Genome-wide association study across five cohorts identifies five novel loci associated with idiopathic pulmonary fibrosis

Richard J Allen[1], Amy Stockwell[2], Justin M Oldham[3], Beatriz Guillen-Guio[1], Carlos Flores[4,5,6], Imre Noth[7], Brian L Yaspan[2], R Gisli Jenkins[8], Louise V Wain[1,9], International IPF Genetics Consortium

[1] Department of Health Sciences, University of Leicester, Leicester, UK
[2] Genentech, South San Francisco, USA
[3] Department of Internal Medicine, University of California Davis, Davis, USA
[4] Research Unit, Hospital Universitario Ntra. Sra. de Candelaria, Santa Cruz de Tenerife, Spain
[5] CIBER de Enfermedades Respiratorias, Instituto de Salud Carlos III, Madrid, Spain
[6] Genomics Division, Instituto Tecnológico y de Energías Renovables (ITER), Santa Cruz de Tenerife, Spain
[7] Division of Pulmonary & Critical Care Medicine, University of Virginia, Charlottesville, USA
[8] National Heart and Lung Institute, Imperial College London, London, UK
[9] National Institute for Health Research, Leicester Respiratory Biomedical Research Centre, Glenfield Hospital, Leicester, UK

Abstract

Idiopathic pulmonary fibrosis (IPF) is a chronic lung condition with poor survival times. We previously published a genome-wide meta-analysis of IPF risk across three studies with independent replication of associated variants in two additional studies. To maximise power and to generate more accurate effect size estimates, we performed a genome-wide meta-analysis across all five studies included in the previous IPF risk GWAS. We utilised the distribution of effect sizes across the five studies to assess the replicability of the results and identified five robust novel genetic association signals implicating mTOR signalling, telomere maintenance and spindle assembly genes in IPF risk.

Introduction

Idiopathic pulmonary fibrosis (IPF) is a chronic lung disease believed to result from an aberrant response to alveolar injury leading to a build-up of scar tissue. This progressive scarring is eventually fatal with half of individuals dying within 3 to 5 years of diagnosis[1]. The cause of IPF is unknown but genetics play an important role in how susceptible an individual is to IPF[2].

Genome-wide association studies (GWAS) are an approach whereby genetic variants from across the genome are tested for their association with a disease. Genetic loci identified by GWAS can implicate genes important in disease pathogenesis and drugs which target the products encoded by these genetically-supported genes are twice as likely to be successful during development. The genetic association statistics from a GWAS are also widely used to identify causal markers of disease through Mendelian randomisation, to conduct heritability estimation and for genetic correlation analyses. It is therefore important that sample sizes are maximised to ensure sufficient statistical power to detect genetic associations and to generate precise effect size estimates.

We recently published a GWAS of IPF risk[2]. The discovery GWAS consisted of three studies (named as the UK, Chicago and Colorado studies) and a replication analysis performed in two independent studies (named as the UUS [USA, UK and Spain] and Genentech studies). This analysis reported 14 genetic signals which implicated host defence, cell-cell adhesion, spindle assembly, TGF-β signalling regulation and telomere maintenance as important biological processes involved in IPF disease risk.

We here present a meta-analysis of genome-wide data from all 5 datasets included in our previous study. The results of this analysis implicate new genetic loci in IPF pathogenesis and provide a unique resource for other studies of IPF risk and pathogenesis.

Methods

Quality control and sample selection have been previously described[2]. In summary, all datasets comprised of unrelated European-ancestry individuals. Individuals in the Genentech study were sequenced using HiSeq X Ten platform (Illumina) and all other individuals were imputed from genotyping data using the HRC reference panel[3]. Genome-wide analyses were performed in each study separately using an additive logistic regression model adjusting for the first 10 genetic principal components to account for population stratification.

The individual GWAS results from the five studies were meta-analysed using an inverse-variance weighted fixed effect meta-analysis using METAL[4]. Variants were included in the meta-analysis if they were available in at least four studies. Genomic control was performed on the meta-analysis results using the LD score regression intercept to account for inflation not explained by polygenic effects[5]. Significant variants were defined as those with meta-analysis $p<5\times10^{-8}$ and conditional analyses were performed using GCTA-COJO to identify additional independent associated variants[6]. Independent associated variants were defined as variants remaining genome-wide significant after conditioning on the most significant variant (sentinel) in the region with consistent effect size estimates in the conditional and non-conditional analysis. Annotation of the sentinel variants was then performed using Variant Effect Predictor[7].

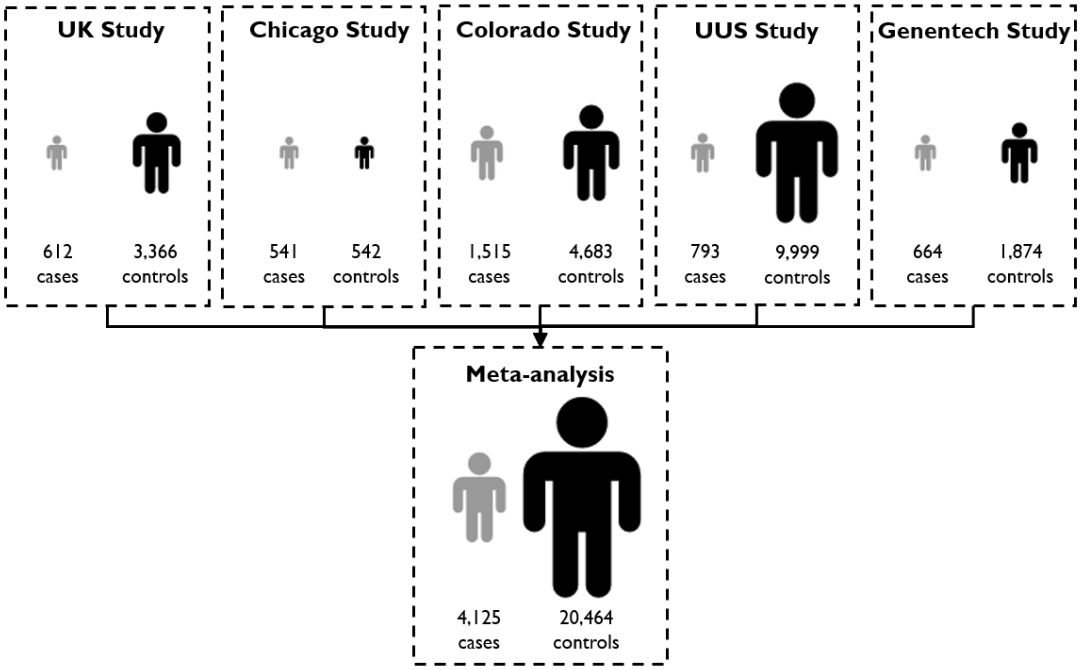
To assess the robustness of novel results, we tested the strength and consistency of results across studies using MAMBA (Meta-Analysis Model-based Assessment of replicability)[8]. Variants with a posterior probability of replicability (PPR)≥90% were considered robust and likely to replicate should additional independent datasets become available.

Genome-wide summary statistics can be accessed at https://github.com/genomicsITER/PFgenetics.

Results

A total of 4,125 cases, 20,464 controls and 7,554,248 genetic variants were included in the analysis (**Figure 1**). The UUS study included one additional case (due to resolving a sample ID issue since the previous publication) and one fewer control (where the individual has since withdrawn consent from UK Biobank) than described in the previous GWAS[2].

Figure 1: Study design and sample sizes



After conditional analyses, there were 23 independent signals with $p<5\times10^{-8}$ in the genome-wide meta-analysis (**Figure 2**). These 23 signals included all 14 associations reported in the previous GWAS (**Supplementary Table 1**). Of the nine novel genetic associations (**Table 1**), five showed evidence of replicability (PPR≥90%). The sentinel variants of these five loci included variants in introns of *KNL1*, *NPRL3*, *STMN3* and *RTEL1,* and an intergenic variant in 10q25.1. All five novel variants had consistent direction of effect across all of the individual studies and reached nominal significance ($p<0.05$) in at least 3 of the studies.

Figure 2: Manhattan plot. Each point shows a genetic variant with chromosomal position on the x axis and the –log(p value) on the y axis. The grey dashed line shows the genome-wide significance level ($p=5\times10^{-8}$). Each signal is labelled with the gene implicated by that signal. Genes in grey are the novel loci that do show evidence of replicability. The plot has been truncated at $p=10^{-30}$.

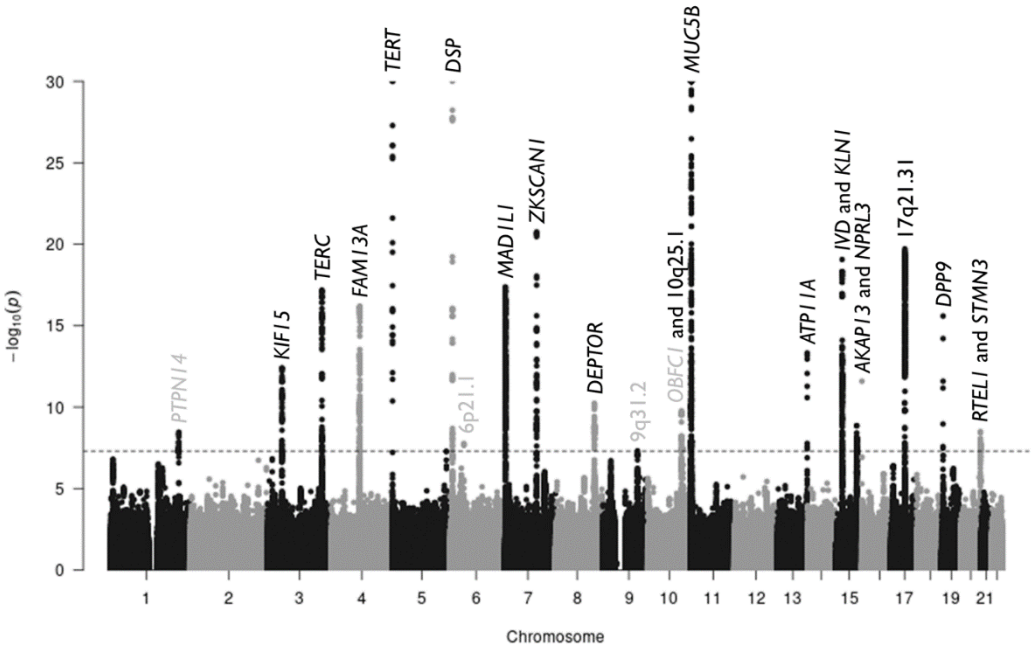


Table 1: Sentinel variants of novel associations. Novel variants are defined as those not reaching significance criteria in previous analysis[2] (the *RTEL1* and *OBFC1* signals have previously shown a possible association – see discussion). Effect sizes and directions are given in terms of the allele that increases risk of IPF. Chr=Chromosome. Position is based on genetic build 37. Annotation obtained from Variant Effect Predictor[7]. EAF=Effect allele frequency calculated across the five studies. The "Direction" column shows the direction of the beta in each of the five individual studies (+ means beta>0, − means beta<0). The "Study p≤0.05" column denotes which individual studies the variant reached nominal significance in (Y means p≤0.05, N means p>0.05). Both the direction and study p<0.05 are given in the order UK, Colorado, Chicago, UUS and then Genentech. OR=Odds ratio. CI=Confidence interval. PPR=posterior probability of replicability calculated using MAMBA[8]. [a] The signal at *KNL1* is independent of the previously reported nearby signal in the *IVD* gene. [b] The *RTEL1* and *STMN3* signals are independent of each other.

Chr	Position	rsid	Annotation	Ref allele	Effect allele	EAF	Direction	Study p≤0.05	OR [95% CI]	p	PPR
i) Novel variants with high posterior probability of replication (PPR≥90%)											
10	111229861	rs79684490	Intergenic (10q25.1)	G	A	4.6%	+ + + + +	YYNYY	1.40 [1.24, 1.57]	3.52×10^{-8}	**94.0%**
15	40931708	rs12912339 [a]	Intron of *KNL1*	G	A	15.9%	+ + + + +	YYNYY	1.30 [1.21, 1.39]	7.41×10^{-13}	**96.5%**
16	162240	rs74614704	Intron of *NPRL3*	G	A	5.6%	+ + + + +	YNNYY	1.49 [1.33, 1.67]	2.57×10^{-12}	**99.4%**
20	62284170	rs112087793 [b]	Intron of *STMN3*	T	C	91.5%	+ + + + +	YYYYY	1.34 [1.21, 1.48]	1.09×10^{-8}	**96.8%**
20	62324391	rs41308092 [b]	Intron of *RTEL1*	G	A	2.1%	+ + + + +	YYYYN	1.75 [1.45, 2.10]	3.13×10^{-9}	**99.9%**
ii) Novel variants not reaching PPR≥90% threshold											
1	214659598	rs4233306	Intron of *PTPN14*	T	C	80.2%	+ + + + +	YYNNN	1.23 [1.15, 1.32]	3.41×10^{-9}	37.4%
6	43352980	rs1214759	Intergenic (6p21.2)	A	G	67.9%	+ + + + +	NYYYN	1.18 [1.11, 1.25]	1.71×10^{-8}	21.9%
9	109480268	rs11788059	Regulatory region variant (9q31.2)	T	C	34.2%	+ + + + +	NYNYY	1.17 [1.10, 1.23]	4.85×10^{-8}	3.1%
10	105640978	rs7100920	Regulatory region of *OBFC1*	C	T	49.0%	+ + − + +	NYNYY	1.19 [1.13, 1.26]	1.67×10^{-10}	32.1%

Discussion

By increasing the number of cases in the discovery analysis by more than 50% compared with the previous IPF risk GWAS, we identified novel genetic signals associated with IPF risk and improved the precision of estimations for previously reported signals. The five novel loci had internal evidence of replicability giving us confidence that these signals are likely to be generalisable.

The signals in *RTEL1* and *OBFC1* have been reported previously but did not meet the significance criteria of the previous three-way GWAS[2]. The new MAMBA analysis suggests that the consistency of effect across studies provides high confidence that the *RTEL1* signal will replicate should an independent dataset become available. This is not the case for the *OBFC1* signal where a low posterior probability of replication suggests that there may be heterogeneity in effect across the contributing studies.

The novel signals require further characterisation to determine the likely causal gene and underlying functional effect of the variants. However, some of the genes that are closest to these new signals have strong candidacy for involvement in IPF pathogenesis. *NPRL3* encodes a GATOR1 complex function component and acts through mTORC1 signalling to inhibit mTOR kinase activity[9]. mTOR regulates TGF-β collagen synthesis and inhibiting mTOR leads to increased deposition of scar tissue[10]. We previously reported an association implicating *DEPTOR*, another mTOR inhibiting gene. We also add to the evidence that cellular ageing plays a key role in IPF pathogenesis through associations at the telomere maintenance genes *TERT*, *TERC* and *RTEL1*. We previously reported associations in spindle assembly genes (*MAD1L1* and *KIF15*) and have identified a novel genetic association in another spindle assembly gene *KNL1* (Kinetochore Scaffold 1 also known as *CASC5*). *STMN3* (Stathmin 3) implicates another cell replication process through tubulin binding[9].

By maximising the statistical power of the analysis, we identified novel genetic associations with IPF risk. These signals may implicate biologically relevant genes that support the importance of TGF-β signalling and cell replication as important processes in disease pathogenesis.

Ethics Statement

This research was conducted using previously published work with appropriate ethics approval. The PROFILE study (which provided samples for the UK and UUS studies) had institutional ethics approval at the University of Nottingham (NCT01134822 – ethics reference 10/H0402/2) and Royal Brompton and Harefield NHS Foundation Trust (NCT01110694 – ethics reference 10/H0720/12). Spanish samples were recruited under ethics approval by ethics committee from the Hospital Universitario N.S. de Candelaria (reference of the approval: PI-19/12). The UUS study also included individuals from clinical trials with ethics approval (ACE [NCT00957242] and PANTHER [NCT00650091]). UK samples were recruited across multiple sites with individual ethics approval (University of Edinburgh Research Ethics Committee [The Edinburgh Lung Fibrosis Molecular Endotyping (ELFMEN) Study NCT04016181] 17/ES/0075, NRES Committee South West – Southmead, Yorkshire and Humber Research Ethics Committee 08/H1304/54 and Nottingham Research Ethics Committee 09/H0403/59). For individuals recruited at the University of Chicago, consenting patients with IPF who were prospectively enrolled in the institutional review board-approved ILD registry (IRB#14163A) were included. Individuals recruited at the University of Pittsburgh Medical Centre had ethics approval from the University of Pittsburgh Human Research Protection Office (reference STUDY20030223: Genetic Polymorphisms in IPF). Individuals from the COMET (NCT01071707) and Lung Tissue Research Consortium (NCT02988388) studies were also included in the Chicago study. All subjects in the Colorado study gave written informed consent as part of IRB-approved protocols for their recruitment at each site and the GWAS study was approved by the National Jewish Health IRB and Colorado Combined Institutional Review Boards (COMIRB). Subjects in the Genentech study provided written informed consent for whole-genome sequencing of their DNA. Ethical approval was provided as per the original clinical trials (INSPIRE [NCT00075998], RIFF [NCT01872689], CAPACITY [NCT00287729 and NCT00287716] and ASCEND [NCT01366209]). For the USCF cohort, sample and data collection were approved by the University of California San Francisco Committee on Human Research and all patients provided written informed consent. For the Vanderbilt cohort, the Institutional Review Boards from Vanderbilt University approved the study and all participants provided written informed consent before enrolment.

Conflicts of Interest and Funding

R Allen is an Action for Pulmonary Fibrosis Mike Bray Research Fellow. A Stockwell and B Yaspan are employees of Genentech/Roche and hold stock and stock options in Roche. J Oldham reports National Institute of Health/National Heart, Lung and Blood Institute grants R56HL158935 and K23HL138190 and personal fees from Boehringer Ingelheim, Genentech, United Therapeutics, AmMax Bio and Lupin pharmaceuticals unrelated to the submitted work. B Guillen-Guio is supported by Wellcome Trust grant 221680/Z/20/Z. G Jenkins is a trustee of Action for Pulmonary Fibrosis and reports personal fees from Astra Zeneca, Biogen, Boehringer Ingelheim, Bristol Myers Squibb, Chiesi, Daewoong, Galapagos, Galecto, GlaxoSmithKline, Heptares, NuMedii, PatientMPower, Pliant, Promedior, Redx, Resolution Therapeutics, Roche, Veracyte and Vicore. L Wain holds a GSK/British Lung Foundation Chair in Respiratory Research (C17-1). The research was partially supported by the National Institute for Health Research (NIHR) Leicester Biomedical Research Centre; the views expressed are those of the author(s) and not necessarily those of the National Health Service (NHS), the NIHR or the Department of Health. The UK and UUS studies selected controls from UK Biobank under application 648. This research used the SPECTRE High Performance Computing Facility at the University of Leicester.

References

1. Lederer DJ, Martinez FJ. Idiopathic pulmonary fibrosis. *N Engl J Med*. 2018;378(19):1811-1823.

2. Allen RJ, Guillen-Guio B, Oldham JM, et al. Genome-wide association study of susceptibility to idiopathic pulmonary fibrosis. *American journal of respiratory and critical care medicine*. 2020;201(5):564-574.

3. McCarthy S, Das S, Kretzschmar W, Durbin R, Abecasis G, Marchini J. A reference panel of 64,976 haplotypes for genotype imputation. *bioRxiv*. 2016:035170.

4. Willer CJ, Li Y, Abecasis GR. METAL: Fast and efficient meta-analysis of genomewide association scans. *Bioinformatics*. 2010;26(17):2190-2191.

5. Bulik-Sullivan BK, Loh P, Finucane HK, et al. LD score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat Genet*. 2015;47(3):291-295.

6. Yang J, Ferreira T, Morris AP, et al. Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nat Genet*. 2012;44(4):369-75, S1-3.

7. McLaren W, Gil L, Hunt SE, et al. The ensembl variant effect predictor. *Genome Biol*. 2016;17(1):122.

8. McGuire D, Jiang Y, Liu M, et al. Model-based assessment of replicability for genome-wide association meta-analysis. *Nature communications*. 2021;12(1):1-14.

9. Stelzer G, Rosen R, Plaschkes I, et al. GeneCards – the human gene database. The GeneCards suite: From gene data mining to disease genome sequence analysis. *Current protocols in bioinformatics*. 2016(54):1.30.1-1.30.33.

10. Woodcock HV, Eley JD, Guillotin D, et al. The mTORC1/4E-BP1 axis represents a critical signaling node during fibrogenesis. *Nature communications*. 2019;10(1):6.

Supplement

International IPF Genetics Consortium

Writing Group

Richard J Allen, Carlos Flores, Beatriz Guillen-Guio, R Gisli Jenkins, Imre Noth, Justin M Oldham, Amy Stockwell, Louise V Wain, Brian L Yaspan

UK Study

Richard J Allen, Helen L Booth, William A Fahy, Ian P Hall, Simon P Hart, Mike R Hill, Nik Hirani, Richard B Hubbard, R Gisli Jenkins, Toby M Maher, Robin J McAnulty, Ann B Millar, Philip L Molyneaux, Vidya Navaratnam, Eunice Oballa, Helen Parfrey, Gauri Saini, Ian Sayers, Martin D Tobin, Louise V Wain, Moira K B Whyte

Chicago Study

Ayodeji Adegunsoye, Carlos Flores, Naftali Kaminski, Shwu-Fan Ma, Imre Noth, Justin M Oldham, Mary E Strek, Yingze Zhang

Colorado Study

Tasha E Fingerlin, David A Schwartz

UUS Study

Richard J Allen, Carlos Flores, Beatriz Guillen-Guio, R Gisli Jenkins, Shwu-Fan Ma, Toby M Maher, Maria Molina-Molina, Philip L Molyneaux, Imre Noth, Justin M Oldham, Louise V Wain

Genentech study

Margaret Neighbors, Xuting Sheng, Amy Stockwell, Brian L Yaspan

Supplementary Table 1: All sentinel variants associated with IPF risk

This table includes the most associated variant (sentinel) for the 19 signals (14 previously reported loci and the five novel loci identified here) associated with IPF risk. The risk allele is the allele associated with increased risk of IPF. Position is for genetic build 37. Chr=chromosome. EAF=Effect allele frequency. OR=Odds ratio. CI=Confidence interval.

Chr	Position	rsid	Implicated gene	Non-effect allele	Risk allele	EAF	OR [95% CI]	p
3	44903434	rs2292181	*KIF15*	G	C	5.2%	1.52 [1.36, 1.70]	3.95×10^{-13}
3	169486271	rs9860874	*TERC*	C	A	27.6%	1.29 [1.22, 1.37]	6.49×10^{-18}
4	89837808	rs2609259	*FAM13A*	C	A	22.4%	1.30 [1.22, 1.39]	6.47×10^{-17}
5	1282414	rs7725218	*TERT*	A	G	67.1%	1.41 [1.33, 1.50]	4.90×10^{-32}
6	7563232	rs2076295	*DSP*	T	G	46.7%	1.49 [1.41, 1.57]	1.50×10^{-48}
7	1868761	rs12537430	*MAD1L1*	A	G	62.5%	1.28 [1.21, 1.35]	4.20×10^{-18}
7	99630342	rs2897075	*ZKSCAN1*	C	T	38.2%	1.30 [1.23, 1.37]	1.77×10^{-21}
8	120940206	rs10808505	*DEPTOR*	G	T	57.3%	1.20 [1.13, 1.26]	6.03×10^{-11}
10	111229861	rs79684490	10q25.1	G	A	4.6%	1.40 [1.24, 1.57]	3.52×10^{-8}
11	1241221	rs35705950	*MUC5B*	G	T	14.5%	5.06 [4.69, 5.47]	9.09×10^{-418}
13	113534984	rs9577395	*ATP11A*	G	C	79.1%	1.29 [1.21, 1.38]	4.78×10^{-14}
15	40716253	rs2304645	*IVD*	G	C	52.6%	1.28 [1.21, 1.35]	8.66×10^{-20}
15	40931708	rs12912339	*KNL1*	G	A	15.9%	1.30 [1.21, 1.39]	7.41×10^{-13}
15	86287910	rs11073517	*AKAP13*	C	T	32.7%	1.19 [1.13, 1.26]	1.36×10^{-9}
16	162240	rs74614704	*NPRL3*	G	A	5.6%	1.49 [1.33, 1.67]	2.57×10^{-12}
17	44214888	rs2077551	17q21.31	C	T	80.7%	1.42 [1.32, 1.53]	1.92×10^{-20}
19	4717672	rs12610495	*DPP9*	A	G	30.6%	1.28 [1.21, 1.36]	2.58×10^{-16}
20	62284170	rs112087793	*STMN3*	T	C	91.5%	1.34 [1.21, 1.48]	1.09×10^{-8}
20	62324391	rs41308092	*RTEL1*	G	A	2.1%	1.75 [1.45, 2.10]	3.13×10^{-9}