DNA LANGUAGE MODELS IDENTIFY VARIANTS PREDICTIVE ACROSS THE HUMAN PHENOME

Benjamin Wild[1*] **Julius Upmeier zu Belzen**[1*] **Luis Hermann**[1] **Paul Kittner**[1]
Sedra Abou Ghaloun[1] **Roland Eils**[1,2]

[1]Digital Health Center, Berlin Institute of Health
[2]Health Data Science Unit, Heidelberg University

{benjamin.wild, julius.upmeier, luis.herrmann, paul.kittner,
sedra.abou-ghaloun, roland.eils}@charite.de

ABSTRACT

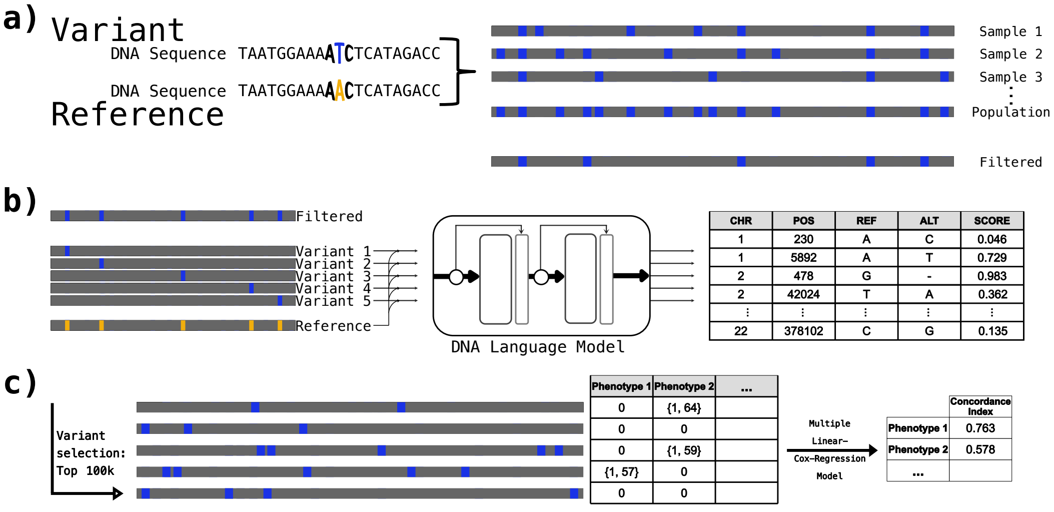
Early identification of individuals at high risk for diseases is crucial to public health, facilitating timely prevention and treatment strategies. Polygenic scores (PGS) offer significant clinical promise by estimating the genetic predisposition to diseases, yet their current impact is limited by insufficient power, especially for rare variants and diseases. While larger cohorts may enhance the power of PGS, advancements in methodology are equally critical. Recently, DNA language models, serving as foundational models for genomic data, have shown impressive capabilities in tasks such as predicting epigenetic marks, identifying regulatory sequences, and annotating variant effects. Yet, their utility beyond local variant effects has not been explored to date. Here, we use the GPN-MSA and Nucleotide Transformer (NT) DNA language models to predict the relationship between genetic variants and disease risk. We use variant-level embeddings to predict the potential of variants to influence a wide range of phenotypes and show that variant sets with high scores are more predictive of diseases across the human phenome than baseline variant sets. While prior work on DNA language models has primarily focused on local variant effects, our work demonstrates their value in genome-wide variant selection, potentially complementing genome-wide association studies (GWAS) and PGS by learning representations that can be used to identify rare variants with large effect sizes. Our results highlight the potential of DNA language models in identifying genotype-phenotype associations.

1 INTRODUCTION

The growing need for effective risk stratification in healthcare, driven by shifting demographics and the rising prevalence of non-communicable diseases, underscores the urgency of identifying individuals at high risk of disease early in the disease trajectory. This is critical not only in industrialized countries grappling with an aging population and increased disease burden, but also in low- and middle-income countries with limited healthcare resources (Vogeli et al., 2007; Girwar et al., 2021).

Genotyping and sequencing technologies have become increasingly affordable, and offer promising avenues for risk stratification by leveraging genetic associations to predict disease risk. GWAS have significantly contributed to our understanding of the genetic underpinnings of complex diseases by identifying variants associated with various traits. However, due to their statistical setup, GWAS alone often fall short in fully capturing the heritability of complex diseases due to their inability to accurately model rare variants (Manolio et al., 2009).

PGS build upon GWAS by aggregating the effects of numerous variants into a singular score that predicts an individual's predisposition for a disease, offering a pathway toward personalized medicine. Nonetheless, the effectiveness of PGS faces challenges such as the need for large sample sizes and the generalizability of scores across different populations (Visscher et al., 2017; Boyle et al., 2017; Khera et al., 2018a; Torkamani et al., 2018; Martin et al., 2019; Hingorani et al., 2023).



CHR	POS	REF	ALT	SCORE
1	230	A	C	0.046
1	5892	A	T	0.729
2	478	G	-	0.983
2	42024	T	A	0.362
⋮	⋮	⋮	⋮	⋮
22	378102	C	G	0.135

Phenotype 1	Phenotype 2	...
0	{1, 64}	
0	0	
0	{1, 59}	
{1, 57}	0	
0	0	

	Concordance Index
Phenotype 1	0.763
Phenotype 2	0.578
...	

Figure 1: Overview of the method: **a)** We use a dataset of 16 111 439 genotyped variants from a large-scale population cohort (UK Biobank) comprised of 487 150 individuals, and filtered for variants also available in the FinnGen dataset for future-validations. **b)** We then evaluate different strategies to extract per-variant scores predictive of disease susceptibility using the GPN-MSA and Nucleotide Transformer DNA language models. **c)** Using these scores, we create variant sets and evaluate their predictive performance in the UK Biobank for a set of polygenic diseases and across the human phenome using linear Cox Regression models.

In PGS, common small-effect variants can be more predictive across a population compared to rare variants with substantial effects. This is further exacerbated by evolutionary processes selecting against high-impact variants, further diluting them in the population. To accurately model genotype-phenotype relations, both types of variants have to be taken into account and likely be addressed with different modeling approaches for optimal results. For the standing height trait for example, a GWAS with N = 5.4 million participants of 281 individual studies was necessary to explain more than 90 % of the estimated variant heritability (Yengo et al., 2022).

Protein language models, such as AlphaMissense, utilize machine learning to predict the effects of missense variants, which alter the amino acid sequence of proteins, potentially affecting their function (Cheng et al., 2023). By leveraging unsupervised learning, structural context from models like AlphaFold, and fine-tuning on population frequency data, AlphaMissense offers state-of-the-art predictions for the pathogenicity of these variants, classifying a significant portion of missense variants as either likely pathogenic or benign. However, a critical limitation of AlphaMissense and similar methods is their focus on the exome, primarily predicting the pathogenicity of variants that result in amino acid substitutions. However, most variants identified in GWAS are located in non-coding regions of the genome, such as intronic and intergenic regions (Chami & Lettre, 2014; Cirillo et al., 2018). These non-coding variants play crucial roles in gene regulation and expression but fall outside the predictive capacity of models focused on protein-coding changes, underscoring a significant gap in our ability to interpret the vast majority of genetic variations implicated in disease through GWAS.

Recent advancements in DNA language models have shown significant promise in enhancing our understanding of genomic data. DNABERT (Ji et al., 2021) leverages a pre-trained bidirectional encoder to capture the nuanced language of non-coding DNA, achieving state-of-the-art performance in identifying gene regulatory elements with improved accuracy and efficiency. The Nucleotide Transformer (Dalla-Torre et al., 2023), through extensive pre-training on diverse genomes, generates transferable, context-specific nucleotide sequence representations, facilitating accurate molecular phenotype prediction even in low-data scenarios. GPN-MSA (Benegas et al., 2023) introduces a novel approach by utilizing whole-genome sequence alignments across multiple species for rapid training, showing remarkable effectiveness in predicting variant effects. HyenaDNA (Nguyen et al., 2023), leveraging implicit convolutions for long-range genomic interactions, significantly extends context length capabilities up to 1 million tokens at the single nucleotide level, setting new benchmarks in genomic modeling by enabling in-context learning and surpassing previous state-of-the-art models in efficiency. These models collectively represent a novel paradigm for rare variant and disease analysis, and highlight the untapped potential of DNA language models in improving clinical outcomes through better models of genotype-phenotype relationships.

In this work, we use DNA language models to predict polygenic variant effects and show that variant sets with high likely effect size outperform baseline variant sets in predicting disease onset across the human phenome in a large real-world population cohort.

2 Methods

2.1 Variant sets and preprocessing

We used the imputed genotyping array dataset from the UK Biobank (UKB)(Sudlow et al., 2015; Bycroft et al., 2018), which includes 93 095 623 variants for 487 150 individuals. To facilitate future validation with the FinnGen cohorts, we selected only those variants present in both the UKB and FinnGen datasets (Kurki et al., 2023). We did not exclude variants based on minor allele frequency or information score.

Variants from FinnGen were converted from the human genome build GRCh38 to GRCh37 (hg19) using CrossMap (Zhao et al., 2014) and then combined with UKB genotypes to create a dataset of 16 111 439 variants for this study. We processed the genetic data using PLINK2 (Purcell et al., 2007), VCFtools (Danecek et al., 2011), and custom workflows for variant set extraction.

For the multivariate regression models, we represented the genotypes as allele dosages ranging from 0 (0 alternative alleles) to 2 (expected 2 alternative alleles), continuously, representing the uncertainty in the imputation process as an expected value (e.g. 0.1 would correspond to most likely two reference alleles) (Collister et al., 2022).

2.2 DNA language model inference

We use the pretrained DNA language model GPN-MSA (Benegas et al., 2023) to compute zero-shot pathogenicity scores ($score = \log p(\text{ALT})/p(\text{REF})$) using the variant effect prediction pipeline provided by the authors. We then extract variant embeddings by extracting the last layer latent representations from the model for the reference and alternative allele sequence with a window size of 128 tokens centered around the respective variant and aggregating them by averaging over the sequence dimension and the forward and reverse strands. We then concatenate these reference and allele embeddings.

Furthermore, we compute zero-shot pathogenicity scores using the largest Nucleotide Transformer model (2.5B parameters) pretrained on multispecies genomes. To compute the score for a given variant, we extract the reference and the alternative genetic sequence centered around the respective variant (window size 6 000) and compute the cosine similarity of alternative and reference embeddings, where lower cosine similarities indicate higher pathogenicity (Dalla-Torre et al., 2023).

The inference for all investigated variants was performed on a cluster running Rocky Linux OS 8.7 (Green Obsidian) and slurm 23.02.7 as workload manager, using four NVLink-connected NVIDIA A100 GPUs. The inference was run in Python virtual environments using PyTorch 2.1.2 and CUDA Toolkit 12.1, with the wall times for inference being less than 1 day for all variants using GPN-MSA and less than 48 hours for the largest chromosome using Nucleotide Transformer.

2.3 Prediction of pathogenicity and polygenic effects

We evaluated three strategies for determining the polygenic relevance of genetic variants through the application of a DNA language model, specifically:

1. **Zero-shot**: The zero-shot scores from GPN-MSA without additional training. The log-likelihood ratio between the alternate and reference allele probabilities as predicted by the

model was used as a quantitative proxy measure of variant effect. Analogously, the cosine similarity scores produced by the Nucleotide Transformer were directly used to measure the variant effect.

2. **ClinVar Pathogenicity Prediction**: We train a logistic regression model to directly predict whether a variant is annotated as pathogenic in ClinVar (Landrum et al., 2014) based on the DNA language model embeddings.
3. **PGS Effect Prediction**: We aggregate scores from *The Polygenic Score Catalog* (Lambert et al., 2021) and train a logistic regression model to predict whether a variant has a significant polygenic effect, defined as a total effect size larger than 1 after summation over all 742 PGS across phenotypes.

2.4 Phenome-wide disease onset prediction

We extracted several datasets, each containing the top 100 000 variants by score from the comprehensive set of 16 million variants, as detailed in Table 1. These datasets comprise a baseline set of randomly selected variants, along with multiple sets based on the top predicted scores from the GPN-MSA and Nucleotide Transformer models. Additionally, we extracted a set of 5 879 variants from the phenotype-genotype reference map (PGRM) that was compiled to include highly robust disease-associated variants, serving as another baseline for phenome-wide predictions (Bastarache et al., 2023).

Subsequently, we generated time-to-event labels for all individuals for all endpoints with a minimum of 100 incident events, as identified in the linked electronic health records of the UK Biobank cohort. We use PheCodeX to phenotype the diseases (Shuey et al., 2023). This process was performed using a data extraction and preprocessing pipeline as outlined in (Steinfeldt et al., 2023).

For each variant set, we fitted a multi-target linear Cox regression model using PyTorch, designed to predict partial log hazards for the 1 727 endpoints with an adapted proportional hazard loss (Kvamme et al., 2019). After training, we restored the subset of parameters for each endpoint to those from the epoch with the lowest validation loss. We then used this model to predict partial log hazards on a held-out test set in a ten-fold cross-validation process, scoring the model for each endpoint and fold using the concordance index (C-index) as the metric. We finally aggregated and reported mean concordance indices for each disease. We filtered for diseases with a significant genetic signal by selecting endpoints where the lower 1% confidence interval of the cross-validation concordance index of any model exceeded 0.5, resulting in a total set of 533 endpoints.

We first limit our analyses to the Caucasian population in the UK Biobank (Field 22006, self-identified white British and not an outlier in genomic principle component space) and exclude individuals with kinship to other study participants (Field 22021, third degree or closer) (Bycroft et al., 2018) to avoid leakage between training and test dataset by partial sequence identity. The remaining data is then split into 10 non-overlapping cross-validation splits of 27k individuals each. To further investigate the generalization of the model to populations of different ancestries, we defined a set of individuals not identified as Caucasian.

Table 1: Variant sets for phenome-wide disease onset prediction.

Name	Variants	DNA language model	Variant scoring strategy
`baseline-random`	100 000	-	-
`baseline-pgrm`	5 879	-	-
`gpn-msa-zero`	100 000	GPN-MSA	Zero-shot (log-likelihood ratio)
`gpn-msa-clinvar`	100 000	GPN-MSA	ClinVar pathogenicity
`gpn-msa-polygenic`	100 000	GPN-MSA	Polygenic effect
`nt-cosine-zero`	100 000	Nucleotide Transformer 2.5b	Zero-shot (cosine distance)

3 Results

To better understand the DNA language models behavior, we performed an enrichment analysis to examine the distribution of mean minor allele frequencies (MAF) and the prevalence of specific variant effects, as categorized by the *most severe consequence* annotation from OpenTargets

Genetics (Ghoussaini et al., 2020), across different variant sets. This analysis showed that the gpn-msa-zero set is biased towards lower minor allele frequency variants compared to the baseline-random set. However, this difference was not evident when the GPN-MSA model was fine-tuned to predict polygenic effects, as observed in the gpn-msa-polygenic set. Additionally, the analysis showed that fine-tuning on PGS effects led to an enrichment of variants in regulatory regions within the gpn-msa-⋆ sets. Notably, the gpn-msa-zero set was enriched with missense variants, a trend not observed in the nt-cosine-zero set from the Nucleotide Transformer (see Table 4).

Selecting variants for genetic risk prediction requires careful consideration of linkage disequilibrium (LD) because high LD can lower the dataset's effective dimensionality and obscure causal signals. We observed that selecting the top 100,000 variants (by effect size) from either a GWAS for coronary artery disease (gwas-cad, Nikpay et al. (2015)) or a PRS for coronary artery disease (prs-cad, Khera et al. (2018b)) tends to include numerous redundant variants due to LD. The PRS slightly mitigates this redundancy. Notably, variant sets derived from the finetuned GPN-MSA models exhibit substantially lower LD than those from GWAS and PRS, suggesting a bias toward causal variants (see Table 5).

3.1 Phenome-wide disease prediction in the UK Biobank

We first focused our analysis on a selection of polygenic diseases spanning different etiologies, namely: Coronary Artery Disease (CAD), Kidney Cancer, Breast Cancer, Type 2 Diabetes (T2D), Rheumatoid Arthritis (RA), and Alzheimer's Disease. We report concordance indices for the disease onset time-to-event prediction task in the UK Biobank cohort in Table 2.

The dataset derived from GPN-MSA scores fine-tuned for polygenic effect (gpn-msa-polygenic) consistently yields the highest concordance index across all the investigated diseases with the only exception of RA. In line with our expectations, it outperforms both the dataset derived from zero-shot GPN-MSA scores (gpn-msa-zero) and the randomly selected baseline dataset (baseline-random) for all disease endpoints. Interestingly, it surpasses the dataset derived from GPN-MSA fine-tuned on ClinVar pathogenicity (gpn-msa-clinvar) by a wide margin for all diseases. Perhaps surprisingly, both the gpn-msa-clinvar dataset and the dataset derived from Nucleotide Transformer scores (nt-cosine-zero) consistently fare worse than the random baseline, with the nt-cosine-zero dataset outperforming the random baseline dataset and the otherwise dominant gpn-msa-polygenic dataset only for RA. For the baseline-pgrm dataset, we consistently obtain the lowest concordance indices across all diseases, likely due to the much smaller number of variants contained in this set.

Table 2: Concordance indices in the UK Biobank for selected polygenic diseases (CAD = Coronary artery disease, T2D = Type 2 Diabetes, RA = Rheumatoid Arthritis).

	CAD	**Kidney cancer**	**Breast cancer**	**T2D**	**RA**	**Alzheimer's**
baseline-pgrm	0.4984	0.5105	0.4991	0.5014	0.5005	0.4860
gpn-msa-clinvar	0.5125	0.5089	0.5044	0.5205	0.5069	0.4993
nt-cosine-zero	0.5520	0.5086	0.5318	0.5735	**0.5514**	0.5232
baseline-random	0.5567	0.4882	0.5342	0.5824	0.5484	0.5271
gpn-msa-zero	0.5567	0.5144	0.5363	0.5887	0.5402	0.5194
gpn-msa-polygenic	**0.5686**	**0.5237**	**0.5442**	**0.5999**	0.5487	**0.5370**

Variant selection using the DNA language models did not result in better predictive performance in the non-Caucasian population than our random baseline (e.g. 0.557 vs 0.543 for CAD, 0.569 vs 0.580 for T2D), however, both outperformed the much smaller PGRM set which only resulted in a concordance index of 0.505 for CAD and 0.496 for T2D.

We then evaluated the predictive performance of the variant sets across the human phenome by calculating mean concordance indices for the endpoint selection described in Section 2.4. Despite the overall low genetic signal and thus also low mean concordance indices, the results are largely consistent with the selection of polygenic diseases (see Table 3). Notably, for this evaluation, we selected the endpoints purely based on their incidence in the UKB without biasing them towards heritable diseases. Overall, the gpn-msa-polygenic model yields the highest average concordance index, followed by the gpn-msa-zero model. While both GPN-MSA informed variant sets have a higher average concordance index than the random variant baseline, the improvement is significantly smaller compared to the well-studied polygenic diseases in Table 2. For the nt-cosine-zero, the average concordance index is slightly lower than for the random baseline model. The gpn-msa-clinvar and baseline-pgrm models exhibit a substantially lower performance compared to the random baseline.

4 Discussion

This study used DNA language models, specifically the GPN-MSA model and the Nucleotide Transformer, to identify genetic variants that predict disease phenotypes across a broad spectrum of diseases. We complement prior work focused on local variant effects by highligthing DNA language models ability to identify variants important for disease susceptibility from both coding and non-coding regions of the genome. This is critical because most genetic associations with diseases are found in non-coding regions, where each variant contributes only slightly to the risk. Notably, this can be applied to variants that occur only once in a population (or even hypothetical ones) since the model only depends on the sequence. The challenge is to aggregate these minor effects to improve polygenic risk models, which is difficult for rare diseases and variants using traditional methods like GWAS and PGS (Boyle et al., 2017; Visscher et al., 2017).

Table 3: Phenome-wide average concordance indices and 95% confidence intervals in the UK Biobank.

	Mean C-index
baseline-pgrm	0.5003 ± 0.0006
gpn-msa-clinvar	0.5059 ± 0.0009
nt-cosine-zero	0.5235 ± 0.0016
baseline-random	0.5256 ± 0.0017
gpn-msa-zero	0.5275 ± 0.0016
gpn-msa-polygenic	**0.5305 ± 0.0017**

Our results suggest DNA language models can enhance our understanding of genotype-phenotype relationships by identifying variants with a broadly predictive power for diseases. However, to refine the predictive accuracy of genetic risk scores, we found it necessary to integrate these models with disease-specific association data from GWAS. We furthermore found instances where the models performed as poorly as a random baseline, likely stemming from the models' focus on high-impact variants, which may not uniformly predict disease susceptibility across a population. This is also evident in the severely decreased performance of the model variant predicting ClinVar pathogenicity annotations, which is likely only valid for coding variants. Adjusting the models to predict PGS effects across the phenome was notably effective, indicating that DNA language models can identify significant variants and be adapted to enhance disease predictions at the population level. However, this requires further validation across diverse datasets and direct optimization of the models to improve performance. Future work will focus on directly using association data in the DNA language model pretraining task.

We do not observe improved generalization of models based on DNA language models to the non-Caucasian population. This is in line with most PGS transferring poorly to other ancestries (Ding et al., 2023), and robust performance estimation is further limited by the lower sample size.

We currently use DNA language models to compute scores for each variant independently, then aggregate these scores with individual genotyping data for downstream analysis. A promising direction for future research involves using these models on sequences from individuals to compute integrated scores that capture polygenic information and potential epistatic effects. This approach, however, would come at a much higher computational cost because it requires inference for each individual and sequence separately, as opposed to only once per variant.

In conclusion, our research shows the potential of DNA language models in improving disease prediction and prevention. By identifying variants with broad predictive power, these models open new possibilities for enhancing polygenic risk assessments and understanding complex genotype-phenotype correlations. Future efforts should focus on refining both the pretraining task and downstream application of DNA language models, integrating them with established PGS approaches (Privé et al., 2021), and validating these findings across different genetic backgrounds in separate external validation cohorts for clinical applications.

ACKNOWLEDGMENTS

This research was conducted using data from UK Biobank, a major biomedical database via application no. 49966. We are grateful to all the participants and all members of the study teams who have enabled this research.

REFERENCES

Lisa Bastarache, Sarah Delozier, Anita Pandit, Jing He, Adam Lewis, Aubrey C. Annis, Jonathon LeFaive, Joshua C. Denny, Robert J. Carroll, Russ B. Altman, Jacob J. Hughey, Matthew Zawistowski, and Josh F. Peterson. The phenotype-genotype reference map: Improving biobank data science through replication. *American Journal of Human Genetics*, 110(9):1522–1533, September 2023. ISSN 1537-6605. doi: 10.1016/j.ajhg.2023.07.012.

Gonzalo Benegas, Carlos Albors, Alan J. Aw, Chengzhong Ye, and Yun S. Song. GPN-MSA: an alignment-based DNA language model for genome-wide variant effect prediction, October 2023. URL `https://www.biorxiv.org/content/10.1101/2023.10.10.561776v1`. Pages: 2023.10.10.561776 Section: New Results.

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, June 2017. ISSN 0092-8674. doi: 10.1016/j.cell.2017.05.038. URL `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5536862/`.

Clare Bycroft, Colin Freeman, Desislava Petkova, Gavin Band, Lloyd T. Elliott, Kevin Sharp, Allan Motyer, Damjan Vukcevic, Olivier Delaneau, Jared O'Connell, Adrian Cortes, Samantha Welsh, Alan Young, Mark Effingham, Gil McVean, Stephen Leslie, Naomi Allen, Peter Donnelly, and Jonathan Marchini. The UK Biobank resource with deep phenotyping and genomic data. *Nature*, 562(7726):203–209, October 2018. ISSN 1476-4687. doi: 10.1038/s41586-018-0579-z. URL `https://www.nature.com/articles/s41586-018-0579-z`. Number: 7726 Publisher: Nature Publishing Group.

Nathalie Chami and Guillaume Lettre. Lessons and Implications from Genome-Wide Association Studies (GWAS) Findings of Blood Cell Phenotypes. *Genes*, 5(1):51–64, January 2014. ISSN 2073-4425. doi: 10.3390/genes5010051. URL `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3978511/`.

Jun Cheng, Guido Novati, Joshua Pan, Clare Bycroft, Akvilė Žemgulytė, Taylor Applebaum, Alexander Pritzel, Lai Hong Wong, Michal Zielinski, Tobias Sargeant, Rosalia G. Schneider, Andrew W. Senior, John Jumper, Demis Hassabis, Pushmeet Kohli, and Žiga Avsec. Accurate proteome-wide missense variant effect prediction with AlphaMissense. *Science*, 381(6664): eadg7492, September 2023. doi: 10.1126/science.adg7492. URL `https://www.science.org/doi/10.1126/science.adg7492`. Publisher: American Association for the Advancement of Science.

Elisa Cirillo, Martina Kutmon, Manuel Gonzalez Hernandez, Tom Hooimeijer, Michiel E. Adriaens, Lars M. T. Eijssen, Laurence D. Parnell, Susan L. Coort, and Chris T. Evelo. From SNPs to pathways: Biological interpretation of type 2 diabetes (T2DM) genome wide association study (GWAS) results. *PLoS ONE*, 13(4), 2018. doi: 10.1371/journal.pone.0193515. URL `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5884486/`. Publisher: PLOS.

Jennifer A. Collister, Xiaonan Liu, and Lei Clifton. Calculating Polygenic Risk Scores (PRS) in UK Biobank: A Practical Guide for Epidemiologists. *Frontiers in Genetics*, 13, 2022. ISSN 1664-8021. URL `https://www.frontiersin.org/journals/genetics/articles/10.3389/fgene.2022.818574`.

Hugo Dalla-Torre, Liam Gonzalez, Javier Mendoza-Revilla, Nicolas Lopez Carranza, Adam Henryk Grzywaczewski, Francesco Oteri, Christian Dallago, Evan Trop, Bernardo P. de Almeida, Hassan Sirelkhatim, Guillaume Richard, Marcin Skwark, Karim Beguir, Marie Lopez, and Thomas Pierrot. The Nucleotide Transformer: Building and Evaluating Robust Foundation Models for Human Genomics, September 2023. URL `https://www.biorxiv.org/content/10.1101/2023.01.11.523679v3`. Pages: 2023.01.11.523679 Section: New Results.

Petr Danecek, Adam Auton, Goncalo Abecasis, Cornelis A. Albers, Eric Banks, Mark A. DePristo, Robert E. Handsaker, Gerton Lunter, Gabor T. Marth, Stephen T. Sherry, Gilean McVean, and Richard Durbin. The variant call format and VCFtools. *Bioinformatics*, 27(15):2156–2158, August 2011. ISSN 1367-4803. doi: 10.1093/bioinformatics/btr330. URL `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3137218/`.

Yi Ding, Kangcheng Hou, Ziqi Xu, Aditya Pimplaskar, Ella Petter, Kristin Boulier, Florian Privé, Bjarni J. Vilhjálmsson, Loes M. Olde Loohuis, and Bogdan Pasaniuc. Polygenic scoring accuracy varies across the genetic ancestry continuum. *Nature*, 618(7966):774–781, June 2023. ISSN 1476-4687. doi: 10.1038/s41586-023-06079-4. URL `https://www.nature.com/articles/s41586-023-06079-4`. Number: 7966 Publisher: Nature Publishing Group.

Maya Ghoussaini, Edward Mountjoy, Miguel Carmona, Gareth Peat, Ellen M Schmidt, Andrew Hercules, Luca Fumis, Alfredo Miranda, Denise Carvalho-Silva, Annalisa Buniello, Tony Burdett, James Hayhurst, Jarrod Baker, Javier Ferrer, Asier Gonzalez-Uriarte, Simon Jupp, Mohd Anisul Karim, Gautier Koscielny, Sandra Machlitt-Northen, Cinzia Malangone, Zoe May Pendlington, Paola Roncaglia, Daniel Suveges, Daniel Wright, Olga Vrousgou, Eliseo Papa, Helen Parkinson, Jacqueline A L MacArthur, John A Todd, Jeffrey C Barrett, Jeremy Schwartzentruber, David G Hulcoop, David Ochoa, Ellen M McDonagh, and Ian Dunham. Open Targets Genetics: systematic identification of trait-associated genes using large-scale genetics and functional genomics. *Nucleic Acids Research*, 49(D1):D1311–D1320, October 2020. ISSN 0305-1048. doi: 10.1093/nar/gkaa840. URL `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7778936/`.

Shelley-Ann M. Girwar, Robert Jabroer, Marta Fiocco, Stephen P. Sutch, Mattijs E. Numans, and Marc A. Bruijnzeels. A systematic review of risk stratification tools internationally used in primary care settings. *Health Science Reports*, 4(3):e329, July 2021. ISSN 2398-8835. doi: 10.1002/hsr2.329. URL `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8299990/`.

Aroon D Hingorani, Jasmine Gratton, Chris Finan, A Floriaan Schmidt, Riyaz Patel, Reecha Sofat, Valerie Kuan, Claudia Langenberg, Harry Hemingway, Joan K Morris, and Nicholas J Wald. Performance of polygenic risk scores in screening, prediction, and risk stratification: secondary analysis of data in the Polygenic Score Catalog. *BMJ Medicine*, 2(1):e000554, October 2023. ISSN 2754-0413. doi: 10.1136/bmjmed-2023-000554. URL `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10582890/`.

Yanrong Ji, Zhihan Zhou, Han Liu, and Ramana V Davuluri. DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome. *Bioinformatics*, 37(15):2112–2120, August 2021. ISSN 1367-4803. doi: 10.1093/bioinformatics/btab083. URL `https://doi.org/10.1093/bioinformatics/btab083`.

Amit V. Khera, Mark Chaffin, Krishna G. Aragam, Mary E. Haas, Carolina Roselli, Seung Hoan Choi, Pradeep Natarajan, Eric S. Lander, Steven A. Lubitz, Patrick T. Ellinor, and Sekar Kathiresan. Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nature Genetics*, 50(9):1219–1224, September 2018a. ISSN 1546-1718. doi: 10.1038/s41588-018-0183-z.

Amit V. Khera, Mark Chaffin, Krishna G. Aragam, Mary E. Haas, Carolina Roselli, Seung Hoan Choi, Pradeep Natarajan, Eric S. Lander, Steven A. Lubitz, Patrick T. Ellinor, and Sekar Kathiresan. Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nature Genetics*, 50(9):1219–1224, September 2018b. ISSN 1546-1718. doi: 10.1038/s41588-018-0183-z. URL `https://www.nature.com/articles/s41588-018-0183-z`. Publisher: Nature Publishing Group.

Mitja I. Kurki, Juha Karjalainen, Priit Palta, Timo P. Sipilä, Kati Kristiansson, Kati M. Donner, Mary P. Reeve, Hannele Laivuori, Mervi Aavikko, Mari A. Kaunisto, Anu Loukola, Elisa Lahtela, Hannele Mattsson, Päivi Laiho, Pietro Della Briotta Parolo, Arto A. Lehisto, Masahiro Kanai, Nina Mars, Joel Rämö, Tuomo Kiiskinen, Henrike O. Heyne, Kumar Veerapen, Sina Rüeger, Susanna Lemmelä, Wei Zhou, Sanni Ruotsalainen, Kalle Pärn, Tero Hiekkalinna, Sami Koskelainen, Teemu Paajanen, Vincent Llorens, Javier Gracia-Tabuenca, Harri Siirtola, Kadri Reis, Abdelrahman G. Elnahas, Benjamin Sun, Christopher N. Foley, Katriina Aalto-Setälä, Kaur Alasoo, Mikko Arvas, Kirsi Auro, Shameek Biswas, Argyro Bizaki-Vallaskangas, Olli

Carpen, Chia-Yen Chen, Oluwaseun A. Dada, Zhihao Ding, Margaret G. Ehm, Kari Eklund, Martti Färkkilä, Hilary Finucane, Andrea Ganna, Awaisa Ghazal, Robert R. Graham, Eric M. Green, Antti Hakanen, Marco Hautalahti, Åsa K. Hedman, Mikko Hiltunen, Reetta Hinttala, Iiris Hovatta, Xinli Hu, Adriana Huertas-Vazquez, Laura Huilaja, Julie Hunkapiller, Howard Jacob, Jan-Nygaard Jensen, Heikki Joensuu, Sally John, Valtteri Julkunen, Marc Jung, Juhani Junttila, Kai Kaarniranta, Mika Kähönen, Risto Kajanne, Lila Kallio, Reetta Kälviäinen, Jaakko Kaprio, Nurlan Kerimov, Johannes Kettunen, Elina Kilpeläinen, Terhi Kilpi, Katherine Klinger, Veli-Matti Kosma, Teijo Kuopio, Venla Kurra, Triin Laisk, Jari Laukkanen, Nathan Lawless, Aoxing Liu, Simonne Longerich, Reedik Mägi, Johanna Mäkelä, Antti Mäkitie, Anders Malarstig, Arto Mannermaa, Joseph Maranville, Athena Matakidou, Tuomo Meretoja, Sahar V. Mozaffari, Mari E. K. Niemi, Marianna Niemi, Teemu Niiranen, Christopher J. O´Donnell, Ma´en Obeidat, George Okafo, Hanna M. Ollila, Antti Palomäki, Tuula Palotie, Jukka Partanen, Dirk S. Paul, Margit Pelkonen, Rion K. Pendergrass, Slavé Petrovski, Anne Pitkäranta, Adam Platt, David Pulford, Eero Punkka, Pirkko Pussinen, Neha Raghavan, Fedik Rahimov, Deepak Rajpal, Nicole A. Renaud, Bridget Riley-Gillis, Rodosthenis Rodosthenous, Elmo Saarentaus, Aino Salminen, Eveliina Salminen, Veikko Salomaa, Johanna Schleutker, Raisa Serpi, Huei-yi Shen, Richard Siegel, Kaisa Silander, Sanna Siltanen, Sirpa Soini, Hilkka Soininen, Jae Hoon Sul, Ioanna Tachmazidou, Kaisa Tasanen, Pentti Tienari, Sanna Toppila-Salmi, Taru Tukiainen, Tiinamaija Tuomi, Joni A. Turunen, Jacob C. Ulirsch, Felix Vaura, Petri Virolainen, Jeffrey Waring, Dawn Waterworth, Robert Yang, Mari Nelis, Anu Reigo, Andres Metspalu, Lili Milani, Tõnu Esko, Caroline Fox, Aki S. Havulinna, Markus Perola, Samuli Ripatti, Anu Jalanko, Tarja Laitinen, Tomi P. Mäkelä, Robert Plenge, Mark McCarthy, Heiko Runz, Mark J. Daly, and Aarno Palotie. FinnGen provides genetic insights from a well-phenotyped isolated population. *Nature*, 613(7944):508–518, January 2023. ISSN 1476-4687. doi: 10.1038/s41586-022-05473-8. URL https://www.nature.com/articles/s41586-022-05473-8. Number: 7944 Publisher: Nature Publishing Group.

Håvard Kvamme, Ørnulf Borgan, and Ida Scheel. Time-to-Event Prediction with Neural Networks and Cox Regression. *Journal of Machine Learning Research*, 20(129):1–30, 2019. ISSN 1533-7928. URL http://jmlr.org/papers/v20/18-424.html.

Samuel A. Lambert, Laurent Gil, Simon Jupp, Scott C. Ritchie, Yu Xu, Annalisa Buniello, Aoife McMahon, Gad Abraham, Michael Chapman, Helen Parkinson, John Danesh, Jacqueline A. L. MacArthur, and Michael Inouye. The Polygenic Score Catalog as an open database for reproducibility and systematic evaluation. *Nature Genetics*, 53(4):420–425, April 2021. ISSN 1546-1718. doi: 10.1038/s41588-021-00783-5. URL https://www.nature.com/articles/s41588-021-00783-5. Number: 4 Publisher: Nature Publishing Group.

Melissa J. Landrum, Jennifer M. Lee, George R. Riley, Wonhee Jang, Wendy S. Rubinstein, Deanna M. Church, and Donna R. Maglott. ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Research*, 42(Database issue):D980–D985, January 2014. ISSN 0305-1048. doi: 10.1093/nar/gkt1113. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3965032/.

Teri A. Manolio, Francis S. Collins, Nancy J. Cox, David B. Goldstein, Lucia A. Hindorff, David J. Hunter, Mark I. McCarthy, Erin M. Ramos, Lon R. Cardon, Aravinda Chakravarti, Judy H. Cho, Alan E. Guttmacher, Augustine Kong, Leonid Kruglyak, Elaine Mardis, Charles N. Rotimi, Montgomery Slatkin, David Valle, Alice S. Whittemore, Michael Boehnke, Andrew G. Clark, Evan E. Eichler, Greg Gibson, Jonathan L. Haines, Trudy F. C. Mackay, Steven A. McCarroll, and Peter M. Visscher. Finding the missing heritability of complex diseases. *Nature*, 461(7265):747–753, October 2009. ISSN 1476-4687. doi: 10.1038/nature08494. URL https://www.nature.com/articles/nature08494. Number: 7265 Publisher: Nature Publishing Group.

Alicia R. Martin, Masahiro Kanai, Yoichiro Kamatani, Yukinori Okada, Benjamin M. Neale, and Mark J. Daly. Clinical use of current polygenic risk scores may exacerbate health disparities. *Nature Genetics*, 51(4):584–591, April 2019. ISSN 1546-1718. doi: 10.1038/s41588-019-0379-x. URL https://www.nature.com/articles/s41588-019-0379-x. Number: 4 Publisher: Nature Publishing Group.

Eric Nguyen, Michael Poli, Marjan Faizi, Armin Thomas, Callum Birch-Sykes, Michael Wornow, Aman Patel, Clayton Rabideau, Stefano Massaroli, Yoshua Bengio, Stefano Ermon, Stephen A. Baccus, and Chris Ré. HyenaDNA: Long-Range Genomic Sequence Modeling at Single Nucleotide Resolution, November 2023. URL `http://arxiv.org/abs/2306.15794`. arXiv:2306.15794 [cs, q-bio].

Majid Nikpay, Anuj Goel, Hong-Hee Won, Leanne M Hall, Christina Willenborg, Stavroula Kanoni, Danish Saleheen, Theodosios Kyriakou, Christopher P Nelson, Jemma C Hopewell, Thomas R Webb, Lingyao Zeng, Abbas Dehghan, Maris Alver, Sebastian M Armasu, Kirsi Auro, Andrew Bjonnes, Daniel I Chasman, Shufeng Chen, Ian Ford, Nora Franceschini, Christian Gieger, Christopher Grace, Stefan Gustafsson, Jie Huang, Shih-Jen Hwang, Yun Kyoung Kim, Marcus E Kleber, King Wai Lau, Xiangfeng Lu, Yingchang Lu, Leo-Pekka Lyytikäinen, Evelin Mihailov, Alanna C Morrison, Natalia Pervjakova, Liming Qu, Lynda M Rose, Elias Salfati, Richa Saxena, Markus Scholz, Albert V Smith, Emmi Tikkanen, Andre Uitterlinden, Xueli Yang, Weihua Zhang, Wei Zhao, Mariza de Andrade, Paul S de Vries, Natalie R van Zuydam, Sonia S Anand, Lars Bertram, Frank Beutner, George Dedoussis, Philippe Frossard, Dominique Gauguier, Alison H Goodall, Omri Gottesman, Marc Haber, Bok-Ghee Han, Jianfeng Huang, Shapour Jalilzadeh, Thorsten Kessler, Inke R König, Lars Lannfelt, Wolfgang Lieb, Lars Lind, Cecilia M Lindgren, Marja-Liisa Lokki, Patrik K Magnusson, Nadeem H Mallick, Narinder Mehra, Thomas Meitinger, Fazal-ur-Rehman Memon, Andrew P Morris, Markku S Nieminen, Nancy L Pedersen, Annette Peters, Loukianos S Rallidis, Asif Rasheed, Maria Samuel, Svati H Shah, Juha Sinisalo, Kathleen E Stirrups, Stella Trompet, Laiyuan Wang, Khan S Zaman, Diego Ardissino, Eric Boerwinkle, Ingrid B Borecki, Erwin P Bottinger, Julie E Buring, John C Chambers, Rory Collins, L Adrienne Cupples, John Danesh, Ilja Demuth, Roberto Elosua, Stephen E Epstein, Tõnu Esko, Mary F Feitosa, Oscar H Franco, Maria Grazia Franzosi, Christopher B Granger, Dongfeng Gu, Vilmundur Gudnason, Alistair S Hall, Anders Hamsten, Tamara B Harris, Stanley L Hazen, Christian Hengstenberg, Albert Hofman, Erik Ingelsson, Carlos Iribarren, J Wouter Jukema, Pekka J Karhunen, Bong-Jo Kim, Jaspal S Kooner, Iftikhar J Kullo, Terho Lehtimäki, Ruth J F Loos, Olle Melander, Andres Metspalu, Winfried März, Colin N Palmer, Markus Perola, Thomas Quertermous, Daniel J Rader, Paul M Ridker, Samuli Ripatti, Robert Roberts, Veikko Salomaa, Dharambir K Sanghera, Stephen M Schwartz, Udo Seedorf, Alexandre F Stewart, David J Stott, Joachim Thiery, Pierre A Zalloua, Christopher J O'Donnell, Muredach P Reilly, Themistocles L Assimes, John R Thompson, Jeanette Erdmann, Robert Clarke, Hugh Watkins, Sekar Kathiresan, Ruth McPherson, Panos Deloukas, Heribert Schunkert, Nilesh J Samani, Martin Farrall, and the CARDIoGRAMplusC4D Consortium. A comprehensive 1000 Genomes–based genome-wide association meta-analysis of coronary artery disease. *Nature Genetics*, 47(10):1121–1130, October 2015. ISSN 1546-1718. doi: 10.1038/ng.3396. URL `https://www.nature.com/articles/ng.3396`. Publisher: Nature Publishing Group.

Florian Privé, Julyan Arbel, and Bjarni J. Vilhjálmsson. LDpred2: better, faster, stronger. *Bioinformatics (Oxford, England)*, 36(22-23):5424–5431, April 2021. ISSN 1367-4811. doi: 10.1093/bioinformatics/btaa1029.

Shaun Purcell, Benjamin Neale, Kathe Todd-Brown, Lori Thomas, Manuel A. R. Ferreira, David Bender, Julian Maller, Pamela Sklar, Paul I. W. de Bakker, Mark J. Daly, and Pak C. Sham. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *American Journal of Human Genetics*, 81(3):559–575, September 2007. ISSN 0002-9297. URL `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1950838/`.

Megan M Shuey, William W Stead, Ida Aka, April L Barnado, Julie A Bastarache, Elly Brokamp, Meredith Campbell, Robert J Carroll, Jeffrey A Goldstein, Adam Lewis, Beth A Malow, Jonathan D Mosley, Travis Osterman, Dolly A Padovani-Claudio, Andrea Ramirez, Dan M Roden, Bryce A Schuler, Edward Siew, Jennifer Sucre, Isaac Thomsen, Rory J Tinker, Sara Van Driest, Colin Walsh, Jeremy L Warner, Quinn S Wells, Lee Wheless, and Lisa Bastarache. Next-generation phenotyping: introducing phecodeX for enhanced discovery research in medical phenomics. *Bioinformatics*, 39(11):btad655, November 2023. ISSN 1367-4811. doi: 10.1093/bioinformatics/btad655. URL `https://doi.org/10.1093/bioinformatics/btad655`.

Jakob Steinfeldt, Benjamin Wild, Thore Buergel, Maik Pietzner, Julius Upmeier zu Belzen, Andre Vauvelle, Stefan Hegselmann, Spiros Denaxas, Harry Hemingway, Claudia Langenberg, Ulf Landmesser, John Deanfield, and Roland Eils. Medical history predicts phenome-wide disease onset and enables the rapid response to emerging health threats, September 2023. URL https://www.medrxiv.org/content/10.1101/2023.03.10.23286918v3. Pages: 2023.03.10.23286918.

Cathie Sudlow, John Gallacher, Naomi Allen, Valerie Beral, Paul Burton, John Danesh, Paul Downey, Paul Elliott, Jane Green, Martin Landray, Bette Liu, Paul Matthews, Giok Ong, Jill Pell, Alan Silman, Alan Young, Tim Sprosen, Tim Peakman, and Rory Collins. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS medicine*, 12(3):e1001779, March 2015. ISSN 1549-1676. doi: 10.1371/journal.pmed.1001779.

Ali Torkamani, Nathan E. Wineinger, and Eric J. Topol. The personal and clinical utility of polygenic risk scores. *Nature Reviews. Genetics*, 19(9):581–590, September 2018. ISSN 1471-0064. doi: 10.1038/s41576-018-0018-x.

Peter M. Visscher, Naomi R. Wray, Qian Zhang, Pamela Sklar, Mark I. McCarthy, Matthew A. Brown, and Jian Yang. 10 Years of GWAS Discovery: Biology, Function, and Translation. *American Journal of Human Genetics*, 101(1):5–22, July 2017. ISSN 0002-9297. doi: 10.1016/j.ajhg.2017.06.005. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5501872/.

Christine Vogeli, Alexandra E. Shields, Todd A. Lee, Teresa B. Gibson, William D. Marder, Kevin B. Weiss, and David Blumenthal. Multiple Chronic Conditions: Prevalence, Health Consequences, and Implications for Quality, Care Management, and Costs. *Journal of General Internal Medicine*, 22(Suppl 3):391–395, December 2007. ISSN 0884-8734. doi: 10.1007/s11606-007-0322-1. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2150598/.

Loïc Yengo, Sailaja Vedantam, Eirini Marouli, Julia Sidorenko, Eric Bartell, Saori Sakaue, Marielisa Graff, Anders U. Eliasen, Yunxuan Jiang, Sridharan Raghavan, Jenkai Miao, Joshua D. Arias, Sarah E. Graham, Ronen E. Mukamel, Cassandra N. Spracklen, Xianyong Yin, Shyh-Huei Chen, Teresa Ferreira, Heather H. Highland, Yingjie Ji, Tugce Karaderi, Kuang Lin, Kreete Lüll, Deborah E. Malden, Carolina Medina-Gomez, Moara Machado, Amy Moore, Sina Rüeger, Xueling Sim, Scott Vrieze, Tarunveer S. Ahluwalia, Masato Akiyama, Matthew A. Allison, Marcus Alvarez, Mette K. Andersen, Alireza Ani, Vivek Appadurai, Liubov Arbeeva, Seema Bhaskar, Lawrence F. Bielak, Sailalitha Bollepalli, Lori L. Bonnycastle, Jette Bork-Jensen, Jonathan P. Bradfield, Yuki Bradford, Peter S. Braund, Jennifer A. Brody, Kristoffer S. Burgdorf, Brian E. Cade, Hui Cai, Qiuyin Cai, Archie Campbell, Marisa Cañadas-Garre, Eulalia Catamo, Jin-Fang Chai, Xiaoran Chai, Li-Ching Chang, Yi-Cheng Chang, Chien-Hsiun Chen, Alessandra Chesi, Seung Hoan Choi, Ren-Hua Chung, Massimiliano Cocca, Maria Pina Concas, Christian Couture, Gabriel Cuellar-Partida, Rebecca Danning, E. Warwick Daw, Frauke Degenhard, Graciela E. Delgado, Alessandro Delitala, Ayse Demirkan, Xuan Deng, Poornima Devineni, Alexander Dietl, Maria Dimitriou, Latchezar Dimitrov, Rajkumar Dorajoo, Arif B. Ekici, Jorgen E. Engmann, Zammy Fairhurst-Hunter, Aliki-Eleni Farmaki, Jessica D. Faul, Juan-Carlos Fernandez-Lopez, Lukas Forer, Margherita Francescatto, Sandra Freitag-Wolf, Christian Fuchsberger, Tessel E. Galesloot, Yan Gao, Zishan Gao, Frank Geller, Olga Giannakopoulou, Franco Giulianini, Anette P. Gjesing, Anuj Goel, Scott D. Gordon, Mathias Gorski, Jakob Grove, Xiuqing Guo, Stefan Gustafsson, Jeffrey Haessler, Thomas F. Hansen, Aki S. Havulinna, Simon J. Haworth, Jing He, Nancy Heard-Costa, Prashantha Hebbar, George Hindy, Yuk-Lam A. Ho, Edith Hofer, Elizabeth Holliday, Katrin Horn, Whitney E. Hornsby, Jouke-Jan Hottenga, Hongyan Huang, Jie Huang, Alicia Huerta-Chagoya, Jennifer E. Huffman, Yi-Jen Hung, Shaofeng Huo, Mi Yeong Hwang, Hiroyuki Iha, Daisuke D. Ikeda, Masato Isono, Anne U. Jackson, Susanne Jäger, Iris E. Jansen, Ingegerd Johansson, Jost B. Jonas, Anna Jonsson, Torben Jørgensen, Ioanna-Panagiota Kalafati, Masahiro Kanai, Stavroula Kanoni, Line L. Kårhus, Anuradhani Kasturiratne, Tomohiro Katsuya, Takahisa Kawaguchi, Rachel L. Kember, Katherine A. Kentistou, Han-Na Kim, Young Jin Kim, Marcus E. Kleber, Maria J. Knol, Azra Kurbasic, Marie Lauzon, Phuong Le, Rodney Lea, Jong-Young Lee, Hampton L. Leonard, Shengchao A. Li, Xiaohui Li, Xiaoyin Li, Jingjing Liang, Honghuang Lin, Shih-Yi Lin, Jun Liu, Xueping Liu, Ken Sin Lo, Jirong Long, Laura Lores-Motta, Jian'an Luan, Valeriya Lyssenko, Leo-Pekka Lyytikäinen, Anubha Mahajan, Vasiliki Mamakou, Massimo Mangino, Ani Manichaikul, Jonathan Marten, Manuel Mattheisen,

Laven Mavarani, Aaron F. McDaid, Karina Meidtner, Tori L. Melendez, Josep M. Mercader, Yuri Milaneschi, Jason E. Miller, Iona Y. Millwood, Pashupati P. Mishra, Ruth E. Mitchell, Line T. Møllehave, Anna Morgan, Soeren Mucha, Matthias Munz, Masahiro Nakatochi, Christopher P. Nelson, Maria Nethander, Chu Won Nho, Aneta A. Nielsen, Ilja M. Nolte, Suraj S. Nongmaithem, Raymond Noordam, Ioanna Ntalla, Teresa Nutile, Anita Pandit, Paraskevi Christofidou, Katri Pärna, Marc Pauper, Eva R. B. Petersen, Liselotte V. Petersen, Niina Pitkänen, Ozren Polašek, Alaitz Poveda, Michael H. Preuss, Saiju Pyarajan, Laura M. Raffield, Hiromi Rakugi, Julia Ramirez, Asif Rasheed, Dennis Raven, Nigel W. Rayner, Carlos Riveros, Rebecca Rohde, Daniela Ruggiero, Sanni E. Ruotsalainen, Kathleen A. Ryan, Maria Sabater-Lleal, Richa Saxena, Markus Scholz, Anoop Sendamarai, Botong Shen, Jingchunzi Shi, Jae Hun Shin, Carlo Sidore, Colleen M. Sitlani, Roderick C. Slieker, Roelof A. J. Smit, Albert V. Smith, Jennifer A. Smith, Laura J. Smyth, Lorraine Southam, Valgerdur Steinthorsdottir, Liang Sun, Fumihiko Takeuchi, Divya Sri Priyanka Tallapragada, Kent D. Taylor, Bamidele O. Tayo, Catherine Tcheandjieu, Natalie Terzikhan, Paola Tesolin, Alexander Teumer, Elizabeth Theusch, Deborah J. Thompson, Gudmar Thorleifsson, Paul R. H. J. Timmers, Stella Trompet, Constance Turman, Simona Vaccargiu, Sander W. van der Laan, Peter J. van der Most, Jan B. van Klinken, Jessica van Setten, Shefali S. Verma, Niek Verweij, Yogasudha Veturi, Carol A. Wang, Chaolong Wang, Lihua Wang, Zhe Wang, Helen R. Warren, Wen Bin Wei, Ananda R. Wickremasinghe, Matthias Wielscher, Kerri L. Wiggins, Bendik S. Winsvold, Andrew Wong, Yang Wu, Matthias Wuttke, Rui Xia, Tian Xie, Ken Yamamoto, Jingyun Yang, Jie Yao, Hannah Young, Noha A. Yousri, Lei Yu, Lingyao Zeng, Weihua Zhang, Xinyuan Zhang, Jing-Hua Zhao, Wei Zhao, Wei Zhou, Martina E. Zimmermann, Magdalena Zoledziewska, Linda S. Adair, Hieab H. H. Adams, Carlos A. Aguilar-Salinas, Fahd Al-Mulla, Donna K. Arnett, Folkert W. Asselbergs, Bjørn Olav Åsvold, John Attia, Bernhard Banas, Stefania Bandinelli, David A. Bennett, Tobias Bergler, Dwaipayan Bharadwaj, Ginevra Biino, Hans Bisgaard, Eric Boerwinkle, Carsten A. Böger, Klaus Bønnelykke, Dorret I. Boomsma, Anders D. Børglum, Judith B. Borja, Claude Bouchard, Donald W. Bowden, Ivan Brandslund, Ben Brumpton, Julie E. Buring, Mark J. Caulfield, John C. Chambers, Giriraj R. Chandak, Stephen J. Chanock, Nish Chaturvedi, Yii-Der Ida Chen, Zhengming Chen, Ching-Yu Cheng, Ingrid E. Christophersen, Marina Ciullo, John W. Cole, Francis S. Collins, Richard S. Cooper, Miguel Cruz, Francesco Cucca, L. Adrienne Cupples, Michael J. Cutler, Scott M. Damrauer, Thomas M. Dantoft, Gert J. de Borst, Lisette C. P. G. M. de Groot, Philip L. De Jager, Dominique P. V. de Kleijn, H. Janaka de Silva, George V. Dedoussis, Anneke I. den Hollander, Shufa Du, Douglas F. Easton, Petra J. M. Elders, A. Heather Eliassen, Patrick T. Ellinor, Sölve Elmståhl, Jeanette Erdmann, Michele K. Evans, Diane Fatkin, Bjarke Feenstra, Mary F. Feitosa, Luigi Ferrucci, Ian Ford, Myriam Fornage, Andre Franke, Paul W. Franks, Barry I. Freedman, Paolo Gasparini, Christian Gieger, Giorgia Girotto, Michael E. Goddard, Yvonne M. Golightly, Clicerio Gonzalez-Villalpando, Penny Gordon-Larsen, Harald Grallert, Struan F. A. Grant, Niels Grarup, Lyn Griffiths, Vilmundur Gudnason, Christopher Haiman, Hakon Hakonarson, Torben Hansen, Catharina A. Hartman, Andrew T. Hattersley, Caroline Hayward, Susan R. Heckbert, Chew-Kiat Heng, Christian Hengstenberg, Alex W. Hewitt, Haretsugu Hishigaki, Carel B. Hoyng, Paul L. Huang, Wei Huang, Steven C. Hunt, Kristian Hveem, Elina Hyppönen, William G. Iacono, Sahoko Ichihara, M. Arfan Ikram, Carmen R. Isasi, Rebecca D. Jackson, Marjo-Riitta Jarvelin, Zi-Bing Jin, Karl-Heinz Jöckel, Peter K. Joshi, Pekka Jousilahti, J. Wouter Jukema, Mika Kähönen, Yoichiro Kamatani, Kui Dong Kang, Jaakko Kaprio, Sharon L. R. Kardia, Fredrik Karpe, Norihiro Kato, Frank Kee, Thorsten Kessler, Amit V. Khera, Chiea Chuen Khor, Lambertus A. L. M. Kiemeney, Bong-Jo Kim, Eung Kweon Kim, Hyung-Lae Kim, Paulus Kirchhof, Mika Kivimaki, Woon-Puay Koh, Heikki A. Koistinen, Genovefa D. Kolovou, Jaspal S. Kooner, Charles Kooperberg, Anna Köttgen, Peter Kovacs, Adriaan Kraaijeveld, Peter Kraft, Ronald M. Krauss, Meena Kumari, Zoltan Kutalik, Markku Laakso, Leslie A. Lange, Claudia Langenberg, Lenore J. Launer, Loic Le Marchand, Hyejin Lee, Nanette R. Lee, Terho Lehtimäki, Huaixing Li, Liming Li, Wolfgang Lieb, Xu Lin, Lars Lind, Allan Linneberg, Ching-Ti Liu, Jianjun Liu, Markus Loeffler, Barry London, Steven A. Lubitz, Stephen J. Lye, David A. Mackey, Reedik Mägi, Patrik K. E. Magnusson, Gregory M. Marcus, Pedro Marques Vidal, Nicholas G. Martin, Winfried März, Fumihiko Matsuda, Robert W. McGarrah, Matt McGue, Amy Jayne McKnight, Sarah E. Medland, Dan Mellström, Andres Metspalu, Braxton D. Mitchell, Paul Mitchell, Dennis O. Mook-Kanamori, Andrew D. Morris, Lorelei A. Mucci, Patricia B. Munroe, Mike A. Nalls, Saman Nazarian, Amanda E. Nelson, Matt J. Neville, Christopher Newton-Cheh, Christopher S. Nielsen, Markus M. Nöthen, Claes Ohlsson, Albertine J. Oldehinkel, Lorena Orozco, Katja Pahkala, Päivi Pajukanta, Colin N. A. Palmer, Esteban J. Parra, Cristian Pattaro, Oluf Ped-

ersen, Craig E. Pennell, Brenda W. J. H. Penninx, Louis Perusse, Annette Peters, Patricia A. Peyser, David J. Porteous, Danielle Posthuma, Chris Power, Peter P. Pramstaller, Michael A. Province, Qibin Qi, Jia Qu, Daniel J. Rader, Olli T. Raitakari, Sarju Ralhan, Loukianos S. Rallidis, Dabeeru C. Rao, Susan Redline, Dermot F. Reilly, Alexander P. Reiner, Sang Youl Rhee, Paul M. Ridker, Michiel Rienstra, Samuli Ripatti, Marylyn D. Ritchie, Dan M. Roden, Frits R. Rosendaal, Jerome I. Rotter, Igor Rudan, Femke Rutters, Charumathi Sabanayagam, Danish Saleheen, Veikko Salomaa, Nilesh J. Samani, Dharambir K. Sanghera, Naveed Sattar, Börge Schmidt, Helena Schmidt, Reinhold Schmidt, Matthias B. Schulze, Heribert Schunkert, Laura J. Scott, Rodney J. Scott, Peter Sever, Eric J. Shiroma, M. Benjamin Shoemaker, Xiao-Ou Shu, Eleanor M. Simonsick, Mario Sims, Jai Rup Singh, Andrew B. Singleton, Moritz F. Sinner, J. Gustav Smith, Harold Snieder, Tim D. Spector, Meir J. Stampfer, Klaus J. Stark, David P. Strachan, Leen M. 't Hart, Yasuharu Tabara, Hua Tang, Jean-Claude Tardif, Thangavel A. Thanaraj, Nicholas J. Timpson, Anke Tönjes, Angelo Tremblay, Tiinamaija Tuomi, Jaakko Tuomilehto, Maria-Teresa Tusié-Luna, Andre G. Uitterlinden, Rob M. van Dam, Pim van der Harst, Nathalie Van der Velde, Cornelia M. van Duijn, Natasja M. van Schoor, Veronique Vitart, Uwe Völker, Peter Vollenweider, Henry Völzke, Niels H. Wacher-Rodarte, Mark Walker, Ya Xing Wang, Nicholas J. Wareham, Richard M. Watanabe, Hugh Watkins, David R. Weir, Thomas M. Werge, Elisabeth Widen, Lynne R. Wilkens, Gonneke Willemsen, Walter C. Willett, James F. Wilson, Tien-Yin Wong, Jeong-Taek Woo, Alan F. Wright, Jer-Yuarn Wu, Huichun Xu, Chittaranjan S. Yajnik, Mitsuhiro Yokota, Jian-Min Yuan, Eleftheria Zeggini, Babette S. Zemel, Wei Zheng, Xiaofeng Zhu, Joseph M. Zmuda, Alan B. Zonderman, John-Anker Zwart, 23andMe Research Team, VA Million Veteran Program, DiscovEHR (DiscovEHR and MyCode Community Health Initiative), eMERGE (Electronic Medical Records and Genomics Network), Lifelines Cohort Study, PRACTICAL Consortium, Understanding Society Scientific Group, Daniel I. Chasman, Yoon Shin Cho, Iris M. Heid, Mark I. McCarthy, Maggie C. Y. Ng, Christopher J. O'Donnell, Fernando Rivadeneira, Unnur Thorsteinsdottir, Yan V. Sun, E. Shyong Tai, Michael Boehnke, Panos Deloukas, Anne E. Justice, Cecilia M. Lindgren, Ruth J. F. Loos, Karen L. Mohlke, Kari E. North, Kari Stefansson, Robin G. Walters, Thomas W. Winkler, Kristin L. Young, Po-Ru Loh, Jian Yang, Tõnu Esko, Themistocles L. Assimes, Adam Auton, Goncalo R. Abecasis, Cristen J. Willer, Adam E. Locke, Sonja I. Berndt, Guillaume Lettre, Timothy M. Frayling, Yukinori Okada, Andrew R. Wood, Peter M. Visscher, and Joel N. Hirschhorn. A saturated map of common genetic variants associated with human height. *Nature*, 610(7933):704–712, October 2022. ISSN 1476-4687. doi: 10.1038/s41586-022-05275-y.

Hao Zhao, Zhifu Sun, Jing Wang, Haojie Huang, Jean-Pierre Kocher, and Liguo Wang. CrossMap: a versatile tool for coordinate conversion between genome assemblies. *Bioinformatics*, 30(7): 1006–1007, April 2014. ISSN 1367-4803. doi: 10.1093/bioinformatics/btt730. URL `https://doi.org/10.1093/bioinformatics/btt730`.

A Appendix

Table 4: Enrichment analysis of variant sets based on the *most severe consequence* annotation in OpenTargets Genetics (Ghoussaini et al., 2020) (MAF = mean minor allele frequency, other columns: proportion of variants with the given annotation in OpenTargets Genetics).

	MAF	**intron**	**intergenic**	**regulatory**	**missense**	**splice**
`baseline-random`	10.4%	49.6%	34.6%	3.3%	0.5%	0.1%
`baseline-pgrm`	26.8%	55.2%	17.5%	4.4%	4.4%	0.3%
`gpn-msa-zero`	6.8%	36.8%	18.1%	3.9%	22.8%	1.1%
`gpn-msa-clinvar`	11.1%	50.5%	36.8%	2.6%	0.4%	0.1%
`gpn-msa-polygenic`	19.6%	48.7%	23.7%	5.0%	4.0%	0.6%
`nt-cosine-zero`	10.3%	49.4%	36.2%	3.5%	0.3%	0.1%

Table 5: Linkage disequilibrium much lower in variant sets based on DNA-language-models rather than GWAS or PRS. We excluded the `pgrm-baseline` dataset from this analysis because comparing the number of variants in LD across sets of different numbers of variants would be misleading. LD pairs per million is defined as the number of variant pairs in $LD{>}0.5$ per one million pairs.

	LD pairs per million
`gpn-msa-zero`	1.6
`baseline-random`	3.5
`nt-cosine-zero`	3.9
`gpn-msa-clinvar`	12.5
`gpn-msa-polygenic`	13.2
`prs-cad`	51.1
`gwas-cad`	86.6