Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

HHS Public Access

Author manuscript

*Nat Med*. Author manuscript; available in PMC 2021 December 07.

Published in final edited form as:

*Nat Med*. 2021 June ; 27(6): 1012–1024. doi:10.1038/s41591-021-01371-0.

# Hematopoietic mosaic chromosomal alterations increase the risk for diverse types of infection

**Seyedeh M. Zekavat, BS**[#1,2,3], **Shu-Hong Lin, PhD**[#4], **Alexander G. Bick, MD PhD**[5,2], **Aoxing Liu, PhD**[6], **Kaavya Paruchuri, MD**[2,3,7], **Chen Wang, PhD**[8,9], **Md Mesbah Uddin, PhD**[2,3], **Yixuan Ye, BS**[1], **Zhaolong Yu, BS**[1], **Xiaoxi Liu, PhD**[10], **Yoichiro Kamatani, PhD**[10], **Romit Bhattacharya, MD**[3], **James P. Pirruccello, MD**[2,3,7], **Akhil Pampana, MS**[2,3], **Po-Ru Loh, PhD**[2,7], **Puja Kohli, MD MMSc**[11,12], **Steven A. McCarroll, PhD**[13,14], **Krzysztof Kiryluk, MD MS**[9,15], **Benjamin Neale, PhD**[13,16], **Iuliana Ionita-Laza, PhD**[8], **Eric A. Engels, MD MPH**[17], **Derek W. Brown, PhD**[4], **Jordan W. Smoller, MD ScD**[13,18,19], **Robert Green, MD MPH**[2,7,14], **Elizabeth W. Karlson, MD MS**[7,20], **Matthew Lebo, PhD**[21,22], **Patrick T. Ellinor, MD PhD**[2,3,7], **Scott T. Weiss, MD MS**[7,23], **Mark J. Daly**[6], **The Biobank Japan Project**[*], **FinnGen Consortium**[*], **Chikashi Terao, MD PhD**[10,24,25], **Hongyu Zhao, PhD**[1,26], **Benjamin L. Ebert, MD PhD**[2,27,28], **Muredach P Reilly, MB MSCE**[15,29], **Andrea Ganna, PhD**[6,16,2], **Mitchell J. Machiela, ScD MPH**[#,4], **Giulio Genovese, PhD**[#,2,13,14], **Pradeep Natarajan, MD MMSc**[#,2,3,7]

[1.]Computational Biology & Bioinformatics Program, Yale University, New Haven, CT, USA [2.]Medical and Population Genetics and Cardiovascular Disease Initiative, Broad Institute of Harvard and MIT, Cambridge, MA, USA [3.]Cardiovascular Research Center, Massachusetts General Hospital, Boston, MA, USA [4.]Integrative Tumor Epidemiology Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, Rockville, MD, USA [5.]Division of Genetic

Please address correspondence to: Pradeep Natarajan MD, MMSc, Massachusetts General Hospital, 185 Cambridge Street, CPZN 3.184, Boston, MA 02114, Office: 617-726-1843, pnatarajan@mgh.harvard.edu, Twitter: @pnatarajanmd.

[#]These authors equally supervised this work.

[*]Lists of authors and their affiliations appear at the end of the paper.

Author contribution:

S.M.Z., S.-H.L., C.W., M.J.M., P.N. performed statistical modeling of UKB, FinnGen, and MGB. C.W. collected and analyzed CUB data. S.M.Z. carried out GWAS and TWAS analyses. P.-R.L. and G.G. called mCAs. M.J.M. and P.N. supervised the study. S.M.Z. and S.-H.L. drafted the manuscript. All authors critically reviewed the manuscript.

Data Availability:

UKB individual-level data are available for request by application (https://www.ukbiobank.ac.uk). The mCA call set was previously returned to the UK Biobank (Return 2062) to enable individual-level linkage to approved UK Biobank applications. Individual-level MGBB data are available from https://personalizedmedicine.partners.org/Biobank/Default.aspx, but restrictions apply to the availability of these data, which were used under IRB approval for the current study, and so are not publicly available. The BBJ genotype data is available from the Japanese Genotype-phenotype Archive (JGA; http://trace.ddbj.nig.ac.jp/jga/index_e.html) under accession code JGAD00000000123. Individual-level linkage of mosaic events can be provided by the BBJ project upon request (https://biobankjp.org/english/index.html). FinnGen data may be accessed through Finnish Biobanks' FinnBB portal (www.finbb.fi). Individual-level CUB COVID-19 data, including mCA call set, are available by application from https://www.ps.columbia.edu/research/core-and-shared-facilities/core-facilities-category/columbia-university-biobank, but consent-related restrictions apply to the availability of these data, and data access requires separate IRB approval for the proposed data use. Aggregate data is also available upon reasonable request. Additionally, the full expanded mCA genome wide association summary statistics have been uploaded onto the LocusZoom website (https://my.locuszoom.org/gwas/525823/). The present article includes all other data generated or analyzed during this study.

Code Availability:

A standalone software implementation (MoChA) of the algorithm used to call mCAs is available at 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 engine as part of MoChA. Code for all other computations are available upon request to the corresponding authors.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Medicine, Department of Medicine, Vanderbilt University Medical Center, Nashville, TN, USA [6]Institute for Molecular Medicine Finland, Helsinki, Finland [7]Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA [8]Department of Biostatistics, Mailman School of Public Health, Columbia University, New York City, NY, USA [9]Division of Nephrology, Department of Medicine, Vagelos College of Physicians and Surgeons, Columbia University, New York City, NY, USA [10]Laboratory for Statistical and Translational Genetics, RIKEN Center for Integrative Medical Sciences, RIKEN, Yokohama, Japan [11]Division of Pulmonary and Critical Care Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA, USA [12]Vertex Pharmaceuticals, Boston, MA, USA [13]Stanley Center, Broad Institute of Harvard and MIT, Cambridge, MA, USA [14]Department of Genetics, Harvard Medical School, Boston, MA, USA [15]Irving Institute for Clinical and Translational Research, Columbia University, New York City, NY, USA [16]Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA, USA [17]Infections and Immunoepidemiology Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, Rockville, MD, USA [18]Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA, USA [19]Department of Psychiatry, Harvard Medical School, Boston, MA, USA [20]Division of Rheumatology, Inflammation and Immunity, Brigham and Women's Hospital, Boston, MA, USA [21]Department of Pathology, Brigham and Women's Hospital and Harvard Medical School, Boston, MA, USA [22]Laboratory for Molecular Medicine, Partners Healthcare, Cambridge, MA, USA [23]Channing Division of Network Medicine, Brigham and Women's Hospital, Boston, MA, USA [24]Clinical Research Center, Shizuoka General Hospital, Shizuoka, Japan [25]The Department of Applied Genetics, The School of Pharmaceutical Sciences, University of Shizuoka, Shizuoka, Japan [26]Department of Biostatistics, Yale School of Public Health, New Haven, CT, USA [27]Department of Medical Oncology, Dana-Farber Cancer Institute, Boston, MA, USA [28]Howard Hughes Medical Institute, Boston, MA, USA [29]Division of Cardiology, Department of Medicine, Vagelos College of Physicians and Surgeons, Columbia University, New York City, NY, USA

[#] These authors contributed equally to this work.

## Abstract

Age is the dominant risk factor for infectious diseases, but the mechanisms linking age to infectious disease risk are incompletely understood. Age-related mosaic chromosomal alterations (mCAs) detected from genotyping of blood-derived DNA, are structural somatic variants indicative of clonal hematopoiesis and are associated with aberrant leukocyte cell counts, hematological malignancy, and mortality. Here we show that mCAs predispose to diverse types of infections. We analyzed mCAs from 768,762 individuals without hematological cancer at the time of DNA acquisition across five biobanks. Expanded autosomal mCAs were associated with diverse incident infections (HR 1.25; 95% CI 1.15–1.36; $P=1.8\times10^{-7}$) including sepsis (HR 2.68; 95% CI 2.25–3.19; $P=3.1\times10^{-28}$), pneumonia (HR 1.76; 95% CI 1.53–2.03; $P=2.3\times10^{-15}$), digestive system infections (HR 1.51; 95% CI 1.32–1.73; $P=2.2–10^{-9}$), and genitourinary infections (HR 1.25; 95% CI 1.11–1.41; $P=3.7\times10^{-4}$). A genome-wide association study of expanded mCAs identified 63 loci, enriched at transcriptional regulatory sites for immune cells. Our results suggest that mCAs are a marker of impaired immunity and confer increased predisposition to infections.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

## Introduction:

With advancing age comes increased susceptibility to infectious diseases[1,2]. Immunosenescence is the age-related erosion of immune function, particularly with respect to adaptive immunity[3–6]. Leukocytes, including T-cells and B-cells, are key mediators of adaptive host defenses against infections, with impaired immune responses increasing risk for infections[7–9]. Age-related mosaic chromosomal alterations (mCAs) detected from blood-derived DNA, are clonal structural somatic alterations (deletions, duplications, or copy neutral loss of heterozygosity) present in a fraction of peripheral leukocytes that can indicate clonal hematopoiesis (CH)[10–12]. mCAs are associated with aberrant leukocyte cell counts, and increased risks for hematological malignancy and mortality[10–18].

While the relationship between mCAs and increased hematologic cancer risk is well established[10–12], the impact of mCAs on age-related diminishment in immune function is poorly understood. We hypothesized that mCAs increase risk of infection since mCAs are somatic variants that increase in abundance with age and are associated with alterations in leukocyte count. In this study, we harnessed DNA genotyping array intensity data and long-range chromosomal phase information inferred from 768,762 individuals across five biobanks to analyze the associations between expanded mCA clones (i.e., mCAs present in at least 10% of peripheral leukocyte DNA indicative of clonal expansion) and diverse infections, including severe coronavirus disease 2019 (COVID-19) from SARS-CoV-2 infection (Figure 1a). To elucidate genetic risk factors for the development of expanded mCA clones, we performed a genome-wide association study (GWAS) in the UK Biobank and subsequent *in silico* cell-specific, transcriptomic, and pathway analyses.

## Results:

### Population characteristics and mCA prevalence

A total of 768,762 unrelated, multi-ethnic individuals across the UK Biobank (UKB) (N=444,199), Mass General Brigham Biobank (MGBB) (N=22,461), FinnGen (N=175,690), BioBank Japan (BBJ) (N=125,541), and Columbia University Biobank (CUB) (N=871) passing genotype and mCA quality control criteria (Supplementary Figures 1–7) were analyzed (Supplementary Table 1). While UKB and BBJ mCA calls were previously performed[10,11], the MoChA pipeline (https://github.com/freeseek/mocha) was implemented to detect mCAs in MGBB, FinnGen, and CUB (Extended Data Figure 1) from genome-wide genotyping of blood DNA in the present study. Among the UKB participants, mean age at DNA collection was 57 (standard deviation [SD] 8) years, 204,579 (46.1%) were male, 188,875 (45.0%) were prior or current smokers, and 66,551 (15.0%) had a history of solid cancer. In the MGBB, mean age was 55 (SD 17) years, 10,306 (45.9%) were male, 9,094 (40.5%) were prior or current smokers, and 6,080 (27.1%) had a history of solid cancer. In FinnGen, mean age was 53 (SD 18) years, 71,000 (40.4%) were male, 42.7% were prior or current smokers (when smoking status was available), and 31,855 (18.1%) had a history of solid cancer. In BBJ, mean age was 65 (SD 12) years, 72,186 (57.5%) were male, and 66,913 (53.3%) were prior or current smokers, and 25,987 (20.7%) had a history of solid cancer. In CUB, mean age was 62.3 (SD 17.9) years, 480 (55.1%) were male, and 221 (25.4%) had a history of solid cancer (Supplementary Table 1).

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

In the UKB, among 444,199 unrelated individuals without a known history of hematologic malignancy, 66,011 (14.9%) carried an mCA (15,350 autosomal) and 12,398 (3.2%) carried an expanded mCA clone, defined as an mCA mutation present in at least 10% of peripheral leukocytes (2,985 autosomal) (Supplementary Table 2). While most of carriers only carried one mCA, 6% of individuals carried between 2 to 22 non-overlapping mCAs (Supplementary Figure 7). In the MGBB, across 22,461 unrelated individuals without a history of hematologic cancer, 3,784 (16.8%) carried an mCA (1,025 autosomal) and 1,026 (5.2%) carried an expanded mCA clone (337 autosomal). In FinnGen, across 175,690 individuals without a history of hematologic cancer, 22,040 (12.5%) carried an mCA (3,164 autosomal), and 9,558 (5.9%) carried an expanded mCA clone (1,620 autosomal). In BBJ, across 125,541 individuals without a history of hematologic cancer, only autosomal mCAs were available, with 20,440 carriers (16.3%) and 1,676 (1.3%) that carried an expanded clone. In CUB COVID-19 cohort, across 871 individuals without a history of hematologic cancer, 258 (29.6%) carried an mCA (168 autosomal), and 177 (20.3%) carried an expanded mCA clone (128 autosomal) (Supplementary Table 2).

Consistent with previous reports, the prevalence of mCAs increased with age and was more common among men (Supplementary Figure 8,9, and Supplementary Table 3). Across the UKB, MGBB, FinnGen, and BBJ cohorts combined, the prevalence of expanded mCAs was 0.5% among individuals <40 years, 1.2% among 40–60 years, 7.8% among 60–80 years, and 26.5% among those greater than 80 years (Figure 1b), the majority of which is due to loss of X in females and loss of Y in males (Supplementary Figure 8). The prevalence of expanded autosomal mCAs was 0.27% among individuals <40 years, 0.52% among 40–60 years, 1.5% among 60–80 years, and 4.6% among those greater than 80 years (Figure 1c).

## Association of mCAs with hematologic traits

We observed a striking association of mCA cell fraction with aberrant cell blood counts acquired at the same visit as blood for genotyping (Figure 2a,b). Increased mCA cell fraction was associated with overall increased white blood cell count with general consistency across the cell differential components, with inflections at around cell fraction of 0.1 (Figure 2b). The strongest association across all mCAs groupings (autosomal/chrX/chrY) with blood counts was between expanded autosomal mCAs and increased lymphocyte count at enrollment (Beta 0.40 SD or $0.25\times10^{9}$ cells/L; 95% CI 0.36 to 0.44 SD; $P=4.2\times10^{-84}$) (Figure 2a, Supplementary Figure 10).

Similarly, incident hematologic cancer risk was also strongly dependent on cell fraction (Figure 2c). We reproduced the associations of mCAs with hematologic cancers with similar effects as previously described in the UKB[11,12]. We found that expanded autosomal mCAs with cell fraction >10% were most strongly associated with incident hematologic cancer (Figure 2d), with the strongest association being for incident chronic lymphocytic leukemia (HR 120.48; 95% CI 92.53 to 156.86; $P=2.2\times10^{-277}$); although an association with polycythemia vera (HR 32.56; 95% CI 22.81 to 46.48; $P=6.0\times10^{-82}$) and myeloid leukemia was also present (HR 11.82; 95% CI 7.29 to 19.18; $P=1.4\times10^{-23}$) (Figure 2d). In comparison, the associations of chrX and chrY mCAs with chronic lymphocytic leukemia

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

were considerably weaker (chrX: HR 27.40, 95% CI 6.58 to 114.16, $P=5.5\times10^{-6}$ and chrY: HR 1.91, 95% CI 0.96 to 3.80, P=0.064) (Figure 2d).

## Associations with diverse infections

mCA presence across the genome was associated with diverse incident infections (defined in Supplementary Tables 4, 5) (HR 1.06; 95% CI 1.04 to 1.09; $P=8.6\times10^{-8}$) (Supplementary Figure 11), independent of age, age$^2$, sex, smoking status, and first 10 principal components of ancestry in the combined UKB, MGBB, and FinnGen meta-analysis. The dependence of this association with mCA cell fraction is further visualized in Figure 3a,b, which shows an increase in proportion of incident infection cases and incident sepsis cases with cell fraction, with greater slopes observed at approximately cell fraction >10%. Accordingly, the associations across diverse infections were stronger for expanded mCA clones, (HR 1.12; 95% CI 1.07 to 1.17; $P=6.3\times10^{-7}$) (Figure 3c). Furthermore, among expanded mCA clones, the strongest association was observed among expanded autosomal mCAs (HR 1.25; 95% CI 1.15 to 1.36; $P=1.8\times10^{-7}$) (Figure 3c). Accounting for multiple hypothesis testing, expanded autosomal mCAs were significantly associated with sepsis (HR 2.68; 95% CI 2.25 to 3.19; $P=3.1\times10^{-28}$), respiratory system infections (HR 1.36; 95% CI 1.24 to 1.50; $P=3.8\times10^{-10}$), digestive system infections (HR 1.51; 95% CI 1.32 to 1.73; $P=2.2\times10^{-9}$), and genitourinary system infections (HR 1.25; 95% CI 1.11 to 1.41; $P=3.7\times10^{-4}$) (Figure 3c). The specific expanded autosomal mCAs implicated for infection were diverse in nature – across all chromosomes, of different sizes, and mixed across gain, loss, and copy-number neutral loss of heterozygosity (CNN-LOH) mCAs (Extended Data Figure 2). Further associations across 20 specific infectious disease subcategories are enumerated in Extended Data Figure 3. For sex chromosome mCAs, none of the incident infections achieved statistical significance (P<0.005) in meta-analysis across the three cohorts; however, respiratory infections were suggestively associated (expanded chrX: HR 1.45; 95% CI 1.11 to 1.90; P=0.0068; expanded chrY: HR 1.09; 95% CI 1.03 to 1.16; P=0.005) (Extended Data Figure 4).

Risks for incident fatal infections were assessed in BBJ since non-fatal incident infectious disease events are currently unavailable in BBJ. Among individuals without any cancer history in BBJ, autosomal mCAs showed nominal associations with fatal incident infections (HR 1.12, 95% CI 1.0 to 1.2 P=0.04), with expanded autosomal mCAs being associated with incident sepsis mortality (HR 2.04; 95% CI 1.04 to 4.16; P=0.05) (Supplementary Table 6, Extended Data Figure 5), as well as pneumonia history (OR 1.40; 95% CI: 1.12 to 1.53; P=0.00080).

Sensitivity analysis for the association of expanded autosomal mCAs and incident sepsis found that the association was consistently significant across different age groups (Supplementary Figure 12), and that it was additionally independent of a 25-factor smoking covariate[17], body mass index, type 2 diabetes mellitus, leukocyte count, lymphocyte count, and lymphocyte percentage (Supplementary Table 7).

Stratified analyses indicated expanded autosomal mCAs in individuals with cancer prior to infection (either any solid tumors, or hematologic malignancy after time of blood draw for genotyping) conferred stronger effects for sepsis (HR 2.79; 95% CI 2.30 to 3.38;

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

$P=9.7\times10^{-26}$) and respiratory system infections (HR 1.60; 95% CI 1.40 to 1.82; $P=6.1\times10^{-12}$) compared to individuals without a prior cancer history (sepsis: HR 1.25; 95% CI 0.80 to 1.95; P=0.33, $P_{interaction}=0.001$; respiratory system infections: HR 1.16; 95% CI 1.00 to 1.34; P=0.045, $P_{interaction}=0.001$) (Figure 4; Supplementary Figure 13–15). This interaction was driven by prevalent solid cancer, not hematologic cancer after DNA acquisition for mCA genotyping (Supplementary Table 8). Further multivariable adjustment indicated that incident sepsis and infection were independent of 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 ICD-10 and ICD-9 phecode groupings[19]) (Supplementary Table 9). We also explored the time difference between cancer diagnosis and specific infections to characterize potential influence from expanded mCA. Univariable analyses showed that expanded mCA carriers tend to have twice higher incidence of post-cancer diagnosis septicemia and pneumonia, and the difference in incidence rate was more prominent in infections occurring > 3 years from cancer diagnosis (Supplementary Table 10; Supplementary Figure 16). Besides cancer patients, we also calculated the univariable association between expanded mCA and diseases in the general public. On average, if we follow individuals without documented cancer, sepsis, or pneumonia history in UKB for 1000 person-years after expanded mCA detection, we would observe 36 individuals developing incident cancer (5 being hematological cancer), 14 individuals developing incident pneumonia, and 8 developing incident sepsis, respectively (Extended Data Figure 6).

## Association with COVID-19 severity

Across 719 COVID-19 hospitalized cases in the UKB, 44 individuals (6%) carried an expanded mCA clone at time of enrollment (in 2010), versus 3% among 337,877 controls. Adjusting for age, $age^2$, sex, prior or current smoking status, and principal components of ancestry, expanded mCAs were associated with COVID-19 hospitalizations (OR 1.59; 95% CI 1.13 to 2.25; P=0.0082), with higher effect estimates from expanded autosomal mCAs (OR 2.17; 95% CI 1.16 to 4.08; P=0.016) (Figure 5a). Analyses in FinnGen showed evidence of independent replication. The meta-analyzed associations across UKB and FinnGen of expanded autosomal mCAs on COVID-19 hospitalization was OR 2.44, 95% CI 1.33 to 4.46, P=0.0038 (Figure 5a).

In the UKB, further sensitivity analysis was performed; the associations persisted with additional adjustment for normalized Townsend deprivation index, normalized body mass index, type 2 diabetes mellitus, hypertension, coronary artery disease, any cancer, asthma, and chronic obstructive pulmonary disease (Extended Data Figure 7a). Additionally, similar associations were observed in the UKB when comparing COVID-19 hospitalization to tested negative controls, COVID-19 positive to all from English provinces and, COVID-19 positive to tested negative controls (Extended Data Figure 7b). Similar to the diverse nature of mCA clones observed in cases of incident infection, specific mCA clones carried by COVID-19 hospitalized individuals were also diverse in nature – across multiple chromosomes, a wide range of sizes, and both gain, loss, and CNN-LOH copy changes (Figure 5b). Similar effects associations effects of expanded mCAs with COVID-19 were also observed with incident pneumonia in the UKB (Extended Data Figure 7c).

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

We next identified 871 patients with COVID-19 from the Columbia University Biobank (CUB) and classified them into mutually exclusive ordinal categories based on COVID-19 outcomes and the World Health Organization's (WHO) COVID-19 progression scales: (1) mild cases (N=52) who are non-hospitalized COVID-19 patients (WHO stage 1–3), (2) moderate cases (N=440) including hospitalization but without intubation or death (WHO stage 4–6), (3) severe cases (N=379) including respiratory failure due to COVID-19 requiring endotracheal intubation (N=140; WHO stage 7–9) or death from COVID-19 (N=239; WHO stage 10). Individuals with prevalent hematologic cancer were excluded from analyses as before. Expanded autosomal mCAs were detected in 5.8% of mild cases, 13.9% of moderate, 16.9% of severe cases (Figure 5c). Expanded autosomal mCAs were associated with these ordinal COVID-19 outcomes with OR of 1.52 (95% CI 1.04 to 2.21, P=0.031), adjusted for age, sex, and self-reported ancestry. Summary statistics for the multivariate logistic regression are shown in Supplementary Table 11. This association was also independent of the status of any other prevalent cancers, validated by a sensitivity analysis that includes adjustment for any cancer diagnosis in Supplementary Table 12.

## Germline genetic predisposition to expanded mCAs

To further elucidate causal factors for expanded mCA clones, we performed a genome-wide association study (GWAS) in the UKB. We identified 63 independent genome-wide significant loci ($r^2 < 0.1$ across 1MB windows of the genome) (Figure 6a, Supplementary Table 13). Across the 63 germline variants, significant correlation was seen between different mCA categories (Extended Data Figure 8), suggesting the presence of shared germline genetic variants predisposing to mCAs across the genome. Follow-up analyses using an additive polygenic risk score comprised of 156 independent genome-wide significant variants associated with mosaic loss-of-chromosome Y (mLOY) from males from a prior study in the UKB[20], found significant associations with expanded autosomal mCAs and expanded ChrX mCAs in females, further highlighting the shared germline contributors towards mCAs across the genome (Extended Data Figure 9). Association of 156 previously identified independent genome-wide significant variants associated with mLOY from Thompson et al. (Nature 2019)[20], with expanded ChrY mCA categories in UKB shows that the two are highly correlated ($r_p$=0.91, P=$3.80\times10^{-57}$), with 1.87x higher effect estimates conferred on expanded ChrY mCAs compared to all mLOY variants as analyzed in Thompson et al. (Supplementary Figure 17). Additionally, strong correlation is seen between germline variants associated with mLOY and their associations with expanded ChrX mCAs, expanded autosomal mCAs, and all expanded mCAs (Supplementary Figure 17). Further analysis of the TP53 variant rs78378222-G identified a particularly strong effect on expanded ChrY mCAs (OR 2.03, 95% CI 1.79 to 2.31, P=$1.33\times10^{-27}$) in addition to all ChrY mCAs (OR 1.79, 95% CI 1.66 to 1.92, P=$8.81\times10^{-53}$), with the ChrY mCA effect being very similar to that previously reported in Thompson et al. Nature 2019[20] (Supplementary Table 14). The TP53 variant rs78378222-G was also associated with expanded autosomal mCAs (OR 1.51, 95% CI 1.21 to 1.88, P=0.00031) and expanded ChrX mCAs (OR 2.26, 95% CI 1.30 to 3.92, P=0.0038). The autosomal mCAs carried by individuals with rs78378222-G were diverse in size, copy-change, and location in the genome (Supplementary Figure 18). TWAS combining the expanded mCA GWAS results with GTEXv8[21] whole blood expression quantitative trait loci (eQTLs) using UTMOST[22]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

prioritized 62 genes ($P<3.2\times10^{-6}$) promoting expanded mCA development (Figure 6b, Supplementary Table 15). While gene enrichment analyses with the Elsevier Pathway Collection did not identify significantly associated pathways after multiple testing correction, top pathways were linked to DNA damage repair and lymphoid processes (Extended Data Figure 10a, Supplementary Table 16). The corresponding GWAS locus-zoom plots for some of these immune-related genes are shown in Extended Data Figure 10b. To prioritize tissues most implicated by these loci, tissue enrichment analyses using GenoSkyline-Plus were performed. Significant enrichment was identified in immune-specific epigenetic and transcriptomic functional regions of the genome ($P=7.1\times10^{-9}$) (Figure 6c). Further stratification of the immune category identified specific enrichment for CD4+ T-cells (P=0.00098) (Figure 6d).

## Discussion:

Across five geographically distinct biobanks comprising 768,762 individuals without known hematologic malignancy, clonal hematopoiesis (CH) represented by expanded mCAs is increasingly prevalent with age but not readily detectable by conventional medical blood tests. In addition to strongly predicting future risk of hematologic malignancy, expanded mCAs were also associated with risk for diverse incident infections, particularly sepsis and respiratory infections. These findings were robust across age, sex, tobacco smoking, and were strongest among those who develop cancer. Consistent with these observations, expanded mCAs were also associated with increased odds for COVID-19 hospitalization.

These results support several conclusions. First, mCA-driven CH is a potential risk factor for infection. Recent work showed that CH with myeloid malignancy driver mutations, also referred to as ‘clonal hematopoiesis of indeterminate potential’ (CHIP), predisposes to myeloid malignancy and coronary artery disease[23–27]. Meanwhile, CH with larger chromosomal alterations (i.e., mCAs) predisposes primarily to lymphoid malignancy but not coronary artery disease[10–12,15,16]. Our observations suggest CH defined by the presence of mCAs is a risk factor for infection. Since the relationship between mCAs and infection risk was not substantially attenuated when adjusting for leukocyte or lymphocyte counts at baseline visit, the impact of mCAs on infection risk likely acts through mechanisms independent of the impact of CH on cell counts. For example, as mCAs alter gene dosage (e.g., via duplications and deletions) and remove allelic heterogeneity (e.g., copy neutral loss-of-heterozygosity events) in leukocytes, potential impacts on the differentiation, function, and survival of leukocytes are mechanisms that could lead to altered infection risk. Our germline analyses specifically implicate lymphoid tissues. In particular, many of the mCA susceptibility loci are the same as those found in chronic lymphocytic leukemia, a condition in which lymphocyte differentiation and function is altered promoting infection risk[28–31]. Therefore, molecular changes in leukocytes that promote clonal expansion may occur at the expense of reduced ability to combat infection.

Second, the infectious disease risk associated with mCAs is exacerbated in the setting of cancer. It is well-established that mCAs in blood-derived DNA increase risk for hematologic cancer[10–12]. Furthermore, recent evidence suggests an association between mCAs detected in blood-derived DNA and increased risk of select solid tumor[14,17,32]. Our analysis

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

identified an interaction between mCAs and prior cancer diagnosis that amplified sepsis and pneumonia risk. Importantly, this interaction was restricted to individuals with solid cancers, not antecedent blood cancer. While this observation could be partially due to synergistic immunosuppressive side effects of cancer therapies[33], the observed associations persisted despite adjustment for many of these treatments. Alternatively, abnormal regulation of immune inflammatory pathways that release cytokines and inflammatory cells may create chronic states of inflammation in individuals with mCAs[34,35]. Based on our analysis, carriers of autosomal mCA are at an increased risk for sepsis (2.7x), pneumonia (1.8x), respiratory system infections (1.4x), digestive system infections (1.5x), and genitourinary system infections (1.3x), and these effects are more prominent in cancer patients. Surveillance for expanded mCA clones, particularly among those who develop solid cancer, may help identify individuals at high risk for infection that could benefit from targeted interventions.

Third, our findings could have particular relevance for the ongoing COVID-19 pandemic. We observed that mCAs are associated with elevated risk for COVID-19 hospitalization, with greater than two-fold risk linked to expanded autosomal mCAs. Maladaptive immune responses, particularly in leukocytes, increase risk for severe COVID-19 infections[36–39]. Awareness of COVID-19 risk associated with mCAs may help with the prioritization of prophylactic treatments. However, whether immune response to current vaccination approaches is altered in the context of mCAs deserves further study.

Fourth, our mCA germline genetic associations replicate many of those previously identified[10,11,20] and additionally suggest a common heritable basis across mCA classes, which may inform therapeutic targets. Genetic variants influencing risk of autosomal mCAs also influence risk of ChrX mCAs in females and ChrY mCAs in males. Furthermore, previously published genetic variants associated with mLOY[20] also influence risk of autosomal mCAs and ChrX mCAs in females. These loci may support putative therapeutic targets that may decrease the risk of mCA development, the rate of mCA clonal expansion, or the risk of progression of mCAs to clinical outcomes.

This analysis of mCAs and infection had some limitations. First, our study only measures mCAs at one time point for each participant. While our sampled mCA time point is likely correlated with CH at time of infection, CH dynamically changes over time potentially leading to differences in cellular fraction or additional undetected events that were acquired prior to infection. Second, we cannot rule out the possibility of undiagnosed hematologic malignancy among individuals with mCAs with only blood DNA. However, given the observed prevalence of mCAs (4% by age 60 years) among individuals without diagnosed hematologic malignancy and general scarcity of hematologic malignancy in the general population, we anticipate undiagnosed hematologic malignancy at DNA acquisition to be uncommon. Third, despite the robust adjustment and sensitivity analyses performed in our statistical analysis, including adjustment for cancer subtype, chemotherapy, bone marrow transplant, radiation, and other features associated with poor cancer prognosis (neutropenia, aplastic anemia, decreased white blood cell count), we cannot completely rule out the impact of residual confounding in our results from unknown or unmeasured sources. Here, consistency across cohorts and infection types and biologic plausibility mitigates this

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

possibility, and the empiric association of mCAs with incident infection may enable improved clinical risk prediction among patient populations as further scientific work is performed to understand the biological mechanisms by which mCAs influence the immune system. Fourth, further causal inference analyses using methods such as Mendelian randomization are limited by the low heritability of autosomal mCAs[11] and low heritability of infectious diseases[40,41]. However, defects in humoral, cell-mediated, and innate immunity have been linked to chronic lymphocytic leukemia (CLL)[28–31]. Whether all of these or specific aspects of immunity are altered for this pre-CLL condition requires further study.

In conclusion, we report evidence for increased susceptibility to a spectrum of infectious diseases in individuals carrying autosomal mCAs in a detectable fraction of leukocytes. The impacts of mCA on infection risk are systemic, with increased susceptibility to infection observed for a variety of organ systems, including severe COVID-19 presentations.

## Online Methods:

### Study samples

The UK Biobank, a population-based cohort of approximately 500,000 participants recruited from 2006–2010, had existing genomic and longitudinal phenotypic data[42]. 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 1$^{st}$ or 2$^{nd}$ 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)[42]. 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 hospitals[43]. 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.finngen.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-

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

derived DNA samples were genotyped using two versions of FinnGen ThermoFisher Axiom custom array (https://www.finngen.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 2003–2007[44]. 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.

The Columbia University Biobank (CUB) COVID-19 Cohort includes multiethnic patients with COVID-19 treated at the Columbia University Irving Medical Center (CUIMC) who underwent SNP genotyping on the Illumina Infinium Global Diversity Array. All patients in the cohort had a PCR-confirmed SARS-CoV-2 infection. All patients who had a blood draw at CUIMC after their positive PCR test were recruited regardless of hospitalization status. These patients were recruited to CUB between March and May 2020, at the peak of the first wave of the New York city pandemic, thus only a small fraction of the cohort were not hospitalized. The Columbia University Biobank COVID-19 studies are reviewed and approved by the Columbia University Medical Center IRB. A subset of patients was included under a public health crisis IRB waiver of consent specifically for COVID-19 studies if patients were deceased, not able to consent, or if the study team was unable to contact them as per Columbia IRB protocols AAAS9552 and AAAS7370. The primary analysis involved 871 patients and excluded individuals who had hematological malignancies. This cohort (N=871) was composed of 480 males and 391 females; the average age was 62 (range 7–101) years; 52% of the participants were self-reported to be Hispanic/Latinx, 14% Black/African American, 11% White/European, and 23% Other or Unknown. All COVID-19 positive patients were classified into exclusive ordinal outcome categories as defined by the WHO: (1) mild cases (N=52) who are non-hospitalized COVID-19 patients (WHO stage 1–3), (2) moderate cases (N=440) including hospitalization but without intubation or death (WHO stage 4–6), (3) severe cases (N=379) including respiratory failure due to COVID-19 requiring endotracheal intubation and mechanical ventilation (N=140; WHO stage 7–9) and death from COVID-19 (N=239; WHO stage 10).

## Mosaic chromosomal alteration detection

Mosaic chromosomal alteration (mCA) detection in the UK Biobank was as described previously[11,12]. Briefly, genotype intensities were transformed to log2(R ratio) (LRR) and B-allele frequency (BAF values) to estimate total and relative allelic intensities, respectively. Rephasing was performed using Eagle2[45] 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 described[12]. UK

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

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 SHAPEIT4[4] 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 engine[46] 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 Figure 1). We removed likely germline copy number polymorphisms (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 Figure 2).

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 SHAPEIT4 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 Figure 3). 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 Figure 4). 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 previously[10]. Briefly, genotyping intensity data was analysed across variants shared between the three primary arrays, and

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

used to compute BAF and LRR. Phasing was performed using the Eagle2 software. Mosaic events were called as previously described[12].

The CUB COVID-19 blood samples were genotyped by the Illumina Infinium Global Diversity Array. Detection of mCAs was performed starting from the probe raw intensity data (IDAT) files of 1,182 samples. The resulting raw intensity data were converted to VCF files using the bcftools gtc2vcf plugin (https://github.com/freeseek/gtc2vcf). Genotype phasing was performed using Eagle2 over the entire cohort. After excluding samples with call rate <0.97 and further removing 1st or 2nd degree relatives, the mCA calling was performed using the MoChA (https://github.com/freeseek/mocha). We excluded 133 samples with phased B-allele frequency (BAF) auto-correlation >0.05, indicative of contamination or other potential sources of poor DNA quality, and 6 samples with phenotype-genotype sex discordance (Supplementary Figure 5). We removed likely germline copy number polymorphisms (LOD_BAF_PHASE <20 for autosomal variants and LOD_BAF_PHASE <5 for sex chromosome variants), constitutional or inborn duplications (0–10Mbp mCAs with relative coverage >2.4) and deletions (filtering out mCAs with relative coverage <0.5) (Supplementary Figure 6). We further excluded 32 individuals with any prevalent hematologic cancer history at time of blood draw for genotyping and had 871 samples remaining for analyses.

## Clinical outcomes

Definitions for infection outcomes are detailed in Supplementary Tables 4,5. In the UKB, the first reported occurrences over median 8-year follow-up in Category 2410 were used as categorized by the UKB which maps primary care data, ICD-9 and ICD-10 codes from hospital inpatient data, ICD-10 codes in death register records, and self-reported medical conditions reported at the baseline, to ICD-10 codes. For each set of phenotypes grouped by organ system or by category, the time to first incident event after baseline examination in individuals free of prevalent history of each disease category was used. In the MGBB, electronic health record data was used to define incident ICD-10 codes grouped in the same fashion after DNA collection date over a median 3-year follow-up. In FinnGen, phenotypes were grouped together across ICD-8, ICD-9, and ICD-10 codes (Supplementary Table 2), with incident infections defined after DNA collection date over a median 3-year follow-up. In BBJ, analyses were performed using fatal incident events attributed to diverse infection outcomes in Supplementary Table 1 since non-fatal incident events were not available; additionally, analyses for pneumonia were performed using history of pneumonia prior to genotyping, based on interviews and medical record reviews[44]. 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 clinical phenotypes defined in the UKB, MGBB, and FinnGen are detailed in Supplementary Tables 17–19. 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

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

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. Smoking status in FinnGen 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.

UKB coronavirus disease 2019 (COVID-19), from SARS-CoV-2 infection, phenotypes used in the present analysis were downloaded on July 27, 2020. SARS-CoV-2 infection was determined by polymerase chain reaction from nasopharygeal, oropharyngeal, or lower respiratory samples obtained between March 16, 2020 and July 17, 2020. COVID-19 hospitalized cases were defined as any individual with at least one positive test who also had evidence for inpatient hospitalization (Field 40100). Controls included two sets: (1) participants from UKB English recruitment centers who were not known to have COVID-19, which were individuals with negative or no known SARS-CoV-2 testing or (2) participants with a negative SARS-CoV-2 test. Individuals with COVID-19 of unknown or low severity (i.e., at least one positive SARS-CoV-2 test without a known hospitalization) were excluded from the primary analyses.

Replication was performed in FinnGen where SARS-CoV-2 infection was determined either by polymerase chain reaction or by antibodies for samples obtained between March 2, 2020 and July 27, 2020. Across both cohorts, individuals who died prior to March 1, 2020, and therefore were not at risk for COVID-19 infection, were excluded from COVID-19 analyses.

### Statistical methods for infection associations

Association analyses of expanded mCAs with primary incident infection across 10 main infectious disease organ system categories (listed under “organ system” in Supplementary Table 1) were performed using Cox proportional hazards models, adjusting for age, age$^2$, sex, ever smoking status, and principal components 1–10 from the genotyping data. The age$^2$ term was added to account for potential quadratic associations between age and disease occurrence, as the association between mCAs and age is also nonlinear. Time since DNA collection was used as the underlying timescale. The proportional hazards assumption was assessed by Schoenfeld residuals and was not rejected. Individuals with a history of hematological cancer prior to DNA collection were excluded. P-value threshold for significance among the primary organ system infection analyses was two-sided Bonferroni threshold, $P<0.05/10=0.005$ to account for multiple hypothesis-testing. Analyses of incident events were performed separately in each biobank using the survival package in R (version 3.5, R Foundation, Vienna, Austria). Meta-analyses of the UKB, MGBB, and FinnGen results were performed using a fixed effects model from the meta package.

For UKB COVID-19 analyses, logistic regression was performed to estimate the association between expanded mCAs and COVID-19 hospitalization using the aforementioned phenotype definition, adjusting for sex, age, age$^2$, smoking status, and the first ten principal components from the genotyping data. As above, individuals with prevalent hematologic cancer were excluded from analyses. For the COVID-19 analyses, statistical significance was assigned at two-sided p-value < 0.05. Secondary multi-variable models were

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

additionally adjusted for normalized Townsend deprivation index[47], inverse rank normalized body mass index at baseline, and any prevalent or incident type 2 diabetes mellitus, hypertension, coronary artery disease, cancer, asthma, and chronic obstructive pulmonary disease.

Further sensitivity analyses were performed in the UK Biobank expanded autosomal mCA and infection associations. First, associations were additionally performed across 20 incident infections among the 10 broader organ system groups, with a Bonferroni threshold used for significance ($P<0.05/20=0.0025$). Second, 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, age$^2$, sex, ever smoking status, and the first ten principal components of genetic ancestry). Third, 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. Fourth, 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 phenotype[17], and 4) adjusting for normalized leukocyte count, lymphocyte count, and lymphocyte percentage at baseline visit. Fifth, 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 phecode groupings[19]), in the same aforementioned models adjusted for age, age$^2$, sex, ever smoking status, and the first ten principal components of genetic ancestry.

## Genome-wide association study

GWAS was performed using Hail-0.2 software (https://hail.is/) on the Google cloud. Variants were filtered to high-quality imputed variants (INFO score >0.4), with minor allele frequency >0.005, and with Hardy-Weinberg Equilibrium $P\geq 1\times 10^{-10}$, as previously performed. A Wald-logistic regression model was used for analysis, adjusting for age, age$^2$, sex, ever smoking, PC1–10, and genotyping array. Significant, independent loci were identified using $P<5\times 10^{-8}$ and clumping in Plink-2.0 using an $r^2$ threshold of 0.1 across 1MB genomic windows using the 1000-Genomes Project European reference panel. An additive mLOY polygenic risk score was developed as such: $\sum_{i=1}^{63} Beta \times SNP_{ij}$, where $Beta$ is the weight for each of the 156 independent genome-wide significant variants previously identified in UKB males[20] and $SNP_{ij}$ is the number of alleles (i.e., 0, 1, or 2) for $SNP_i$ in female $j$ in the UKB.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

### Cell-type enrichment analyses

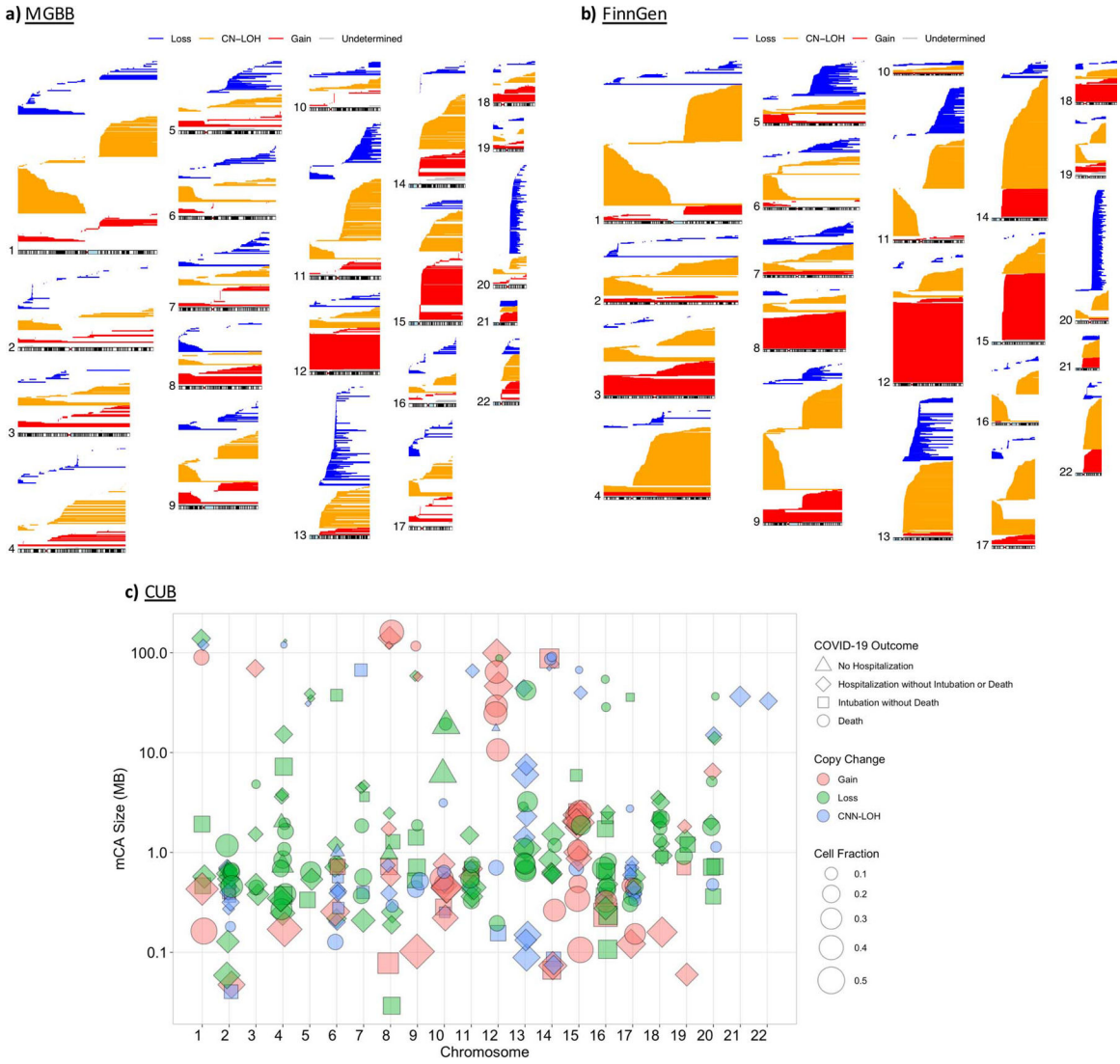
We applied partitioned LD score regression using the LDSC software[48] (v1.0.1) to perform enrichment analysis using the expanded mCA GWAS summary statistics in combination with tissue-specific epigenetic and transcriptomic functionality annotations from GenoSkyline-Plus[22]. In addition to the baseline annotations for diverse genomic features as suggested in the LDSC user manual, we specifically examined the enrichment signals on two tiers of annotations of different resolutions: GenoSkyline-Plus functionality scores of 7 broad tissue clusters (immune, brain, cardiovascular, muscle, gastrointestinal tract, epithelial, and others); and GenoSkyline-Plus functionality scores of 11 tissue and cell types within the immune cluster (listed in Figure 6d).

### Transcriptome-wide association and pathway enrichment analysis

Transcriptome-wide association was performed using the expanded mCA GWAS summary statistics in combination with the UTMOST[49] whole blood model updated to GTEXv8 (N=670). Significant genes were identified using a Bonferroni cutoff of $P<0.05/15,625$ or $3.2\times10^{-6}$. Pathway enrichment analyses was performed using genes with TWAS $P<0.001$ using the Elsevier Pathways through the EnrichR web server[50].

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

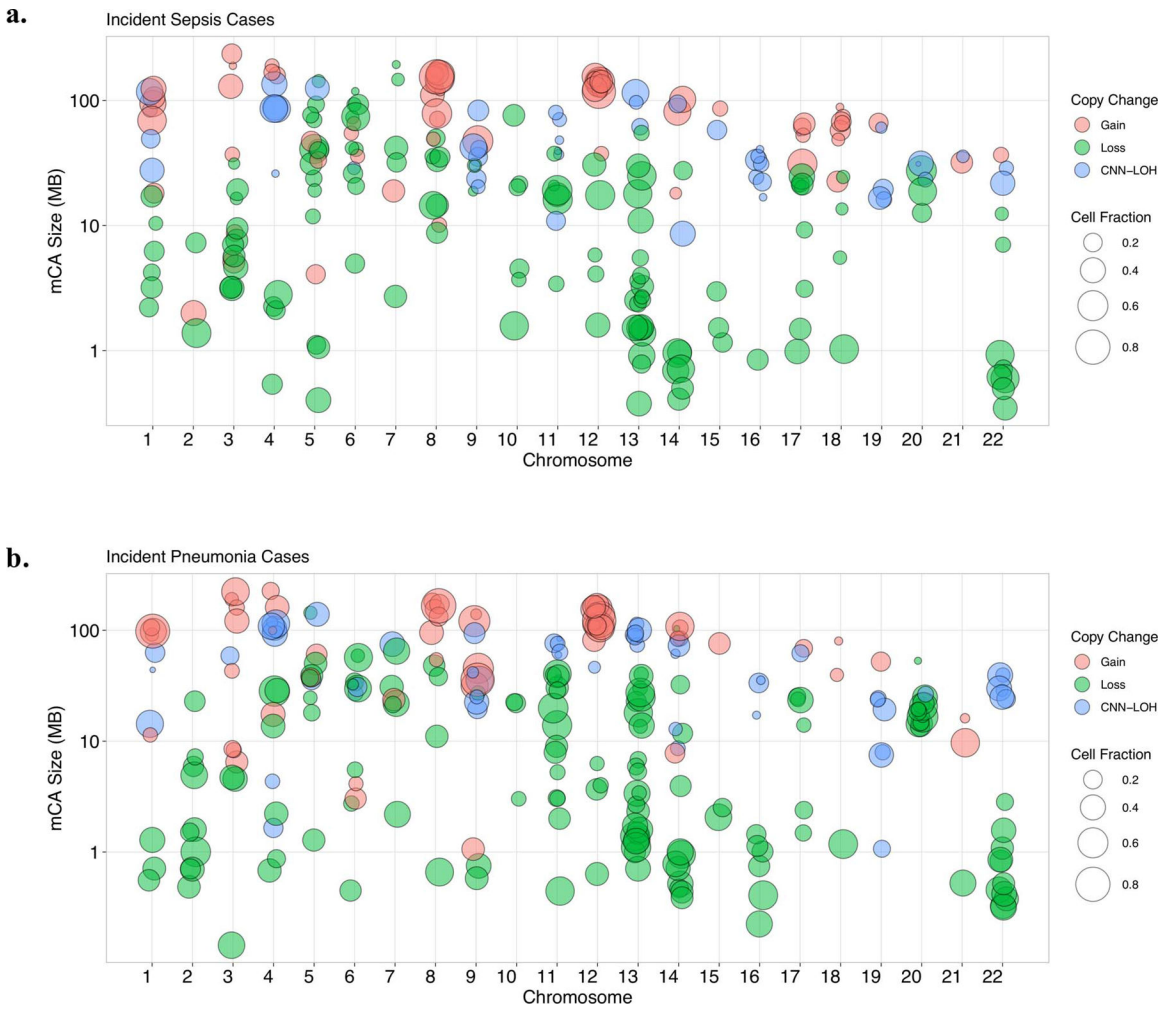
## Extended Data




**Extended Data Fig. 1. mCA calls by chromosome.**

mCA calls by chromosome in the a) MGBB b) FinnGen, and c) CUB. CN-LOH = copy neutral loss of heterozygosity, CUB=Columbia University Biobank, MGBB=Mass-General Brigham Biobank

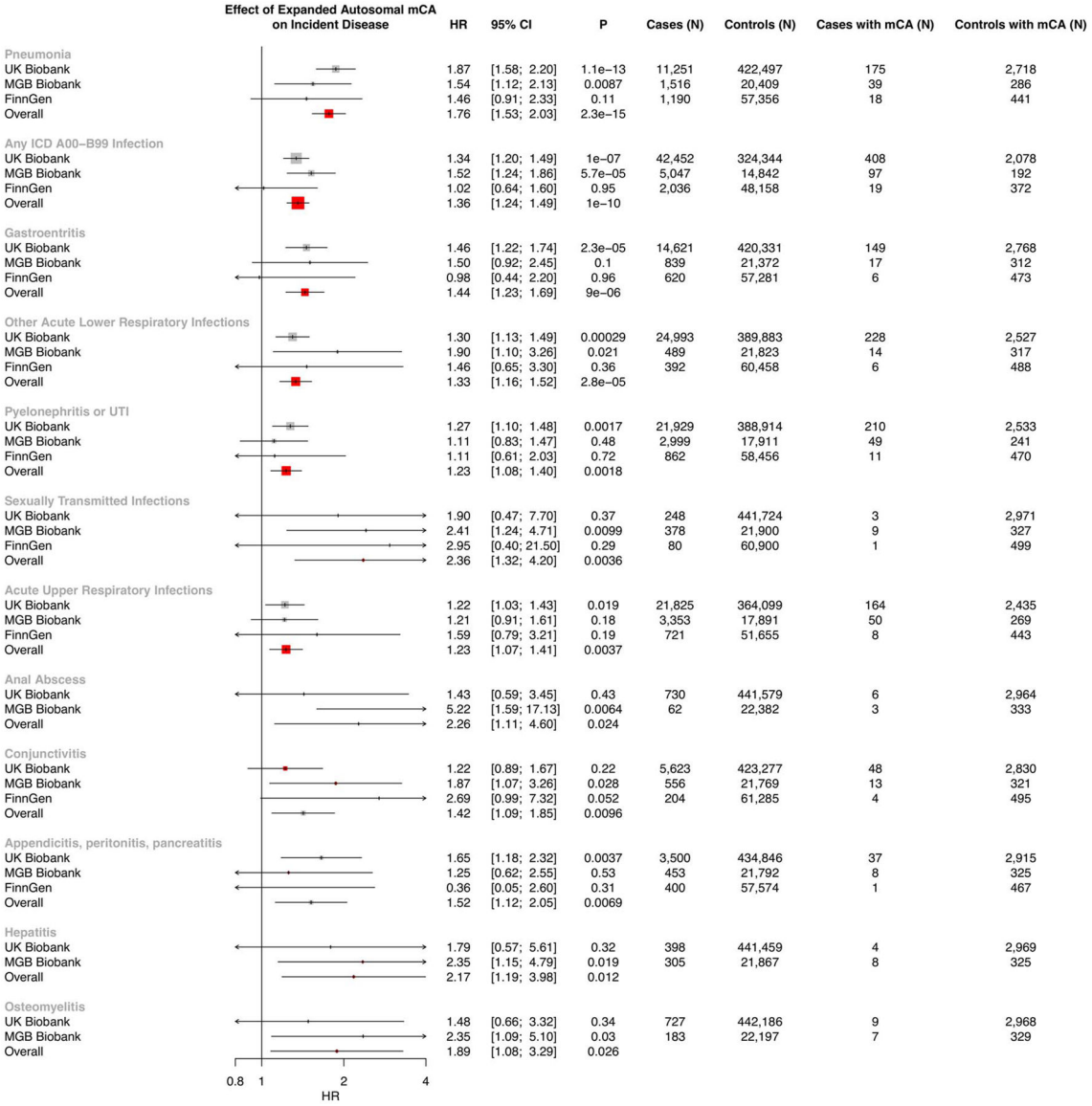
Author Manuscript Author Manuscript Author Manuscript Author Manuscript




**Extended Data Fig. 2. Visualization of the diverse range of expanded autosomal mCAs detected across the genome among individuals with a. incident sepsis and b. incident pneumonia in the UKB.**

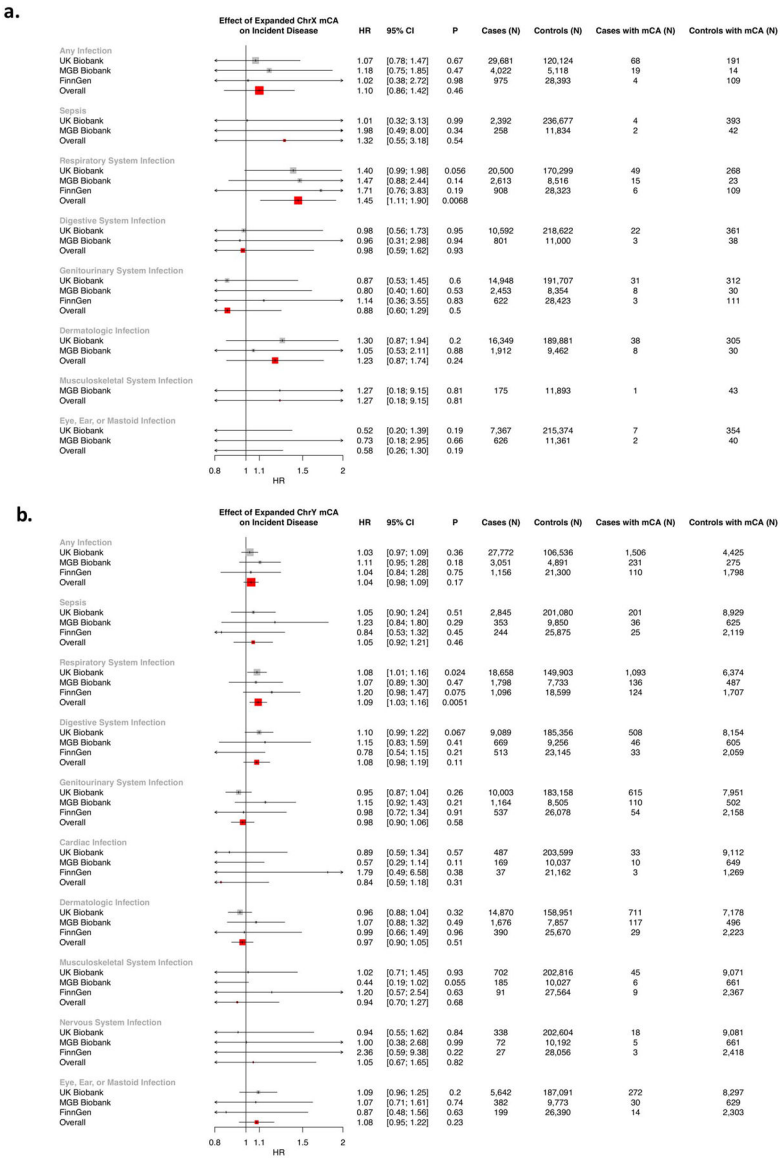
Each point represents one mCA carried by a case, with the x-axis as the chromosome, y-axis as the mCA size in mega-bases of DNA (MB), color as the copy change, and size of the point as the cell fraction of that mCA. CNN-LOH=copy number neutral loss of heterozygosity, MB = megabases of DNA, mCA = mosaic chromosomal alterations

Author Manuscript Author Manuscript Author Manuscript Author Manuscript




**Extended Data Fig. 3. Suggestive associations (P<0.05) of expanded autosomal mCAs with specific incident infections by Cox proportional-hazards models.**

Analyses are adjusted for age, age$^2$, sex, smoking status, and principal components 1–10 of ancestry. Bonferroni correction was used to determine the level of statistical significance (0.05/20 or P<0.0025). Overall estimates across studies are generated via fixed effect meta-analysis. Error bars show 95% confidence intervals. mCA = mosaic chromosomal alterations.




**Extended Data Fig. 4. Associations of a) expanded ChrY and b) expanded ChrX mCAs with incident infections.**

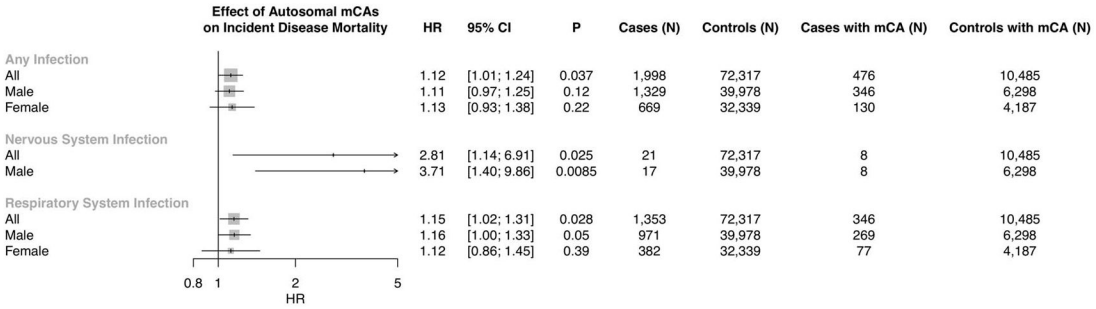
Both panels employ Cox proportional-hazards model adjusting for age, age$^2$, sex, smoking status, and principal components 1–10 of ancestry. Error bars show 95% confidence intervals. Bonferroni correction was used to determine the level of statistical significance for each mCA category (P<0.005). mCA = mosaic chromosomal alterations.



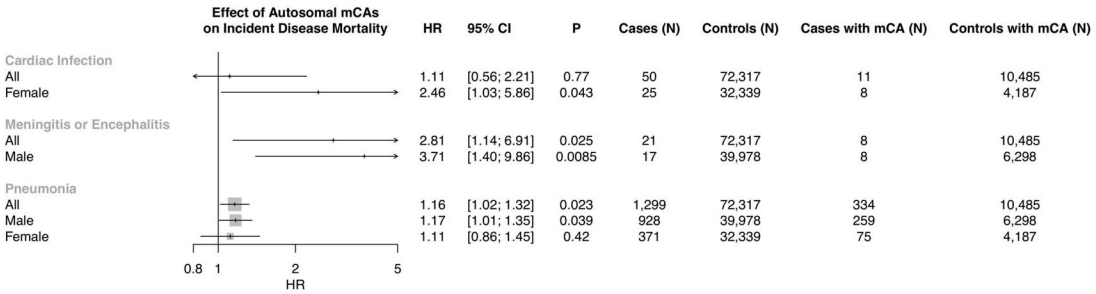
Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

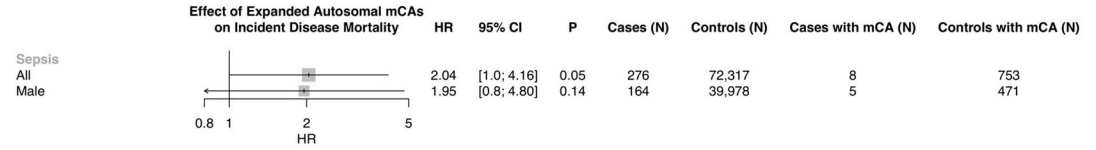
**a.**




**b.**



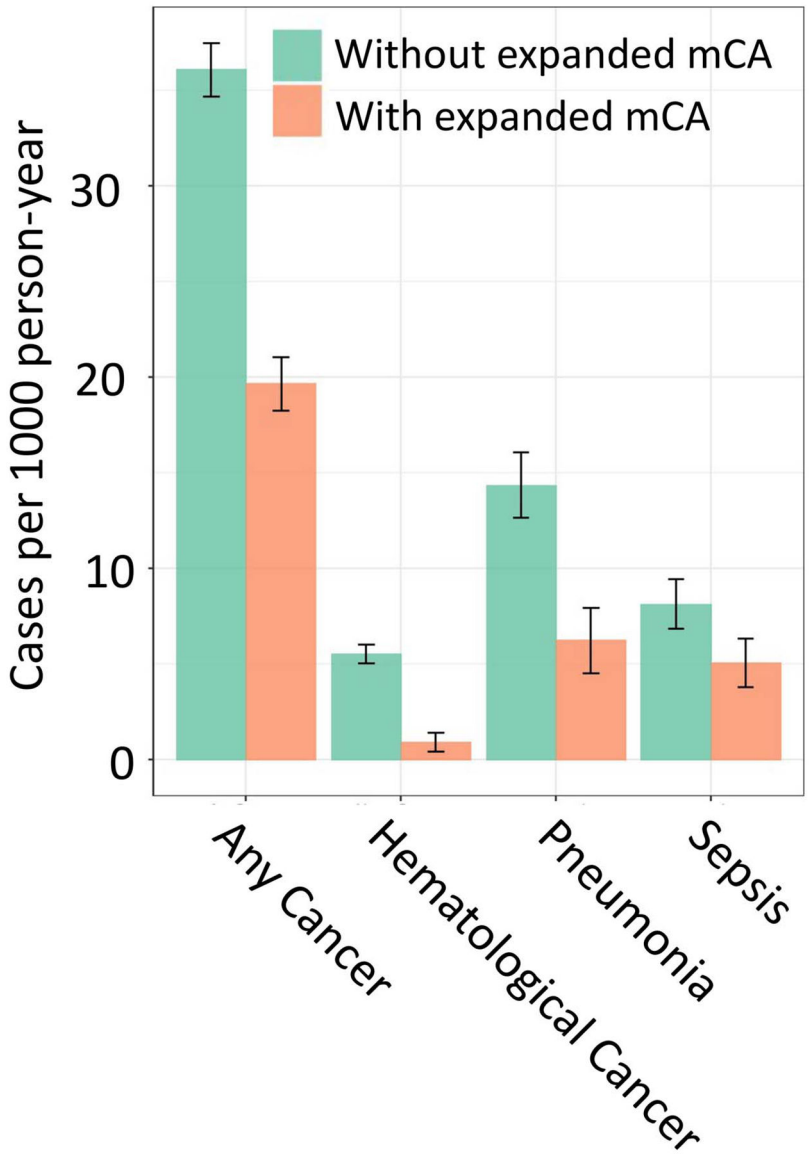

**c.**




**Extended Data Fig. 5. Suggestive associations (P<0.05) of mCAs with incident infection-related mortality in Biobank Japan**

Associations of autosomal mCAs with a) organ-system level infections and b) specific infection categories. c) Association of expanded autosomal mCAs with Sepsis. All panels employ Cox proportional-hazards model adjusting for age, age$^2$, sex, smoking status, and principal components 1–10 of ancestry. Error bars show 95% confidence intervals. Bonferroni correction was used to determine the level of statistical significance. Full results are in Supplementary Table 6. Associations are presented among individuals without any cancer history. mCA = mosaic chromosomal alterations.

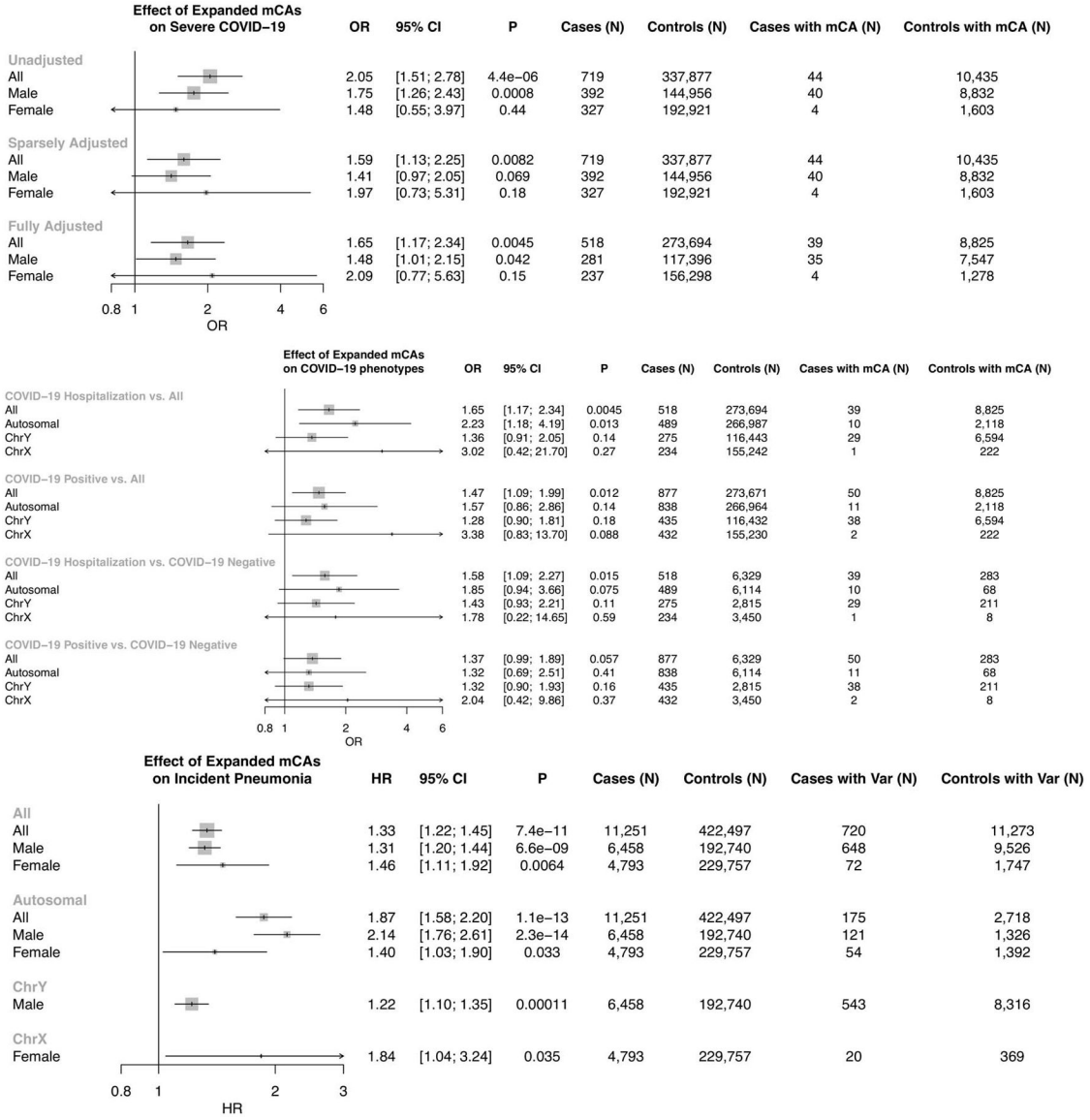
Author Manuscript Author Manuscript Author Manuscript Author Manuscript




**Extended Data Fig. 6. Incidence rate of at risk population developing each disease (N=445,101 UKB participants).**

95% confidence intervals were calculated based on normal approximation. mCA = mosaic chromosomal alterations

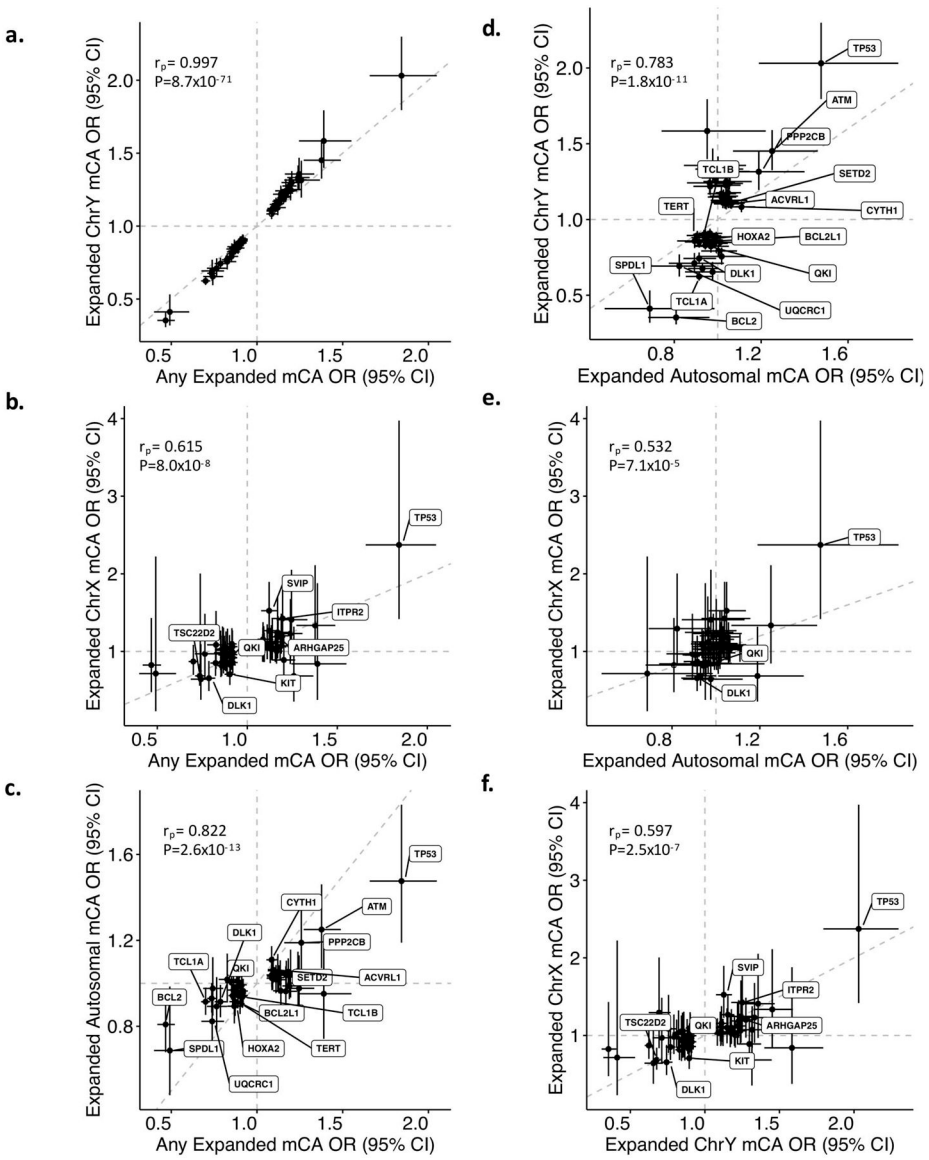
Author Manuscript Author Manuscript Author Manuscript Author Manuscript




**Extended Data Fig. 7. Associations of expanded mCAs in the UK Biobank with COVID-19 and incident pneumonia.**

Associations of expanded mCAs with **a.** COVID-19 hospitalization across different adjustment models, and **b.** different COVID-19 phenotypes in fully adjusted logistic regression models. Adjustment models include 1) an unadjusted model, 2) a sparsely adjusted model which adjusts for age, age2, sex, smoking status, and principal components of ancestry, and 3) a fully adjusted model which additionally adjusts for Townsend deprivation index, BMI, and the following comorbidities: Asthma, COPD, CAD, T2D, any cancer, and HTN. Bonferroni correction was used to determine the level of statistical significance. mCA = mosaic chromosomal alterations, COPD = chronic obstructive pulmonary disease, CAD = coronary artery disease, T2D = type 2 diabetes mellitus. **c.** Association of expanded mCAs with incident pneumonia stratified by sex, adjusted for age, age2, sex (in the All model only), smoking status, and principal components of ancestry. Error bars show 95% confidence intervals. mCA = mosaic chromosomal alterations

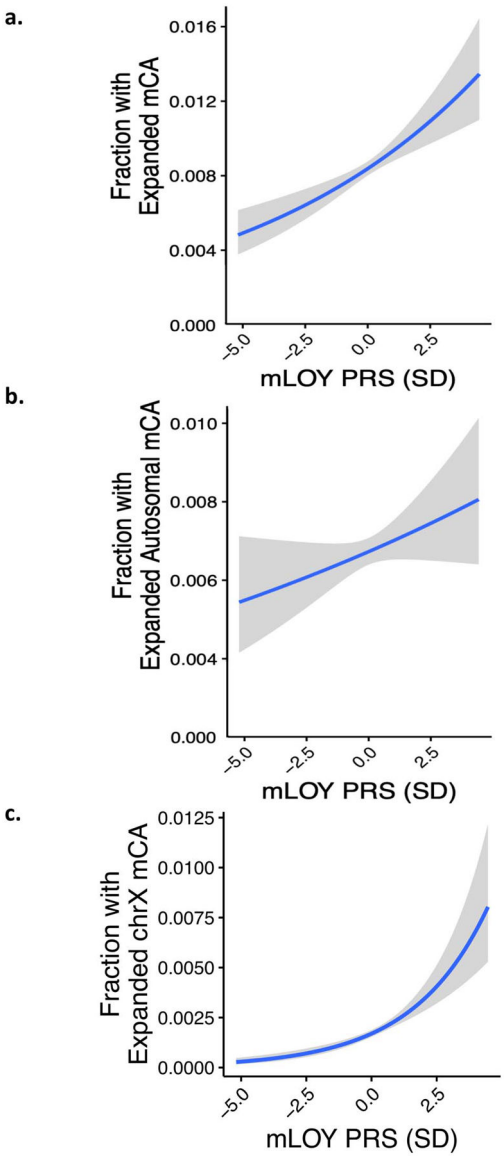
Author Manuscript Author Manuscript Author Manuscript Author Manuscript




**Extended Data Fig. 8. Correlated associations of 63 independent genome-wide significant variants associated with expanded mCAs between different mCA categories in the UKB.**

Bonferroni correction was used to determine the level of statistical significance for the correlation analyses (P<0.05/6=0.0083). 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

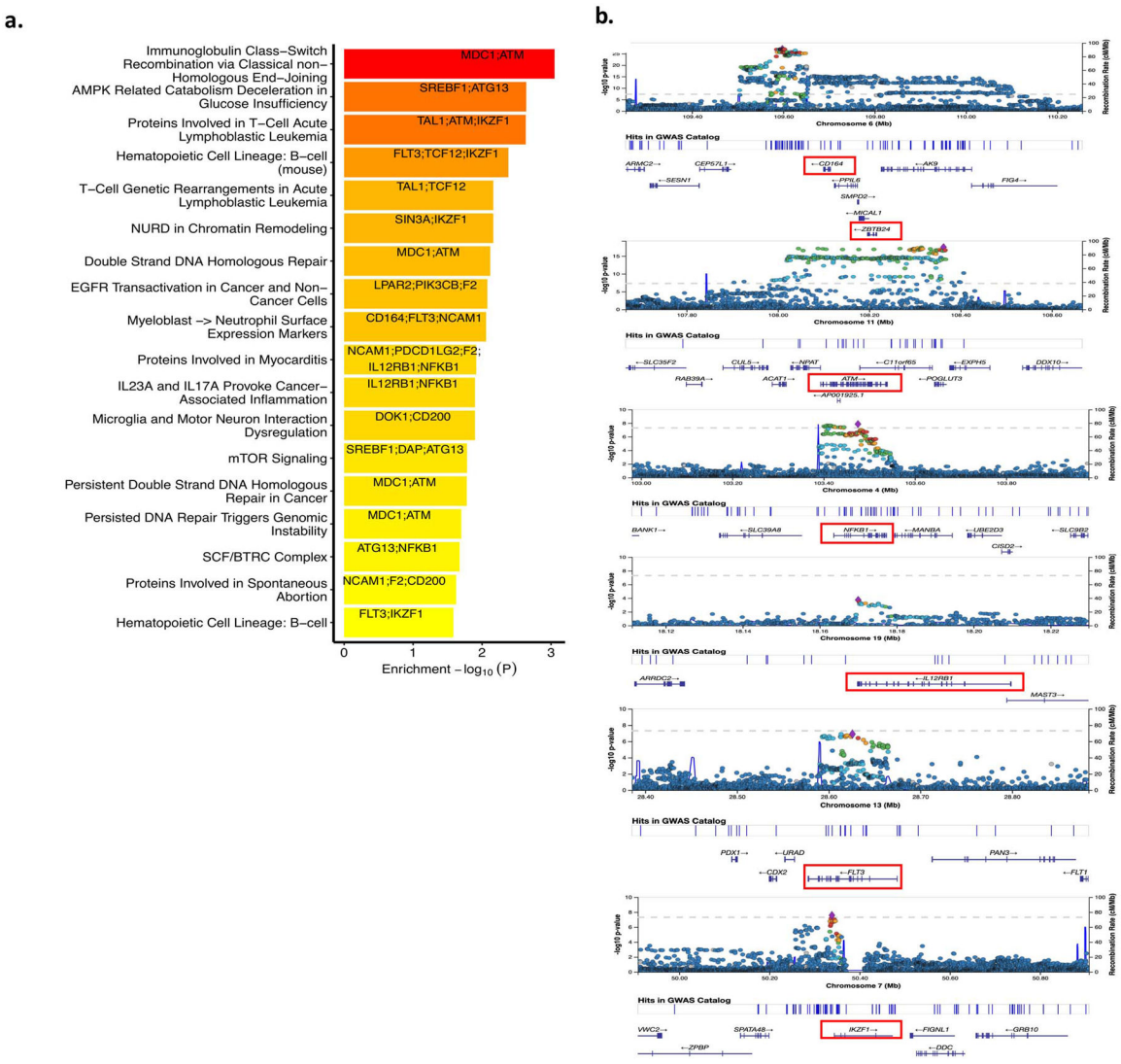
Author Manuscript Author Manuscript Author Manuscript Author Manuscript




**Extended Data Fig. 9. Association of a mLOY PRS consisting of 156 previously identified[20] independent genome-wide significant variants associated with mLOY, with different expanded mCA categories in UKB Females.**

Error bands were derived from binomial proportion 95% confidence intervals. mCA = mosaic chromosomal alterations, mLOY = mosaic Loss-of-chromosome Y, PRS = polygenic risk score

Author Manuscript Author Manuscript Author Manuscript Author Manuscript




**Extended Data Fig. 10. Pathway enrichment of TWAS results using the Elsevier Pathways.**

a. Top results from pathway enrichment analysis of the TWAS results using the Elsevier Pathways. b. Highlighting the GWAS locus-zoom plots for some of the TWAS genes implicated in the top pathways from panel a. Red boxes highlight the gene(s) with strongest association in the TWAS analyses. GWAS = genome-wide association study, TWAS = transcriptome-wide association study

## Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

## Acknowledgements:

Thanks to Chris Whelan, Chris Llanwarne, Jason Cerrato, Kyle Vernest, and Khalid Shakir and many other members of the Terra/Cromwell team for their help and advice in the development of the MoChA pipeline. Thanks to Petr Danecek for implementing critical features needed in BCFtools. Thanks to Stephen Chanock for critical input and comments. Thanks to Erikka Loftfield for assistance with the 25-level smoking adjustment variable.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Thanks to the participants and staff of the UKB, MGBB, and BBJ. UKB analyses were conducted using Applications 7089 and 21552. We would like to thank all study participants and their families for contributing to the Columbia University Biobank (CUB) COVID-19 Cohort. The genotyping was made possible by the Columbia University Biobank and its COVID-19 Genomics Workgroup members, including Andrea Califano, Wendy Chung, Christine K. Garcia, David B. Goldstein, Iuliana Ionita-Laza, Krzysztof Kiryluk, Richard Mayeux, Sheila M. O'Byrne, Danielle Pendrick, Muredach P. Reilly, Soumitra Sengupta, Peter Sims, and Anne-Catrin Uhlemann. We also acknowledge the COVID-19 Host Genetics Initiative consortium for providing infrastructure for collaboration.

**Funding:**

P.N. is supported by a Hassenfeld Scholar Award from the Massachusetts General Hospital, and grants from the National Heart, Lung, and Blood Institute (R01HL1427, R01HL148565, and R01HL148050). P.N. and B.L.E. are supported by a grant from Fondation Leducq (TNE-18CVD04). S.M.Z is supported by the NIH National Heart, Lung, and Blood Institute (1F30HL149180-01) and the NIH Medical Scientist Training Program Training Grant (T32GM136651). A.G.B. is supported by a Burroughs Wellcome Fund Career Award for Medical Scientists. G.G is supported by NIH grant R01 HG006855, NIH grant R01 MH104964, and the Stanley Center for Psychiatric Research. J.P.P is supported by a John S LaDue Memorial Fellowship. K.P. is supported by NIH grant 5-T32HL007208-43. P.T.E. is supported by supported grants from the National Institutes of Health (1RO1HL092577, R01HL128914, K24HL105780), the American Heart Association (18SFRN34110082), and by the Foundation Leducq (14CVD01). P.-R.L. is supported by NIH grant DP2 ES030554 and a Burroughs Wellcome Fund Career Award at the Scientific Interfaces. This work was supported by the Intramural Research Program of the National Cancer Institute, National Institutes of Health, extramural grants from the National Heart, Lung, and Blood Institute, and Fondation Leducq. The opinions expressed by the authors are their own and this material should not be interpreted as representing the official viewpoint of the U.S. Department of Health and Human Services, the National Institutes of Health, or the National Cancer Institute. The Columbia University Biobank is supported by the Vagelos College of Physicians & Surgeons as well as the Precision Medicine Resource and Biomedical Informatics Resource of Irving Institute for Clinical and Translational Research, home of the Columbia University's Clinical and Translational Science Award (CTSA), funded by the National Center for Advancing Translational Sciences, National Institutes of Health, through Grant Number UL1TR001873.

**Competing Interests:**

P.N. reported grants from Amgen during the conduct of the study and grants from Boston Scientific; grants and personal fees from Apple; personal fees from Novartis and Blackstone Life Sciences; and other support from Vertex outside the submitted work. P.T.E. has received grant support from Bayer AG and has served on advisory boards or consulted for Bayer AG, Quest Diagnostics, MyoKardia and Novartis, outside of the present work. S.M.Z., S-H.L., M.J.M., G.G., and P.N. have filed a patent application (serial no. 63/079,74) on the prediction of infection from mCAs. G.G. and S.A.M. have filed a patent application (PCT/WO2019/079493) for the MoChA mCA detection method employed in the present study. The remaining authors declare no competing interests.

## The Biobank Japan Project

Satoshi Koyama[30], Kaoru Ito[30], Yukihide Momozawa[31], Koichi Matsuda[32,33], Yuji Yamanashi[34], Yoichi Furukawa[35], Takayuki Morisaki[36], Yoshinori Murakami[37], Kaori Muto[38], Akiko Nagai[38], Wataru Obara[39], Ken Yamaji[40], Kazuhisa Takahashi[41], Satoshi Asai[42], Yasuo Takahashi[43], Takao Suzuki[44], Nobuaki Sinozaki[44], Hiroki Yamaguchi[45], Shiro Minami[46], Shigeo Murayama[47], Kozo Yoshimori[48], Satoshi Nagayama[49], Daisuke Obata[50], Masahiko Higashiyama[51], Akihide Masumoto[52], Yukihiro Koretsune[53]

[30]Laboratory for Cardiovascular Genomics and Informatics, RIKEN Center for Integrative Medical Sciences, Yokohama, Japan.

[31]Laboratory for Genotyping Development, RIKEN Center for Integrative Medical Sciences, Yokohama, Japan.

[32]Laboratory of Genome Technology, Human Genome Center, Institute of Medical Science, The University of Tokyo, Tokyo, Japan.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

[33]Laboratory of Clinical Genome Sequencing, Graduate School of Frontier Sciences, The University of Tokyo, Tokyo, Japan.

[34]Division of Genetics, The Institute of Medical Science, The University of Tokyo, Tokyo, Japan.

[35]Division of Clinical Genome Research, Institute of Medical Science, The University of Tokyo, Tokyo, Japan.

[36]Division of Molecular Pathology IMSUT Hospital, Department of Internal Medicine Project Division of Genomic Medicine and Disease Prevention, The Institute of Medical Science The University of Tokyo, Tokyo, Japan.

[37]Department of Cancer Biology, Institute of Medical Science, The University of Tokyo, Tokyo, Japan.

[38]Department of Public Policy, Institute of Medical Science, The University of Tokyo, Tokyo, Japan.

[39]Department of Urology, Iwate Medical University, Iwate, Japan.

[40]Department of Internal Medicine and Rheumatology, Juntendo University Graduate School of Medicine, Tokyo, Japan.

[41]Department of Respiratory Medicine, Juntendo University Graduate School of Medicine, Tokyo, Japan.

[42]Division of Pharmacology, Department of Biomedical Science, Nihon University School of Medicine, Tokyo, Japan.

[43]Division of Genomic Epidemiology and Clinical Trials, Clinical Trials Research Center, Nihon University School of Medicine, Tokyo, Japan.

[44]Tokushukai Group, Tokyo, Japan.

[45]Departmentof Hematology, Nippon Medical School, Tokyo, Japan.

[46]Department of Bioregulation, Nippon Medical School, Kawasaki, Japan.

[47]Tokyo Metropolitan Geriatric Hospital and Institute of Gerontology, Tokyo, Japan.

[48]Fukujuji Hospital, Japan Anti-Tuberculosis Association, Tokyo, Japan.

[49]The Cancer Institute Hospital of the Japanese Foundation for Cancer Research, Tokyo, Japan.

[50]Center for Clinical Research and Advanced Medicine, Shiga University of Medical Science, Shiga, Japan.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

[51]Department of General Thoracic Surgery, Osaka International Cancer Institute, Osaka, Japan.

[52]Iizuka Hospital, Fukuoka, Japan.

[53]National Hospital Organization Osaka National Hospital, Osaka, Japan.

## FinnGen Consortium

Aarno Palotie[54], Adam Ziemann[55], Adele Mitchell[56], Adriana Huertas-Vazquez[57], Aino Salminen[58], Airi Jussila[59], Aki Havulinna[54], Alex Mackay[60], Ali Abbasi[55], Amanda Elliott[54,61], Amy Cole[62], Anastasia Shcherban[54], Anders Mälarstig[63], Andrea Ganna[54], Andrey Loboda[57], Anna Podgornaia[57], Anne Lehtonen[55], Anne Pitkäranta[64], Anne Remes[65], Annika Auranen[59], Antti Hakanen[66], Antti Palomäki[67], Anu Jalanko[54], Anu Loukola[64], Aparna Chhibber[57], Apinya Lertratanakul[55], Arto Lehisto[54], Arto Mannermaa[68], Åsa Hedman[63], Audrey Chu[69], Aviv Madar[62], Awaisa Ghazal[54], Benjamin Challis[60], Benjamin Sun[56], Beryl Cummings[70], Bridget Riley-Gillis[55], Bridget Riley-Gills[55], Caroline Fox[57], Chia-Yen Chen[56], Clarence Wang[71], Clement Chatelain[71], Daniel Gordin[58], Danjuma Quarless[55], Danny Oh[72], David Choy[72], David Close[60], David Pulford[69], David Rice[58], Dawn Waterworth[73], Deepak Raipal[71], Deepak Rajpal[71], Denis Baird[56], Dhanaprakash Jambulingam[74], Diana Chang[72], Diptee Kulkarni[69], Dirk Paul[60], Dongyu Liu[71], Edmond Teng[72], Eero Punkka[64], Eeva Ekholm[67], Eeva Kangasniemi[75], Eija Laakkonen[76], Eleonor Wigmore[60], Elina Järvensivu[77], Elina Kilpeläinen[54], Elisabeth Widen[54], Ellen Tsai[56], Elmutaz Mohammed[78], Erich Strauss[72], Erika Kvikstad[78], Esa Pitkänen[54], Essi Kaiharju[77], Ethan Xu[71], Fanli Xu[69], Fedik Rahimov[55], Felix Vaura[79], Franck Auge[71], Georg Brein[54], Glenda Lassi[60], Graham Heap[55], Hannele Laivuori[54], Hannele Mattsson[77], Hannele Uusitalo-Järvinen[59], Hannu Kankaanranta[59], Hannu Uusitalo[59], Hao Chen[72], Harri Siirtola[80], Heikki Joensuu[58], Heiko Runz[56], Heli Lehtonen[63], Henrike Heyne[54], Hilkka Soininen[81], Howard Jacob[55], Hubert Chen[72], Huei-Yi Shen[54], Huilei Xu[62], Iida Vähätalo[80], Ilkka Kalliala[58], Ioanna Tachmazidou[60], Jaakko Kaprio[54], Jaakko Parkkinen[63], Jaison Jacob[62], Janet Kumar[69], Janet van Adelsberg[78], Jari Laukkanen[82], Jarmo Ritari[83], Javier Garcia-Tabuenca[80], Javier Gracia-Tabuenca[80], Jeffrey Waring[55], Jennifer Schutzman[72], Jimmy Liu[69], Jiwoo Lee[54,61], Joanna Betts[69], Joel Rämö[54], Johanna Huhtakangas[65], Johanna Mäkelä[75], Johanna Mattson[58], Johanna Schleutker[66], Johannes Kettunen[84], John Eicher[69], Jonas Zierer[62], Jonathan Chung[62], Joni A Turunen[58], Jorge Esparza Gordillo[69], Joseph Maranville[78], Juha Karjalainen[54,61], Juha Mehtonen[54], Juha Rinne[67], Juha Sinisalo[58], Juhani Junttila[84], Jukka Koskela[58], Jukka Partanen[85], Jukka Peltola[59], Julie Hunkapiller[72], Jussi Pihlajamäki[81], Justin Wade[55], Juulia Partanen[54], Kaarin Mäkikallio[67], Kai Kaarniranta[81], Kaisa Tasanen[65], Kaj Metsärinne[67], Kalle Pärn[54], Karen S King[69], Kari Eklund[58], Kari Linden[63], Kari Nieminen[59], Katariina Hannula-Jouppi[58], Katherine Call[71], Katherine Klinger[71], Kati Donner[54], Kati Hyvärinen[83], Kati Kristiansson[77], Katja Kivinen[54], Katri Kaukinen[59], Katri Pylkäs[86], Katrina de Lange[62], Keith Usiskin[78], Kimmo Palin[87], Kirill Shkura[57], Kirsi Auro[69], Kirsi Kalpala[63], Kirsi Sipilä[65], Klaus Elenius[67], Kristin Tsuo[54,61], L. Elisa Lahtela[54], Laura Addis[69], Laura Huilaja[65], Laura Kotaniemi-Talonen[59], Laura Mustaniemi[88], Laura Pirilä[67], Laure Morin-Papunen[65], Lauri Aaltonen[58], Leena Koulu[67], Liisa Suominen[81], Lila Kallio[66], Linda McCarthy[69], Liu Aoxing[54], Lotta

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Männikkö[77], Maen Obeidat[62], Manuel Rivas[89], Marco Hautalahti[88], Margit Pelkonen[81], Mari Kaunisto[54], Mari E Niemi[54], Maria Siponen[81], Marika Crohns[71], Marita Kalaoja[86], Marja Luodonpää[65], Marja Vääräsmäki[65], Marja-Riitta Taskinen[58], Marjo Tuppurainen[81], Mark Daly[54], Mark McCarthy[72], Markku Laakso[81], Markku Laukkanen[77], Markku Voutilainen[67], Markus Juonala[67], Markus Perola[77], Marla Hochfeld[78], Martti Färkkilä[58], Mary Pat Reeve[54], Masahiro Kanai[9], Matt Brauer[70], Matthias Gossel[71], Matti Peura[54], Meg Ehm[69], Melissa Miller[63], Mengzhen Liu[55], Mervi Aavikko[54], Miika Koskinen[64], Mika Helminen[80], Mika Kähönen[59], Mikko Arvas[85], Mikko Hiltunen[81], Mikko Kiviniemi[81], Minal Caliskan[78], Minna Karjalainen[86], Minna Raivio[58], Mirkka Koivusalo[88], Mitja Kurki[54,61], Mutaamba Maasha[9], Nan Bing[63], Natalie Bowers[72], Neha Raghavan[57], Nicole Renaud[62], Niko Välimäki[87], Nina Hautala[65], Nina Mars[54], Nina Pitkänen[66], Nizar Smaoui[55], Oili Kaipiainen-Seppänen[81], Olli Carpén[64], Oluwaseun A. Dada[54], Onuralp Soylemez[57], Oskari Heikinheimo[58], Outi Tuovila[90], Outi Uimari[65], Padhraig Gormley[69], Päivi Auvinen[81], Päivi Laiho[77], Päivi Mäntylä[81], Päivi Polo[67], Paola Bronson[56], Paula Kauppi[58], Peeter Karihtala[65], Pekka Nieminen[58], Pentti Tienari[58], Petri Virolainen[66], Pia Isomäki[59], Pietro Della Briotta Parolo[54], Pirkko Pussinen[58], Priit Palta[54], Raimo Pakkanen[90], Raisa Serpi[84], Rajashree Mishra[69], Reetta Hinttala[84], Reetta Kälviäinen[81], Regis Wong[77], Relja Popovic[55], Richard Siegel[62], Riitta Lahesmaa[67], Risto Kajanne[54], Robert Graham[70], Robert Plenge[78], Robert Yang[73], Roosa Kallionpää[67], Ruoyu Tian[56], Russell Miller[63], Sahar Esmaeeli[55], Saila Kauppila[65], Sally John[56], Sami Heikkinen[91], Sami Koskelainen[77], Samir Wadhawan[78], Sampsa Pikkarainen[58], Samuel Heron[74], Samuli Ripatti[54], Sanna Seitsonen[58], Sanni Lahdenperä[56], Sanni Ruotsalainen[54], Sarah Pendergrass[72], Sarah Smith[88], Sauli Vuoti[71], Shabbeer Hassan[54], Shameek Biswas[78], Shuang Luo[54], Sina Rüeger[54], Sini Lähteenmäki[77], Sirkku Peltonen[67], Sirpa Soini[77], Slavé Petrovski[60], Soumitra Ghosh[69], Stefan McDonough[63], Stephanie Loomis[56], Steven Greenberg[78], Susan Eaton[56], Susanna Lemmelä[54], Tai-He Xia[71], Tarja Laitinen[75], Taru Tukiainen[54], Teea Salmi[59], Teemu Niiranen[79], Teemu Paajanen[77], Teijo Kuopio[82], Terhi Kilpi[77], Terhi Ollila[58], Tero Hiekkalinna[77], Tero Jyrhämä[54], Terttu Harju[65], Tiina Luukkaala[80], Tiinamaija Tuomi[58], Tim Behrens[70], Tim Lu[72], Timo Blomster[65], Timo P. Sipilä[54], Tom Southerington[88], Tomi Mäkelä[92], Tuomo Kiiskinen[54], Tuomo Mantere[84], Tuomo Meretoja[58], Tushar Bhangale[72], Tuula Salo[58], Tuuli Sistonen[77], Ulla Palotie[58], Ulvi Gursoy[67], Urho Kujala[82], Valtteri Julkunen[81], Veikko Salomaa[79], Veli-Matti Kosma[68], Venkat Subramaniam Rathinakannan[74], Venla Kurra[59], Vesa Aaltonen[67], Victor Neduva[57], Vincent Llorens[54], Vishal Sinha[54], Vuokko Anttonen[65], Wei Zhou[9], Wilco Fleuren[73], Xing Chen[63], Xinli Hu[63], Ying Wu[63], Yunfeng Huang[56]

[54]Institute for Molecular Medicine Finland, HiLIFE, University of Helsinki, Finland

[55]AbbVie, Chicago, IL, United States

[56]Biogen, Cambridge, MA, United States

[57]Merck, Kenilworth, NJ, United States

[58]Hospital District of Helsinki and Uusimaa, Helsinki, Finland

[59]Pirkanmaa Hospital District, Tampere, Finland

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

[60]AstraZeneca, Cambridge, United Kingdom

[61]Broad Institute, Cambridge, MA, United States

[62]Novartis, Basel, Switzerland

[63]Pfizer, New York, NY, United States

[64]Helsinki Biobank, Helsinki University and Hospital District of Helsinki and Uusimaa, Helsinki

[65]Northern Ostrobothnia Hospital District, Oulu, Finland

[66]Auria Biobank, University of Turku, Hospital District of Southwest Finland, Turku, Finland

[67]Hospital District of Southwest Finland, Turku, Finland

[68]Biobank of Eastern Finland / University of Eastern Finland / Northern Savo Hospital District, Kuopio, Finland

[69]GlaxoSmithKline, Brentford, United Kingdom

[70]Maze Therapeutics, San Francisco, CA, United States

[71]Sanofi, Paris, France

[72]Genentech, San Francisco, CA, United States

[73]Janssen Biotech, Beerse, Belgium

[74]University of Turku, Turku, Finland

[75]Finnish Clinical Biobank Tampere, University of Tampere, Pirkanmaa Hospital District, Tampere, Finland

[76]University of Jyväskylä, Jyväskylä, Finland

[77]THL Biobank, The National Institute of Health and Welfare Helsinki, Finland

[78]Celgene, Summit, NJ, United States/ Bristol Myers Squibb, New York, NY, United States

[79]The National Institute of Health and Welfare Helsinki, Finland

[80]University of Tampere, Tampere, Finland

[81]Northern Savo Hospital District, Kuopio, Finland

[82]Central Finland Biobank, University of Jyväskylä / Central Finland Health Care District, Jyväskylä, Finland

[83]Finnish Red Cross Blood Service, Helsinki, Finland

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

[84]Northern Finland Biobank Borealis, University of Oulu, Northern Ostrobothnia Hospital District, Oulu, Finland

[85]Finnish Red Cross Blood Service, Finnish Hematology Registry and Clinical Biobank, Helsinki, Finland

[86]University of Oulu, Oulu, Finland

[87]University of Helsinki, Helsinki, Finland

[88]Finnish Biobank Cooperative, Turku, Finland

[89]University of Stanford, Stanford, CA, United States

[90]Business Finland, Helsinki, Finland

[91]University of Eastern Finland, Kuopio, Finland

[92]HiLIFE, University of Helsinki, Finland, Finland

## References

1. Gardner ID The effect of aging on susceptibility to infection. Rev Infect Dis 2, 801–810 (1980). [PubMed: 6763306]
2. Gavazzi G & Krause KH Ageing and infection. Lancet Infect Dis 2, 659–666 (2002). [PubMed: 12409046]
3. Aw D, Silva AB & Palmer DB Immunosenescence: emerging challenges for an ageing population. Immunology 120, 435–446 (2007). [PubMed: 17313487]
4. Franceschi C, Bonafe M & Valensin S Human immunosenescence: the prevailing of innate immunity, the failing of clonotypic immunity, and the filling of immunological space. Vaccine 18, 1717–1720 (2000). [PubMed: 10689155]
5. Ongradi J & Kovesdi V Factors that may impact on immunosenescence: an appraisal. Immun Ageing 7, 7 (2010). [PubMed: 20546588]
6. Panda A, et al. Human innate immunosenescence: causes and consequences for immunity in old age. Trends Immunol 30, 325–333 (2009). [PubMed: 19541535]
7. Aoshi T, Koyama S, Kobiyama K, Akira S & Ishii KJ Innate and adaptive immune responses to viral infection and vaccination. Curr Opin Virol 1, 226–232 (2011). [PubMed: 22440781]
8. Holly MK, Diaz K & Smith JG Defensins in Viral Infection and Pathogenesis. Annu Rev Virol 4, 369–391 (2017). [PubMed: 28715972]
9. Pallett LJ, Schmidt N & Schurich A T cell metabolism in chronic viral infection. Clin Exp Immunol 197, 143–152 (2019). [PubMed: 31038727]
10. Terao C, et al. Chromosomal alterations among age-related haematopoietic clones in Japan. Nature (2020).
11. Loh PR, Genovese G & McCarroll SA Monogenic and polygenic inheritance become instruments for clonal selection. Nature (2020).
12. Loh PR, et al. Insights into clonal haematopoiesis from 8,342 mosaic chromosomal alterations. Nature 559, 350–355 (2018). [PubMed: 29995854]
13. Lin SH, et al. Mosaic chromosome Y loss is associated with alterations in blood cell counts in UK Biobank men. Sci Rep 10, 3655 (2020). [PubMed: 32108144]
14. Forsberg LA, et al. Mosaic loss of chromosome Y in peripheral blood is associated with shorter survival and higher risk of cancer. Nat Genet 46, 624–628 (2014). [PubMed: 24777449]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript


15. Jacobs KB, et al. Detectable clonal mosaicism and its relationship to aging and cancer. Nat Genet 44, 651–658 (2012). [PubMed: 22561519]
16. Laurie CC, et al. Detectable clonal mosaicism from birth to old age and its relationship to cancer. Nat Genet 44, 642–650 (2012). [PubMed: 22561516]
17. Loftfield E, et al. Predictors of mosaic chromosome Y loss and associations with mortality in the UK Biobank. Sci Rep 8, 12316 (2018). [PubMed: 30120341]
18. Machiela MJ, et al. Characterization of large structural genetic mosaicism in human autosomes. Am J Hum Genet 96, 487–497 (2015). [PubMed: 25748358]
19. Wu P, et al. Mapping ICD-10 and ICD-10-CM Codes to Phecodes: Workflow Development and Initial Evaluation. JMIR Med Inform 7, e14325 (2019). [PubMed: 31553307]
20. Thompson DJ, et al. Genetic predisposition to mosaic Y chromosome loss in blood. Nature 575, 652–657 (2019). [PubMed: 31748747]
21. Consortium GT The GTEx Consortium atlas of genetic regulatory effects across human tissues. Science 369, 1318–1330 (2020). [PubMed: 32913098]
22. Lu Q, et al. Systematic tissue-specific functional annotation of the human genome highlights immune-related DNA elements for late-onset Alzheimer's disease. PLoS Genet 13, e1006933 (2017). [PubMed: 28742084]
23. Bick AG, et al. Genetic Interleukin 6 Signaling Deficiency Attenuates Cardiovascular Risk in Clonal Hematopoiesis. Circulation 141, 124–131 (2020). [PubMed: 31707836]
24. Genovese G, et al. Clonal hematopoiesis and blood-cancer risk inferred from blood DNA sequence. N Engl J Med 371, 2477–2487 (2014). [PubMed: 25426838]
25. Jaiswal S, et al. Age-related clonal hematopoiesis associated with adverse outcomes. N Engl J Med 371, 2488–2498 (2014). [PubMed: 25426837]
26. Jaiswal S, et al. Clonal Hematopoiesis and Risk of Atherosclerotic Cardiovascular Disease. N Engl J Med 377, 111–121 (2017). [PubMed: 28636844]
27. Xie M, et al. Age-related mutations associated with clonal hematopoietic expansion and malignancies. Nat Med 20, 1472–1478 (2014). [PubMed: 25326804]
28. Wang L, et al. Integrated single-cell genetic and transcriptional analysis suggests novel drivers of chronic lymphocytic leukemia. Genome Res 27, 1300–1311 (2017). [PubMed: 28679620]
29. de Weerdt I, et al. Innate lymphoid cells are expanded and functionally altered in chronic lymphocytic leukemia. Haematologica 101, e461–e464 (2016). [PubMed: 27662009]
30. Bartik MM, Welker D & Kay NE Impairments in immune cell function in B cell chronic lymphocytic leukemia. Semin Oncol 25, 27–33 (1998).
31. Arruga F, et al. Immune Response Dysfunction in Chronic Lymphocytic Leukemia: Dissecting Molecular Mechanisms and Microenvironmental Conditions. Int J Mol Sci 21(2020).
32. Zhou W, et al. Mosaic loss of chromosome Y is associated with common variation near TCL1A. Nat Genet 48, 563–568 (2016). [PubMed: 27064253]
33. Galluzzi L, Buque A, Kepp O, Zitvogel L & Kroemer G Immunological Effects of Conventional Chemotherapy and Targeted Anticancer Agents. Cancer Cell 28, 690–714 (2015). [PubMed: 26678337]
34. Balkwill F & Mantovani A Inflammation and cancer: back to Virchow? Lancet 357, 539–545 (2001). [PubMed: 11229684]
35. de Visser KE, Eichten A & Coussens LM Paradoxical roles of the immune system during cancer development. Nat Rev Cancer 6, 24–37 (2006). [PubMed: 16397525]
36. Lucas C, et al. Longitudinal analyses reveal immunological misfiring in severe COVID-19. Nature (2020).
37. Giamarellos-Bourboulis EJ, et al. Complex Immune Dysregulation in COVID-19 Patients with Severe Respiratory Failure. Cell Host Microbe 27, 992–1000 e1003 (2020). [PubMed: 32320677]
38. Huang C, et al. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. Lancet 395, 497–506 (2020). [PubMed: 31986264]
39. Cunha LL, Perazzio SF, Azzi J, Cravedi P & Riella LV Remodeling of the Immune Response With Aging: Immunosenescence and Its Potential Impact on COVID-19 Immune Response. Front Immunol 11, 1748 (2020). [PubMed: 32849623]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

40. Zekavat SM, et al. Elevated Blood Pressure Increases Pneumonia Risk: Epidemiological Association and Mendelian Randomization in the UK Biobank. Med (N Y) (2020).
41. Tian C, et al. Genome-wide association and HLA region fine-mapping studies identify susceptibility loci for multiple common infections. Nat Commun 8, 599 (2017). [PubMed: 28928442]
42. Bycroft C, et al. The UK Biobank resource with deep phenotyping and genomic data. Nature 562, 203–209 (2018). [PubMed: 30305743]
43. Smoller JW, et al. An eMERGE Clinical Center at Partners Personalized Medicine. J Pers Med 6(2016).
44. Nagai A, et al. Overview of the BioBank Japan Project: Study design and profile. J Epidemiol 27, S2–S8 (2017). [PubMed: 28189464]
45. Loh PR, et al. Reference-based phasing using the Haplotype Reference Consortium panel. Nat Genet 48, 1443–1448 (2016). [PubMed: 27694958]
46. Voss K, Auwera GVD & Gentry J Full-stack genomics pipelining with GATK4 + WDL + Cromwell. (2017).
47. Townsend P, Phillimore P, Beattie A Health and deprivation. Inequality and the North. Health Policy 10(1989).
48. Finucane HK, et al. Partitioning heritability by functional annotation using genome-wide association summary statistics. Nat Genet 47, 1228–1235 (2015). [PubMed: 26414678]
49. Hu Y, et al. A statistical framework for cross-tissue transcriptome-wide association analysis. Nat Genet 51, 568–576 (2019). [PubMed: 30804563]
50. Kuleshov MV, et al. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. Nucleic Acids Res 44, W90–97 (2016). [PubMed: 27141961]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

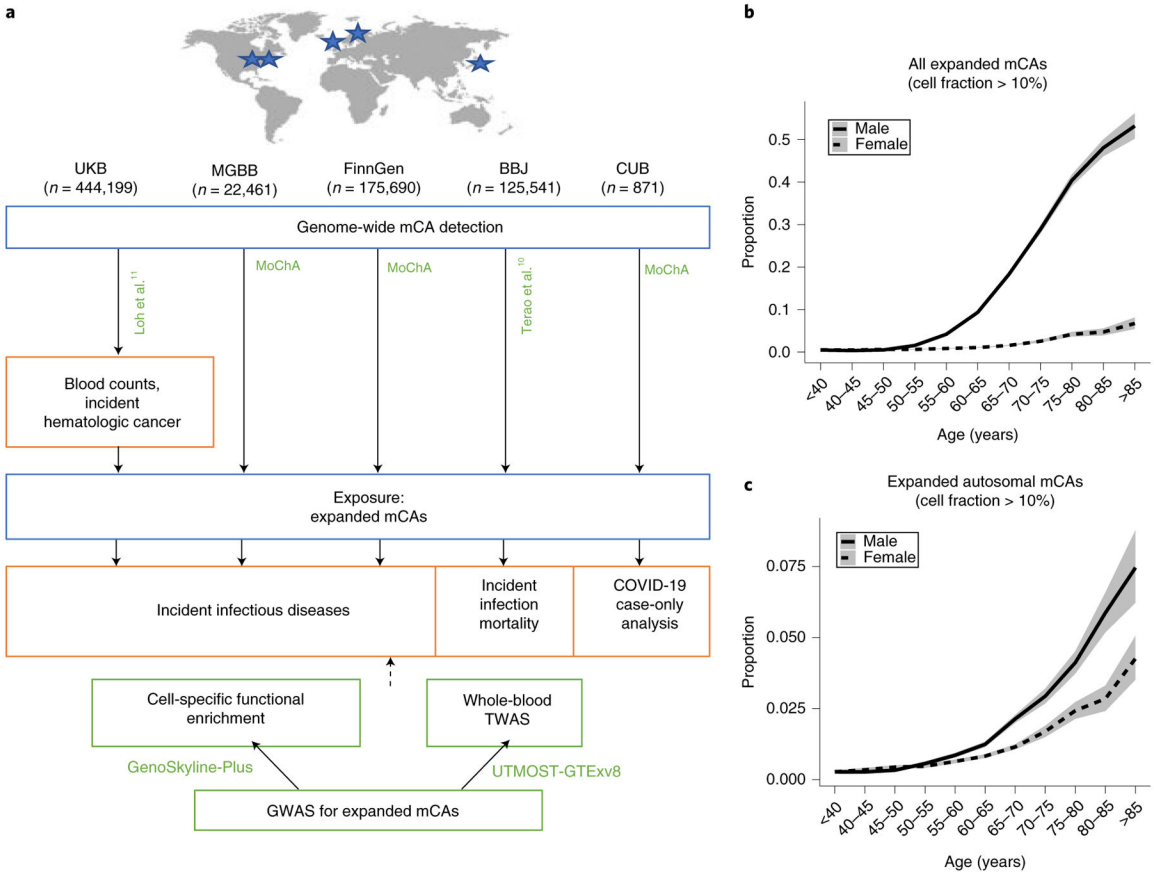



**Figure 1:**

Study schematic. **a.** Genome-wide mCAs were detected across the UKB[11], MGBB (via the MoChA pipeline), FinnGen (via the MoChA pipeline), BBJ[10], and CUB. Association of expanded mCAs (cell fraction >10%) with incident infectious diseases in UKB, MGBB, and FinnGen, with incident infectious disease mortality in BBJ, and with COVID-19 severity among COVID-19 positive cases in the CUB, was performed. A GWAS for expanded mCAs was then performed in the UKB to discover causal factors for expanded mCAs. Using the GWAS results, cell-specific functional enrichment analyses were performed using GenoSkyline-Plus, which combines epigenetic and transcriptomic annotations with GWAS summary statistics to estimate the relative contribution of cell-specific functional markers to the GWAS results. Additionally, to prioritize putative causal genes and pathways promoting the development of expanded mCAs, whole blood TWAS was performed using UTMOST via GTEx v8. Association of **b**. all expanded mCAs with cell fraction >10%, and **c**. all expanded autosomal mCAs, with age using 5-year age bins stratified by sex among individuals in the UKB, MGBB, FinnGen, and BBJ combined. Error bands were derived from binomial proportion 95% confidence intervals. Plots by cohort and across other mCA groupings are available in Supplementary Figure 8, 9. BBJ = BioBank Japan, CUB = Columbia University Biobank, GTEx v8 = Genotype-Tissue Expression project version 8, GWAS=genome-wide association study, MGBB = Mass General Brigham Biobank, mCA = mosaic chromosomal alterations, MoChA = Mosaic Chromosomal Alterations software

(https://github.com/freeseek/mocha), TWAS = transcriptome-wide association study, UKB = UK Biobank, UTMOST = Unified Test for MOlecular SignaTures.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

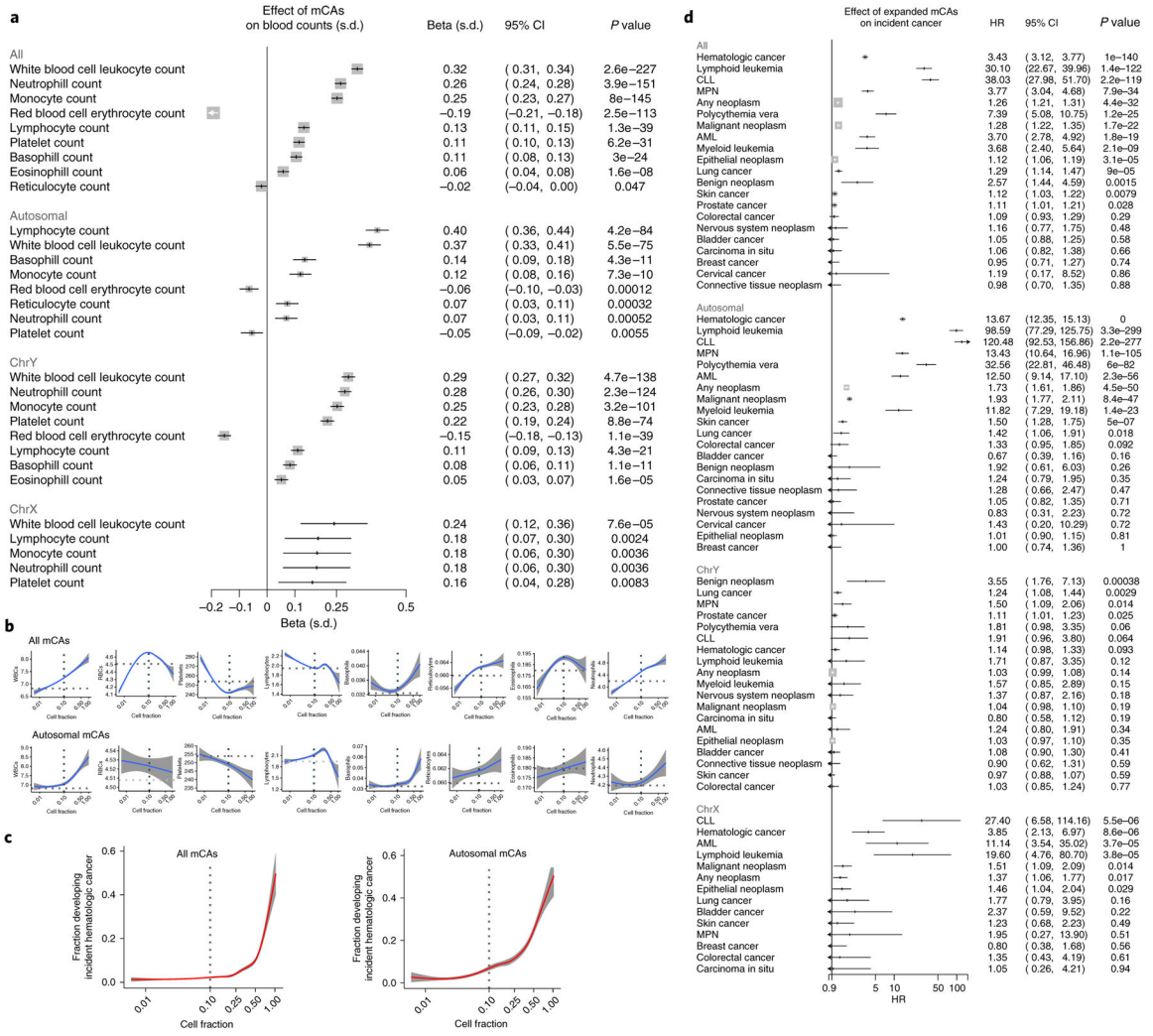



**Figure 2:**

Associations of mCAs with hematologic traits. **a.** Linear regression is employed to explore the association between blood counts and expanded mCAs. Associations are adjusted for age, age$^2$, sex, smoking status, and principal components of ancestry. Error bars show the 95% confidence interval for estimates. Bonferroni correction was used to determine the level of statistical significance. **b.** Relationship of mCA cell fraction with blood counts (in units of 10^9 cells/L) in the UKB among individuals without prevalent hematologic cancer at time of blood draw for genotyping and cell count measurement. The dotted horizontal lines reflect the mean blood count for individuals without an mCA. The dotted vertical lines at cell fraction of 0.10 represents the cutoff for the expanded mCA definition. Individuals with known hematologic cancer at time of or prior to blood draw for genotyping were excluded. Error bands were derived from binomial proportion 95% confidence intervals. **c.** Association of expanded mCA categories (with cell fraction>10%) with incident cancer in the UK Biobank. Analyses are adjusted for age, age$^2$, sex, smoking status, and principal components of ancestry. Individuals with a history of hematologic cancer at enrollment were removed from analysis. Error bands were derived from binomial proportion 95% confidence intervals.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

**d.** Association of expanded mCA categories (with cell fraction>10%) with incident cancer in the UK Biobank is accessed by Cox proportional-hazards model with time-on-study as the underlying time scale. Analyses are adjusted for age, age$^2$, sex, smoking status, and principal components of ancestry. Error bars show the 95% confidence interval for estimates. Bonferroni correction was used to determine the level of statistical significance. Individuals with a history of hematologic cancer at enrollment were removed from analysis.
CLL = chronic lymphocytic leukemia, MPN = myeloproliforative neoplasm, mCA = mosaic chromosomal alterations

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

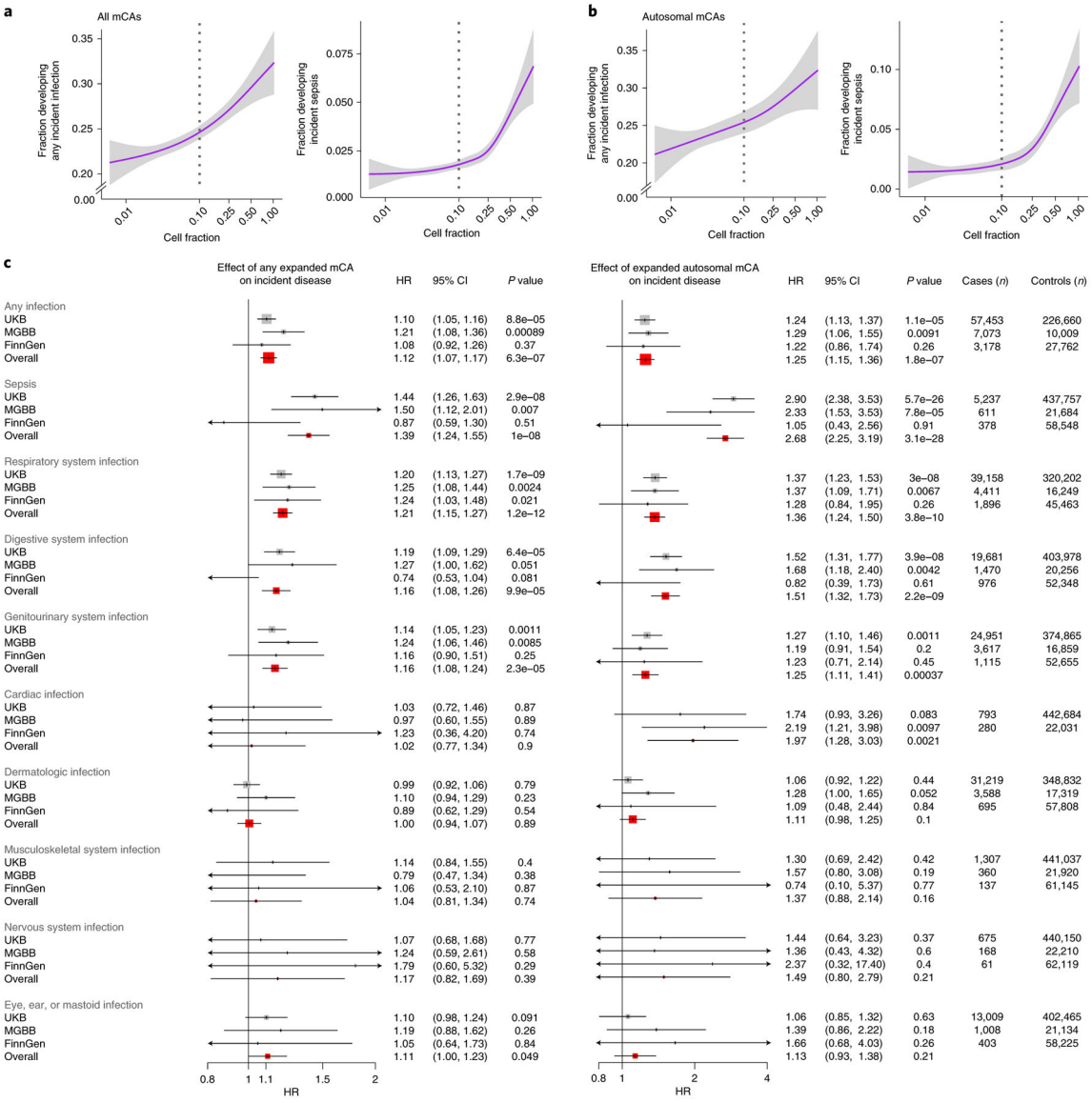



**Figure 3:**

Associations of expanded mCAs with incident infections. Visualizing the dependence with cell fraction among **a.** all mCAs, and **b.** autosomal mCAs, of any incident infection and incident sepsis in the UKB among individuals without prevalent hematologic cancer at time of blood draw for genotyping across. The dotted vertical lines at cell fraction of 0.10 represents the cutoff for the expanded mCA definition. Error bands were derived from binomial proportion 95% confidence intervals. **c.** Association of all expanded mCAs, and separately, expanded autosomal mCAs with incident infections across individuals in the UKB, MGBB, and FinnGen by Cox proportional-hazards models with the underlying time scale of time-on-study. Analyses are adjusted for age, age$^2$, sex, smoking status, and principal components 1–10 of ancestry. Error bars show the 95% confidence interval for estimates. Bonferroni correction was used to determine the level of statistical significance. Individuals with prevalent hematologic cancer were excluded from analysis. Association

analyses for other groupings of mCAs (including across all mCAs regardless of cell fraction, as well as chrX and chrY mCAs are provided in Supplementary Figures 11, 12). BBJ = BioBank Japan, MGBB = Mass General Brigham Biobank, mCA = mosaic chromosomal alterations, UKB = UK Biobank

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

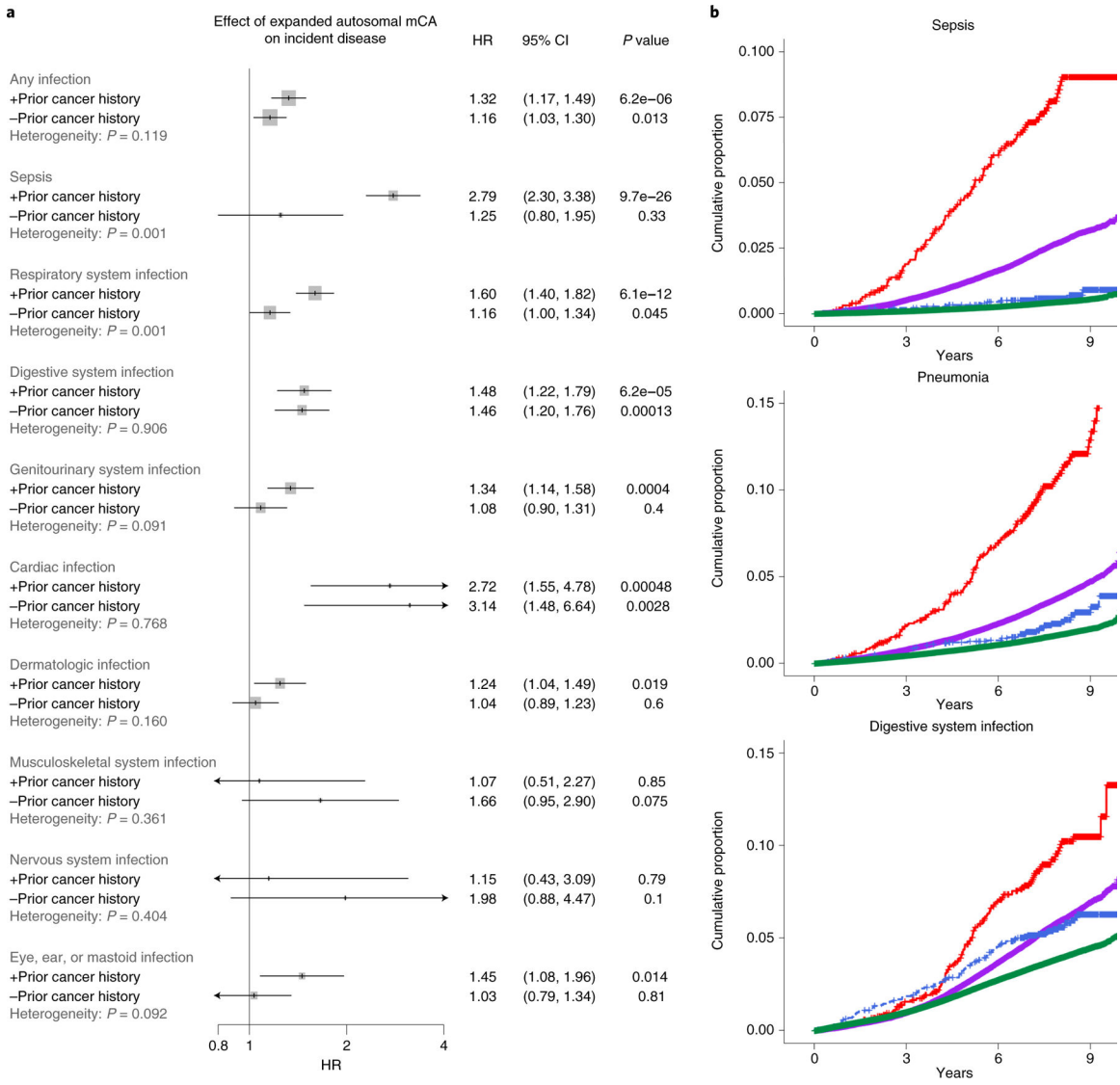



**Figure 4:**

Association of expanded autosomal mCAs and incident infections, stratified by antecedent cancer history. **a.** Association of expanded autosomal mCAs with incident infections across individuals with and without a cancer history before their incident infection, meta-analyzed across UKB, MGBB, and FinnGen combined (cohort-specific analyses are available in Supplementary Figure 14) assuming a fixed effect. Error bars show the 95% confidence interval for estimates. Bonferroni correction was used to determine the level of statistical significance. Individuals with known hematologic cancer at time of or prior to blood draw for genotyping were excluded. Analyses are adjusted for age, age2, sex, smoking status, and principal components of ancestry. **b.** Cumulative incidence curves for various infections in UKB. Top: sepsis, middle: pneumonia, bottom: digestive system infection. Results from MGBB and FinnGen are available in Supplmentary Figure 16. **Red:** mCA+ Cancer+, **Purple:** mCA− Cancer+, **Blue:** mCA+ Cancer−, **Green:** mCA− cancer−. Individuals with known hematologic cancer at time of or prior to blood draw for genotyping were excluded.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

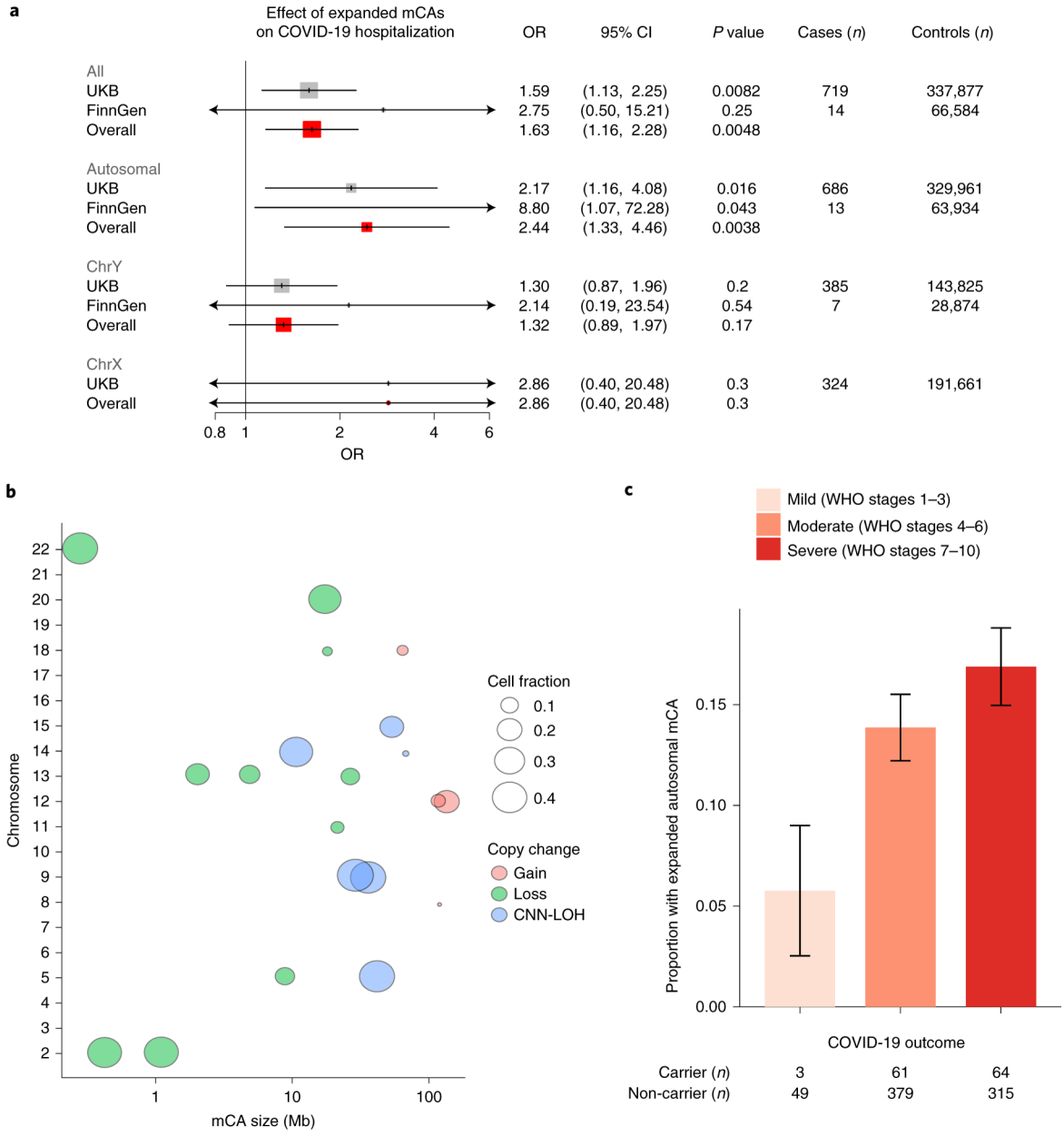



**Figure 5:**

Association of expanded mCAs with COVID-19 severity. **a.** Association of expanded mCAs with COVID-19 Hospitalization across the UKB and FinnGen determined by logistic regression. Error bars show the 95% confidence interval for estimates. Bonferroni correction was used to determine the level of statistical significance. Individuals with known hematologic cancer at time of or prior to blood draw for genotyping were excluded. Analyses are adjusted for age, age$^2$, sex, ever smoking status, and principal components of ancestry. **b.** Visualization of the diverse range of expanded autosomal mCAs detected across the genome among individuals hospitalized with COVID-19 in the UK Biobank. Each point represents one mCA carried by a case, with the x-axis as the chromosome, y-axis as the mCA size in mega-bases of DNA (MB). **c.** Proportion of expanded autosomal mCAs in each

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

category of COVID-19 outcomes for the CUB COVID-19 cohort, defined using the WHO COVID-19 scale (n=871 participants). 95% binomial proportion confidence intervals are shown. The table below the bar chart shows the counts of expanded autosomal mCA carriers and non-carriers in each outcome category. In CUB, the adjusted association between expanded autosomal mCAs and these ordinal COVID-19 outcomes is evaluated by ordinal regression and has OR of 1.52 (CI 95% 1.04 to 2.21, P= 0.031, two-tailed); summary statistics for the covariates included in the adjusted model for CUB are in Supplementary Table 11. MGBB = Mass General Brigham Biobank, UKB = UK Biobank, MB=megabase, CNN-LOH = copy number neutral loss of heterozygosity, CUB = Columbia University Biobank, WHO = World Health Organization

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

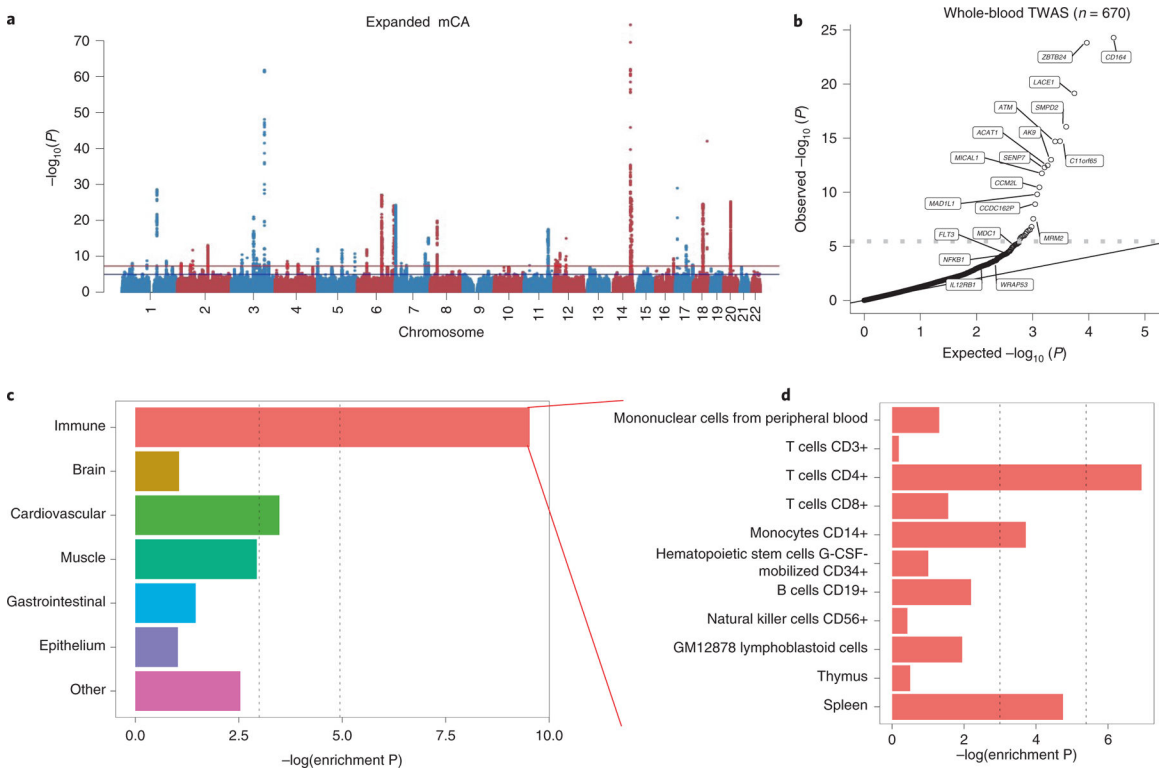



**Figure 6:**

Inherited risk factors for expanded mCAs: GWAS, TWAS, and Cell Type Enrichment. **a.** GWAS for expanded mCA identified 63 independent loci. **b.** Quantile-quantile plot of the whole blood TWAS of the expanded mCA GWAS using 670 samples from GTExv8 shows enrichment across 62 genes. The horizontal dotted line reflects the Bonferroni-adjusted p-value for significance. Genes with TWAS $P<5\times10^{-8}$ or those important in the pathway-enrichment analyses from Extended Data Figure 10 are labeled. **c.** cell-type enrichment results from the Expanded mCA GWAS across immune, brain, cardiovascular (CV), muscle, gastrointestinal (GI), epithelium, and other tissues as annotated using GenoSkyline-Plus annotations. **d.** Zooming in to show the stratified enrichment by specific categories of immune cells and tissues. Across panels C. and D., the vertical dotted lines indicate (1) P=0.05 for suggestive enrichment, and (2) the Bonferroni-adjusted P-value for significant enrichment. GWAS = genome wide association study, TWAS = transcriptome-wide association study, CV = cardiovascular, GI = Gastrointestinal