

Supplementary Information:

Hematopoietic mosaic chromosomal alterations and risk for infection among 767,891 individuals without blood cancer

Seyedeh M. Zekavat BS^{*1-3}, Shu-Hong Lin PhD^{*4}, Alexander G. Bick MD PhD^{5,2}, Aoxing Liu PhD⁶, Kaavya Paruchuri MD^{2,3,7}, Md Mesbah Uddin PhD^{2,3}, Yixuan Ye BS¹, Zhaolong Yu BS¹, Xiaoxi Liu PhD⁸, Yoichiro Kamatani PhD⁸, James P. Pirruccello MD^{2,3,7}, Akhil Pampana MS^{2,3}, Po-Ru Loh PhD^{2,7}, Puja Kohli MD MMSc^{10,11}, Steven A. McCarroll PhD^{12,13}, Benjamin Neale PhD^{12,14}, Eric A. Engels MD MPH⁴, Derek W. Brown PhD⁴, Jordan W. Smoller MD ScD^{12,15,16}, Robert Green MD MPH^{2,7,9}, Elizabeth W. Karlson MD MS^{7,17}, Matthew Lebo PhD^{18,19}, Patrick T. Ellinor MD PhD^{2,3,7}, Scott T. Weiss MD MS^{7,20}, Mark J. Daly⁶, The Biobank Japan Project, FinnGen Consortium, Chikashi Terao MD PhD^{8,21,22}, Hongyu Zhao PhD^{1,23}, Benjamin L. Ebert MD PhD^{2,24,25}, COVID-19 Host Genetics Initiative, Andrea Ganna PhD^{6,14,2}, Mitchell J. Machiela ScD MPH^{#4}, Giulio Genovese PhD^{#2,12,9}, Pradeep Natarajan MD MMSc^{#2,3,7}

1. Computational Biology & Bioinformatics Program, Yale University, New Haven, CT
2. Medical and Population Genetics and Cardiovascular Disease Initiative, Broad Institute of Harvard and MIT, Cambridge, MA
3. Cardiovascular Research Center, Massachusetts General Hospital, Boston, MA
4. Division of Cancer Epidemiology and Genetics, National Cancer Institute, Rockville, MD
5. Division of Genetic Medicine, Department of Medicine, Vanderbilt University Medical Center
6. Institute for Molecular Medicine Finland, Helsinki, Finland.
7. Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA
8. Laboratory for Statistical and Translational Genetics, RIKEN Center for Integrative Medical Sciences, RIKEN, Yokohama, Japan
9. Department of Genetics, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA
10. Division of Pulmonary and Critical Care Medicine, Massachusetts General Hospital, Boston, MA
11. Vertex Pharmaceuticals, Boston, MA
12. Stanley Center, Broad Institute of Harvard and MIT, Cambridge, MA
13. Department of Genetics, Harvard Medical School, Boston, MA
14. Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA
15. Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA
16. Department of Psychiatry, Harvard Medical School, Boston, MA
17. Division of Rheumatology, Inflammation and Immunity, Brigham and Women's Hospital, Boston, MA
18. Department of Pathology, Brigham and Women's Hospital and Harvard Medical School, Boston, MA
19. Laboratory for Molecular Medicine, Partners Healthcare, Cambridge, MA
20. Channing Division of Network Medicine, Brigham and Women's Hospital, Boston, MA
21. Clinical Research Center, Shizuoka General Hospital, Shizuoka, Japan
22. The Department of Applied Genetics, The School of Pharmaceutical Sciences, University of Shizuoka, Shizuoka, Japan
23. Department of Biostatistics, Yale School of Public Health, New Haven, CT
24. Department of Medical Oncology, Dana-Farber Cancer Institute, Boston, MA, USA
25. Howard Hughes Medical Institute, Boston, MA, USA

* These authors equally contributed to this work.

These authors equally supervised this work.

Table of Contents

- 1) **Supplementary Note 1:** Extended Methods (pg.4-10)
 - a) Cohorts
 - b) Mosaic chromosomal alteration detection
 - c) Other phenotype definitions
 - d) Secondary statistical analyses
 - e) Sensitivity analyses
- 2) **Supplementary Figures** (pg.11-26)
 - a) **Supplementary Note 2:** MGB Biobank mCA sample quality control analyses.
 - b) **Supplementary Note 3:** MGB Biobank mCA variant quality control analyses.
 - c) **Supplementary Note 4:** FinnGen mCA sample quality control analyses.
 - d) **Supplementary Note 5:** FinnGen mCA variant quality control analyses.
 - e) **Supplementary Note 6:** Total number of mCAs (A) and expanded mCAs (B) per individual in the UK Biobank for mCA carriers.
 - f) **Supplementary Note 7:** Prevalence of mCA categories by age bin across cohorts.
 - g) **Supplementary Note 8:** Prevalence of expanded mCA categories by age bin across cohorts.
 - h) **Supplementary Note 9:** Proportion of expanded autosomal mCAs carriers across lymphocyte count and percentage by age bin and sex.
 - i) **Supplementary Note 10:** Associations of any mCA with incident infections.
 - j) **Supplementary Note 11:** Suggestive associations ($P < 0.05$) of expanded autosomal mCAs with incident infection categories.
 - k) **Supplementary Note 12:** Associations of A) expanded ChrY and B) expanded ChrX mCAs with incident infections.
 - l) **Supplementary Note 13:** Associations of expanded autosomal mCAs with incident sepsis and among different age strata in the UK Biobank.
 - m) **Supplementary Note 14:** Associations of expanded autosomal mCAs with incident infections among A) those with antecedent cancer (ie: cancer prior to their infection) B) those without antecedent cancer.
 - n) **Supplementary Note 15:** Interaction of expanded autosomal mCAs with antecedent cancer prior to infection in the UK Biobank and MGB Biobank.
 - o) **Supplementary Note 16:** Correlated associations of 63 independent genome-wide significant variants associated with expanded mCAs (from Supplementary Data 3) between different mCA categories (expanded autosomal mCAs, expanded ChrX mCAs, expanded ChrY mCAs) in the UKB.
 - p) **Supplementary Note 17:** Association of a mLOY PRS consisting of 156 previously identified¹ independent genome-wide significant variants associated with mLOY, with different expanded mCA categories in UKB dFemales.
- 3) **Supplementary Tables** (pg.27-33)
 - a) **Supplementary Table 1:** Baseline summary statistics across the UK Biobank, MGB Biobank, and Biobank Japan among individuals analyzed.
 - b) **Supplementary Table 2:** mCA counts by cohort.
 - c) **Supplementary Table 3:** Association of potential risk factors with expanded autosomal mCAs in the UK Biobank.
 - d) **Supplementary Table 4:** Association of mCAs with mortality from incident infection in Biobank Japan.

- e) **Supplementary Table 5:** Sensitivity analysis of incident sepsis association in the UK Biobank with the addition of a 25-factor smoking covariate, lymphocyte count and percentage, BMI, and prevalent type 2 diabetes.
 - f) **Supplementary Table 6:** Sensitivity analysis of incident sepsis and pneumonia association in the UK Biobank among populations of individuals with different types of cancer prior to incident infection, where solid cancer is defined as any non-hematologic cancer.
- 4) **Supplementary Data:** (provided in external excel sheet)
- a) **Supplementary Data 1:** Infection phenotypic grouping categories across organ systems used in the UK Biobank, MGB Biobank, and Biobank Japan
 - b) **Supplementary Data 2:** Infection phenotypic grouping categories across organ systems used in FinnGen (includes ICD10, 9, and 8 codes as specified).
 - c) **Supplementary Data 3:** 63 independent loci identified in the expanded mCA GWAS
 - d) **Supplementary Data 4:** Transcriptome-wide association using GTExv8-whole blood (N=670)
 - e) **Supplementary Data 5:** Gene set pathway enrichment analysis of the transcriptome-wide analyses (from Supplementary Data 4) using the Elsevier Pathway Collection through EnrichR
 - f) **Supplementary Data 6:** Other phenotype definitions used in the UK Biobank COVID-19 sensitivity analyses
 - g) **Supplementary Data 7:** Other phenotype definitions used in MGB Biobank
 - h) **Supplementary Data 8:** Other phenotype definitions used in FinnGen
- 5) Supplementary References (pg. 34)

Supplementary Note 1: **Extended Methods**

Cohorts:

The UK Biobank, a population-based cohort of approximately 500,000 participants recruited from 2006-2010, had existing genomic and longitudinal phenotypic data². Baseline assessments were conducted at 22 assessment centres across the UK with sample collections including blood-derived DNA. Of 488,377 genotyped individuals, we analyzed 445,101 participants consenting to genetic analyses and who passed sample quality control criteria for mCA calling, had genotypic-phenotypic sex concordance, no 1st or 2nd degree relatives (random exclusion of one from each pair), and no prevalent hematologic cancer at time of blood draw. Genome-wide genotyping of blood-derived DNA was performed by UK Biobank using two genotyping arrays sharing 95% of marker content: Applied Biosystems UK BiLEVE Axiom Array (807,411 markers in 49,950 participants) and Applied Biosystems UK Biobank Axiom Array (825,927 markers in 438,427 participants) both by Affymetrix (Santa Clara, CA)². Secondary use of the data was approved by the Massachusetts General Hospital Institutional Review Board (protocol 2013P001840) and facilitated through UK Biobank Applications 7089 and 21552.

The MGB Biobank (MGBB) contains genotypic and clinical data from >105,000 patients who consented to broad-based research across 7 regional hospitals³. Baseline phenotypes were ascertained from the electronic medical record (EMR) and surveys on lifestyle, environment, and family history. Of the approximately 36,000 genotyped individuals, 27,778 samples had available probe raw intensity data (IDAT) files for mCA calling. Blood-derived DNA samples were genotyped using three versions of the Multi-Ethnic Genotyping Array (MEGA) SNP array

offered by Illumina. Secondary use of the data was approved by the Massachusetts General Hospital Institutional Review Board (protocol 2020P000904).

The FinnGen project (<https://www.finnngen.fi/en>), launched in 2017, covers the whole of Finland and aims to improve health of people around the world through genetic studies. The latest released version (R6) contains genotypic, demographic, and extensive health (e.g. national inpatient/outpatient registers since 1969/1998, cancer register since 1953, and drug reimbursement register since 1964) information from 269,077 Finnish individuals. Blood-derived DNA samples were genotyped using two versions of FinnGen ThermoFisher Axiom custom array (<https://www.finnngen.fi/en/researchers/genotyping>) provided by the Thermo Fisher genotyping service facility.

Biobank Japan (BBJ) is a hospital-based registry that collected clinical, DNA, and serum samples from approximately 200,000 consented patients with one or more of 47 target diseases at a total of 66 hospitals between 2003-2007⁴. Blood DNA was genotyped in three batches using different arrays or set of arrays, namely: (1) a combination of Illumina Infinium Omni Express and Human Exome; (2) Infinium Omni Express Exome v.1.0; and (3) Infinium Omni Express Exome v.1.2, which capture very similar SNPs. These analyses were approved by the ethics committees of RIKEN Center for Integrative Medical Sciences and the Institute of Medical Sciences, the University of Tokyo.

Mosaic chromosomal alteration detection:

Mosaic chromosomal alteration (mCA) detection in the UK Biobank was as described previously^{5,6}. Briefly, genotype intensities were transformed to $\log_2(\text{R ratio})$ (LRR) and B-allele frequency (BAF values) to estimate total and relative allelic intensities, respectively. Re-phasing was performed using Eagle2⁷ and mCA calling was performed by leveraging long-range phase information to search for allelic imbalances between maternal and paternal allelic fractions across contiguous genomic segments. Constitutional duplications and low-quality calls were filtered out and cell fraction was estimated as previously described⁵. UK Biobank mCA calls were obtained from dataset Return 2062 generated from UK Biobank application 19808.

Detection of mCAs in the MGB Biobank was performed starting from raw IDAT intensity files from the Illumina Multi-Ethnic Global Array (MEGA). Genotype clustering was performed using the Illumina GenCall algorithm. The resulting GTC genotype files were converted to VCF files using the bcftools gtc2vcf plugin (<https://github.com/freeseek/gtc2vcf>). Genotype phasing across the whole cohort was performed using SHAPEIT4⁴ in windows of a maximum of 20 centimorgans with 2 centimorgans of overlap between consecutive windows. Phased genotypes were ligated across overlapping windows using bcftools concat (<https://github.com/samtools/bcftools>). mCA detection in the MGB Biobank was performed with MoChA^{1,2} (<https://github.com/freeseek/mocha>). A pipeline to execute the whole workflow from raw files all the way to final mCA calls is available in WDL format for the Cromwell execution engine⁸ as part of MoChA. We excluded 160 samples with phased B-allele frequency (BAF) auto-correlation >0.05 , indicative of contamination or other potential sources of poor DNA quality, and 67 samples with phenotype-genotype sex discordance (**Supplementary Note 2**). We removed likely germline copy number polymorphisms ($\text{lod_baf_phase} < 20$ for autosomal

variants and lod_baf_phase <5 for sex chromosome variants), constitutional or inborn duplications (mCAs <2Mb with relative coverage >2.25, and mCAs 2-10Mb with relative coverage >2.4) and deletions (filtering out mCAs with relative coverage <0.5) (**Supplementary Note 3**).

FinnGen blood samples are genotyped by two versions of FinnGen ThermoFisher Axiom custom array. The detection of mCAs in FinnGen was performed starting from the genotype/intensity tables of 201,322 samples by using the “txt” mode of the MoChA WDL pipeline (<https://github.com/freeseek/mocha>). The input genotype/intensity tables for mCA detection were directly provided by the Thermo Fisher genotyping service, which performed genotype calling from the raw CEL files for each batch using the apt-probeset-genotype tool. Genotype phasing across the whole cohort was performed using SHAPEIT44 in windows of a maximum of 20 centimorgans with 2 centimorgans of overlap between consecutive windows. Phased genotypes were ligated across overlapping windows using bcftools concat (<https://github.com/samtools/bcftools>). We excluded 215 samples with phased B-allele frequency (BAF) auto-correlation >0.05, indicative of contamination or other potential sources of poor DNA quality, and 83 samples with phenotype-genotype sex discordance (**Supplementary Note 4**). We removed likely germline copy number polymorphisms (LOD_BAF_PHASE <20 for autosomal variants and LOD_BAF_PHASE <5 for sex chromosome variants, and LOD_BAF_PHASE <10 unless they are larger than 5Mbp (or 10 Mbp if they span the centromere)), constitutional or inborn duplications (0.5-5Mbp mCAs with relative coverage >2.5 and Bdev<0.1 and 5-10Mbp mCAs with relative coverage >2.75) and deletions (filtering out mCAs with relative coverage <0.5) (**Supplementary Note 5**). After further removing 1st or 2nd

degree relatives, individuals with any prevalent hematologic cancer history at time of blood draw for genotyping, there were 175,690 samples remaining for analyses.

The detection of mCAs in the BBJ was as described previously⁹. Briefly, genotyping intensity data was analysed across variants shared between the three primary arrays, and used to compute BAF and LRR. Phasing was performed using the Eagle2 software. Mosaic events were called as previously described⁵.

Other phenotype definitions:

Covariate definitions in the UK Biobank, including type 2 diabetes mellitus, hypertension, coronary artery disease, asthma, and chronic obstructive pulmonary disease, are provided in **Supplementary Data 6**. Cancer cases in the UK Biobank were identified using the cancer register (Category 100092) in combination with inpatient ICD-10 registry (Field IDs 41270/41280) and hematologic cancer cases were identified using the cancer registry's Field ID 40011 (hematological cancer identified from biopsy), Field ID 40005/40006 C81-96, D45-47, and inpatient ICD-10 registry (Field ID 41270/41280 C81-96, D45-47).

In the MGBB, cancer cases were identified using ICD-10 C00-D49, and hematologic cancer cases were identified using C81-96, D45-47. Other phenotype definitions in the MGBB are provided in **Supplementary Data 7**. Smoking status in MGBB was defined using a combination of electronic health record data and survey data. Follow-up time was coded as time from blood draw for genotyping to event (development of incident phenotype) or, for controls, time from

sample collection to either censor date (10/31/19) or date of death if the patient died prior to the last censor.

Cancer and covariate phenotype definitions from FinnGen grouped across ICD 8, 9, and 10 codes are provided in **Supplementary Data 6**. Smoking status in FinGen was defined based on survey data. Follow-up time was coded as time from blood draw for genotyping to event (development of incident phenotype) or, for controls, time from sample collection to either censor date (12/31/19) or date of death if the patient died prior to the last censor.

Secondary statistical analyses:

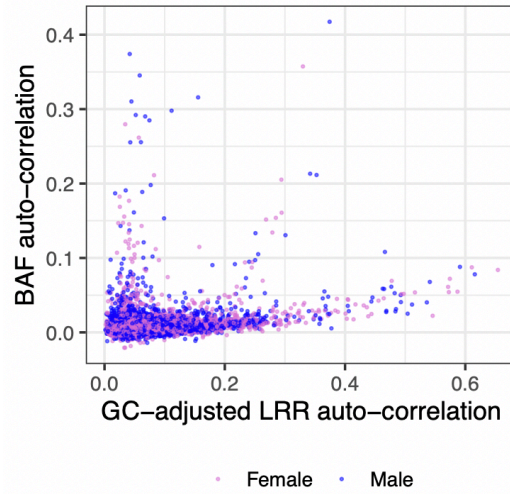
Secondary associations were performed across other mCA exposures: all mCAs, all expanded autosomal mCAs, all autosomal mCAs, all ChrX mCAs, expanded ChrX mCAs, all ChrY mCAs, and expanded ChrY mCAs. Of note, approximately 99% of ChrX and ChrY mCA calls were loss of ChrX and ChrY (mLOX and mLOY, respectively). P-value threshold for significance among the secondary exposures was $0.05/10/7=0.0007$. Additionally, associations of expanded autosomal mCAs with 20 secondary sub-outcomes separate from the organ system infections (listed under “category” in **Supplementary Data 1**) were performed to detect infection-specific associations. All association analyses were performed adjusted for age, age², sex, current or prior smoking history, and principal components 1-10 of genetic ancestry. P-value threshold for significance across these secondary analyses accounting for multiple hypothesis-testing was $0.05/20=0.0025$.

Sensitivity analyses:

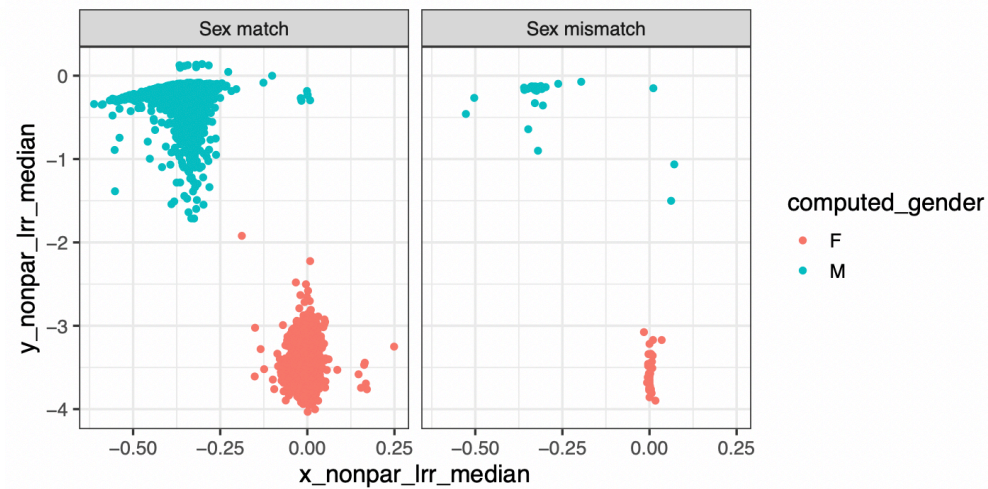
Further sensitivity analyses were performed in the UK Biobank expanded autosomal mCA and infection associations. First, stratified cancer analyses were performed among individuals with antecedent cancer prior to their incident infection in both the UK and MGB Biobanks, additionally stratifying for the same aforementioned covariates (age, age², sex, ever smoking status, and the first ten principal components of genetic ancestry). Secondly, interaction analysis was performed using a mCA x antecedent cancer term in the model to analyze the interaction between mCAs and antecedent cancer prior to incident infection. Thirdly, for the incident sepsis association, adding four sets of covariates to the Cox proportional hazards model: 1) normalized body mass index and type 2 diabetes mellitus, 2) any antecedent cancer prior to incident infection, 3) adjusting for a more comprehensive 25-factor smoking phenotype¹⁰, and 4) adjusting for normalized leukocyte count, lymphocyte count, and lymphocyte percentage at baseline visit. Fourthly, we evaluated the association of expanded autosomal mCAs incident sepsis and pneumonia associations among subgroups of individuals with cancer prior to infection including: prevalent solid cancer, incident hematologic cancer, and incident solid cancer prior to infection, in models adjusted for age, age², sex, ever smoking status, and the first ten principal components of genetic ancestry. Lastly, we further evaluated the association of expanded autosomal mCAs with incident pneumonia and sepsis in separate models adjusted for different predictors of cancer morbidity including chemotherapy, neutropenia, aplastic anemia, decreased white blood cell count, bone marrow or stem cell transplant, and radiation effects prior to infection (with these phenotypes defined using the Vanderbilt ICD-10 and ICD-9 phencode groupings¹¹), in the same aforementioned models adjusted for age, age², sex, ever smoking status, and the first ten principal components of genetic ancestry.

Supplementary Figures:

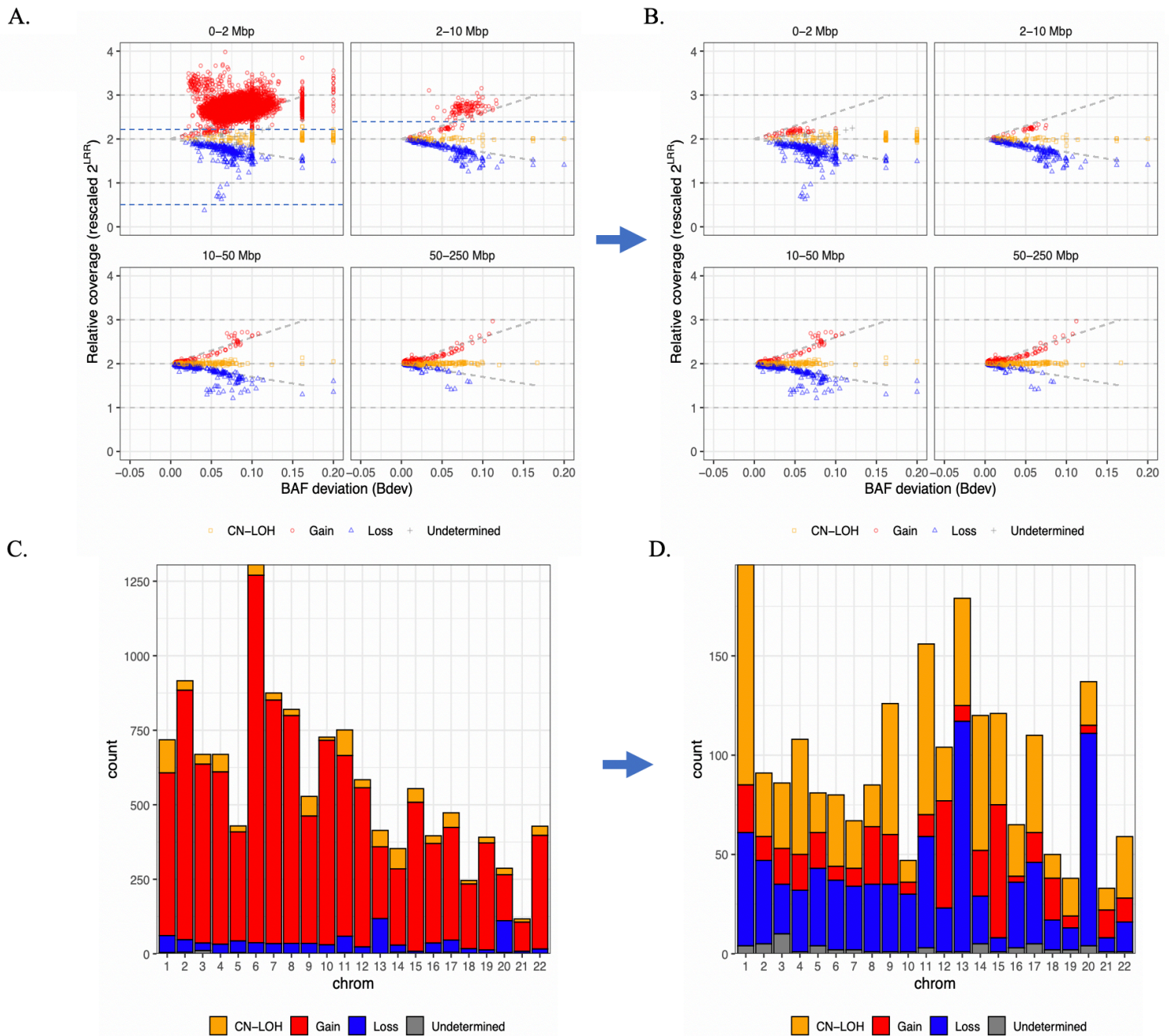
A.



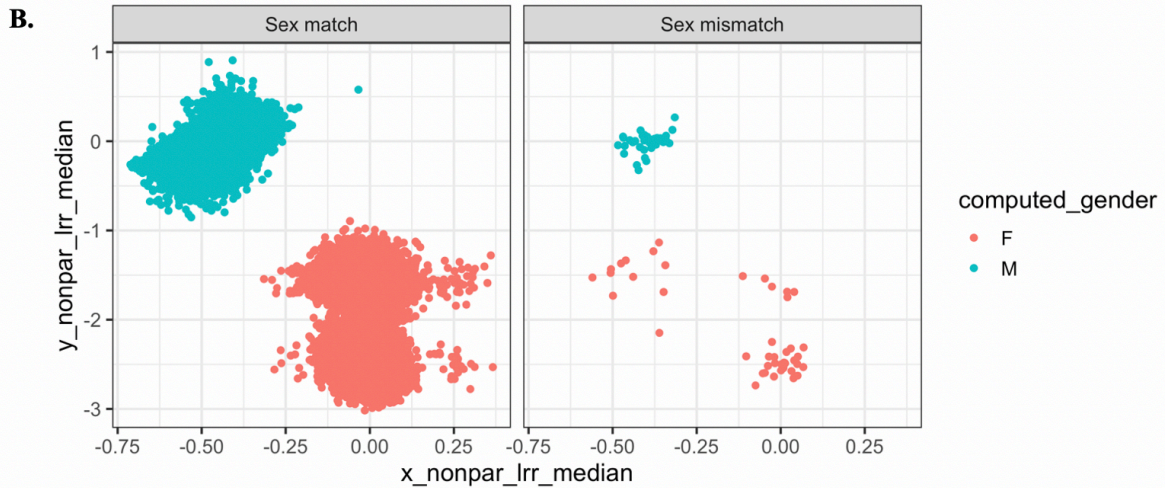
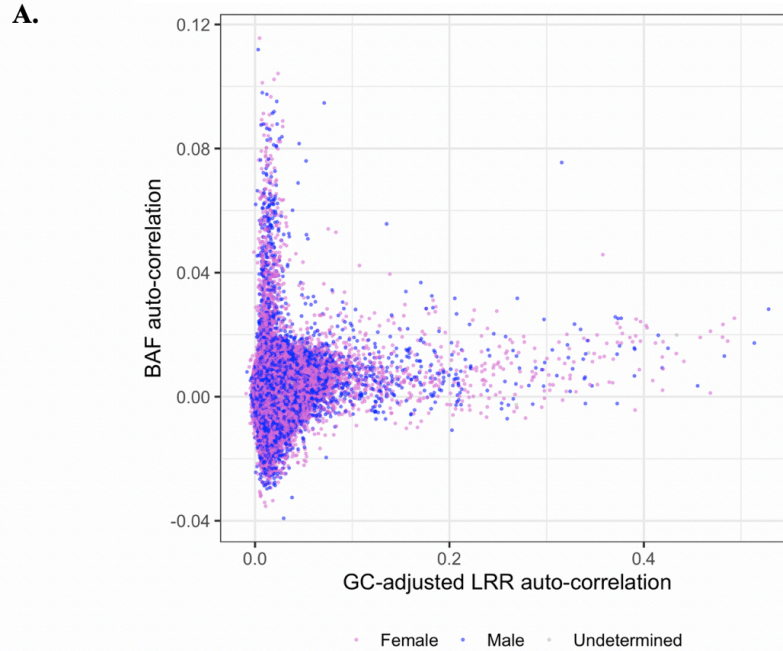
B.



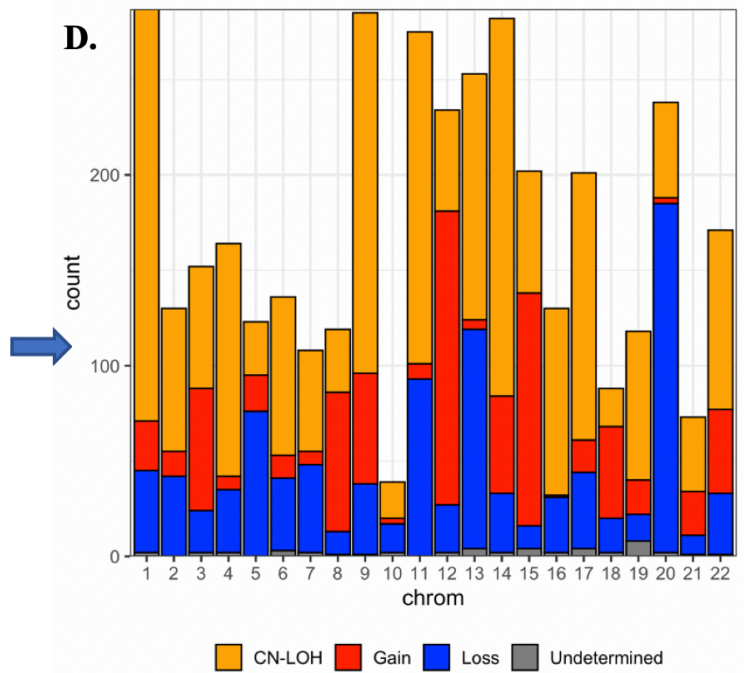
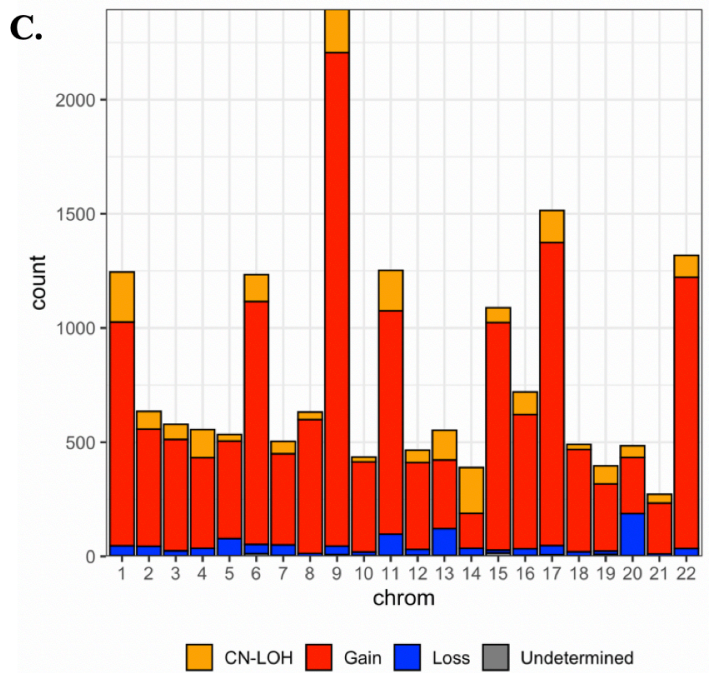
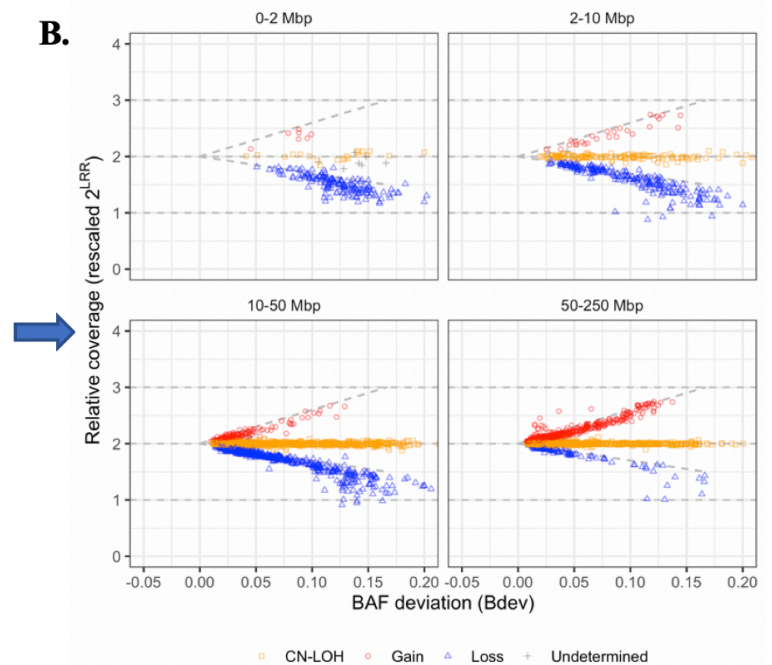
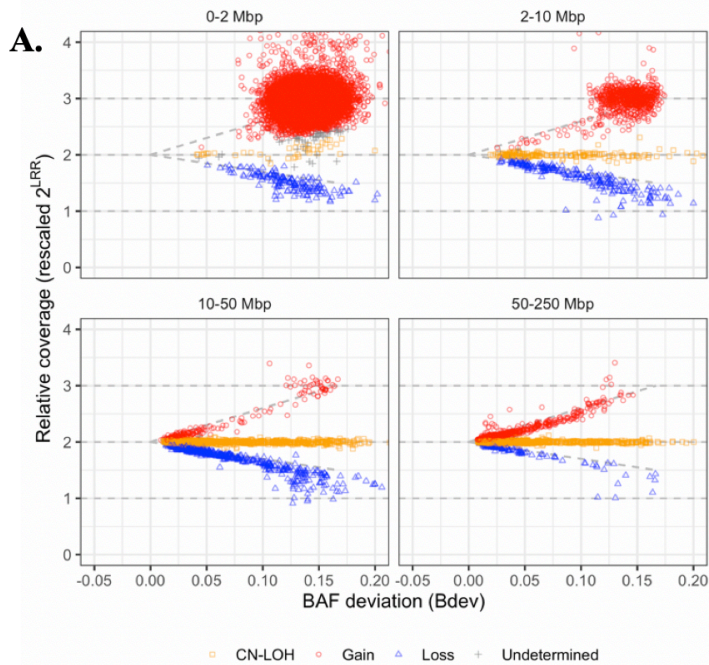
Supplementary Note 2: MGB Biobank mCA sample quality control analyses. A. plotting sample-level phased B allele frequency (BAF) auto-correlation across consecutive phased heterozygous sites versus Log R Ratio (LRR) of intensities using local GC content. B. Showing sex mismatches between MoChA-derived sex computed using the chrX nonPAR region versus reported sex. LRR=Log R Ratio, MoChA = mosaic chromosomal alterations caller, BAF=B allele frequency, nonPAR = Non-pseudoautosomal region



Supplementary Note 3: MGB Biobank mCA variant quality control analyses. Plots A. and C. represent mCAs carried among the quality-control filtered sample set, and after removal of likely germline variants ($\text{LOD_BAF_PHASE} < 20$ for autosomal or mCAs annotated as known germline copy number polymorphisms). Plots B. and D. reflect additional variant quality control filters to remove constitutional duplications (0-2Mbp mCAs with relative coverage > 2.25 and 2-10Mbp mCAs with relative coverage > 2.4) and remove constitutional deletions (mCAs with relative coverage < 0.5). CN-LOH = copy-neutral loss of heterozygosity, LOD_BAF_PHASE = log odds score for mCA model based on BAF and genotype phase, mCA = mosaic chromosomal alteration

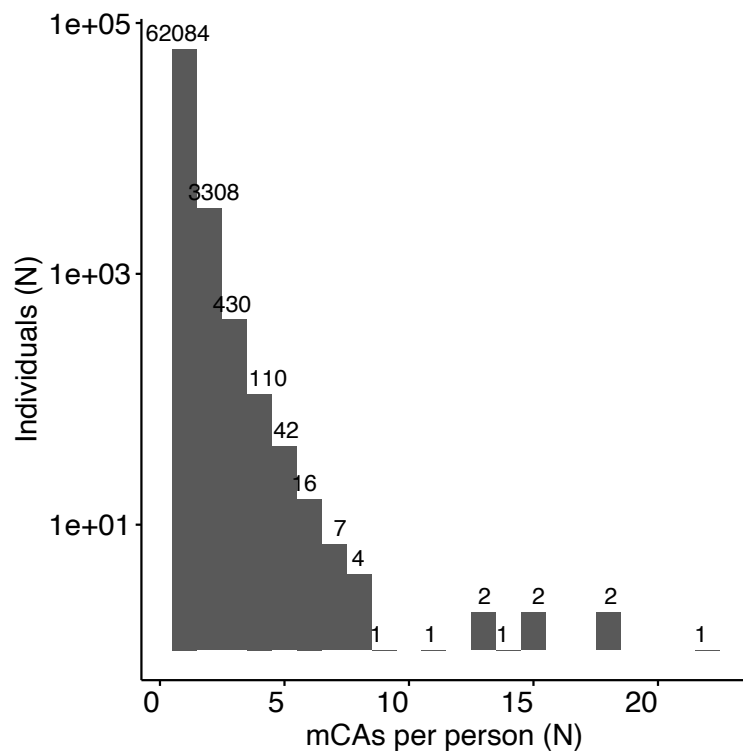


Supplementary Note 4: FinnGen mCA sample quality control analyses. A. plotting sample-level phased B allele frequency (BAF) auto-correlation across consecutive phased heterozygous sites versus Log R Ratio (LRR) of intensities using local GC content. B. Showing sex mismatches between MoChA-derived sex computed using the chrX nonPAR region versus reported sex. LRR=Log R Ratio, MoChA = mosaic chromosomal alterations caller, BAF=B allele frequency, nonPAR = Non-pseudoautosomal region

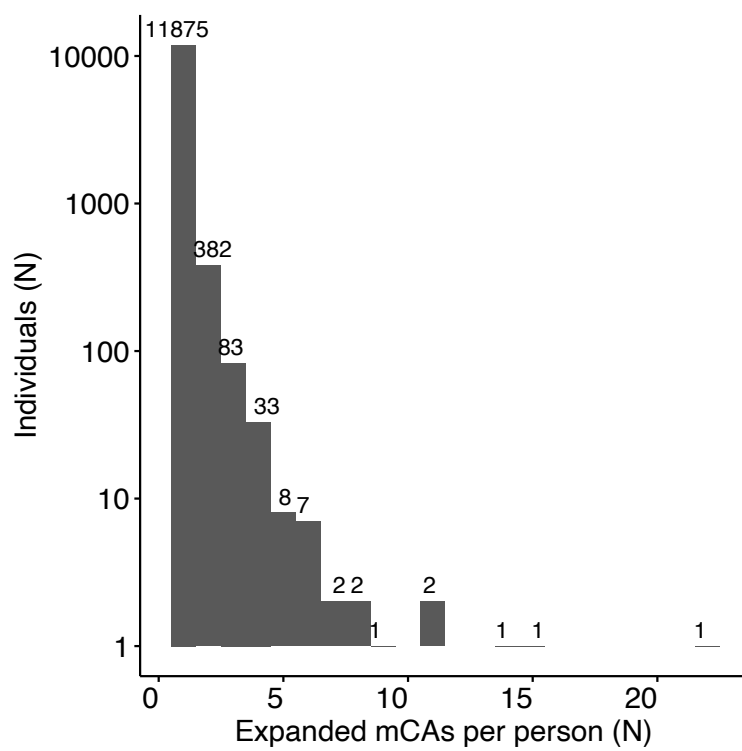


Supplementary Note 5: FinnGen mCA variant quality control analyses. Plots A. and C. represent mCAs after removal of likely germline variants ($LOD_BAF_PHASE < 20$ for autosomal variants or mCAs annotated as known germline copy number polymorphisms). Plots B. and D. reflect additional sample quality control filters including removal of samples with call rate < 0.97 and $BAF_AUTO > 0.03$, and removal of calls with $LOD_BAF_PHASE < 10$ unless they are larger than 5Mbp (or 10 Mbp if they span the centromere) and calls that are likely constitutional duplications (0.5-5Mbp mCAs with relative coverage > 2.5 and $Bdev < 0.1$ and 5-10Mbp mCAs with relative coverage > 2.75) and constitutional deletions (mCAs with relative coverage < 0.5). CN-LOH = copy-neutral loss of heterozygosity, LOD_BAF_PHASE = log odds score for mCA model based on BAF and genotype phase, mCA = mosaic chromosomal alteration

A.

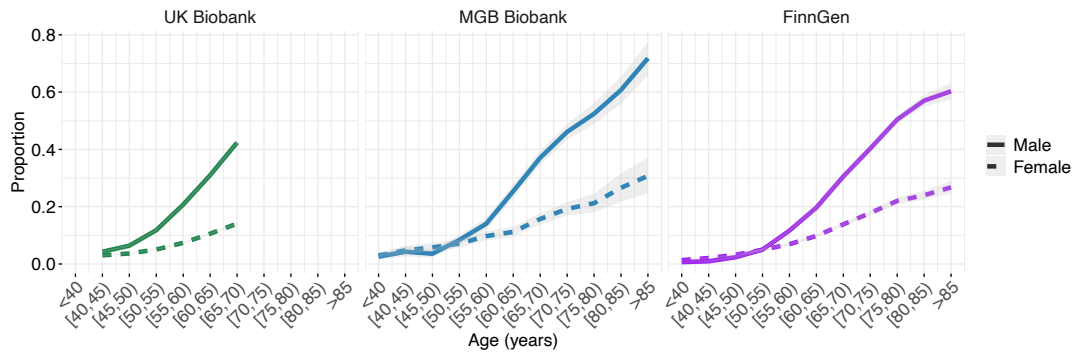


B.

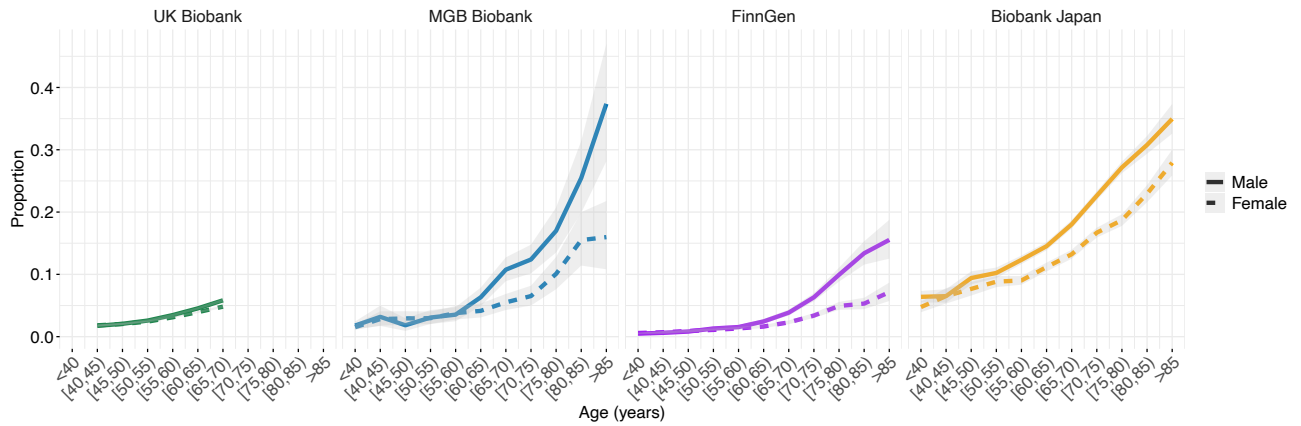


Supplementary Note 6: Total number of mCAs (A) and expanded mCAs (B) per individual in the UK Biobank for mCA carriers. mCAs = mosaic chromosomal alterations

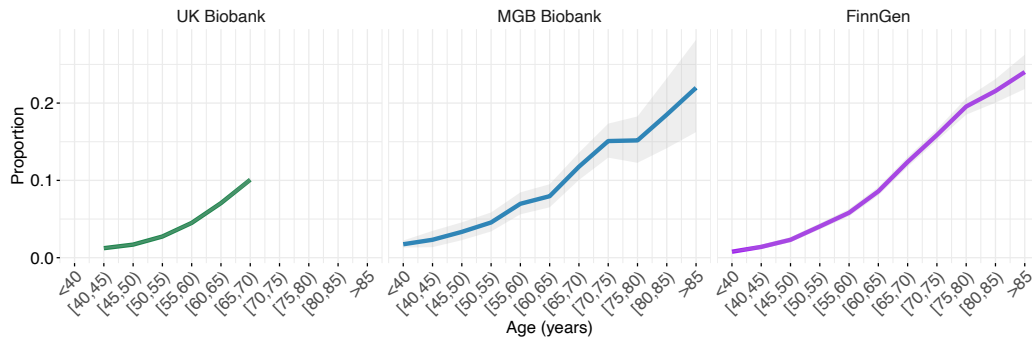
A. Any mCA:



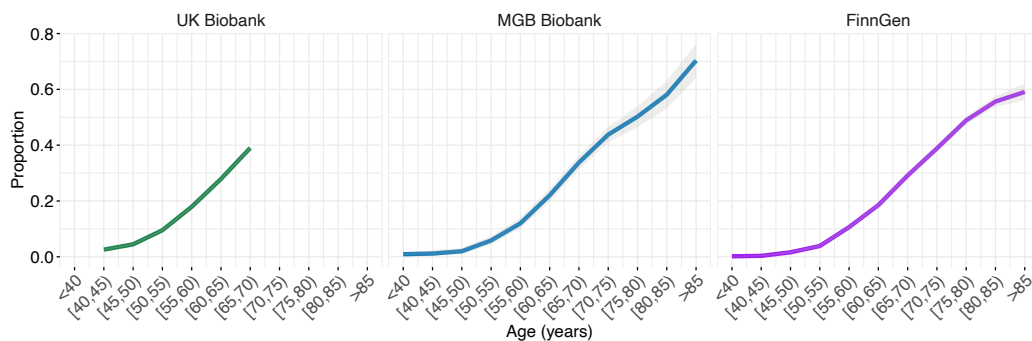
B. Autosomal mCA:



C. ChrX:

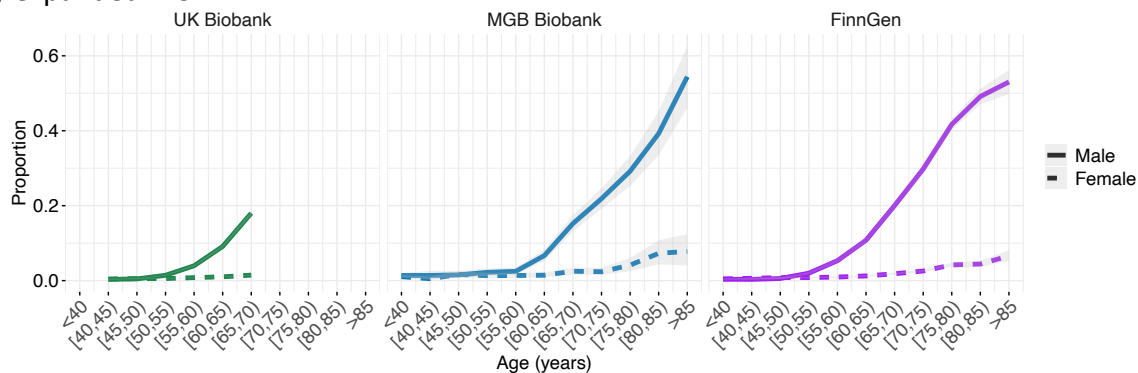


D. ChrY:

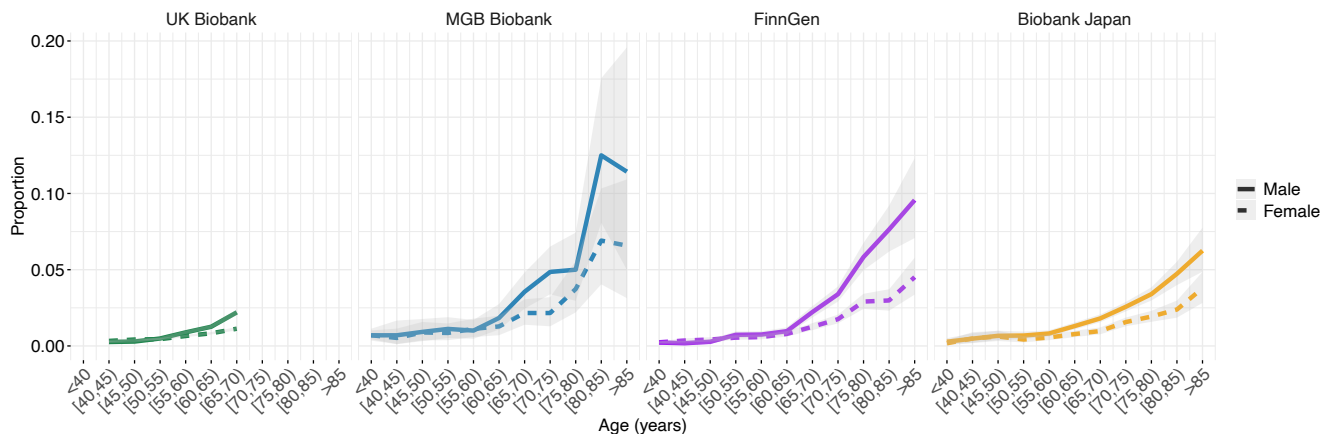


Supplementary Note 7: Prevalence of mCA categories by age bin across cohorts. mCA = mosaic chromosomal alterations.

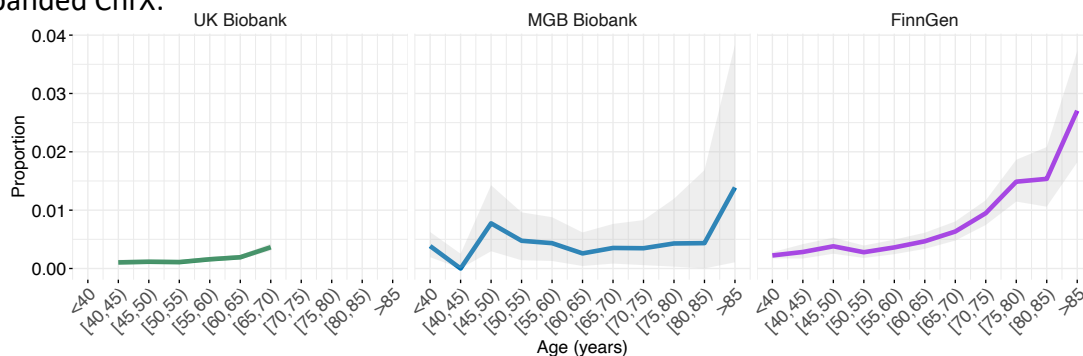
A. Any expanded mCA:



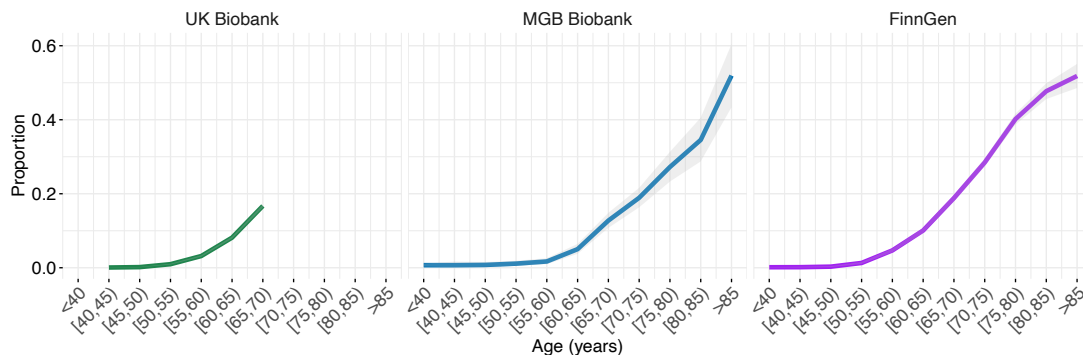
B. Expanded Autosomal mCA:



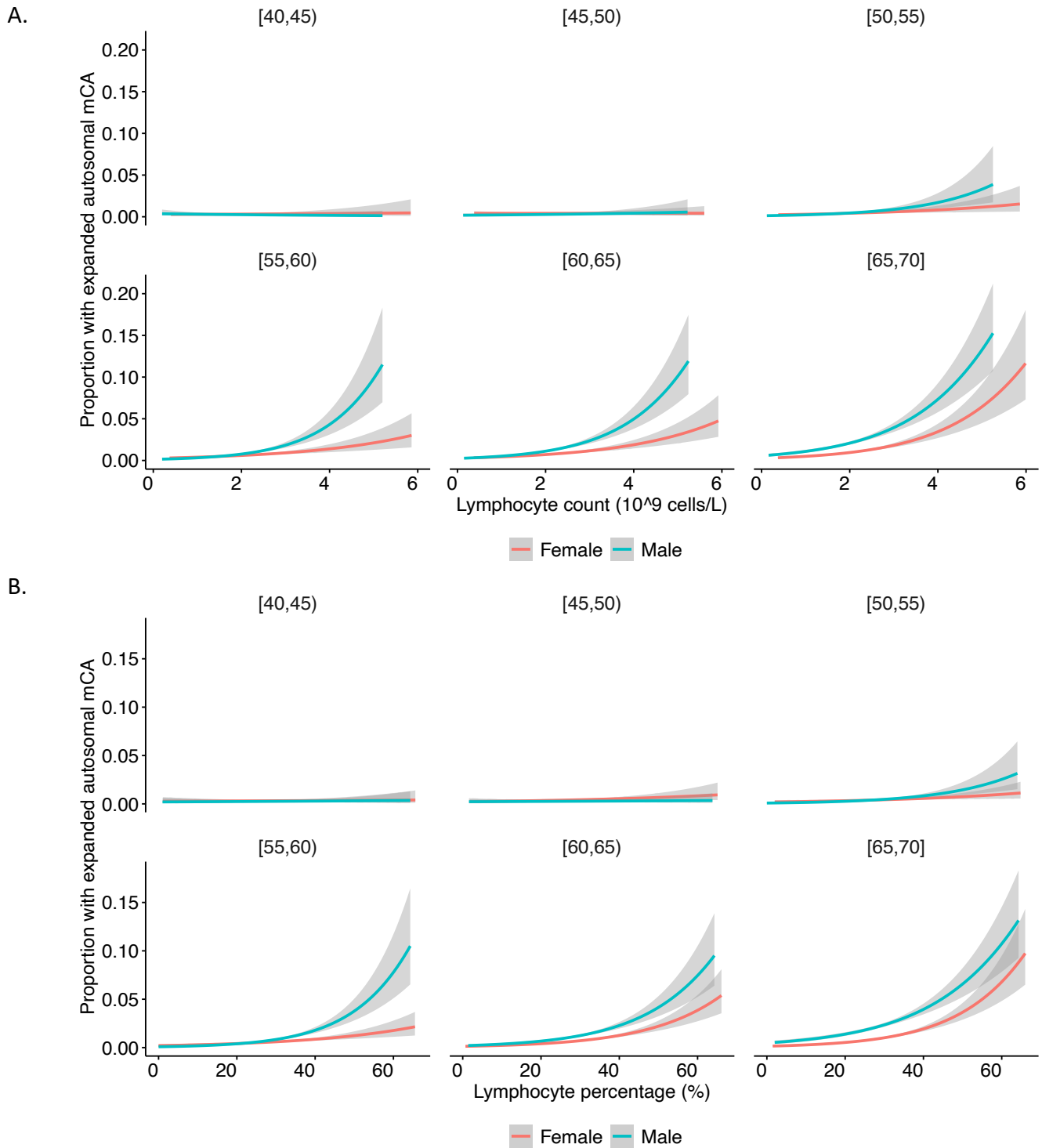
C. Expanded ChrX:



D. Expanded ChrY:

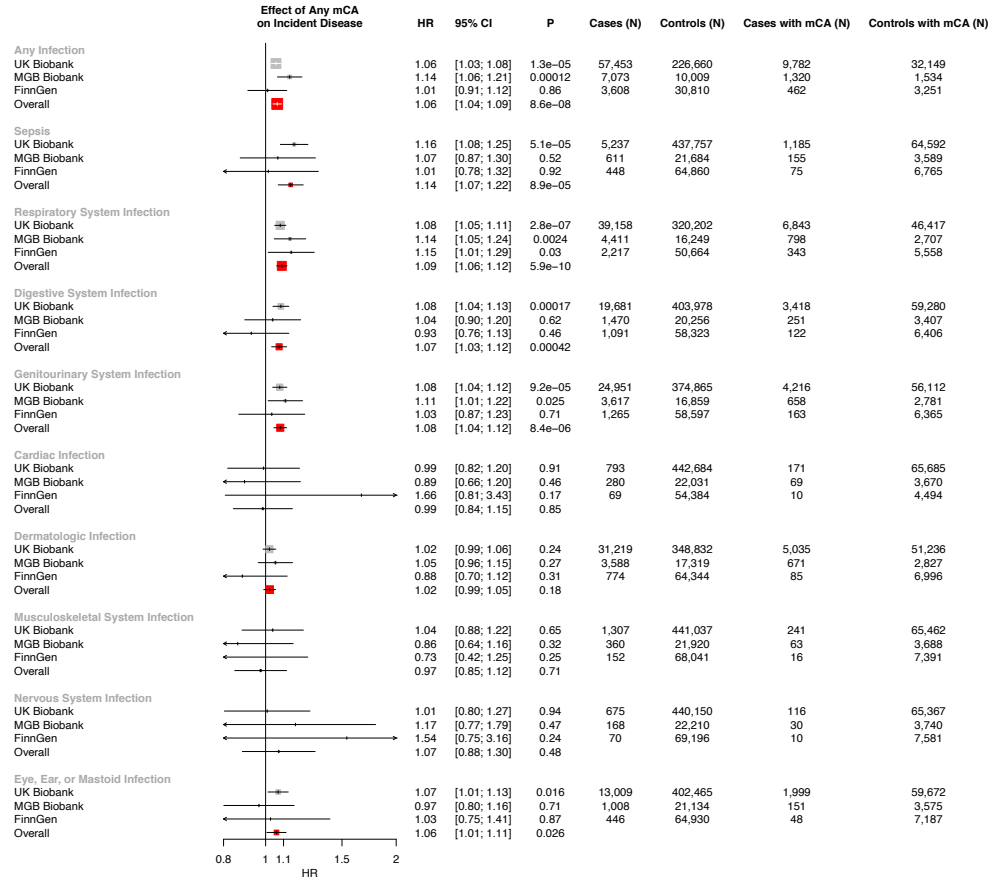


Supplementary Note 8: Prevalence of expanded mCA categories by age bin across cohorts.
mCA = mosaic chromosomal alterations.

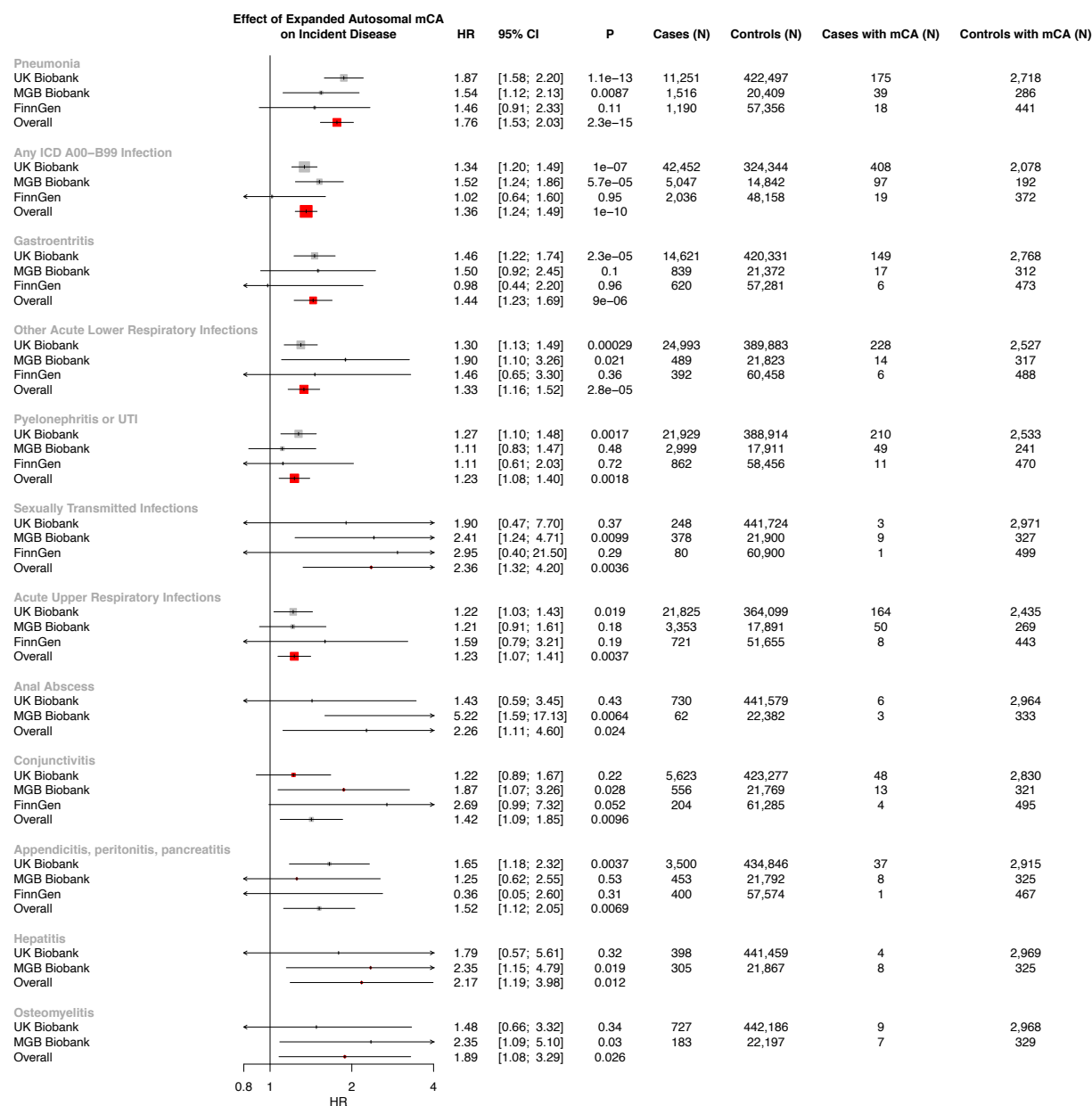


Supplementary Note 9: Proportion of expanded autosomal mCAs carriers across lymphocyte count and percentage by age bin and sex. Association of expanded autosomal mCA prevalence with A. Lymphocyte count and B. Lymphocyte percentage across bins of age and sex. mCA = mosaic chromosomal alterations.

A.

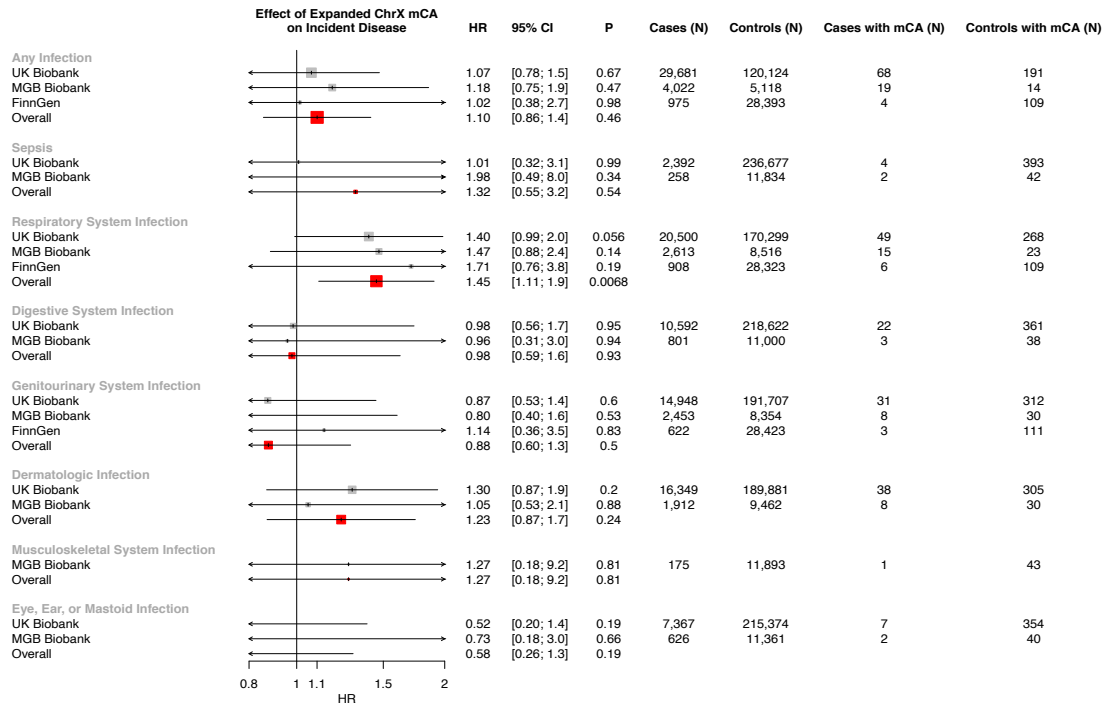


Supplementary Note 10: Associations of A) any mCA and B) any expanded mCA with incident infections. mCA = mosaic chromosomal alterations.

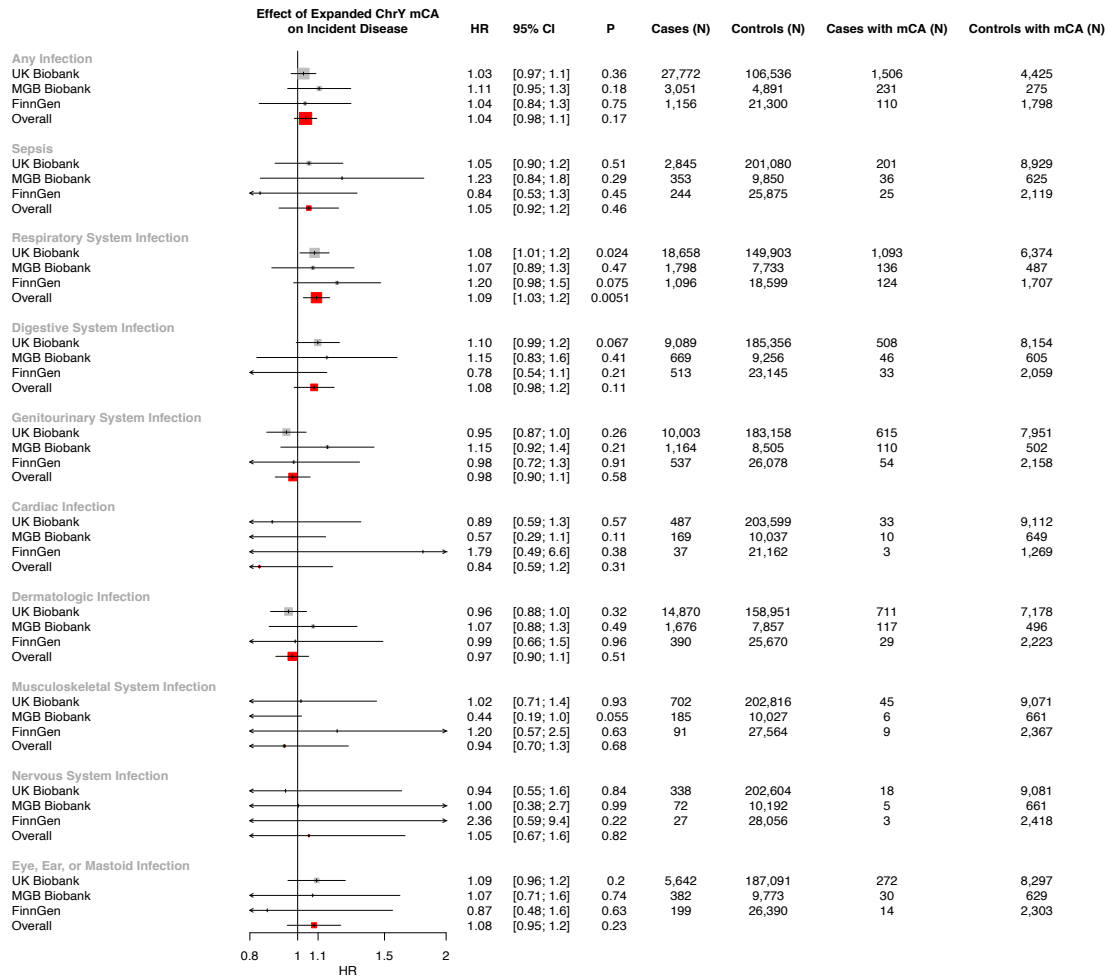


Supplementary Note 11: Suggestive associations ($P < 0.05$) of expanded autosomal mCAs with incident infection categories. mCA = mosaic chromosomal alterations.

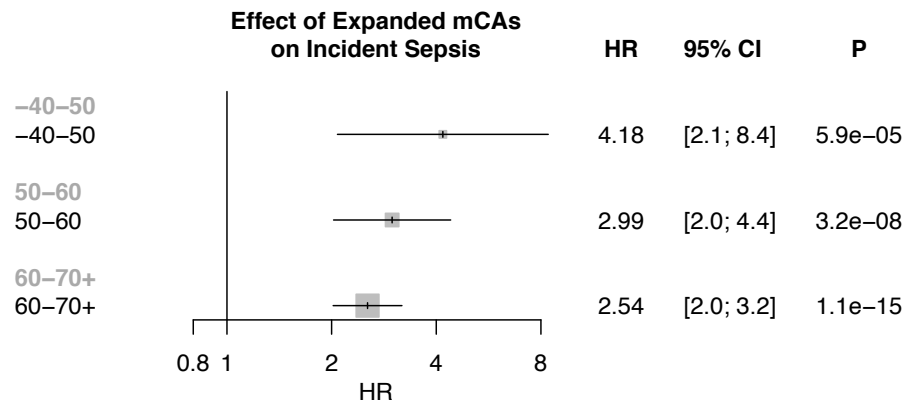
A.



B.

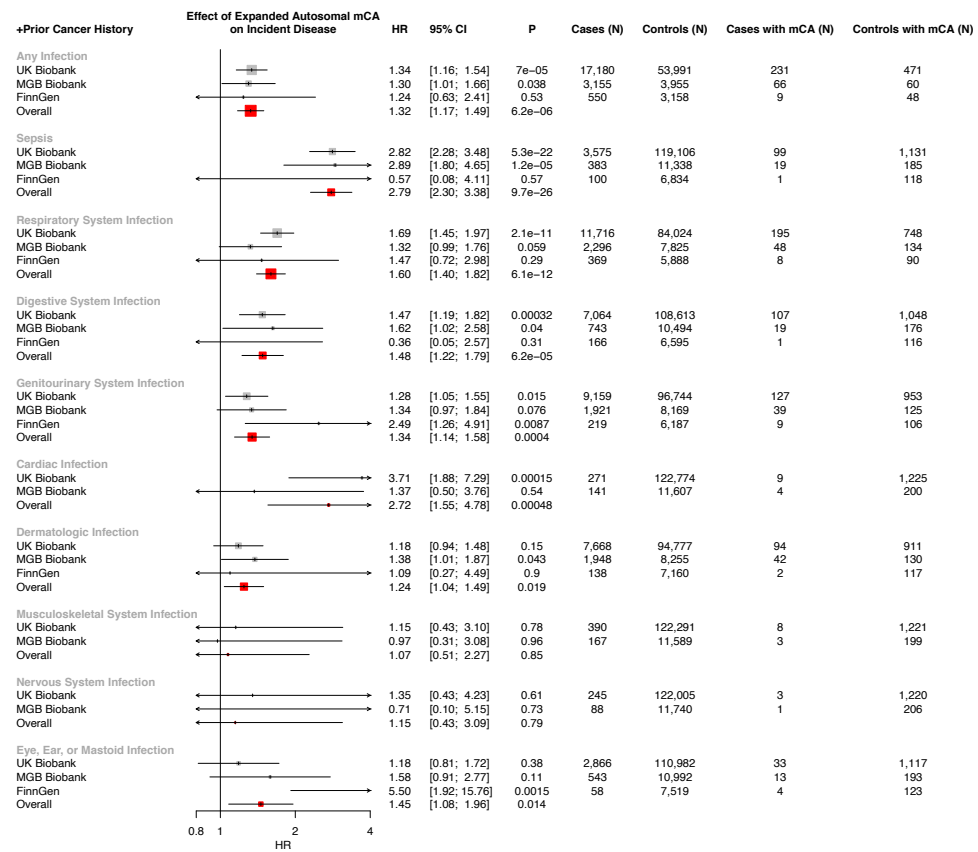


Supplementary Note 15: Associations of A) expanded ChrY and B) expanded ChrX mCAs with incident infections. mCA = mosaic chromosomal alterations.

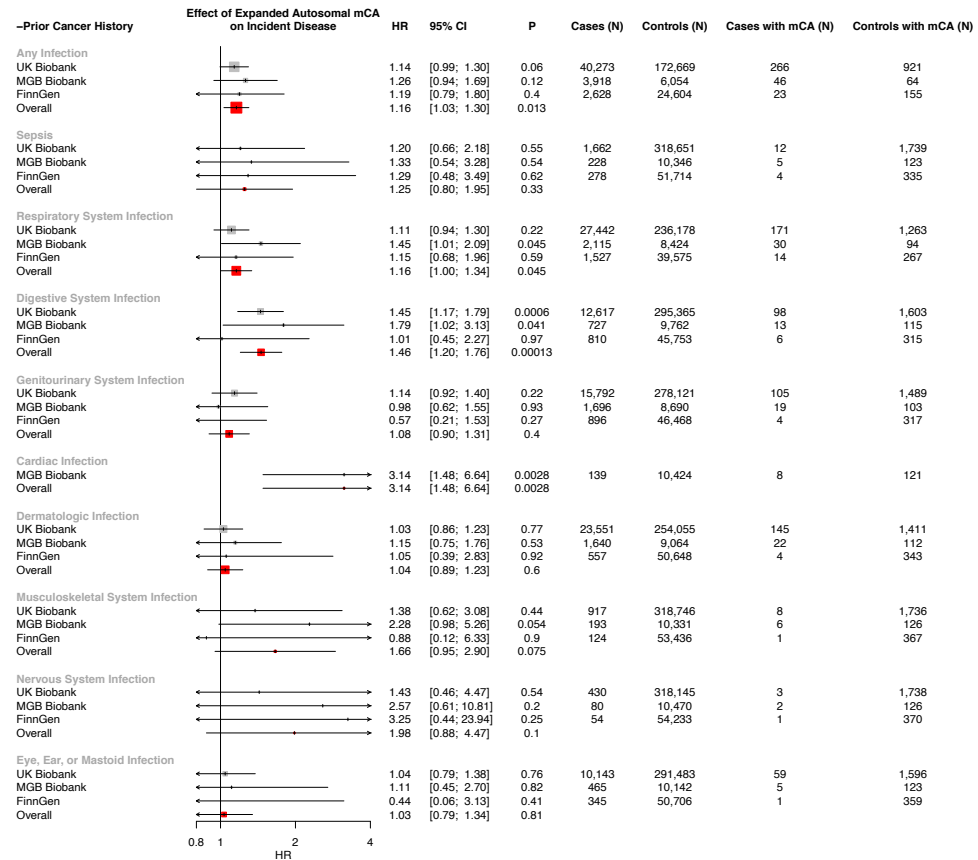


Supplementary Note 13: Associations of expanded autosomal mCAs with incident sepsis and among different age strata in the UK Biobank. Individuals with prevalent hematologic cancer were excluded from analyses. Associations were adjusted for sex, ever smoking status, and principal components 1-10 of ancestry. mCA = mosaic chromosomal alterations.

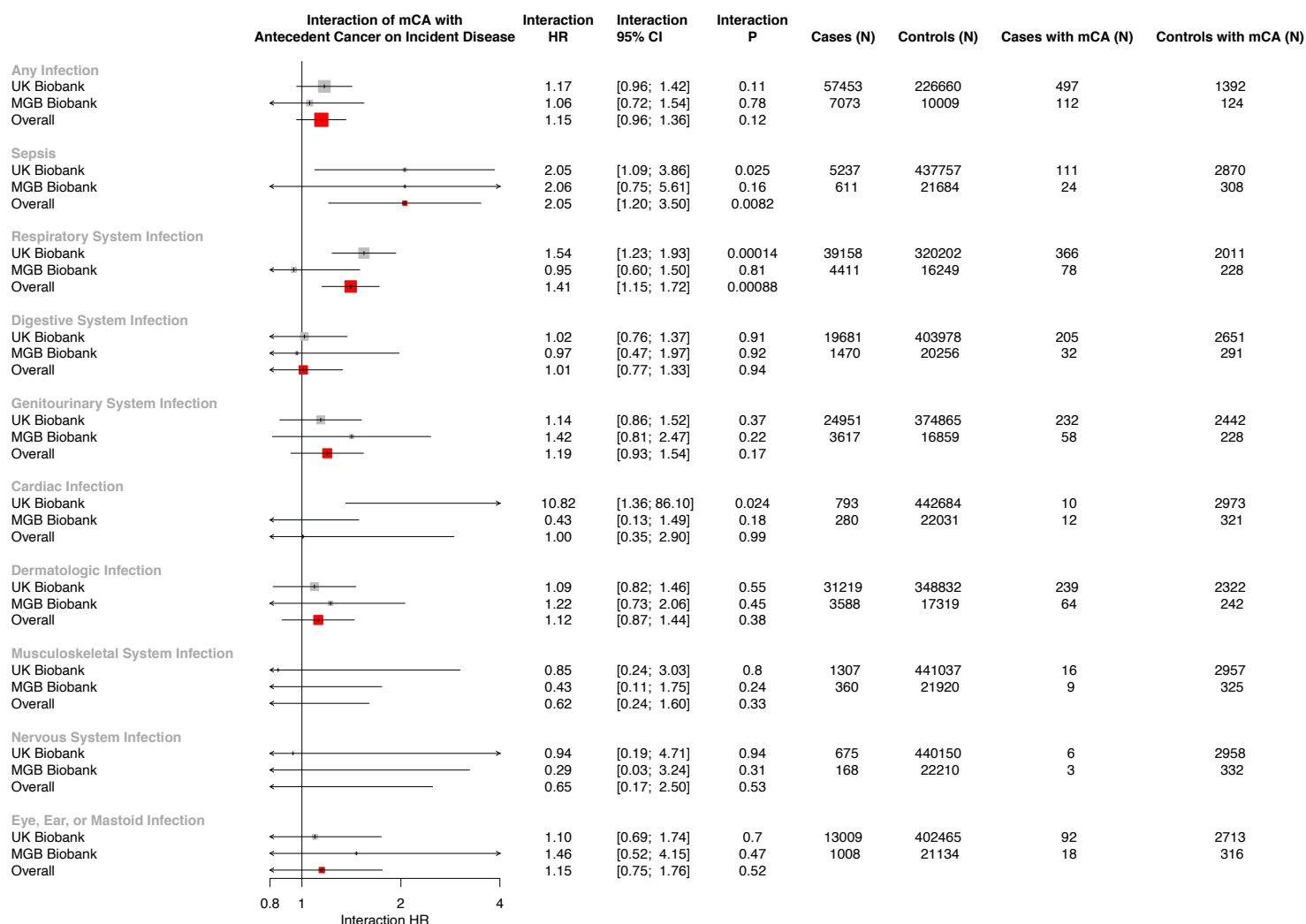
A.



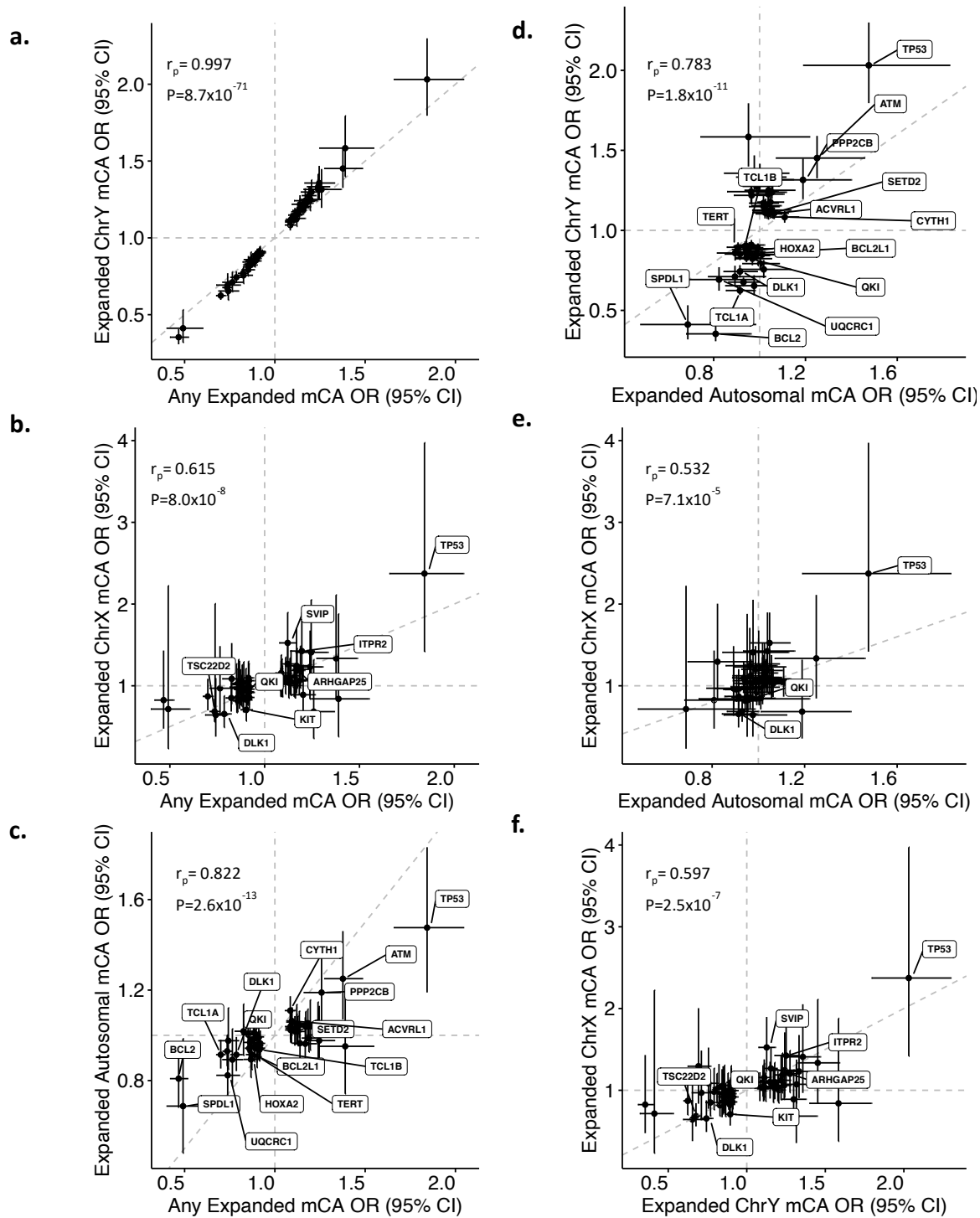
B.



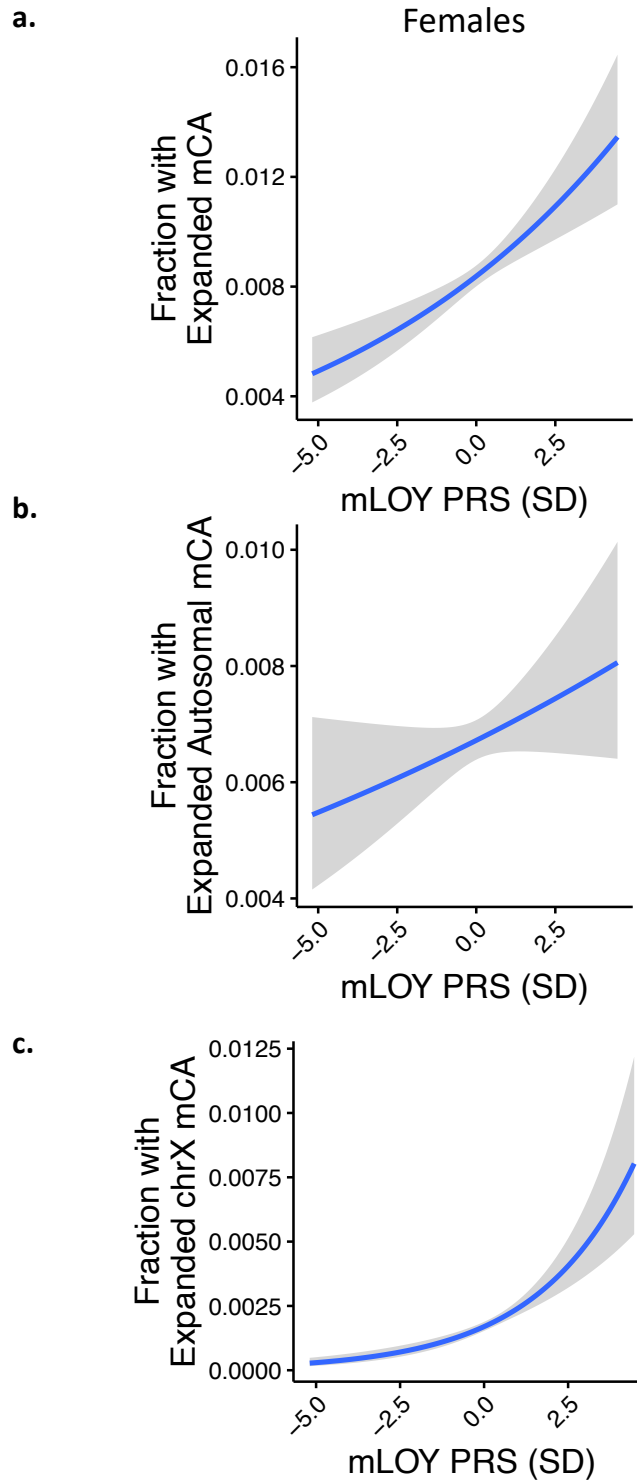
Supplementary Note 14: Associations of expanded autosomal mCAs with incident infections among A) those with antecedent cancer (ie: cancer prior to their infection) B) those without antecedent cancer. mCA = mosaic chromosomal alterations.



Supplementary Note 15: Interaction of expanded autosomal mCAs with antecedent cancer prior to infection in the UK Biobank and MGB Biobank. mCA = mosaic chromosomal alterations.



Supplementary Note 16: Correlated associations of 63 independent genome-wide significant variants associated with expanded mCAs (from Supplementary Data 3) between different mCA categories (expanded autosomal mCAs, expanded ChrX mCAs, expanded ChrY mCAs) in the UKB. Across all panels except for panel (a), the labeled genes represent genes attributed to variants that have $P < 0.05$ across the mCA categories in both axes. mCA = mosaic chromosomal alterations, r_p = Pearson correlation



Supplementary Note 17: Association of a mLOY PRS consisting of 156 previously identified¹ independent genome-wide significant variants associated with mLOY, with different expanded mCA categories in UKB Females. mCA = mosaic chromosomal alterations, mLOY = mosaic Loss-of-chromosome Y, PRS = polygenic risk score.

	<i>UK Biobank</i>	<i>MGB Biobank</i>	<i>FinnGen*</i>	<i>Biobank Japan</i>
<i>N</i>	444,199	22,461	175,690	125,541
<i>Age of DNA collection (mean (SD))</i>	56.5 (8)	55.0 (16.8)	53.4 (18.4)	64.6 (12.4)
<i>Sex (Male (%))</i>	204,579 (46.1%)	10,306 (45.9%)	71,000 (40.4)	72,186 (57.5%)
<i>Prior or Current Smoker (%)</i>	188,875 (45.0%)	9,094 (40.5%)	30,554 (42.7)	66,913 (53.3%)
<i>Race</i>	White: 417,828 (94.1%)	White: 18,933 (84.3%)	White: 175,690 (100%)	Asian: 125,541 (100%)
	Asian: 10,277 (2.3%)	Asian: 569 (2.5%)		
	Black: 7,173 (1.6%)	Black: 1,056 (4.7%)		
	Mixed: 2,634 (0.6%)	Other: 744 (3.3%)		
	Other: 4,160 (0.9%)	Unknown: 1,159 (5.2%)		
<i>BMI (mean (SD))</i>	27.4 (4.8)	28.5 (6.2)	NA	23.4 (3.7)
<i>Prevalent Solid Cancer</i>	66,551 (15.0%)	6,080 (27.1%)	31,855 (18.1%)	25,987 (20.7%)
<i>Prevalent Type 2 Diabetes Mellitus</i>	10,835 (2.4%)	1,782 (7.9%)	22,326 (13.2%)	31,636 (25.2%)
<i>Prevalent Coronary Artery Disease</i>	25,287 (5.7%)	3,908 (17.4%)	19,474 (11.1%)	23,099 (18.4%)
<i>Prevalent Hypertension</i>	129,888 (29.2%)	11,010 (49.0%)	NA	37,913 (30.2%)
<i>Prevalent Hypercholesterolemia</i>	66,483 (15.0%)	9,881 (44.0%)	8,583 (5.2%)	35,026 (27.9%)

Supplementary Table 1: Baseline summary statistics across the UK Biobank, MGB Biobank, FinnGen, and Biobank Japan among individuals analyzed.

* of note, 104,175 individuals from FinnGen had missing smoking status; BMI and hypertension were not available

BMI = body mass index; MGB = Mass General Brigham; SD = standard deviation.

	<i>UK Biobank</i>	<i>MGB Biobank</i>	<i>FinnGen</i>	<i>Biobank Japan</i>
<i>N</i>	444,199	22,461	175,690	125,541
<i>Any mCA (%)</i>	66,011 (14.9)	3,784 (16.8)	22,040 (12.5)	NA
<i>Autosomal mCA (%)</i>	15,350 (3.5)	1,025 (5.2)	3,164 (2.0)	20,440 (16.3)
<i>ChrX (%)</i>	12,265 (5.1)	820 (7.0)	7,058 (6.8)	NA
<i>ChrY (%)</i>	41,284 (20.1)	2,201 (22.0)	12,599 (18.0)	NA
<i>Any expanded mCA (%)</i>	12,398 (3.2)	1,026 (5.2)	9,558 (5.9)	NA
<i>Expanded autosomal mCA (%)</i>	2,985 (0.8)	337 (1.8)	1,620 (1.0)	1,676 (1.3%)
<i>Expanded ChrX (%)</i>	397 (0.2)	44 (0.2)	479 (0.5)	NA
<i>Expanded ChrY (%)</i>	9168 (4.5)	669 (3.4)	7663 (11.8)	NA

Supplementary Table 2: mCA counts by cohort.

Chr = chromosome; mCA = mosaic chromosomal alteration; MGB = Mass General Brigham; N = number.

	<i>OR</i>	<i>Lower 95% CI</i>	<i>Upper 95% CI</i>	<i>P</i>
<i>Age</i>	0.92	0.86	1.005	0.068
<i>Age²</i>	1.0013	1.0006	1.002	2.84E-05
<i>Sex (Male)</i>	1.44	1.33	1.56	1.33E-19
<i>Prior or Current Smoking</i>	1.09	1.01	1.18	0.03
<i>Prevalent Solid Cancer</i>	0.95	0.82	1.09	0.44
<i>Prevalent Type 2 Diabetes Mellitus</i>	1.05	0.84	1.32	0.66

Supplementary Table 3: Association of potential risk factors with expanded autosomal mCAs in the UK Biobank.

CI = confidence interval; OR = odds ratio

<i>Phenotype</i>	<i>Exposure</i>	<i>Sex</i>	<i>HR</i>	<i>P</i>	<i>Controls (N)</i>	<i>Incident Cases (N)</i>	<i>Incident Cases with mCA (N)</i>	<i>Controls with mCA (N)</i>
<i>Nervous System Infection</i>	Autosomal mCA	Male	3.71	0.0085	39,978	17	8	6,298
<i>Meningitis or Encephalitis</i>		Male	3.71	0.0085	39,978	17	8	6,298
<i>Pneumonia</i>		All	1.16	0.023	72,317	1,299	334	10,485
<i>Nervous System Infection</i>		All	2.81	0.025	72,317	21	8	10,485
<i>Meningitis or Encephalitis</i>		All	2.81	0.025	72,317	21	8	10,485
<i>Respiratory System Infection</i>		All	1.15	0.028	72,317	1,353	346	10,485
<i>Any Infection</i>		All	1.12	0.037	72,317	1,998	476	10,485
<i>Pneumonia</i>		Male	1.17	0.039	39,978	928	259	6,298
<i>Endocarditis or Myocarditis</i>		Female	2.46	0.043	32,339	25	8	4,187
<i>Respiratory System Infection</i>		Male	1.16	0.050	39,978	971	269	6,298
<i>Sepsis</i>	Expanded	All	2.04	0.050	72,317	276	8	753
<i>Sepsis</i>	Autosomal mCA	All	2.04	0.050	72,317	276	8	753

Supplementary Table 4: Association of mCAs with mortality from incident infection in Biobank Japan. Suggestive associations ($P < 0.05$) are presented among individuals without antecedent cancer prior to the infection phenotype.

HR = hazard ratio; mCA = mosaic chromosomal alteration; N = number

<i>Incident Sepsis Multivariate Model Component</i>	<i>HR</i>	<i>P</i>	<i>Lower 95% CI</i>	<i>Upper 95% CI</i>
<i>Expanded Autosomal mCA</i>	2.00	2.64E-08	1.56	2.54
<i>Antecedent Cancer Prior to Incident Sepsis</i>	5.49	<1E-300	5.09	5.93
<i>Age</i>	1.02	0.64	0.95	1.09
<i>Age²</i>	1.00	0.41	1.00	1.00
<i>Sex (Male)</i>	1.27	9.56E-11	1.18	1.37
<i>current cigar pipe smoker, former cigarette smoker</i>	1.90	0.011	1.16	3.12
<i>current cigar pipe smoker, not former cigarette smoker</i>	1.81	0.077	0.94	3.49
<i>current cigarette smoker, 10 to <20/day</i>	1.83	1.79E-10	1.52	2.20
<i>current cigarette smoker, 20 to <40/day</i>	2.06	8.23E-14	1.70	2.49
<i>current cigarette smoker, <10/day</i>	1.39	0.047	1.00	1.92
<i>current cigarette smoker, ≥40/day</i>	3.33	3.65E-05	1.88	5.89
<i>current occasional smoker, smoked cigarettes daily in past, <20/day</i>	1.37	0.19	0.86	2.18
<i>current occasional smoker, smoked cigarettes daily in past, ≥20/day</i>	1.83	0.01	1.13	2.96
<i>current occasional smoker, smoked cigars or pipes daily in past</i>	0.51	0.50	0.07	3.60
<i>current occasional smoker, smoked ≥100 cigarettes in lifetime</i>	0.99	0.94	0.70	1.39
<i>former cigarette smoker, <20/day, quit 1-5 year ago</i>	1.14	0.48	0.80	1.63
<i>former cigarette smoker, <20/day, quit 10-20 year ago</i>	1.15	0.27	0.90	1.46
<i>former cigarette smoker, <20/day, quit 5-10 year ago</i>	1.29	0.11	0.94	1.78
<i>former cigarette smoker, <20/day, quit <1 year ago</i>	1.90	0.06	0.98	3.65
<i>former cigarette smoker, <20/day, quit ≥20 year ago</i>	1.08	0.34	0.92	1.26
<i>former cigarette smoker, ≥20/day, quit 1-5 year ago</i>	2.25	1.07E-13	1.82	2.79
<i>former cigarette smoker, ≥20/day, quit 10-20 year ago</i>	1.54	7.00E-07	1.30	1.82
<i>former cigarette smoker, ≥20/day, quit 5-10 year ago</i>	1.64	1.11E-05	1.31	2.04
<i>former cigarette smoker, ≥20/day, quit <1 year ago</i>	2.08	0.0064	1.23	3.53
<i>former cigarette smoker, ≥20/day, quit ≥20 year ago</i>	1.25	0.00080	1.10	1.42
<i>former daily cigar pipe smoker</i>	1.23	0.19	0.91	1.66
<i>former occasional cigarette smoker, lifetime cigarette smoking unknown</i>	1.07	0.76	0.70	1.63
<i>former occasional cigarette smoker, smoked <100 cigarettes in lifetime</i>	1.27	0.085	0.97	1.66
<i>former occasional cigarette smoker, smoked ≥100 cigarettes in lifetime</i>	1.02	0.72	0.90	1.17
<i>Missing smoking</i>	1.55	0.0085	1.12	2.15
<i>BMI (SD)</i>	1.21	7.08E-31	1.17	1.25
<i>Prevalent Type 2 Diabetes Mellitus</i>	1.84	1.75E-16	1.59	2.12
<i>Leukocyte count (SD)</i>	1.31	1.38E-07	1.19	1.45
<i>Lymphocyte count (SD)</i>	0.71	4.57E-07	0.63	0.81
<i>Lymphocyte percentage (SD)</i>	1.12	0.054	1.00	1.27
<i>PC1</i>	0.98	0.18	0.96	1.01
<i>PC2</i>	0.99	0.42	0.97	1.01
<i>PC3</i>	0.99	0.59	0.97	1.02
<i>PC4</i>	0.99	0.37	0.98	1.01
<i>PC5</i>	1.01	0.07	1.00	1.01
<i>PC6</i>	1.01	0.46	0.99	1.03
<i>PC7</i>	0.99	0.23	0.97	1.01
<i>PC8</i>	1.01	0.58	0.99	1.03
<i>PC9</i>	1.00	0.58	0.99	1.01
<i>PC10</i>	1.00	0.61	0.99	1.02

Supplementary Table 5: Sensitivity analysis of incident sepsis association in the UK Biobank with the addition of a 25-factor smoking covariate, leukocyte count, lymphocyte count and percentage, BMI, and prevalent type 2 diabetes mellitus.

BMI = body mass index; CI = confidence interval; HR = hazard ratio; PC = principal component; SD = standard deviation

<i>Outcome</i>	<i>Population of people with cancer prior to infection</i>	<i>HR</i>	<i>P</i>	<i>Lower 95% CI</i>	<i>Upper 95% CI</i>	<i>Cases (N)</i>	<i>Controls (N)</i>	<i>Cases with mCA (N)</i>	<i>Controls with mCA (N)</i>
<i>Sepsis</i>	Prevalent Solid Cancer	1.86	0.0097	1.16	2.96	1258	64921	20	521
	Incident Solid Cancer Prior to Infection	0.76	0.44	0.38	1.53	1619	51867	10	361
	Incident Hematologic Cancer Prior to Infection	0.98	0.88	0.77	1.25	833	2864	83	312
	Incident Hematologic Cancer and Prevalent Solid Cancer Prior to Infection	0.94	0.85	0.52	1.72	144	546	15	63
	Any Cancer Prior to Infection	2.82	5.28E-22	2.28	3.48	3575	119106	99	1131
<i>Pneumonia</i>	Prevalent Solid Cancer	1.68	0.0057	1.16	2.43	2382	62325	40	480
	Incident Solid Cancer Prior to Infection	1.33	0.18	0.87	2.03	2369	49466	24	323
	Incident Hematologic Cancer Prior to Infection	1.19	0.18	0.92	1.54	655	2886	80	300
	Incident Hematologic Cancer and Prevalent Solid Cancer Prior to Infection	1.73	0.076	0.94	3.18	119	528	16	55
	Any Cancer Prior to Infection	2.26	5.08E-17	1.86	2.73	5295	114149	130	1048

Supplementary Table 6: Sensitivity analysis of incident sepsis and pneumonia association in the UK Biobank among populations of individuals with different types of cancer prior to incident infection, where solid cancer is defined as any non-hematologic cancer. Other covariates in the model included age, age², sex, smoking status, and PC1-10 of ancestry.

CI = confidence interval; HR = hazard ratio; mCA = mosaic chromosomal alteration; N = number

References:

- 1 Thompson, D. J. *et al.* Genetic predisposition to mosaic Y chromosome loss in blood. *Nature* **575**, 652-657, doi:10.1038/s41586-019-1765-3 (2019).
- 2 Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203-209, doi:10.1038/s41586-018-0579-z (2018).
- 3 Smoller, J. W. *et al.* An eMERGE Clinical Center at Partners Personalized Medicine. *J Pers Med* **6**, doi:10.3390/jpm6010005 (2016).
- 4 Nagai, A. *et al.* Overview of the BioBank Japan Project: Study design and profile. *J Epidemiol* **27**, S2-S8, doi:10.1016/j.je.2016.12.005 (2017).
- 5 Loh, P. R. *et al.* Insights into clonal haematopoiesis from 8,342 mosaic chromosomal alterations. *Nature* **559**, 350-355, doi:10.1038/s41586-018-0321-x (2018).
- 6 Loh, P. R., Genovese, G. & McCarroll, S. A. Monogenic and polygenic inheritance become instruments for clonal selection. *Nature*, doi:10.1038/s41586-020-2430-6 (2020).
- 7 Loh, P. R. *et al.* Reference-based phasing using the Haplotype Reference Consortium panel. *Nat Genet* **48**, 1443-1448, doi:10.1038/ng.3679 (2016).
- 8 Voss, K., Auwera, G. V. D. & Gentry, J. Full-stack genomics pipelining with GATK4 + WDL + Cromwell. . doi:doi:10.7490/F1000RESEARCH.1114634.1 (2017).
- 9 Terao, C. *et al.* Chromosomal alterations among age-related haematopoietic clones in Japan. *Nature*, doi:10.1038/s41586-020-2426-2 (2020).
- 10 Loftfield, E. *et al.* Predictors of mosaic chromosome Y loss and associations with mortality in the UK Biobank. *Sci Rep* **8**, 12316, doi:10.1038/s41598-018-30759-1 (2018).
- 11 Wu, P. *et al.* Mapping ICD-10 and ICD-10-CM Codes to Phecodes: Workflow Development and Initial Evaluation. *JMIR Med Inform* **7**, e14325, doi:10.2196/14325 (2019).