Systematic single-variant and gene-based association testing of 3,700 phenotypes in 281,850 UK Biobank exomes

Konrad J. Karczewski[1,2,3,†], Matthew Solomonson[1,2,†], Katherine R. Chao[1,2,†], Julia K. Goodrich[1,2], Grace Tiao[1,2], Wenhan Lu[1,2,3], Bridget M. Riley-Gillis[4], Ellen A. Tsai[5], Hye In Kim[6], Xiuwen Zheng[4], Fedik Rahimov[4], Sahar Esmaeeli[4], A. Jason Grundstad[4], Mark Reppell[4], Jeff Waring[4], Howard Jacob[4], David Sexton[5], Paola G. Bronson[5], Xing Chen[6], Xinli Hu[6], Jacqueline I. Goldstein[1,2,3], Daniel King[1,2,3], Christopher Vittal[1,2,3], Timothy Poterba[1,2,3], Duncan S. Palmer[1,2,3], Claire Churchhouse[1,2,3], Daniel P. Howrigan[1,2,3], Wei Zhou[1,2], Nicholas A. Watts[1,2], Kevin Nguyen[1,2], Huy Nguyen[1,2], Cara Mason[7], Christopher Farnham[7], Charlotte Tolonen[7], Laura D. Gauthier[7], Namrata Gupta[7], Daniel G. MacArthur[1,2,8,9], Heidi L. Rehm[1,2], Cotton Seed[1,2,3], Anthony A. Philippakis[7], Mark J. Daly[1,2,3,10], J. Wade Davis[4,*], Heiko Runz[5,*], Melissa R. Miller[6,*], Benjamin M. Neale[1,2,3,*]

[1]Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, Massachusetts 02142, USA
[2]Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, Massachusetts 02114, USA
[3]Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, Massachusetts 02142, USA
[4]Genomics Research Center, AbbVie, North Chicago, Illinois, 60064, USA
[5]Biogen, Inc, Cambridge, Massachusetts 02142, USA
[6]Worldwide Research Development and Medical, Pfizer, Inc, Cambridge, Massachusetts 02139, USA
[7]Data Sciences Platform, Broad Institute of MIT and Harvard, Cambridge, Massachusetts 02142, USA
[8]Present address: Centre for Population Genomics, Garvan Institute of Medical Research and UNSW, Sydney, NSW, Australia
[9]Present address: Murdoch Children's Research Institute, Parkville, VIC, Australia
[10]Institute for Molecular Medicine Finland, Helsinki, Finland
†Denotes equal contribution
*Denotes equal contribution

Abstract (125 words)

Genome-wide association studies have successfully discovered thousands of common variants associated with human diseases and traits, but the landscape of rare variation in human disease has not been explored at scale. Exome sequencing studies of population biobanks provide an opportunity to systematically evaluate the impact of rare coding variation across a wide range of phenotypes to discover genes and allelic series relevant to human health and disease. Here, we present results from systematic association analyses of 3,700 phenotypes using single-variant and gene tests of 281,850 individuals in the UK Biobank with exome sequence data. We find that the discovery of genetic associations is tightly linked to frequency as well as correlated with metrics of deleteriousness and natural selection. We highlight biological findings elucidated by these data and release the dataset as a public resource alongside a browser framework for rapidly exploring rare variant association results.

Introduction

Coding variation has been the most readily interpretable class of genomic variation since the development of the gene model and mapping of the human genome. As such, it has facilitated the mapping and interpretation of variants with immediate clinical importance such as the American College of Medical Genetics actionable variant list (*1*). More recently, exome sequencing has yielded the discovery of specific causal variants for hundreds of rare diseases, particularly dominant acting *de novo* variants for severe diseases (*2*).

As the sample sizes of exome sequencing datasets continue to grow, so do the opportunities to identify associations between rare variants and phenotypes (both complex traits and diseases). In complex diseases, identifying causal genetic factors for a given disease can provide direct insight into the potential for therapeutic avenues. For instance, gain-of-function variants in *PCSK9* have been demonstrated to increase LDL levels and thus risk for cardiovascular disease (*3*). Accordingly, loss-of-function (LoF) variants are protective for cardiovascular disease (*4*), and less than 15 years after the discovery of this effect, therapeutic approaches to inhibit PCSK9 have been brought to market (*5*).

Deeply phenotyped biobanks present a unique opportunity to simultaneously analyze multiple diseases and traits within a single cohort. These datasets enable the discovery of new disease genes with therapeutic potential at a large scale across phenotypes. For instance, a recent joint association analysis of the UK Biobank and the FinnGen study cohorts revealed rare variants in *ANGPTL7* that protect against glaucoma (*6*). The UK Biobank is a collection of over 500,000 participants with standardized, detailed phenotypic data on which GWAS have been run extensively. Recent studies have leveraged previous releases of the exome sequence data to explore various aspects of rare variant associations in this dataset (*7*–*9*). Here, we present and publicly release results from a systematic, large-scale rare variant association analysis of 3,700 phenotypes in more than 300,000 exome sequenced individuals.

Generating high-quality exome data for rare variant associations

We built an end-to-end pipeline for read mapping, processing, joint variant calling, quality control, and mixed model association analysis, and applied this pipeline to 302,325 individuals with exome sequence data from the UK Biobank. The read mapping and processing pipeline adopted the GATK Best Practices pipeline (GRCh38), and the resulting gVCF files were joint-called using a scalable implementation in Hail (Supplementary Information; Fig. S1) (*10*). We processed a set of 3,700 phenotypes including 1,117 quantitative traits as well as 2,583 binary traits with at least 200 cases, which included 681 disease endpoints based on ICD-10 codes (Fig. S2).

After performing quality control (QC) in a similar but augmented (e.g. array concordance; see Supplementary Information) manner as for the Genome Aggregation Database (gnomAD) (*11*), we generated a high-quality dataset of 286,310 individuals (Figs. S3 to S5; table S1), including 281,850 individuals of European ancestry in which we find 20,343,543 high-quality variants (Fig. S6). For each of 19,591 protein-coding genes, we considered up to three functional annotation categories: predicted LoF (pLoF), missense (including low-confidence pLoF variants and in-frame indels), and synonymous, resulting in 57,650 groups for association testing (i.e., one group per gene and functional annotation category).

Creating a high-quality set of rare variant associations

We performed group tests using the mixed model framework SAIGE-GENE (*12*), which includes single-variant tests and gene-based burden (mean) and SKAT-O (hybrid variance/mean) tests (Fig. S7). In total, we performed up to 7,575,993 single-variant tests and 57,650 group tests for each of 3,700 phenotypes (Fig. 1). Additionally, we randomly generated 314 heritable phenotypes to test the asymptotic properties of the mixed-model association testing framework (Figs. S8 to S9), and to determine empirical p-value thresholds for Type I error control. Based on this analysis, for each phenotype, we consider genome-wide p-value

thresholds of 2.5 x 10^{-8} for SKAT-O tests, 6.7 x 10^{-7} for burden tests, and 8 x 10^{-9} for single-variant tests (see Supplementary Information; Fig. S10), corresponding to approximately 0.05 expected false positives per phenotype.

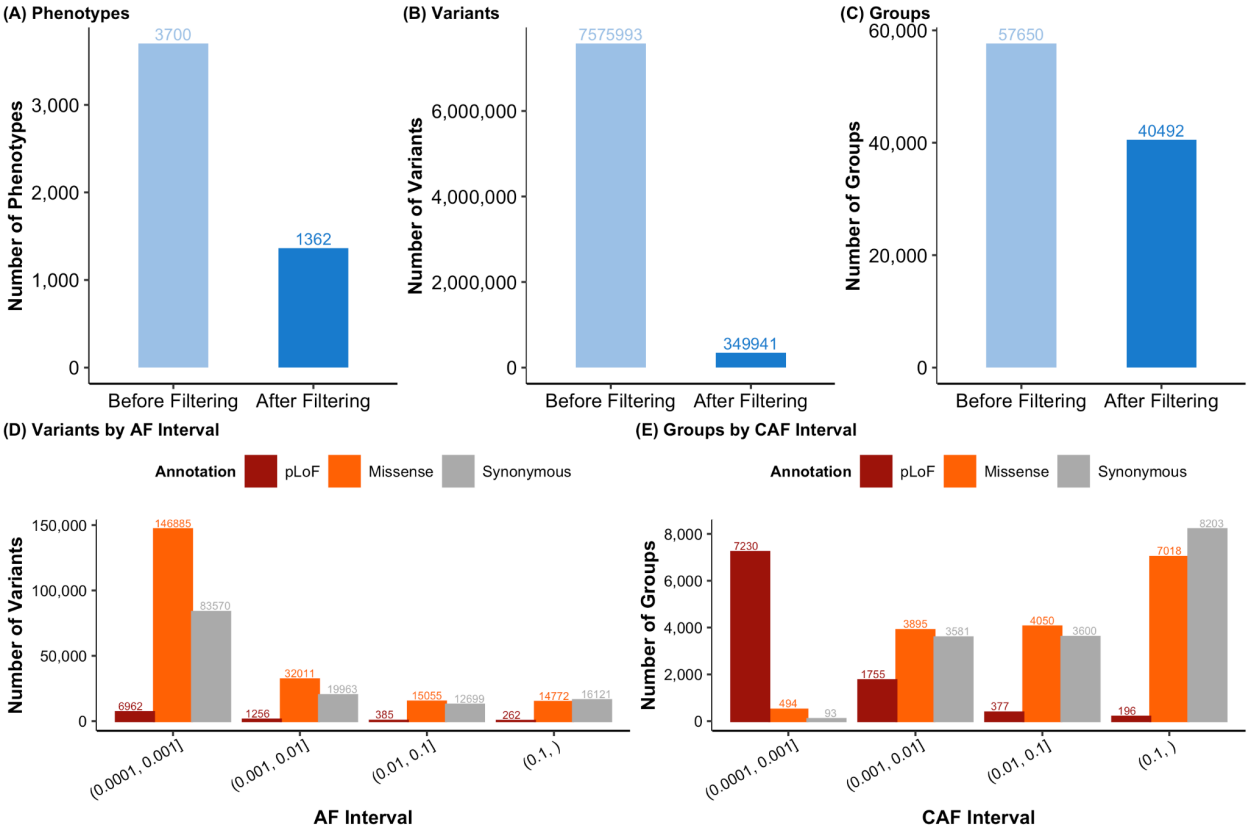


Figure 1 | Quality control of rare variant association tests. The number of phenotypes (**A**), variants (**B**), and groups (i.e., gene-annotation pairs; **C**) before and after quality control. After quality control, the number of variants (**D**) and genes (**E**) are broken down by annotation and frequency bin (alternate allele frequency for variants, cumulative allele frequency for genes).

We performed extensive quality control on these summary statistics (Fig. 1; Table S2; Supplementary Information), including an 0.01% minor and cumulative allele frequency filter for variants and genes, respectively, as well as genomic control (lambda GC) for each phenotype and each gene (Figs. S11 to S15). Further, we pruned to a set of 1,362 high-quality independent phenotypes encompassing 559 continuous traits and 803 binary traits, including 310 ICD codes (Figs. 1A, S16; Table S2). We confirm the robustness of our results by comparing them to a previous large-scale study of height (Tables S3 to S5, Fig. S17) and red

blood cell phenotypes (Table S6), for which our analysis replicates the majority of associations with consistent direction of effect (*13*, *14*).

We filtered to 349,941 variants, including 8,865 pLoF variants, 208,723 missense variants, and 132,353 synonymous variants with a cohort frequency of at least 0.01% (corresponding to an allele count of approximately 60; Fig. 1B). For group tests, we filter to a high-quality set of 40,492 gene tests with at least 20X coverage (Fig. S13) and an aggregate allele frequency of at least 0.01% for pLoF (N=9,558 genes), missense (N=15,457), and synonymous (N=15,477) (Fig. 1C).

Using these criteria, we identified a total of 27,421, and 4,560 associations meeting our p-value threshold, with a mean of 20.1 and 3.35 associations per phenotype, for single-variant tests and group tests, respectively (disease results shown in Fig. 2A-B). Comparing the group test results to single-variant association test results, we find 1,069 associations (on average 0.8 per phenotype) from group tests where no association reached our p-value threshold for any single variant in the corresponding gene (Fig. 2C).

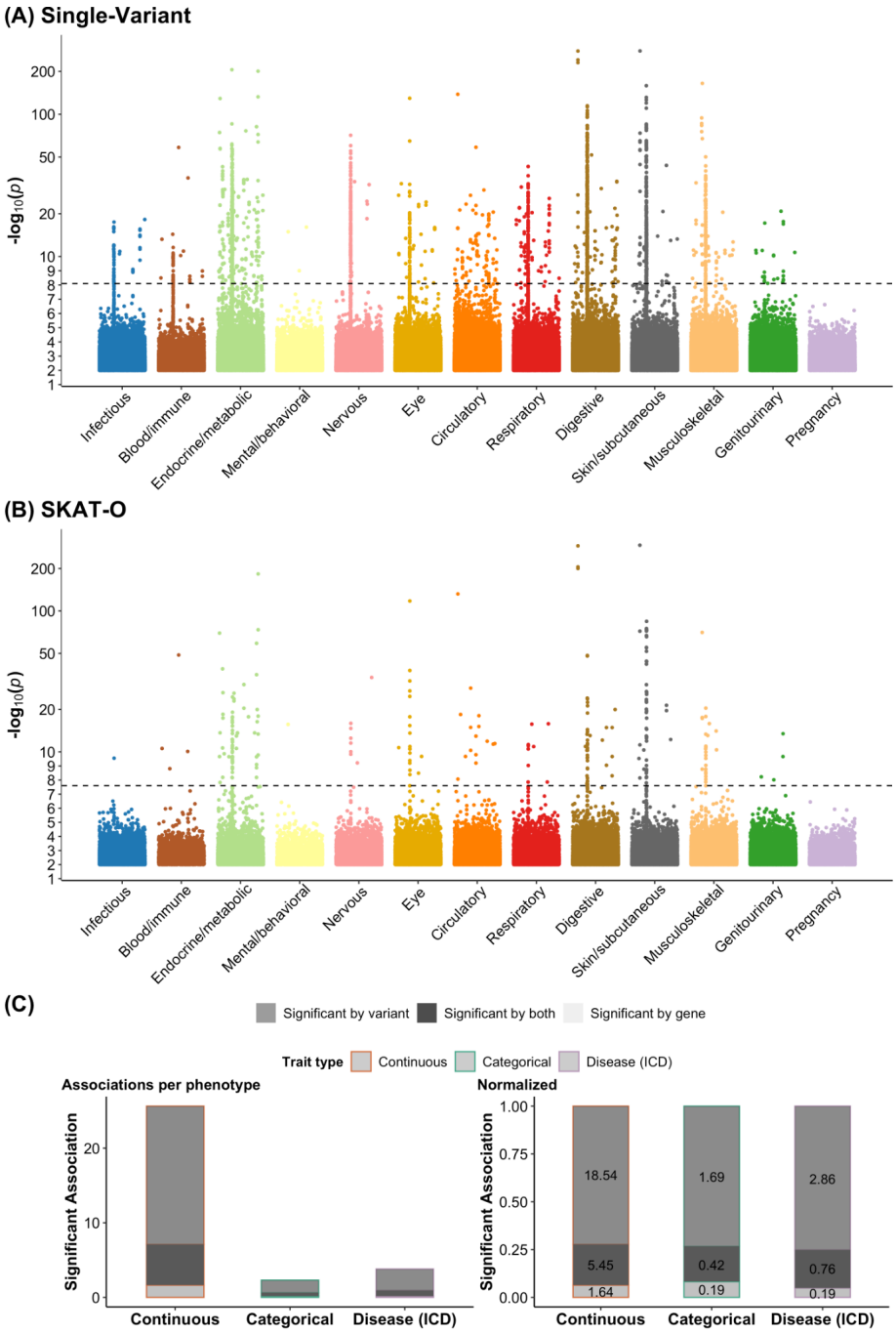


Figure 2 | Rare variant association testing is enhanced by group tests. (**A-B),** Manhattan plots depict the distribution of p-values for all single-variant (**A**) and SKAT-O gene-based (**B**) associations, where the minimum p-value across phenotypes in each ICD chapter is shown. (**C**): The number of gene-level associations per phenotype is shown as a barplot, broken down by trait type (left) and normalized within each trait type (right). The single-variant tests are grouped into genes where at least one associated variant is necessary to be "Significant by variant" which is shown alongside group tests ("Significant by gene") as well as genes where an association is found both for group and single-variant tests.

Displaying rare variant associations

The utility of human genetic variation datasets are substantially enhanced when made accessible in the form of online portals that enable non-technical domain experts to quickly browse, interpret, and export results for downstream follow-up (*15*). We extended our gnomAD browser toolkit to create the genebass (gene-biobank association summary statistics) browser (https://genebass.org), a new, highly interactive tool for exploring large numbers of gene-based PheWAS analysis results. This resource provides users with direct access to all 3,700 phenotypes, serving up 609,538,421 gene-level association statistics (across 19,591 genes, 3 annotation sets, and 3 burden tests) and 20,800,574,337 single variant association statistics across 7,575,993 exome variants. For completeness, the released dataset includes all association statistics, including pre-quality controlled data, but we provide functionality to filter to only the highest quality data presented herein. Our web application features a novel layout and navigational scheme for rapidly browsing phenome-wide associations by integrating results across genes and variants. Customizable controls, plots, and tables enable flexible filtering and visualization of phenotypes, genes, and variants of interest; results can be exported for downstream analyses; and variant associations across traits can be compared to inform pathways associated with complex traits and develop therapeutic hypotheses (see Supplementary Information).

Frequency and selection affect the landscape of rare variant associations

A major complicating factor in the analysis and interpretation of association statistics, particularly from rare variants, is the relationship between natural selection, allele frequency, effect size, and power for discovery. Sham et al. showed that the power to detect association is proportional to the variance explained of a biallelic variant (*16*). Specifically, for a continuous trait the variance explained of a biallelic variant that is purely additive is $2pqa^2$ where p is the allele frequency, q = 1-p and a is the allelic effect of the variant. Thus, for a fixed effect size, a more common variant will capture more variance and by extension show stronger association.

However, the process of negative selection will tend to decrease the frequency of functional damaging variants, suggesting that variants with large effect sizes are more likely to be rare. Indeed, partitioned heritability analyses for common variants support the presence of these countervailing forces, as comparatively lower frequency variants have larger absolute effect sizes but this growth in effect size is slower than the loss in variance explained from their lower frequency (*17*). In evaluating the landscape of rare variant association, we observe a similar pattern with increasing proportion of variants associated with at least one phenotype as frequency increases (Fig. 3A); however, within each frequency category, we observe the effect of functional annotation, a known correlate for deleteriousness, on the proportion of variants with at least one association.

Comparing the number of associations by variant annotation within each allele frequency category, we find that pLoF variants have a larger number of associations than missense variants, followed by synonymous variants for single-variant tests (Fig. 3A) as well as group tests (Fig. 3B). Within missense variants, variant deleteriousness as predicted by PolyPhen2 (*18*) is correlated with the number of associations meeting our p-value threshold (Fig. S18). For splice donor variants, we find a correlation between the proportion expressed across transcripts (pext) (*19*) and the number of associations (Fig. 3C). Additionally, the pathogenicity level of ClinVar variants is correlated with phenotypic association (Fig. 3D).

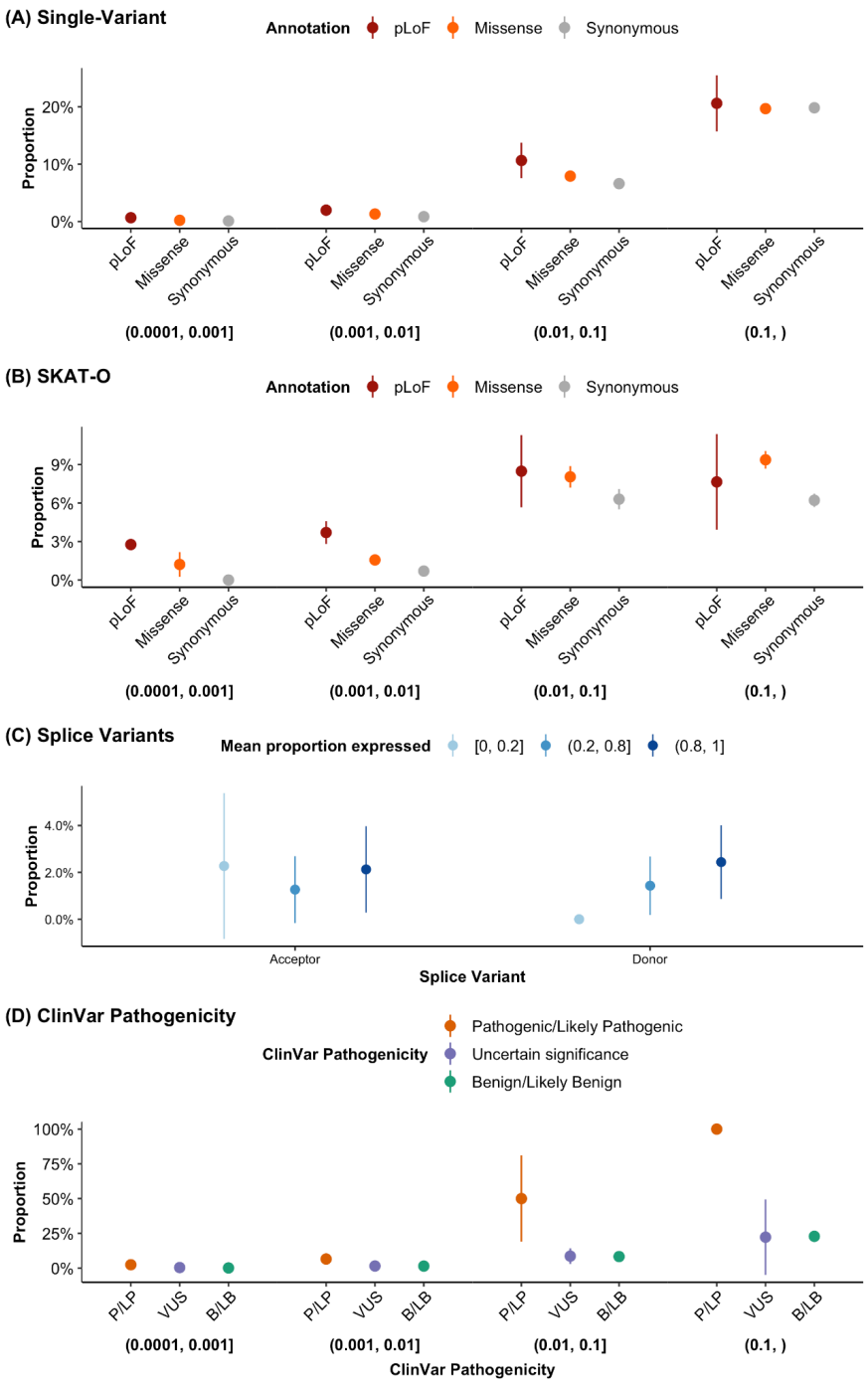


Figure 3 | The influence of variant allele frequency and functional annotation in exome association testing. The proportion of single variants (**A**) and genes (**B**) with at least one suggestive hit is shown broken down by allele frequency category (**A**) or cumulative allele frequency category (**B**), each shown below the plot, broken down by functional annotation. This metric is also plotted by the proportion expressed across transcripts for splice variants (**C**), and ClinVar pathogenicity status (**D**).

Gene function influences association statistics

We examined the phenotypic impact of gene categories previously known to have functional relevance and/or a role in disease. In particular, we find that 470 genes previously implicated in developmental delay (*20*) are more likely to be associated with a phenotype in the UK Biobank (Fisher's exact $p = 2 \times 10^{-3}$; Fig. 4). Further, we observe a correlation between selection against pLoFs in a gene and the phenotypic impact of pLoFs in that gene: specifically, constrained genes (i.e., those in the highest decile of LOEUF, a metric of loss-of-function intolerance) are more likely to be associated with a phenotype (5.9%) than a frequency-matched set of genes in the genome (2.0%; Fisher's exact $p = 2.1 \times 10^{-6}$; Fig. 4). Similarly, autosomal dominant and autosomal recessive genes, as well as genes with previously established hits in the GWAS catalog and FDA approved drug targets, show an increased phenotypic impact of pLoFs and missense variants. Finally, cell essential genes show an increased proportion of associated phenotypes through a pLoF mechanism, while non-essential genes do not show an enrichment.

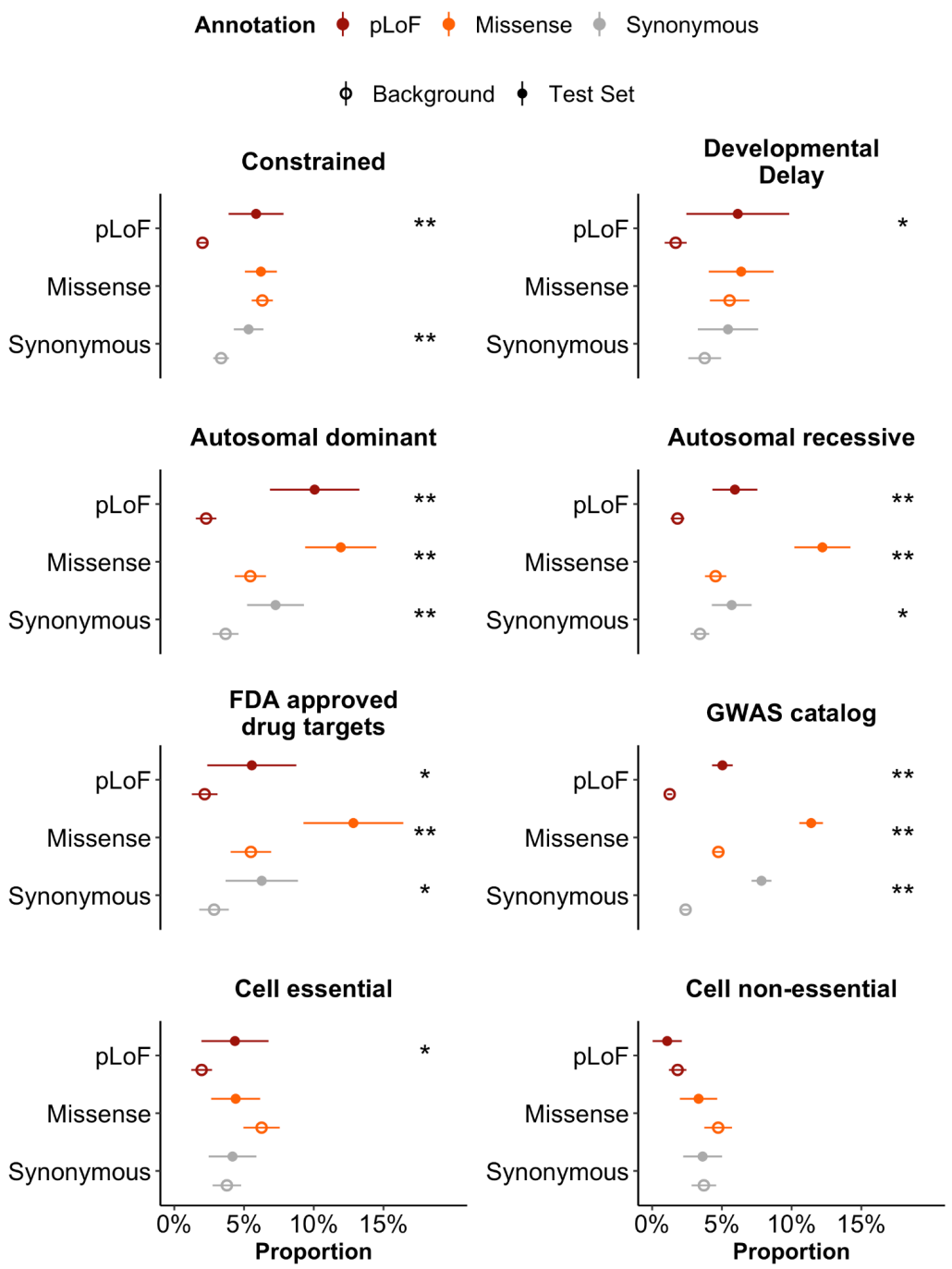


Figure 4 | The effect of gene function on the landscape of rare variant associations. The proportion of gene-annotation pairs with at least one association (SKAT-O $p < 2.5 \times 10^{-8}$) is shown for a number of gene categories, each compared to a background set of genes matched on cumulative allele frequency. Error bars represent 95% confidence intervals. Asterisks denote a significant difference between the background set and test set (* and ** indicate $p < 0.05$ and $p < 0.001$, respectively).

Biological insights from rare variant association results

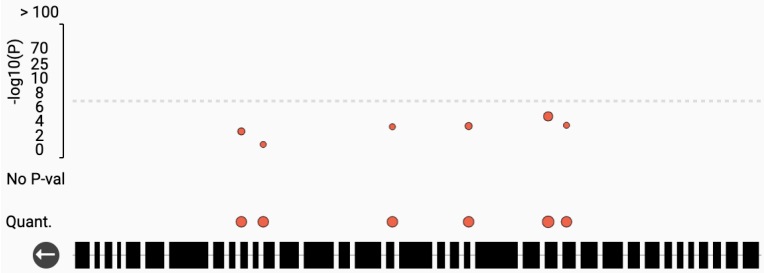
The biological information encapsulated in this dataset is extremely high-dimensional, and we release the full dataset of results for the benefit of the community. Here, we highlight a set of known and putative associations as examples of the power of this dataset. First, we recapitulate many known associations from previous studies, including associations between *PCSK9* and LDL cholesterol (pLoF burden $p = 3 \times 10^{-94}$), *COL1A1* and bone density (pLoF burden $p = 1.5 \times 10^{-8}$) (*21*), *KLF1* and several red blood cell traits (pLoF burden $< 2 \times 10^{-8}$) (*22*), and *LRP5* (Wnt coreceptor) and bone density and osteoporosis phenotypes (pLoF burden $< 5 \times 10^{-7}$) (*23*).

Finally, we highlight novel biological signals identified in the exome dataset, enabled by the results browser. In particular, we find an association between predicted loss-of-function of *SCRIB* and white matter integrity of tapetum (Fig. 5). Notably, this association is not significant at any single pLoF variant, but when aggregated into a SKAT-O or burden group test, the overall ablation of the transcript is associated at a p-value of 6×10^{-13} (Fig. 5A). This provides additional context to a signal observed in a recent GWAS of white matter integrity (*24*) averaged across regions of the brain, as well as in the body of corpus callosum (Fig. 5B). To our knowledge, this gene has not been associated in previous genome-wide association studies, although it is a constrained gene (pLI = 0.93) that shows evidence for neural tube defects in mice (*25*) with ultra-rare occurrences in humans (*26*).

SCRIB gene burden associations with **Mean OD in tapetum on FA skeleton (left)**

Category	P-value SKATO	P-value burden	P-value SKAT	Total variants	Filter
pLoF	5.92e-13	6.94e-13	1.31e-9	6	☑
Missense, LC	0.203	0.939	0.114	233	☐
Synonymous	0.24	0.478	0.135	157	☐

SCRIB single variant associations with **Mean OD in tapetum on FA skeleton (left)**



Variant ID	CSQ	HGVSp	P-Value	Beta	AC Trait
8-143810514-G-A	stop gained	p.Gln499Ter	4.98e-5	2.78e+0	2
8-143810727-T-TA	frameshift	p.Ile455TyrfsT...	9.02e-4	3.22e+0	1
8-143807599-G-A	stop gained	p.Gln711Ter	1.14e-3	3.16e+0	1
8-143804997-GCG...	frameshift	p.Val894Hisfs...	1.44e-3	3.09e+0	1
8-143795049-TGC...	frameshift	p.Val1274Cysf...	6.44e-3	2.64e+0	1
8-143795418-A-G	splice donor		4.38e-1	7.52e-1	1

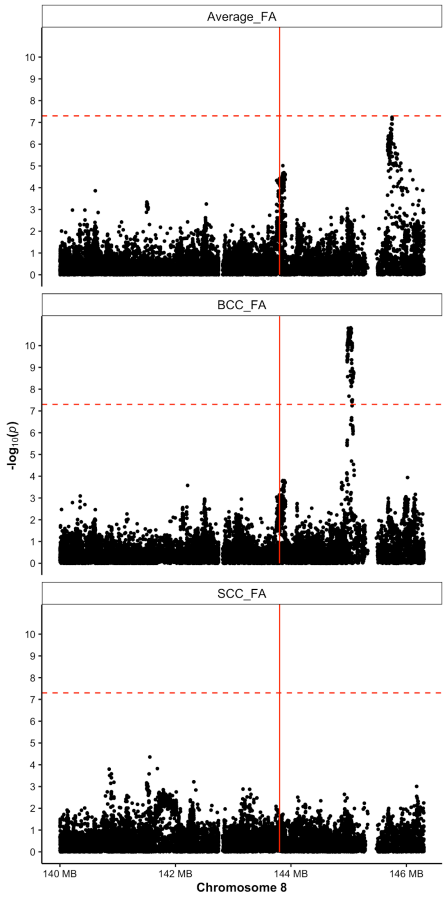


Figure 5 | Refined association between SCRIB and white matter integrity of tapetum. (**A**): The association between pLoF variants in SCRIB with mean OD (orientation dispersion index) in tapetum on FA (fractional anisotropy) skeleton (from dMRI data). 6 rare pLoF variants are discovered, all of which have a positive beta value (bottom). In aggregate, these variants show an association at $p = 6 \times 10^{-13}$. (**B**): A Manhattan plot of a previous GWAS of FA averaged across brain regions (top), body of corpus callosum (middle), and splenium of corpus callosum (bottom). Horizontal dashed line indicates a GWAS genome-wide significance threshold (5×10^{-8}), and vertical line indicates the location of SCRIB.

Discussion

We have generated rare variant association analysis summary statistics for 3,700 phenotypes and made these data available to the public, via bulk data downloads as well as a public-facing browser (https://genebass.org).

There are a number of limitations to our analysis. Although we have performed extensive QC to improve the reliability of these results, we urge caution in interpreting association results, particularly for the rarest binary traits (prevalence < 10^{-4}) and for ultra-rare variants (frequency <

10^{-4}). For pLoF variants, the median cumulative allele frequency across genes is approximately 1.5×10^{-4}, suggesting that group tests at current sample sizes are only powered to detect individual gene effects for quantitative traits that capture at least 0.02% of variance, as well as diseases and traits that have a high prevalence (well above 10%; Fig. S10). This is further observed in the lack of asymptotic properties of the mixed-model tests for rarer binary traits (Fig. S9). Nonetheless, global biological trends are apparent, such as the relative ordering of functional impact (pLoF > missense > synonymous; Fig. 3), highlighting that the ability to accurately annotate variants with the functional consequences on a gene is critical to powering discovery in rare variant analysis. Further, measures of natural selection at the gene level continue to highlight that certain classes of genes, such as LoF-intolerant genes, are clearly enriched for phenotypic associations.

Finally, these association analyses were only performed for individuals of European ancestry, the largest group in the dataset. Notably, these analyses only interrogate a slice of human genetic diversity, and expanding to additional ancestries has been shown to increase power and resolution for genetic discovery (*27–29*); however, as the sample sizes of non-European individuals in the UK Biobank dataset are very limited, these analyses would be underpowered for most binary traits including many disease outcomes. Additional datasets with large sample sizes of diverse individuals will be required to overcome these limitations.

Biobank teams

Abbvie:

- Steering team: Jeff Waring, Howard Jacob, J. Wade Davis
- Data management team: A. Jason Grundstad, Silvia Orozco
- Extended Scientific team: Bridget Riley-Gillis, Sahar Esmaeeli, Fedik Rahimov, Ali Abbasi, Nizar Smaoui, Xiuwen Zheng, Emily King, John Lee, Reza Hammond, Mark Reppell, Hyun Ji Noh

Biogen:

- Steering team: Ellen Tsai, Christopher D. Whelan, Paola Bronson, David Sexton, Sally John, Heiko Runz
- Data management team: Eric Marshall, Mehool Patel, Saranya Duraisamy, Timothy Swan
- Extended Scientific team: Denis Baird, Chia-Yen Chen, Susan Eaton, Jake Gagnon, Feng Gao, Cynthia Gubbels, Yunfeng Huang, Varant Kupelian, Kejie Li, Dawei Liu, Stephanie Loomis, Helen McLaughlin, Adele Mitchell, Nilanjana Sadhu, Benjamin Sun, Ruoyu Tian

Pfizer:

- Steering team: Hye In Kim, Xinli Hu, Morten Sogaard, A. Katrina Loomis, Eric Fauman, Melissa R. Miller
- Data Management team: Jay Bergeron, Andrew Hill, Juha Sarimaa, Zhan Ye, Xing Chen
- Extended Scientific team: Yi-Pin Lai, Jean-Philippe Fortin, Joanne Berghout, Robert Moccia, Craig L. Hyde

Acknowledgements

We thank Danielle Ciofani and Cathy Marshall for their efforts in launching this project. We thank the participants and leadership of the UK Biobank: this work was done under UK Biobank applications 26041 and 48511.

References

1. S. S. Kalia, K. Adelman, S. J. Bale, W. K. Chung, C. Eng, J. P. Evans, G. E. Herman, S. B. Hufnagel, T. E. Klein, B. R. Korf, K. D. McKelvey, K. E. Ormond, C. S. Richards, C. N. Vlangos, M. Watson, C. L. Martin, D. T. Miller, Recommendations for reporting of secondary findings in clinical exome and genome sequencing, 2016 update (ACMG SF v2.0): a policy statement of the American College of Medical Genetics and Genomics. *Genet. Med.* **19**, 249–255 (2017).

2. M. J. Bamshad, D. A. Nickerson, J. X. Chong, Mendelian Gene Discovery: Fast and Furious with No End in Sight. *Am. J. Hum. Genet.* **105**, 448–455 (2019).

3. M. Abifadel, M. Varret, J.-P. Rabès, D. Allard, K. Ouguerram, M. Devillers, C. Cruaud, S. Benjannet, L. Wickham, D. Erlich, A. Derré, L. Villéger, M. Farnier, I. Beucler, E. Bruckert, J. Chambaz, B. Chanu, J.-M. Lecerf, G. Luc, P. Moulin, J. Weissenbach, A. Prat, M. Krempf, C. Junien, N. G. Seidah, C. Boileau, Mutations in PCSK9 cause autosomal dominant hypercholesterolemia. *Nat. Genet.* **34**, 154–156 (2003).

4. J. C. Cohen, E. Boerwinkle, T. H. Mosley Jr., H. H. Hobbs, Sequence Variations in PCSK9,Low LDL, and Protection against Coronary Heart Disease. *N. Engl. J. Med.* **354**, 1264–1272 (2006).

5. M. S. Sabatine, R. P. Giugliano, A. C. Keech, N. Honarpour, S. D. Wiviott, S. A. Murphy, J. F. Kuder, H. Wang, T. Liu, S. M. Wasserman, P. S. Sever, T. R. Pedersen, FOURIER Steering Committee and Investigators, Evolocumab and Clinical Outcomes in Patients with Cardiovascular Disease. *N. Engl. J. Med.* **376**, 1713–1722 (2017).

6. Y. Tanigawa, M. Wainberg, J. Karjalainen, T. Kiiskinen, G. Venkataraman, S. Lemmelä, J. A. Turunen, R. R. Graham, A. S. Havulinna, M. Perola, A. Palotie, FinnGen, M. J. Daly, M. A. Rivas, Rare protein-altering variants in ANGPTL7 lower intraocular pressure and protect against glaucoma. *PLoS Genet.* **16**, e1008682 (2020).

7. Q. Wang, R. S. Dhindsa, K. Carss, A. Harper, A. Nag, I. Tachmazidou, D. Vitsios, S. V. V. Deevi, A. Mackay, D. Muthas, M. Hühn, S. Monkley, H. Olsson, S. Wasilewski, K. R. Smith, R. March, A. Platt, C. Haefliger, S. Petrovski, AstraZeneca Genomics Initiative, Surveying the contribution of rare variants to the genetic architecture of human disease through exome sequencing of 177,882 UK Biobank participants. *biorXiv* (2020), p. 2020.12.13.422582.

8. S. J. Jurgens, S. H. Choi, V. N. Morrill, M. Chaffin, J. P. Pirruccello, J. L. Halford, L.-C. Weng, V. Nauffal, C. Roselli, A. W. Hall, K. G. Aragam, K. L. Lunetta, S. A. Lubitz, P. T. Ellinor, Rare Genetic Variation Underlying Human Diseases and Traits: Results from 200,000 Individuals in the UK Biobank. *Cold Spring Harbor Laboratory* (2020), p. 2020.11.29.402495.

9. A. M. Deaton, M. M. Parker, L. D. Ward, A. O. Flynn-Carroll, L. BonDurant, G. Hinkle, P. Nioi, Gene-level analysis of rare variants in 363,977 whole exome sequences reveals an association of GIGYF1 loss of function with diabetes. *medRxiv* (2021) (available at https://www.medrxiv.org/content/10.1101/2021.01.19.21250105v1.abstract).

10. Hail Team, *Hail 0.2.54. https://github.com/hail-is/hail/releases/tag/0.2.54* (2020;

https://github.com/hail-is/hail/releases/tag/0.2.54).

11. K. J. Karczewski, L. C. Francioli, G. Tiao, B. B. Cummings, J. Alföldi, Q. Wang, R. L. Collins, K. M. Laricchia, A. Ganna, D. P. Birnbaum, L. D. Gauthier, H. Brand, M. Solomonson, N. A. Watts, D. Rhodes, M. Singer-Berk, E. M. England, E. G. Seaby, J. A. Kosmicki, R. K. Walters, K. Tashman, Y. Farjoun, E. Banks, T. Poterba, A. Wang, C. Seed, N. Whiffin, J. X. Chong, K. E. Samocha, E. Pierce-Hoffman, Z. Zappala, A. H. O'Donnell-Luria, E. V. Minikel, B. Weisburd, M. Lek, J. S. Ware, C. Vittal, I. M. Armean, L. Bergelson, K. Cibulskis, K. M. Connolly, M. Covarrubias, S. Donnelly, S. Ferriera, S. Gabriel, J. Gentry, N. Gupta, T. Jeandet, D. Kaplan, C. Llanwarne, R. Munshi, S. Novod, N. Petrillo, D. Roazen, V. Ruano-Rubio, A. Saltzman, M. Schleicher, J. Soto, K. Tibbetts, C. Tolonen, G. Wade, M. E. Talkowski, Genome Aggregation Database Consortium, B. M. Neale, M. J. Daly, D. G. MacArthur, The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*. **581**, 434–443 (2020).

12. W. Zhou, Z. Zhao, J. B. Nielsen, L. G. Fritsche, J. LeFaive, S. A. Gagliano Taliun, W. Bi, M. E. Gabrielsen, M. J. Daly, B. M. Neale, K. Hveem, G. R. Abecasis, C. J. Willer, S. Lee, Scalable generalized linear mixed model for region-based association tests in large biobanks and cohorts. *Nat. Genet.* **52**, 634–639 (2020).

13. E. Marouli, M. Graff, C. Medina-Gomez, K. S. Lo, A. R. Wood, T. R. Kjaer, R. S. Fine, Y. Lu, C. Schurmann, H. M. Highland, S. Rüeger, G. Thorleifsson, A. E. Justice, D. Lamparter, K. E. Stirrups, V. Turcot, K. L. Young, T. W. Winkler, T. Esko, T. Karaderi, A. E. Locke, N. G. D. Masca, M. C. Y. Ng, P. Mudgal, M. A. Rivas, S. Vedantam, A. Mahajan, X. Guo, G. Abecasis, K. K. Aben, L. S. Adair, D. S. Alam, E. Albrecht, K. H. Allin, M. Allison, P. Amouyel, E. V. Appel, D. Arveiler, F. W. Asselbergs, P. L. Auer, B. Balkau, B. Banas, L. E. Bang, M. Benn, S. Bergmann, L. F. Bielak, M. Blüher, H. Boeing, E. Boerwinkle, C. A. Böger, L. L. Bonnycastle, J. Bork-Jensen, M. L. Bots, E. P. Bottinger, D. W. Bowden, I. Brandslund, G. Breen, M. H. Brilliant, L. Broer, A. A. Burt, A. S. Butterworth, D. J. Carey, M. J. Caulfield, J. C. Chambers, D. I. Chasman, Y.-D. I. Chen, R. Chowdhury, C. Christensen, A. Y. Chu, M. Cocca, F. S. Collins, J. P. Cook, J. Corley, J. C. Galbany, A. J. Cox, G. Cuellar-Partida, J. Danesh, G. Davies, P. I. W. de Bakker, G. J. de Borst, S. de Denus, M. C. H. de Groot, R. de Mutsert, I. J. Deary, G. Dedoussis, E. W. Demerath, A. I. den Hollander, J. G. Dennis, E. Di Angelantonio, F. Drenos, M. Du, A. M. Dunning, D. F. Easton, T. Ebeling, T. L. Edwards, P. T. Ellinor, P. Elliott, E. Evangelou, A.-E. Farmaki, J. D. Faul, M. F. Feitosa, S. Feng, E. Ferrannini, M. M. Ferrario, J. Ferrieres, J. C. Florez, I. Ford, M. Fornage, P. W. Franks, R. Frikke-Schmidt, T. E. Galesloot, W. Gan, I. Gandin, P. Gasparini, V. Giedraitis, A. Giri, G. Girotto, S. D. Gordon, P. Gordon-Larsen, M. Gorski, N. Grarup, M. L. Grove, V. Gudnason, S. Gustafsson, T. Hansen, K. M. Harris, T. B. Harris, A. T. Hattersley, C. Hayward, L. He, I. M. Heid, K. Heikkilä, Ø. Helgeland, J. Hernesniemi, A. W. Hewitt, L. J. Hocking, M. Hollensted, O. L. Holmen, G. K. Hovingh, J. M. M. Howson, C. B. Hoyng, P. L. Huang, K. Hveem, M. A. Ikram, E. Ingelsson, A. U. Jackson, J.-H. Jansson, G. P. Jarvik, G. B. Jensen, M. A. Jhun, Y. Jia, X. Jiang, S. Johansson, M. E. Jørgensen, T. Jørgensen, P. Jousilahti, J. W. Jukema, B. Kahali, R. S. Kahn, M. Kähönen, P. R. Kamstrup, S. Kanoni, J. Kaprio, M. Karaleftheri, S. L. R. Kardia, F. Karpe, F. Kee, R. Keeman, L. A. Kiemeney, H. Kitajima, K. B. Kluivers, T. Kocher, P. Komulainen, J. Kontto, J. S. Kooner, C. Kooperberg, P. Kovacs, J. Kriebel, H. Kuivaniemi, S. Küry, J. Kuusisto, M. La Bianca, M. Laakso, T. A. Lakka, E. M. Lange, L. A. Lange, C. D. Langefeld, C. Langenberg, E. B. Larson, I.-T. Lee, T. Lehtimäki, C. E. Lewis, H. Li, J. Li, R. Li-Gao, H. Lin, L.-A. Lin, X. Lin, L. Lind, J. Lindström, A. Linneberg, Y. Liu, Y. Liu, A. Lophatananon, J. 'an Luan, S. A. Lubitz, L.-P. Lyytikäinen, D. A. Mackey, P. A. F. Madden, A. K. Manning, S.

Männistö, G. Marenne, J. Marten, N. G. Martin, A. L. Mazul, K. Meidtner, A. Metspalu, P. Mitchell, K. L. Mohlke, D. O. Mook-Kanamori, A. Morgan, A. D. Morris, A. P. Morris, M. Müller-Nurasyid, P. B. Munroe, M. A. Nalls, M. Nauck, C. P. Nelson, M. Neville, S. F. Nielsen, K. Nikus, P. R. Njølstad, B. G. Nordestgaard, I. Ntalla, J. R. O'Connel, H. Oksa, L. M. O. Loohuis, R. A. Ophoff, K. R. Owen, C. J. Packard, S. Padmanabhan, C. N. A. Palmer, G. Pasterkamp, A. P. Patel, A. Pattie, O. Pedersen, P. L. Peissig, G. M. Peloso, C. E. Pennell, M. Perola, J. A. Perry, J. R. B. Perry, T. N. Person, A. Pirie, O. Polasek, D. Posthuma, O. T. Raitakari, A. Rasheed, R. Rauramaa, D. F. Reilly, A. P. Reiner, F. Renström, P. M. Ridker, J. D. Rioux, N. Robertson, A. Robino, O. Rolandsson, I. Rudan, K. S. Ruth, D. Saleheen, V. Salomaa, N. J. Samani, K. Sandow, Y. Sapkota, N. Sattar, M. K. Schmidt, P. J. Schreiner, M. B. Schulze, R. A. Scott, M. P. Segura-Lepe, S. Shah, X. Sim, S. Sivapalaratnam, K. S. Small, A. V. Smith, J. A. Smith, L. Southam, T. D. Spector, E. K. Speliotes, J. M. Starr, V. Steinthorsdottir, H. M. Stringham, M. Stumvoll, P. Surendran, L. M. 't Hart, K. E. Tansey, J.-C. Tardif, K. D. Taylor, A. Teumer, D. J. Thompson, U. Thorsteinsdottir, B. H. Thuesen, A. Tönjes, G. Tromp, S. Trompet, E. Tsafantakis, J. Tuomilehto, A. Tybjaerg-Hansen, J. P. Tyrer, R. Uher, A. G. Uitterlinden, S. Ulivi, S. W. van der Laan, A. R. Van Der Leij, C. M. van Duijn, N. M. van Schoor, J. van Setten, A. Varbo, T. V. Varga, R. Varma, D. R. V. Edwards, S. H. Vermeulen, H. Vestergaard, V. Vitart, T. F. Vogt, D. Vozzi, M. Walker, F. Wang, C. A. Wang, S. Wang, Y. Wang, N. J. Wareham, H. R. Warren, J. Wessel, S. M. Willems, J. G. Wilson, D. R. Witte, M. O. Woods, Y. Wu, H. Yaghootkar, J. Yao, P. Yao, L. M. Yerges-Armstrong, R. Young, E. Zeggini, X. Zhan, W. Zhang, J. H. Zhao, W. Zhao, W. Zhao, H. Zheng, W. Zhou, EPIC-InterAct Consortium, CHD Exome+ Consortium, ExomeBP Consortium, T2D-Genes Consortium, GoT2D Genes Consortium, Global Lipids Genetics Consortium, ReproGen Consortium, MAGIC Investigators, J. I. Rotter, M. Boehnke, S. Kathiresan, M. I. McCarthy, C. J. Willer, K. Stefansson, I. B. Borecki, D. J. Liu, K. E. North, N. L. Heard-Costa, T. H. Pers, C. M. Lindgren, C. Oxvig, Z. Kutalik, F. Rivadeneira, R. J. F. Loos, T. M. Frayling, J. N. Hirschhorn, P. Deloukas, G. Lettre, Rare and low-frequency coding variants alter human adult height. *Nature*. **542**, 186–190 (2017).

14. Y. Hu, A. M. Stilp, C. P. McHugh, S. Rao, D. Jain, X. Zheng, J. Lane, S. Méric de Bellefon, L. M. Raffield, M.-H. Chen, L. R. Yanek, M. Wheeler, Y. Yao, C. Ren, J. Broome, J.-Y. Moon, P. S. de Vries, B. D. Hobbs, Q. Sun, P. Surendran, J. A. Brody, T. W. Blackwell, H. Choquet, K. Ryan, R. Duggirala, N. Heard-Costa, Z. Wang, N. Chami, M. H. Preuss, N. Min, L. Ekunwe, L. A. Lange, M. Cushman, N. Faraday, J. E. Curran, L. Almasy, K. Kundu, A. V. Smith, S. Gabriel, J. I. Rotter, M. Fornage, D. M. Lloyd-Jones, R. S. Vasan, N. L. Smith, K. E. North, E. Boerwinkle, L. C. Becker, J. P. Lewis, G. R. Abecasis, L. Hou, J. R. O'Connell, A. C. Morrison, T. H. Beaty, R. Kaplan, A. Correa, J. Blangero, E. Jorgenson, B. M. Psaty, C. Kooperberg, R. T. Walton, B. P. Kleinstiver, H. Tang, R. J. F. Loos, N. Soranzo, A. S. Butterworth, D. Nickerson, S. S. Rich, B. D. Mitchell, A. D. Johnson, P. L. Auer, Y. Li, R. A. Mathias, G. Lettre, N. Pankratz, C. C. Laurie, C. A. Laurie, D. E. Bauer, M. P. Conomos, A. P. Reiner, NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium, Whole-genome sequencing association analysis of quantitative red blood cell phenotypes: The NHLBI TOPMed program. *Am. J. Hum. Genet.* **108**, 874–893 (2021).

15. K. J. Karczewski, B. Weisburd, B. Thomas, M. Solomonson, D. M. Ruderfer, D. Kavanagh, T. Hamamsy, M. Lek, K. E. Samocha, B. B. Cummings, D. Birnbaum, The Exome Aggregation Consortium, M. J. Daly, D. G. MacArthur, The ExAC browser: displaying reference data information from over 60 000 exomes. *Nucleic Acids Res.* **45**, D840–D845 (2017).

16. P. C. Sham, S. S. Cherny, S. Purcell, J. K. Hewitt, Power of linkage versus association analysis of quantitative traits, by use of variance-components models, for sibship data. *Am. J. Hum. Genet.* **66**, 1616–1630 (2000).

17. S. Gazal, H. K. Finucane, N. A. Furlotte, P.-R. Loh, P. F. Palamara, X. Liu, A. Schoech, B. Bulik-Sullivan, B. M. Neale, A. Gusev, A. L. Price, Linkage disequilibrium-dependent architecture of human complex traits shows action of negative selection. *Nat. Genet.* **49**, 1421–1427 (2017).

18. I. A. Adzhubei, S. Schmidt, L. Peshkin, V. E. Ramensky, A. Gerasimova, P. Bork, A. S. Kondrashov, S. R. Sunyaev, A method and server for predicting damaging missense mutations. *Nat. Methods*. **7**, 248–249 (2010).

19. B. B. Cummings, K. J. Karczewski, J. A. Kosmicki, E. G. Seaby, N. A. Watts, M. Singer-Berk, J. M. Mudge, J. Karjalainen, F. K. Satterstrom, A. H. O'Donnell-Luria, T. Poterba, C. Seed, M. Solomonson, J. Alföldi, Genome Aggregation Database Production Team, Genome Aggregation Database Consortium, M. J. Daly, D. G. MacArthur, Transcript expression-aware annotation improves rare variant interpretation. *Nature*. **581**, 452–458 (2020).

20. J. Kaplanis, K. E. Samocha, L. Wiel, Z. Zhang, K. J. Arvai, R. Y. Eberhardt, G. Gallone, S. H. Lelieveld, H. C. Martin, J. F. McRae, P. J. Short, R. I. Torene, E. de Boer, P. Danecek, E. J. Gardner, N. Huang, J. Lord, I. Martincorena, R. Pfundt, M. R. F. Reijnders, A. Yeung, H. G. Yntema, Deciphering Developmental Disorders Study, L. E. L. M. Vissers, J. Juusola, C. F. Wright, H. G. Brunner, H. V. Firth, D. R. FitzPatrick, J. C. Barrett, M. E. Hurles, C. Gilissen, K. Retterer, Evidence for 28 genetic disorders discovered by combining healthcare and research data. *Nature*. **586**, 757–762 (2020).

21. V. Mann, E. E. Hobson, B. Li, T. L. Stewart, S. F. Grant, S. P. Robins, R. M. Aspden, S. H. Ralston, A COL1A1 Sp1 binding site polymorphism predisposes to osteoporotic fracture by affecting bone density and quality. *J. Clin. Invest.* **107**, 899–907 (2001).

22. A. Perkins, X. Xu, D. R. Higgs, G. P. Patrinos, L. Arnaud, J. J. Bieker, S. Philipsen, KLF1 Consensus Workgroup, Krüppeling erythropoiesis: an unexpected broad spectrum of human red blood cell disorders due to KLF1 variants. *Blood*. **127**, 1856–1862 (2016).

23. R. Baron, G. Rawadi, Targeting the Wnt/beta-catenin pathway to regulate bone formation in the adult skeleton. *Endocrinology*. **148**, 2635–2643 (2007).

24. B. Zhao, T. Li, Y. Yang, X. Wang, T. Luo, Y. Shan, Z. Zhu, D. Xiong, M. E. Hauberg, J. Bendl, J. F. Fullard, P. Roussos, Y. Li, J. L. Stein, H. Zhu, Common genetic variation influencing human white matter microstructure. *bioRxiv* (2020), p. 2020.05.23.112409.

25. J. N. Murdoch, D. J. Henderson, K. Doudney, C. Gaston-Massuet, H. M. Phillips, C. Paternotte, R. Arkell, P. Stanier, A. J. Copp, Disruption of scribble (Scrb1) causes severe neural tube defects in the circletail mouse. *Hum. Mol. Genet.* **12**, 87–98 (2003).

26. Y. Lei, H. Zhu, C. Duhon, W. Yang, M. E. Ross, G. M. Shaw, R. H. Finnell, Mutations in planar cell polarity gene SCRIB are associated with spina bifida. *PLoS One*. **8**, e69262 (2013).

27. S. Sakaue, M. Kanai, Y. Tanigawa, J. Karjalainen, M. Kurki, S. Koshiba, A. Narita, T.

Konuma, K. Yamamoto, M. Akiyama, K. Ishigaki, A. Suzuki, K. Suzuki, W. Obara, K. Yamaji, K. Takahashi, S. Asai, Y. Takahashi, T. Suzuki, N. Shinozaki, H. Yamaguchi, S. Minami, S. Murayama, K. Yoshimori, S. Nagayama, D. Obata, M. Higashiyama, A. Masumoto, Y. Koretsune, K. I. FinnGen, C. Terao, T. Yamauchi, I. Komuro, T. Kadowaki, G. Tamiya, M. Yamamoto, Y. Nakamura, M. Kubo, Y. Murakami, K. Yamamoto, Y. Kamatani, A. Palotie, M. A. Rivas, M. J. Daly, K. Matsuda, Y. Okada, A global atlas of genetic associations of 220 deep phenotypes. *bioRxiv* (2020), , doi:10.1101/2020.10.23.20213652.

28. L. Majara, A. Kalungi, N. Koen, H. Zar, D. J. Stein, E. Kinyanda, E. G. Atkinson, A. R. Martin, Low generalizability of polygenic scores in African populations due to genetic and environmental diversity. *bioRxiv* (2021), p. 2021.01.12.426453.

29. J. Morales, D. Welter, E. H. Bowler, M. Cerezo, L. W. Harris, A. C. McMahon, P. Hall, H. A. Junkins, A. Milano, E. Hastings, C. Malangone, A. Buniello, T. Burdett, P. Flicek, H. Parkinson, F. Cunningham, L. A. Hindorff, J. A. L. MacArthur, A standardized framework for representation of ancestry data in genomics studies, with application to the NHGRI-EBI GWAS Catalog. *Genome Biol.* **19**, 21 (2018).