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

Daniel J. Wright*[1], Felix R. Day*[1], Nicola D. Kerrison*[1], Florian Zink*[2], Alexia Cardona[1], Patrick Sulem[2], Deborah J. Thompson[3], Svanhvit Sigurjonsdottir[2], Daniel F Gudbjartsson[2], Agnar Helgason[2], J. Ross Chapman[4], Steve P. Jackson[5,6], Claudia Langenberg[1], Nicholas J. Wareham[1], Robert A. Scott[1], Unnur Thorsteindottir[2,7], Ken K. Ong[1,8], Kari Stefansson*[2,7] and John R.B. Perry*[1]

1. MRC Epidemiology Unit, School of Clinical Medicine, University of Cambridge, Cambridge, UK

2. deCODE genetics/Amgen, Inc., IS-101 Reykjavik, Iceland

3. Centre for Cancer Genetic Epidemiology, Department of Public Health and Primary Care, University of Cambridge, Cambridge, UK

4. Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, UK

5. Wellcome Trust and Cancer Research UK Gurdon Institute, University of Cambridge, Cambridge, UK

6. Department of Biochemistry, University of Cambridge, Cambridge, UK

7. Faculty of Medicine, University of Iceland, Reykjavik, Iceland

8. Department of Paediatrics, University of Cambridge, Cambridge, UK

* denotes equal contribution

Correspondence to John R.B. Perry (John.perry@mrc-epid.cam.ac.uk)

ABSTRACT

The Y-chromosome is frequently lost in hematopoietic cells, representing the most common somatic mutation in men. However, the mechanisms regulating mosaic loss of chromosome-Y (mLOY), and its clinical relevance, are unknown. Using genotype array intensity data and sequence reads in 85,542 men, we identify 19 genomic regions ($P<5x10^{-8}$) associated with mLOY. Cumulatively, these loci also predicted X-chromosome loss in women (N=96,123, $P=4x10^{-6}$). Additional epigenome-wide methylation analyses in whole blood highlighted 36 differentially methylated sites associated with mLOY. Identified genes converge on aspects of cell proliferation and cell-cycle regulation, including DNA synthesis (*NPAT*), DNA damage response (*ATM*), mitosis (*PMF1-CENPN-MAD1L1*) and apoptosis (*TP53*). We highlight 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 enable a measure of cell-cycle efficiency at population scale, identifying genes implicated in aneuploidy, genome instability and cancer susceptibility.

INTRODUCTION

For over a century, errors in cell division have been described which result in too few or too many chromosomes in daughter cells, a cytogenetic feature termed aneuploidy. 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 focussed largely on the role of mitosis and mitotic checkpoint signalling, primarily in cellular and animal models[1,2]. Recent human genomic studies have shown that aneuploidy can be estimated using intensity data from standard genotyping arrays; an approach validated by DNA sequencing[3–5]. These population-based studies demonstrate that mLOY is more frequent than other mosaic chromosomal and structural mutations: indeed, around 1 in 5 men over 80 years of age has detectable Y mosaicism in whole blood-derived DNA[4], reflecting the capacity of some cells to survive without this chromosome.

Although a common feature in the general population, it remains unclear whether mLOY is relevant to disease susceptibility, or whether cells in tissues other than peripheral blood undergo similar rates of chromosomal loss. Population studies have identified correlations between mLOY and smoking status, an association which appears transient and reversible after smoking cessation[6]. Such epidemiological studies have also identified associations with non-hematological cancers[4,5] and Alzheimer's disease[7]; however, these observations are inconsistent[3] and possibly subject to confounding or reverse-causality.

The ability to assay a common measure of aneuploidy in large array-genotyped populations could enable systematic identification of variants/genes involved in cell division errors. This would in turn enable a better understanding of the mechanisms involved, and the potential causal consequences of aneuploidy on cancer risk, inferred using Mendelian randomisation approaches. To date, a single genomic association with mLOY near *TCL1A* has been reported (N=12,369), suggesting that germline variation influencing mosaic chromosome loss can be detected[3]. Here, we use data in up to 85,542 men, highlighting widespread genomic, transcriptomic and epigenetic signatures of mosaic Y chromosome loss. We also demonstrate that this approach can successfully identify genes implicated in cell cycle regulation, genome instability and cancer susceptibility.

RESULTS

As a proxy for mLOY, we estimated mean intensity log-R ratio of all array-genotyped Y-chromosome SNPs (mLRR-Y) in a sample of 67,034 male participants from the UK Biobank cohort (UKBB)[8]. A normal distribution centred around zero was observed (standard deviation = 0.067), with negative values indicating reduced Y chromosome abundance in the clonal blood cell population **(Supplementary Figure 1)**.

Consistent with previous reports[3,6], we observed a strong negative correlation between mLRR-Y and age (r=-0.21). A strong association with 'ever smoking' status was also observed ($P=3.05\times10^{-82}$), which in combination with age explained 4.74% of the trait variance (age alone = 4.45%). We sought to demonstrate the causal relationship between smoking and mLOY through the principle of Mendelian randomization, using a reported and widely used genetic instrument for smoking frequency[9]. By modelling genetic variants robustly associated with cigarettes smoked per day at the *CHRNA5-CHRNA3-CHRNB4* nicotinic receptor locus, we inferred a causal effect of smoking on decreased mLRR-Y (increased Y loss) (rs1051730 P=0.03 [$P_{never\text{-}smokers}$=0.41, $P_{ever\text{-}smokers}$=0.04]). This genetic association was confirmed in independent replication samples (EPIC Norfolk and deCODE combined N=18,508, P=0.009, overall combined P=0.004).

Many autosomal genetic variants are associated with mLOY

To identify novel genetic variants associated with mLOY, we performed a genome-wide association study of mLRR-Y as a quantitative trait in UKB. After stringent quality control (see **Methods**), the most significantly-associated SNPs were located at the previously reported[3] mLOY locus*, TCL1A* ($P=3.6\times10^{-23}$). In addition, we identified a further 18 novel signals at genome-wide significance ($P<5\times10^{-8}$), with no evidence for significant inflation of test statistics genome-wide (lambda=1.05) (**Supplementary Figures 2 and 3)**. Replication was subsequently performed in an independent set of 9,793 men with array intensity data, in addition to 8,715 men from deCODE with Y loss estimated using sequence reads (see **Methods**). Both replication datasets 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 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 many additional associated variants remain to be discovered.

We next used HaploReg[10] and sequence data from the deCODE study to functionally annotate identified variants and genes. This highlighted four signals containing highly correlated missense variants, implicating *MAD1L1* (rs1801368, $r^2>0.98$), *PMF1* (rs1052053, $r^2=1$), *NREP* (rs11559, $r^2=0.74$) and *NPAT* (rs2070661, $r^2=0.97$) as potential candidates.

To ascertain whether the identified signals are 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 nineteen loci exhibited consistently stronger associations with the bottom 5% of mLRR-Y (greatest mLOY) than with the top 5% (**Supplementary Table 1**), suggesting their impact was on mosaic Y chromosome loss rather than 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.

Genome-wide pathway analyses conducted on association results for continuous mLRR-Y highlighted five pre-defined biological pathways enriched for association (study-wise significant FDR<0.05), the most significant of which was 'Apoptosis' genes defined per the Kyoto Encyclopaedia of Genes and Genomes (KEGG)[11] (**Supplementary Table 2**). Other significant pathways included sulphur metabolism, susceptibility to colorectal, prostate and thyroid cancers, and progesterone-mediated oocyte maturation.

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

We next sought to understand whether our identified variants acted only on the Y chromosome, or 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, **Figure 1**). Chromosome X copy number was estimated to have a heritability of 26% (17.4-36.2%) in the deCODE data; comparable to that of Y chromosome loss. Cumulatively, the 19 Y loss SNPs significantly predicted X loss in women, with the expected direction of effect (**Figure 2**, $P=4x10^{-6}$).

Identifying transcriptomic and epigenetic signatures of mLOY

To identify potential functional transcripts mediating Y chromosome loss, we performed summary statistic approaches to infer gene expression associations using three analytical imputation approaches[12–14] in independent whole-blood expression datasets (**Supplementary Tables 3-5**). Across these datasets, eight genes (*HM13, SMPD2, TCL1A, SENP7, NPAT, ATM, ACAT1, CENPN*) were significantly associated with mLRR-Y, all of which mapped near to one of the 19 associated genetic signals from GWAS.

We additionally identified 36 methylation variable positions (MVPs) correlated with mLRR-Y levels in 569 whole-blood samples from the European Prospective Investigation of Cancer (EPIC)-Norfolk cohort[15] (**Supplementary Table 6**). All significant MVPs were in genomic regions distinct (>500kb) from the 19 mLOY loci, with the exception of four correlated methylation probes within the *TP53* gene region. To ascertain if any of the methylation changes represented causal drivers of mLOY, we next identified *cis*-methylation quantitative trait loci (meQTLs) in publicly available data[16] for all associated probes. In total, 20 probes had one or more genetic variants in *cis* which 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 $P=5.6x10^{-31}$, mLRR-Y $P=9x10^{-4}$). 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[17].

Genetic overlap with cancer susceptibility

Three mLOY signals are correlated with signals previously reported for basal cell carcinoma[18], glioma[19], neuroblastoma[20] (*TP53*), or testicular cancer[21,22] (*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 previously reported for prostate cancer susceptibility[23,24], 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 ($P_{EGGER\text{-}MR}$ = 0.26, **Supplementary Table 8, Supplementary Figure 4**). Under the hypothesis that susceptibility to many types of cancer may have a common basis in mitotic error, we performed a GWAS in UKB defining men with *any* diagnosed cancer as a case (N= 7,745 cases, 58,562 controls). This approach was recently used for multiple reproductive cancers, yielding several novel loci[25]. Applying the 19 mLRR-Y signals as an additive genetic instrument, there was no evidence of a dose-response relationship between genetically-modelled mLOY and cancer risk in men ($P_{EGGER\text{-}MR}$ = 0.94, **Supplementary Table 9 and Supplementary Figure 5**). To test the relationship between cancer risk and mLOY more comprehensively, we estimated the extent of shared genetic architecture across the whole genome using LD score regression[26]. This revealed an overall significant inverse relationship between mLRR-Y and cancer risk (rg=-0.42, P=0.02), which was not significant when considering only female cancer cases (rg=-0.06, P=0.64).

DISCUSSION

Our findings, together with previous reports, demonstrate that loss of the Y-chromosome in peripheral blood likely represents a proxy trait for the study of aneuploidy in large-scale populations, which can be readily estimated from sequencing reads or array-based genotyping data. The nature of the genes identified by our analyses suggests that genetic determinants of mLOY reflect general mechanisms of aneuploidy, which we speculate most frequently manifest in mLOY due to the higher capacity of cells to tolerate Y-chromosome loss. This hypothesis is supported by the observation that these same SNPs also predicted X chromosome loss in women, the second most frequent large-scale mosaic event[27].

Pathway analyses identified enrichment for cancer and apoptosis pathways associated with mLOY. This is further supported by the many well-established cell cycle regulation genes which we observed either as the closest gene to the association signal, or which were 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 (**Figure 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 segregation[28,29]. In particular, *TPX2* recruits the crucial mitotic enzyme, Aurora Kinase A, to the spindle[30], whilst *ATMIN* regulates expression of a dynein motor component (*DYNLL1*) which critically mediates spindle positioning[31–33] and also modulates Nek9 kinase signalling required for correct spindle formation and function[34]. Similarly, Rho-GEF 10 (*ARHGEF10,* for which we observe a nearby methylated signal) regulates centrosome duplication and prevents formation of multipolar spindles[35]. We identified a missense variant in *MAD1L1* (MAD1 mitotic arrest deficient like 1), a major component of the spindle assembly checkpoint (SAC). This represents a key cellular safeguard against chromosome mis-segregation (and subsequent ploidy errors), supressing metaphase-anaphase progression until chromatids are bi-orientated on a bipolar spindle at the metaphase plate[1]. During cytokinesis, *SEPT5* (septin

5, implicated in our methylation analysis) encodes a conserved cell cycle regulator required for effective cell division[36], while activation of signalling by Rho-GEF 10 (*ARHGEF10*) facilitates contractile ring ingression to separate the two daughter cells[37].

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 fragility[38–40]. G_1 to S-phase transition is dependent on *NPAT*, at least in part through it promoting histone gene transcription[41], while *ATM*, at least in part in association with *ATMIN*[42], acts as major cell cycle checkpoint kinase dedicated to maintaining genome stability throughout interphase, with particular importance at the G_1/S and G_2/M checkpoints[40]. In response to double-stranded DNA breaks (DSBs) 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 senescence[38–40,43]. *TREX1* encodes 3' Repair Exonuclease 1, which digests aberrant replication intermediates and single stranded DNA from genotoxic stress to prevent chronic checkpoint activation[44]. Predicted deleterious missense variants in this gene were recently identified in a mouse GWAS for micronucleus formation, a biomarker of chromosomal breaks, whole chromosome loss and extranuclear DNA[45].

At the later stages of the cell lifespan, several genes implicated by our GWAS findings – including *TP53, TCL1A, SMPD2, BCL2* and *BCL2L1* – functionally impact on apoptotic events[46–50]. Apoptosis is a prime mechanism by which cells with detected DNA damage or ploidy errors may be eliminated[51]: indeed, p53 drives multiple cell-cycle exit responses in response to aberrant mitosis, including G_1 arrest[43,52,53]. The *TP53* variant associated with mLOY in our analyses is the one previously reported for basal cell carcinoma: for this trait, the risk allele changes the AATAAA polyadenylation signal to AATACA, resulting in impaired 3′-end processing of *TP53* mRNA[18]. Our findings also implicated genes involved in spermatogenesis[54,55] (*HENMT1* and *DAZAP1*), and cellular growth and differentiation[56] (*DLK1)*.

The genes directly involved in mitotic prophase-metaphase and the SAC have clear roles in averting chromosomal mis-segregation and preventing these from persisting unchecked, however how the broader set of genes we identify here may act to promote mLOY remains less clear. We speculate that either many of these genes act in ways that are not currently recognised, 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 apoptotic regulatory genes and cascades may play 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 haematopoietic 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 play in determining

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 prioritise novel 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 surprising given that findings from mouse studies in which mitotic checkpoint components are experimentally down-regulated demonstrate an inconsistent relationship between aneuploidy and spontaneous tumorigenesis[1]. It is possible, therefore, that some of our identified genes may promote benign aneuploidy, whereas others may play a role more generally 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 to genes that promote aneuploidy in a more stable way. This of course does not preclude identifying causal risk factors for mLOY, 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 non-genetic modifiable factors which 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 may thereby add a new approach to the study of cellular ageing and its associations with disease, particularly cancer.

Data availability statement

The genome-wide discovery data used is from UK Biobank and can be obtained via application from 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. Methylation data is available from the same EPIC-Norfolk resource and gene expression datasets are publically available from three resources: MetaXcan (https://github.com/hakyimlab/MetaXcan) , SMR (http://cnsgenomics.com/software/smr/) and TWAS (http://gusevlab.org/projects/fusion/).

Acknowledgements

This research has been conducted using the UK Biobank Resource under Application Number 9905. This work was supported by the UK Medical Research Council (Unit Programme numbers MC_UU_12015/1 and MC_UU_12015/2). Research in the S. Jackson laboratory is funded by Cancer Research UK (CRUK; programme grant C6/A18796), with Institute core funding provided by CRUK (C6946/A14492) and the Wellcome Trust (WT092096). S. Jackson receives salary from the University of Cambridge, supplemented by CRUK. We thank the MRC Epidemiology genetics group members for useful Friday morning discussions.

Author contributions

All authors reviewed the original and revised manuscripts. Statistical analysis: D.J.W., F.R.D., N.D.K., F.Z., A.C., P.S., R.A.S., J.R.B.P. Individual study sample collection, genotyping and phenotyping: S.S., D.F.G., A.H., N.D.K., A.C., F.Z. Individual study principal investigators: C.L., N.J.W., U.T., K.K.O., K.S., J.R.B.P. Project design and interpretation of results: D.J.W., F.R.D., N.D.K., P.S., D.J.T., J.R.C., S.P.J., C.L., N.J.W., U.T., K.K.O., K.S., J.R.B.P.

Competing financial interests statement

F.Z., P.S., S.S., D.F.G., A.H., U.T. and K.S. are employees of deCODE Genetics/Amgen Inc. (Reykjavik, Iceland). R.A.S. is an employee of GlaxoSmithKline plc.

REFERENCES – Main Text

1. Holland, A. J. & Cleveland, D. W. Boveri revisited: chromosomal instability, aneuploidy and tumorigenesis. *Nat. Rev. Mol. Cell Biol.* **10,** 478–487 (2009).

2. Thompson, S. L., Bakhoum, S. F. & Compton, D. A. Mechanisms of chromosomal instability. *Curr. Biol.* **20,** R285–95 (2010).

3. Zhou, W. *et al.* Mosaic loss of chromosome Y is associated with common variation near TCL1A. *Nat. Genet.* **48,** 563–8 (2016).

4. 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,** 624–8 (2014).

5. Jacobs, K. B. *et al.* Detectable clonal mosaicism and its relationship to aging and cancer. *Nat. Genet.* **44,** 651–658 (2012).

6. Dumanski, J. P. *et al.* Mutagenesis. Smoking is associated with mosaic loss of chromosome Y. *Science* **347,** 81–3 (2015).

7. Dumanski, J. P. *et al.* Mosaic Loss of Chromosome Y in Blood Is Associated with Alzheimer Disease. *Am. J. Hum. Genet.* **98,** 1208–19 (2016).

8. Sudlow, C. *et al.* UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. *PLoS Med.* **12,** e1001779 (2015).

9. Thorgeirsson, T. E. *et al.* Sequence variants at CHRNB3-CHRNA6 and CYP2A6 affect smoking behavior. *Nat. Genet.* **42,** 448–53 (2010).

10. Ward, L. D. & Kellis, M. HaploReg: A resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. *Nucleic Acids Res.* **40,** (2012).

11. Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M. & Tanabe, M. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res.* **44,** D457–D462 (2016).

12. Gusev, A. *et al.* Integrative approaches for large-scale transcriptome-wide association studies. *Nat. Genet.* **48,** 245–52 (2016).

13. Gamazon, E. R. *et al.* A gene-based association method for mapping traits using

reference transcriptome data. *Nat. Genet.* **47,** 1091–1098 (2015).

14. Zhu, Z. *et al.* Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nat. Genet.* **48,** 481–487 (2016).

15. Day, N. *et al.* EPIC-Norfolk: study design and characteristics of the cohort. European Prospective Investigation of Cancer. *Br. J. Cancer* **80 Suppl 1,** 95–103 (1999).

16. Bonder, M. J. *et al.* Disease variants alter transcription factor levels and methylation of their binding sites. *Nat. Genet.* (2016). doi:10.1038/ng.3721

17. Henderson, M. C. *et al.* High-throughput RNAi screening identifies a role for TNK1 in growth and survival of pancreatic cancer cells. *Mol. Cancer Res.* **9,** 724–32 (2011).

18. Stacey, S. N. *et al.* A germline variant in the TP53 polyadenylation signal confers cancer susceptibility. *Nat. Genet.* **43,** 1098–1103 (2011).

19. Walsh, K. M. *et al.* Analysis of 60 reported glioma risk SNPs replicates published GWAS findings but fails to replicate associations from published candidate-gene studies. *Genet. Epidemiol.* **37,** 222–8 (2013).

20. Diskin, S. J. *et al.* Rare variants in TP53 and susceptibility to neuroblastoma. *J. Natl. Cancer Inst.* **106,** dju047 (2014).

21. Ruark, E. *et al.* Identification of nine new susceptibility loci for testicular cancer, including variants near DAZL and PRDM14. *Nat. Genet.* **45,** 686–9 (2013).

22. Chung, C. C. *et al.* Meta-analysis identifies four new loci associated with testicular germ cell tumor. *Nat. Genet.* **45,** 680–5 (2013).

23. Al Olama, A. A. *et al.* A meta-analysis of 87,040 individuals identifies 23 new susceptibility loci for prostate cancer. *Nat. Genet.* **46,** 1103–9 (2014).

24. Eeles, R. *et al.* The genetic epidemiology of prostate cancer and its clinical implications. *Nat. Rev. Urol.* **11,** 18–31 (2014).

25. Kar, S. P. *et al.* Genome-Wide Meta-Analyses of Breast, Ovarian, and Prostate Cancer Association Studies Identify Multiple New Susceptibility Loci Shared by at Least Two Cancer Types. *Cancer Discov.* 1–17 (2016). doi:10.1158/2159-8290.CD-15-1227

26. Bulik-Sullivan, B. *et al.* An atlas of genetic correlations across human diseases and traits. *Nat. Genet.* **47,** 1236–41 (2015).

27. Machiela, M. J. *et al.* Female chromosome X mosaicism is age-related and preferentially affects the inactivated X chromosome. *Nat. Commun.* **7,** 11843 (2016).

28. Cheeseman, I. M. & Desai, A. Molecular architecture of the kinetochore-microtubule interface. *Nat. Rev Mol. Cell Biol.* **9,** 33–46 (2008).

29. Kline, S. L., Cheeseman, I. M., Hori, T., Fukagawa, T. & Desai, A. The human Mis12 complex is required for kinetochore assembly and proper chromosome segregation. *J. Cell Biol.* **173,** 9–17 (2006).

30. Kufer, T. A. *et al.* Human TPX2 is required for targeting Aurora-A kinase to the

spindle. *J. Cell Biol.* **158,** 617–623 (2002).

31. Jurado, S. *et al.* ATM Substrate Chk2-interacting Zn2+ Finger (ASCIZ) Is a Bi-functional Transcriptional Activator and Feedback Sensor in the Regulation of Dynein Light Chain (DYNLL1) Expression. *J. Biol. Chem.* **287,** 3156–3164 (2012).

32. Dunsch, A. K. *et al.* Dynein light chain 1 and a spindle-associated adaptor promote dynein asymmetry and spindle orientation. *J. Cell Biol.* **198,** 1039–1054 (2012).

33. Zaytseva, O. *et al.* The Novel Zinc Finger Protein dASCIZ Regulates Mitosis in Drosophila via an Essential Role in Dynein Light-Chain Expression. *Genetics* **196,** 443–453 (2014).

34. Regue, L. *et al.* DYNLL/LC8 Protein Controls Signal Transduction through the Nek9/Nek6 Signaling Module by Regulating Nek6 Binding to Nek9. *J. Biol. Chem.* **286,** 18118–18129 (2011).

35. Aoki, T., Ueda, S., Kataoka, T. & Satoh, T. Regulation of mitotic spindle formation by the RhoA guanine nucleotide exchange factor ARHGEF10. *BMC Cell Biol.* **10,** 56 (2009).

36. Beites, C. L., Xie, H., Bowser, R. & Trimble, W. S. The septin CDCrel-1 binds syntaxin and inhibits exocytosis. *Nat. Neurosci.* **2,** 434–9 (1999).

37. Zuo, Y., Oh, W. & Frost, J. A. Controlling the switches: Rho GTPase regulation during animal cell mitosis. *Cell. Signal.* **26,** 2998–3006 (2014).

38. Mazouzi, A., Velimezi, G. & Loizou, J. I. DNA replication stress: Causes, resolution and disease. *Exp. Cell Res.* **329,** 85–93 (2014).

39. Zeman, M. K. & Cimprich, K. A. Causes and consequences of replication stress. *Nat. Cell Biol.* **16,** 2–9 (2014).

40. Osborn, A. J., Elledge, S. J. & Zou, L. Checking on the fork: the DNA-replication stress-response pathway. *Trends Cell Biol.* **12,** 509–16 (2002).

41. Gao, G. *et al.* NPAT expression is regulated by E2F and is essential for cell cycle progression. *Mol. Cell. Biol.* **23,** 2821–33 (2003).

42. 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).

43. Santaguida, S. & Amon, A. Short- and long-term effects of chromosome mis-segregation and aneuploidy. *Nat. Rev. Mol. Cell Biol.* **16,** 473–485 (2015).

44. 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).

45. McIntyre, R. E. *et al.* A Genome-Wide Association Study for Regulators of Micronucleus Formation in Mice. *G3 (Bethesda).* **6,** 2343–54 (2016).

46. Bieging, K. T., Mello, S. S. & Attardi, L. D. Unravelling mechanisms of p53-mediated tumour suppression. *Nat. Rev. Cancer* **14,** 359–70 (2014).

47. Yabu, T. *et al.* Stress-induced ceramide generation and apoptosis via the phosphorylation and activation of nSMase1 by JNK signaling. *Cell Death Differ.* **22,** 258–73 (2015).

48. Laine, J., Künstle, G., Obata, T., Sha, M. & Noguchi, M. The protooncogene TCL1 is an Akt kinase coactivator. *Mol. Cell* **6,** 395–407 (2000).

49. Czabotar, P. E., Lessene, G., Strasser, A. & Adams, J. M. Control of apoptosis by the BCL-2 protein family: implications for physiology and therapy. *Nat. Rev. Mol. Cell Biol.* **15,** 49–63 (2014).

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

51. Zhivotovsky, B. & Kroemer, G. Apoptosis and genomic instability. *Nat. Rev. Mol. Cell Biol.* **5,** 752–762 (2004).

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

53. Ganem, N. J. *et al.* Cytokinesis failure triggers hippo tumor suppressor pathway activation. *Cell* **158,** 833–848 (2014).

54. Lim, S. L. *et al.* HENMT1 and piRNA Stability Are Required for Adult Male Germ Cell Transposon Repression and to Define the Spermatogenic Program in the Mouse. *PLOS Genet.* **11,** e1005620 (2015).

55. Hsu, L. C.-L. *et al.* DAZAP1, an hnRNP protein, is required for normal growth and spermatogenesis in mice. *RNA* **14,** 1814–22 (2008).

56. Falix, F. A., Aronson, D. C., Lamers, W. H. & Gaemers, I. C. Possible roles of DLK1 in the Notch pathway during development and disease. *Biochim. Biophys. Acta - Mol. Basis Dis.* **1822,** 988–995 (2012).

Figure Legends

Figure 1
Estimated X and Y chromosome loss with age in the Icelandic deCODE study.
(A) Y chromosome copy number estimated in 8703 males from whole genome sequencing. (B) X chromosome copy-number for 9280 females. In each case, the black line indicates the line of best fit with age at blood collection as a linear predictor.

Figure 2
Association of 19 SNP mLOY genetic risk score on X loss in women.
The genetic risk score is additive, based on mLRR-Y increasing allele dosage.

Figure 3
Overview of identified genes implicated in Y chromosome loss.
Genes falling within GWAS loci are shown in blue, those implicated by methylation analyses in green. Grey boxes highlight specific checkpoints, signalling cascades, or enzymes of note. Green arrows denote activation of a target by phosphorylation, blue arrows a signalling cascade and its ultimate effect.

Table 1 | Genome-wide significant associations with Y chromosome loss.

SNP	Location	Alleles[1]	UK Biobank (N=67,034)		EPIC Norfolk (N=9,793)		deCODE (N=8,715)		Replication P	Overall P	Gene[4]
			Effect[2]	P	Effect[2]	P	Effect[3]	P			
rs17758695	18q21.33	C/T/0.97	-0.01	$6.4x10^{-21}$	-0.014	$3.7x10^{-04}$	-0.020	$9.1x10^{-13}$	$2.7x10^{-17}$	$1.3x10^{-33}$	*BCL2* [NC]
rs1122138	14q32.13	C/A/0.84	-0.005	$3.6x10^{-23}$	-0.006	$4.3x10^{-04}$	-0.007	$1.5x10^{-04}$	$8.0x10^{-10}$	$6.3x10^{-31}$	*TCL1A* [NEC]
rs78378222	17p13.1	G/T/0.01	-0.013	$1.3x10^{-15}$	-0.032	$1.8x10^{-06}$	-0.026	$3.8x10^{-10}$	$7.3x10^{-18}$	$3.4x10^{-28}$	*TP53* [CN]
rs59633341	3q25.1	A/AT/0.16	-0.004	$2.6x10^{-18}$	-0.009	$7.5x10^{-07}$	-0.007	$1.1x10^{-05}$	$8.5x10^{-14}$	$4.1x10^{-28}$	*TSC22D2* [N]
rs2736609	1q22	T/C/0.36	-0.003	$1.9x10^{-12}$	-0.003	$4.9x10^{-02}$	-0.006	$2.5x10^{-07}$	$2.4x10^{-10}$	$2.0x10^{-19}$	*PMF1* [CFN], *SEMA4A* [CE]
rs13191948	6q21	C/T/0.54	-0.002	$1.2x10^{-11}$	-0.006	$5.4x10^{-06}$	-0.005	$3.8x10^{-05}$	$4.5x10^{-12}$	$2.2x10^{-19}$	*SMPD2* [E], *CCDC162P* [NE]
rs60084722	20q11.21	CT/C/0.79	-0.003	$6.6x10^{-13}$	-0.002	$2.5x10^{-01}$	-0.006	$9.4x10^{-05}$	$1.5x10^{-6}$	$1.6x10^{-17}$	*TPX2* [NEC], *BCL2L1* [C], *HM13* [E]
rs381500	6q26	C/A/0.55	-0.002	$5.7x10^{-11}$	-0.002	$1.9x10^{-01}$	-0.005	$1.1x10^{-07}$	$1.8x10^{-7}$	$5.0x10^{-16}$	*QKI* [N]
rs56084922	5q22.1	G/A/0.08	-0.005	$2.9x10^{-13}$	-0.004	$1.2x10^{-01}$	-0.005	$1.6x10^{-03}$	$2.8x10^{-3}$	$3.0x10^{-15}$	*NREP* [N]
rs137952017	14q32.2	C/CT/0.85	-0.003	$1.2x10^{-09}$	-0.01	$1.3x10^{-07}$	-0.004	$4.0x10^{-04}$	$2.4x10^{-8}$	$4.0x10^{-15}$	*DLK1* [N]
rs4721217	7p22.3	T/C/0.4	-0.002	$6.5x10^{-10}$	-0.005	$2.8x10^{-04}$	-0.003	$1.1x10^{-05}$	$1.7x10^{-6}$	$3.5x10^{-14}$	*MAD1L1* [NFC]
rs35091702	8p12	C/CAAAAAAG/0.74	-0.002	$4.2x10^{-10}$	-0.004	$6.0x10^{-03}$	-0.002	$3.9x10^{-02}$	$6.5x10^{-3}$	$9.5x10^{-12}$	*RBPMS* [N]
rs4754301	11q22.3	A/G/0.55	-0.002	$1.3x10^{-09}$	-0.001	$5.4x10^{-01}$	-0.002	$2.8x10^{-02}$	$1.5x10^{-2}$	$6.5x10^{-11}$	*NPAT* [NF], *ATM* [C], *ACAT1* [E]
rs12448368	16q23.2	C/T/0.13	-0.003	$9.8x10^{-10}$	-0.002	$2.5x10^{-01}$	-0.003	$2.4x10^{-02}$	$2.2x10^{-2}$	$7.1x10^{-11}$	*CENPN* [NEC], *ATMIN* [CE]
rs11082396	18q12.3	C/T/0.13	-0.003	$3.3x10^{-09}$	-0.004	$6.7x10^{-02}$	-0.003	$1.2x10^{-01}$	$1.1x10^{-2}$	$1.2x10^{-10}$	*SETBP1* [N]
rs13088318	3q12.3	G/A/0.34	-0.002	$4.1x10^{-09}$	-0.0004	$7.7x10^{-01}$	-0.003	$1.7x10^{-02}$	$2.1x10^{-2}$	$2.7x10^{-10}$	*SENP7* [E]
rs77522818	17q21.33	A/T/0.96	-0.005	$1.3x10^{-09}$	-0.004	$3.0x10^{-01}$	-0.002	$2.4x10^{-01}$	$1.6x10^{-1}$	$8.8x10^{-10}$	*FAM117A* (N)
rs10687116	13q14.11	AGATG/A/0.8	-0.002	$2.6x10^{-08}$	-0.001	$5.8x10^{-01}$	-0.003	$5.8x10^{-02}$	$1.0x10^{-2}$	$8.8x10^{-10}$	*WBP4* [N]
rs115854006	3p21.31	C/T/0.96	-0.006	$3.7x10^{-08}$	-0.007	$5.4x10^{-02}$	0.002	$9.3x10^{-01}$	$3.4x10^{-1}$	$4.5x10^{-08}$	*TREX1* [C], *PLXNB1* [C]

1. mLRR-Y lowering allele / increasing allele / lowering allele frequency
2. Effect estimates in per-allele decreases in raw mean intensity log-R ratio units
3. Effect estimate per allele for copy number transformed log2(chrY copy-number)
4. Labelled gene where preceding nomenclature refers to [N] nearest (default), [C] biological candidate, [E] expression mediated by mLRR-Y associated SNPs, [F] non-synonymous variant in gene.

ONLINE METHODS

Estimating Y chromosome mosaicism in UK Biobank

We analysed data from the May 2015 release of imputed genetic data from UK Biobank[8], containing ~73M SNPs, short indels and large structural variants in 152,249 individuals. Full details have been published elsewhere[57]. Briefly, the samples were genotyped on two slightly different arrays - approximately 50,000 on the custom UK BiLEVE study array, and the remainder (~100,000) on the UK Biobank Axiom array (Affymetrix), which was specifically designed to optimize imputation performance in GWAS studies. Removal of SNPs with missing data, multi-allelic SNPs, SNPs with a minor allele frequency (MAF) <1%, and 1,037 sample outliers, resulted in a dataset with 641,018 autosomal SNPs in 152,256 samples for phasing and imputation. Imputation was performed using a reference panel created by merging the UK10K haplotype panel with the 1000 Genomes Phase 3 reference panel.

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

mLOY was estimated by calculating the mean log-R ratio (normalised signal intensity) of SNPs on the male-specific region of the Y chromosome. Signal intensity, genotype call and confidence files from Affymetrix Power Tools software were analysed using the PennCNV-Affy pipeline[58] to produce a log-R ratio (LRR) for each SNP. SNPs without LRR calculable on both arrays, or those flagged by UKB as failing QC, were excluded. Whole Y chromosome fluorescence signal intensity was summarised by calculation of mean LRR across all Y chromosome SNPs (mLRR-Y). After omission of monomorphic SNPs, genotyping and QC failures, 253 SNPs were available across all participants for derivation of mLRR-Y.

Association testing and signal selection

Autosomal SNPs were analysed by linear mixed models implemented in BOLT-LMM[59] 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 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 (N=32,539] vs ever N=34,329] smoking 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 following genome-wide significant signals were excluded from further consideration due to concerns of technical artefacts: rs61737590 (Chr1-27Mb), rs115979215 (Chr2-54Mb), rs1857807 (Chr2-115Mb), rs115722056 (Chr2-171Mb), rs73191481 (Chr3-105Mb), rs9289877 (Chr3-152Mb), rs77306208 (Chr3-194Mb), rs9269173 (Chr6-32Mb), rs117810108 (Chr7-130Mb), rs117941885 (Chr12-90Mb), rs118031436 (Chr15-57Mb), rs16961626 (Chr16-84Mb), rs58108384 (Chr20-7Mb), rs73892829 (Chr21-19Mb), rs116446488 (Chr22-24Mb). All were excluded due to fulfilment of 2 or more of the following criteria: a) singletons in regional association plots, b) significantly associated with genotype array status, c) associated with mLRR-Y in women (reflecting technical background intensity).

Replication

Replication was performed in two independent studies using two separate techniques.

The first comprised 9,793 men from the EPIC-Norfolk study[15], following the same protocol using GWAS array intensity data as described above.

Secondly, we analyzed whole-blood genome sequences of 8,715 Icelandic males[60] (age range 41-105 years, mean 63 years), that had been whole-genome sequenced by Illumina method to a mean depth of 37x.

As an estimate of chromosome Y copy-number we used the average read depth over chromosome Y, using exclusively X-degenerate regions. This was computed by samtools from bam files aligned to hg38 and normalized by genome-wide sequencing coverage for the subject. A total of 12 outlier individuals (copy-number greater than 1.25) were excluded.

Chromosome Y copy-number had a strong negative correlation with age at bleeding (Spearman correlation r=-0.50). For individuals older than 60 years at the time of sample collection, the distribution of chromosome Y copy-number has a heavy left tail with copy-numbers as low as 0.08.

Association analysis was performed using BOLT-LMM[59] after inverse normal transformation and adjustment for age at bleeding. To enable comparison with the estimates obtained from GWAS array intensity data, effect sizes for log2(chrY copy-number) were estimated using robust linear regression (rlm from R package MASS).

The fraction of variance explained by a given variant was calculated using the formula 2f(1-f)a^2, in which f denotes the minor allele frequency of the variant and a is the additive effect in standard deviations. Heritability estimates were calculated using the spearman rank correlation of the traits between sibling pairs (max N=1488).

X chromosome loss

Similarly to mLOY, X chromosome loss was estimated using two complementary methods. Firstly, mLRR-X was calculated in UK Biobank (N=75,595) and EPIC Norfolk (N=11,248), using the same methodology described for X loss. Secondly, a similar analysis was performed using whole blood genome sequences of 9,302 Icelandic females (age range 41-106 years, mean 63 years) whole-genome sequenced to a mean depth of 36x. The chromosome X copy-number was estimated from the average read depth over chromosome X, excluding paralogous regions PAR1 and PAR2, the X-transposed region, and the centromere. This estimate was normalized by genome-wide sequencing coverage for the subject and adjusted for the sequencing protocol. A total of 22 outlier individuals (copy-number greater than 2.5 or less than 1.5) were excluded. We observed a Spearman correlation of -0.28 between the chromosome X copy-number and age at bleeding.

Cancer GWAS

To understand the genomic relationship between cancer and mLOY, we defined an ‘any prevalent cancer’ variable in UKB using linked UK cancer registrations. Individuals with a reported age of diagnosis in the cancer registry were coded 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 (rg) were calculated between mLRR-Y and cancer using LD Score Regression[26].

In order to assess the possible causal links between cancer and mLOY we applied Mendelian Randomization methods, which have been described extensively elsewhere[61]. In order to be as conservative as possible we preferentially report results from the Egger regression method, though inverse weighted, median weighted and penalised median weighted analyses were also calculated.

Gene expression

To identify specific eQTL linked genes, we utilised three complementary approaches – SMR, TWAS and MetaXcan – enabling systematic integration of publicly available gene expression data with our genome-wide dataset.

Summary Mendelian Randomization (SMR) uses summary-level gene expression data to map potentially functional genes to trait-associated SNPs[14]. We ran this approach against the publicly available whole-blood eQTL dataset published by Westra *et al*[62], providing association statistics for 5,952 transcripts. A conservative significance threshold was set at $P<4.9x10^{-6}$ reflecting the number of genes tested genome-wide.

MetaXcan, a meta-analysis extension of the PrediXcan method[13], 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 which 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 transcriptome datasets with genome-wide SNP data and then regresses the imputed transcriptome levels with 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 utilized 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 transcriptome data from the GTEx Project[63]. Threshold for statistical significance was estimated using the Bonferroni correction for number of tested genes.

Finally, we used the recently described Transcriptome-wide Association Study (TWAS) approach[12] to infer gene expression association using two whole blood datasets (Young Finns Study and Netherlands Twin Registry cohorts). The threshold for significance was set to correct for the number of studies and genes ($P<1x10^{-5}$). Each of the three approaches described in this section were compared by estimating the correlation (r) of association Z scores across genes present in all three datasets. There was strong concordance between the 2,326 transcripts analysed across the three approaches/datasets; SMR vs. TWAS r=0.72, SMR vs. MetaXcan r=0.54, TWAS vs. 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 Methylation 450k BeadChip platform. After setting methylation markers with detection p-value ≥ 0.01 to missing, methylation beta values were calculated for each marker. Quantile normalisation of methylation betas was applied separately to different marker groups based on colour channel, probe type and M/U subtypes[64]. Samples with a sample call rate ≤0.99 were removed (n=77). Methylation beta

value distributions of the X, Y and autosomal chromosome markers were analysed separately and a further 11 sample outliers were excluded. Within each sample, markers with a marker call rate ≤ 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 (marker call rate) and 8,775 CpGs with a call rate ≤ 0.95 were excluded. 3,295 CpGs with multimodal distributions of methylation intensities, identified by the R package ENmix[65], which typically arise from technical artefacts were also excluded. A further 18,874 CpG sites which were previously identified as mapping to more than 1 genomic location[66] were also excluded. The final cleaned dataset comprised 442,920 autosomal CpG sites. To account for cell composition variability, we estimated counts of T lymphocyte subtypes, natural killer cells, monocytes, granulocytes and B lymphocytes using the minfi R package[67,68]. These were included as covariates in subsequent epigenome-wide regression models.

To examine the association between methylation markers and mLOY, we performed an epigenome-wide association analysis in all male EPIC-Norfolk methylation samples (n=569). mLRR-Y was regressed separately on each methylation marker, adjusted for type 2 diabetes status, age, current smoking status, estimated cell counts, and sample plate. Bonferroni correction was applied, accounting for the number of markers tested ($p=1\times10^{-7}$). Furthermore, we checked that no significant CpG sites had sequences which also mapped to the Y chromosome.

Association statistics for genetic variants within the probe vicinity and corresponding methylation levels (i.e cis-meQTLs) were available from the BIOS QTL browser (http://www.genenetwork.nl/biosqtlbrowser/)

Pathway analyses

Meta-Analysis Gene-set Enrichment of variaNT Associations (MAGENTA) was used to explore pathway-based associations in the full GWAS dataset. MAGENTA implements a gene set enrichment analysis (GSEA) based approach, as previously described[69]. Briefly, each gene in the genome is mapped to a single index SNP with the lowest P-value within a 110 kb upstream, 40 kb downstream window. This P-value, representing 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 false discovery rate (FDR) <0.05 in either analysis. In total, 3216 pathways from Gene Ontology, PANTHER, KEGG and Ingenuity were tested for enrichment of multiple modest associations with mLRR-Y.

REFERENCES – Online Methods

57. Allen, N. E., Sudlow, C., Peakman, T. & Collins, R. UK Biobank Data: Come and Get It. *Sci. Transl. Med.* **6,** 224ed4–224ed4 (2014).

58. Wang, K. *et al.* PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Res.* **17,** 1665–74 (2007).

59. Loh, P.-R. *et al.* Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nat. Genet.* **47,** 284–290 (2015).

60. Gudbjartsson, D. F. *et al.* Large-scale whole-genome sequencing of the Icelandic population. *Nat. Genet.* **47,** 435–44 (2015).

61. Bowden, J., Davey Smith, G., Haycock, P. C. & Burgess, S. Consistent Estimation in Mendelian Randomization with Some Invalid Instruments Using a Weighted Median Estimator. *Genet. Epidemiol.* **40,** 304–314 (2016).

62. Westra, H.-J. *et al.* Systematic identification of trans eQTLs as putative drivers of known disease associations. *Nat. Genet.* **45,** 1238–43 (2013).

63. Lonsdale, J. *et al.* The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.* **45,** 580–585 (2013).

64. Lehne, B. *et al.* A coherent approach for analysis of the Illumina HumanMethylation450 BeadChip improves data quality and performance in epigenome-wide association studies. *Genome Biol.* **16,** 37 (2015).

65. Xu, Z., Niu, L., Li, L. & Taylor, J. A. ENmix: a novel background correction method for Illumina HumanMethylation450 BeadChip. *Nucleic Acids Res.* **44,** e20–e20 (2016).

66. Naeem, H. *et al.* Reducing the risk of false discovery enabling identification of biologically significant genome-wide methylation status using the HumanMethylation450 array. *BMC Genomics* **15,** 51 (2014).

67. Houseman, E. A. *et al.* DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC Bioinformatics* **13,** 86 (2012).

68. Aryee, M. J. *et al.* Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinformatics* **30,** 1363–9 (2014).

69. Segrè, A. V, Groop, L., Mootha, V. K., Daly, M. J. & Altshuler, D. Common inherited variation in mitochondrial genes is not enriched for associations with type 2 diabetes or related glycemic traits. *PLoS Genet.* **6,** e1001058 (2010).