Cornell University, United States

Reviewed by:
Jinlin Guo,
Chengdu University of Traditional Chinese Medicine, China
Leli Zeng,
Sun Yat-sen University, China

***Correspondence:**
Ni Xie
xn100@szu.edu.cn

[†]*These authors share co-first authorship*

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

Received: *15 February 2022*
Accepted: *02 June 2022*
Published: *19 July 2022*

Citation:
Ning L, Huo Q and Xie N (2022) Comprehensive Analysis of the Expression and Prognosis for Tripartite Motif-Containing Genes in Breast Cancer. Front. Genet. 13:876325. doi: 10.3389/fgene.2022.876325

1 INTRODUCTION

Breast cancer (BC) is a complex disease with a high incidence worldwide, especially in women (Siegel et al., 2021). An approximate of 2.3 million new cases were diagnosed in 2020 (Sung et al., 2021). Based on the gene expression pattern, BC is classified into five intrinsic subtypes: luminal A, luminal B, HER2+, basal-like, and normal-like (Sørlie et al., 2003). Different subtypes usually require different treatment strategies (Rouzier et al., 2005). Although great progress has been made recently in molecular diagnostic and neoadjuvant chemotherapy (Waks and Winer, 2019), there are still ~680,000 BC patients who die each year (Sung et al., 2021). Therefore, discovering new biomarkers for stratified BC patients and new therapeutic targets is urgent.

TABLE 1 | Differentially expressed TRIMs in BC.

Genes	Official full name	Locus	Discovery dataset (TCGA-BRCA)				Validation dataset (SRP324699)	
			Non-paired		Paired		Fc	*p*
			Fc	*p*	Fc	*p*		
TRIM59	tripartite motif containing 59	3q25.33	4.47	8.55e-129	4.98	1.89e-47	3.59	3.59e-08
TRIM46	tripartite motif containing 46	1q22	3.17	5.38e-38	2.88	1.67e-18	1.61	0.021
TRIM66	tripartite motif containing 66	11p15.4	0.64	5.16e-14	0.57	2.49e-12	0.66	0.026
TRIM52-AS1	TRIM52 antisense RNA 1	5q35.3	0.57	7.42e-29	0.55	3.14e-17	0.66	0.022
TRIM68	tripartite motif containing 68	11p15.4	0.55	2.61e-25	0.54	1.42e-16	0.64	0.025
TRIM7	tripartite motif containing 7	5q35.3	0.57	1.34e-4	0.60	3.62e-04	0.53	0.030
TRIM2	tripartite motif containing 2	4q31.3	0.26	2.76e-42	0.28	8.00e-23	0.36	1.45e-05
TRIM29	tripartite motif containing 29	11q23.3	0.17	2.14e-24	0.17	2.47e-16	0.041	1.71e-06
TRIM9	tripartite motif containing 9	14q22.1	0.43	2.70e-10	0.46	4.52e-06	0.30	6.94e-04

Fc, fold change; p, p-*value; TRIMs, tripartite motif.*

The tripartite motif-containing genes (TRIMs) are involved in many cellular functions by acting as E3 ubiquitin ligases (Ozato et al., 2008; Lv et al., 2017; van Gent et al., 2018). As a result, the dysregulation of TRIMs results in a lot of diseases (Jefferies et al., 2011; Borlepawar et al., 2019; Kumarasinghe et al., 2021), including various types of cancer (Cambiaghi et al., 2012; Crawford et al., 2018; Jaworska et al., 2020). More than 80 members from this conserved protein superfamily are reported in humans (Sardiello et al., 2008). Several TRIMs play an oncogenic role in cancers, while some exert tumor suppressor functions (Venuto and Merla, 2019). Earlier, our group found that rare variants in *TRIM31* have contributed to the genetic susceptibility of nasopharyngeal carcinoma (NPC), and higher TRIM31 expression is associated with poor overall survival (OS) of NPC (Ning et al., 2020). We recently reported that TRIM58 can promote drug resistance in BC (Wang et al., 2022). TRIM28 (Wei et al., 2016), TRIM44 (Kawabata et al., 2017), and TRIM59 (Tan et al., 2018) were reported to play an oncogene role, and TRIM16 (Bell et al., 2013) was a tumor suppressor gene in BC. The important role of TRIMs in BC suggests that some TRIMs may have a prognostic value in BC. Higher expression of TRIM44 (Kawabata et al., 2017) and lower expression of TRIM13 (Chen et al., 2019)/TRIM21 (Zhou et al., 2018) are poor prognostic factors in BC.

Although the expression levels of some TRIMs have been reported to have prognostic values, the expression pattern of TRIMs and their prognostic values are not comprehensively clarified in BC. The functional study of this family is still limited. In this study, by integrating several datasets, we explored the expression profile of all TRIMs in BC, prioritized several not well-studied TRIMs that may contribute to BC tumorigenesis, and identified *TRIM45* as a potential new prognostic marker for BC.

2 MATERIALS AND METHODS

2.1 Dataset Acquisition and Expression Level Quantification

Expression profiles (read counts) of 1,291 The Cancer Genome Atlas (TCGA)-BRCA samples (Ciriello et al., 2015) were downloaded from the Genomic Data Commons portal (https://portal.gdc.cancer.gov). After removing redundant samples and samples with incomplete clinical information, 1194 samples including 1075 tumor, 112 adjacent normal, and 7 metastatic were kept. In this dataset, 111 individuals have normal and tumor paired samples.

In addition, an RNA-seq dataset (SRP324699) was downloaded from The National Center for Biotechnology Information (NCBI) Sequence Read Archive database (https://www.ncbi.nlm.nih.gov/sra), which included 14 paired normal and tumor BC samples. The FASTQ files were aligned with the human reference genome hg38 using TopHat2 (Kim et al., 2013). Reads were counted using htseq-count (Anders et al., 2015) for each gene based on the GENCODE annotation (v38).

Quantile normalization and 10-based log transformations were performed on the raw read counts of the expression data. The following analyses were based on the normalized reads.

Single-cell RNA-seq data were used to characterize the cellular expression specificity in BC. These data were generated from three samples, including a primary tumor, a positive lymph node, and a negative lymph node. The expression matrix was downloaded from the NCBI Gene Expression Omnibus (GEO) database under accession ID (GSE158399) and then processed using the Scanpy pipeline (Wolf et al., 2018). The data matrix from the three samples was merged, cells with expression genes less than 200 were filtered, and filtered genes were expressed in fewer than five cells. We further filtered out the cells that expressed high-level mitochondrial genes (>20%) and cells that expressed more than 4,000 genes. Marker genes were found using rank_genes_groups with parameters "leiden" and "wilcoxon." Cell types are annotated according to the expression of marker genes (B cells: CD79A, CD79B; T cells: CD3D, IL7R; tumor cells: KRT19; myeloid cells: LYZ; stromal cells: others).

2.2 Differential Analysis of Tripartite Motif-Containing Genes

From GENCODE annotation, 115 TRIMs were extracted. The expressions of TRIMs were compared between BC and normal samples. Two-tailed Student's *t*-test was used to do the comparison. One gene is significantly changed if the absolute log fold change (|logFc|) was >0.58 and the Benjamini–Hochberg false discovery rate–corrected *p*-value <0.05. In addition, paired Student's t-test

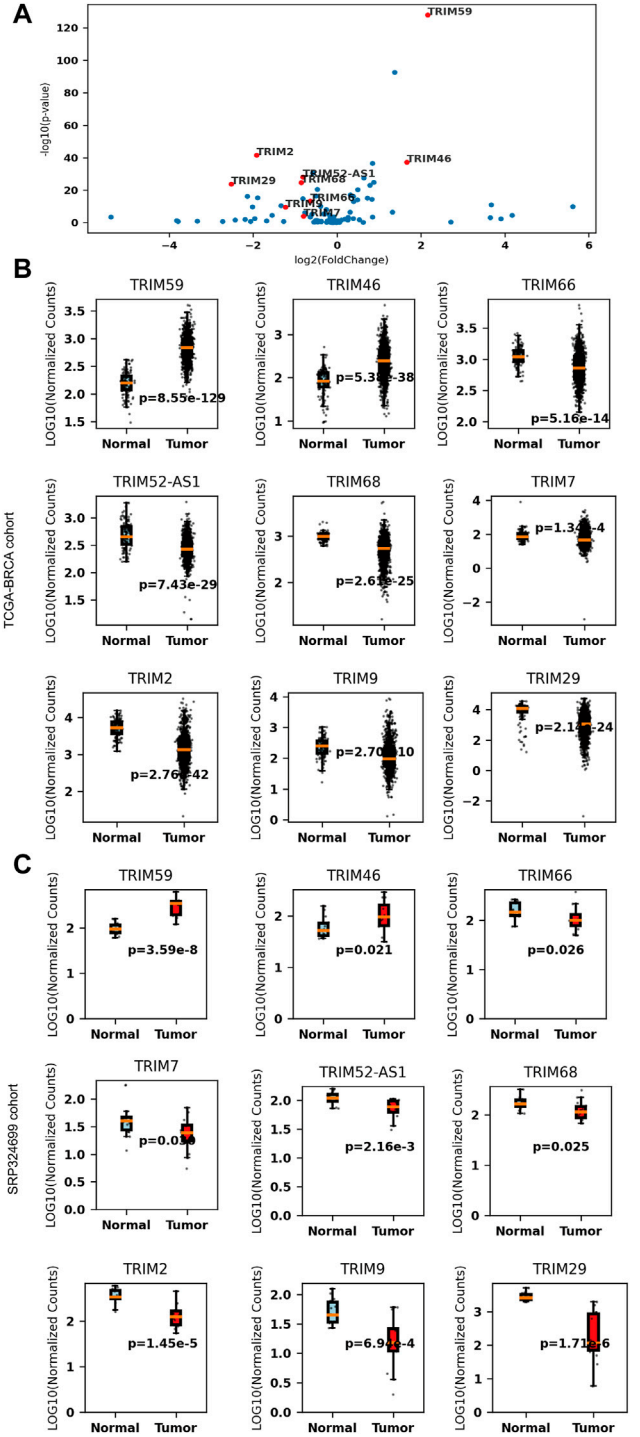


FIGURE 1 | The expression of nine differentially expressed tripartite motif-containing genes (TRIMs) in breast cancer (BC). **(A)** volcano plot of TRIMs in The Cancer Genome Atlas (TCGA)-BRCA cohort **(B)** boxplots of the expressions of upregulated *TRIM59/46* and downregulated *TRIM66/52-AS1/68/7/2/9/29* in the TCGA-BRCA cohort (N_{Normal} = 112, N_{tumor} = 1075). **(C)** boxplots of the expressions of upregulated *TRIM59/46* and downregulated *TRIM66/52-AS1/68/7/2/9/29* in the SRP324699 cohort (N_{Normal} = 14, N_{tumor} = 14).

was used for paired comparison among the normal tumor samples from the same individuals. In the validation, an RNA-seq dataset (SRP324699) was used, which included 14 paired tumor and adjacent normal samples. The *p*-value cut-off of 0.05 was used.

2.3 Correlation Analysis Between Tripartite Motif-Containing Genes and Clinical Features

The clinical information of the TCGA-BRCA samples was downloaded from cBioPortal (https://www.cbioportal.org/) (Gao et al., 2013). Three types of information were considered: molecular subtypes, stages, and age groups. Samples were divided into five molecular subtypes (normal-like, luminal A, luminal B, HER2+, and basal) based on the expression profiles, four stages (I, II, III, IV, and V) according to the American Joint Committee on Cancer (AJCC) pathologic tumor stage, and four age-at-diagnosis groups (<35, 35–50, 50–70, >70).

One-way ANOVA was used to test if there is any difference among the expressions of the candidate gene in different subgroups. A *p*-value after the Bonferroni correction was used to define the significance cut-off ($p < 0.05/43 = 0.0011$). If there is any significant difference in the subgroups, Tukey's honestly significant difference (HSD) test is further performed to find the significantly different pairs in the subgroups ($p < 0.05$).

2.4 Analysis of the Prognostic Value of Tripartite Motif-Containing Genes

The relationship between TRIMs' expression and the OS of BC patients is analyzed using the Kaplan–Meier curve and log-rank test using Python package lifelines (https://lifelines.readthedocs.io/en/latest/, version 0.26.4; Davidson-Pilon, 2019). Genes with $p < 0.05$ were considered to have prognostic values.

In the validation phase, an independent dataset including 327 individuals was downloaded from the NCBI GEO database (GSE20685). The Kaplan–Meier curve and log-rank test were performed similarly. A statistically significant difference was considered when a log-rank *p* was <0.05.

Univariate and multivariate Cox proportional hazards regression analyses using lifelines were performed on TRIMs and clinicopathological characteristics on the OS of patients. A statistically significant difference was considered when the *p* was <0.05.

2.5 Functional Analysis of Gene Set

We conducted a protein–protein interaction network analysis of TRIMs to explore the interactions among them using the STRING database (version 11.5, https://string-db.org/; Szklarczyk et al., 2021), which is a database that collects all publicly available sources of protein–protein interaction. Genes were analyzed using the full STRING network, and the confidence score of 0.4 was used for filtering interactions. In addition to the TRIMs we analyzed, directly interacted genes were added to the network by clicking "More" once from the web interface.

TABLE 2 | Association of TRIMs' expression and breast cancer clinical features.

Characteristics	Genes	F test		Significant pairs (Tukey's honestly significant difference test)[a]
		F-value	*p*	
Molecular subtypes[b]	*TRIM3*	155.70	2.07e-98	G1 vs. (G2, G3, G3, G4, G5); G2 vs. G5; G3 vs. G5; G4 vs. G5
	TRIM29	125.05	7.74e-83	G1 vs. (G2, G3, G4); G2 vs. (G3, G5); G3 vs. (G4, G5); G4 vs. G5
	TRIM2	106.59	9.74e-73	G1 vs. (G2, G3, G4); G2 vs. (G3, G5); G3 vs. (G4, G5); G4 vs. G5
	TRIM23	96.40	6.65e-60	G1 vs. (G2, G3, G5); G2 vs. (G4, G5); G3 vs. (G4, G5); G4 vs. G5
	TRIM8	84.50	7.78e-60	G1 vs. G5; G2 vs. (G3, G4, G5); G3 vs. G5; G4 vs. G5
	TRIM65	74.82	7.27e-54	G1 vs. G5; G2 vs. (G3, G5); G3 vs. (G4, G5); G4 vs. G5
Age-at-diagnosis groups	*TRIM29*	8.23	1.38e-05	G2 vs. G3; G2 vs. G4
	TIRM47	8.22	2.06e-05	G1 vs. G3; G1 vs. G4; G2 vs. G3; G2 vs. G4
	TRIM24	7.62	4.80e-05	G1 vs. G2; G1 vs. G3; G1 vs. G4; G2 vs. G4
	TRIM65	6.57	2.13e-04	G2 vs. G3; G2 vs. G4
	TRIM52	5.70	7.20e-04	G2 vs. G3; G2 vs. G4
Stage (I, II, III, IV, and V)	*TRIM3*	6.37	0.00028	G2 vs. G3; G2 vs. G4

Tested using one-way ANOVA.

[a]*Tumor subtypes: G1: normal-like, G2: luminal A, G3: luminal B, G4: HER2+, G5: basal; age-at-diagnosis groups: G1: <35, G2: 35–50, G3: 50–70, G4: >70; stages: G1: stage I, G2: stage II, G3: stage III, G4: stage IV and V.*

[b]*Only six top associated TRIMs were shown in the table.*

The coexpressed genes are identified using the TCGA-BRCA dataset. The Spearman correlation coefficients were calculated between the target genes and other genes. One gene was considered a coexpressed gene if the |coefficient| was >0.6. As a gene family, we performed the correlation analysis among the selected key genes. The coefficients among the TRIMs were calculated and visualized using a heatmap.

Gene Ontology (GO), Kyoto Encyclopedia of Genes and Genomes (KEGG), pathway and Reactome pathway enrichment analyses were performed using g:Profiler (https://biit.cs.ut.ee/gprofiler; Raudvere et al., 2019) to elucidate the function, biological process, and enrichment pathway of the selected genes. The adjusted enrichment *p*-value was reported by g:Profiler as its default. Biological processes, cellular components, and molecular functions were included in the GO enrichment analysis, while KEGG and Reactome analysis defined the pathways related to the TRIMs.

2.6 Statistical Analysis

Statistical analysis and data plotting were performed using Python (version3). The value of 0.05 was the cut-off for the statistical significance unless otherwise specified. The Kaplan–Meier method with a log-rank test was used for survival analysis. Univariable and multivariable analyses were performed using the Cox proportional hazards model. One-way ANOVA was used to compare multiple groups; the HSD test was used to test the difference among subgroups. The Mann–Whitney *U* test was used for the comparison of the TRIM45-high/low groups.

3 RESULTS

3.1 Differential Analysis of Tripartite Motif-Containing Genes in Breast Cancer

According to the GENCODE, 115 TRIM-related genes were extracted. Of them, 114 TRIMs can be identified from the TCGA-BRCA dataset. There are 110 TRIMs expressed in BC with at least one read detected in one of the samples, while four pseudo-TRIMs (*TRIM60P11Y*, *TRIM60P5Y*, *TRIM60P10Y*, and *TRIM60P9Y*) are not expressed at all in this dataset. Overall, the three most highly expressed genes are *TRIM28*, *TRIM44*, and *TRIM8* in the primary and metastatic tumors and normal tissues.

To explore TRIMs dysregulated in BC, we compared the expression profiles between normal and tumor samples using the TCGA-BRCA dataset (N_{tumor} = 1075 and N_{normal} = 112). The analysis identified 35 significantly differentially expressed genes (DEGs) from 110 TRIMs ($p_{adjusted} < 0.05$, $|logFc| > 0.58$). In total, 16 TRIMs (*TRIM15/L2/39-RPP21/71/80P/59/46/11/72/60P18/62/67/3/14/45/28*) are upregulated, while 19 TRIMs (*TRIM61/66/31/63/7/52-AS1/22/68/9/31-AS1/51BP/60P17/2/43B/55/50/51JP/29/75*) are downregulated in BC.

To validate these results, we further analyzed another dataset (SRP324699) that included 14 tumor and adjacent normal tissues. Nine of the 35 DEGs are still significant in the independent cohort ($p < 0.05$, $|logFc| > 0.58$). *TRIM59/46* are consistently upregulated and *TRIM66/52-AS1/68/7/2/9/29* are consistently downregulated in BC. The comparison details are listed in **Table 1**. These significant TRIMs were highlighted in a volcano plot in **Figure 1A**, and the boxplots of the expression of the TRIMs are shown in **Figure 1B** and **Figure 1C** for TCGA and the validation cohort, respectively.

Nine genes are still significant when we do a paired comparison among the 111 individuals with normal and tumor samples ($p_{adjusted} < 0.05$, $|logFc|>0.58$) (**Table 1**). Six (*TRIM59/46/68/2/9/29*) of the nine TRIMs are also significantly altered in the metastatic tissues when compared with the normal samples ($p_{adjusted} < 0.05$), except TRIM66/52-AS1/7.

As a result, we conclude that *TRIM59/46* are upregulated and *TRIM66/52-AS1/68/7/2/9/29* are downregulated in BC. Abnormal

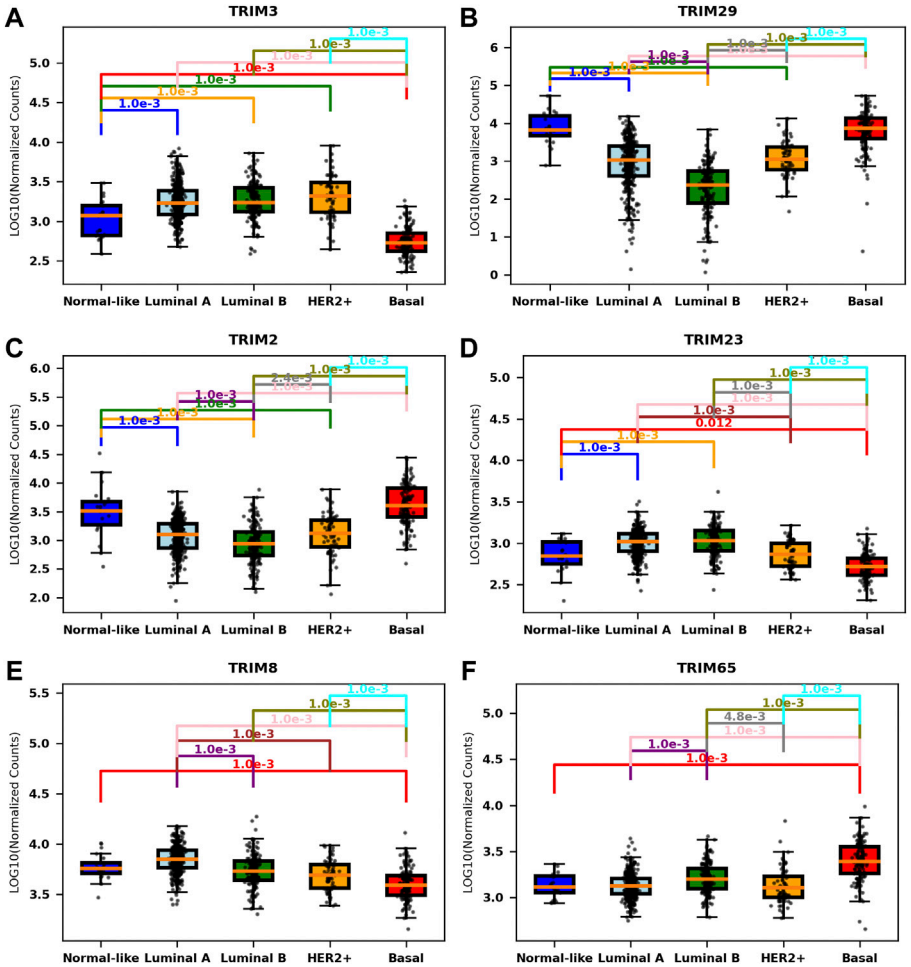


FIGURE 2 | Association of TIRMs' expression and molecular subtypes in BC. Boxplots of the expressions of the most significant TRIMs in different molecular subtypes. **(A)** *TRIM3*, expressed significantly lower in the normal-like and basal subtypes. **(B)** *TRIM29*, expressed significantly higher in the normal-like and basal subtypes. **(C)** *TRIM2*, expressed significantly higher in the normal-like and basal subtypes. **(D)** *TRIM23*, expressed significantly lower in the basal subtype. **(E)** *TRIM8*, expressed significantly lower in the basal subtype. **(F)** *TRIM65*, expressed significantly higher in the basal subtype.

expression of these four TRIMs suggested they may be involved in the pathogenesis of BC.

3.2 Expression of Tripartite Motif-Containing Genes Correlated to the Molecular Subtypes of Breast Cancer

Next, we analyzed the association between the expression level of TRIMs and clinical features using the TCGA-BRCA dataset. One-way ANOVA was used to test if there was any expression difference of TRIMs in the molecular subtypes (normal-like, luminal A, luminal B, HER2+, and basal), tumor stages (I, II, III, IV, and V), and age-at-diagnosis groups (age groups; <35, 35–50, 50–70, and >70). Tukey's HSD test was further performed to find significant pairs.

Overall, approximately 33.67% of TRIMs (37/110) are correlated with molecular subtypes, but only a small portion of TRIMs are associated with age groups (5/110 or 4.55%) or with the clinical stages (1/110 or 1.82%) ($p < 0.001$, F test, Bonferroni correction). The statistic details for the significantly correlated genes are listed in **Table 2** (Only six genes with the least *p*-value are listed for molecular subtypes). The most significant one in molecular subtypes is *TRIM3* (p = 2.07e−98), which is expressed significantly lower in normal-like and basal BC than in other subtypes (**Figure 2A**). A different scenario was found for *TRIM29* (p = 7.74e−83) and *TRIM2* (p = 9.74e−73); they are highly expressed in normal-like and basal BC than in other subtypes (**Figure 2B** and **Figure 2C**). The expressions of *TRIM23/TRIM8* are significantly lower in basal than in other subtypes (**Figure 2D** and **Figure 2E**), while *TRIM65* is expressed significantly higher in basal than in other subtypes (**Figure 2F**). The top three genes associated with age groups are *TRIM29* (p = 1.38e−05), *TRIM47* (p = 2.06e−05), and *TRIM24* (4.80e−05).

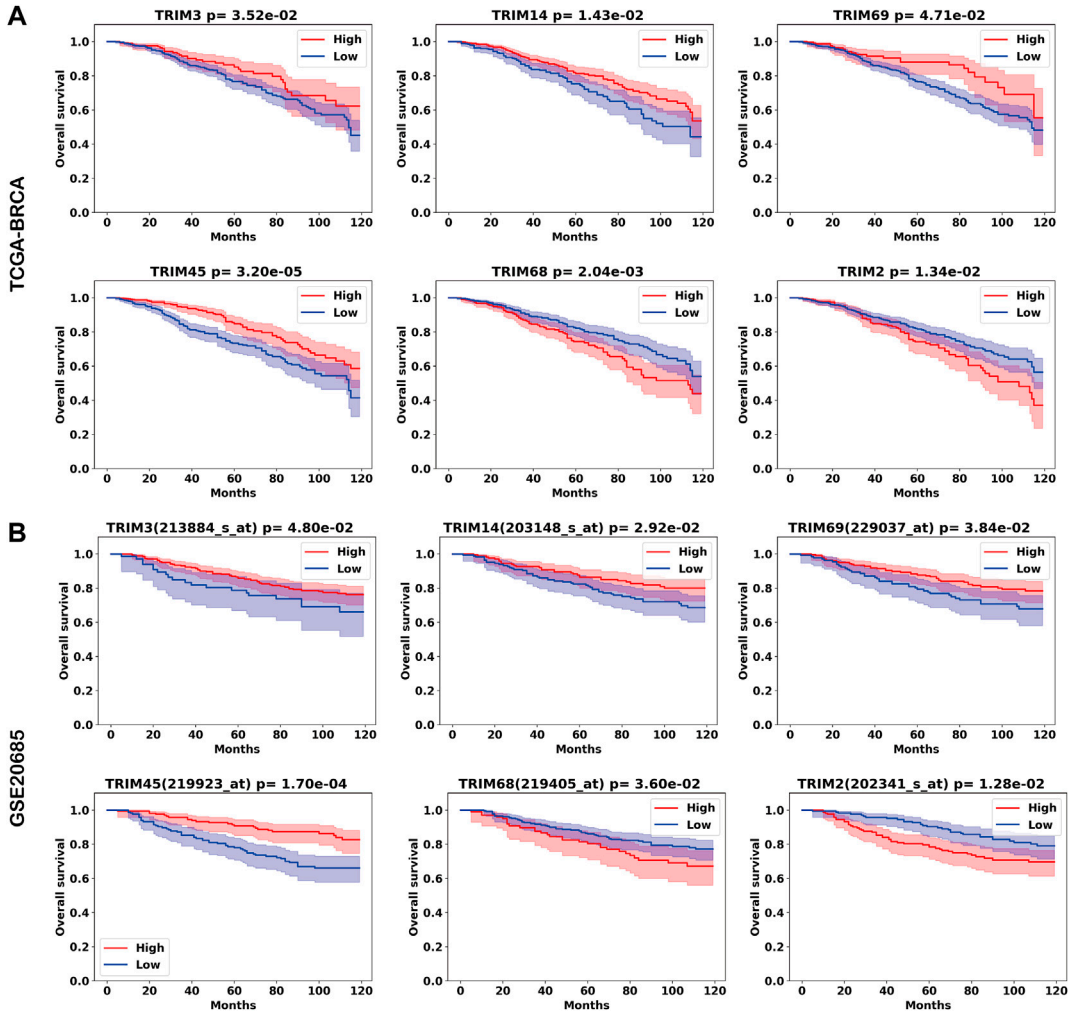


FIGURE 3 | Tripartite motif (TRIMs) associated with the prognosis of BC. The Kaplan–Meier curves for genes *TRIM3/14/69/45/68/2* in **(A)** TCGA-BRCA cohort and **(B)** the GSE20685 cohort.

In conclusion, many TRIMs are significantly correlated with molecular subtypes, but not with tumor stages or patient age at diagnosis.

3.3 Tripartite Motif-Containing Genes Associated With Prognosis of Breast Cancer

We next investigated the prognosis value of TRIMs' expression in BC. The relationship between the expression of TRIMs and the OS of BC patients was analyzed using the Kaplan–Meier curve and log-rank test using the TCGA-BRCA dataset. The analysis revealed that the expressions of 37 TRIMs are significantly associated with the OS of BC ($p < 0.05$, log-rank test).

To validate these results, we performed Kaplan–Meier analysis using an independent dataset. Six TRIMs are validated from 24 genes from the TCGA-BRCA cohort. Higher expressions of *TRIM3*, *TRIM14*, *TRIM69*, and *TRIM45* are associated with a favorable OS, while higher expressions of *TRIM68* and *TRIM2* are associated with a worse OS in BC. The Kaplan–Meier curves for the six genes are shown in **Figures 3A,B** for the TCGA-BRCA and GSE20685 cohorts.

Of the nine DEGs, *TRIM2* and *TRIM68* are significantly associated with worse OS in both the TCGA-BRCA cohort and the validation cohort. Moreover, while *TRIM2* and *TRIM68* are downregulated in BC, higher expressions of *TRIM2* and *TRIM68* are associated with worse OS in BC. Of the seven upregulated genes, there is no significant gene identified.

In the previous analysis, TRIMs' expressions are associated with BC subtypes. Of the six TRIMs (*TRIM3/14/69/45/2/68*) identified from Kaplan–Meier analysis, 5/6 (83.3%) were associated with molecular subtypes except for *TRIM14*; *TRIM3* was associated with stages. To investigate the independent role of TRIMs, we applied a univariate and multivariate Cox proportional hazards model to the expression data. In the univariant analysis, ages, luminal A subtype, HER2+ subtype, stages, and the expression of *TRIM69/45* are associated with the OS of BC (**Table 3**). In the multivariate analysis, high expression of *TRIM45* demonstrates an

TABLE 3 | Univariate and multivariate Cox proportional hazards analyses for TRIMs, ages, stages, and molecular subtypes in breast cancer.

Characteristics[a]		Univariate analysis		Multivariate analysis	
		Hazard ratio (95% CI)	*p*	Hazard ratio (95% CI)	*p*
Age		1.03 (1.02-1.05)	<**0.005**	1.04 (1.02-1.05)	<**0.005**
Molecular subtypes	Lumina A	0.66 (0.46-0.94)	**0.02**		
	Lumina B	1.43 (0.93-2.21)	0.10		
	Her2+	2.29 (1.31-4.01)	<**0.005**	1.92 (1.04-3.55)	**0.04**
	Basal	0.89 (0.54-1.45)	0.63		-
	Normal-like	0.99 (0.36-2.67)	0.98		-
AJCC stages	I	0.37 (0.21-0.68)	<**0.005**	0.19 (0.10-0.36)	<**0.005**
	II	0.66 (0.46-0.95)	**0.03**	0.38 (0.26-0.56)	<**0.005**
	III	2.12 (1.44-3.12)	<**0.005**	1.94 (1.30-2.88)	<**0.005**
	IV and V	3.88 (2.24-6.69)	<**0.005**		
TRIM3		1.40 (0.76-2.60)	0.28	-	-
TRIM14		0.73 (0.32-1.67)	0.46	-	-
TRIM69		0.36 (0.13-0.96)	**0.04**	-	-
TRIM45		0.40 (0.20-0.80)	**0.01**	0.47 (0.23-0.99)	**0.05**
TRIM68		1.18 (0.59-2.35)	0.65	-	-
TRIM2		1.08 (0.68-1.71)	0.74	-	-

[a]*Age, numeric; molecular subtypes: (five) categories were encoded into binary categories; AJCC stages: (four) categories were encoded into binary categories.*
AJCC, American Joint Committee on Cancer; CI, confidence interval;.p, p-*value; TRIMs, tripartite motif-containing genes.*
The bold value means "significant p *value". Also need bold other 3* p *values less than 0.05.*

independent role (from molecular subtypes, ages, and stages) with a better prognosis in BC (**Table 3**).

Based on the analysis, the expression of *TRIM45* may be used as a prognosis marker for BC, while the prognosis value of *TRIM3/14/69/68/2* may need to be further investigated.

3.4 Expression Specificity of Tripartite Motif-Containing Genes in Cell Types

BC is shown to have high cellular heterogeneity; next, we profiled a single-cell dataset (GSE158399) to investigate if there is any cell-type expression specificity in TRIMs. This piece of data includes three samples: primary BC, positive lymph node, and negative lymph node. Based on cell markers, these cells were annotated into five categories: tumor cells, T cells, B cells, stromal cells, and myeloid cells. The cell number in each cell type were counted and visualized using a pie plot (**Figurea 4A–C**). There are 48 TRIMs expressed in this dataset. The expression levels of these genes are summarized according to cell types and sample types (**Figures 4D–F**). We found that most TRIMs are expressed in the tumor and stromal cells in the primary BC, and the expression in the nontumor cell types are accounting for a large proportion of the total expression. Some genes (*TRIM22/TRIM38/TRIM5/TRIM59*) are expressed lower in tumor cells than in other cell types (**Figures 4D–F**). In addition, numerous TRIMs are highly expressed in the T cells and B cells in the BC lymph nodes (negative and positive), which may correspond to their important role in immunity.

3.5 Interaction, Co-expression, and Functional Analysis of Tripartite Motif-Containing Genes

The interactions of TRIMs were investigated using the STRING database. Nine DEGs (*TRIM59/46/66/52-AS1/68/7/2/9/29*) and six genes (*TRIM3/14/69/45/68/2*) in the Kaplan–Meier analysis were searched in the STRING database (https://string-db.org/). TRIM52-AS1 is a non-coding RNA, and the other 12 genes and 5 directly interacted non-TRIM genes formed three disconnected networks (**Figure 5A**). TRIM69 itself, which did not interact with any other genes, formed an isolated network. One small network was formed by three TRIMs (TRIM45, TRIM68, TRIM9) and a non-TRIM gene DCC netrin 1 receptor (DCC). DCC is considered a tumor suppressor; downregulation of TRIM9 and TRIM68 in BC may lead to changes in BC, although we did not see the correlation in the mRNA level (Spearman coefficient r = 0.022, 0.043). The third network includes eight TRIMs (TRIM2/7/46/3/66/59/29/14) that interacted with four non-TRIM proteins: actinin alpha 4 (ACTN4), inhibitor of nuclear factor kappa B kinase regulatory subunit gamma (IKBKG), mitochondrial antiviral signaling protein (MAVS), and ubiquitin-specific peptidase 14 (USP14). TRIM14 interacted with USP14, MAVS, and IKBKG, while TRIM14 is downregulated in BC, although we also did not see the correlations in the mRNA level (r = 0.10, 0.026, 0.052). In addition, we also did observe the correlation between *TRIM3* and *ACTN4* in the mRNA level (r = -0.075).

To investigate the possible function of the 13 TRIMs, their co-expression genes were calculated using Spearman correlation in the TCGA-BRCA dataset. The correlations among these TRIMs were limited (**Figure 5B**). We further checked the correlations in the top 30 expressed TRIMs in BC, and the heatmap suggests the low correlation in this family (**Figure 5C**). In addition, the correlations of 32 TRIMs with protein expression information in the TCGA-BRCA cohort are shown in **Figure 5D**. In conclusion, low expression correlations were found among TRIMs. Including the 13 TRIMs, there were 405 genes coexpressed with the targeted TRIMs (r > 0.6). No gene was correlated with *TRIM68/9/46/52-AS1/14/45/59* under the threshold r > 0.6. GO and KEGG/Reactome pathway enrichment was performed based on the coexpressed genes using the Gprofiler. In the biological process (BP) category, the three most enriched

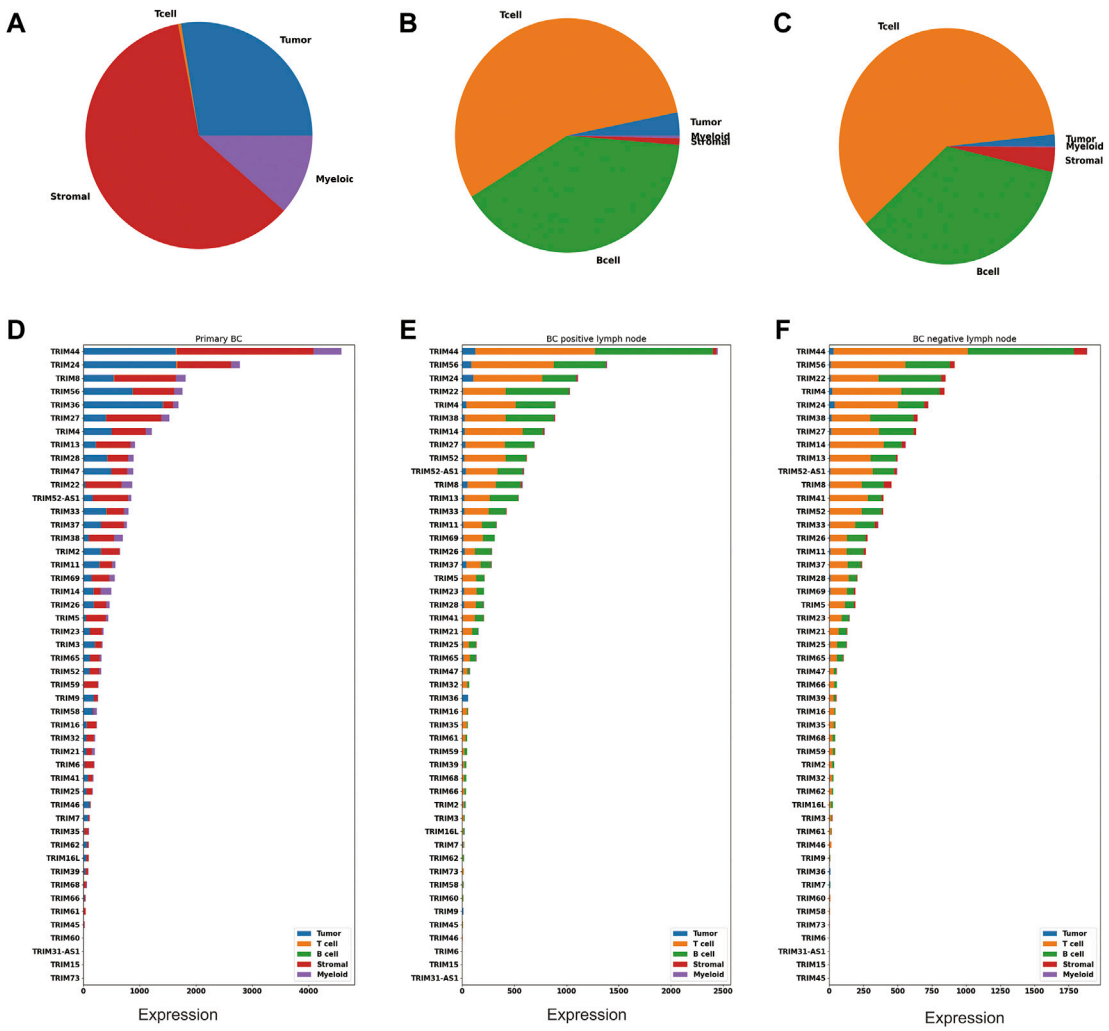


FIGURE 4 | Cell-type expression specificity of TRIMs in BC. Pie plots showing the percentage of the tumor, myeloid, stromal, B, and T cells in **(A)** the primary tumor, **(B)** the positive lymph node, and **(C)** the negative lymph node using GSE158399. Bar plots showing the expression of TRIMs in different cell types for **(D)** the primary tumor, **(E)** the positive lymph node, and **(F)** the negative lymph node.

items are the muscle system process, muscle contraction, and muscle structure development (**Figure 5E**). Contractile fiber, myofibril, and sarcomere are the three most enriched cellular component (CC) items (**Figure 5F**). The three most significant GO items in the molecular function (MF) category are the actin binding/structural constituent of muscle/cytoskeletal protein binding (**Figure 5G**). The KEGG pathway enrichment analysis indicated that these genes were enriched in the cardiac muscle contraction pathway. Reactome pathway enrichment suggested that these genes were enriched in several immune-related pathways (interferon-gamma signaling/antigen processing and presentation/adaptive immune system) beside the pathways related to the muscle contraction (**Figure 5H**).

Since we did not see any gene coexpressed with *TRIM45*, we divided the samples from the TCGA-BRCA cohort into TRIM45-high and TRIM45-low groups according to the median expression of *TRIM45*. A comparison of the expression profile between the two groups identified 961 significantly DEGs. The estrogen signaling pathway was found enriched in these genes by KEGG pathway enrichment analysis (**Figure 6A**). In addition, we found that the TRIM45-high group has a significantly lower proliferation score, apoptosis score, cell cycle score, and TSC mTor score and a significantly higher hormone a score and hormone b score (Mann–Whitney U test, $p < 0.05$) (**Figure 6B**). The results suggested that TRIM45 may be involved in the hormone response and other cellular functions like proliferation. The TRIM45-high group with a lower proliferation score, cell cycle score, and mTor score and higher hormone a/b score may partially explain the better OS of this group.

4 DISCUSSION

The TRIM family has many members, some of which are reported to contribute to the pathogenicity of BC recently. In

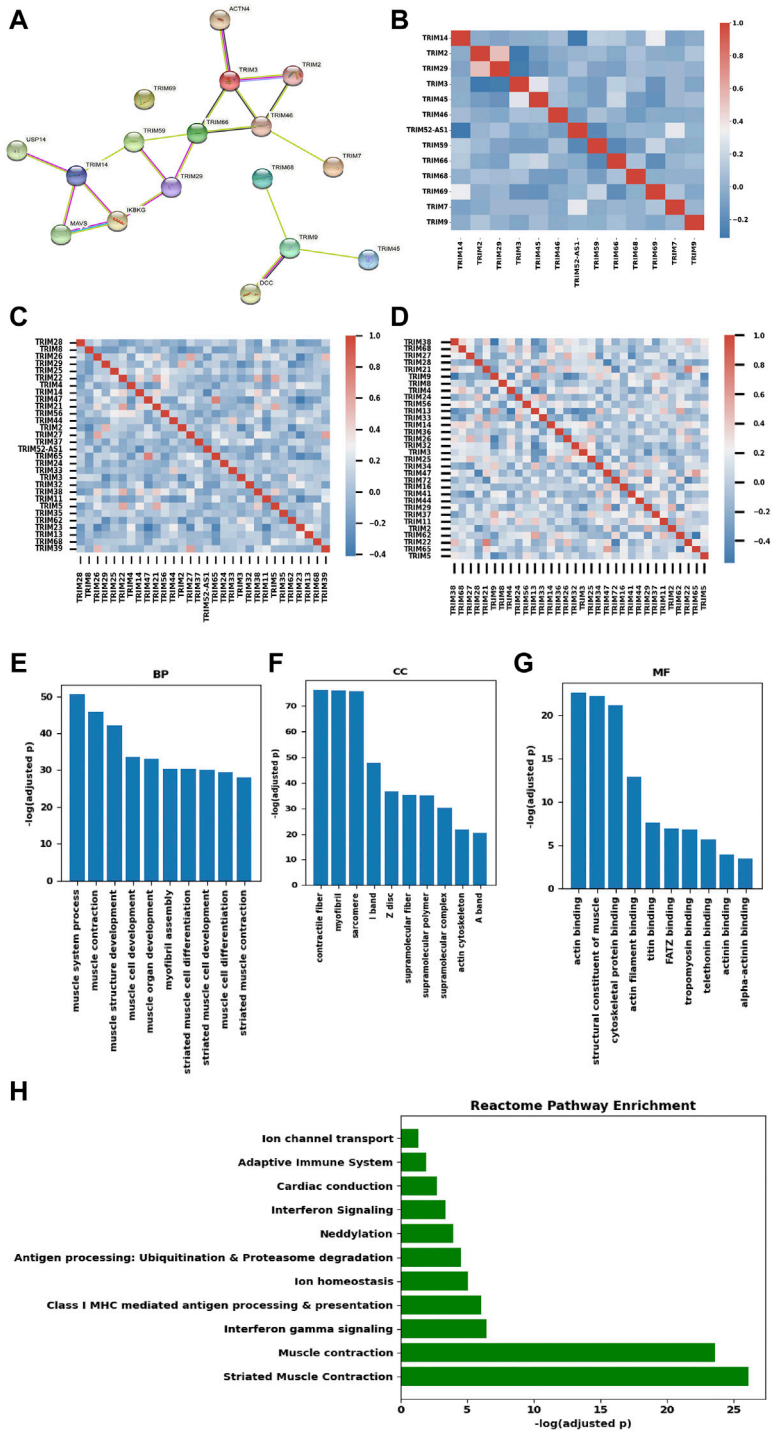


FIGURE 5 | Interaction, co-expression, and functional analysis of TRIMs. **(A)** the protein–protein interaction network of TRIMs generated using STRING. **(B)** heatmap of the Pearson's correlation efficiency of the 13 candidate TRIMs. **(C)** heatmap of the correlation efficiency of top 30 expressed TRIMs in BC. **(D)** heatmap of the correlation efficiency of 32 TRIMs with protein expression in TCGA-BRCA cohort. **(E)** bar plot of the enriched biological process (BP) items. **(F)** bar plot of the enriched cellular component (CC) items. **(G)** bar plot of the enriched molecular function (MF) items. **(H)** Bar plot of the enriched Reactome pathways.

this study, we integrated several public datasets to explore the expression profile of all TRIMs in BC. Nine DEGs (*TRIM59/46/66/52-AS1/68/7/2/9/29*) were identified and validated using two independent datasets. In addition, we found that six TRIMs have prognosis values using two independent datasets. Higher *TRIM3/14/69/45* expression and lower *TRIM68/2* expression were

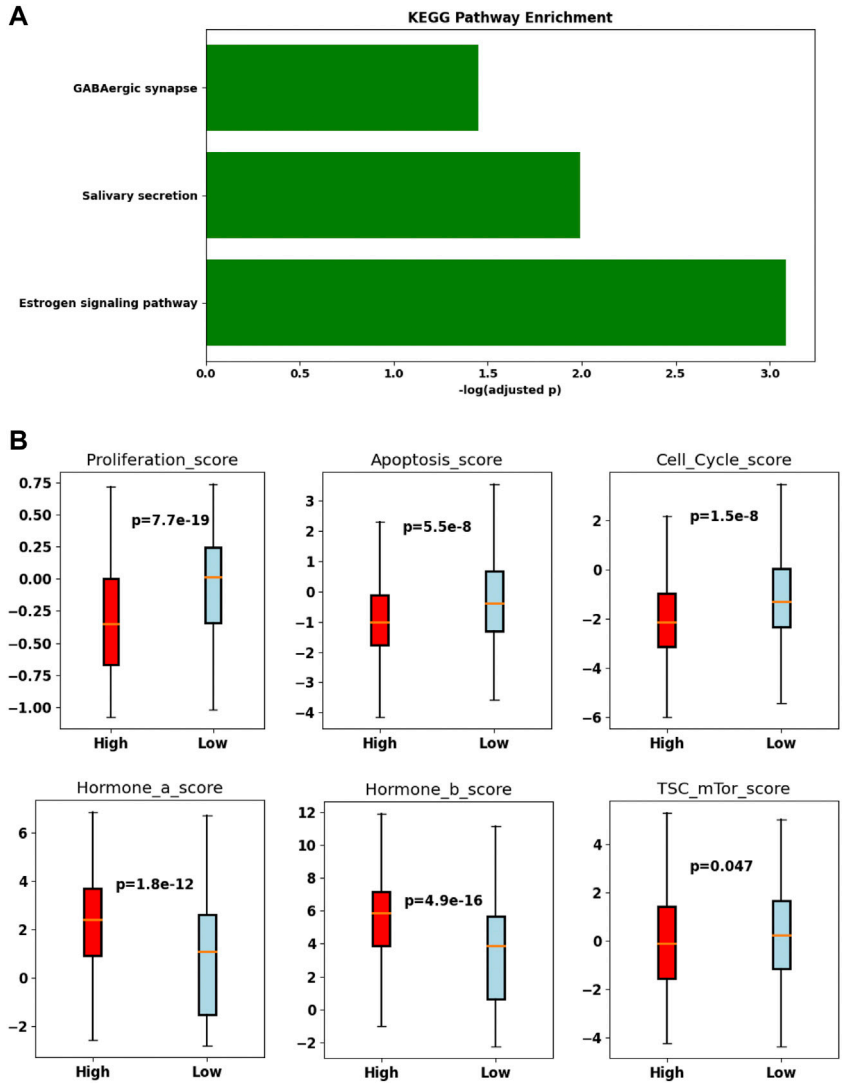


FIGURE 6 | Difference in the TRIM45-group and the TRIM45-low group. **(A)** bar plot of the enriched Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway. **(B)** Boxplot of proliferation score, apoptosis score, cell cycle score, hormone a score, hormone b score, TSC mTor score.

associated with favorable OS in BC patients. The expression of TRIM45 is an independent prognosis factor from patient age, tumor stages, and molecular subtypes.

The alterations of the expression levels of the nine genes may suggest their important role in BC progression. *TRIM59* overexpression can promote BC progression (Liu et al., 2018), while genetic depletion of this gene suppresses BC metastasis (Tan et al., 2019). Knockdown of TRIM66 in the BC cell line can suppress proliferation (Zhang et al., 2021). *TRIM29* is functional as a tumor suppressor in BC (Ai et al., 2014). The functions of *TRIM2/TRIM46/TRIM68/TRIM9/TRIM52-AS1/TRIM7* are not clear in BC yet. In other cancers, *TRIM7* can suppress hepatocellular carcinoma progression (Zhu et al., 2020), while an oncogenic function was reported in colorectal cancer (Cao et al., 2019) and pancreatic cancer (Sun et al., 2020) for *TRIM2* and in hepatocellular carcinoma for *TRIM52-AS1* (Zhou et al., 2021). Of the nine genes, *TRIM2* is downregulated in BC, and it is also found to be associated with a better OS in BC.

Six genes were identified to have prognostic values in BC. Higher *TRIM3/14/69/45* expression and lower *TRIM68/2* expression are associated with favorable OS in BC. The three TRIMs [*TRIM21* (Zhou et al., 2018), *TRIM44* (Kawabata et al., 2017), and *TRIM13* (Chen et al., 2019)] reported to have prognostic values in BC were not validated in our analysis. *TRIM14* was reported to promote BC proliferation (Hu et al., 2019), while the role of *TRIM3* is somewhat contradictory in BC (Liu et al., 2014; Wang et al., 2020). Depletion of TRIM68 can inhibit colorectal cancer cell proliferation (Tan et al., 2017). Limited information is found for *TRIM69*.

TRIM45 is the only gene from the TRIM family with an independent effect on the OS of BC using multivariate Cox

regression analysis. Higher expression of TRIM45 was found in the luminal A and luminal B subtypes than in other molecular subtypes. Higher expression was also found in the stage I group than in the stage III group. In addition, we found that the TRIM45-high group was with higher hormone scores and lower proliferation/cell cycle scores. These may help explain the better OS in the higher group. The single-cell RNA-seq data demonstrated that it is expressed in nontumor cells. It is reported as a tumor suppressor gene in glioma (Zhang et al., 2017), but few studies have reported it can be found in BC. Further functional study of *TIRM45* in hormone response may help clarify its role in BC.

Correlation analysis demonstrates that TRIMs have low expression correlations with each other and with other genes in BC. The low correlations may suggest that they prefer working alone or may be because they function as an ubiquitin ligase, which involves protein-level modifications. Pathway enrichment analysis of the co-expression genes indicates their functions are related to muscle contraction and immunity. The dysregulation of muscle-related pathway in BC was reported before (Bohlen et al., 2018). The single-cell-level expression of TRIMs revealed that most TRIMs are also expressed in the immune and stromal cells, suggesting that they have an important role in the BC microenvironment. Functional studies of their role in the immune or stromal cells may help elucidate their roles in BC progression.

In conclusion, through the bioinformatics analysis, we report nine DEGs and six candidate genes linked to BC prognosis in the TRIM family. Some TRIMs (*TRIM45/46/2/69*) are not well-studied in BC, and *TRIM45* may serve as a novel biomarker for BC.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article, and further inquiries can be directed to the corresponding author.

ETHICS STATEMENT

All data were downloaded from the public database and followed the data access policies. This study was exempt from ethical review by the ethics committee of Shenzhen Second People's Hospital. This study did not involve individual information, so there was no requirement for informed consent.

AUTHOR CONTRIBUTIONS

LN: study design and writing–original draft; QH: analysis and interpretation of data; NX: review. All authors approved the submitted version. LN and QH have contributed equally to this work.

FUNDING

This project was supported by the National Natural Science Foundation of China, China (No. 82172356 and No. 81972003); the Natural Science Foundation of Guangdong, China (Nos. 2021A1515012144 and 2020A1515111165); and the Science, Technology and Innovation Commission of Shenzhen Municipality (No. JCYJ20180507182025817).

REFERENCES

Ai, L., Kim, W.-J., Alpay, M., Tang, M., Pardo, C. E., Hatakeyama, S., et al. (2014). TRIM29 Suppresses TWIST1 and Invasive Breast Cancer Behavior. *Cancer Res.* 74 (17), 4875–4887. doi:10.1158/0008-5472.can-13-3579

Anders, S., Pyl, P. T., and Huber, W. (2015). HTSeq--A Python Framework to Work with High-Throughput Sequencing Data. *Bioinformatics* 31 (2), 166–169. doi:10.1093/bioinformatics/btu638

Bell, J. L., Malyukova, A., Kavallaris, M., Marshall, G. M., and Cheung, B. B. (2013). TRIM16 Inhibits Neuroblastoma Cell Proliferation through Cell Cycle Regulation and Dynamic Nuclear Localization. *Cell Cycle* 12 (6), 889–898. doi:10.4161/cc.23825

Bohlen, J., McLaughlin, S. L., Hazard-Jenkins, H., Infante, A. M., Montgomery, C., Davis, M., et al. (2018). Dysregulation of Metabolic-Associated Pathways in Muscle of Breast Cancer Patients: Preclinical Evaluation of Interleukin-15 Targeting Fatigue. *J. Cachexia Sarcopenia Muscle* 9 (4), 701–714. doi:10.1002/jcsm.12294

Borlepawar, A., Frey, N., and Rangrez, A. Y. (2019). A Systematic View on E3 Ligase Ring TRIMmers with a Focus on Cardiac Function and Disease. *Trends Cardiovasc. Med.* 29 (1), 1–8. doi:10.1016/j.tcm.2018.05.007

Cambiaghi, V., Giuliani, V., Lombardi, S., Marinelli, C., Toffalorio, F., and Pelicci, P. G. (2012). "TRIM Proteins in Cancer," in *TRIM/RBCC Proteins*. Editor G. Meroni (New York, NY: Springer), 77–91. doi:10.1007/978-1-4614-5398-7_6

Cao, H., Fang, Y., Liang, Q., Wang, J., Luo, B., Zeng, G., et al. (2019). TRIM2 is a Novel Promoter of Human Colorectal Cancer. *Scand. J. Gastroenterology* 54 (2), 210–218. doi:10.1080/00365521.2019.1575463

Chen, W. X., Cheng, L., Xu, L. Y., Qian, Q., and Zhu, Y. L. (2019). Bioinformatics Analysis of Prognostic Value of TRIM13 Gene in Breast Cancer. *Biosci. Rep.* 39 (3), BSR20190285. doi:10.1042/BSR20190285

Ciriello, G., Gatza, M. L., Beck, A. H., Wilkerson, M. D., Rhie, S. K., Pastore, A., et al. (2015). Comprehensive Molecular Portraits of Invasive Lobular Breast Cancer. *Cell* 163 (2), 506–519. doi:10.1016/j.cell.2015.09.033

Crawford, L. J., Johnston, C. K., and Irvine, A. E. (2018). TRIM Proteins in Blood Cancers. *J. Cell Commun. Signal.* 12 (1), 21–29. doi:10.1007/s12079-017-0423-5

Davidson-Pilon, C. (2019). Lifelines: Survival Analysis in Python. *J. Open Source Softw.* 4 (40), 1317. doi:10.21105/joss.01317

Gao, J., Aksoy, B. A., Dogrusoz, U., Dresdner, G., Gross, B., Sumer, S. O., et al. (2013). Integrative Analysis of Complex Cancer Genomics and Clinical Profiles Using the cBioPortal. *Sci. Signal* 6 (269), pl1. doi:10.1126/scisignal.2004088

Hu, G., Pen, W., and Wang, M. (2019). TRIM14 Promotes Breast Cancer Cell Proliferation by Inhibiting Apoptosis. *Oncol. Res.* 27 (4), 439–447. doi:10.3727/096504018x15214994641786

Jaworska, A. M., Wlodarczyk, N. A., Mackiewicz, A., and Czerwinska, P. (2020). The Role of TRIM Family Proteins in the Regulation of Cancer Stem Cell Self-Renewal. *Stem Cells* 38 (2), 165–173. doi:10.1002/stem.3109

Jefferies, C., Wynne, C., and Higgs, R. (2011). Antiviral TRIMs: Friend or Foe in Autoimmune and Autoinflammatory Disease? *Nat. Rev. Immunol.* 11 (9), 617–625. doi:10.1038/nri3043

Kawabata, H., Azuma, K., Ikeda, K., Sugitani, I., Kinowaki, K., Fujii, T., et al. (2017). TRIM44 is a Poor Prognostic Factor for Breast Cancer Patients as a Modulator of NF-κB Signaling. *Int. J. Mol. Sci.* 18 (9), 1931. doi:10.3390/ijms18091931

Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., and Salzberg, S. L. (2013). TopHat2: Accurate Alignment of Transcriptomes in the Presence of Insertions,

Deletions and Gene Fusions. *Genome Biol.* 14 (4), R36. doi:10.1186/gb-2013-14-4-r36

Kumarasinghe, L., Xiong, L., Garcia-Gimeno, M. A., Lazzari, E., Sanz, P., and Meroni, G. (2021). TRIM32 and Malin in Neurological and Neuromuscular Rare Diseases. *Cells* 10 (4), 820. doi:10.3390/cells10040820

Liu, Y., Dong, Y., Zhao, L., Su, L., Diao, K., and Mi, X. (2018). TRIM59 Overexpression Correlates with Poor Prognosis and Contributes to Breast Cancer Progression through AKT Signaling Pathway. *Mol. Carcinog.* 57 (12), 1792–1802. doi:10.1002/mc.22897

Liu, Y., Raheja, R., Yeh, N., Ciznadija, D., Pedraza, A. M., Ozawa, T., et al. (2014). TRIM3, a Tumor Suppressor Linked to Regulation of p21Waf1/Cip1. *Oncogene* 33 (3), 308–315. doi:10.1038/onc.2012.596

Lv, D., Li, Y., Zhang, W., Alvarez, A. A., Song, L., Tang, J., et al. (2017). TRIM24 Is an Oncogenic Transcriptional Co-activator of STAT3 in Glioblastoma. *Nat. Commun.* 8 (1), 1–13. doi:10.1038/s41467-017-01731-w

Ning, L., Ko, J. M.-Y., Yu, V. Z., Ng, H. Y., Chan, C. K.-C., Tao, L., et al. (2020). Nasopharyngeal Carcinoma MHC Region Deep Sequencing Identifies HLA and Novel Non-HLA TRIM31 and TRIM39 Loci. *Commun. Biol.* 3 (1), 759. doi:10.1038/s42003-020-01487-y

Ozato, K., Shin, D.-M., Chang, T.-H., and Morse, H. C. (2008). TRIM Family Proteins and Their Emerging Roles in Innate Immunity. *Nat. Rev. Immunol.* 8 (11), 849–860. doi:10.1038/nri2413

Raudvere, U., Kolberg, L., Kuzmin, I., Arak, T., Adler, P., Peterson, H., et al. (2019). g:Profiler: A Web Server for Functional Enrichment Analysis and Conversions of Gene Lists (2019 Update). *Nucleic Acids Res.* 47 (W1), W191–W198. doi:10.1093/nar/gkz369

Rouzier, R., Perou, C. M., Symmans, W. F., Ibrahim, N., Cristofanilli, M., Anderson, K., et al. (2005). Breast Cancer Molecular Subtypes Respond Differently to Preoperative Chemotherapy. *Clin. Cancer Res.* 11 (16), 5678–5685. doi:10.1158/1078-0432.ccr-04-2421

Sardiello, M., Cairo, S., Fontanella, B., Ballabio, A., and Meroni, G. (2008). Genomic Analysis of the TRIM Family Reveals Two Groups of Genes with Distinct Evolutionary Properties. *BMC Evol. Biol.* 8 (1), 225. doi:10.1186/1471-2148-8-225

Siegel, R. L., Miller, K. D., Fuchs, H. E., and Jemal, A. (2021). Cancer Statistics, 2021. *CA A Cancer J. Clin.* 71 (1), 7–33. doi:10.3322/caac.21654

Sørlie, T., Tibshirani, R., Parker, J., Hastie, T., Marron, J. S., Nobel, A., et al. (2003). Repeated Observation of Breast Tumor Subtypes in Independent Gene Expression Data Sets. *Proc. Natl. Acad. Sci.* 100 (14), 8418–8423. doi:10.1073/pnas.0932692100

Sun, Q., Ye, Z., Qin, Y., Fan, G., Ji, S., Zhuo, Q., et al. (2020). Oncogenic Function of TRIM2 in Pancreatic Cancer by Activating ROS-Related NRF2/ITGB7/FAK Axis. *Oncogene* 39 (42), 6572–6588. doi:10.1038/s41388-020-01452-3

Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., et al. (2021). Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA A Cancer J. Clin.* 71 (3), 209–249. doi:10.3322/caac.21660

Szklarczyk, D., Gable, A. L., Nastou, K. C., Lyon, D., Kirsch, R., Pyysalo, S., et al. (2021). The STRING Database in 2021: Customizable Protein-Protein Networks, and Functional Characterization of User-Uploaded Gene/Measurement Sets. *Nucleic acids Res.* 49 (D1), D605–D612. doi:10.1093/nar/gkaa1074

Tan, P., He, L., and Zhou, Y. (2019). TRIM59 Deficiency Curtails Breast Cancer Metastasis through SQSTM1-Selective Autophagic Degradation of PDCD10. *Autophagy* 15 (4), 747–749. doi:10.1080/15548627.2019.1569951

Tan, P., Ye, Y., He, L., Xie, J., Jing, J., Ma, G., et al. (2018). TRIM59 Promotes Breast Cancer Motility by Suppressing P62-Selective Autophagic Degradation of PDCD10. *PLoS Biol.* 16 (11), e3000051. doi:10.1371/journal.pbio.3000051

Tan, Z., Liu, X., Yu, E., Wang, H., Tang, L., Wang, H., et al. (2017). Lentivirus-Mediated RNA Interference of Tripartite Motif 68 Inhibits the Proliferation of Colorectal Cancer Cell Lines SW1116 and HCT116 *In Vitro*. *Oncol. Lett.* 13 (4), 2649–2655. doi:10.3892/ol.2017.5787

van Gent, M., Sparrer, K. M. J., and Gack, M. U. (2018). TRIM Proteins and Their Roles in Antiviral Host Defenses. *Annu. Rev. Virol.* 5, 385–405. doi:10.1146/annurev-virology-092917-043323

Venuto, S., and Merla, G. (2019). E3 Ubiquitin Ligase TRIM Proteins, Cell Cycle and Mitosis. *Cells* 8 (5), 510. doi:10.3390/cells8050510

Waks, A. G., and Winer, E. P. (2019). Breast Cancer Treatment: A Review. *JAMA* 321 (3), 288–300. doi:10.1001/jama.2018.19323

Wang, J., Yang, F., Zhuang, J., Huo, Q., Li, J., and Xie, N. (2022). TRIM58 Inactivates P53/p21 to Promote Chemoresistance via Ubiquitination of DDX3 in Breast Cancer. *Int. J. Biochem. Cell Biol.* 143, 106140. doi:10.1016/j.biocel.2021.106140

Wang, X., Zhang, Y., Pei, X., Guo, G., Xue, B., Duan, X., et al. (2020). TRIM3 Inhibits P53 Signaling in Breast Cancer Cells. *Cancer Cell Int.* 20 (1), 559. doi:10.1186/s12935-020-01630-z

Wei, C., Cheng, J., Zhou, B., Zhu, L., Khan, M. A., He, T., et al. (2016). Tripartite Motif Containing 28 (TRIM28) Promotes Breast Cancer Metastasis by Stabilizing TWIST1 Protein. *Sci. Rep.* 6 (1), 29822. doi:10.1038/srep29822

Wolf, F. A., Angerer, P., and Theis, F. J. (2018). SCANPY: Large-Scale Single-Cell Gene Expression Data Analysis. *Genome Biol.* 19 (1), 15. doi:10.1186/s13059-017-1382-0

Zhang, H., Zheng, Y., and Zhang, Y. (2021). Knockdown of TRIM66 in MDA-MB-468 Triple Negative Breast Cancer Cell Line Suppresses Proliferation and Promotes Apoptosis through EGFR Signaling. *Pol. J. Pathol.* 72 (2), 160–166. doi:10.5114/pjp.2021.109518

Zhang, J., Zhang, C., Cui, J., Ou, J., Han, J., Qin, Y., et al. (2017). TRIM45 Functions as a Tumor Suppressor in the Brain via its E3 Ligase Activity by Stabilizing P53 through K63-Linked Ubiquitination. *Cell Death Dis.* 8 (5), e2831. doi:10.1038/cddis.2017.149

Zhou, C., Chen, Z., Peng, C., Chen, C., and Li, H. (2021). Long Noncoding RNA TRIM52-AS1 Sponges miR-514a-5p to Facilitate Hepatocellular Carcinoma Progression through Increasing MRPS18A. *Cancer Biother. Radiopharm.* 36 (2), 211–219. doi:10.1089/cbr.2019.3271

Zhou, W., Zhang, Y., Zhong, C., Hu, J., Hu, H., Zhou, D., et al. (2018). Decreased Expression of TRIM21 Indicates Unfavorable Outcome and Promotes Cell Growth in Breast Cancer. *Cancer Manag. Res.* 10, 3687–3696. doi:10.2147/cmar.s175470

Zhu, L., Qin, C., Li, T., Ma, X., Qiu, Y., Lin, Y., et al. (2020). The E3 Ubiquitin Ligase TRIM7 Suppressed Hepatocellular Carcinoma Progression by Directly Targeting Src Protein. *Cell Death Differ.* 27 (6), 1819–1831. doi:10.1038/s41418-019-0464-9

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.

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

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