











# *GIGYF1* loss of function is associated with clonal mosaicism and adverse metabolic health

Yajie Zhao <sup>1</sup>, Stasa Stankovic <sup>1</sup>, Mine Koprulu<sup>1</sup>, Eleanor Wheeler <sup>1</sup>, Felix R. Day <sup>1</sup>, Hana Lango Allen<sup>1</sup>, Nicola D. Kerrison<sup>1</sup>, Maik Pietzner <sup>1</sup>, Po-Ru Loh <sup>2,3</sup>, Nicholas J. Wareham <sup>1</sup>, Claudia Langenberg <sup>1</sup>, Ken K. Ong <sup>1</sup> & John R. B. Perry <sup>1</sup>✉

Mosaic loss of chromosome Y (LOY) in leukocytes is the most common form of clonal mosaicism, caused by dysregulation in cell-cycle and DNA damage response pathways. Previous genetic studies have focussed on identifying common variants associated with LOY, which we now extend to rarer, protein-coding variation using exome sequences from 82,277 male UK Biobank participants. We find that loss of function of two genes—*CHEK2* and *GIGYF1*—reach exome-wide significance. Rare alleles in *GIGYF1* have not previously been implicated in any complex trait, but here loss-of-function carriers exhibit six-fold higher susceptibility to LOY (OR = 5.99 [3.04–11.81],  $p = 1.3 \times 10^{-10}$ ). These same alleles are also associated with adverse metabolic health, including higher susceptibility to Type 2 Diabetes (OR = 6.10 [3.51–10.61],  $p = 1.8 \times 10^{-12}$ ), 4 kg higher fat mass ( $p = 1.3 \times 10^{-4}$ ), 2.32 nmol/L lower serum IGF1 levels ( $p = 1.5 \times 10^{-4}$ ) and 4.5 kg lower handgrip strength ( $p = 4.7 \times 10^{-7}$ ) consistent with proposed *GIGYF1* enhancement of insulin and IGF-1 receptor signalling. These associations are mirrored by a common variant nearby associated with the expression of *GIGYF1*. Our observations highlight a potential direct connection between clonal mosaicism and metabolic health.

<sup>1</sup>MRC Epidemiology Unit, Institute of Metabolic Science, University of Cambridge, Cambridge, UK. <sup>2</sup>Division of Genetics, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA, USA. <sup>3</sup>Broad Institute of MIT and Harvard, Cambridge, MA, USA. ✉email: [john.perry@mrc-epid.cam.ac.uk](mailto:john.perry@mrc-epid.cam.ac.uk)

**M**osaic loss of the Y chromosome in leukocytes (LOY) is the most common form of clonal mosaicism, first noted over fifty years ago<sup>1,2</sup>. It has been associated with the risk of a number of complex diseases and traits, however, the biological mechanisms underpinning these observations are unclear<sup>3–10</sup>. Like other forms of clonal mosaicism, LOY is strongly associated with age, reflecting greater opportunity for mitotic errors in hemopoietic stem cell division and subsequent clonal expansion to occur. Predisposition to LOY also has a heritable component and to date, over 150 associated common genetic variants have been identified<sup>11–14</sup>. These loci have implicated genes involved in cell-cycle fidelity and DNA damage response (DDR), supporting the idea that LOY is a readily detectable manifestation of subtle defects in these processes<sup>12,13</sup>. We have hypothesized that the predisposition to genomic instability that is shared across multiple cell types, including leukocytes, may explain the observational associations between LOY and other health outcomes<sup>13</sup>. This concept is most apparent for *CHEK2* loss of function, which both promotes LOY in men and extends reproductive life in women through the shared mechanism of inhibiting DNA damage sensing and apoptosis. Identifying novel genetic determinants associated with LOY has the potential, therefore, to not only increase our knowledge of clonal hematopoiesis but also to identify loci that underlie susceptibility to other complex traits through shared biological mechanisms. We previously demonstrated this with Type 2 Diabetes (T2D), where overlap with LOY highlights loci which likely impact cellular homeostasis in metabolic tissues. For example, alleles in *CCND2* increase the risks of both T2D and LOY<sup>13</sup>, with this gene encoding the major D-type cyclin that is expressed in pancreatic  $\beta$ -cells and is essential for adult  $\beta$  cell growth<sup>15</sup>.

To date, genetic studies for LOY have focussed on genotype-array imputed common genetic variation, which largely misses the contributions of rarer, often more deleterious, alleles<sup>11–13</sup>. Here, we report an exome-sequence GWAS for LOY, assessing the role of rare protein-coding variation. We extend and confirm previous observations supporting the role of *CHEK2* and additionally identify an association with *GIGYF1* loss of function, highlighting an intriguing link between LOY and metabolic health.

## Results

Previous studies have quantified LOY using either a quantitative measure derived from the mean log<sub>2</sub>-transformed R ratio of signal intensity (mLRR-Y)<sup>11</sup> or more recently a more-powered dichotomous measure (PAR-LOY) using allele-specific genotyping intensities in the sex chromosome pseudo-autosomal region (PAR)<sup>13</sup>. We note that both measures are proxies for the abundance of Y chromosome genetic material in the measured biological samples, derived from intensity data which contains much experimental ‘noise’. As these measures are independent—one relies on PAR genotypes only whilst the other excludes them—we hypothesized that an aggregate of the two would further help improve the signal-to-noise ratio of these measures and therefore increase statistical power to detect genetic associations. We name this combined quantitative measure PAR-LOYq (Online Methods) and estimated it in the same UK Biobank participants who were previously studied for PAR-LOY ( $N = 205,011$  men). As expected PAR-LOYq calls provided a more powerful measure for discovery analysis, with a median 10.6% increase in chi-square association statistic for the 156 LOY loci previously identified by PAR-LOY (Supplementary Data 1)<sup>13</sup>.

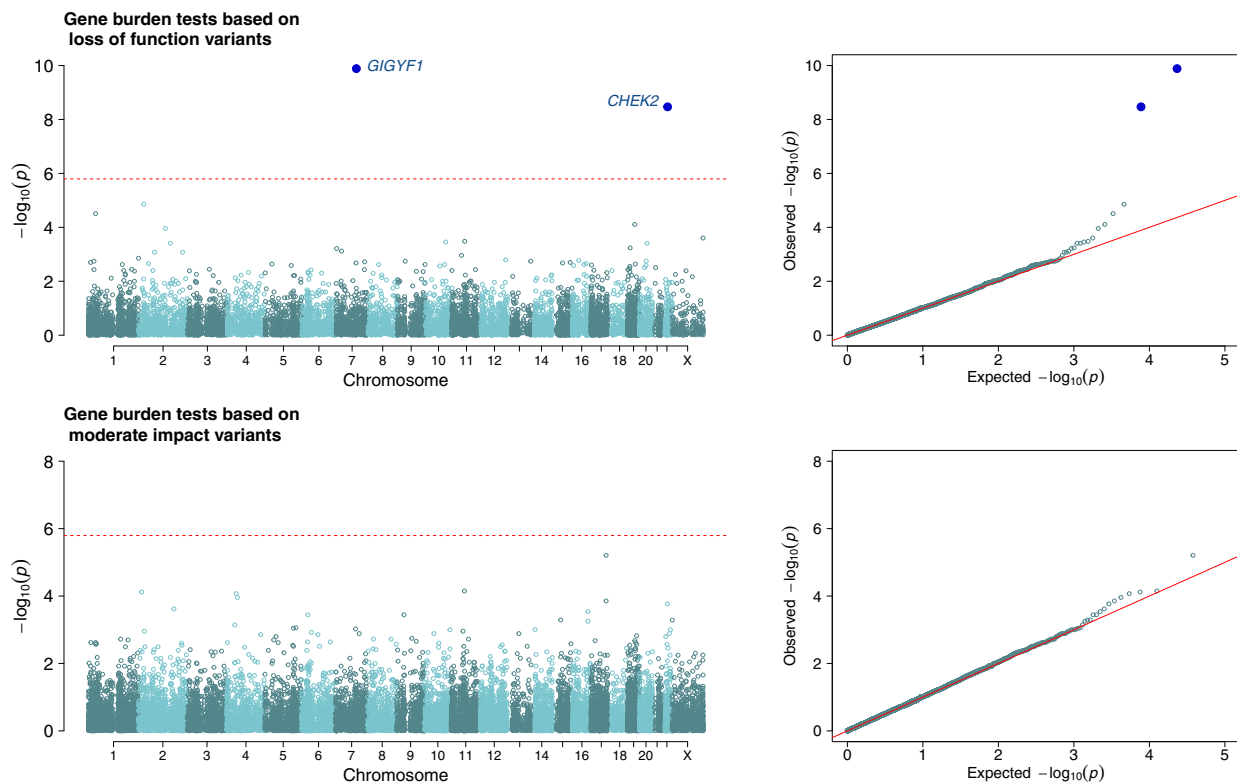
To identify genes associated with LOY, we performed gene burden analyses for PAR-LOYq in 82,277 male UK Biobank

participants with exome sequence data (Online Methods). Two models were tested exome-wide, by collapsing together rare ( $MAF < 0.5\%$ ) loss of function or moderate-impact variants in each individual gene. The association of the burden test in two genes, *CHEK2* and *GIGYF1*, were statistically significant exome-wide ( $p < 1.6 \times 10^{-6}$ ) across these analyses (Supplementary Data 2, Fig. 1). Loss of function variants in *CHEK2* ( $N = 543$  carriers, effect = 0.23 SD higher PAR-LOYq between rare allele carriers vs. non-carriers,  $p = 3.4 \times 10^{-9}$ ) have previously been implicated with LOY as the most common frameshift variant (1100delC,  $MAF \sim 0.2\%$ ) is well captured by GWAS imputation and directly genotyped on the UKBB array. This single variant accounted for 76% of loss of function carriers and the *CHEK2* association was nominally significant when it was excluded ( $p = 0.02$ , effect = 0.18 SD). An independent burden test of rare moderate-impact alleles in *CHEK2* (not including 1100delC and other loss of function alleles) was also associated with PAR-LOYq (Supplementary Data 2,  $N = 1057$  carriers, effect = 0.11 SDs,  $p = 1.7 \times 10^{-4}$ ).

*GIGYF1* loss of function ( $N = 40$  male carriers) was associated with a 0.93 SD (0.64–1.21,  $p = 1.3 \times 10^{-10}$ ) higher PAR-LOYq. This burden signal combined the effects of 27 rare variants (Supplementary Data 3); a single base insertion frameshift with 10 carriers, 4 doubletons, and 22 singleton rare alleles. No individual variant was more significant than the overall *GIGYF1* test result, which remained significant in a leave-one-out analysis of each variant (Supplementary Data 4). Rare moderate-impact alleles were not associated with LOY in aggregate ( $p = 0.70$ ), however, several individual moderate-impact variants exhibited nominally significant associations (Supplementary Data 3). We note that missense alleles likely represent a heterogeneous collection of loss of function, gain of function, and benign effects. As with *CHEK2*, bioinformatic filters poorly predicted which missense variants in *GIGYF1* were associated with LOY (Fig. S1). Further genome-wide burden analyses in STAAR (see methods), weighting each variant by its CADD score, did not identify additional LOY-associated genes (Supplementary Data 5).

We next performed several sensitivity analyses to further explore the genetic architecture of this *GIGYF1*-LOY association. Firstly, we observed consistent effects using the two previous LOY traits, with a 6-fold (OR = 5.99 [3.04–11.81],  $p = 6 \times 10^{-7}$ ) higher risk of a PAR-LOY dichotomous call and a  $-0.038$  ( $\sim 0.81$  SD,  $p = 8.8 \times 10^{-9}$ ) reduction in mLRRy. Secondly, in a sensitivity analysis, PAR-LOYq association results were highly consistent when excluding multi-allelic sites ( $p = 8.4 \times 10^{-9}$ ) or indels ( $p = 9.9 \times 10^{-3}$ ) and when restricting to high-confidence loss of function variants defined by LOFTEE ( $p = 4.1 \times 10^{-13}$ )<sup>16</sup>. Sequencing quality control parameters for each individual variant appeared robust (Supplementary Data 3). Thirdly, we reproduced the same association signal using a second analytical pipeline implemented in STAAR ( $p = 1.73 \times 10^{-10}$ )<sup>17</sup>. Finally, we showed that *GIGYF1* loss of function was not associated with any genetically derived principal component and carriers were geographically dispersed across the UK (Figs. S2–S3).

*GIGYF1* is named after its known binding to growth factor receptor-bound protein 10 (GRB10) and interacts with both the insulin and *IGF1* receptors<sup>18</sup>. We, therefore, postulated that loss of function alleles may also impact metabolic health, and, therefore, tested *GIGYF1* burden test association analyses across 17 metabolic-health related traits in men and women (Supplementary Data 6). *GIGYF1* loss of function ( $N = 64$  carriers) was associated with higher susceptibility to T2D (OR = 6.10 [3.51–10.61],  $p = 1.8 \times 10^{-12}$ ) and higher acute and longer-term average levels of glycemia in non-diabetic individuals (random glucose  $p = 2.6 \times 10^{-5}$  and HbA1c  $p = 6.6 \times 10^{-7}$ ). Of the 64 carriers, 19 (30%) had T2D, compared to 7.1% in the population



**Fig. 1** Manhattan and Quantile-Quantile (QQ) plots for exome-wide gene burden test statistics. The dashed red line denotes the exome-wide significance threshold ( $p < 1.6 \times 10^{-6}$ ). Burden tests performed in  $N = 82,277$  males.

of UK Biobank in whom sequence data was available. Carrier status was also associated with a  $1.85 \text{ kg/m}^2$  higher body mass index ( $p = 5.3 \times 10^{-4}$ ),  $4 \text{ kg}$  higher fat mass ( $p = 1.3 \times 10^{-4}$ ),  $1.85 \text{ kg}$  higher lean mass ( $p = 5.2 \times 10^{-3}$ ),  $0.04$  higher waist-to-hip ratio ( $p = 1.8 \times 10^{-6}$ ),  $-0.01$  lower sitting to standing height ratio ( $p = 4.3 \times 10^{-7}$ ),  $4.5 \text{ kg}$  lower grip strength ( $p = 4.7 \times 10^{-7}$ ) and  $2.32 \text{ nmol/L}$  lower serum IGF1 levels ( $p = 1.5 \times 10^{-4}$ ). The T2D association was largely unattenuated by adjustment for BMI (OR  $5.07$  [ $2.78$ – $9.27$ ]  $p = 8.9 \times 10^{-11}$ ) and the clinical characteristics of the rare allele carriers with T2D did not provide any evidence of a phenotype distinct from typical T2D (Supplementary Data 7). Notably, *GIGYF1* loss of function was not associated with birthweight, puberty timing, childhood body size, or adult height ( $p > 0.05$ ).

We next examined whether common genetic variation in *GIGYF1* was also associated with LOY and metabolic health parameters. We observed that an intergenic variant (rs221781, MAF = 11% Supplementary Data 8 and Fig. S4)  $\sim 25 \text{ kb}$  from *GIGYF1* was significantly associated with higher glucose ( $P = 4.80 \times 10^{-15}$ ) and HbA1c ( $P = 3.40 \times 10^{-10}$ ) in UK Biobank. This same allele was associated with a higher risk of T2D<sup>19</sup> (OR adj BMI =  $1.06$  ( $1.04$ – $1.09$ ),  $p = 8.50 \times 10^{-8}$ ) and LOY ( $p = 3.00 \times 10^{-6}$ ), but with lower circulating LDL ( $p = 3.40 \times 10^{-10}$ ) and HDL ( $p = 1.90 \times 10^{-18}$ ) levels. The variant was not associated with BMI ( $p = 0.09$ ). The lead signal for T2D (rs221781) is also the lead conditionally independent eQTL for *GIGYF1* across a number of GTEx tissues including subcutaneous adipose (Fig. S5), in which we observed that higher expression of *GIGYF1* was associated with a lower risk of T2D. The lead eQTL for *GIGYF1* is rs221792 in cultured fibroblasts ( $p = 1.3 \times 10^{-32}$ ) which is in high LD ( $r^2 = 0.71$ ,  $D' = 1$ ) with rs221781. The association of common *GIGYF1* variants with T2D was also confirmed in Million Veteran Program data, in which we found

a previously reported lead SNP for T2D was in high LD with rs221781 (rs534043,  $r^2 = 1$ ,  $P = 8.03 \times 10^{-10}$ ) with a consistent direction of effect<sup>20</sup>.

## Discussion

In summary, this exome-wide approach identified rare loss of function alleles in *GIGYF1* exhibiting an effect on LOY  $\sim 5$  times larger than any genetic variants previously identified by GWAS. Similarly, these alleles confer effect sizes on a number of metabolic outcomes far larger than those previously identified by imputed GWAS and other smaller sequencing studies. For example, rare variants in *PDX1*, *CCND2*, *SLC30A8*, and *PAM* are associated with double the odds of T2D<sup>21–23</sup>, whereas *GIGYF1* loss of function is associated with a six-fold increased risk (OR =  $5.96$  [ $3.43$ – $10.38$ ]). The majority of common variants associated with T2D confer much more modest effects (OR  $< 1.5$ )<sup>19</sup>.

*GIGYF1* encodes a member of the *gyf* family of adaptor proteins. It binds growth factor receptor-bound 10 (GRB10), which is another adaptor protein that binds activated insulin receptors and insulin-like growth factor-1 (IGF-1) receptors to negatively regulate receptor signaling, metabolic responses, and IGF1-induced mitogenesis<sup>18,24,25</sup>. Transfection of cells with GRB10-binding fragments of *GIGYF1* leads to greater activation of both the insulin receptor and the IGF-1 receptor<sup>26</sup>. Our findings relating loss of function variants in *GIGYF1* to metabolic and anthropometric outcomes are broadly consistent with the notion that in individuals carrying two functional copies of this gene, *GIGYF1* enhances insulin and IGF-1 receptor signaling, leading to greater handgrip strength (relative to loss of function carriers), sitting height and circulating IGF-1 levels (due to increased insulin signaling), and lower % body fat, WHR, HbA1c, glucose levels, and susceptibility to T2D. We previously highlighted the potential

role of IGF signaling in promoting chromosomal instability and the cellular accumulation of DNA damage and reported that genetically higher IGF-1 levels are related to greater LOY<sup>27</sup>. It may therefore appear paradoxical that here we find that loss of function in *GIGYF1* (putatively leading to decreased IGF-1 signaling) should be associated with increased rather than decreased LOY. We hypothesize that *GIGYF1* might enhance DDR mechanisms to protect DNA integrity in the face of IGF-1-mediated tissue growth and differentiation. *GIGYF1* and the related protein *GIGYF2* are implicated in translational repression<sup>28</sup> and translation-coupled mRNA decay<sup>29</sup>, which suggests that they may have biological roles beyond insulin and IGF-1 receptor signaling. Although *GIGYF1* is broadly expressed<sup>30</sup>, the lack of associations in our data with some established IGF-1-related traits, such as birth weight and adult height, might reflect tissue or developmental specificity in its effects. We anticipate that future experimental work will shed light on these questions to better understand the links between clonal mosaicism and metabolic health.

## Methods

**Phenotype definitions.** Until now, there were two established mLOY estimation methods based on SNP-array data: (1) the median or mean of log R ratio (mLRR-Y) genotyping intensity values of the probes on the male-specific regions of chromosome Y (MSY); and (2) the phase-based computational method that estimates allelic imbalance using only the pseudoautosomal regions (PAR-LOY) detailed previously<sup>13</sup>. The mLRR-Y and PAR-LOY are independent approaches as they are estimated from non-overlapping regions of the Y chromosome. Although there is a considerable correlation in the LOY estimates produced by these two methods, we sought to combine the independent information considered by the two approaches to increase power for genetic association analyses. We combined PAR-LOY and mLRR-Y with an additional measure, the estimated fraction of cells with LOY (AF-LOY) which was estimated when generating PAR-LOY<sup>13</sup>. Our new combined call of LOY (PAR-LOYq) is defined as  $PAR-LOY + (3 * AF-LOY) - (3 * mLRR-Y)$  (cropped to the range [0,2]). The intuition behind this formula is to augment the binary PAR-LOY variable by up-weighting individuals who have a larger LOY cell fraction (as estimated by AF-LOY and mLRR-Y), which may be more strongly associated with risk alleles.

We compared the performance of the three LOY estimates derived from the genotyping array data using the full set of male UKBB participants<sup>13</sup>. All UK Biobank participants provided written informed consent, the study was approved by the National Research Ethics Service Committee North West—Haydock, and all study procedures were performed in accordance with the ethical principles for medical research from the World Medical Association Declaration of Helsinki. We performed association testing with age and ever smoking status, which are two established risk factors for LOY<sup>31,32</sup>, and the 156 previously reported LOY-associated loci<sup>13</sup>. For both age and smoking status, PAR-LOYq outperformed the two established mLOY estimation methods using the same sample; the *t*-test statistic of PAR-LOYq for age increased by 65.4% and 5.2%, respectively, and the *t*-test statistic of PAR-LOYq for ever smoking status increased by 44.9% and 11.1%, respectively. Improvement of PAR-LOYq over PAR-LOY was also evaluated for the 156 previously identified variants by assessing the median improvement in chi-square statistic.

Participants were classified as cases of Type 2 diabetes (T2D) according to the previously published UKBB probable T2D algorithm<sup>33</sup> based on baseline self-reported diabetes or medications, in addition to evidence from electronic health records (Hospital Episode Statistics or Death Registration) consistent with T2D (International Statistical Classification of Diseases and Related Health Problems Tenth Revision code E11). Any possible or probable Type 1 diabetes cases were excluded. Controls were participants without evidence of T2D as defined above. The GWAS on random glucose and HbA1C—using the BOLT-LMM pipeline described below—was performed after excluding individuals with our defined T2D criteria. The T2D test statistic for the common variant was taken from the DIAMANTE consortium GWAS meta-analysis<sup>19</sup>. All other phenotypes used in this study were available from UK Biobank and any applied transformations are described in the relevant results tables.

**UK Biobank exome-sequence data processing and QC.** We downloaded VCF and PLINK format files for whole-exome sequencing (WES) data of 200,643 UK Biobank participants, which were made available in October 2020. The overview of this 200 K WES release is described at <https://biobank.ndph.ox.ac.uk/ukb/label.cgi?id=170>. Details of sequence data processing (read alignment, variant calling, etc.) are described in papers of Szustakowski et al. [<https://doi.org/10.1101/2020.11.02.2022232>] and Yun et al. [<https://doi.org/10.1101/2020.02.10.942086>]

We merged individual VCF files into a single VCF file of each chromosome using BCFtools v1.9<sup>34</sup>. We converted each chromosome file losslessly to a GDS (Genomic Data Structure) format file (an RData object) using the seqVCF2GDS() function from the R package SeqArray v1.30.0<sup>35</sup>. We used SeqArray package and GDS data object to extract the dosage matrix and perform additional variant and genotype level filtering below. Such genotype data processing is faster than using a flat text VCF file because GDS is implemented using an optimized C++ library and a high-level R interface is provided by the platform-independent R package gdsfmt<sup>35,36</sup>.

We used SeqArray package to calculate and extract the QC metrics. Firstly, we identified and flagged 7,913,671 on-target variants (those defined by the *xgen\_plus\_spikein.GRCh38.bed* file genomic coordinates) among the total of 15,916,398 called variants on autosomes and chromosome X. The UKBB released VCF file has a number of QC metrics that can be used for variant site and individual genotype filtering: QUAL (variant site-level quality score, Phred scale); AQ (variant site-level allele quality score reflecting evidence for each alternate allele, Phred scale); DP (individual genotype call-level approximate read depth (reads with MQ = 255 or with bad mates had already been filtered out)); AD (individual genotype call-level allelic depths for the ref and alt alleles in the order listed); GQ (individual genotype call-level Genotype Quality, Phred scale). We additionally calculated the site-level genotype missingness (the number of samples at each site without genotype call).

After generating the summary statistics of QUAL and AQ metrics, we noted that the released UKBB 200 K WES data already had some QC filters applied. The values of QUAL and AQ ranged from 20 (error rate = 1%) to 99 (error rate <0.0001%) with mean 44.5 (error rate <0.01%). For all chromosomes, the distributions of the values of QUAL and AQ are nearly the same. We decided not to apply additional stricter filters on these two site-level metrics. We calculated summary statistics (minimum, maximum, mean, and 1st, 2nd, and 3rd quartile) for DP and GQ for each variant based on all 200,643 samples for autosomes and 110,438 female samples for the X chromosome. We recorded the number of samples with GQ < 20 at each variant. We calculated allelic balance for each heterozygous genotype calls at on-target bi-allelic sites (ABratio), defined as the number of alternate allele's reads (provided in the AD field) divided by the total depth which equals the sum of reading depths of reference allele and alternative allele. We then generated the same per-site summary statistics as above for ABratio. We defined and excluded a heterozygous genotype call as imbalanced if  $ABratio \leq 0.25$  or  $ABratio \geq 0.8$ .

In our sensitivity analysis, we applied three variant-level filters to exclude variants at potentially poorly performing sites: filter 1: >5% missingness (samples without genotype calls); filter 2: the maximum of the read depth of genotype calls (DP) across samples <10; and filter 3: >20% genotype calls with GQ < 20. After applying these three filters, 1,161,679 (7.3%) of the initial 15,916,398 variants, and 96,640 (1.2%) of the 7,913,671 on-target variants were excluded. For the variants included in our variant-set analysis, we also generated the same QC metrics restricted only to rare allele carriers. Ultimately all of these metrics were used to filter out variants in sensitivity analyses that were initially performed using the default QC parameters applied to the UKBB released dataset.

**Variant annotation and definition of gene burden sets.** We annotated variants released in UK Biobank (UKBB) 200 K whole-exome sequencing (WES) VCF files using the Ensembl Variant Effect Predictor tool release 99 based on build hg38<sup>37</sup>. For each uploaded variant, the default VEP features include consequence and impact of the variant, overlapping gene, position at cDNA and protein level, and amino acid change, if applicable. In addition to the default features, the following plugins from VEP were used: (i) SIFT<sup>38</sup>, which predicts whether an amino acid substitution affects protein function based on sequence homology and the physical properties of the amino acid, (ii) Polyphen-2<sup>39</sup>, which predicts possible impact of an amino acid substitution on the structure and function of a protein, (iii) CADD<sup>40</sup> which provides deleteriousness prediction scores for all variants based on diverse genomic features, and (iv) LOFTEE<sup>16</sup> which provides loss of function prediction for variants. The variants were annotated for every available overlapping transcript in Ensembl. We used the most severe variant definition for each variant-gene pair, which provides the annotation of the variant for the transcript it has the most severe consequence on.

We defined loss of function variants as those with 'high impact' prediction by VEP. This includes: frameshift variants, transcript ablating or transcript amplifying variants, splice acceptor or splice donor variants, stop lost, start gained, or stop gained variants. 'Moderate impact' variants include missense variants, inframe deletion or insertions, missense variants, and protein-altering variants.

**Gene association testing.** Gene burden scores were created by collapsing all annotated rare alleles together to define a binary call denoting whether an individual carries none vs. one or more rare alleles at a given gene. Reported effect estimates therefore represent the trait difference between carriers and non-carriers of these alleles. These dummy variables were then transformed into BGEN file format genotype call files for association testing using a linear mixed model implemented in BOLT-LMM<sup>41</sup> to account for cryptic population structure and relatedness. Only autosomal genetic variants that were common (minor allele frequency (MAF) >1%), passed quality control in all 106 batches, and were present

on both genotyping arrays were included in the genetic relationship matrix (GRM). Genotyping chip, age at baseline, and ten genetically derived principal components were included as covariates. Samples were excluded from analysis if they failed UK Biobank quality control parameters, were of non-European ancestry or if the participant withdrew consent from the study.

**Secondary association testing.** We applied STAAR (variant-Set Test for Association using Annotation information)<sup>17</sup> as a secondary analytical approach for associated genes. STAAR is a general framework for performing a rare variants association study at scale, suitable for whole exome or genome population-level datasets such as UKBB. STAAR accounts for population structure and relatedness, by fitting linear and logistic mixed models for quantitative and dichotomous traits. It takes as input individual data frames for genotypes, phenotypes, covariates including age, age<sup>2</sup>, sex, chip, PC1-PC10 were generated from the SNP array data and (sparse) GRM.

We used the basic function of STAAR (with CADD-score weighting additionally performed in a sensitivity analysis) and set the thresholds of MAF  $\leq 0.5\%$  and  $\geq 2$  rare variants count in a gene. The output of STAAR provides *p*-values for a number of different rare variant set burden tests including SKAT (sequence kernel association test), Burden test, and ACAT-V (set-based aggregated Cauchy association test). In addition, STAAR provides an omnibus test result by using the combined Cauchy association test to aggregate the association across the different tests.

To ensure that the individual gene-level result is not disproportionately affected by a single variant of considerably larger effect and that the others are part of the same variant set, we performed a drop-one-out analysis using STAAR for our target gene.

Effect estimates for dichotomous traits were estimated by using logistic regression performed in R (3.3.3). Where these are reported they include the *p*-value obtained from the linear mixed-model generated by BOLT-LMM.

**Reporting summary.** Further information on research design is available in the Nature Research Reporting Summary linked to this article.

## Data availability

All individual-level data used in this study are available from the UK Biobank study upon application ([www.ukbiobank.ac.uk](http://www.ukbiobank.ac.uk)). Full exome-wide summary statistics are reported in the supplement. The exome sequence data resource is described here: <https://biobank.ndph.ox.ac.uk/ukb/label.cgi?id=170> and mosaic LOY calls here: <https://biobank.ndph.ox.ac.uk/ukb/dset.cgi?id=3094>. European GWAS meta-analysis summary results from the Million Veteran Program and other biobanks are available via dbGaP (NCBI dbGaP analysis accession pha004945, available to download from <https://ftp.ncbi.nlm.nih.gov/dbgap/studies/phs001672/analyses/>). DIAMANTE consortium GWAS meta-analysis results are available to download from the DIAGRAM consortium website (<https://www.diagram-consortium.org/downloads.html>).

Received: 24 February 2021; Accepted: 21 June 2021;

Published online: 07 July 2021

## References

- Jacobs, P. A., Brunton, M., Court Brown, W. M., Doll, R. & Goldstein, H. Change of human chromosome count distribution with age: evidence for a sex differences. *Nature* **197**, 1080–1081 (1963).
- Jacobs, P. A., Court Brown, W. M. & Doll, R. Distribution of human chromosome counts in relation to age. *Nature* **191**, 1178–1180 (1961).
- Noveski, P. et al. Loss of Y chromosome in peripheral blood of colorectal and prostate cancer patients. *PLoS ONE* **11**, e0146264 (2016).
- Ganster, C. et al. New data shed light on Y-loss-related pathogenesis in myelodysplastic syndromes. *Genes Chromosomes Cancer* **54**, 717–724 (2015).
- Persani, L. et al. Increased loss of the Y chromosome in peripheral blood cells in male patients with autoimmune thyroiditis. *J. Autoimmun.* **38**, J193–6 (2012).
- Haitjema, S. et al. Loss of Y chromosome in blood is associated with major cardiovascular events during follow-up in men after carotid endarterectomy. *Circ. Cardiovasc. Genet.* **10**, e001544 (2017).
- Dumanski, J. P. et al. Mosaic loss of chromosome Y in blood is associated with Alzheimer disease. *Am. J. Hum. Genet.* **98**, 1208–1219 (2016).
- Grassmann, F. et al. Y chromosome mosaicism is associated with age-related macular degeneration. *Eur. J. Hum. Genet.* <https://doi.org/10.1038/s41431-018-0238-8>. (2018).
- Lofffield, E. et al. Predictors of mosaic chromosome Y loss and associations with mortality in the UK Biobank. *Sci. Rep.* **8**, 12316 (2018).
- Forsberg, L. A. et al. Mosaic loss of chromosome Y in peripheral blood is associated with shorter survival and higher risk of cancer. *Nat. Genet.* **46**, 624–628 (2014).
- Zhou, W. et al. Mosaic loss of chromosome Y is associated with common variation near TCLA. *Nat. Genet.* **48**, 563–568 (2016).
- Wright, D. J. et al. Genetic variants associated with mosaic Y chromosome loss highlight cell cycle genes and overlap with cancer susceptibility. *Nat. Genet.* **49**, 674–679 (2017).
- Thompson, D. J. et al. Genetic predisposition to mosaic Y chromosome loss in blood. *Nature* **575**, 652–657 (2019).
- Terao, C. et al. GWAS of mosaic loss of chromosome Y highlights genetic effects on blood cell differentiation. *Nat. Commun.* **10**, 4719 (2019).
- He, L. M. et al. Cyclin D2 protein stability is regulated in pancreatic beta-cells. *Mol. Endocrinol.* **23**, 1865–1875 (2009).
- Karczewski, K. J. et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).
- Li, X. et al. Dynamic incorporation of multiple in silico functional annotations empowers rare variant association analysis of large whole-genome sequencing studies at scale. *Nat. Genet.* **52**, 969–983 (2020).
- Giovannone, B. et al. Two novel proteins that are linked to insulin-like growth factor (IGF-I) receptors by the Grb10 adapter and modulate IGF-I signaling. *J. Biol. Chem.* **278**, 31564–31573 (2003).
- Mahajan, A. et al. Fine-mapping type 2 diabetes loci to single-variant resolution using high-density imputation and islet-specific epigenome maps. *Nat. Genet.* **50**, 1505–1513 (2018).
- Vujkovic, M. et al. Discovery of 318 new risk loci for type 2 diabetes and related vascular outcomes among 1.4 million participants in a multi-ancestry meta-analysis. *Nat. Genet.* **52**, 680–691 (2020).
- Steinthorsdottir, V. et al. Identification of low-frequency and rare sequence variants associated with elevated or reduced risk of type 2 diabetes. *Nat. Genet.* **46**, 294–298 (2014).
- Huyghe, J. R. et al. Exome array analysis identifies new loci and low-frequency variants influencing insulin processing and secretion. *Nat. Genet.* **45**, 197–201 (2013).
- Flannick, J. et al. Exome sequencing of 20,791 cases of type 2 diabetes and 24,440 controls. *Nature* **570**, 71–76 (2019).
- Dufresne, A. M. & Smith, R. J. The adapter protein GRB10 is an endogenous negative regulator of insulin-like growth factor signaling. *Endocrinology* **146**, 4399–4409 (2005).
- Holt, L. J. & Siddle, K. Grb10 and Grb14: enigmatic regulators of insulin action—and more? *Biochem. J.* **388**, 393–406 (2005).
- Preston, E., Butler, K. & Haas, N. Does magnetic resonance imaging compromise integrity of the blood-brain barrier? *Neurosci. Lett.* **101**, 46–50 (1989).
- Stankovic, S. et al. Elucidating the genetic architecture underlying IGF1 levels and its impact on genomic instability and cancer risk. *Wellcome Open Res.* **6**, 20 (2021).
- Peter, D. et al. GIGYF1/2 proteins use auxiliary sequences to selectively bind to 4EHP and repress target mRNA expression. *Genes Dev.* **31**, 1147–1161 (2017).
- Weber, R. et al. 4EHP and GIGYF1/2 mediate translation-coupled messenger RNA decay. *Cell Rep.* **33**, 108262 (2020).
- GTEX Consortium. et al. Genetic effects on gene expression across human tissues. *Nature* **550**, 204–213 (2017).
- Dumanski, J. P. et al. Mutagenesis. Smoking is associated with mosaic loss of chromosome Y. *Science* **347**, 81–83 (2015).
- Forsberg, L. A., Gisselsson, D. & Dumanski, J. P. Mosaicism in health and disease—clones picking up speed. *Nat. Rev. Genet.* **18**, 128–142 (2017).
- Eastwood, S. V. et al. Algorithms for the capture and adjudication of prevalent and incident diabetes in UK biobank. *PLoS ONE* **11**, e0162388 (2016).
- Li, H. et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
- Zheng, X. et al. SeqArray—a storage-efficient high-performance data format for WGS variant calls. *Bioinformatics* **33**, 2251–2257 (2017).
- Zheng, X. et al. A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics* **28**, 3326–3328 (2012).
- McLaren, W. et al. The Ensembl variant effect predictor. *Genome Biol.* **17**, 122 (2016).
- Sim, N.-L. et al. SIFT web server: predicting effects of amino acid substitutions on proteins. *Nucleic Acids Res.* **40**, W452–W457 (2012).
- Adzhubei, I., Jordan, D. M. & Sunyaev, S. R. Predicting functional effect of human missense mutations using PolyPhen-2. *Curr. Protoc. Hum. Genet.* **76**, 7–20 (2013).
- Rentzsch, P., Witten, D., Cooper, G. M., Shendure, J. & Kircher, M. CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res.* **47**, D886–D894 (2019).
- Loh, P.-R. et al. Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nat. Genet.* **47**, 284–290 (2015).

## Acknowledgements

All Cambridge (UK) authors are supported by the Medical Research Council (Unit programs: MC\_UU\_12015/2, MC\_UU\_00006/2, MC\_UU\_12015/1, and MC\_UU\_00006/1). N.J.W. is an NIHR Senior Investigator. P.-R.L. is supported by NIH grant DP2 ES030554 and a Burroughs Wellcome Fund Career Award at the Scientific Interfaces. This research was conducted using the UK Biobank Resource under application 9905.

## Author contributions

All authors contributed to manuscript writing and approved the final version. Statistical analysis: Y.Z., S.S., M.K., E.W., F.R.D., H.L.A., N.D.K., M.P., P.R.L., J.R.B.P. Study oversight: N.J.W., C.L., K.K.O., J.R.B.P.

## Competing interests

The authors declare no competing interests.

## Additional information

**Supplementary information** The online version contains supplementary material available at <https://doi.org/10.1038/s41467-021-24504-y>.

**Correspondence** and requests for materials should be addressed to J.R.B.P.

**Peer review information** *Nature Communications* thanks Meredith Yeager and the other, anonymous, reviewers for their contribution to the peer review of this work.

**Reprints and permission information** is available at <http://www.nature.com/reprints>

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2021