Common genetic associations between age-related diseases

Handan Melike Dönertaş[1*], Daniel K. Fabian[1], Matías Fuentealba Valenzuela[1,2], Linda Partridge[2,3], Janet M. Thornton[1*]

[1]European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD, UK.
[2]Institute of Healthy Aging, Department of Genetics, Evolution and Environment, University College London, London, UK
[3]Max Planck Institute for Biology of Aging, Cologne, Germany
*Correspondence: donertas.melike@gmail.com; thornton@ebi.ac.uk

List of Supplementary Tables

Table S1: List of diseases in each age-of-onset cluster.

Table S2: Partially genetic causal relationship between diseases for all significant pairs (FDR corrected $p \leq 0.01$ and mean GCP>0.6).

Table S3: Enrichment results comparing GWAS hits for the UKBB diseases with the GWAS-Catalog traits.

Table S4: List of multidisease and multicategory genes associated with each cluster or cluster combinations.

Table S5: List of genes overlapping between aging databases and the multidisease or multicategory genes associated with each cluster.

Table S6: GO enrichment results.

Table S7: List of drugs specifically targeting multicategory clusters 1, 2, or '1 & 2' genes.

Table S8: Summary of Fisher's exact test result, testing the agonist to antagonist ratio within vs. across clusters.

Table S9: List of antagonistic associations between cluster 1 and cluster 2 diseases and the corresponding genes. 'Median RAF Difference' corresponds to the median value of the risk allele frequencies (cluster1 minus cluster2) for each antagonistic SNP.

Supplementary Information

UK Biobank Data

Using the samples that passed quality control (see Methods, n = 484,598), we first did an exploratory analysis using the basic demographics, disease data, and aging-related data fields.

There were more females (n=262,758) than males (n = 221,840) (Figure S1a). The age range of participants during the first visit was between 37 (minimum age of males = 37, females = 39) to 73 (maximum age of males = 73, females = 71) with a median value of 58 (median age of males = 58, females = 57) (Figure S1b). There were 13,697 participants who died after participating in the study and the death rate was higher in males (Figure S1c). As expected, height, weight, and BMI also differed in females and males (Figure S1d-f).

Participants were also asked how they rate their health, how satisfied they are with their health, smoking status, alcohol drinker status, if other people generally say they look i) younger than they are, ii) about their age, or iii) older than they are, and if they had a close relative who had non-accidental sudden death. Overall, more people rated their health high and were happy with their health (Figure S2a-b). Most of the UKBB participants either never smoked or were previous smokers (Figure S2c) and are current alcohol drinkers (Figure S2d). Most of the participants also reported that people generally think they are either younger than their age or about the same age (Figure S2e). Most of the participants did not have any close relatives who died suddenly from non-accidental causes (Figure S2f).

We also checked the distribution of other aging-related fields, namely parents' age at death, age at menarche, and age at menopause. There were 391,842 participants with at least one parent idead. The distribution was wide (10 to 117), but the majority of the data (between the first and third quantiles) lie between 65 and 79.5 (average age at death) (Figure S3a). The age at menarche differed between 5 and 25, with a median of 13 (Figure S3b). The age at menopause differed between 18 and 68, with a median of 50 (Figure S3c).

The number of self-reported operations ranged between 1 to 32, with a median of 1 and the number of self-reported medications ranged between 0 to 48, with a median of 2 (Figure S4a). Among 39,910 participants with cancer, most of them had only one cancer, while there was also a participant with 6 cancers (Figure S4b).

We then checked the correlations between these traits (Figure S5). Age when attended assessment center was very strongly correlated with age at death. It also showed a correlation with parental age at death, the number of non-cancer diseases, and the number of medications taken. "Overall health rating" and "health satisfaction" were also correlated with the number of diseases and medications. Moreover, these values also showed a correlation with BMI and weight. While 'sex' and 'standing height' were correlated with 'weight', they were not correlated with 'BMI' and 'overall health rating' which are both correlated with 'weight'. BMI was also correlated with 'number of medications taken'.

Although we did not use these traits directly in our analysis, we performed an exploratory analysis to decide on the potential covariates to use in the GWAS model (see Methods).

Disease Categories

The UKBB includes disease information from two sources: i) disease ICD10/9 codes based on hospital episode statistics (HES) and ii) the self-reported (SR) diseases. We have a particular interest in the self-reported diseases as the participants also report the age at diagnosis. Since they report the diseases they had at earlier ages as well, this data is less biased by the age distribution of UK Biobank participants. Moreover, a previous study using the UKBB suggested that GWAS using self-reported diseases and ICD-10 codes were sufficiently similar[1].

Like ICD10 codes, the UKBB SR Diseases are defined in a hierarchical structure (Figure S6). This tree is constructed by the UKBB nurses and it mostly reflects the system or the tissue in which that disease is most symptomatic in. Participants enter SR disease data with a trained nurse, who guides them. However, we some participants did not consider the disease hierarchy while some did. For example, some patients having 'essential hypertension' also reported having 'hypertension' which is the parent node, while some did not. In order not to bias data, we propagated disease data towards upper levels, so that a participant with a disease at a lower level is always annotated with the connected nodes at upper levels.

Importantly, we only considered 116 non-cancer diseases with at least 2,000 cases and that were not sex-specific. Although we exclude sex-limited diseases, we included the ones that were more prevalent in females (thyroid problem, hypothyroidism, bone disorder, osteoporosis) or in males (abdominal hernia, gout, heart attack) (Figure S7).

Disease Co-occurrences

We next calculated disease co-occurrences, using relative risk score to calculate associations and ϕ values as a measure of robustness[2–5] (Methods). There were five major clusters with high relative risk scores and robustness and they seemingly cluster by disease categories: i) musculoskeletal/trauma diseases and early-onset gastrointestinal diseases such as appendicitis, ii) other musculoskeletal/trauma diseases such as sciatica and disc problems, iii) respiratory/ENT diseases such as bronchitis and pneumonia, iv) cardiovascular diseases and diabetes, v) retinal problem, glaucoma, and cataract (Figure S8). While most of these clusters are biologically plausible, some could be explained by reporting bias, e.g. it is plausible that only a fraction of people reported their childhood diseases, resulting in an artificial association between bone fractures and appendicitis. Moreover, we saw a strong negative correlation between osteoarthritis and arthritis (nos). The disease 'arthritis (nos)' does not include osteoarthritis by definition (nos = not osteoarthritis) and seeing this association suggests that we can detect co-occurrences reliably.

1. Cortes, A., Dendrou, C., Fugger, L. & McVean, G. Systematic classification of shared components of genetic risk for common human diseases. *bioRxiv* (2018).
2. Gutiérrez-Sacristán, A. *et al.* comoRbidity: an R package for the systematic analysis of disease comorbidities. *Bioinformatics* **34**, 3228–3230 (2018).
3. Jiang, Y., Ma, S., Shia, B.-C. & Lee, T.-S. An Epidemiological Human Disease Network Derived from Disease Co-occurrence in Taiwan. *Sci. Rep.* **8**, 4557 (2018).
4. Park, J., Lee, D.-S., Christakis, N. A. & Barabási, A.-L. The impact of cellular networks on disease comorbidity. *Mol. Syst. Biol.* **5**, 262 (2009).
5. Sanchez-Valle, J. *et al.* Unveiling the molecular basis of disease co-occurrence: towards personalized comorbidity profiles. *bioRxiv* 431312 (2018) doi:10.1101/431312.

Supplementary Figures

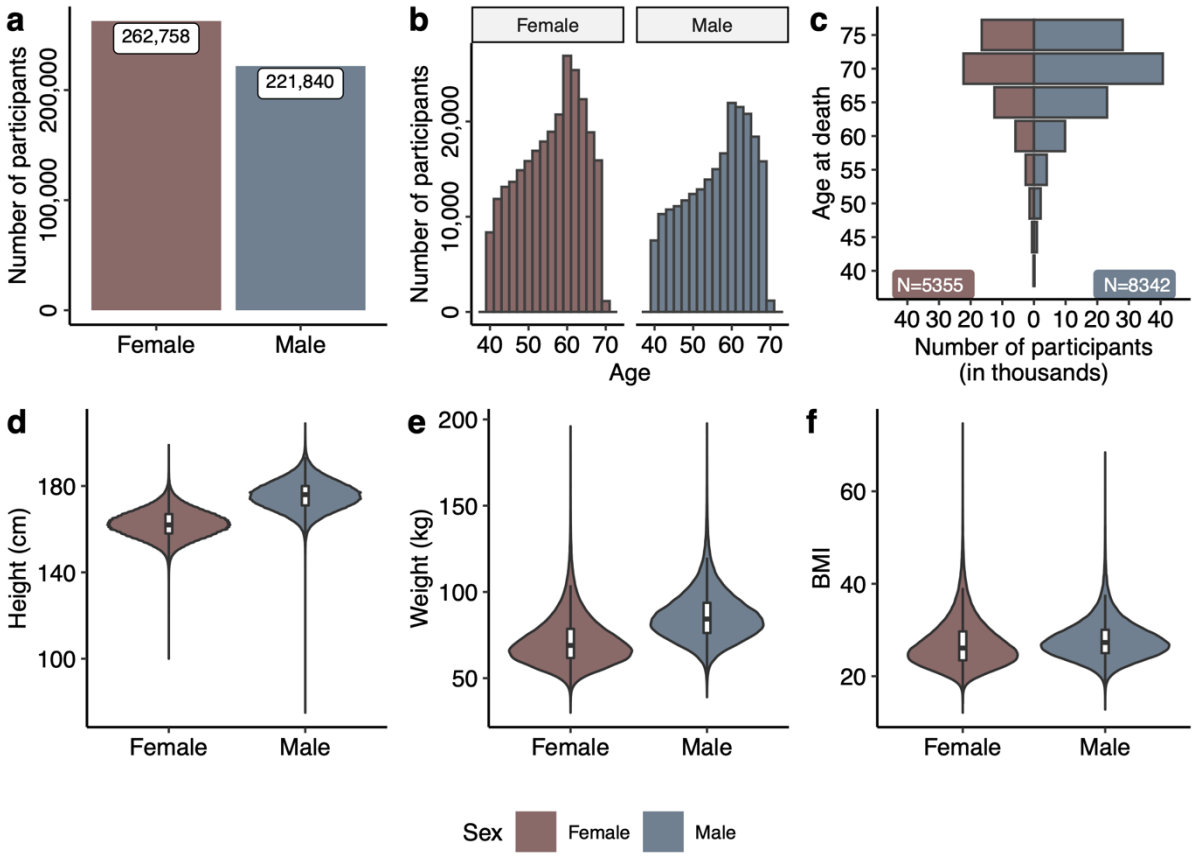

Figure S1: Participant data in UK Biobank after quality control steps. a) The number of female and male participants, b) Age distribution when participants first attended the UKBB assessment center and answered self-reported questions, c) Age at death (every 5 years are binned together) for the participants who died after attending the UKBB assessment center. The values are corrected for the number of female and male participants who passed the ages specified in the y-axis, d) Distributions of 'standing height' field in the UKBB, e) Distributions of 'weight' field in the UKBB, f) Distributions of BMI field calculated using 'standing height' and 'weight' fields in the UKBB (see Methods).

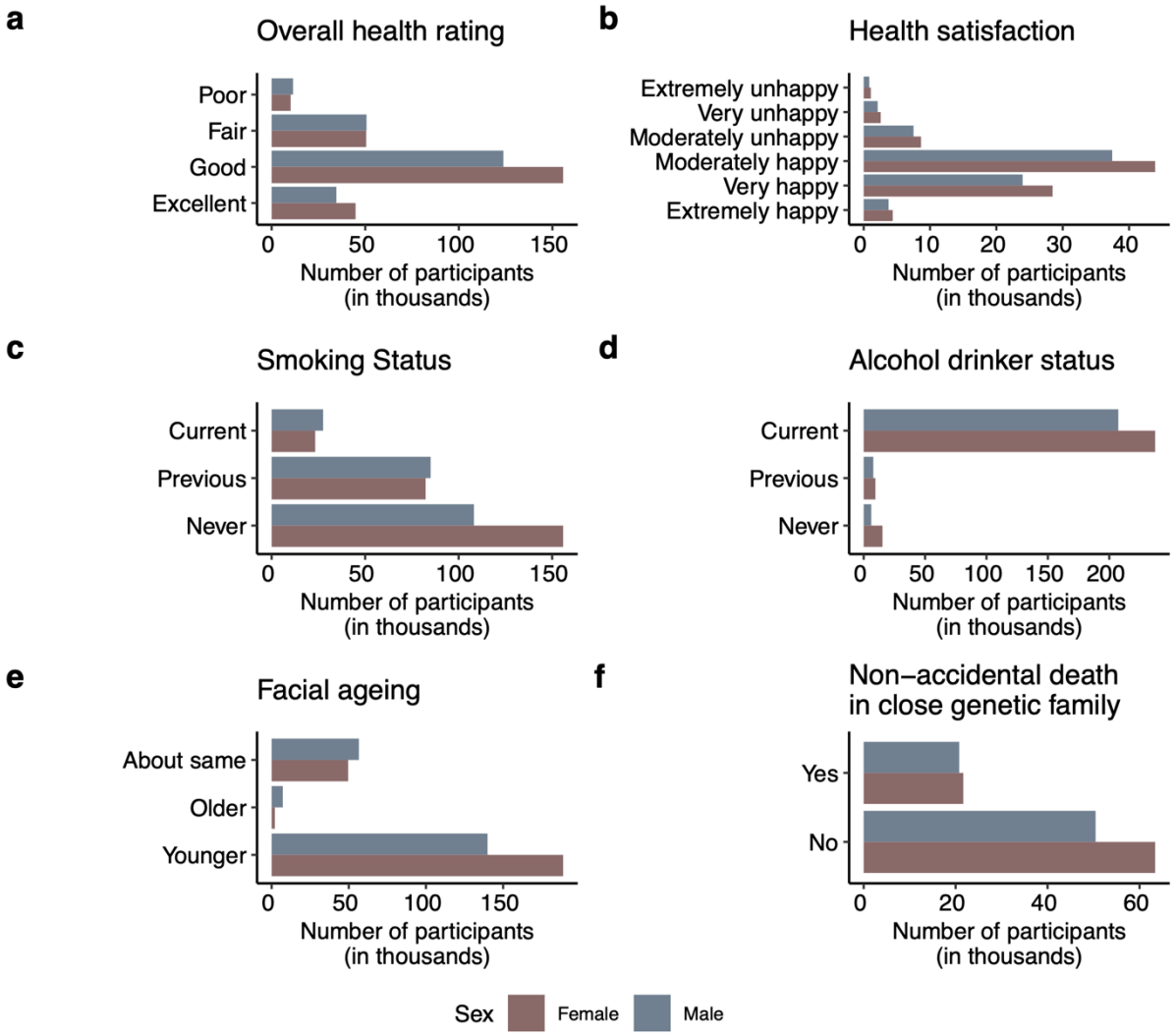

Figure S2: The distribution of a) Overall health rating, b) Health satisfaction, c) Smoking status, d) Alcohol drinker status, e) Facial aging, f) Non-accidental death in close genetic family fields in the UKBB. x-axes show the number of participants, while y-axes are the answers given by the participants.

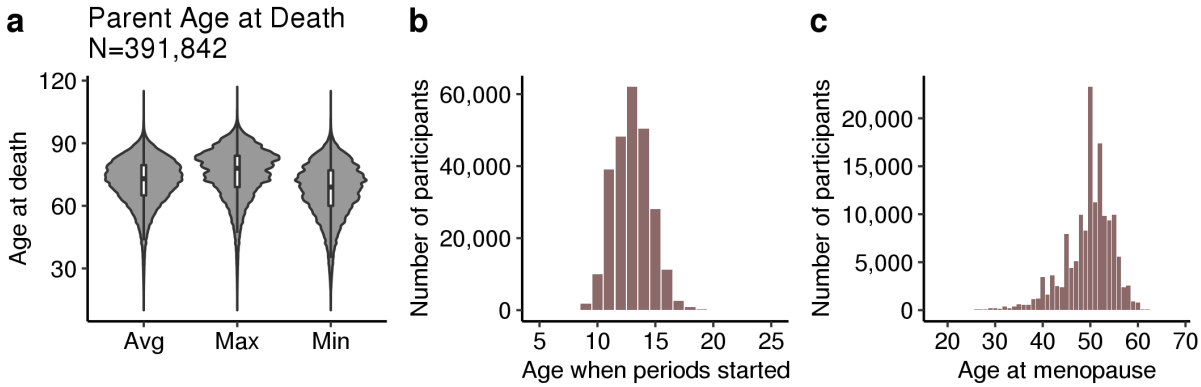

Figure S3: Distributions of a) parents' age at death, b) age when periods started (menarche), and c) Age at menopause (last menstrual period).

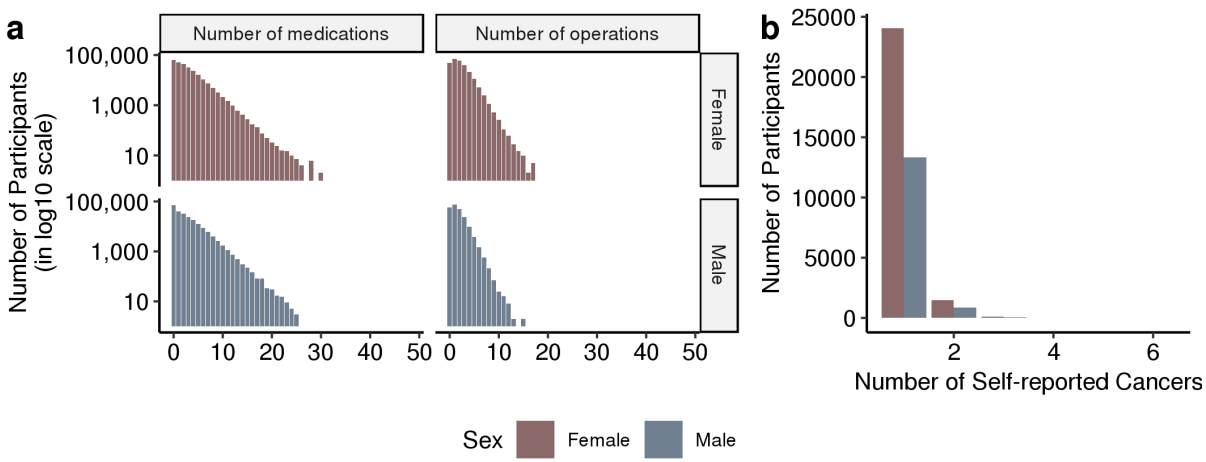

Figure S4: Self-reported health data. a) The number of self-reported medications and operations (x-axes) for the participants in the UK Biobank. The y-axis shows the number of participants on a log10 scale. b) The number of self-reported cancers (x-axis). Y-axis shows the number of participants.

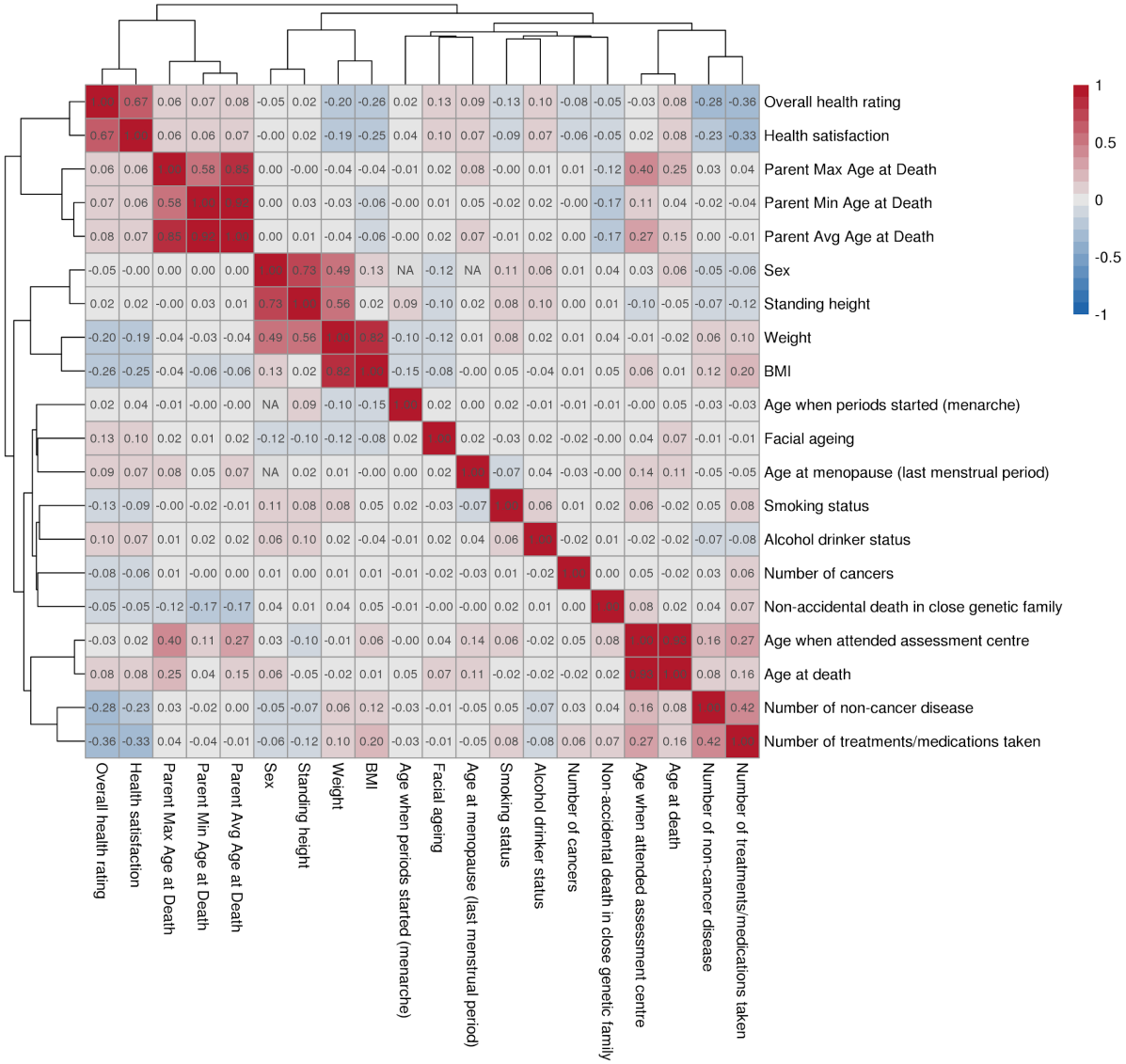

Figure S5: Pairwise correlations between traits. Each row and column shows a trait in the UKBB, and the color shows the pairwise Spearman correlation coefficients between traits. Dark red denotes a strong positive correlation, while dark blue indicates strong negative values. Traits are ordered based on the hierarchical clustering of the correlation coefficients.

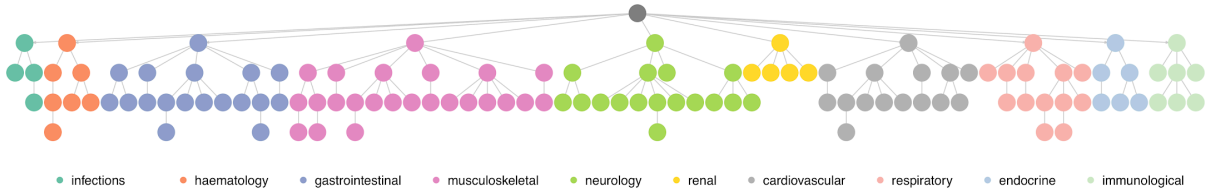

Figure S6: Disease hierarchy for the 116 diseases included in the analysis. The nodes are colored by the disease categories as indicated in the legend.

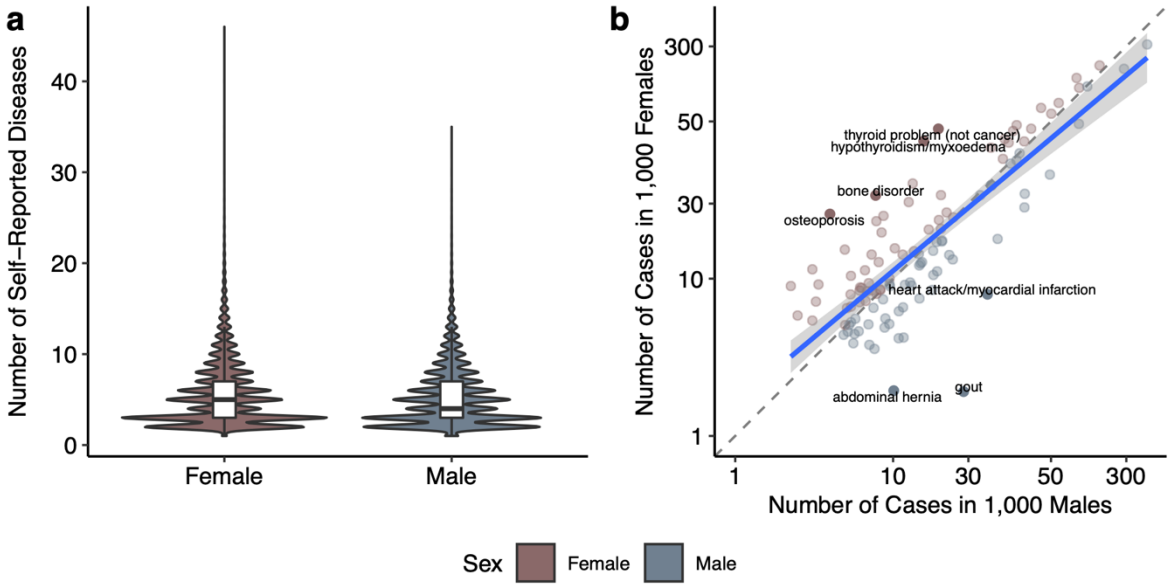

Figure S7: Sex-stratified statistics for 116 selected diseases. a) The distribution of the number of self-reported diseases (y-axis) stratified by sex (x-axis). b) The distribution of disease prevalence in males and females. The x- and y-axes show the number of cases in 1,000 males and females (on a log scale), respectively. The color of each point denotes diseases with a higher prevalence in females (rosy brown, above the dashed line) or males (slate grey, below the dashed line). The linear regression line is depicted as blue. Diseases having a residual value bigger than 3 standard deviations are labeled but not excluded as they are also common in the other sex.

Figure S8: Disease association matrix summarizing relative risk scores and correlations. Each row and column denote diseases, ordered by hierarchical clustering of risk scores. The color is defined by relative risk scores while the size is determined by ϕ value, indicating the robustness of the association (see Methods). The diagonal tiles are colored by the UK Biobank's disease hierarchy to visualize if diseases from the same category cluster together. Associations for the 62 diseases that have at least one relative risk ratio higher than four ($log_2 RR \geq 2$) or lower than minus four ($log_2 RR \leq -2$) are plotted.

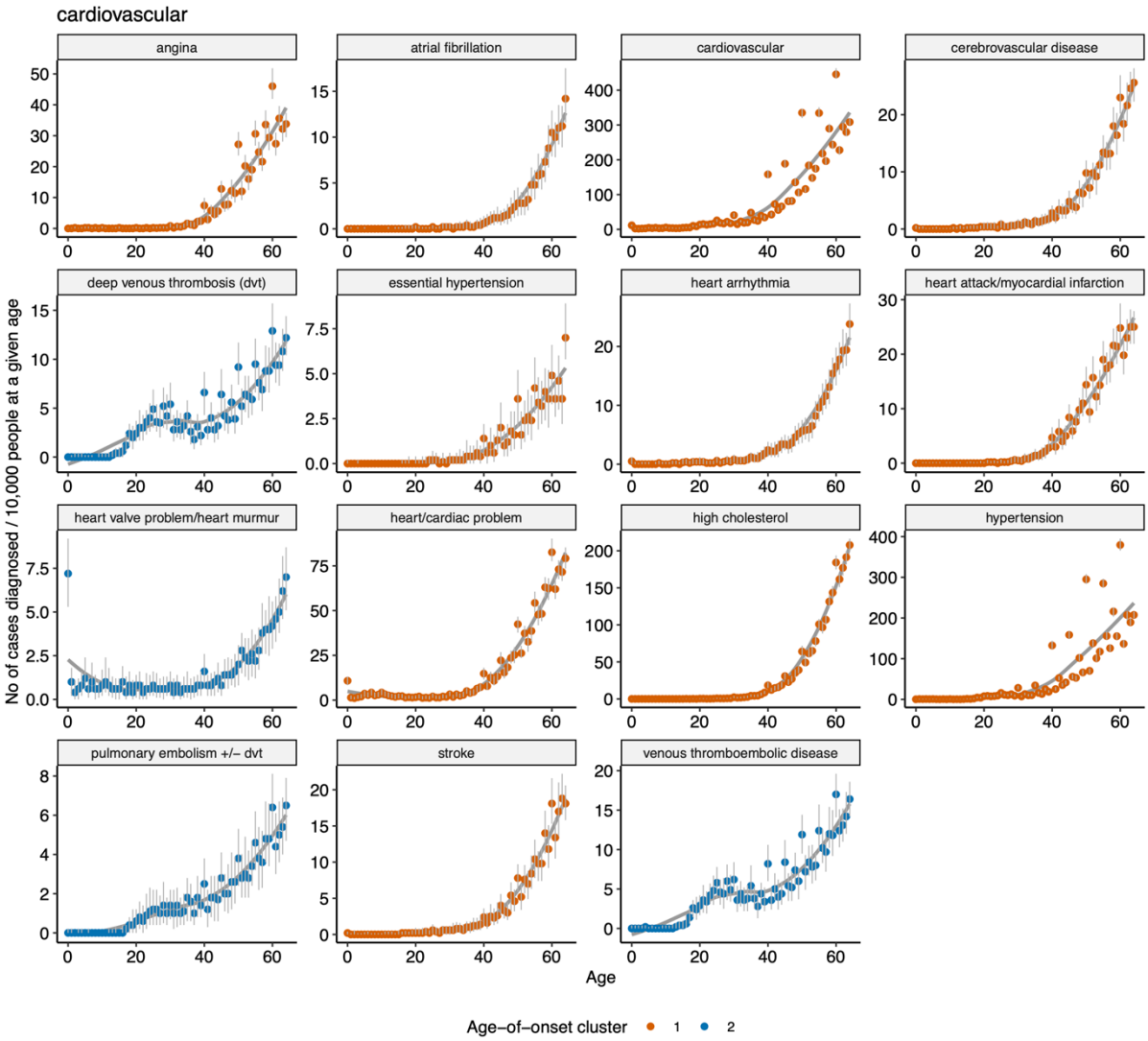

Figure S9: Age-of-onset distributions for the cardiovascular diseases. The y-axis shows how many people in 10,000 are diagnosed with that disease at a certain age (x-axis). The plots are also normalized by the number of people that are older than a given age so that it is unaffected by the distribution of ages in the UKBB (Figure S1b). We ran permutations to define confidence intervals for the disease onset rates. We thus down-sampled the UKBB population using 50,000 participants for 100 times and calculated the median (points, colored by the age-of-onset cluster in Figure 2) and 95% range of the all points (gray error lines). A best-fit curve (calculated using loess regression between the medians and age-of-onset) is also displayed.

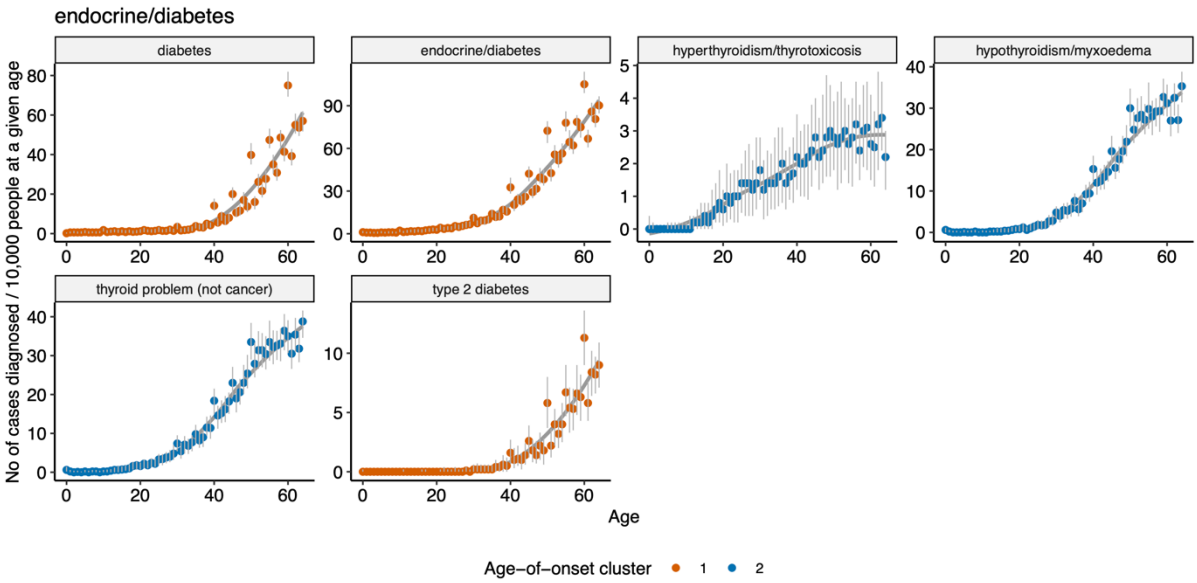

Figure S10: Same as Figure S9, but for endocrine / diabetes diseases.

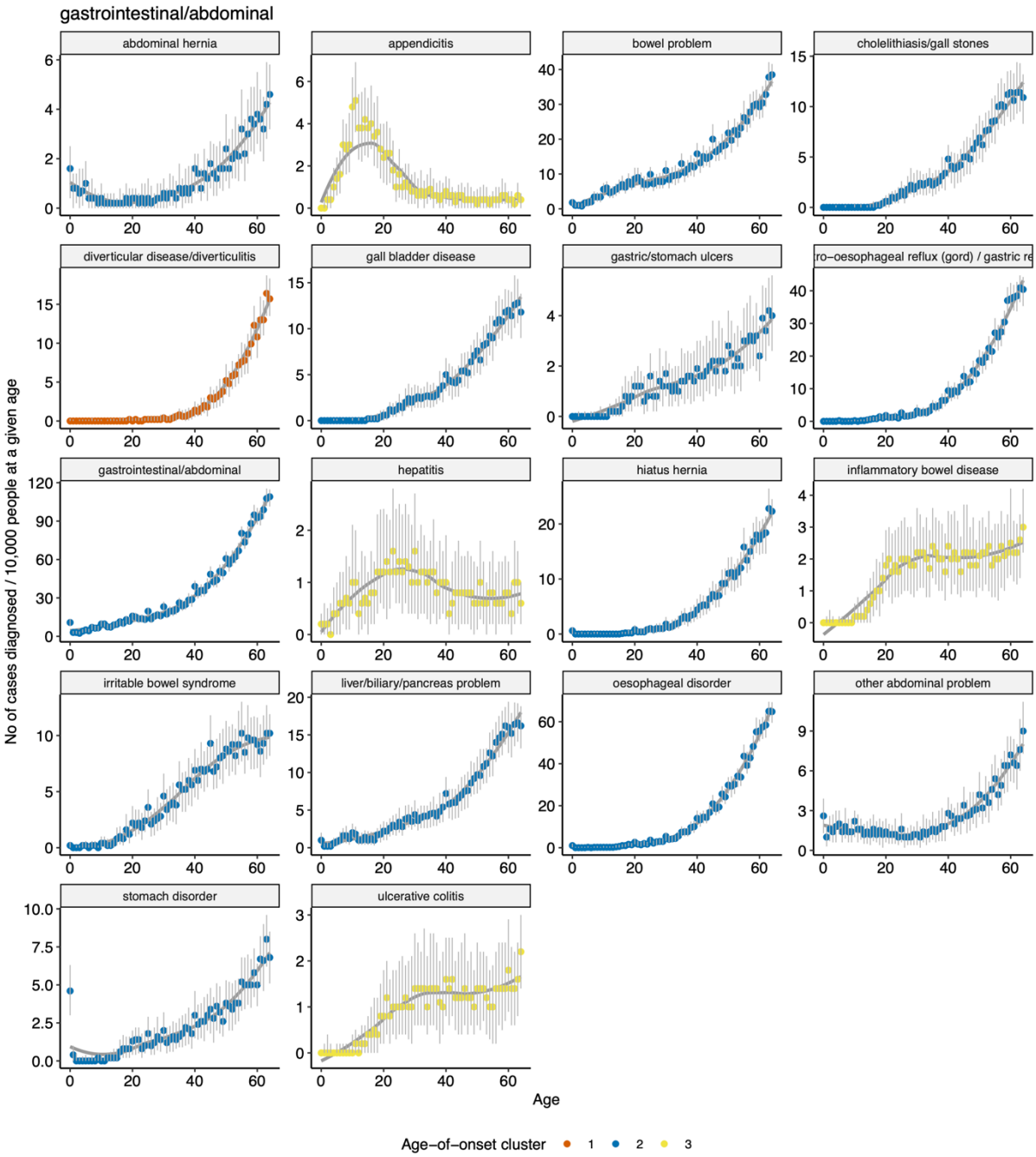

Figure S11: Same as Figure S9, but for gastrointestinal / abdominal diseases.

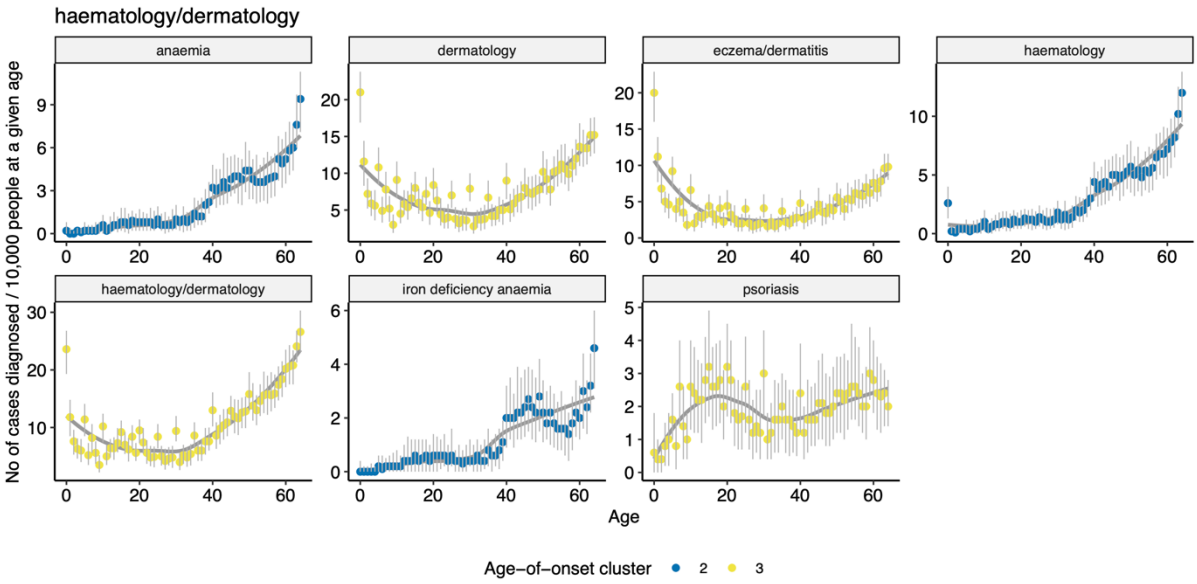

Figure S12: Same as Figure S9, but for haematology / dermatology diseases.

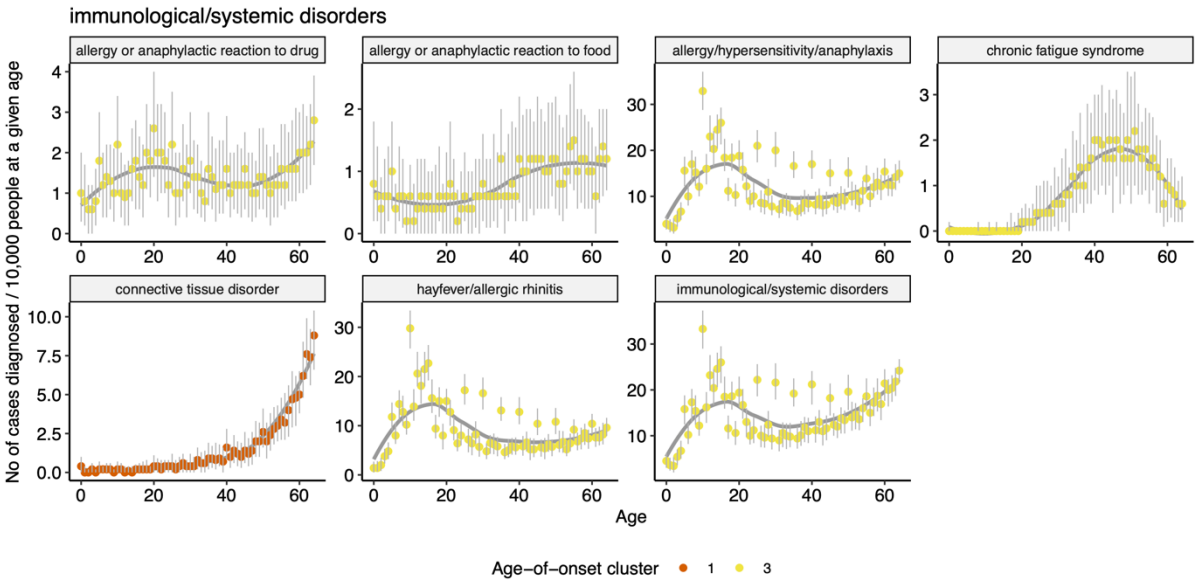

Figure S13: Same as Figure S9, but for immunological / systemic disorders.

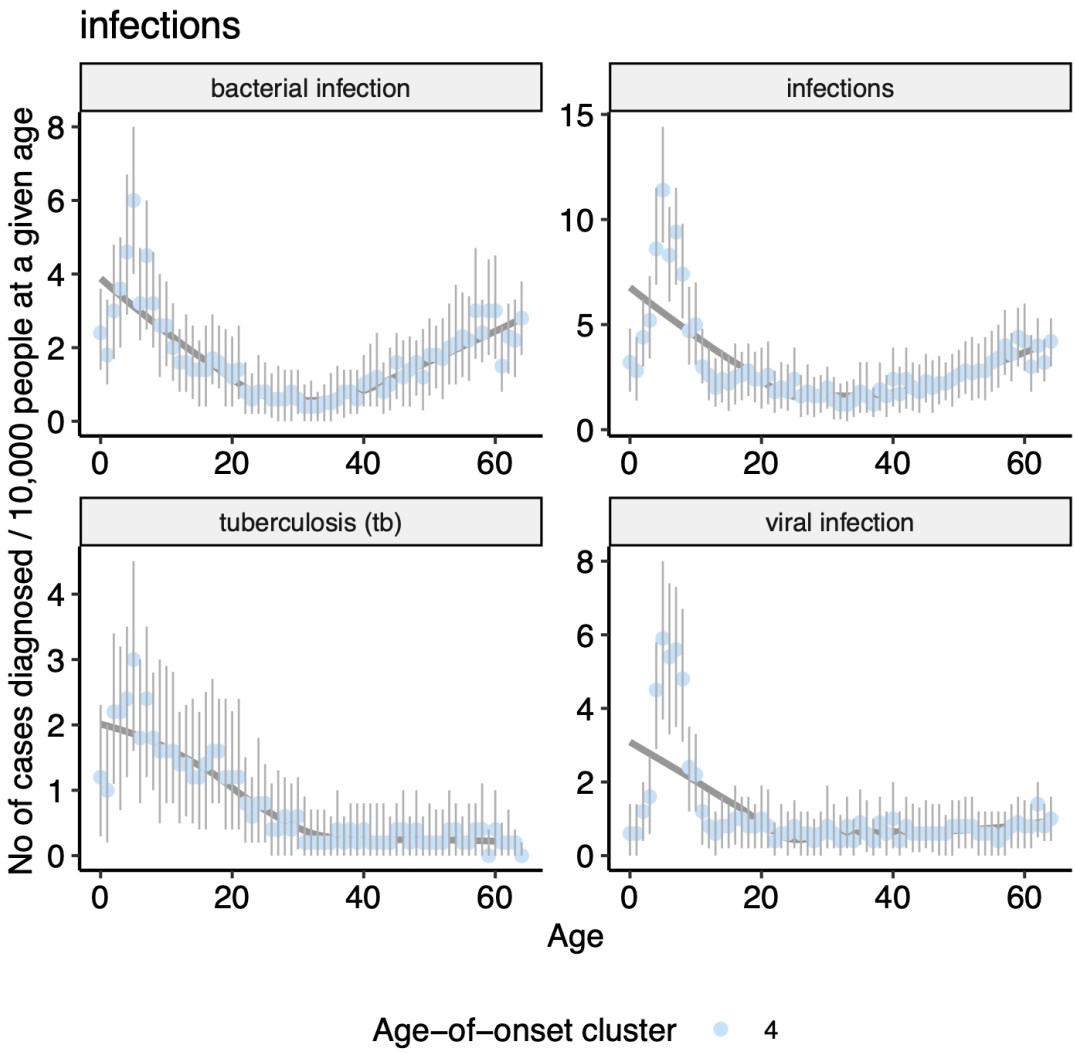

Figure S14: Same as Figure S9, but for infections.

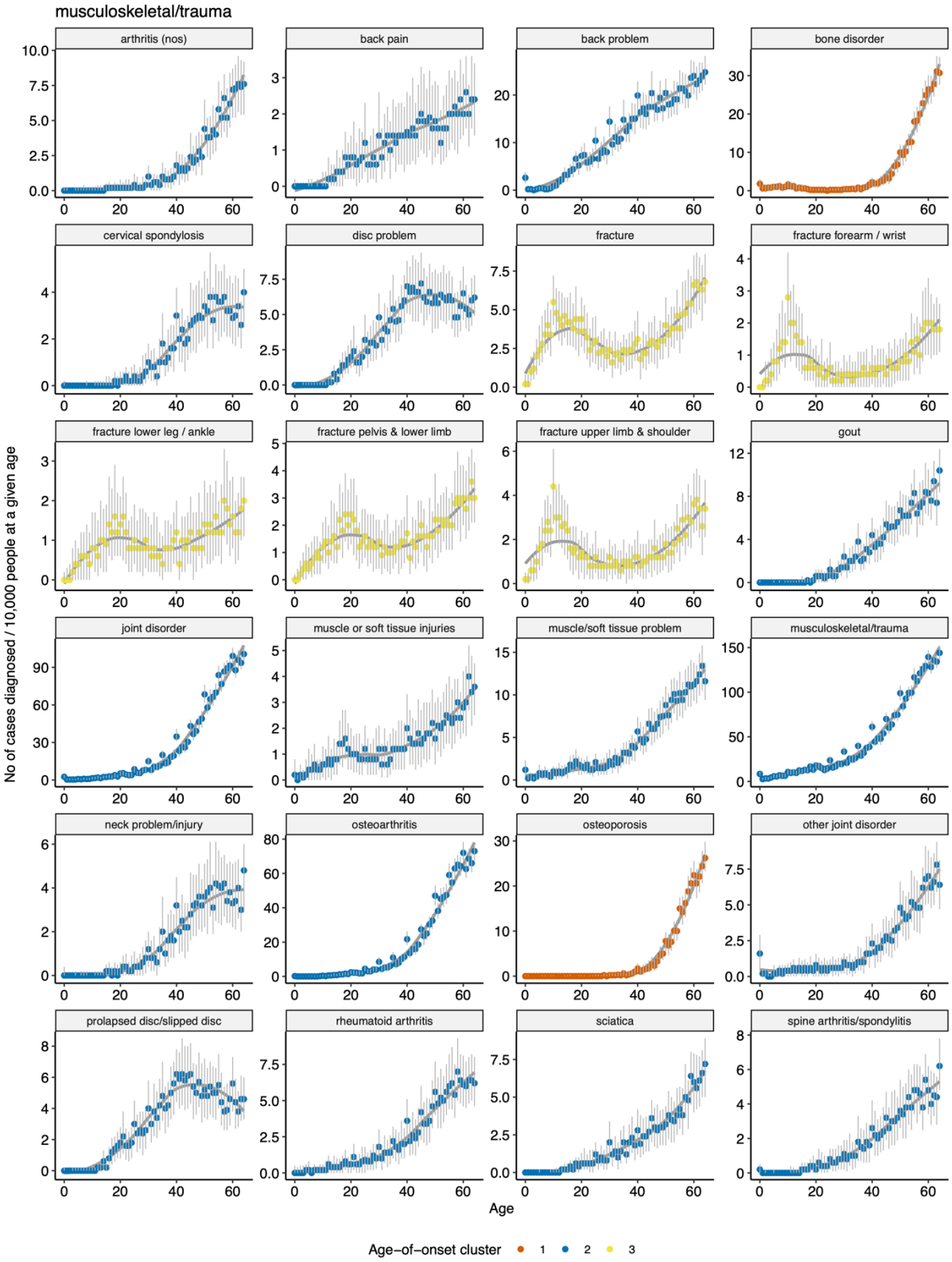

Figure S15: Same as Figure S9, but for musculoskeletal / trauma diseases.

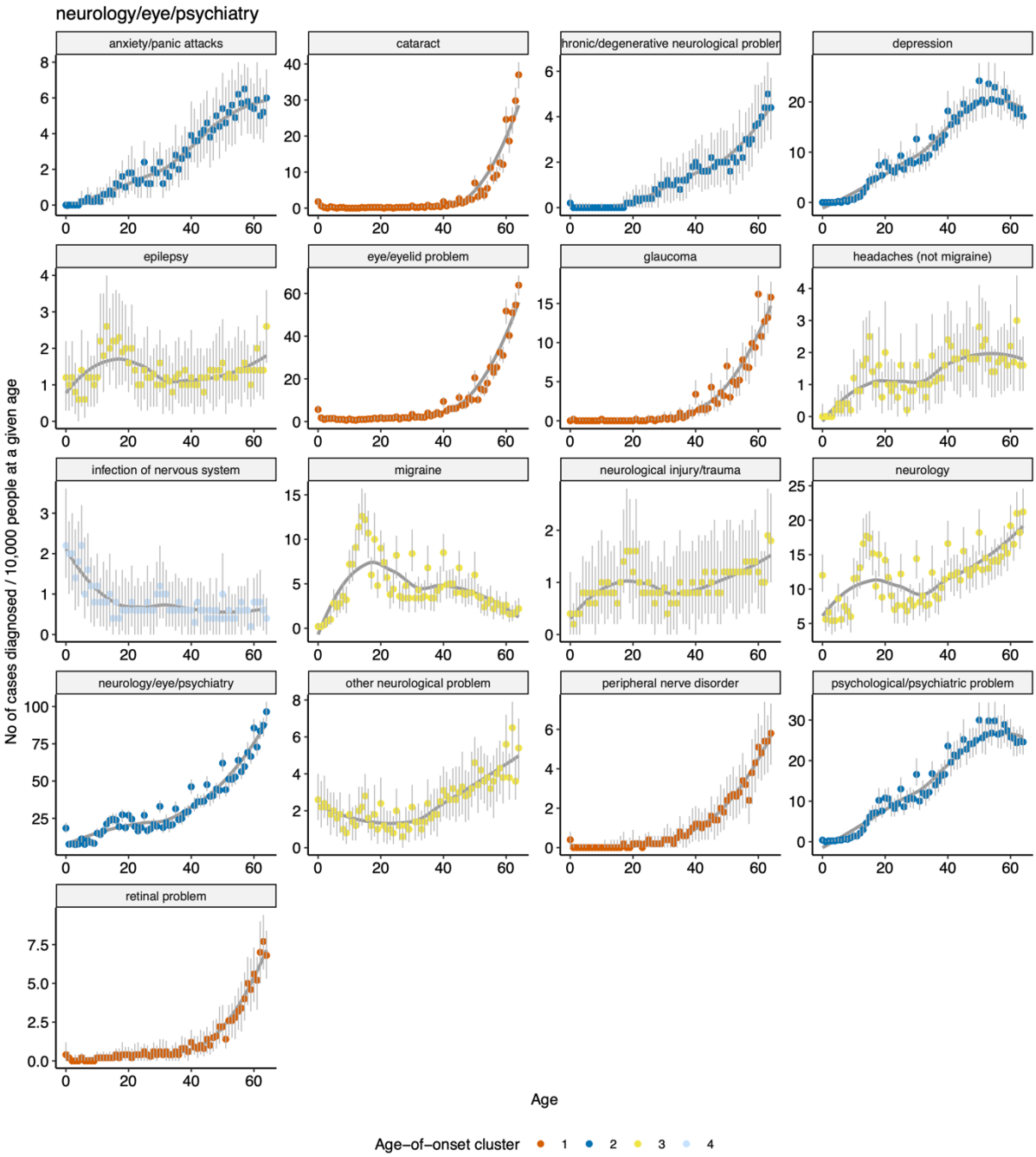

Figure S16: Same as Figure S9, but for neurology / eye / psychiatry diseases.

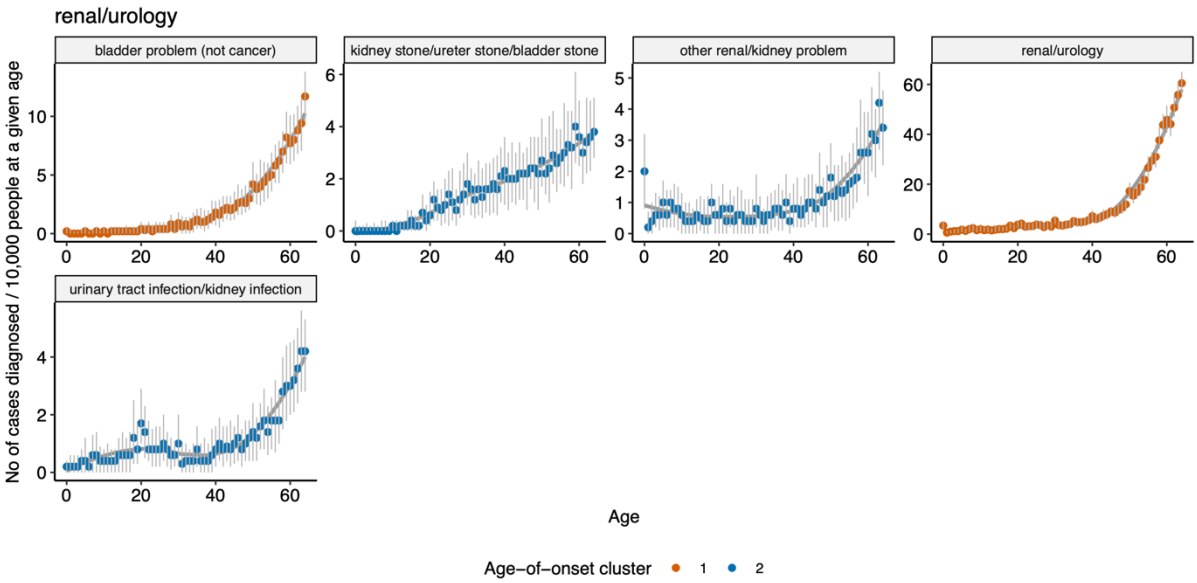

Figure S17: Same as Figure S9, but for renal / urology diseases.

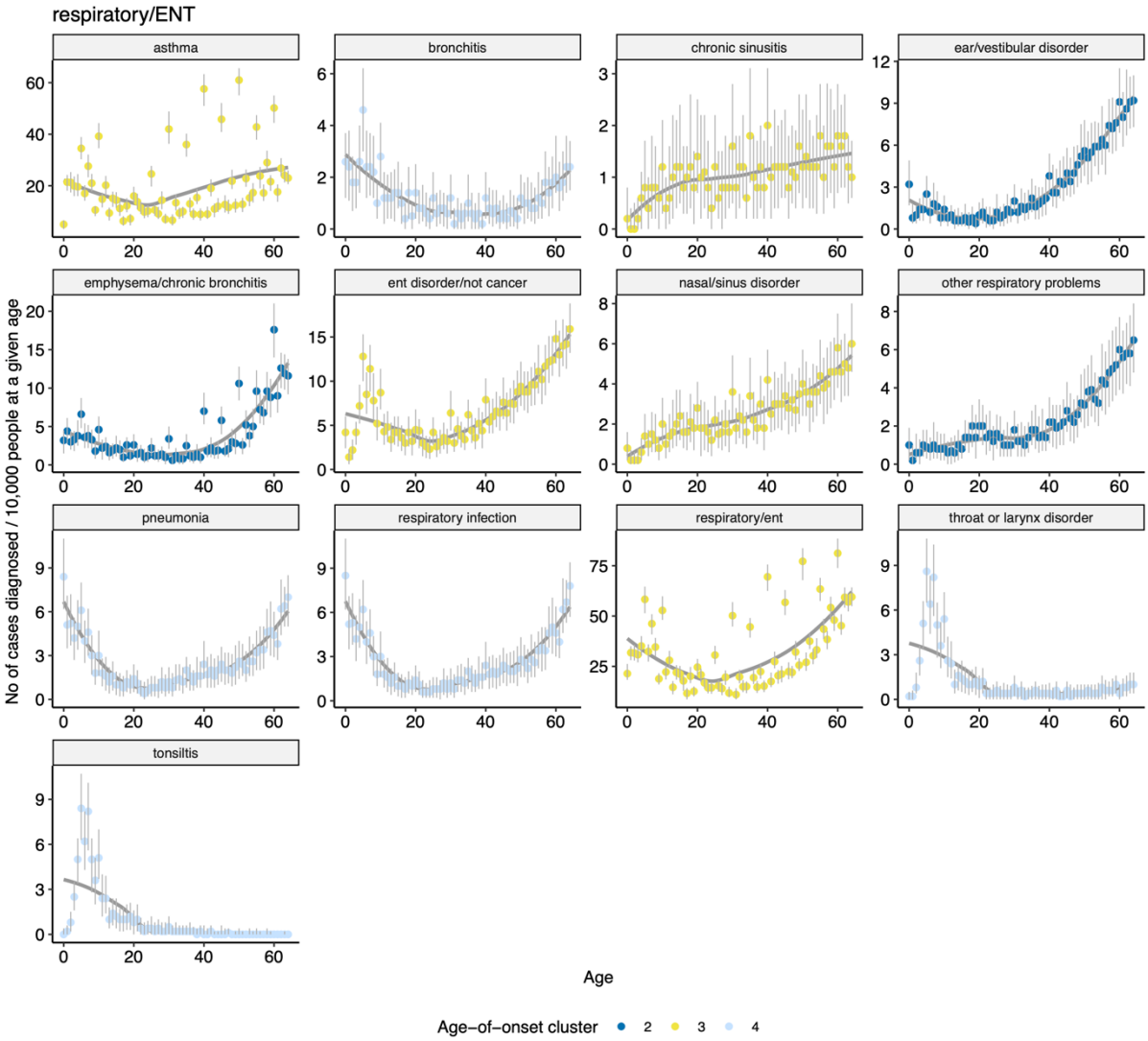

Figure S18: Same as Figure S9, but for respiratory / ENT diseases.

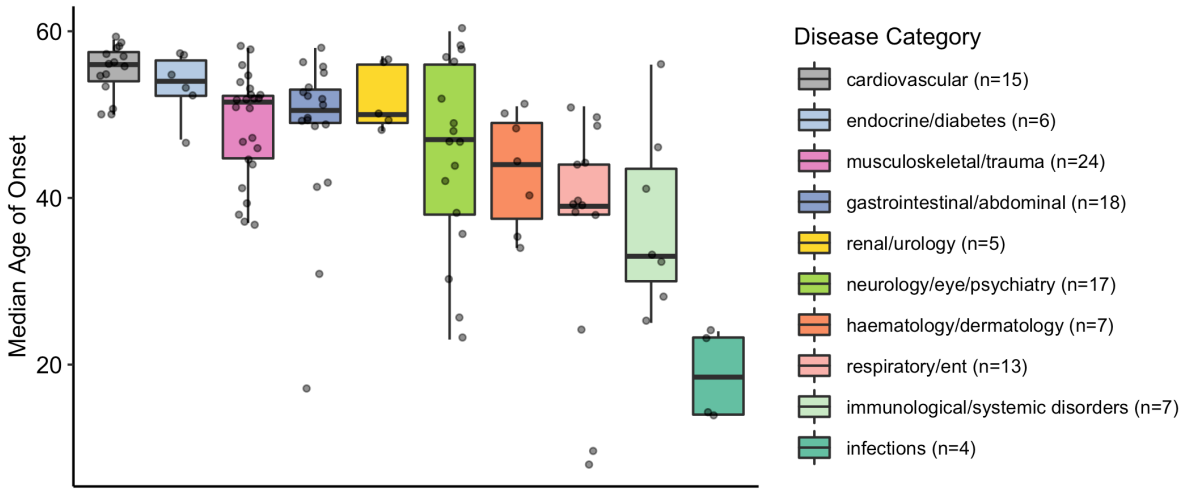

Figure S19: Distribution of median age-of-onset (y-axis) across categories (x-axis). Points show diseases, grouped by the categories (individual boxplots). Categories are ordered by the median value of the median age-of-onset.

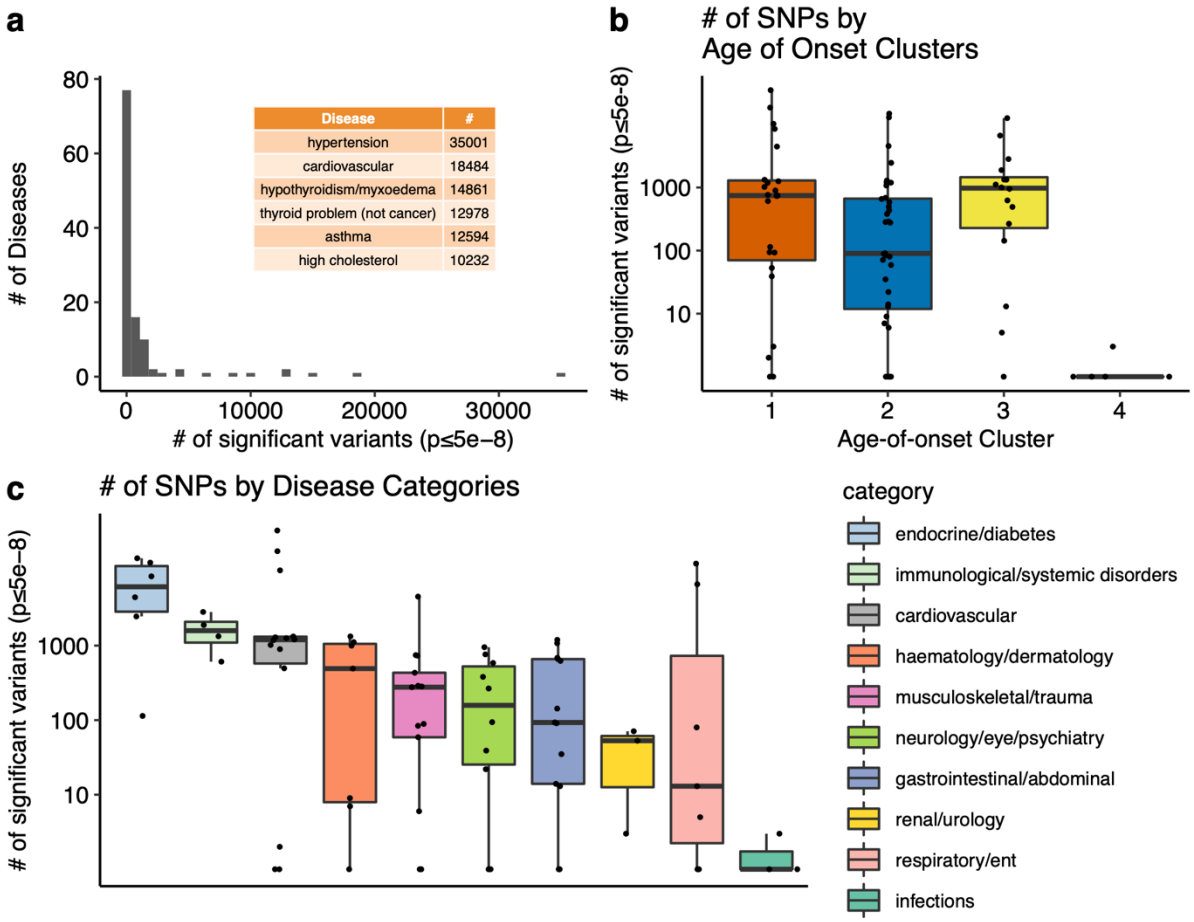

Figure S20: a) Number of diseases for different number of significant variants (p≤5e-8). Diseases with the highest number of associations (N≥10,000) are given as an inset table. b) Comparison of the number of significant associations (y-axis, on a log scale) across age-of-onset clusters (x-axis) (ANOVA after excluding cluster 4, p = 0.06). Since the y-axis is on a log scale, diseases with zero significant associations are not shown on the graph. c) The same as b) but for disease categories. Categories are ordered by the median number of significant SNPs.

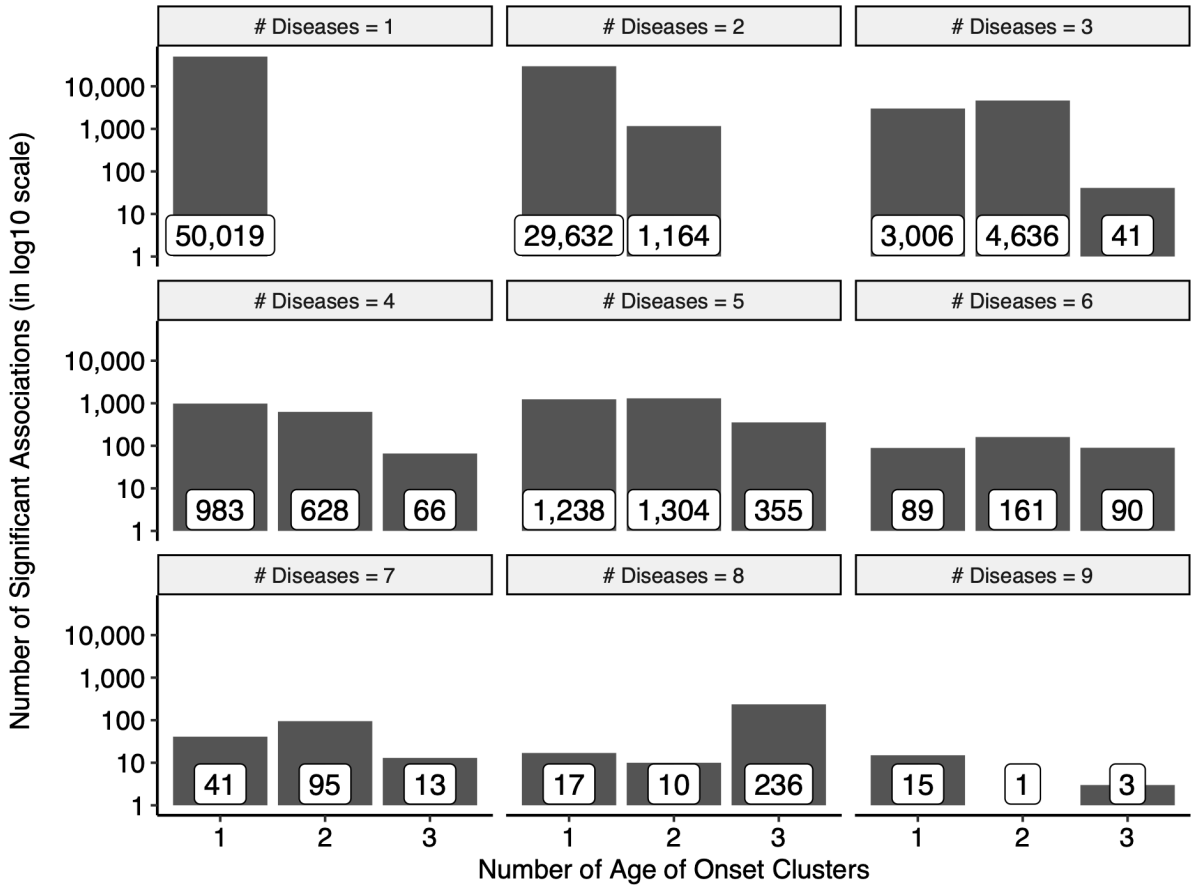

Figure S21: Distributions of the number of significant associations (y-axis) according to the number of diseases associated with a given SNP and the number of age-of-onset clusters (x-axis). For example, the upper left plot indicates that 50,019 polymorphisms are significantly associated with one disease in one age-of-onset cluster, while the lower right plot shows that there are 15, 1, and 3 significant SNPs associated with 9 diseases in one, two, or three age-of-onset clusters, respectively.

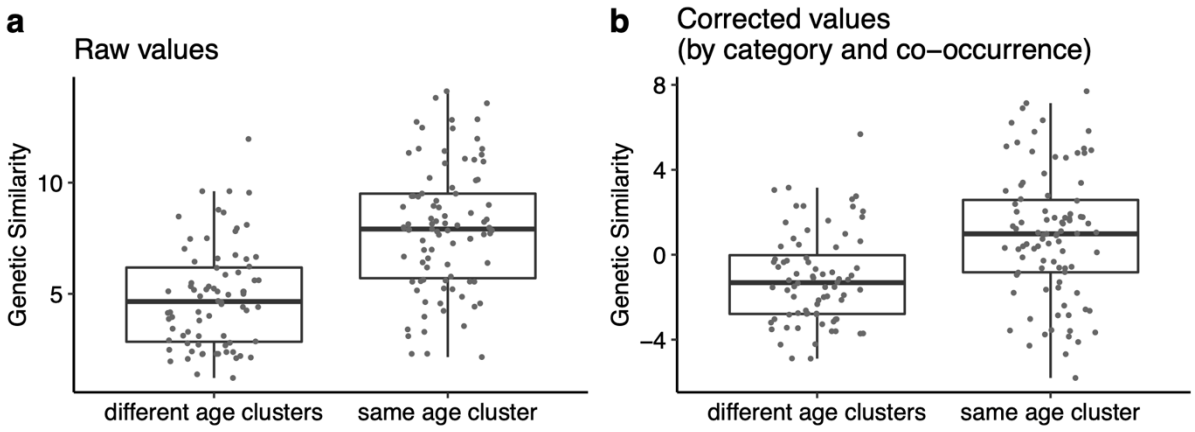

Figure S22: a) The difference between genetic similarity within and across age-of-onset clusters. Y-axis shows the genetic similarity (see Methods). b) The same as a) but the y-axis is corrected for disease category and co-occurrence using a linear model. This panel is the same as Figure 3b and given here only for an easier comparison.

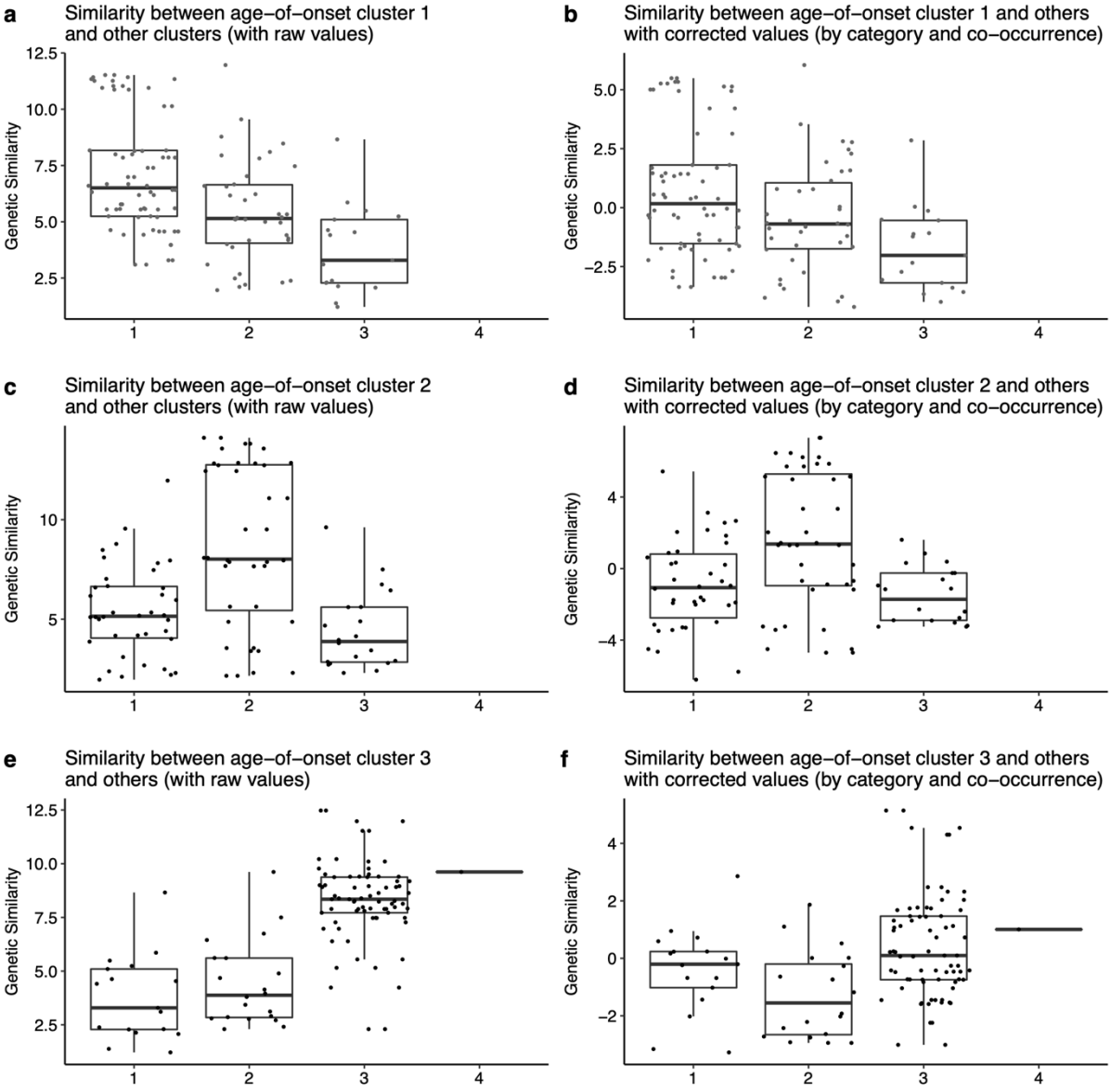

Figure S23: Genetic similarities between cluster 1 (a, b), 2 (c, d), 3 (e, f) and other age-of-onset clusters. The y-axis shows the genetic similarity on a log2 scale as the raw values (a, c, e) or as values corrected for disease category and co-occurrence using a linear model (b, d, f) (see Methods for details).

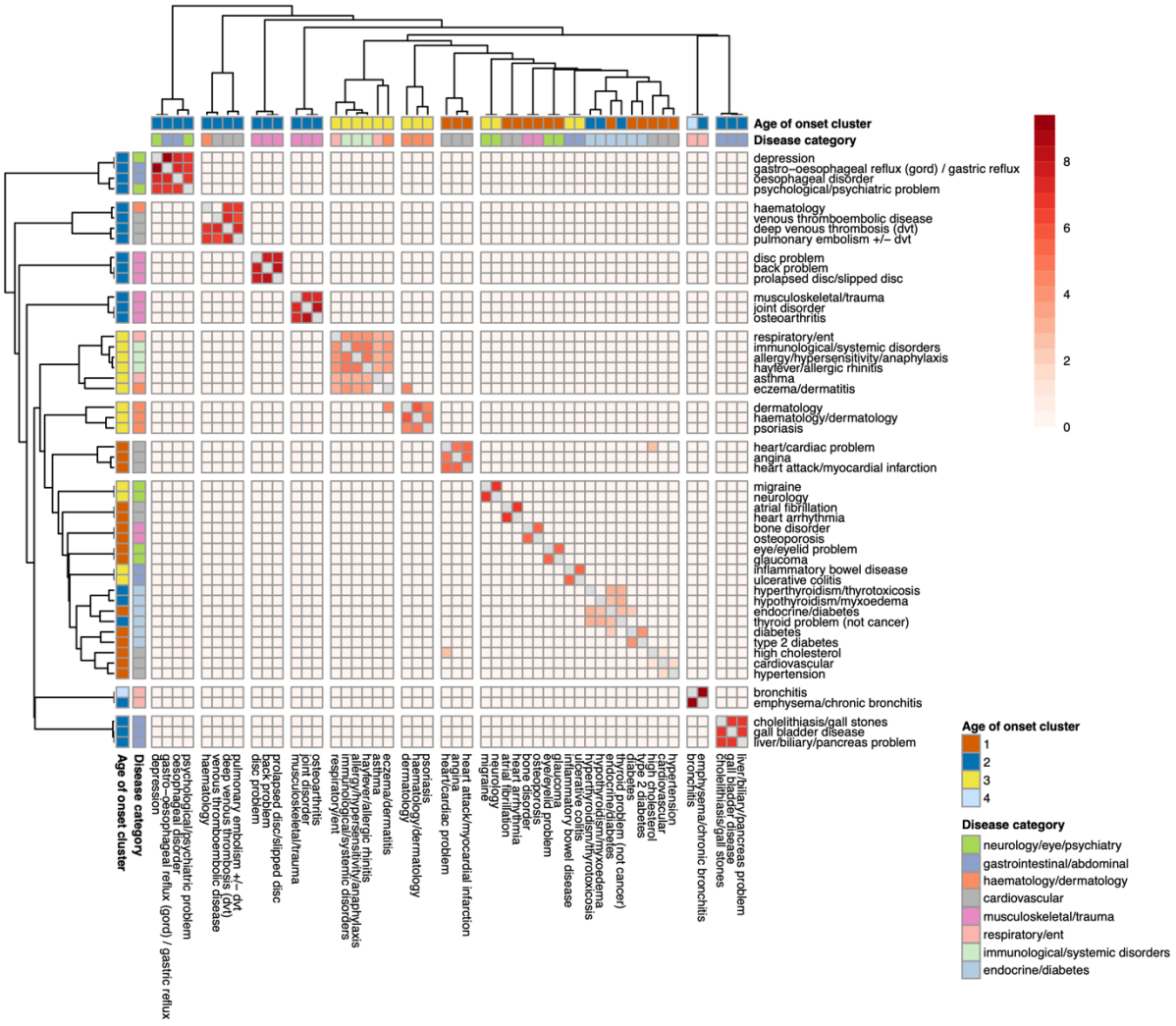

Figure S24: Significant genetic similarities (p≤0.01) calculated using independent LD blocks. Diseases (n=50) with at least one significant genetic similarity are displayed. The color shows the genetic similarity score, darker red means a higher score. Annotation columns show the age-of-onset clusters and disease categories. The diseases are clustered by the hierarchical clustering of genetic similarity scores.

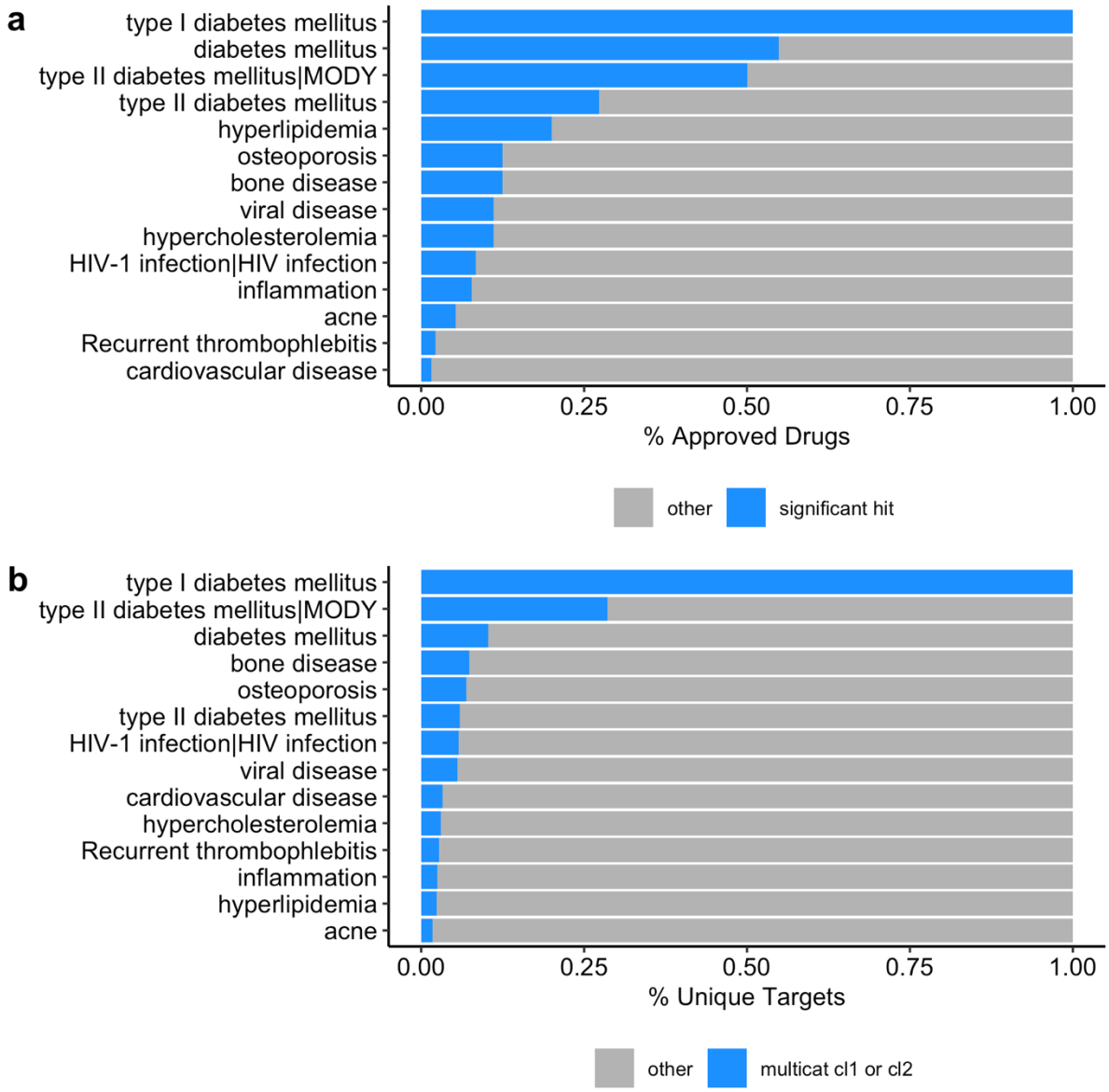

Figure S25: Distribution of a) the drugs and b) their targets approved for 14 conditions treated with the significant hits for drug repurposing . a) X-axis shows the proportion of the significant hits in the drug repurposing study approved for the treatment of the conditions listed on y-axis. b) The same as a) but showing the proportion of unique targets of the approved drugs (x-axis) for the conditions listed on the y-axis.

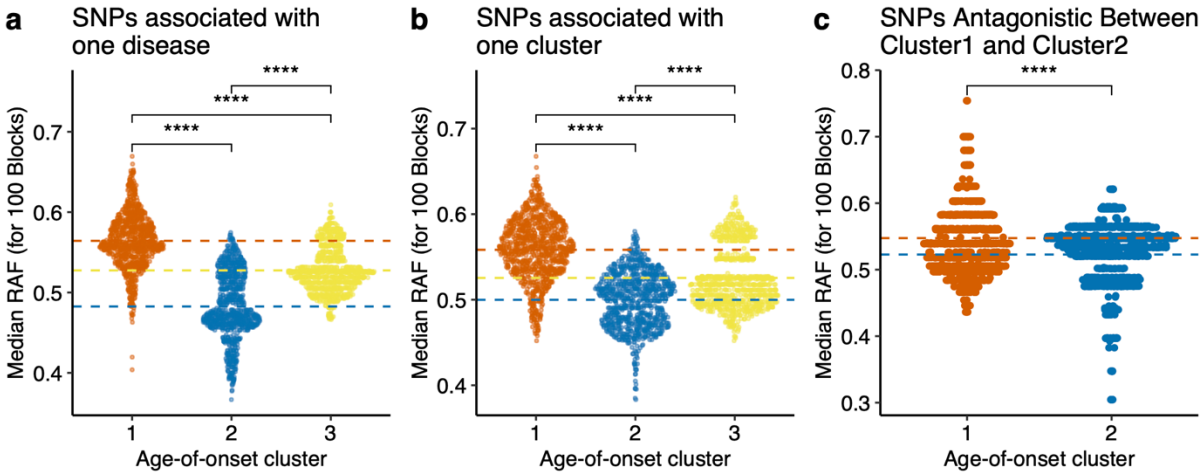

Figure S26: The distribution of Median Risk Allele Frequencies (RAF, y-axis) for 100 randomly sampled LD blocks, for 1,000 times, using variants a) associated with one disease, b) associated with one cluster, c) with antagonistic association between cluster 1 and cluster 2. *ns: p>0.05, *:p≤0.05, **: p≤0.01, ***: p≤0.001, ****: p≤0.00001*

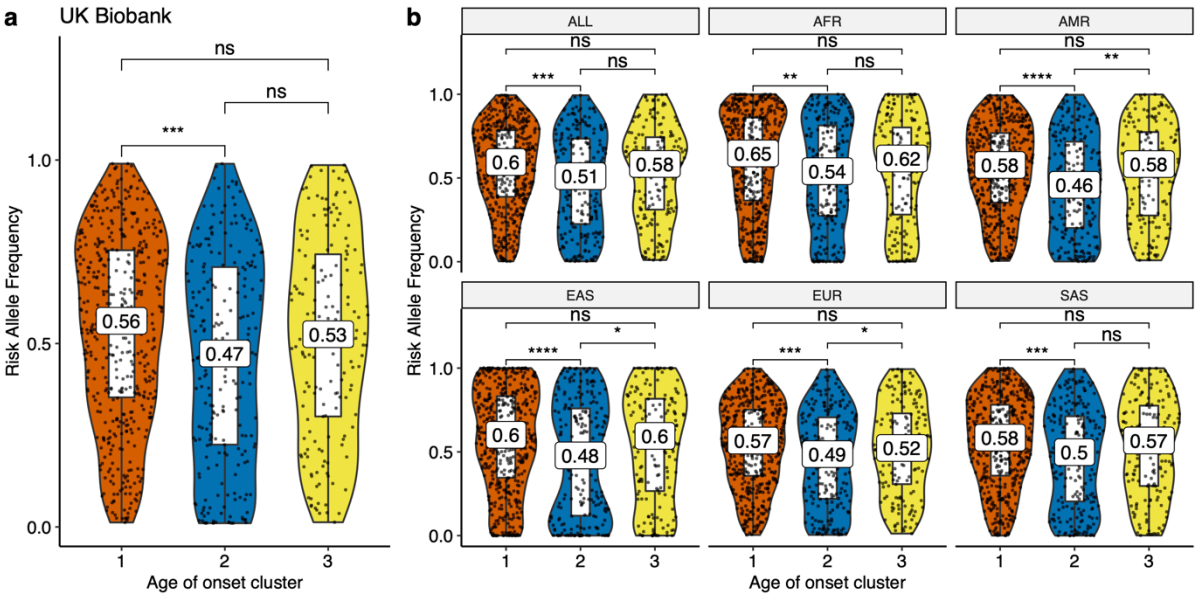

Figure S27: a) Risk allele frequency distributions (y-axis) of different age-of-onset clusters (x-axis) in UK Biobank for SNPs associated with one disease. This plot is the same as Figure 6a, and included here for an easier comparison. b) The same as panel a but for different 1000 Genomes super-populations (ALL: complete 1000 Genomes cohort, AFR: African, AMR: Ad Mixed American, EAS: East Asian, EUR: European, SAS: South Asian). *ns: p>0.05, *:p≤0.05, **: p≤0.01, ***: p≤0.001, ****: p≤0.00001*

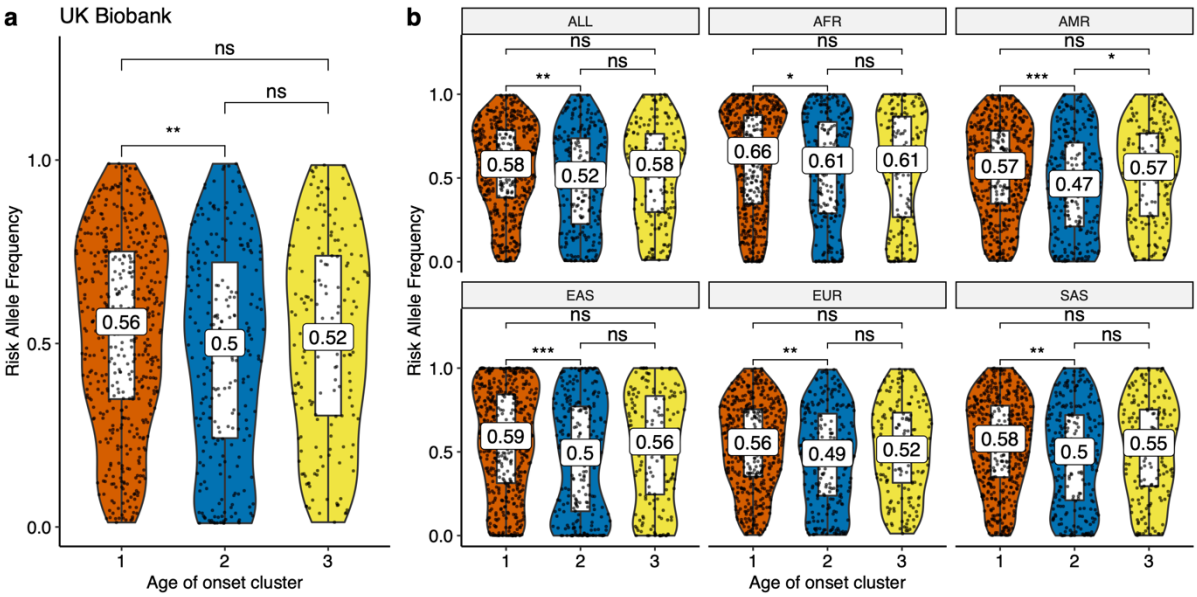

Figure S28: a) Risk allele frequency distributions (y-axis) of different age-of-onset clusters (x-axis) in UK Biobank for SNPs associated with one cluster, excluding antagonistic associations. This plot is the same as Figure 6b, and included here for an easier comparison. b) The same as a) but for different 1000 Genomes super-populations (ALL: complete 1000 Genomes cohort, AFR: African, AMR: Ad Mixed American, EAS: East Asian, EUR: European, SAS: South Asian). *ns: p>0.05, *:p≤0.05, **: p≤0.01, ***: p≤0.001, ****: p≤0.00001*

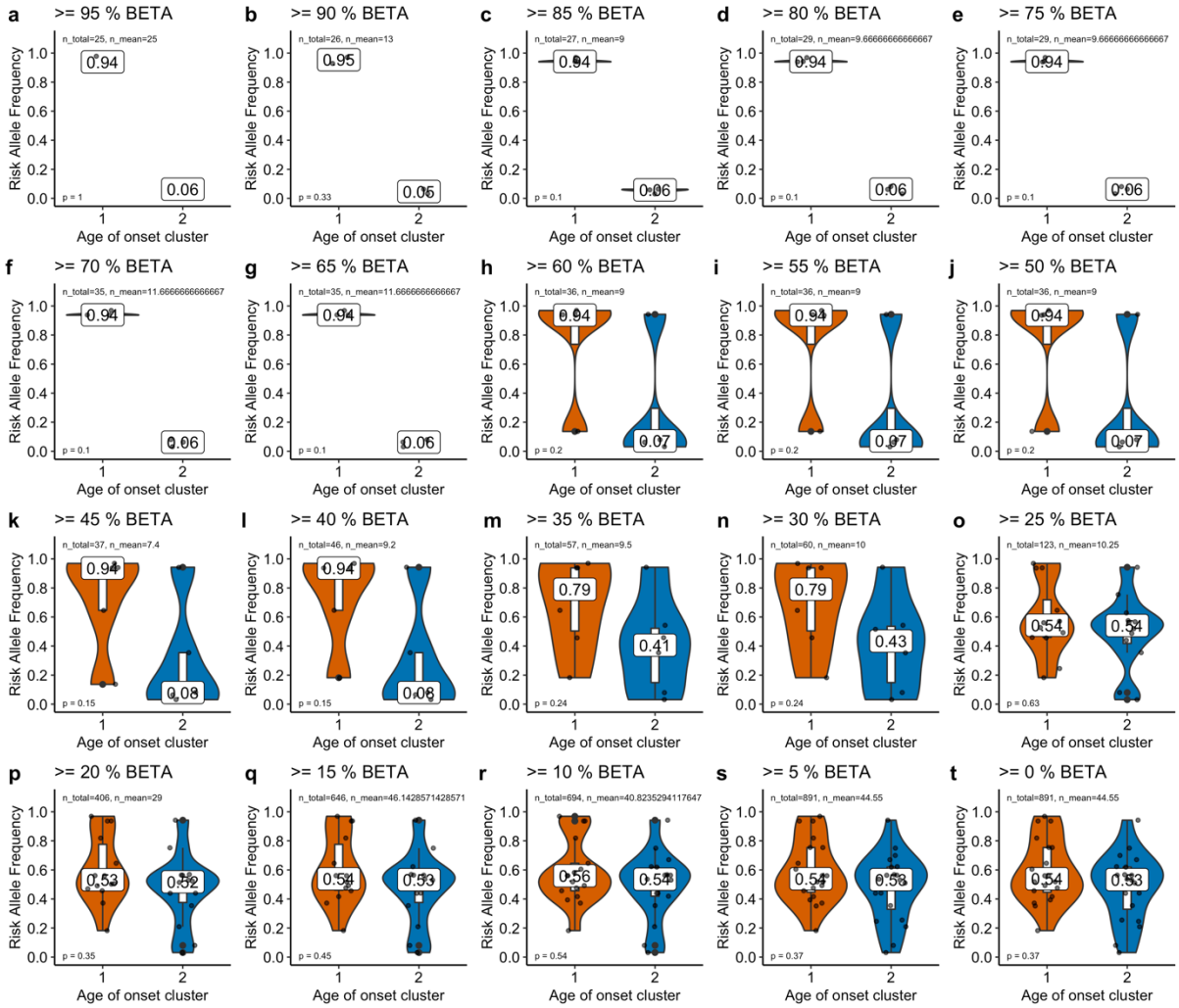

Figure S29: Risk allele frequencies in UK Biobank for the loci showing antagonistic associations between cluster 1 and cluster 2 filtered by different effect size cutoffs. The title of each plot shows the cutoff, where e.g. >=95 % BETA means only the SNPs with a BETA (effect size) value higher than 95% of all other antagonistic SNPs are used. >=0 % BETA means no filtering.

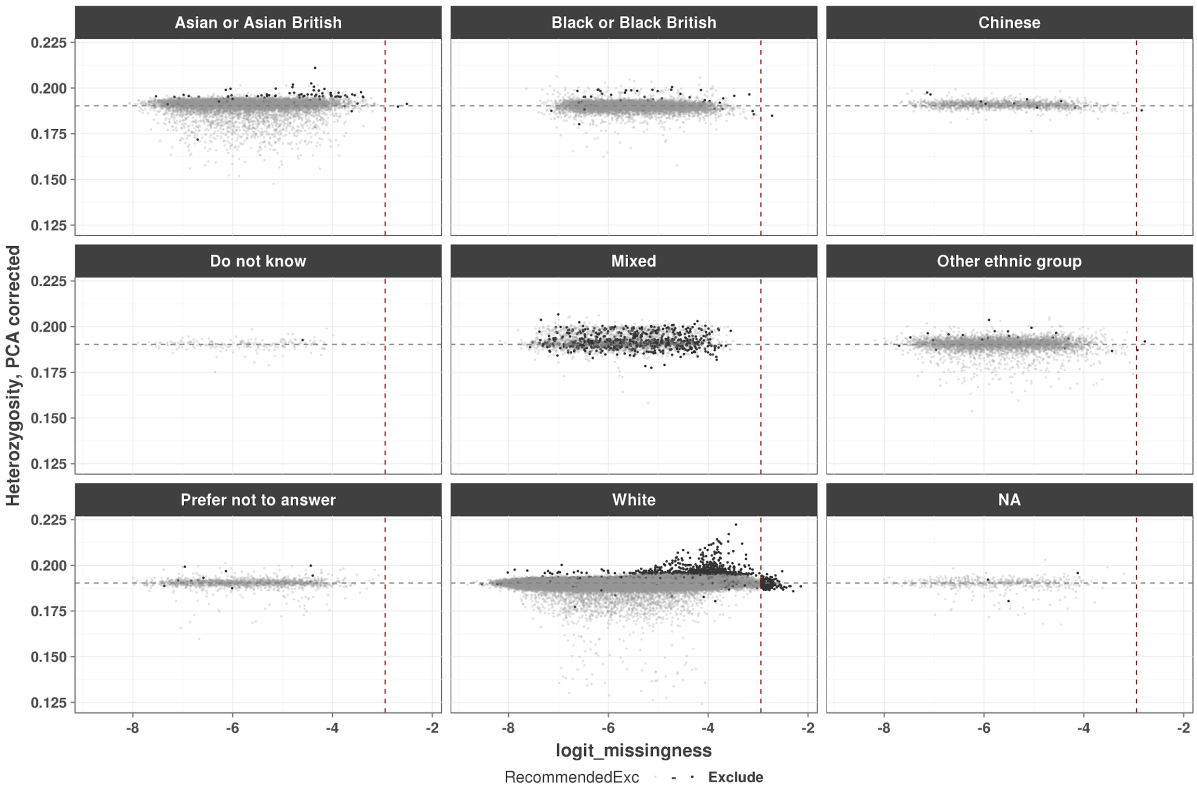

Figure S30: Scatter plot between logit(missingness) and PCA corrected heterozygosity measures. Each panel shows a self-declared ethnic background. Vertical red lines show the missing rate of 0.05, and horizontal grey lines show the average heterozygosity in UK Biobank.

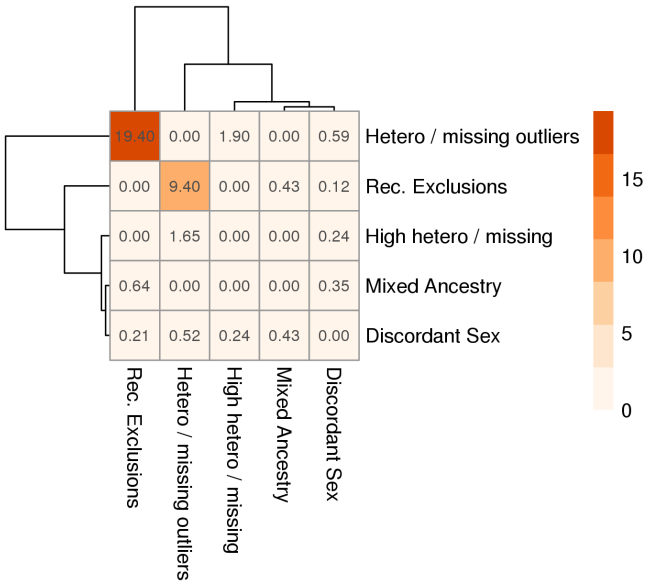

Figure S31: Heatmap showing the percent overlap between exclusions based on different criteria. Values show the percent of the column in the row, e.g. 19.4% of “Rec. Exclusions” are in “Hetero / missing outliers” i) “Hetero / missing outliers”: ‘22027-0.0’ (Outliers for heterozygosity or missing rate), ii) “Rec. Exclusions”: field ‘22010-0.0’ (Recommended genomic analysis exclusions), iii) “High hetero / missing”: ‘22018-0.0’, High heterozygosity rate (after correcting for ancestry) or high missing rate, iv) “Mixed Ancestry”: ‘22018-0.0’, Participant self-declared as having a mixed ancestral background, and v) “Discordant Sex”: as described in the sample QC methods.

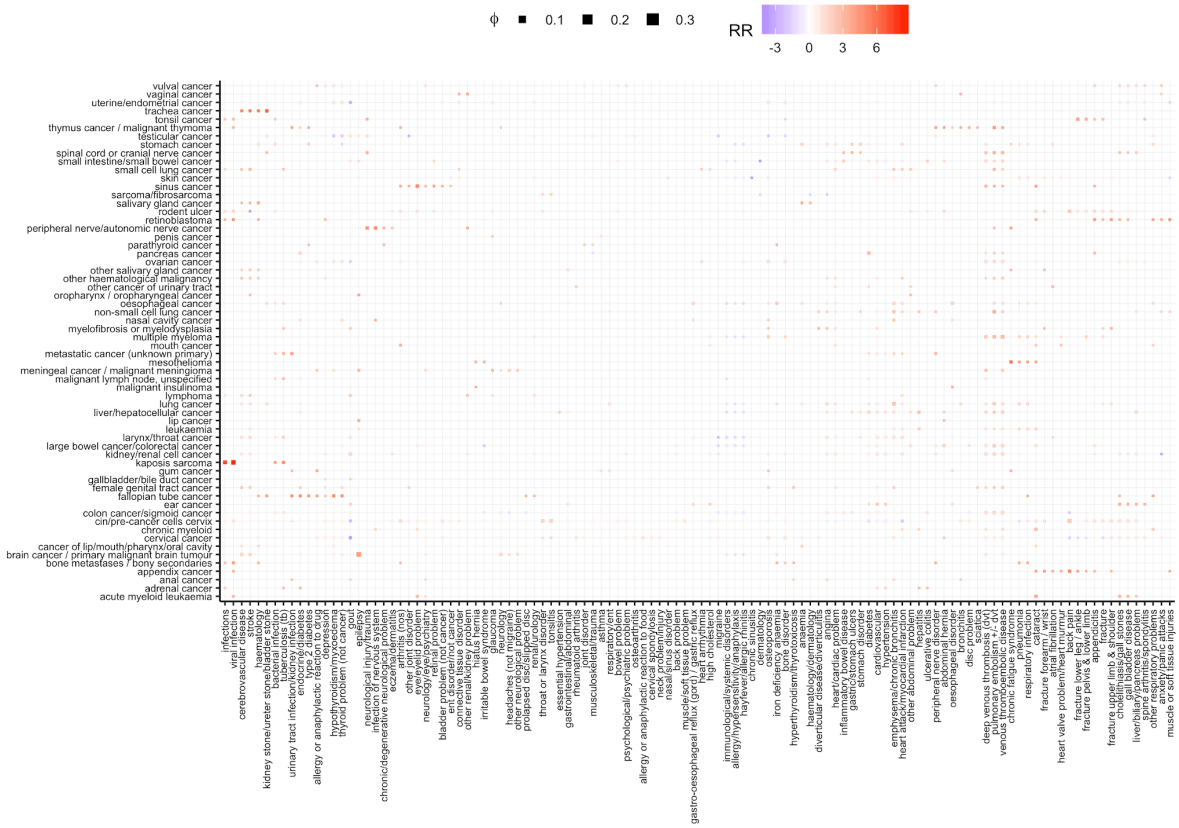

Figure S32: Cancer - disease co-occurrence matrix summarizing relative risk scores and correlations. Each row shows a cancer type and column shows a disease. The color is defined by relative risk scores while the size is determined by ϕ value (in the same scale as Figure S8 for better comparison), indicating the robustness. Associations for the 114 diseases and 62 cancers that have at least one relative risk ratio higher than four ($log_2 RR \geq 2$) or lower than minus four ($log_2 RR \leq -2$) are plotted.

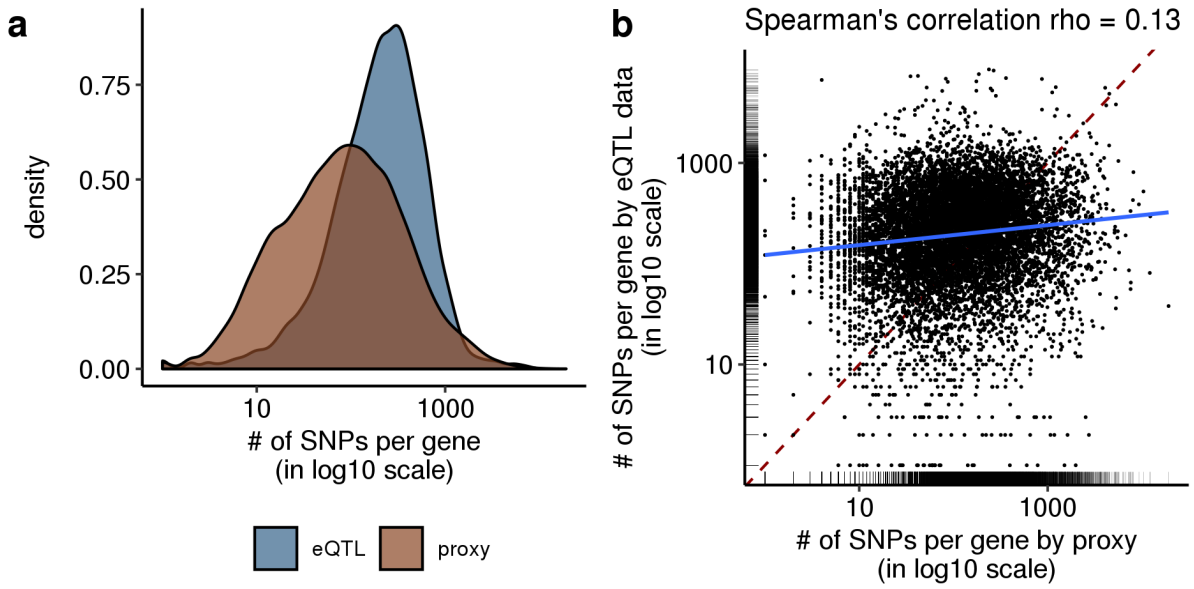

Figure S33: a) Density plots showing the number of SNPs per gene, based on eQTL data (blue) and proximity (brown). b) Scatter plot between the number of SNPs per gene mapped using genomic proximity (x-axis) or eQTL data (y-axis). Each dot represents a gene and the blue line shows the linear model. Dashed red line shows one-to-one relationship. The rug-plots on the axes show the marginal distribution of genes.