Supplemental Material

S1 UK Biobank data processing

1. Genotype and phenotype data from the UK Biobank release [8] were extracted (488,377 individuals, 784,256 variants) and filtered as follows:
   (a) Genotype data were extracted from the chrom*.cal files using the UK Biobank gconv tool
   (b) Phenotype data were taken from our application-specific csv file for application 22419
2. Only individuals who self-identified as white British were included in the study cohort (57,275 individuals removed)
3. All monomorphic variants were removed (19,189 variants removed)
4. Individuals identified by the UK Biobank to have high heterozygosity, excessive relatedness, or aneuploidy were removed (1,550 individuals removed)
5. Variants with a minor allele frequency less than 2.5% were not included (253,939 variants removed)
6. Only variants found to be Hardy-Weinberg Equilibrium (Fisher's exact test p-value $> 10^{-6}$) using plink 2.0 [36] were included (40,433 variants removed)
7. Variants with missingness greater than 1% were removed (60,523 variants removed)
8. Individuals with greater than 5% genotype missingness were removed (38 individuals removed)
9. Individuals who were third-degree relatives or closer were removed using the following process: One individual was removed at random from any pair of individuals with a kinship coefficient greater than 0.0442, calculated using KING (version 2.0; [106])

S2 WINGS applied to seven continuous and 81 binary phenotypes in the UK Biobank

Figure S1 displays results from simultaneously applying WINGS to seven continuous and 81 binary phenotypes. The binary phenotypes and continuous phenotypes cluster separately, with the exception of nucleated red blood cells (NRB). We note that the NRB phenotype is only partially continuous in that there is a continuous spectrum of nucleated red blood cells for unhealthy individuals, but all healthy individuals will have a zero value. Thus, it is not surprising that NRB phenotype does not belong to a prioritized cluster.

Ignoring the NRB phenotype, the cluster of continuous phenotypes (represented in purple on the top of the dendrogram in Figure S1) remains completely disjoint from the case-control phenotypes until there is only a single cluster containing all phenotypes. We observe that the [BMI, WHR] cluster has 3,634 shared significant genes (p-value $< 2.83 \times 10^{-6}$); the [PLC, MCV, MPV] cluster has 1,746 shared significant genes; and, the full continuous cluster with phenotypes [BMI, WHR, PLC, MCV, MCV, Height] has 541 shared significant genes. This is unsurprising as complex continuous phenotypes have been shown to be highly polygenic [3, 107, 108].

S3 Robustness to clustering criterion

In this paper, we present WINGS, a thresholded clustering algorithm based on Ward Hierarchical Clustering. While the Ward linkage criterion works well to cluster phenotypes, WINGS can easily be adapted to use other linkage criteria. To test the robustness of WINGS with respect to the choice of linkage criterion, we applied our method using single linkage, average linkage, and complete linkage clustering to the 81 phenotypes we analyzed from the UK Biobank in the main text (see [34] for more information on single linkage, average linkage, and complete linkage clustering). We used the same branch length thresholding algorithm described in Section 2.2 with each linkage criterion to identify prioritized clusters, and the clustering algorithms were all applied to the -$\log_{10}$-transformed PEGASUS gene scores.

The dendrograms and sorted branch length plots for these results are demonstrated in Figures S2-S7. For reference, the Ward-based WINGS results are presented in Figure 3 and Figure S11. In partial agreement with the Ward-based WINGS results, single linkage clustering paired with the branch length thresholding algorithm prioritizes the polyp cluster and kidney cluster, as well as two-phenotype subsets of the metabolic cluster (I20, I25) and immunological 2 cluster (G35, M06) (Figures S2-S3). The prioritized clusters identified via complete linkage clustering are in better agreement with our Ward-based WINGS results (Figures S4-S5). Specifically, complete linkage clustering paired with the branch length thresholding algorithm prioritizes the liver cluster, kidney cluster, polyp cluster, and alzheimer's/dementia cluster, where each of these clusters contains the same phenotypes as the corresponding Ward-based prioritized clusters. Moreover, combinations of the metabolic cluster and immunological clusters are prioritized using complete linkage clustering. Lastly, the average linkage prioritized clusters contain a two-phenotype subset of the metabolic cluster (I20, I25) and combination of phenotypes from the immunological 1 and immunological 2 clusters (Figures S6-S7).

While the prioritized clusters vary depending on the linkage criterion, there is significant overlap between these clusters. Moreover, the dissimilarity scores (from [92]) between the corresponding trees are relatively low; the dissimilarity index between the Ward tree in Figure 3 and the single linkage Tree in Figure S2 is $Z = 0.1892$, the dissimilarity index between the Ward tree in Figure 3 and the complete linkage Tree in Figure S4 is $Z = 0.1497$, and the dissimilarity index between the Ward tree in Figure 3 and the average linkage Tree in Figure S6 is $Z = 0.1412$.

We reiterate that previous work on comparing different agglomerative hierarchical clustering algorithms suggests that Ward clustering performs the best when applied to high dimensional, noisy data and is therefore particularly useful for its application to the high dimensional gene score matrices studied in this work (as long as cluster sizes are assumed to be approximately equal) [40, 41]. Future studies will be dedicated to fully understanding the differences between the prioritized clusters identified by WINGS, single linkage clustering, average linkage clustering, and complete linkage clustering.

S4 Analysis of 26 case-control phenotypes

Here we present results from applying WINGS to 26 binary chronic illness phenotypes in the UK Biobank. Figure S9 displays the branch length outputs of WINGS (see Methods, section 2) applied to the -$\log_{10}$ transformed PEGASUS gene scores computed using cases and controls from the UK Biobank for 26 binary chronic illness phenotypes that were also studied by Shi et al. [2] and Pickrell et al. [11].

The prioritized -$\log_{10}$ clusters identified by WINGS in Figure S9 can be annotated as metabolic [E11, I25, E78], immunological [K900, J45, K51, L40, M06, G35, M05, M07], and Alzheimer's/dementia [G30, F01] (see Table S1 for common disease names, as well as the shared significant genes in a cluster). The prioritized clusters identified when WINGS is applied to these 26 -$\log_{10}$ transformed phenotypes in the UK Biobank are similar to the prioritized clusters identified from WINGS applied to 81 case-control -$\log_{10}$ transformed phenotypes in the UK Biobank (see Figure 3 in the main text).

Figure S8 displays the dendrogram output of WINGS applied to the -$\log_{10}$-transformed PEGASUS gene scores for these 26 binary chronic illness phenotypes in the UK Biobank. The dendrogram displays the hierarchical nature of the immunological cluster (orange branches in Figure S8), and it demonstrates the proximity of the [G30, F01] cluster to other phenotypes.

Table S1: Phenotypes analyzed in this study sorted by International Classication of Disease (ICD10) codes. * denotes that the phenotype was included in the initial analysis of 26 case-control phenotypes that were also studied by Pickrell et al. [11] and Shi et al. [2]

Disease	ICD10 Code	Number of Cases	Heritability
Iron deficiency anemia	D50	6284	0.0041
Other anemias	D64	9522	0.0026
Other coagulation defects	D68	809	0.0067
Neutropenia	D70	2636	0.0019
*Sarcoidosis	D86	449	0.0052
Other hypothyroidism	E03	11691	0.0384
Type 1 diabetes mellitus	E10	2373	0.0071
*Type 2 diabetes mellitus	E11	15080	0.0526
Other disorders of pancreatic internal secretion	E16	764	0.0003
Overweight and obesity	E66	8950	0.0267
*Disorders of lipoprotein metabolism and other lipidemias	E78	29778	0.0498
Disorders of mineral metabolism	E83	1758	0.001
*Vascular dementia	F01	156	0.0017
Alcohol related disorders	F10	4313	0.0105
*Schizophrenia	F20	425	0.0011
*Bipolar disorder	F31	791	0.0042
*Major depressive disorder	F32	9714	0.0143
Other anxiety disorders	F41	4881	0.0067
*Parkinson's disease	G20	972	0.005
*Alzheimer's disease	G30	331	0.0008
*Multiple sclerosis	G35	1124	0.0029
Epilepsy and recurrent seizures	G40	3071	0.0049
*Migraine	G43	2263	0.0013
Sleep disorders	G47	4410	0.0089
Age-realted cataract	H25	6814	0.0078
Glaucoma	H40	3729	0.013
Hypertension	I10	64135	0.0932
Hypertensive chronic kidney disease	I12	1274	0.004
Angina pectoris	I20	15063	0.0393
Acute myocradial infarction	I21	6655	0.022
*Chronic ischemic heart disease	I25	20958	0.0466
Cardiomyopathy	I42	1035	0.0037
Heart failure	I50	4423	0.0121
Atherosclerosis	I70	1025	0.0047
Varicose veins of lower extremities	I83	8988	0.0277
Hypotension	I95	4072	0.0037
Other and unspecified disorders of nose and nasal sinuses	J34	5393	0.0012
Emphysema	J43	1388	0.0103
Other chronic obstructive pulmonary disease	J44	6833	0.0223
*Asthma	J45	21758	0.043
Gastro-esophageal reflux disease	K21	19132	0.0188
Gastric ulcer	K25	3467	0.005
Duodenal ulcer	K26	2517	0.0028
Functional dyspepsia	K30	9696	0.0054
*Crohn's disease	K50	1436	0.0068
*Ulcerative colitis	K51	2661	0.0079
Diverticular disease of large intestine without perforation or abscess	K573	19462	0.0321

Irritable bowel syndrome	K58	4563	0.0089
Rectal polyp	K621	5210	0.0075
Polyp of colon	K635	9306	0.0105
Fibrosis and cirrhosis of liver	K74	676	0.0026
Other diseases of liver	K76	2791	0.0035
Other diseases of gallbladder	K82	1482	0.0011
Other diseases of pancreas	K86	896	0.0014
*Celiac disease	K900	1522	0.0051
Gastrointestinal hemorrhage	K922	4387	0.0008
*Psoriasis	L40	1836	0.0047
*Lupus erythematosus	L93	105	0.0033
*Rheumatoid arthritis with rheumatoid factor	M05	465	0.0063
*Other rheumatoid arthritis	M06	3581	0.0072
Gout	M10	2661	0.0132
Other arthritis	M13	9500	0.0109
Osteoarthritis of hip	M16	9876	0.0208
Osteoarthritis of knee	M17	16612	0.031
Other and unspecified osteoarthritis	M19	13548	0.0156
Scoliosis	M41	838	0.0037
Other disorders of muscle	M62	746	0.0042
Synovitis and tenosynovitis	M65	4311	0.0081
Fibroblastic disorders	M72	3267	0.0231
Osteoporosis	M81	4884	0.0115
Chronic kidney disease	N18	3714	0.0055
Other disorders of kidney and ureter	N28	1996	0.0052
Other disorders of urinary system	N39	15870	0.0112
Benign prostatic hyperplasia	N40	9471	0.0108
Inflammatory diseases of prostate	N41	1334	0.004
Endometriosis	N80	3235	0.0094
Abnormalities of heart beat	R00	7018	0.0016
*Allergy status to penicillin	Z880	13436	0.0132
*Allergy status to narcotic agent status	Z885	983	0.0011
*Allergy status to analgesic agent status	Z886	3586	0.0033
*Allergy status to serum and vaccine status	Z887	157	0.0014

691

Table S2: For each prioritized cluster identified in our analysis of the UK Biobank (Figures 3 and S11), we list shared significant genes that have been previously associated with at least one phenotype in that cluster and that shared significant genes have a PEGASUS gene-level association p-value that is significant after Bonferroni correction for 17,651 autosomal genes for all phenotypes in the prioritized cluster of interest. Starred gene names indicate that a gene occurs in at least one significantly enriched pathway in our gene set enrichment analysis (see PEGASUS-WINGS GitHub repository for all pathways passing FDR < 0.05 for each prioritized cluster). In the first column, the number in parenthesis corresponds to the number of phenotypes in that cluster; the second column (N) corresponds to the number of shared significant genes (PEGASUS
p-value $< 2.83 \times 10^{-6}$).

Cluster label	N	Genes from column (2) associated in GWAS Catalog for ≥ 1 phenotype in cluster	GWAS Catalog associated phenotype and relevant references for gene in column (3)
Metabolic (8)	2	*NCR3*	IgG Glycolysation [52] Diabetes mellitus [53]
Immunological 1 (3)	181	*C6orf10* *HCP5* *HLA-DQA1** *MICA** *NOTCH4* *POU5F1*	Psoriasis [54] Psoriasis [54, 55, 56] Celiac disease [59, 60] Psoriasis [54] Celiac disease [61] Psoriasis [57]
Immunological 2 (6)	96	*APOM* *BRD2* *BTNL2* *CDSN* *CFB** *HCP5* *HLA-DOA** *HLA-DQA1** *HLA-DQB1** *HLA-DRA** *HLA-DRB5** *MICA** *MICB** *NCR3** *NOTCH4** *PBX2* *PSORS1C1* *TAP2**	Rheumatoid arthritis [69] Asthma [67] Asthma [68] Asthma [68] Asthma [52] Asthma [68, 66] Asthma [67] Asthma [62] Rheumatoid arthritis [63] Type 1 diabetes [64] Asthma [65] Hypothyroidism [11] Asthma [66] Rheumatoid arthritis [63] Rheumatoid arthritis [70] Type 1 diabetes [55] Asthma [66] Type 1 diabetes [52] Asthma [67] Asthma [67] Asthma [68] Type 1 diabetes [52]
Polyp (2)	43	*UTP23* *GREM1* *SCG5* *SMAD7** *CABLES2* *LAMA5** *PREX1*	Colorectal cancers [74] Colorectal cancers [75] Colorectal cancers [76] Colorectal cancers [77] Colorectal cancers [78] Colorectal cancers [79] Colorectal cancers [80]
Alzheimer's/Dementia (2)	8	*PALM2* *APOC1* *APOC2*	Frontotemporal dementia [81] Alzheimer's disease [82] Alzheimer's disease [82]

		APOC4	Alzheimer's disease [82]
		APOE	Alzheimer's disease [83]
		CLPTM1	Alzheimer's disease [84]
		PVRL2	Alzheimer's disease [85]
		TOMM40	Alzheimer's disease [86]
Kidney (2)	9	*OVOL1*	Urate levels [88]
Liver (2)	25	*GATAD2A*	Nonalcoholic fatty liver disease [89]
		PNPLA3	Nonalcoholic fatty liver disease [90]
		SAMM50	Nonalcoholic fatty liver disease [91]

Phenotypes in Simulation	Shared Genetic Architecture											
	A: Power				**B**: Precision				**C**: F1			
	0.1	0.25	0.5	0.75	0.1	0.25	0.5	0.75	0.1	0.25	0.5	0.75
25	99.91	100	100	100	78.76	80.92	85.32	87.67	0.87	0.87	0.86	0.86
50	99.97	100	100	100	71.22	76.06	79.18	84.27	0.74	0.75	0.74	0.73
75	99.94	100	100	100	67.62	73.09	76.45	82.14	0.66	0.66	0.67	0.66
100	99.97	100	100	100	65.05	72.52	76.61	81.81	0.64	0.64	0.62	0.62

Table S3: WINGS performance on simulated data generated using the empirical distribution of PASCAL [24] sum gene scores for Crohn's disease (17,582 genes). Power (A), precision (B), and F1 score (C) of WINGS across a range of phenotypes included as well as shared genetic architecture. "Shared genetic architecture" denotes the percentage of the 175 significant genes in each phenotype that are shared across all phenotypes in a cluster. Every entry in the table represents 1,000 simulations under the corresponding parameters. Power and precision are defined explicitly in Table 2. F1 score is twice the product of precision and recall divided by the sum of precision and recall; in this context, recall is the percentage of ground truth clusters prioritized by WINGS.

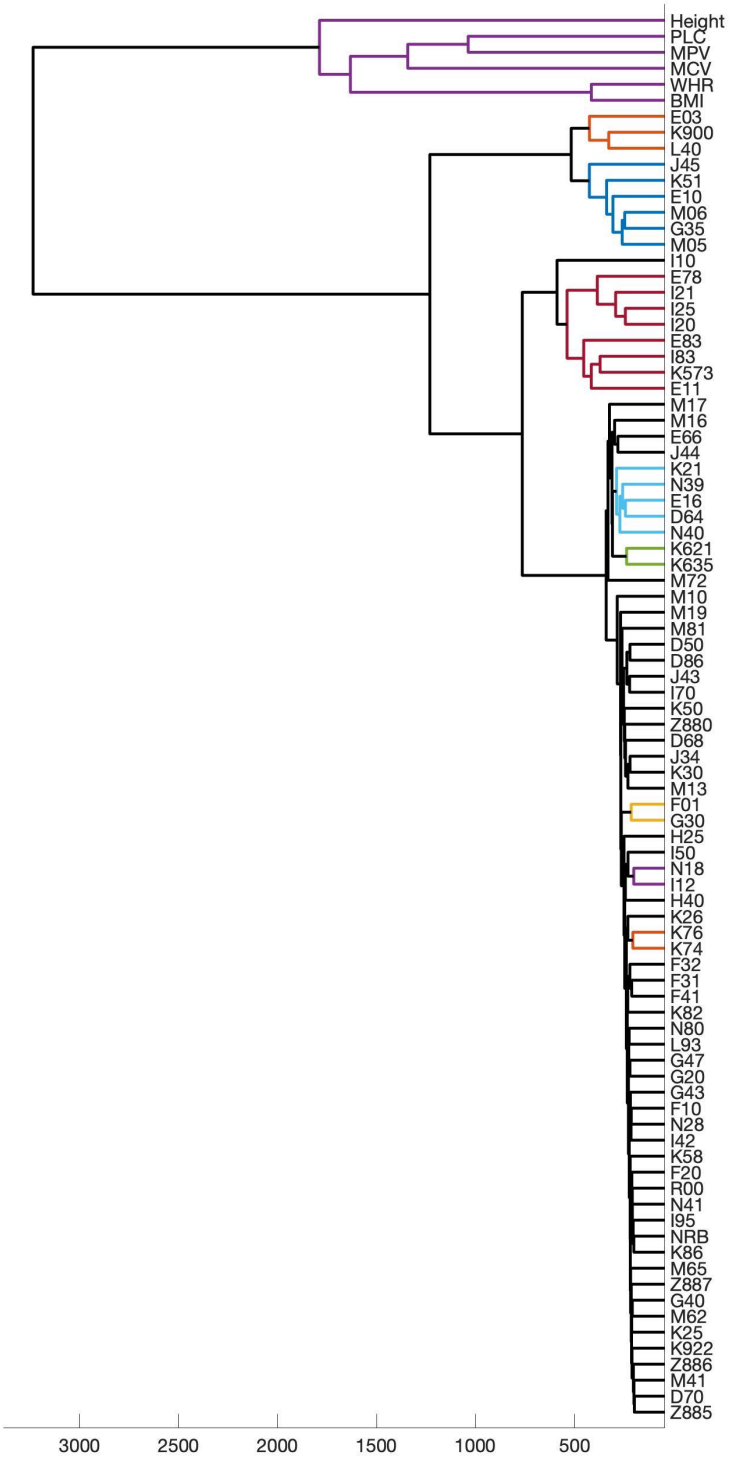

Figure S1: **WINGS dendrogram from 81 case-control phenotypes and seven continuous phenotypes in the UK Biobank separates continuous and binary phenotypes.** We show the dendrogram output of Ward hierarchical clustering applied to the -$\log_{10}$ transformed PEGASUS scores of the empirical continuous and binary phenotypes. The branches are color coded by the largest prioritized clusters identified by the branch length thresholding algorithm. The continuous phenotypes (except for the nucleated red blood cells (NRB) phenotype) cluster together on the top of the dendrogram (in purple), remaining disjoint from the remaining binary phenotypes until there is a single cluster.

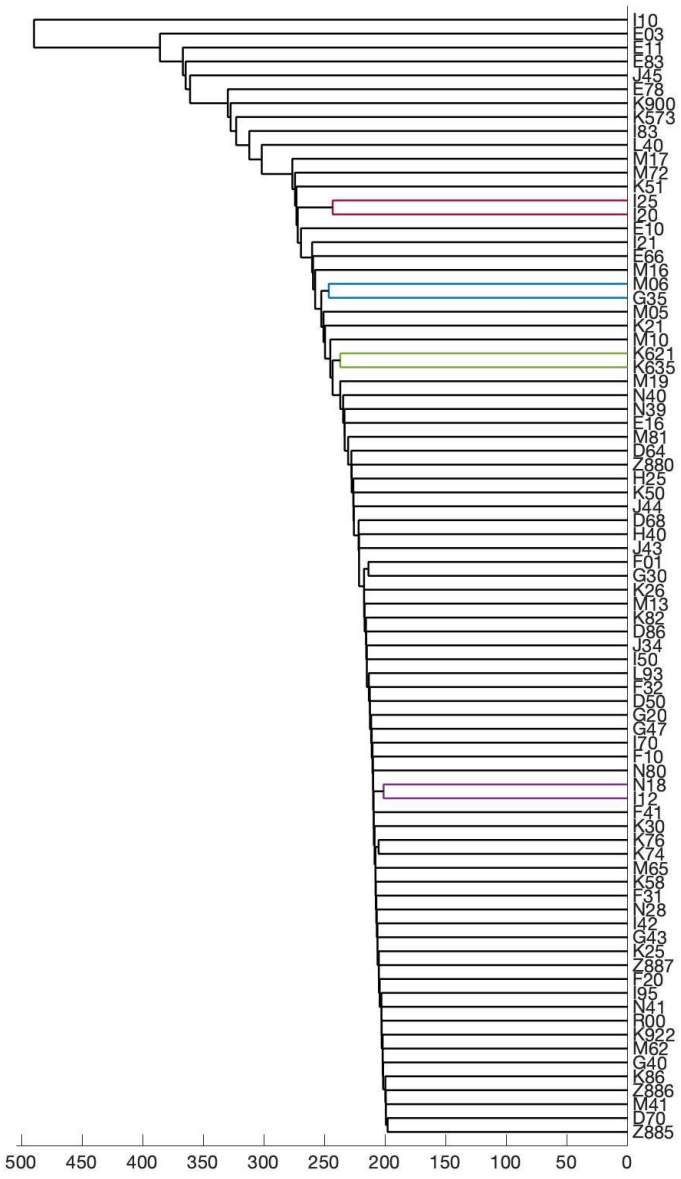

Figure S2: **Single linkage clustering applied to -$\log_{10}$ transformed PEGASUS p-values of 81 phenotypes from the UK Biobank.** We show the dendrogram corresponding to the output of single linkage hierarchical clustering applied to the -$\log_{10}$ transformed PEGASUS scores of the 81 phenotypes from the UK Biobank. The branches are color coded by the largest prioritized clusters identified by the branch length thresholding algorithm. The dissimilarity index (from [92]) between this tree and Figure 3 is $Z = 0.1892$.

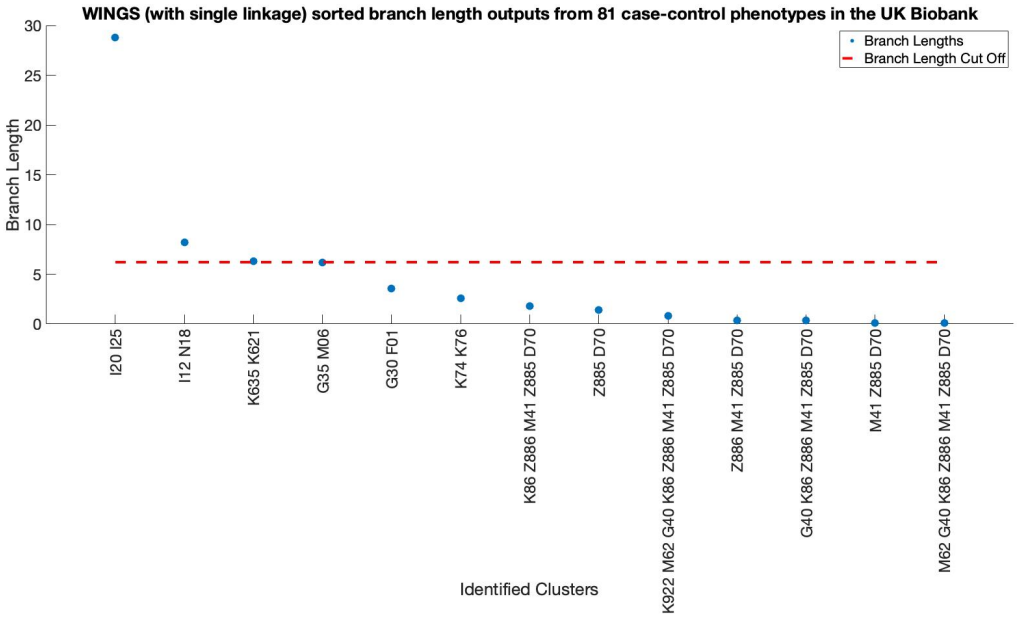

Figure S3: **Single linkage clustering applied to -$\log_{10}$ transformed PEGASUS p-values of 81 phenotypes from the UK Biobank.** We show the sorted branch lengths corresponding to the output of single linkage hierarchical clustering applied to the -$\log_{10}$ transformed PEGASUS scores of the 81 phenotypes from the UK Biobank. The dashed red horizontal line on the right figure corresponds to the branch length threshold, where the prioritized clusters are those lying above the dashed line.

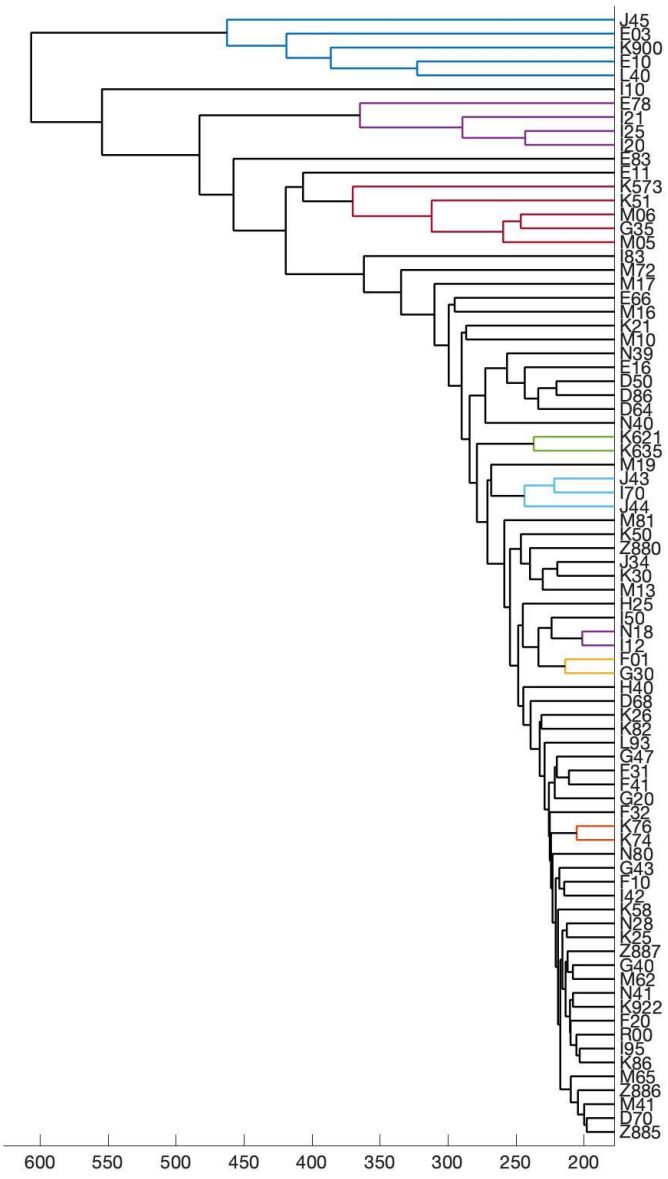

Figure S4: **Complete linkage clustering applied to -$\log_{10}$ transformed PEGASUS p-values of 81 phenotypes from the UK Biobank.** We show the dendrogram corresponding to the output of complete linkage hierarchical clustering applied to the -$\log_{10}$ transformed PEGASUS scores of the 81 phenotypes from the UK Biobank. The branches are color coded by the largest prioritized clusters identified by the branch length thresholding algorithm. The dissimilarity index (from [92]) between this tree and Figure 3 is $Z = 0.1497$.

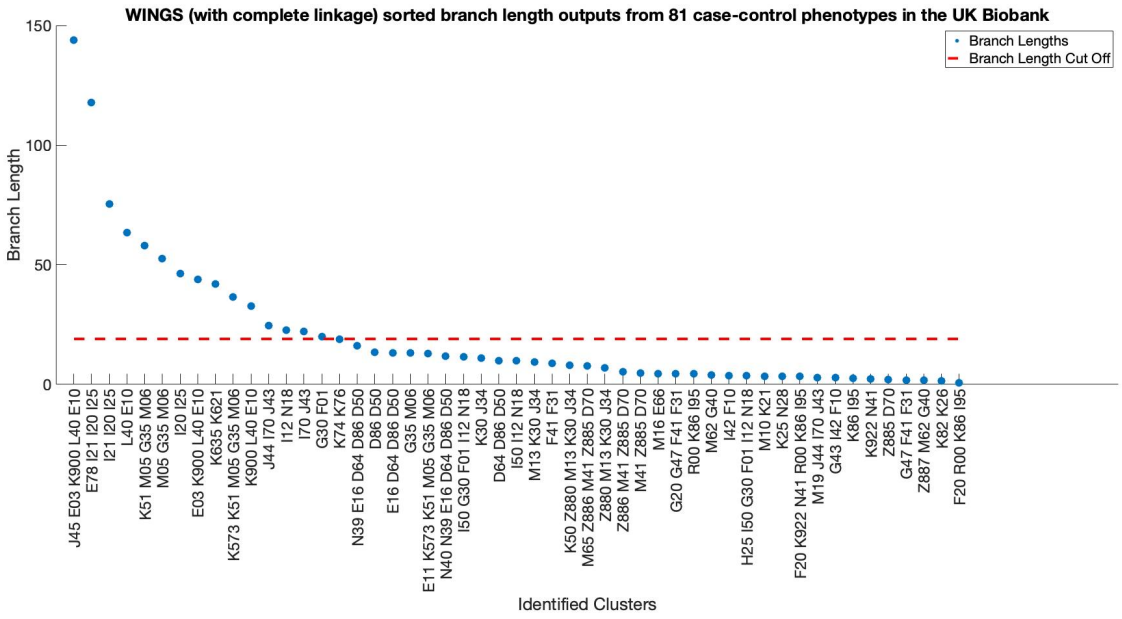

Figure S5: **Complete linkage clustering applied to -$\log_{10}$ transformed PEGASUS p-values of 81 phenotypes from the UK Biobank.** We show the sorted branch lengths corresponding to the output of complete linkage hierarchical clustering applied to the -$\log_{10}$ transformed PEGASUS scores of the 81 phenotypes from the UK Biobank. The dashed red horizontal line on the right figure corresponds to the branch length threshold, where the prioritized clusters are those lying above the dashed line.

Figure S6: **Average linkage clustering applied to -$\log_{10}$ transformed PEGASUS p-values of 81 phenotypes from the UK Biobank.** We show the dendrogram corresponding to the output of average linkage hierarchical clustering applied to the -$\log_{10}$ transformed PEGASUS scores of the 81 phenotypes from the UK Biobank. The branches are color coded by the largest prioritized clusters identified by the branch length thresholding algorithm. The dissimilarity index (from [92]) between this tree and Figure 3 is $Z = 0.1412$.

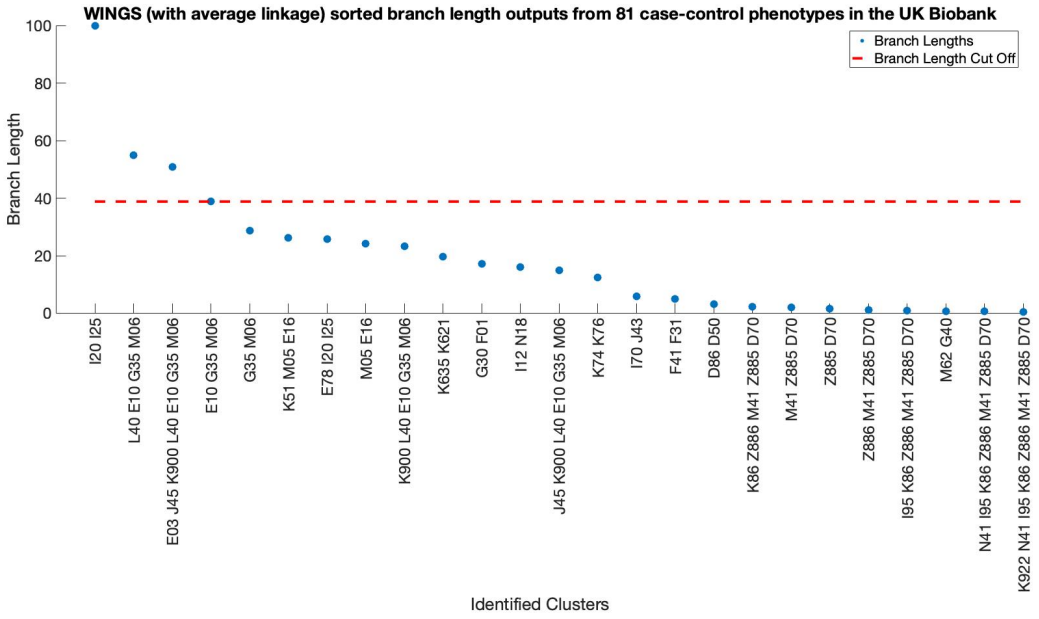

Figure S7: **Average linkage clustering applied to -$\log_{10}$ transformed PEGASUS p-values of 81 phenotypes from the UK Biobank.** We show the sorted branch lengths corresponding to the output of average linkage hierarchical clustering applied to the -$\log_{10}$ transformed PEGASUS scores of the 81 phenotypes from the UK Biobank. The dashed red horizontal line on the right figure corresponds to the branch length threshold, where the prioritized clusters are those lying above the dashed line.

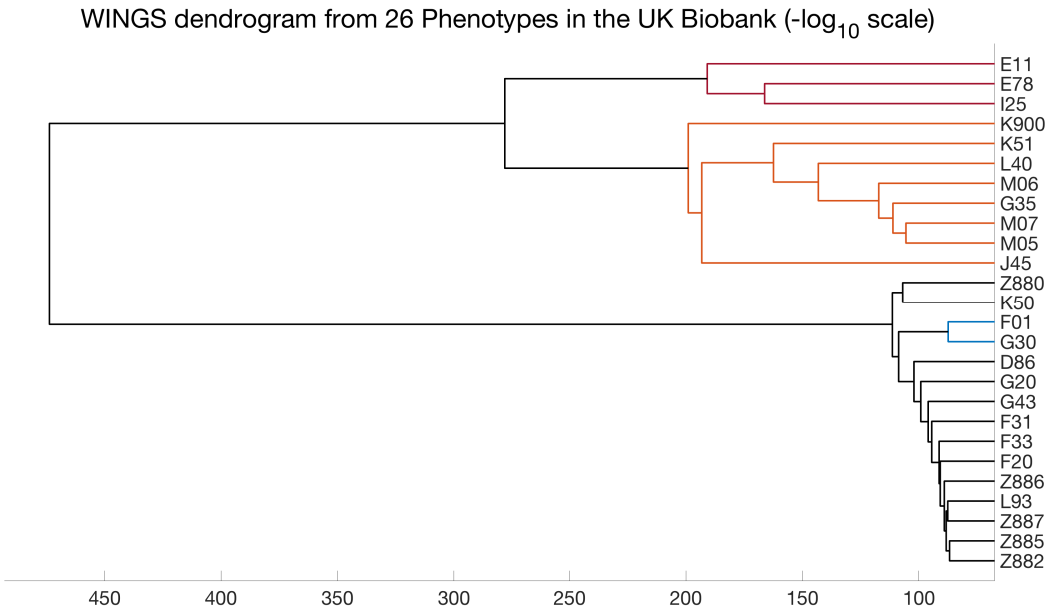

Figure S8: **WINGS dendrogram applied to -$\log_{10}$ transformed PEGASUS scores for 26 binary chronic illness phenotypes from the UK Biobank.** We show the dendrogram output of WINGS applied to the -$\log_{10}$ transformed PEGASUS scores of the 26 binary chronic illness phenotypes from the UK Biobank data. The color coded branches correspond to prioritized clusters identified by WINGS. The corresponding sorted branch lengths are presented in Figure S9 (B) in the paper.

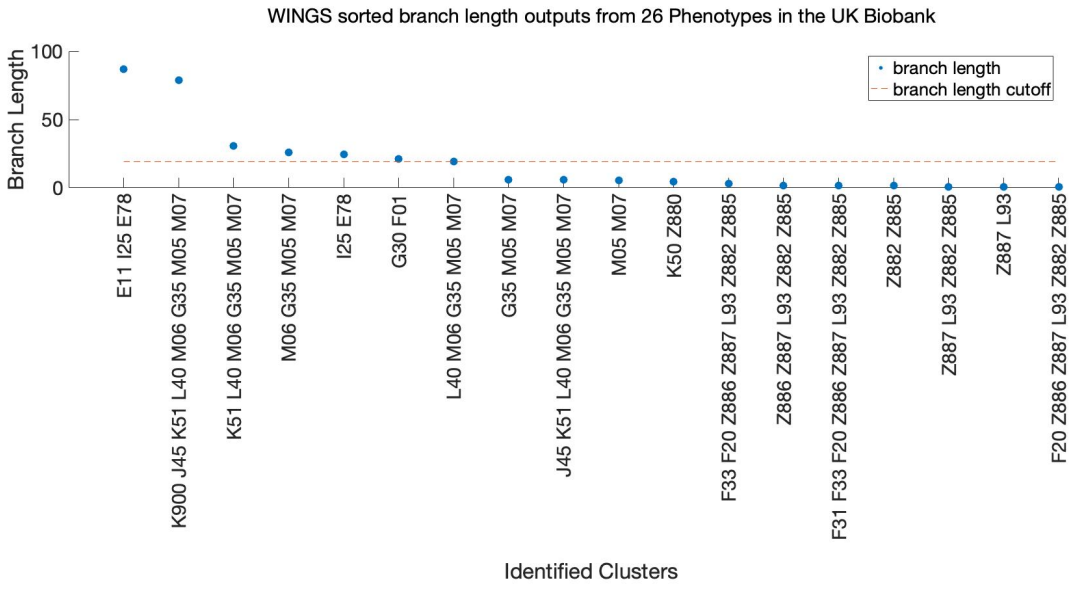

Figure S9: **WINGS sorted branch lengths applied to 26 binary chronic illness phenotypes from the UK Biobank on the -$\log_{10}$ scale.** We show the sorted branch lengths corresponding to the branches in the dendrogram output of WINGS applied to -$\log_{10}$-transformed PEGASUS gene scores for 26 case-control phenotypes in the UK Biobank. The dashed red horizontal line corresponds to the branch length threshold, where the identified prioritized clusters are those lying above the dashed line (boxed). Here, the x-axis shows the ICD10 codes; see Table S1 for the corresponding common disease names.

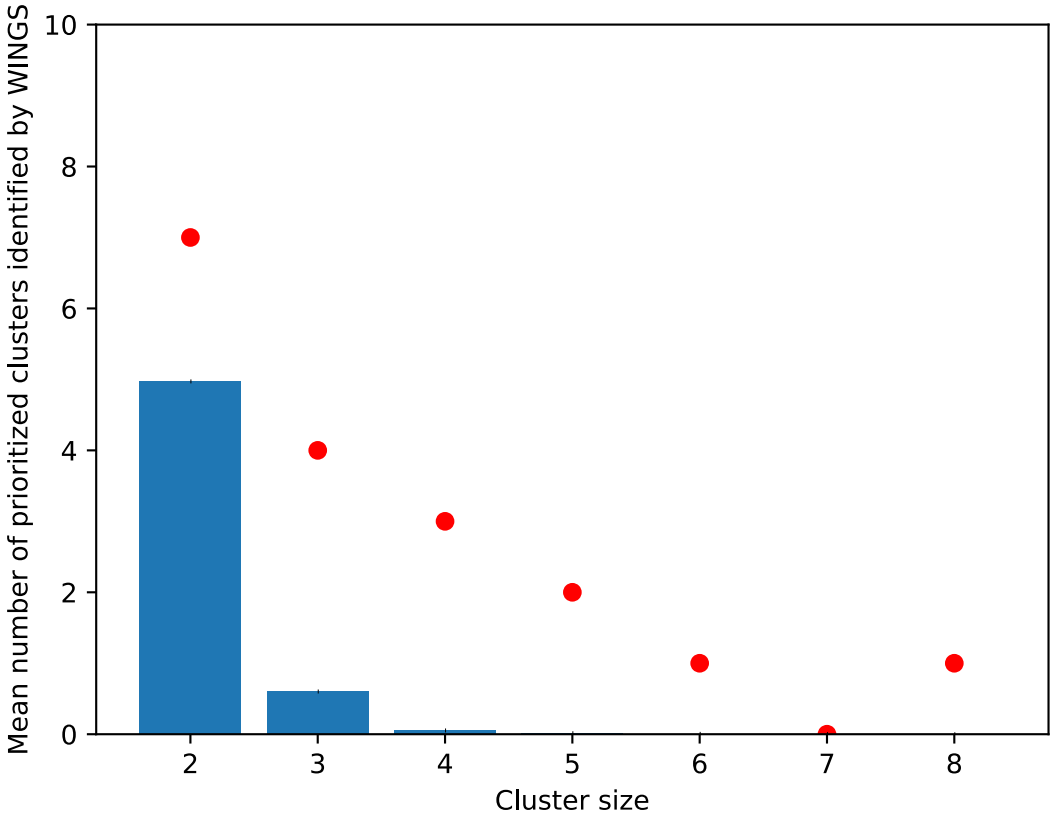

Figure S10: **Average number of prioritized clusters identified in 1,000 permutations, where each permutation shuffles gene scores for each of the 81 phenotypes analyzed in the UK Biobank.** For each cluster size shown, the number of prioritized clusters identified in the empirical matrix of gene scores (red dots) for the same phenotypes exceeds those observed in this permutation test. We note that we set the cluster size threshold to eight in our empirical analyses.

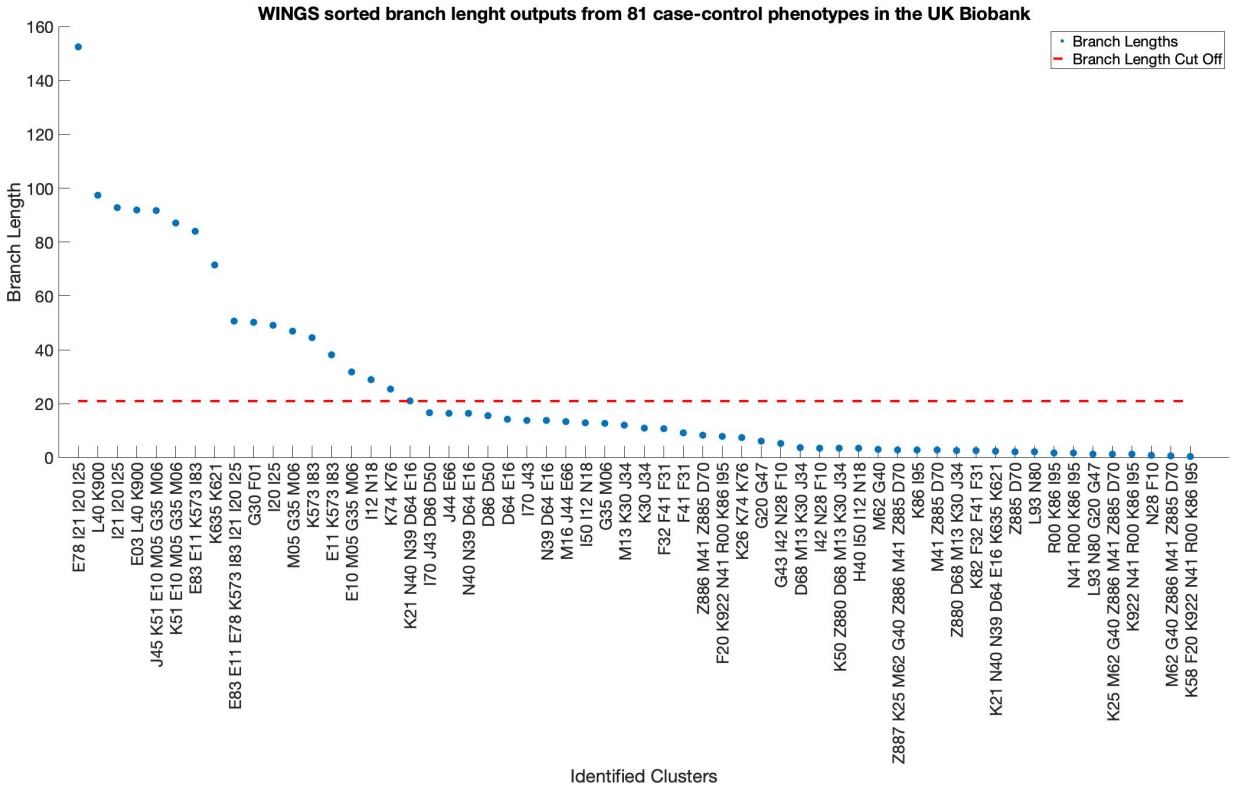

Figure S11: **WINGS sorted branch lengths from 81 case-control phenotypes in the UK Biobank reveals clusters of phenotypes with shared significant genetic architecture.** We show the sorted branch lengths corresponding to the dendrogram branches generated by WINGS when applied to the -$\log_{10}$ transformed PEGASUS gene scores from 81 case-control phenotypes in the UK Biobank. The dashed red horizontal line corresponds to the branch length threshold, where the prioritized clusters are those lying above the dashed line. The corresponding dendrogram is presented in Figure 3.

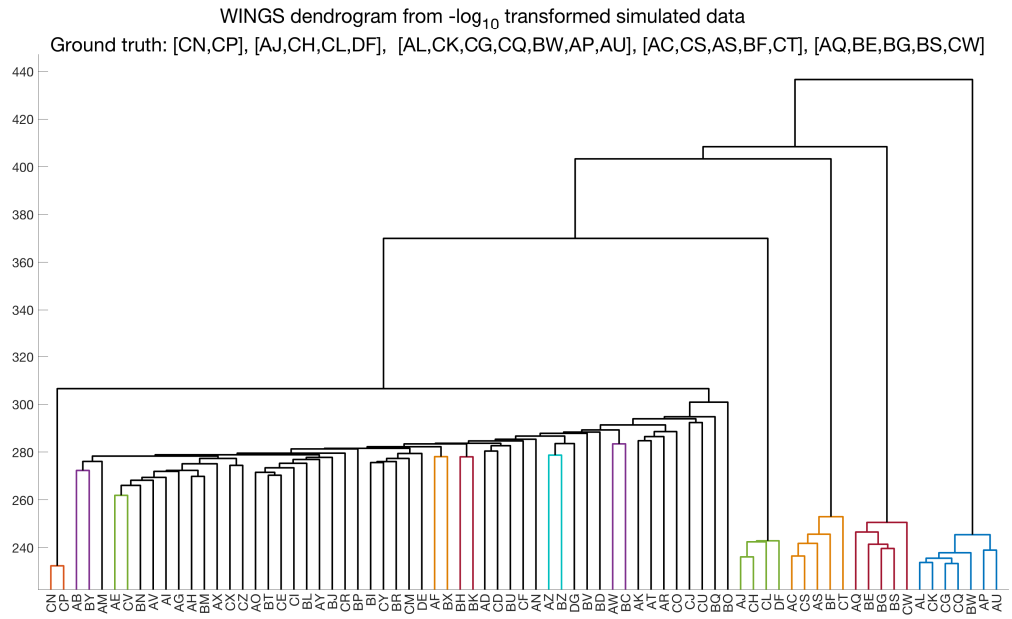

Figure S12: **WINGS dendrogram from a simulation on the -$\log_{10}$ scale.** We show the dendrogram output of Ward hierarchical clustering applied to the -$\log_{10}$ transformed PEGASUS scores of a simulation with 75 phenotypes and 75% shared genes. The branches are color coded by the largest prioritized clusters identified by the branch length thresholding algorithm. The corresponding sorted branch lengths are presented in Figure 2 in the paper.

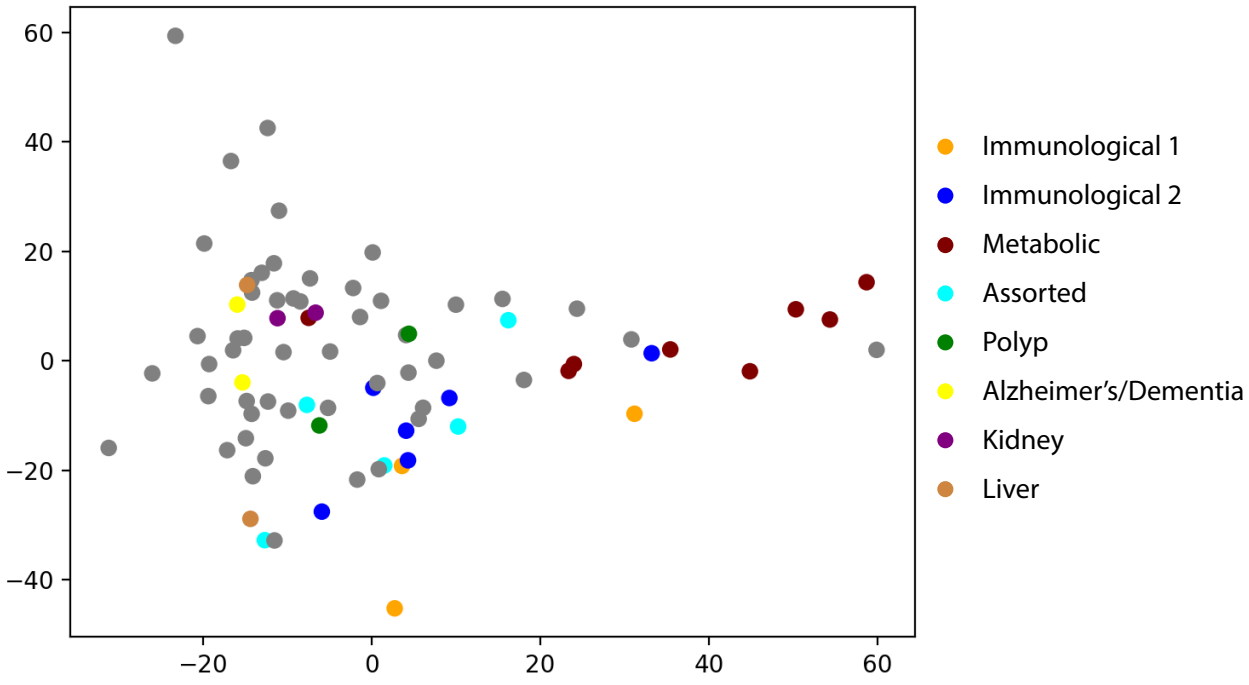

Figure S13: **Position of 81 case-control phenotypes resulting from a disPCA analysis.** Each point in the plot represents a single phenotype. The Euclidean distances between the points were used to apply WINGS and form the dendrogram displayed in Figure 4. The points are colored corresponding to their prioritized cluster status in Figure 3. The disPCA analysis does not result in those phenotype clusters identified by WINGS being easily differentiable.

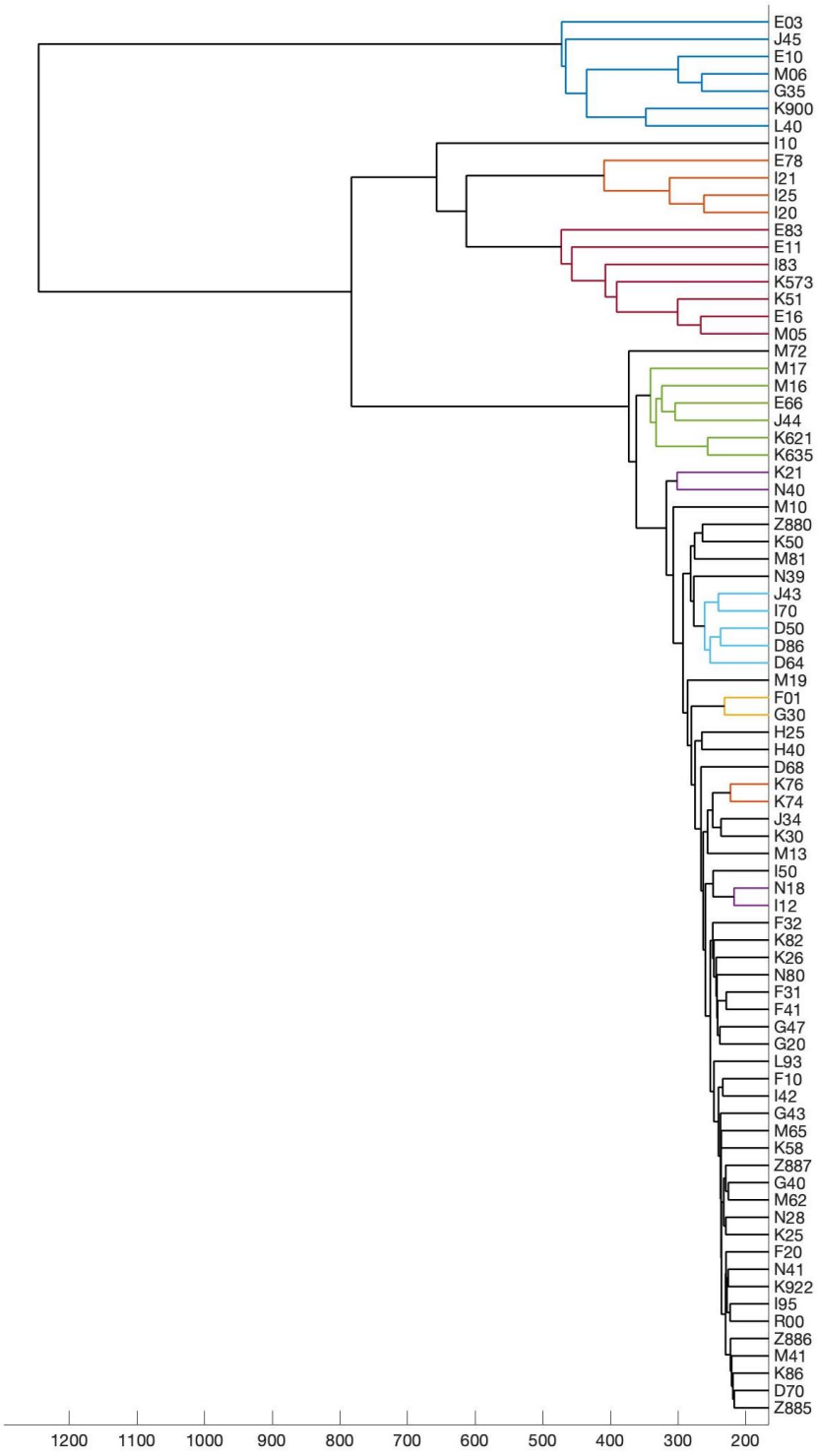

Figure S14: **WINGS dendrogram from 81 case-control phenotypes using both genes and intergenic regions as features.** We analyzed a matrix of PEGASUS p-values for 17,651 genes with a +/- 50kb buffer region and 2,960 intergenic regions on the -$\log_{10}$ scale using both genes and intergenic regions as features. Compared to the dendrogram shown in Figure 3, the topology of the tree is preserved and the resulting dissimilarity index (from [92]) between these two trees is $Z = 0.1091$. The corresponding sorted branch length diagram is presented in Figure S15.

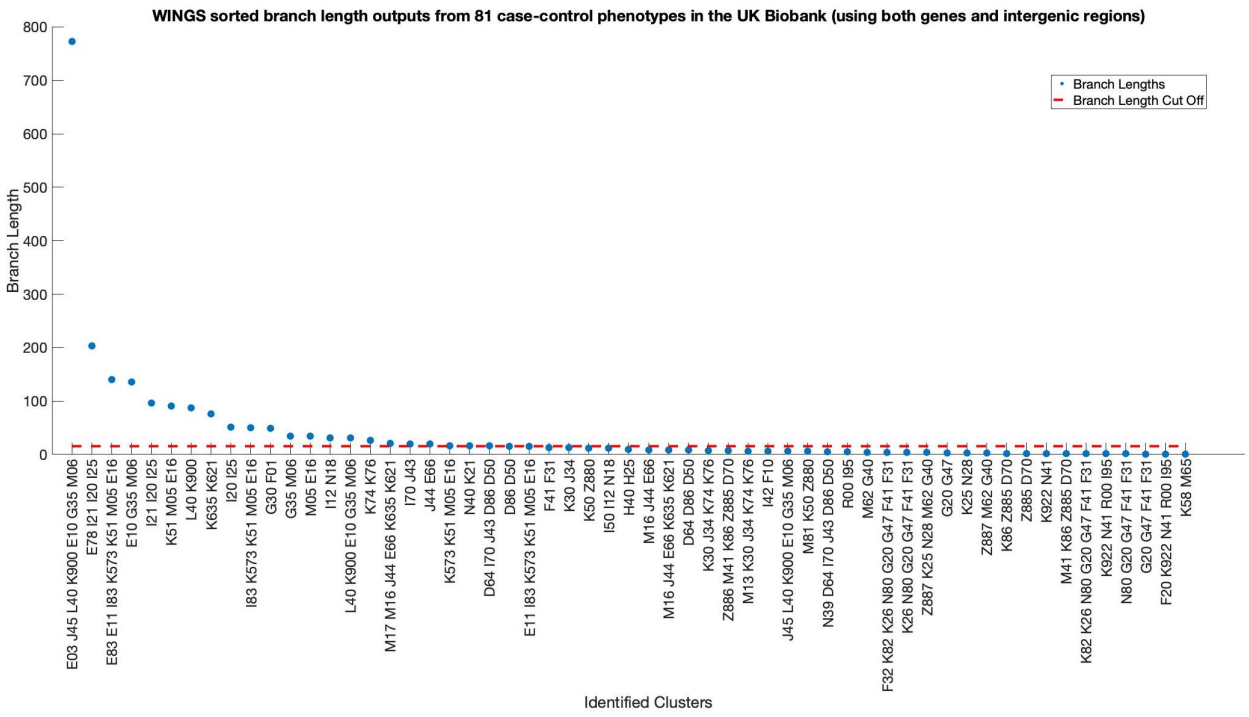

Figure S15: **WINGS sorted branch lengths from 81 case-control phenotypes using both genes and intergenic regions as features.** We show the sorted branch lengths corresponding to the dendrogram branches generated by WINGS when applied to a matrix of PEGASUS p-values for 17,651 genes with a +/- 50kb buffer region and 2,960 intergenic regions on the -$\log_{10}$ scale using both genes and intergenic regions as features. The dashed red horizontal line corresponds to the branch length threshold, where the prioritized clusters are those lying above the dashed line. The corresponding dendrogram is presented in Figure S14.

Figure S16: **WINGS dendrogram from 81 case-control phenotypes using imputed data.** We analyzed a matrix of PEGASUS p-values on the -$\log_{10}$ scale using the imputed data. The dissimilarity index (from [92]) between this tree and Figure 3 is $Z = 0.1379$. The corresponding sorted branch length diagram is presented in Figure S17.

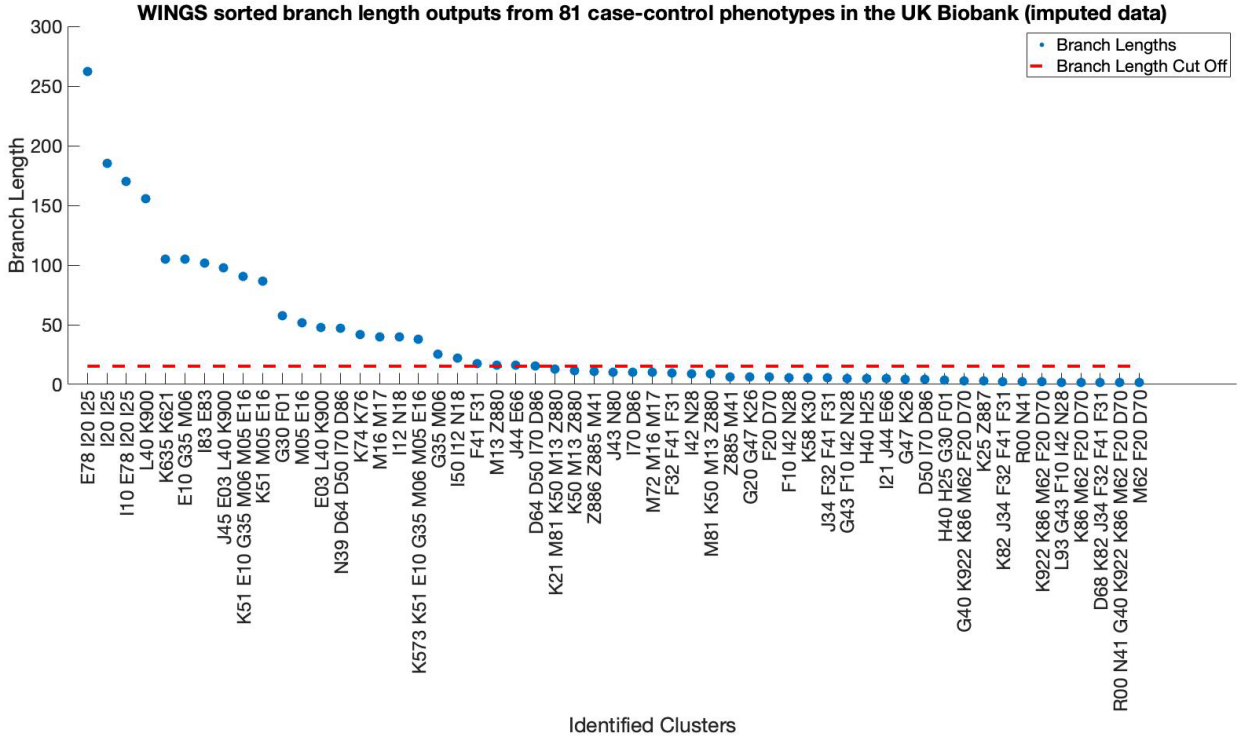

Figure S17: **WINGS sorted branch lengths from 81 case-control phenotypes using imputed data.** We show the sorted branch lengths corresponding to the dendrogram branches generated by WINGS when applied to the -$\log_{10}$ transformed PEGASUS gene scores from 81 case-control phenotypes in the UK Biobank with the imputed data. The dashed red horizontal line corresponds to the branch length threshold, where the prioritized clusters are those lying above the dashed line. The corresponding dendrogram is presented in Figure S16.

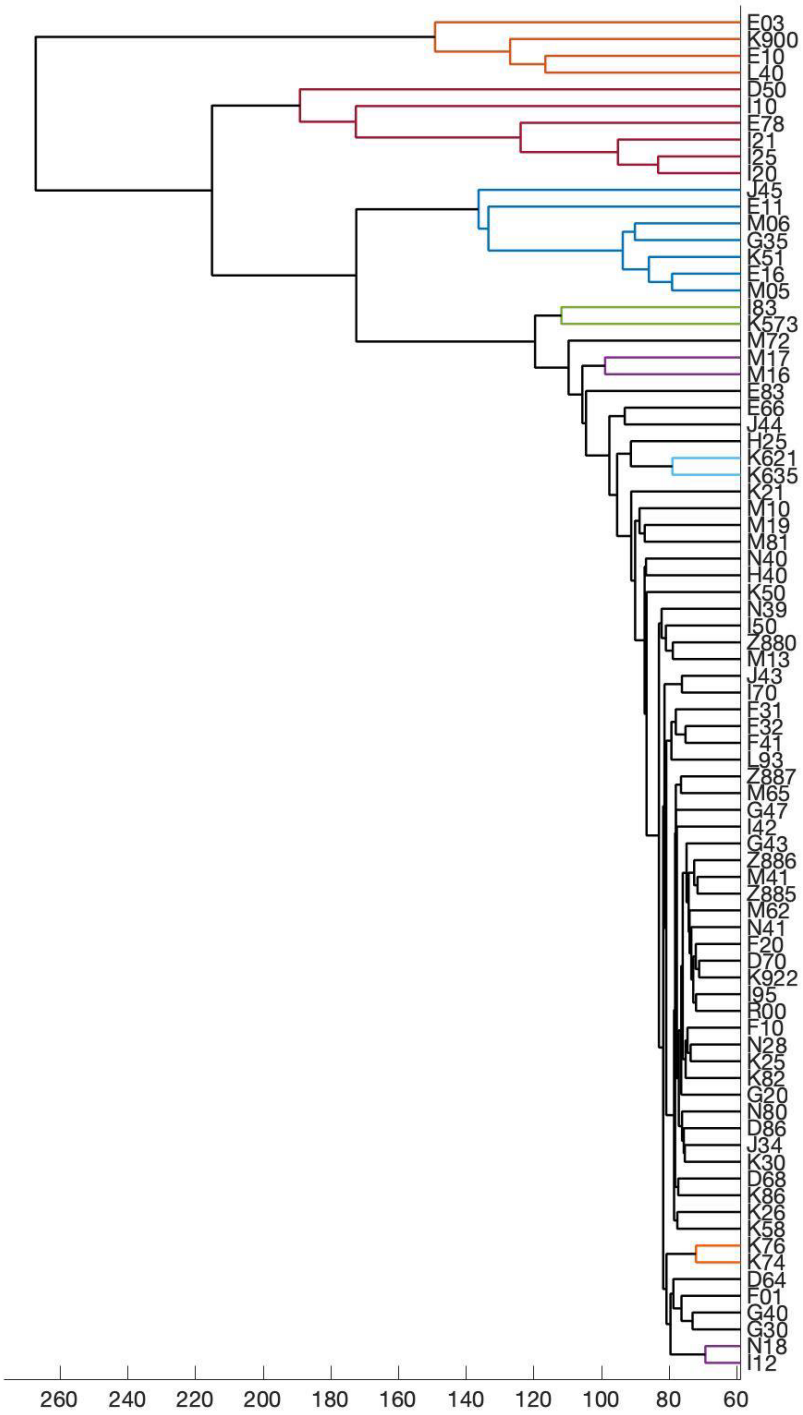

Figure S18: **WINGS dendrogram from 81 case-control phenotypes in the UK Biobank using gene scores with no upstream or downstream buffer region.** We analyzed a matrix of PEGASUS p-values on the -$\log_{10}$ scale using gene scores with no upstream or downstream buffer region. The dissimilarity index (from [92]) between this tree and Figure 3 is $Z = 0.1240$. The corresponding sorted branch length diagram is presented in Figure S19.

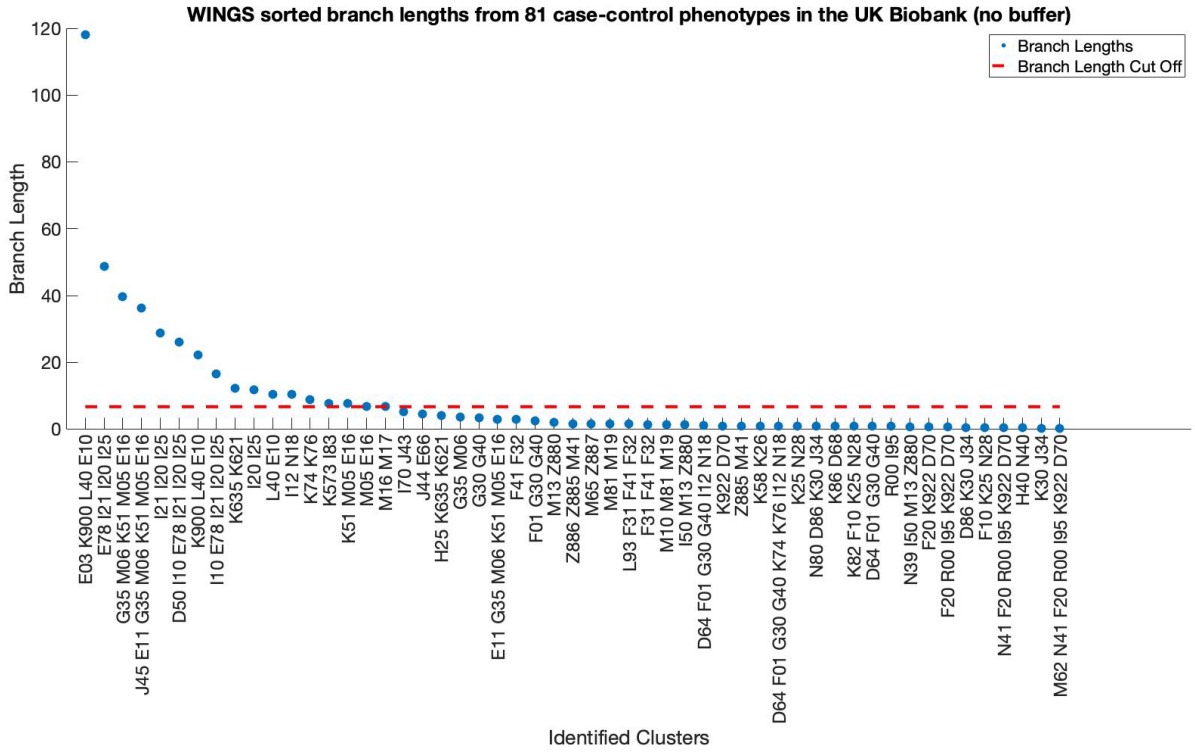

Figure S19: **WINGS sorted branch lengths from 81 case-control phenotypes in the UK Biobank using gene scores with no upstream or downstream buffer region.** We show the sorted branch lengths corresponding to the dendrogram branches generated by WINGS when applied to the -$\log_{10}$ transformed PEGASUS gene scores with no upstream or downstream butter region from 81 case-control phenotypes in the UK Biobank. The dashed red horizontal line corresponds to the branch length threshold, where the prioritized clusters are those lying above the dashed line. The corresponding dendrogram is presented in Figure S18.

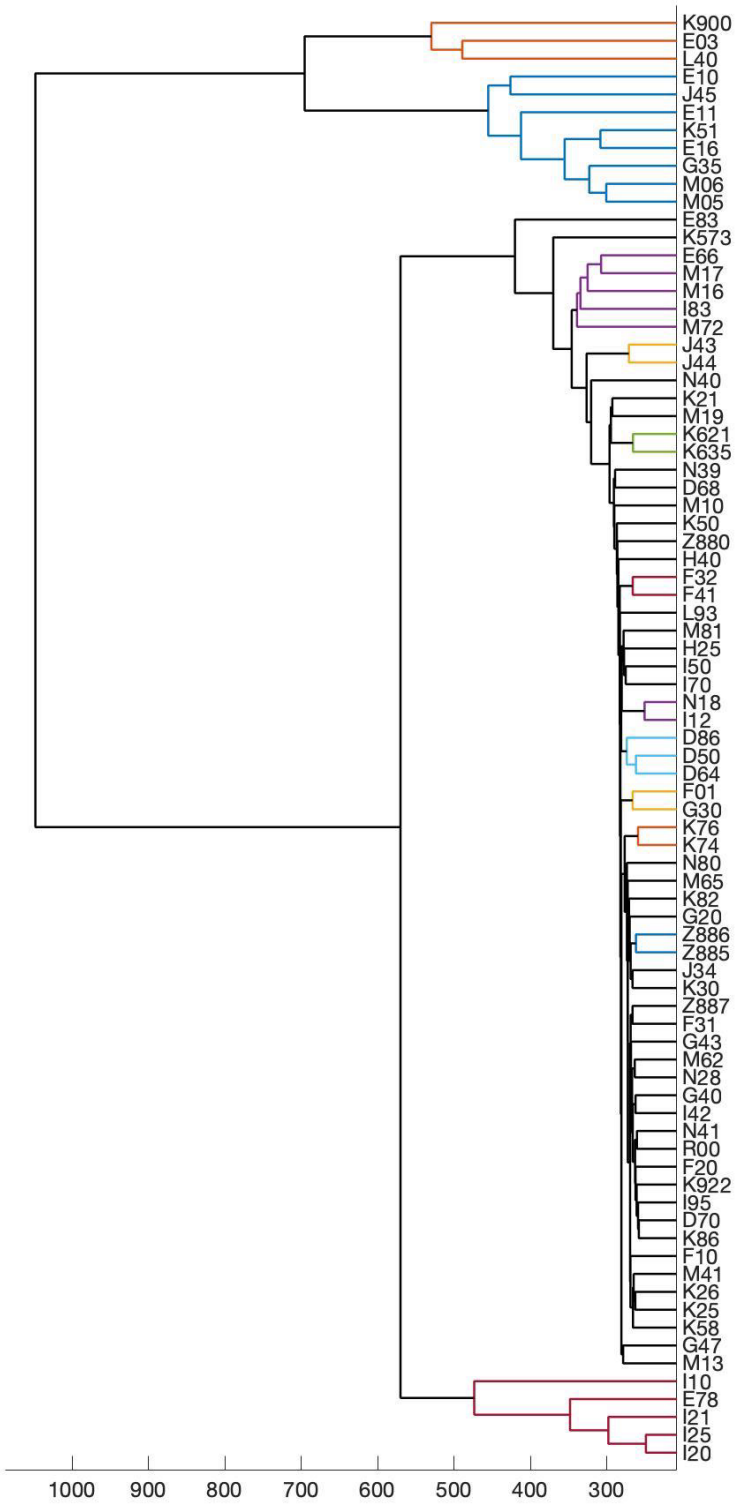

Figure S20: **WINGS dendrogram from 81 case-control phenotypes in the UK Biobank after partitioning the genome into independent haplotype blocks.** We analyzed a matrix of PEGASUS p-values on the -$\log_{10}$ scale using all independent haplotype blocks as features (33,686). The dissimilarity index (from [92]) between this tree and Figure 3 is $Z = 0.1276$. The corresponding sorted branch length diagram is presented in Figure S21.

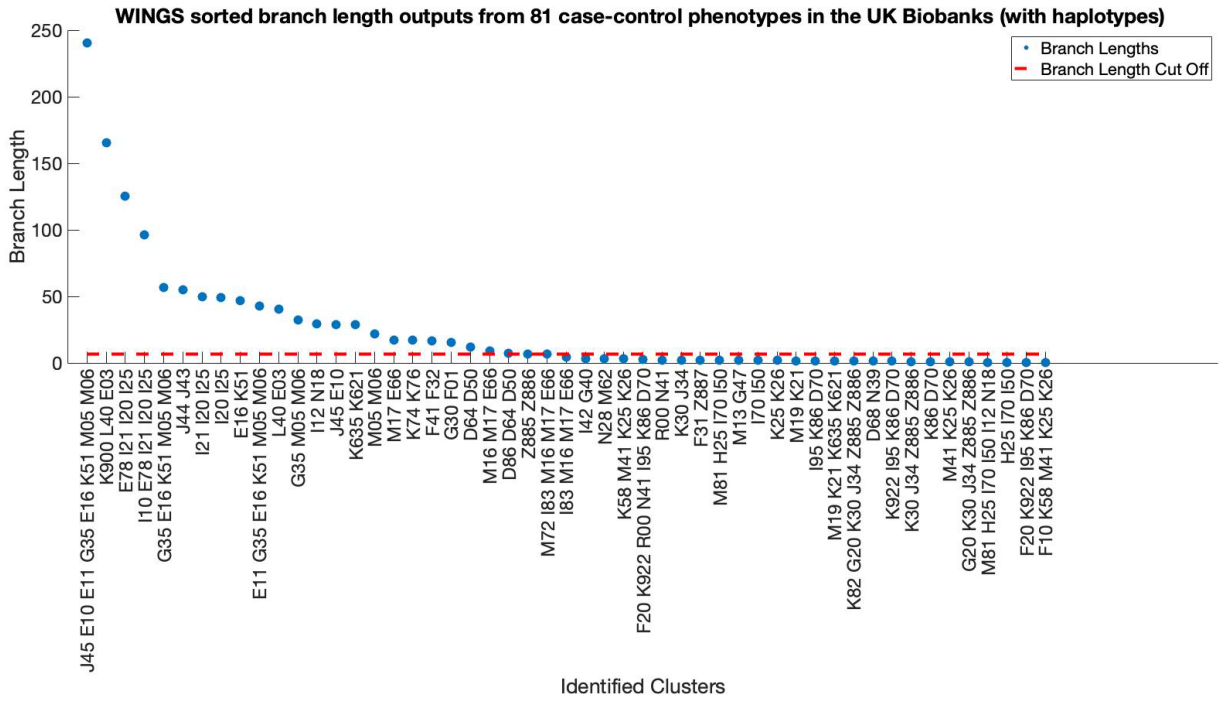

Figure S21: **WINGS sorted branch lengths from 81 case-control phenotypes in the UK Biobank using haplotype blocks as regions.** We show the sorted branch lengths corresponding to the dendrogram branches generated by WINGS when applied to the -$\log_{10}$ transformed PEGASUS gene scores from 81 case-control phenotypes in the UK Biobank for 33,686 independent haplotype regions. The dashed red horizontal line corresponds to the branch length threshold, where the prioritized clusters are those lying above the dashed line. The corresponding dendrogram is presented in Figure S20.

References

[1] Farhad Hormozdiari, Gleb Kichaev, Wen-Yun Yang, Bogdan Pasaniuc, and Eleazar Eskin. Identification of causal genes for complex traits. *Bioinformatics*, 31(12):206–213, 2015.

[2] Huwenbo Shi, Gleb Kichaev, and Bogdan Pasaniuc. Contrasting the genetic architecture of 30 complex traits from summary association data. *The American Journal of Human Genetics*, 99(1):139–153, 2016.

[3] Evan A Boyle, Yang I Li, and Jonathan K Pritchard. An expanded view of complex traits: from polygenic to omnigenic. *Cell*, 169(7):1177–1186, 2017.

[4] Marcus W Feldman and Sohini Ramachandran. Missing compared to what? revisiting heritability, genes and culture. *Phil. Trans. R. Soc. B*, 373(1743):20170064, 2018.

[5] JE Huffman. Examining the current standards for genetic discovery and replication in the era of mega-biobanks. *Nature communications*, 9(1):5054, 2018.

[6] Evangelos Evangelou, Helen R Warren, David Mosen-Ansorena, Borbala Mifsud, Raha Pazoki, He Gao, Georgios Ntritsos, Niki Dimou, Claudia P Cabrera, Ibrahim Karaman, et al. Genetic analysis of over 1 million people identifies 535 new loci associated with blood pressure traits. *Nature genetics*, 50(10):1412, 2018.

[7] Yan Zhang, Guanghao Qi, Ju-Hyun Park, and Nilanjan Chatterjee. Estimation of complex effect-size distributions using summary-level statistics from genome-wide association studies across 32 complex traits. *Nature genetics*, 50(9):1318, 2018.

[8] Clare Bycroft, Colin Freeman, Desislava Petkova, Gavin Band, Lloyd T Elliott, Kevin Sharp, Allan Motyer, Damjan Vukcevic, Olivier Delaneau, Jared O'Connell, et al. Genome-wide genetic data on˜ 500,000 uk biobank participants. *BioRxiv*, page 166298, 2017.

[9] Dan M Roden, Jill M Pulley, Melissa A Basford, Gordon R Bernard, Ellen W Clayton, Jeffrey R Balser, and Dan R Masys. Development of a large-scale de-identified dna biobank to enable personalized medicine. *Clinical Pharmacology & Therapeutics*, 84(3):362–369, 2008.

[10] Joshua C Denny, Marylyn D Ritchie, Melissa A Basford, Jill M Pulley, Lisa Bastarache, Kristin Brown-Gentry, Deede Wang, Dan R Masys, Dan M Roden, and Dana C Crawford. Phewas: demonstrating the feasibility of a phenome-wide scan to discover gene–disease associations. *Bioinformatics*, 26(9):1205–1210, 2010.

[11] Joseph K Pickrell, Tomaz Berisa, Jimmy Z Liu, Laure Ségurel, Joyce Y Tung, and David A Hinds. Detection and interpretation of shared genetic influences on 42 human traits. *Nature genetics*, 48(7):709, 2016.

[12] Farhad Hormozdiari, Martijn van de Bunt, Ayellet V Segrè, Xiao Li, Jong Wha J Joo, Michael Bilow, Jae Hoon Sul, Sriram Sankararaman, Bogdan Pasaniuc, and Eleazar Eskin. Colocalization of gwas and eqtl signals detects target genes. *The American Journal of Human Genetics*, 99(6):1245–1260, 2016.

[13] Joshua C Denny, Lisa Bastarache, Marylyn D Ritchie, Robert J Carroll, Raquel Zink, Jonathan D Mosley, Julie R Field, Jill M Pulley, Andrea H Ramirez, Erica Bowton, et al. Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. *Nature biotechnology*, 31(12):1102, 2013.

[14] Joshua C Denny, Lisa Bastarache, and Dan M Roden. Phenome-wide association studies as a tool to advance precision medicine. *Annual review of genomics and human genetics*, 17:353–373, 2016.

[15] Changjian Jiang and Zhao-Bang Zeng. Multiple trait analysis of genetic mapping for quantitative trait loci. *Genetics*, 140(3):1111–1127, 1995.

[16] Jonathan Marchini, Bryan Howie, Simon Myers, Gil McVean, and Peter Donnelly. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nature genetics*, 39(7):906, 2007.

[17] Manuel AR Ferreira and Shaun M Purcell. A multivariate test of association. *Bioinformatics*, 25(1):132–133, 2008.

[18] Matthew Stephens. A unified framework for association analysis with multiple related phenotypes. *PloS one*, 8(7):e65245, 2013.

[19] Patrick Turley, Raymond K Walters, Omeed Maghzian, Aysu Okbay, James J Lee, Mark Alan Fontana, Tuan Anh Nguyen-Viet, Robbee Wedow, Meghan Zacher, Nicholas A Furlotte, et al. Multi-trait analysis of genome-wide association summary statistics using mtag. *Nature genetics*, 50(2):229, 2018.

[20] Mark DM Leiserson, Jonathan V Eldridge, Sohini Ramachandran, and Benjamin J Raphael. Network analysis of gwas data. *Current Opinion in Genetics Development*, 23(6):602 – 610, 2013. Genetics of system biology.

[21] Peter M Visscher, Sarah E Medland, Manuel AR Ferreira, Katherine I Morley, Gu Zhu, Belinda K Cornes, Grant W Montgomery, and Nicholas G Martin. Assumption-free estimation of heritability from genome-wide identity-by-descent sharing between full siblings. *PLoS genetics*, 2(3):e41, 2006.

[22] Jimmy Z Liu, Allan F Mcrae, Dale R Nyholt, Sarah E Medland, Naomi R Wray, Kevin M Brown, Nicholas K Hayward, Grant W Montgomery, Peter M Visscher, Nicholas G Martin, et al. A versatile gene-based test for genome-wide association studies. *The American Journal of Human Genetics*, 87(1):139–145, 2010.

[23] Michael C Wu, Seunggeun Lee, Tianxi Cai, Yun Li, Michael Boehnke, and Xihong Lin. Rare-variant association testing for sequencing data with the sequence kernel association test. *The American Journal of Human Genetics*, 89(1):82–93, 2011.

[24] David Lamparter, Daniel Marbach, Rico Rueedi, Zoltán Kutalik, and Sven Bergmann. Fast and rigorous computation of gene and pathway scores from snp-based summary statistics. *PLoS computational biology*, 12(1):e1004714, 2016.

[25] Or Zuk, Eliana Hechter, Shamil R Sunyaev, and Eric S Lander. The mystery of missing heritability: Genetic interactions create phantom heritability. *Proceedings of the National Academy of Sciences*, 109(4):1193–1198, 2012.

[26] Diana Chang and Alon Keinan. Principal component analysis characterizes shared pathogenetics from genome-wide association studies. *PLoS computational biology*, 10(9):e1003820, 2014.

[27] Priyanka Nakka, Benjamin J Raphael, and Sohini Ramachandran. Gene and network analysis of common variants reveals novel associations in multiple complex diseases. *Genetics*, 204(2):783–798, 2016.

[28] Priyanka Nakka, Natalie P Archer, Heng Xu, Philip J Lupo, Benjamin J Raphael, Jun J Yang, and Sohini Ramachandran. Novel gene and network associations found for lymphoblastic leukemia using case-control and family-based studies in multi-ethnic populations. *Cancer Epidemiology and Prevention Biomarkers*, pages cebp–0360, 2017.

[29] Nicola Aceto, Aditya Bardia, David T. Miyamoto, Maria C. Donaldson, Ben S. Wittner, Joel A. Spencer, Min Yu, Adam Pely, Amanda Engstrom, Huili Zhu, Brian W. Brannigan, Ravi Kapur, Shannon L. Stott, Toshi Shioda, Sridhar Ramaswamy, David T. Ting, Charles P. Lin, Mehmet Toner, Daniel A. Haber, and Shyamala Maheswaran. Circulating tumor cell clusters are oligoclonal precursors of breast cancer metastasis. *Cell*, 158(5):1110 – 1122, 2014.

[30] Jessica C.S. Brown, Justin Nelson, Benjamin VanderSluis, Raamesh Deshpande, Arielle Butts, Sarah Kagan, Itzhack Polacheck, Damian J. Krysan, Chad L. Myers, and Hiten D. Madhani. Unraveling the biology of a fungal meningitis pathogen using chemical genetics. *Cell*, 159(5):1168 – 1187, 2014.

[31] Inti A. Pagnuco, Juan I. Pastore, Guillermo Abras, Marcel Brun, and Virginia L. Ballarin. Analysis of genetic association using hierarchical clustering and cluster validation indices. *Genomics*, 109(5):438 – 445, 2017.

[32] Peter Langfelder, Bin Zhang, and Steve Horvath. Defining clusters from a hierarchical cluster tree: the dynamic tree cut package for r. *Bioinformatics*, 24(5):719–720, 2008.

[33] Antoine E. Zambelli. A data-driven approach to estimating the number of clusters in hierarchical clustering. *ISCB Comm J*, 5(2809), 2016.

[34] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer, 2009.

[35] Gad Abraham, Yixuan Qiu, and Michael Inouye. Flashpca2: principal component analysis of biobank-scale genotype datasets. *Bioinformatics*, 33(17):2776–2778, 2017.

[36] Christopher C Chang, Carson C Chow, Laurent CAM Tellier, Shashaank Vattikuti, Shaun M Purcell, and James J Lee. Second-generation plink: rising to the challenge of larger and richer datasets. *Gigascience*, 4(1):7, 2015.

[37] Brendan K Bulik-Sullivan, Po-Ru Loh, Hilary K Finucane, Stephan Ripke, Jian Yang, Nick Patterson, Mark J Daly, Alkes L Price, Benjamin M Neale, Schizophrenia Working Group of the Psychiatric Genomics Consortium, et al. Ld score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nature genetics*, 47(3):291, 2015.

[38] Joe H. Ward Jr. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58(301):236–244, 1963.

[39] Joe H. Ward, Jr. and Marion E. Hook. Application of an hierarchical grouping procedure to a problem of grouping profiles. *Educational and Psychological Measurement*, 23(1):69–81, 1963.

[40] Bipul Hossen, Siraj-Ud Doulah, and Aminul Hoque. Methods for evaluating agglomerative hierarchical clustering for gene expression data: A comparative study. *Computational Biology and Bioinformatics*, 3(6):88–94, 2015.

[41] Laura Ferreira and David B. Hitchcock. A comparison of hierarchical methods for clustering functional data. *Communications in Statistics - Simulation and Computation*, 38(9):1925–1949, 2009.

[42] Christophe Leys, Christophe Ley, Olivier Klein, Philippe Bernard, and Laurent Licata. Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *Journal of Experimental Social Psychology*, 49(4):764 – 766, 2013.

[43] Peter J. Rousseeuw and Christophe Croux. Alternatives to the median absolute deviation. *Journal of the American Statistical Association*, 88(424):1273–1283, 1993.

[44] Matlab data import and analysis, 2018. The MathWorks, Natick, MA, USA.

[45] Peter J. Huber. *Robust Statistics*, pages 1248–1251. Springer Berlin Heidelberg, Berlin, Heidelberg, 2011.

[46] Genevieve L Wojcik, WH Linda Kao, and Priya Duggal. Relative performance of gene-and pathway-level methods as secondary analyses for genome-wide association studies. *BMC genetics*, 16(1):34, 2015.

[47] David M. Jordan, Marie Verbanck, and Ron Do. The landscape of pervasive horizontal pleiotropy in human genetic variation is driven by extreme polygenicity of human traits and diseases. *bioRxiv*, 2019.

[48] Brendan Bulik-Sullivan, Hilary K Finucane, Verneri Anttila, Alexander Gusev, Felix R Day, Po-Ru Loh, Laramie Duncan, John RB Perry, Nick Patterson, Elise B Robinson, et al. An atlas of genetic correlations across human diseases and traits. *Nature genetics*, 47(11):1236, 2015.

[49] Gordon Fehringer, Geoffrey Liu, Laurent Briollais, Paul Brennan, Christopher I Amos, Margaret R Spitz, Heike Bickeböller, H Erich Wichmann, Angela Risch, and Rayjean J Hung. Comparison of pathway analysis approaches using lung cancer gwas data sets. *PloS one*, 7(2):e31816, 2012.

[50] Edward Y Chen, Christopher M Tan, Yan Kou, Qiaonan Duan, Zichen Wang, Gabriela Vaz Meirelles, Neil R Clark, and Avi Ma'ayan. Enrichr: interactive and collaborative html5 gene list enrichment analysis tool. *BMC bioinformatics*, 14(1):128, 2013.

[51] Minoru Kanehisa, Miho Furumichi, Mao Tanabe, Yoko Sato, and Kanae Morishima. Kegg: new perspectives on genomes, pathways, diseases and drugs. *Nucleic acids research*, 45(D1):D353–D361, 2016.

[52] Yaron Tomer, Lawrence M Dolan, George Kahaly, Jasmin Divers, Ralph B D'Agostino Jr, Giuseppina Imperatore, Dana Dabelea, Santica Marcovina, Mary Helen Black, Catherine Pihoker, et al. Genome wide identification of new genes and pathways in patients with both autoimmune thyroiditis and type 1 diabetes. *Journal of autoimmunity*, 60:32–39, 2015.

[53] Gordan Lauc, Jennifer E Huffman, Maja Pučić, Lina Zgaga, Barbara Adamczyk, Ana Mužinić, Mislav Novokmet, Ozren Polašek, Olga Gornik, Jasminka Krištić, et al. Loci associated with n-glycosylation of human immunoglobulin g show pleiotropy with autoimmune diseases and haematological cancers. *PLoS genetics*, 9(1):e1003225, 2013.

[54] Kwan-Yeung Lee, Kwong-Sak Leung, Nelson LS Tang, and Man-Hon Wong. Discovering genetic factors for psoriasis through exhaustively searching for significant second order snp-snp interactions. *Scientific reports*, 8(1):15186, 2018.

[55] Adrià Aterido, Juan D Cañete, Jesús Tornero, Carlos Ferrándiz, José Antonio Pinto, Jordi Gratacós, Rubén Queiró, Carlos Montilla, Juan Carlos Torre-Alonso, José J Pérez-Venegas, et al. Genetic variation at the glycosaminoglycan metabolism pathway contributes to the risk of psoriatic arthritis but not psoriasis. *Annals of the rheumatic diseases*, 78(3):355–364, 2019.

[56] Ying Liu, Cynthia Helms, Wilson Liao, Lisa C Zaba, Shenghui Duan, Jennifer Gardner, Carol Wise, Andrew Miner, MJ Malloy, Clive R Pullinger, et al. A genome-wide association study of psoriasis and psoriatic arthritis identifies new disease loci. *PLoS genetics*, 4(4):e1000041, 2008.

[57] Xue-Jun Zhang, Wei Huang, Sen Yang, Liang-Dan Sun, Feng-Yu Zhang, Qi-Xing Zhu, Fu-Ren Zhang, Chi Zhang, Wen-Hui Du, Xiong-Ming Pu, et al. Psoriasis genome-wide association study identifies susceptibility variants within lce gene cluster at 1q21. *Nature genetics*, 41(2):205, 2009.

[58] Patrick Coit, Prashant Kaushik, Liron Caplan, Gail S Kerr, Jessica A Walsh, Maureen Dubreuil, Andreas Reimold, and Amr H Sawalha. Genome-wide dna methylation analysis in ankylosing spondylitis identifies hla-b* 27 dependent and independent dna methylation changes in whole blood. *Journal of autoimmunity*, 2019.

[59] Patrick CA Dubois, Gosia Trynka, Lude Franke, Karen A Hunt, Jihane Romanos, Alessandra Curtotti, Alexandra Zhernakova, Graham AR Heap, Róza Ádány, Arpo Aromaa, et al. Multiple common variants for celiac disease influencing immune gene expression. *Nature genetics*, 42(4):295, 2010.

[60] David A Van Heel, Lude Franke, Karen A Hunt, Rhian Gwilliam, Alexandra Zhernakova, Mike Inouye, Martin C Wapenaar, Martin CNM Barnardo, Graeme Bethel, Geoffrey KT Holmes, et al. A genome-wide association study for celiac disease identifies risk variants in the region harboring il2 and il21. *Nature genetics*, 39(7):827, 2007.

[61] Malin Östensson, Caroline Montén, Jonas Bacelis, Audur H Gudjonsdottir, Svetlana Adamovic, Johan Ek, Henry Ascher, Elisabet Pollak, Henrik Arnell, Lars Browaldh, et al. A possible mechanism behind autoimmune disorders discovered by genome-wide linkage and association analysis in celiac disease. *PLoS One*, 8(8):e70174, 2013.

[62] Amber Dahlin, Joanne E Sordillo, John Ziniti, Carlos Iribarren, Meng Lu, Scott T Weiss, Kelan G Tantisira, Quan Lu, Mengyuan Kan, Blanca E Himes, et al. Large-scale, multiethnic genome-wide association study identifies novel loci contributing to asthma susceptibility in adults. *Journal of Allergy and Clinical Immunology*, 143(4):1633–1635, 2019.

[63] Xia Jiang, Henrik Källberg, Zuomei Chen, Lisbeth Ärlestig, Solbritt Rantapää-Dahlqvist, Sonia Davila, Lars Klareskog, Leonid Padyukov, and Lars Alfredsson. An immunochip-based interaction study of contrasting interaction effects with smoking in acpa-positive versus acpa-negative rheumatoid arthritis. *Rheumatology*, 55(1):149–155, 2015.

[64] Jason D Cooper, Deborah J Smyth, Adam M Smiles, Vincent Plagnol, Neil M Walker, James E Allen, Kate Downes, Jeffrey C Barrett, Barry C Healy, Josyf C Mychaleckyj, et al. Meta-analysis of genome-wide association study data identifies additional type 1 diabetes risk loci. *Nature genetics*, 40(12):1399, 2008.

[65] Nick Shrine, Michael A Portelli, Catherine John, María Soler Artigas, Neil Bennett, Robert Hall, Jon Lewis, Amanda P Henry, Charlotte K Billington, Azaz Ahmad, et al. Moderate-to-severe asthma in individuals of european ancestry: a genome-wide association study. *The Lancet Respiratory Medicine*, 7(1):20–34, 2019.

[66] Florence Demenais, Patricia Margaritte-Jeannin, Kathleen C Barnes, William OC Cookson, Janine Altmüller, Wei Ang, R Graham Barr, Terri H Beaty, Allan B Becker, John Beilby, et al. Multiancestry association study identifies new asthma risk loci that colocalize with immune-cell enhancer marks. *Nature genetics*, 50(1):42, 2018.

[67] Tomomitsu Hirota, Atsushi Takahashi, Michiaki Kubo, Tatsuhiko Tsunoda, Kaori Tomita, Satoru Doi, Kimie Fujita, Akihiko Miyatake, Tadao Enomoto, Takehiko Miyagawa, et al. Genome-wide association study identifies three new susceptibility loci for adult asthma in the japanese population. *Nature genetics*, 43(9):893, 2011.

[68] Berta Almoguera, Lyam Vazquez, Frank Mentch, John Connolly, Jennifer A Pacheco, Agnes S Sundaresan, Peggy L Peissig, James G Linneman, Catherine A McCarty, David Crosslin, et al. Identification of four novel loci in asthma in european american and african american populations. *American journal of respiratory and critical care medicine*, 195(4):456–463, 2017.

[69] Eun-Heui Jin Seon-Hee Yim So-Young Yang Seung-Hyun Jung Seung-Hun Shin Wan-Uk Kim Seung-Cheol Shim Tai-Gyu Kim Hu, Hae-Jin and Yeun-Jun Chung. Common variants at the promoter region of the apom confer a risk of rheumatoid arthritis. *Experimental molecular medicine*, 43(11):613, 2011.

[70] Lei Jiang, Jian Yin, Lingying Ye, Jian Yang, Gibran Hemani, Ai-jun Liu, Hejian Zou, Dongyi He, Lingyun Sun, Xiaofeng Zeng, et al. Novel risk loci for rheumatoid arthritis in han chinese and congruence with risk variants in europeans. *Arthritis & rheumatology*, 66(5):1121–1132, 2014.

[71] Akul Singhania, Hitasha Rupani, Nivenka Jayasekera, Simon Lumb, Paul Hales, Neil Gozzard, Donna E Davies, Christopher H Woelk, and Peter H Howarth. Altered epithelial gene expression in peripheral airways of severe asthma. *PloS one*, 12(1):e0168680, 2017.

[72] Edith Chen, GE Miller, HA Walker, JM Arevalo, CY Sung, and SW Cole. Genome-wide transcriptional profiling linked to social class in asthma. *Thorax*, 64(1):38–43, 2009.

[73] Gwan Gyu Song and Young Ho Lee. Pathway analysis of genome-wide association study on asthma. *Human immunology*, 74(2):256–260, 2013.

[74] Nada A Al-Tassan, Nicola Whiffin, Fay J Hosking, Claire Palles, Susan M Farrington, Sara E Dobbins, Rebecca Harris, Maggie Gorman, Albert Tenesa, Brian F Meyer, et al. A new gwas and meta-analysis with 1000genomes imputation identifies novel risk variants for colorectal cancer. *Scientific reports*, 5:10442, 2015.

[75] Nicola Whiffin, Fay J Hosking, Susan M Farrington, Claire Palles, Sara E Dobbins, Lina Zgaga, Amy Lloyd, Ben Kinnersley, Maggie Gorman, Albert Tenesa, et al. Identification of susceptibility loci for colorectal cancer in a genome-wide meta-analysis. *Human molecular genetics*, 23(17):4729–4737, 2014.

[76] COGENT Study et al. Meta-analysis of genome-wide association data identifies four new susceptibility loci for colorectal cancer. *Nature genetics*, 40(12):1426, 2008.

[77] Philipp Hofer, Michael Hagmann, Stefanie Brezina, Erich Dolejsi, Karl Mach, Gernot Leeb, Andreas Baierl, Stephan Buch, Hedwig Sutterlüty-Fall, Judith Karner-Hanusch, et al. Bayesian and frequentist analysis of an austrian genome-wide association study of colorectal cancer and advanced adenomas. *Oncotarget*, 8(58):98623, 2017.

[78] Stephanie L Schmit, Christopher K Edlund, Fredrick R Schumacher, Jian Gong, Tabitha A Harrison, Jeroen R Huyghe, Chenxu Qu, Marilena Melas, David J Van Den Berg, Hansong Wang, et al. Novel common genetic susceptibility loci for colorectal cancer. *JNCI: Journal of the National Cancer Institute*, 111(2):146–157, 2018.

[79] Richard S Houlston, Jeremy Cheadle, Sara E Dobbins, Albert Tenesa, Angela M Jones, Kimberley Howarth, Sarah L Spain, Peter Broderick, Enric Domingo, Susan Farrington, et al. Meta-analysis of three genome-wide association studies identifies susceptibility loci for colorectal cancer at 1q41, 3q26. 2, 12q13. 13 and 20q13. 33. *Nature genetics*, 42(11):973, 2010.

[80] Yingchang Lu, Sun-Seog Kweon, Chizu Tanikawa, Wei-Hua Jia, Yong-Bing Xiang, Qiuyin Cai, Chenjie Zeng, Stephanie L Schmit, Aesun Shin, Keitaro Matsuo, et al. Large-scale genome-wide association study of east asians identifies loci associated with risk for colorectal cancer. *Gastroenterology*, 156(5):1455–1466, 2019.

[81] Cyril Pottier, Xiaolai Zhou, Ralph B Perkerson III, Matt Baker, Gregory D Jenkins, Daniel J Serie, Roberta Ghidoni, Luisa Benussi, Giuliano Binetti, Adolfo López de Munain, et al. Potential genetic modifiers of disease risk and age at onset in patients with frontotemporal lobar degeneration and grn mutations: a genome-wide association study. *The Lancet Neurology*, 17(6):548–558, 2018.

[82] Riccardo E Marioni, Sarah E Harris, Qian Zhang, Allan F McRae, Saskia P Hagenaars, W David Hill, Gail Davies, Craig W Ritchie, Catharine R Gale, John M Starr, et al. Gwas on family history of alzheimer's disease. *Translational psychiatry*, 8, 2018.

[83] Vijay K Ramanan, Shannon L Risacher, Kwangsik Nho, Sungeun Kim, Shanker Swaminathan, Li Shen, Tatiana M Foroud, Hakon Hakonarson, Matthew J Huentelman, Paul S Aisen, et al. Apoe and bche as modulators of cerebral amyloid deposition: a florbetapir pet genome-wide association study. *Molecular psychiatry*, 19(3):351, 2014.

[84] Iris E Jansen, Jeanne E Savage, Kyoko Watanabe, Julien Bryois, Dylan M Williams, Stacy Steinberg, Julia Sealock, Ida K Karlsson, Sara Hägg, Lavinia Athanasiu, et al. Genome-wide meta-analysis identifies new loci and functional pathways influencing alzheimer's disease risk. 2019.

[85] Gyungah R Jun, Jaeyoon Chung, Jesse Mez, Robert Barber, Gary W Beecham, David A Bennett, Joseph D Buxbaum, Goldie S Byrd, Minerva M Carrasquillo, Paul K Crane, et al. Transethnic genome-wide scan identifies novel alzheimer's disease loci. *Alzheimer's & Dementia*, 13(7):727–738, 2017.

[86] Carlos Cruchaga, Petra Nowotny, John SK Kauwe, Perry G Ridge, Kevin Mayo, Sarah Bertelsen, Anthony Hinrichs, Anne M Fagan, David M Holtzman, John C Morris, et al. Association and expression analyses with single-nucleotide polymorphisms in tomm40 in alzheimer disease. *Archives of neurology*, 68(8):1013–1019, 2011.

[87] Takahisa Kanekiyo, Huaxi Xu, and Guojun Bu. Apoe and aβ in alzheimer's disease: accidental encounters or partners? *Neuron*, 81(4):740–754, 2014.

[88] Anna Köttgen, Eva Albrecht, Alexander Teumer, Veronique Vitart, Jan Krumsiek, Claudia Hundertmark, Giorgio Pistis, Daniela Ruggiero, Conall M O'Seaghdha, Toomas Haller, et al. Genome-wide association analyses identify 18 new loci associated with serum urate concentrations. *Nature genetics*, 45(2):145, 2013.

[89] Takahisa Kawaguchi, Toshihide Shima, Masayuki Mizuno, Yasuhide Mitsumoto, Atsushi Umemura, Yoshihiro Kanbara, Saiyu Tanaka, Yoshio Sumida, Kohichiro Yasui, Meiko Takahashi, et al. Risk estimation model for nonalcoholic fatty liver disease in the japanese using multiple genetic markers. *PloS one*, 13(1):e0185490, 2018.

[90] Takuya Kitamoto, Aya Kitamoto, Masato Yoneda, Hideyuki Hyogo, Hidenori Ochi, Takahiro Nakamura, Hajime Teranishi, Seiho Mizusawa, Takato Ueno, Kazuaki Chayama, et al. Genome-wide scan revealed that polymorphisms in the pnpla3, samm50, and parvb genes are associated with development and progression of nonalcoholic fatty liver disease in japan. *Human genetics*, 132(7):783–792, 2013.

[91] Goh Eun Chung, Young Lee, Jeong Yoon Yim, Eun Kyung Choe, Min-Sun Kwak, Jong In Yang, Boram Park, Jong-Eun Lee, Jeong A Kim, and Joo Sung Kim. Genetic polymorphisms of pnpla3 and samm50 are associated with nonalcoholic fatty liver disease in a korean population. *Gut and liver*, 12(3):316, 2018.

[92] Isabella Morlini and Sergio Zani. Dissimilarity and similarity measures for comparing dendrograms and their applications. *Advances in Data Analysis and Classification*, 6(2):85–105, Jul 2012.

[93] Frank R Grubbs. Sample criteria for testing outlying observations. *The Annals of Mathematical Statistics*, 21(1):27–58, 1950.

[94] Xiang Zhu and Matthew Stephens. Large-scale genome-wide enrichment analyses identify new trait-associated genes and pathways across 31 human phenotypes. *Nature communications*, 9(1):4361, 2018.

[95] Minghui Wang, Jianfei Huang, Yiyuan Liu, Li Ma, James B Potash, and Shizhong Han. Combat: a combined association test for genes using summary statistics. *Genetics*, 207(3):883–891, 2017.

[96] Peter Carbonetto and Matthew Stephens. Integrated enrichment analysis of variants and pathways in genome-wide association studies indicates central role for il-2 signaling genes in type 1 diabetes, and cytokine signaling genes in crohn's disease. *PLoS genetics*, 9(10):e1003770, 2013.

[97] Sasha Bozeat, Carol A Gregory, Matthew A Lambon Ralph, and John R Hodges. Which neuropsychiatric and behavioural features distinguish frontal and temporal variants of frontotemporal dementia from alzheimer's disease? *Journal of Neurology, Neurosurgery & Psychiatry*, 69(2):178–186, 2000.

[98] Richard J Perrin, Anne M Fagan, and David M Holtzman. Multimodal techniques for diagnosis and prognosis of alzheimer's disease. *Nature*, 461(7266):916, 2009.

[99] Matthew Stephens. False discovery rates: a new deal. *Biostatistics*, 18(2):275–294, 2016.

[100] George Hripcsak, Matthew E Levine, Ning Shang, and Patrick B Ryan. Effect of vocabulary mapping for conditions on phenotype cohorts. *Journal of the American Medical Informatics Association*, 25(12):1618–1625, 2018.

[101] Chaitanya Shivade, Preethi Raghavan, Eric Fosler-Lussier, Peter J Embi, Noemie Elhadad, Stephen B Johnson, and Albert M Lai. A review of approaches to identifying patient phenotype cohorts using electronic health records. *Journal of the American Medical Informatics Association*, 21(2):221–230, 2013.

[102] Bhautesh Dinesh Jani, Peter Hanlon, Barbara I. Nicholl, Ross McQueenie, Katie I. Gallacher, Duncan Lee, and Frances S. Mair. Relationship between multimorbidity, demographic factors and mortality: findings from the uk biobank cohort. 17(1):74, 2019.

[103] Alexandra Havdahl, Ruth Mitchell, Lavinia Paternoster, and George Smith Davey. Investigating causality in the association between vitamin d status and self-reported tiredness. *Scientific reports*, 9(1):2880, 2019.

[104] Mihaela E Sardiu, Joshua M. Gilmore, Brad D. Groppe, Arnob Dutta, Laurence Florens, and Michael P. Washburn. Topological scoring of protein interaction networks. *Nature Communications*, 10(1118), 2019.

[105] Alicia R Martin, Christopher R Gignoux, Raymond K Walters, Genevieve L Wojcik, Benjamin M Neale, Simon Gravel, Mark J Daly, Carlos D Bustamante, and Eimear E Kenny. Human demographic history impacts genetic risk prediction across diverse populations. *The American Journal of Human Genetics*, 100(4):635–649, 2017.

[106] Ani Manichaikul, Josyf C Mychaleckyj, Stephen S Rich, Kathy Daly, Michèle Sale, and Wei-Min Chen. Robust relationship inference in genome-wide association studies. *Bioinformatics*, 26(22):2867–2873, 2010.

[107] Jian Yang, Beben Benyamin, Brian P McEvoy, Scott Gordon, Anjali K Henders, Dale R Nyholt, Pamela A Madden, Andrew C Heath, Nicholas G Martin, Grant W Montgomery, et al. Common snps explain a large proportion of the heritability for human height. *Nature genetics*, 42(7):565, 2010.

[108] Andrew R Wood, Tonu Esko, Jian Yang, Sailaja Vedantam, Tune H Pers, Stefan Gustafsson, Audrey Y Chu, Karol Estrada, Jian'an Luan, Zoltán Kutalik, et al. Defining the role of common variation in the genomic and biological architecture of adult human height. *Nature genetics*, 46(11):1173, 2014.