Genetic variants associated with mosaic Y chromosome loss highlight cell cycle genes and overlap with cancer susceptibility

Daniel J Wright1,9, Felix R Day1,9, Nicola D Kerrison1,9, Florian Zink2,9, Alexia Cardona1, Patrick Sulem2, Deborah J Thompson3, Svanhvit Sigurjonsdottir2, Daniel F Gudbjartsson2, Agnar Helgason2, J Ross Chapman4, Steve P Jackson5,6, Claudia Langenberg1, Nicholas J Wareham1, Robert A Scott1, Unnur Thorsteindottir2,7, Ken K Ong1,8, Kari Stefansson2,7,9 & John R B Perry1,9

The Y chromosome is frequently lost in hematopoietic cells, which represents the most common somatic alteration in men. However, the mechanisms that regulate mosaic loss of chromosome Y (mLOY), and its clinical relevance, are unknown. We used genotype-array-intensity data and sequence reads from 85,542 men to identify 19 genomic regions ($P < 5 \times 10^{-8}$) that are associated with mLOY. Cumulatively, these loci also predicted X chromosome loss in women ($n = 96,123$; $P = 4 \times 10^{-6}$). Additional epigenome-wide methylation analyses using whole blood highlighted 36 differentially methylated sites associated with mLOY. The genes identified converge on aspects of cell proliferation and cell cycle regulation, including DNA synthesis (NPAT), DNA damage response (ATM), mitosis (PMF1, CENPN and MAD1L1) and apoptosis (TP53). We highlight the shared genetic architecture between mLOY and cancer susceptibility, in addition to inferring a causal effect of smoking on mLOY. Collectively, our results demonstrate that genotype-array-intensity data enables a measure of cell cycle efficiency at population scale and identifies genes implicated in aneuploidy, genome instability and cancer susceptibility.

For over a century, errors in cell division that result in too few or too many chromosomes in daughter cells, a cytogenetic feature termed aneuploidy, have been described. Although a well-established feature of human cancer cells, it remains unclear whether acquired aneuploidy is a cause or consequence of tumorigenesis. Research into the molecular mechanisms of aneuploidy has focused largely on the role of mitosis and mitotic checkpoint signaling, primarily in cellular and animal models1,2. Recent human genomic studies have shown that aneuploidy can be estimated using intensity data from standard genotyping arrays, an approach that has been validated by DNA sequencing3-5. These population-based studies have demonstrated that mLOY is more frequent than other mosaic chromosomal and structural alterations; indeed, about one in five men over 80 years of age has detectable mLOY. Such epidemiological studies have also identified associations with non-hematological cancers4,5 and Alzheimer’s disease7; however, these observations are inconsistent5 and possibly subject to confounding or reverse-causality.

The ability to assay a common measure of aneuploidy in large array-genotyped populations could facilitate the systematic identification of variants and genes involved in cell division errors. This would, in turn, enable a better understanding of the mechanisms involved and of the potential causal consequences of aneuploidy on cancer risk, which may subsequently be inferred using Mendelian randomization approaches. To date, a single genetic association with mLOY near TCL1A has been reported ($n = 12,369$), suggesting that germline variation influencing mosaic chromosome loss can be detected3. Here we used data from up to 85,542 men to highlight widespread genomic, transcriptomic and epigenetic signatures of mosaic Y chromosome loss, and we demonstrate that this approach successfully identifies genes implicated in cell cycle regulation, genome instability and cancer susceptibility.

RESULTS

As a proxy for mLOY, we estimated the mean intensity log-R ratio of all array-genotyped Y chromosome single-nucleotide polymorphisms

---

1MRC Epidemiology Unit, School of Clinical Medicine, University of Cambridge, Cambridge, UK. 2deCODE Genetics–Amgen, Inc., Reykjavik, Iceland. 3Center for Cancer Genetic Epidemiology, Department of Public Health and Primary Care, University of Cambridge, Cambridge, UK. 4Wellcome Trust Center for Human Genetics, University of Oxford, Oxford, UK. 5Wellcome Trust and Cancer Research UK Gurdon Institute, University of Cambridge, Cambridge, UK. 6Department of Biochemistry, University of Cambridge, Cambridge, UK. 7Faculty of Medicine, University of Iceland, Reykjavik, Iceland. 8Department of Pediatrics, University of Cambridge, Cambridge, UK. 9These authors contributed equally to this work. Correspondence should be addressed to J.R.B.P. (john.perry@mrc-epid.cam.ac.uk).

Received 30 September 2016; accepted 26 February 2017; published online 27 March 2017; doi:10.1038/ng.3821
Many autosomal genetic variants are associated with mLOY

To identify the genetic variants associated with mLOY, we performed a genome-wide association study (GWAS) of mLRR-Y as a quantitative trait in the UKB cohort. After stringent quality control (Online Methods), the most significantly associated SNPs were located at the previously reported mLOY locus on chromosome 14, TCLA (P = 3.6 × 10−23) (Supplementary Fig. 2). In addition, we identified an additional 18 previously unknown signals at genome-wide significance (P < 5 × 10−8), with no evidence for significant inflation of test statistics genome wide (λ = 1.05) (Supplementary Figs. 2 and 3). We subsequently performed replication in an independent set of 9,793 men with array intensity data, and in a further 8,715 men from Iceland (deCODE cohort) with Y chromosome loss estimated using sequence reads (Online Methods). Both replication data sets provided strong statistical support for the identified loci, with all 19 loci retaining genome-wide significance in a combined model (Table 1). As evaluated in the deCODE cohort data, these loci cumulatively explained 2.7% of the total variance in Y chromosome copy number. We estimated an overall heritability of 34% (25.2–42.4%), suggesting that many additional associated variants remain to be discovered.

We next used HaploReg10 and sequence data from the deCODE study to functionally annotate the variants and genes identified. This analysis highlighted four signals containing highly correlated missense variants, implicating MAD1L1 (rs1801368, r² = 0.98), PMF1 (rs1052053, r² = 1), NREP (rs11559, r² = 0.74) and NPAT (rs2070661, r² = 0.97) as potential candidates.

To ascertain whether the identified signals were more likely to reflect gain or loss of Y chromosome material, we performed two analyses comparing the bottom and top 5% of mLRR-Y-ranked individuals to the median 25%, as a dichotomous indicator of extreme Y chromosome loss or gain. All 19 loci showed consistently stronger associations with the bottom 5% of mLRR-Y (greatest mLOY) than with the top 5% (Supplementary Table 1), suggesting that their effect was on mosaic Y chromosome loss rather than its gain. Analysis of mLRR-Y as a continuous trait across all individuals was, however, the most powerful approach for variant discovery, as only two of the signals reached genome-wide significance in the stratified analysis.
Articles

Genome-wide pathway analyses conducted on association results for continuous mLRR-Y highlighted five pre-defined biological pathways that were enriched for association (study-wise significant false discovery rate (FDR) < 0.05), the most significant of which was ‘apoptosis’ as per the Kyoto Encyclopedia of Genes and Genomes (KEGG)\(^\text{11}\) (Supplementary Table 2). Other pathways included sulfur metabolism, susceptibility to colorectal, prostate and thyroid cancers, and progesterone-mediated oocyte maturation.

The effect of mLOY variants on X-chromosome loss in women

We next sought to understand whether the variants we identified acted only on the Y chromosome or whether they promoted aneuploidy of other chromosomes more generally. Using a combined sample of 96,123 women from three studies, we ascertained X chromosome loss via both array-intensity data (\(n = 86,843\)) and sequence reads (\(n = 9,280\)) (Fig. 1). Chromosome X copy number was estimated to have a heritability of 26% (17.4–36.2%) in the deCODE data, which was comparable to that of Y chromosome loss. Cumulatively, the 19 mLOY SNPs significantly (\(P = 4 \times 10^{-6}\)) predicted X chromosome loss in women, with the expected direction of effect (Fig. 2).

Identification of transcriptomic and epigenetic signatures of mLOY

To identify potential functional transcripts that mediate Y chromosome loss, we applied summary statistic approaches to infer gene expression associations using three analytical imputation methods\(^\text{12–14}\) in independent whole-blood expression data sets (Supplementary Tables 3–5). Across these data sets, eight genes (HM13, SMPD2, TCL1A, SENP7, NPAT, ATM, ACAT1 and CENPN) were significantly associated with mLRR-Y (\(\alpha \leq 0.05\)) after accounting for multiple testing (Online Methods), all of which mapped near to one of the 19 associated genetic signals from the GWAS.

We additionally identified 36 methylation variable positions (MVPs) that correlated with mLRR-Y levels in 569 whole-blood samples from the EPIC-Norfolk cohort\(^\text{15}\) (Supplementary Table 6). All significant (\(P \leq 1 \times 10^{-7}\)) MVPs were in genomic regions that were distinct (>500 kb) from the 19 mLOY loci, with the exception of four correlated methylation probes within the TP53 gene region. To determine whether any of the methylation changes represented causal drivers of mLOY, we next identified cis-methylation quantitative trait loci (meQTLs) in publicly available data\(^\text{16}\) for all of the associated probes. In total, 20 probes had one or more genetic variants in cis that were associated with methylation levels of the corresponding site (Supplementary Table 7). None of these genetic variants were correlated with the 19 genomic loci; however, one cis-meQTL survived multiple-test correction for association with mLRR-Y (rs7208523, cg20116579 methylation \(P = 5.6 \times 10^{-31}\); mLRR-Y \(P = 9 \times 10^{-4}\)).

![Figure 1](https://example.com/figure1.png)

**Figure 1** Estimated X chromosome and Y chromosome loss with age in members of the Icelandic deCODE study. (a) Y chromosome copy number estimated in 8,703 males from whole-genome sequencing. (b) X chromosome copy number for 9,280 females. In each case, the black line indicates the line of best fit with age at the time of blood collection as a linear predictor.

![Figure 2](https://example.com/figure2.png)

**Figure 2** Association of the 19-SNP mLOY genetic risk score with X chromosome loss in women. The genetic risk score is additive, based on mLRR-Y increasing allele dosage. CI, confidence interval.
This suggests that genetic variation at the TNK1 locus, a gene with known involvement in tumor growth and survival, may be associated with increased mLOY via an epigenetic mechanism.

**Genetic overlap with cancer susceptibility**

Three mLOY signals are correlated with signals previously reported for basal cell carcinoma, glioma, neuroblastoma or testicular cancer (SEMA4A–PMF1 and MAD1L1). In each case, the mLRR-Y-decreasing allele (i.e., increased mLOY) was associated with increased cancer susceptibility. We performed a reciprocal lookup of 90 loci that were previously reported for prostate cancer susceptibility, the most common male non-skin cancer in western populations. There was no obvious enrichment of signal across these loci and no apparent dose-response relationship between the allelic effects on prostate cancer and mLOY. We used a proxy trait for the study of aneuploidy in large-scale populations, loss of the Y chromosome in peripheral blood likely represents a frequent large-scale mosaic event.

Under the hypothesis that susceptibility to many types of cancer may have a common basis in mitotic error, we performed a GWAS in the UKB cohort, in which we defined men with any diagnosed cancer as a case. We performed a GWAS in the UKB cohort, in which we defined men with many types of cancer may have a common basis in mitotic error, particularly cancer, prostate cancer and mLOY. We applied the mLRR-Y decreasing allele (i.e., increased mLOY) was associated with increased cancer susceptibility. We performed a reciprocal lookup of 90 loci that were previously reported for prostate cancer susceptibility, the most common male non-skin cancer in western populations. There was no obvious enrichment of signal across these loci and no apparent dose-response relationship between the allelic effects on prostate cancer and mLOY. We performed a GWAS in the UKB cohort, in which we defined men with any diagnosed cancer as a case.

Pathway analyses identified enrichment for cancer and apoptosis pathways that were associated with mLOY. This is further supported by the many genes we observed that encode well-established cell cycle regulation factors, either as the closest gene to the association signal or implicated via altered expression or protein-coding changes. Major mechanistic aspects of the cell cycle and key regulators of cell cycle...
progression were represented by these findings (Fig. 3), including elements of three cell cycle checkpoints and several genes with complementary functional roles in mitosis. TPX2, CENPN, PMF1 and ATMIN are involved in aspects of chromosome alignment during metaphase, spindle assembly, orientation and attachment to chromatids ahead of segregation38,29. In particular, TPX2 recruits the crucial mitotic enzyme, Aurora kinase A, to the spindle30, whereas ATMIN regulates expression of a dynein motor component (DYNLL1) that critically mediates spindle positioning31–33 and also modulates Nek9 kinase signaling, which is required for correct spindle formation and function34. Similarly, Rho guanine-exchange factor 10 (encoded by ARHGEF10; for which we observed a nearby methylated signal) regulates centrosome duplication and prevents the formation of multipolar spindles35. We identified a missense-encoding variant in MAD1L1 (MAD1 mitotic arrest deficient like 1), a major component of the spindle assembly checkpoint. This checkpoint represents a key cellular safeguard against chromosome mis-segregation (and subsequent ploidy errors), as it suppresses metaphase–anaphase progression until the chromatids are bi-orientated on a bipolar spindle at the metaphase plate1. During cytokinesis, SEPT5 (septin 5; implicated in our methylation analysis) encodes a conserved cell cycle regulator that is required for effective cell division36, whereas activation of signaling by Rho guanine-exchange factor 10 (ARHGEF10) facilitates contractile ring ingression to separate the two daughter cells37.

We also implicated a number of genes with established roles in the replication and stability of nuclear DNA in interphase; replication errors are a key cause of genomic instability and chromosomal fragility38–40. G1-to-S-phase transition is dependent on NAP1, at least in part through its promotion of histone gene transcription41, whereas ATM, at least in part in association with ATMIN42, functions as major cell cycle checkpoint kinase that is dedicated to maintaining genome stability throughout interphase, with particular importance at the G1-S and G2-M checkpoints40. In response to double-stranded DNA breaks (DSBs), which are indicative of genomic instability, ATM promotes various responses via p53 and other factors to promote DNA repair, arrest cell cycle progression, or otherwise initiate cell cycle exit strategies, including apoptosis and senescence38–40,43. TREC1 encodes 3’ repair exonuclease 1, which digests aberrant replication intermediates and single-stranded DNA arising from genotoxic stress, to prevent chronic checkpoint activation by these factors44. Predicted deleterious missense-encoding variants in this gene were recently identified in a mouse GWAS for micronucleus formation, a biomarker of chromosomal breaks, whole-chromosome loss and extranuclear DNA45.

At the later stages of the cellular lifespan, several genes implicated by our GWAS findings—including TP53, TCL1A, SMPD2, BCL2 and BCL2L1—functionally affect apoptotic events46–50. Apoptosis is a prime mechanism by which cells with detected DNA damage or ploidy errors may be eliminated51; indeed, p53 drives multiple cell cycle exit responses in response to aberrant mitosis, including G1 arrest43,52,53. The TP53 variant associated with mLOY in our analyses is the one that was previously associated with basal cell carcinoma risk: for this trait, the risk allele changes the AATAAA polyadenylation signal to AATACA, which results in impaired 3’-end processing of the TP53 mRNA18. Our findings also implicated genes involved in spermatogenesis54,55 (HENMT1 and DAZAP1) and in cellular growth and differentiation56 (DLK1).

The genes directly involved in mitotic prophase–metaphase and in the spindle assembly checkpoint have clear roles in averting chromosomal mis-segregation and preventing these from persisting unchecked; however, how the broader set of genes we identified here may function to promote mLOY remains less clear. We speculate that many of these genes act in ways that are not currently recognized or, alternatively, that the other highlighted processes outside of cell cycle control and mitosis are important. In particular, as a major mode of cell cycle exit, our observed enrichment of genes involved in apoptosis and its regulation may have a more passive permissive role in enabling mis-segregated cells to survive with ploidy errors rather than being directly causative of them.

Although an initial defect during the cell cycle process is required to generate an aneuploid daughter cell, clonal expansion is likely required to drive the lineage to a detectable frequency in the circulating white blood cell population. It is possible that mLOY in hematopoietic precursors confers a proliferative advantage to such cells, leading to a relative enrichment of assayable mLOY progeny. We therefore speculate that some loci may operate through this pathway to further facilitate or promote clonal expansion of these cells. Additional functional experimentation in cellular and animal systems is ultimately required to fully elucidate this issue and the role individual associated genes may have in promoting mLOY. We also acknowledge that there are likely other, currently unknown, mechanisms by which our associated loci exert their effects.

We observed a substantial shared genetic architecture between mLOY and cancer susceptibility, suggesting that bivariate analyses of these two traits may help to prioritize new cancer susceptibility loci and elucidate their functions. We could not, however, find evidence of a dose-response relationship between these two traits. This is perhaps not unexpected given that findings from mouse studies in which mitotic checkpoint components were experimentally downregulated demonstrate an inconsistent relationship between aneuploidy and spontaneous tumorigenesis1. It is possible, therefore, that some of the genes we identified may promote benign aneuploidy, whereas others may have a more general role in genome instability. This makes the use of genetic variants associated with mLOY difficult within a Mendelian randomization framework, as genes with general roles in instability may have different phenotypic consequences than genes that promote aneuploidy in a more stable way. This, of course, does not preclude the identification of causal risk factors for mLOY, which were exemplified by our positive causal inference for smoking on mLOY, using a genetic instrument for cigarettes per day. More generally, the association between smoking and mLOY suggests that care should be taken to avoid confounding influences, such as socioeconomic patterning, in epidemiological observations between mLOY and disease. In addition to fully evaluating the broader disease relevance of mLOY, future epidemiological studies should look to assess the differential rates at which mLOY changes in individuals over time, its relevance in other tissue types and further nongenetic modifiable factors that may influence it.

In conclusion, our study highlights that estimation of mLOY, using genotype-array–intensity data may serve as a useful quantitative measure of cell cycle efficiency and genome stability, and it may thereby add a new approach to the study of cellular aging and its associations with disease, particularly cancer.

METHODS

Methods, including statements of data availability and any associated accession codes and references, are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We thank members of the MRC Epidemiology Unit genetics group for useful Friday morning discussions. This research was conducted using the UK Biobank Resource
and tumorigenesis.

35. Aoki, T., Ueda, S., Kataoka, T. & Sluder, G. Prolonged pro-metaphase blocks daughter cell proliferation via an essential role in dynein light-chain expression. *Nature*, 563–568 (2016).

36. Beites, C.L., Xie, H., Bowser, R. & Trimble, W.S. The septin CDCrel-1 binds syntaxin and affects nocodazole-induced cell death. *Mol. Cell. Biol.*, 434–439 (1999).

37. Zuo, Y., Oh, W. & Frost, J.A. Controlling the switches: Rho GTPase regulation during animal cell mitosis. *Cell Biol.*, 2998–3006 (2014).

38. Mazouzi, A., Velimir, M. & Zuo, L. DNA replication stress: causes, resolution and disease. *Exp. Cell Res.*, 509–516 (2002).

39. Gao, G. et al. Ataxia telangiectasia mutated (ATM) expression is regulated by EZF and is essential for cell cycle progression. *Cell Biol.*, 2821–2833 (2003).

40. Osborn, A.J., Elledge, S.J. & Zou, L. The fork: the DNA-replication stress-response pathway. *Trends Cell Biol.*, 509–516 (2002).

41. Schmidt, L. et al. ATMIN is required for the ATM-mediated signaling and recruitment of 53BP1 to DNA damage sites upon replication stress. *DNA Repair (Amst).* 24, 122–130 (2014).

42. Santaguida, S. & Amon, A. Short- and long-term effects of chromosome mis-segregation and aneuploidy. *Nat. Rev. Mol. Cell Biol.*, 473–485 (2015).

43. Christmann, M. & Kaina, B. Transcriptional regulation of human DNA repair genes following genotoxic stress: trigger mechanisms, inducible responses and genotoxic adaptation. *Nucleic Acids Res.*, 41, 8403–8420 (2013).

44. McNicholas, R.E. A genome-wide association study for regulators of micronucleus formation in mice. *G3 (Bethesda)*, 5, 2334–2354 (2016).

45. Beigang, R. et al. Human chromosome X mosaicism is age related and associated with several human disease types. *Nat. Genet.*, 47, 145–146 (2015).

46. Haimovitz-Friedman, A., Kolesnick, R.N. & Fuks, Z. Ceramide signaling in apoptosis. *Br. Med. Bull.*, 53, 539–553 (1997).

47. Zhivotovsky, B. & Kroemer, G. Apoptosis and genomic instability. *Nat. Rev. Mol. Cell Biol.*, 752–762 (2006).

48. Uetake, Y. & Sluder, G. Prolonged pro-metaphase blocks daughter cell proliferation despite normal completion of mitosis. *Curr. Biol.*, 20, 1666–1671 (2010).

49. Ganem, N.J. et al. Cytokinesis failure triggers hipo tumor suppressor pathway activation. *Cell*, 158, 833–844 (2014).

50. Lim, S.L. et al. HNFMT and piRNA stability are required for adult male germ-cell transvection repression and to define the spermatogenic program in the mouse. *PLoS Genet.*, 11, e1005620 (2015).

51. Hsu, L.C.-L. et al. DAZAP1, an hnrNP protein, is required for normal growth and spermatogenesis in mice. *RNA*, 14, 1814–1822 (2008).

52. Falci, F.A., Aronson, D.C., Viger, L.S. & Galán, J.E. Possible roles of DLK1 in the Notch pathway during development and disease. *Biochim. Biophys. Acta*, 1822, 988–995 (2012).
ONLINE METHODS
Estimating Y chromosome mosaicism in the UK Biobank cohort. We analyzed data from the May 2015 release of imputed genetic data from UK Biobank, which contains ~73 million SNPs, short insertion and deletions (indels) and large structural variants from 152,249 individuals. Full details have been published elsewhere. Briefly, the samples were genotyped on two slightly different arrays: approximately 50,000 were analyzed on the custom UK BiLEVE study array, and the remainder (~100,000) were analyzed on the UK Biobank Axiom array (Affymetrix), which was specifically designed to optimize imputation performance in GWAS. Removal of SNPs that had missing data, were multi-allelic or had a minor allele frequency (MAF) <1%, as well as removal of 1,037 sample outliers, resulted in a data set of 641,018 autosomal SNPs in 152,256 samples for phasing and imputation. Imputation was performed using a reference panel that was created by merging the UK10K haplotype panel with the 100 Genomes phase 3 reference panel.

In addition to the quality-control (QC) metrics performed centrally by UK Biobank, we defined a subset of samples of ‘white European’ ancestry using a K-means clustering approach that was applied to the first four principal components calculated from genome-wide SNP genotypes. All of the individuals defined in this group also self-identified as being of white ancestry in questionnaires.

mLOY was estimated by calculating the mean log-R ratio (normalized signal intensity) of SNPs on the male-specific region of the Y chromosome. Signal intensity, genotype call and confidence files from the Affymetrix Power Tools software were analyzed using the PennCNV-Affy pipeline to produce a log-R ratio file for each SNP. SNPs without LRR calculable on both arrays, or those flagged by UKB as failing QC, were excluded. Fluorescence signal intensity of the whole Y chromosome was summarized by calculation of the mean LRR across all Y chromosome SNPs (mLRR-Y). After omission of monomorphic SNPs and of genotyping and QC failures, 253 SNPs were available across all participants for the derivation of mLRR-Y.

Association testing and signal selection. Autosomal SNPs were analyzed by linear mixed models implemented in BOLT–LMM to account for cryptic population structure and relatedness within this group in our genetic association tests. The regression model included age and genotyping array as covariates. SNPs with an imputation quality <0.4 or with a MAF <0.1% were excluded post analysis. After application of QC criteria, a maximum of 67,034 men were available for analysis with genotype and phenotype data. Samples were subdivided by ‘never smoking’ (n = 32,539) versus ‘ever smoking’ (n = 34,329) for the Mendelian randomization analysis using the CHRNA5–CHRNA3–CHRNB4 rs1051730 locus. Genomic loci were defined on the basis of physical proximity using a 1-Mb window. The chromosome Y copy number had a heavy left tail with copy number >1.25) were excluded. We observed a Spearman correlation of −0.28 between the chromosome X copy number and age at the time of blood collection.

Cancer genome-wide association study. To understand the genomic relationship between cancer and mLOY, we defined an ‘any prevalent cancer’ variable in the UKB cohort using linked UK cancer registrations. Individuals with a reported age of diagnosis in the cancer registry were censored as a case. Individuals with inconsistent cancer diagnosis (i.e., a reported cancer but not age at diagnosis) were set to ‘missing’, and controls were defined as any individual with no self-reported or registry-defined cancer. GWAS analysis was performed as described above, including age, sex and genotyping array as covariates. Genetic correlations (g) were calculated between mLRR-Y and cancer using LD score regression. To assess the possible causal links between cancer and mLOY, we applied Mendelian randomization methods, which have been described extensively elsewhere.

Gene expression. To identify specific eQTL-linked genes, we used three complementary approaches, summary Mendelian randomization (SMR), transcriptome-wide association study (TWAS) and MetaXcan, enabling systematic integration of publicly available gene expression data with our genome-wide data set. SMR uses summary-level gene expression data to map potentially functional genes to trait-associated SNPs. We ran this approach against the publicly available whole-blood eQTL data set published by Westra et al., resulting in association statistics for 5,952 transcripts. A conservative significance threshold was set at P < 4.9 × 10−6, reflecting the number of genes tested genome-wide.

MetaXcan, a meta-analysis extension of the PrediXcan method, was used to infer the association between genetically predicted gene expression (GPGE) and mLRR-Y. PrediXcan is a gene-based data aggregation and integration method that incorporates information from gene expression data and GWAS data to translate evidence of association with a phenotype from the SNP level to the gene. Briefly, PrediXcan first imputes gene expression at an individual level using prediction models trained on measured transcriptomic data sets with
genome-wide SNP data and then regresses the imputed transcriptome levels with the phenotype of interest. MetaXcan extends its application to allow inference of the direction and magnitude of GPGE–phenotype associations with only summary GWAS statistics, which is advantageous when SNP–phenotype associations result from a meta-analysis setting, and also when individual-level data are not available. As input, we used GWAS meta-analysis summary statistics for mLRR-Y, LD matrix from the 1000 Genomes project, and, as weights, gene expression regression coefficients for SNPs from models trained with whole-blood transcriptomic data from the GTEx Project. The threshold for statistical significance was estimated using the Bonferroni correction for the number of tested genes.

Finally, we used the recently described TWAS approach to infer gene expression association using two whole-blood data sets (Young Finns Study and The Netherlands Twin Registry cohorts). The threshold for significance was set to correct for the number of studies and genes \((P < 1 \times 10^{-3})\). Each of the three approaches described in this section was compared by estimating the correlation \((r)\) of association \(z\)-scores across genes present in all three data sets. There was strong concordance between the 2,326 transcripts analyzed across the three approaches and data sets: SMR versus TWAS, \(r = 0.72\); SMR versus MetaXcan, \(r = 0.54\); TWAS versus MetaXcan, \(r = 0.55\).

**Methylation.** DNA methylation in whole blood was measured for 1,378 individuals in the EPIC-Norfolk cohort using the Illumina Human Methylation450k BeadChip platform. After setting methylation markers with detection \(P > 0.01\) to ‘missing’, methylation \(\beta\)-values were calculated for each marker. Quantile normalization of methylation \(\beta\)-values was applied separately to different marker groups based on color channel, probe type and methylated/ unmethylated intensity (M/U) subtype of probes (type-I M red, type-I U red, type-I M green, type-I U green, type-II red, type-II green). Samples with a sample call rate \(\leq 0.99\) were removed \((n = 77)\). Methylation \(\beta\)-value distributions of the X, Y and autosomal chromosome markers were analyzed separately, and a further 11 sample outliers were excluded. Within each sample, markers with a marker call rate \(\leq 0.95\) were excluded \((n = 4,423)\).

All further downstream analyses were restricted to autosomal methylation markers. Signal detection of methylation intensities can be affected by several factors, including SNPs on the probe, repetitive DNA and cross-reactive probes. We thus calculated the proportion of missing data at each CpG site \((P \geq 0.01)\) to ‘missing’, methylation \(\beta\)-values were calculated for each marker. Quantile normalization of methylation \(\beta\)-values was applied separately to different marker groups based on color channel, probe type and methylated/unmethylated intensity (M/U) subtype of probes (type-I M red, type-I U red, type-I M green, type-I U green, type-II red, type-II green). Samples with a sample call rate \(\leq 0.99\) were removed \((n = 77)\). Methylation \(\beta\)-value distributions of the X, Y and autosomal chromosome markers were analyzed separately, and a further 11 sample outliers were excluded. Within each sample, markers with a marker call rate \(\leq 0.95\) were excluded \((n = 4,423)\).

All further downstream analyses were restricted to autosomal methylation markers. Signal detection of methylation intensities can be affected by several factors, including SNPs on the probe, repetitive DNA and cross-reactive probes. We thus calculated the proportion of missing data at each CpG site \((P \geq 0.01)\) to ‘missing’, methylation \(\beta\)-values were calculated for each marker. Quantile normalization of methylation \(\beta\)-values was applied separately to different marker groups based on color channel, probe type and methylated/unmethylated intensity (M/U) subtype of probes (type-I M red, type-I U red, type-I M green, type-I U green, type-II red, type-II green). Samples with a sample call rate \(\leq 0.99\) were removed \((n = 77)\). Methylation \(\beta\)-value distributions of the X, Y and autosomal chromosome markers were analyzed separately, and a further 11 sample outliers were excluded. Within each sample, markers with a marker call rate \(\leq 0.95\) were excluded \((n = 4,423)\).

**Pathway analyses.** Meta-analysis gene set enrichment of variant associations (MAGENTA) was used to explore pathway-based associations in the full GWAS data set. MAGENTA implements a gene set enrichment analysis (GSEA)-based approach, as previously described. Briefly, each gene in the genome is mapped to a single index SNP with the lowest \(P\) value within a 110-kb upstream and 40-kb downstream window. This \(P\) value, which represents a gene score, is then corrected for confounding factors such as gene size, SNP density and LD-related properties in a regression model. Genes within the HLA region were excluded from analysis due to difficulties in accounting for gene density and LD patterns. Each mapped gene in the genome is then ranked by its adjusted gene score. At a given significance threshold (95th and 75th percentiles of all gene scores), the observed number of gene scores in a given pathway, with a ranked score above the specified threshold percentile, is calculated. This observed statistic is then compared to 1,000,000 randomly permuted pathways of identical size. This generates an empirical GSEA \(P\) value for each pathway. Study-wise significance was determined when an individual pathway reached a FDR \(< 0.05\) in either analysis. In total, 3,216 pathways from Gene Ontology, PANTHER, KEGG and Ingenuity were tested for enrichment of multiple modest associations with mLRR-Y.

**Data availability statement.** The genome-wide discovery data used are from UK Biobank cohort can be obtained via application from http://www.ukbiobank.ac.uk. Requests for access to the underlying replication data is limited by participant consent and data sharing agreements; requests should be directed via http://www.srl.cam.ac.uk/epic/ or the corresponding author. The methylation data are available from the same EPIC-Norfolk resource (http://www.srl.cam.ac.uk/epic/), and the expression data sets are publicly available from MetaXcan (https://github.com/hakyimlab/MetaXcan), SMR (http://csgenomics.com/software/smr/) and TWAS (http://gusevlab.org/projects/fusion/).