SCIENTIFIC REPORTS
natureresearch

OPEN

Genetic architecture of complex traits and disease risk predictors

Soke Yuen Yong[1]✉, Timothy G. Raben[1], Louis Lello[1,2] & Stephen D. H. Hsu[1,2]

Genomic prediction of complex human traits (e.g., height, cognitive ability, bone density) and disease risks (e.g., breast cancer, diabetes, heart disease, atrial fibrillation) has advanced considerably in recent years. Using data from the UK Biobank, predictors have been constructed using penalized algorithms that favor sparsity: i.e., which use as few genetic variants as possible. We analyze the specific genetic variants (SNPs) utilized in these predictors, which can vary from dozens to as many as thirty thousand. We find that the fraction of SNPs in or near genic regions varies widely by phenotype. For the majority of disease conditions studied, *a large amount* of the variance is accounted for by SNPs outside of coding regions. The state of these SNPs cannot be determined from exome-sequencing data. This suggests that exome data alone will miss much of the heritability for these traits—i.e., existing PRS cannot be computed from exome data alone. We also study the fraction of SNPs and of variance that is in common between pairs of predictors. The DNA regions used in disease risk predictors so far constructed seem to be largely disjoint (with a few interesting exceptions), suggesting that individual genetic disease risks are largely uncorrelated. It seems possible in theory for an individual to be a low-risk outlier in all conditions simultaneously.

Genomic prediction of complex traits and disease risks has advanced considerably thanks to the recent advent of large data sets and improved algorithms. These algorithms range from simple regression, applied to one SNP at a time to estimate statistical significance and effect size (e.g., as used in GWAS), to high dimensional optimization methods such as compressed sensing or sparse learning[1–4]. They produce Polygenic Risk Scores (PRS) or Polygenic Scores (PGS): functions that map the state of an individual's DNA at specific locations (SNPs), to a risk score or predicted quantitative trait value.

Predictors (PGS or PRS) now exist for a number of important traits and risks, many of which have undergone out-of-sample testing (i.e., validation in groups of individuals not used in training and from other data sets or from separate ancestries)[5–7]. The genetic architectures (i.e., the properties of the SNPs activated in the predictors, which are sparse) uncovered vary significantly: the number of SNPs required to capture most of the predictor variance ranges from a few dozen to many thousands. In contrast, traditional Genome Wide Association studies (GWAS) can implicate the entire genome[8,9], making them unwieldy to analyze.

In the case of disease risk, the predictors are already good enough to identify *genetic risk outliers*. That is, individuals with unusually high (or low) *genetic* risk of a specific condition. There are many clinical applications for such predictors[5,10–19] (although there is still much work to be done to overcome sampling and algorithmic biases and disparity[20,21]). Below we mention two possible future examples.

Breast Cancer: Certain variants in the BRCA1 and BRCA2 genes are known to elevate Breast Cancer risk significantly[22,23]. However, these mutations affect no more than a few women per thousand in the general population[24–26]. By contrast, PRS using thousands of common SNP variants can now identify of order *ten times* as many women who are in the high-risk category[5,7,10,27,28]. Standard of Care for high-risk women typically includes additional screening, such as mammograms beginning a decade earlier than for normal risk women. Early detection can also lead to significant cost savings[29]. What can we say about the thousands of common SNPs used in the PRS? Do they overlap with SNPs used in PRS for other conditions (e.g., other cancers)?

Height: Idiopathic Short Stature (ISS) refers to extreme short stature that does not have a diagnostic explanation (e.g., height below 5 foot 2 inches in adult males)[30]. Human growth hormone (HGH) treatment is sometimes prescribed for children who are at risk for this condition, at a cost in the $100k range. Typically, these would be children in the bottom percentiles for height within their age group[31,32]. However, it is difficult for pediatric endocrinologists, whose responsibility it is to prescribe HGH for these children, to know whether the child is simply passing through a temporary phase of slow growth (and will, by adulthood, reach normal height)[33,34].

[1]Department of Physics and Astronomy, Michigan State University, East Lansing, USA. [2]Genomic Prediction, North Brunswick, NJ, USA. ✉email: yongsoke@msu.edu

Adult height prediction from DNA (with 95 percent confidence interval roughly ±2 inches) will allow physicians to avoid expensive HGH treatment (with significant potential side-effects) for children who are merely short for their age (late-developing) and are likely to be in the normal range in adulthood.

For the first time, we can begin to address some general questions concerning the genetic architectures of complex traits. In this paper we address the following questions:

1. What is the (qualitative) genetic architecture of specific disease risks? How many SNPs, where are they, how many genes?
2. How much of the total risk is controlled by loci in coding vs non-coding regions?
3. Is exome sequencing data sufficient for computation of Polygenic Risk Scores (PRS)?
4. How much genetic overlap exists between different disease architectures? With millions of SNPs in the genome it is entirely possible that different diseases have nearly disjoint genetic architectures—i.e., risk is mostly controlled by distinct regions of DNA. On the other hand, we might uncover overlap regions of DNA which affect multiple disease risks simultaneously.

In this paper we consider predictors for a selection of disease conditions/traits: asthma, atrial fibrillation, basal cell carcinoma, breast cancer, coronary artery disease, type-1 diabetes, type-2 diabetes, diastolic blood pressure, educational years, gallstones, glaucoma, gout, heart attack, height, high cholesterol, hypertension, hypothyroidism, malignant melanoma, menopause, pulse rate, and systolic blood pressure. All predictors, except the coronary artery disease predictor, were built by training on case-control phenotype data from the UK Biobank[35,36] that relied on custom array genotyping (see Appendix C in the Supplementary Information for details). This array was designed to have detailed coverage in areas known to be associated with certain phenotypes, and to contain a wide sampling of the entire human genome. More details of the array design can be found on the UK Biobank website https://www.ukbiobank.ac.uk/scientists-3/uk-biobank-axiom-array/ and in Appendix C. The predictors were derived using the L1 penalized regression (sparse learning) methods found in[5,6]. Recent algorithmic benchmarking for complex trait prediction in plants and animals has shown that linear methods work as well if not better than non-linear, Bayesian, or deep learning approaches for current data set sizes[37]. Predictors for diastolic blood pressure, systolic blood pressure, and pulse rate are reported for the first time here, but were designed as described in[5]. On the other hand, the coronary artery predictor originated from[7], and the associated minor allele frequencies were obtained using Ensembl's minor allele frequency calculator[38].

Of course, not all of the genetic variants affecting disease risk have been discovered. The PRS continue to improve as more training data become available. However, the SNPs used in the existing PRS tend to be those that account for the most risk variance. Equivalently, the statistical evidence supporting their association with the disease risk is highest. Relevant SNPs that have yet to be discovered are either common SNPs with very small effect size, or very rare SNPs that are not probed using existing gene arrays.

Methods

Most of the predictors from our research group used in this work were previously designed and published in[5,6]. The predictors for menopause, pulse rate, diastolic and systolic blood pressure, and education years have not been previously published, but were generated using the same methods as described in[5,6]. Here, we summarize how the predictors were trained, and refer the reader to the indicated manuscripts for complete details.

The data set used for training was taken from the 2018 release of genotyping and imputation data for all 500,000 participants in the UK Biobank study[39]. Only genetically British individuals, as defined by the UK Biobank in[40], were included in our training set. (Briefly speaking, the UK Biobank defines "genetically British" by clustering the genomes in principal component space). The UK Biobank reports data for case-control phenotypes via a combination of self-reported data and various ICD-x diagnosis codes, and for continuous phenotypes via physical measurements.

To build predictors based on a simple linear model of genetic predisposition, i.e. $\vec{Y} = \bar{X} \cdot \vec{\beta} + \vec{\epsilon}$, we use both a custom implementation of L1-penalized LASSO as well as the Python scikit-learn version. This approach is derived from compressed sensing research[41–44]. The result is a *sparse* active set of SNPs with weights derived from minimizing the objective function:

$$\mathcal{O}_{\lambda}(\vec{\beta}) = \frac{1}{2}||\vec{Y} - \bar{X} \cdot \vec{\beta}||_{L2}^{2} + n\lambda|\beta|_{L1} \rightarrow \vec{\beta}^{*} = argmin\Big[\mathcal{O}_{\lambda}(\vec{\beta})\Big] \tag{1}$$

where λ is a hyper-parameter for the minimization algorithm. The penalization term corrects for some amount of Linkage Disequilibrium (LD): if a SNP is given a non-zero β value by the algorithm, the minimization will disfavor choosing other SNPs in high LD with the first SNP to be in the active set. Repeating the procedure (e.g. when performing cross-validation) will yield slightly different active SNP sets, but with similar separation of cases and controls or similar correlations.

The relationship between the resulting polygenic score $\bar{X} \cdot \vec{\beta}^{*}$ and estimated risk of disease is determined empirically using a validation dataset. We find that the estimated risk varies non-linearly with $\bar{X} \cdot \vec{\beta}^{*}$—rising and falling rapidly for outliers in the population distribution for $\bar{X} \cdot \vec{\beta}^{*}$.

Note that because of the minimization procedure in Eq. (1), the final scores, $\bar{X} \cdot \vec{\beta}^{*}$, will not necessarily be of order 1. As a matter of fact, the typical scores for case-control conditions are of maximum order 10^{-3}. Because of this, the sum of the variances of the scores from each of our predictors is not obviously directly related to estimates of SNP-based heritability for that phenotype, e.g. as estimated using REML methods. For reference, in Supplementary Table S60 in Appendix D, we provide a list of SNP heritabilities for most of the phenotypes studied here, which were derived by other groups via REML methods.

Variance and effect sizes. The majority of this work deals with characterizing the relative sizes of the effect of particular SNPs on the performance of a polygenic predictor. A predictor in this case is a set of weights, $\{\beta_i\}$, for a set of SNPs, S. An individual's phenotype, y, can be described by

$$y = \tilde{y} + \epsilon = \sum_{i \in S} x_i \beta_i + \epsilon \,, \tag{2}$$

where $x_i \in \{0, 1, 2\}$ is the number of minor alleles for SNP i, $\tilde{y}$ is the predicted value of the phenotype, and ϵ is an error term. Our primary object of interest is the variance of this prediction.

The contribution of a single SNP to this variance is expressed in terms of the β_i and the minor allele frequency (MAF), f_i, as:

$$\mathrm{var}(x_i \beta_i) = \beta_i^2 \mathrm{var}(x_i) = \beta_i^2 2(1 - f_i) f_i \,. \tag{3}$$

In the limit of small MAF, this becomes $2\beta_i^2 f_i + o(f_i^2) \approx 2\beta_i^2 f_i$. The overall variance of the individual's predictor score (for continuous traits, this is a predicted phenotype) can thus be described as:

$$\begin{aligned} \mathrm{var}(\tilde{y}) = \mathrm{var}\left(\sum_{i \in S} x_i \beta_i\right) &= \sum_{i \in S}\left(\mathrm{var}(x_i \beta_i) + 2\sum_{j<i} \mathrm{cov}(\mathrm{x_i}\beta_\mathrm{i}, \mathrm{x_j}\beta_\mathrm{j})\right) \\ &= \sum_{i \in S}\left(2\beta_i^2(1 - f_i) f_i + 2\sum_{j<i} \mathrm{cov}(\mathrm{x_i}\beta_\mathrm{i}, \mathrm{x_j}\beta_\mathrm{j})\right) \approx \sum_{i \in S} 2\beta_i^2 (1 - f_i) f_i \,, \end{aligned} \tag{4}$$

where the final approximation holds when the SNPs are largely uncorrelated. This is true for most minor allele frequencies, and has been checked empirically in particular instances (see for example[6]). For the predictors in this work, it is unusual for SNPs more than 2,000 kilo base pairs apart to have correlation higher than 0.01 or so. L1 penalization generally results in mostly uncorrelated SNPs in the predictor. In this sense, the variance due to each SNP can be considered as a linear effect. Supplementary Table S59 in Appendix C lists the total variance accounted for by all predictor SNPs in the active set for every one of our twenty-one disease conditions.

We can then calculate the portion of predicted variance accounted for by any subset, $\mathscr{S} \subset S$, of the active SNPs, as a fraction of the total predictor variance of the phenotype:

$$\frac{\mathrm{var}(\tilde{y}')}{\mathrm{var}(\tilde{y})} = \frac{\mathrm{var}\left(\sum_{i \in \mathscr{S}} x_i \beta_i\right)}{\mathrm{var}\left(\sum_{j \in S} x_j \beta_j\right)} \approx \frac{\sum_{i \in \mathscr{S}} 2\beta_i^2 (1 - f_i) f_i}{\sum_{j \in S} 2\beta_j^2 (1 - f_j) f_j} \,, \tag{5}$$

where again, the final approximation holds when the SNPs are uncorrelated.

Effect of predictor SNPs located in genic regions. A SNP may be regarded as being within a genic region if its genomic coordinates fall between the start-point and end-point coordinates of any protein-coding gene. The GENCODE Release 19 annotation of the human genome[45] (based on the GRCh37.p13 reference human genome assembly[46]) was chosen as the source of our reference set of gene boundary coordinates. Currently, it is still unclear where exactly genic regions end and intergenic regions begin[47–49], and so there is a possibility that this choice of gene boundary coordinates may not be definitive for our purposes. This motivated us to analyze (for our selection of disease conditions and traits) how the number of the predictor SNPs categorized as located in genic regions and the variance accounted for by these predictor SNPs located in genic regions changes as all gene boundaries (according to GENCODE Release 19) are expanded by an increasing number k of kilo base pairs at both ends.

We then investigated how this genic variance is distributed between individual (protein-coding) genes. When considering a specific disease condition, for each gene, the variance accounted for by all predictor SNPs located within the (effective) gene boundary coordinates is summed and expressed as a percentage of the total variance accounted for by all predictor SNPs for that condition. For the purposes of this calculation, the boundaries of the genic regions were chosen to be at $k = 30$. Since each genic predictor SNP may lie within the boundaries of more than one gene due to the expanded gene boundaries overlapping, multiple genes may share the exact same set of predictor SNPs and therefore the same value of total variance accounted for by single genes. We also identified the predictor SNPs located within the high-variance genes named above, together with the values of variance accounted for by each SNP, and the particular gene(s) with which the SNP is associated.

Limits of exome sequencing data when probing disease conditions and traits. Modern whole-exome sequencing techniques are expected to be able to access about 85% of known disease-related variants[50,51]. Assuming this is correct, we would expect about the same fraction of genic SNPs belonging to our predictors for disease conditions / traits to be identifiable via exome-sequencing data. To verify this, we compared our sets of predictor SNPs against the whole-exome sequencing data released by the UK Biobank in March 2019[52] (to be specific, the version of the whole-exome sequencing data set generated using a Functionally Equivalent (FE) pipeline[53]). Once again, genic SNPs were defined as those SNPs located within the GENCODE Release 19 protein-coding gene boundaries extended by 30 kilo base pairs at both ends (or $k = 30$). We calculated for each disease condition the percentage of predictor SNPs located in genic regions which are also found in the UK Biobank exome data, as well as the percentage of variance accounted for by these genic predictor SNPs which are also found in the UK Biobank exome data.

Pairwise comparison of predictors. To compare pairs of predictors we use two methods: (1) overlap of SNPs and (2) overlap of variance accounted for. For (1), a pair of predictors can be compared by identifying SNPs which the predictors have in common. Here, pairs of SNPs less than 4,000 base pairs apart were considered to be essentially the same SNP (where this separation was chosen so that most or all SNP pairs with high linkage disequilibrium levels are expected to be identified with one another[54]). For (2) we define a new correlation measure: for each SNP, *i*, in the first predictor (corresponding to the condition labelling the row in Tables 1 and 2), all SNPs, *j*, from the second predictor (corresponding to the condition labelling the column in Tables 1 and 2) located less than 4,000 base pairs away were identified. Every such associated SNP from the second predictor was assigned a weight of uniform magnitude, with the sign of the weight based on the sign of the SNP's effect size relative to the sign of the effect size of the SNP from the first predictor—positive when the signs were the same, and negative when the signs were different. These weights were then multiplied by the variance due to the SNP from the first predictor and summed. If we label each set of SNPs within 4,000 base pairs away as $\mathcal{C}_i$, this correlation estimate, $\tilde{r}$, can be expressed as

$$\tilde{r} \approx \lambda \sum_i \sum_{j \in \mathcal{C}_i} \mathrm{sgn}\left(\beta_\mathrm{i}^{(1)} \beta_\mathrm{j}^{(2)}\right) 2(\beta_\mathrm{i}^{(1)})^2 (1 - \mathrm{f_i})\mathrm{f_i}\,, \tag{6}$$

where the normalization is chosen to be the total predictor variance of the row predictor,

$$\lambda = \left(\sum_i 2(\beta_i^{(1)})^2 (1 - f_i) f_i \right)^{-1} .$$

This produced a weighted overlap in terms of variance, that accounts for whether the associated SNP pairs are positively correlated (effect sizes have equal signs) or negatively correlated (effect sizes have differing signs).

Note: In this portion of the analysis, the coronary artery disease predictor was restricted to the top twenty thousand predictor SNPs as ranked by value of variance accounted for.

Statement on methods. No human participants were directly involved in this study. Details of the data collected by the UK Biobank are outlined in the Supplementary Methods section (Appendix C) of the Supplementary Information.

Results and analysis

Predictor SNPs in genic regions. We are interested in investigating how predictor SNPs located in genic regions impact our predictors.

The most obvious way to identify predictor SNPs located inside genic regions is to define a SNP as being within a genic region if its genomic coordinates fall between the start-point and end-point coordinates of any protein-coding gene. The GENCODE Release 19 annotation of the human genome[45] (based on the GRCh37.p13 reference human genome assembly[46]) was chosen as the source of our reference set of gene boundary coordinates. Currently, it is still unclear where exactly genic regions end and intergenic regions begin[47–49], and so there is a possibility that this choice of gene boundary coordinates may not be definitive for our purposes. We asked these questions: as these reference gene boundary coordinates are varied, by how much does the separation of predictor SNPs into genic and non-genic categories vary, and how significantly does the influence of the (increasingly large) genic section of the predictor SNPs change?

Figure 1 shows for a selection of disease conditions how the number of the predictor SNPs categorized as located in genic regions—expressed as a percentage of the total number of predictor SNPs for that disease condition—changes as all gene boundaries (according to GENCODE Release 19) are expanded by an increasing number k of kilo base pairs at both ends. At the reference gene boundaries ($k = 0$), the percentage of predictor SNPs which are genic ranges from about 50% (many disease conditions) to about 60% (gallstones, malignant melanoma, atrial fibrillation); while at $k = 30$, this percentage rises to between 60% (breast cancer, type-1 diabetes, education years) to 75% (gallstones, malignant melanoma). This description excludes the coronary artery disease predictor (42.5–55%), which has a distinctly low proportion of genic predictor SNPs.

For all disease conditions that we consider here, the increase in the genic percentage of predictor SNPs with k between $k = 0$ and $k = 30$ occurs at roughly the same rate and appears to be almost linear. This implies that directly outside the reference gene boundary coordinates, the predictor SNPs for these twenty-one disease conditions are approximately uniformly distributed by distance for up to 30 kilo bp from the reference boundaries.

We mention that the coronary artery disease predictor shows distinctly separate behavior when compared to the other conditions in Fig. 1, probably due to the different method used in its construction, as detailed in[7]. This predictor was also trained on the UK Biobank, but used an implementation of the LDPred algorithm on a collection of associated phenotypes. This predictor involves 6,630,150 active SNPs, which is *orders of magnitude* larger than for the other predictors mentioned in this work (which typically contain a few thousand active SNPs).

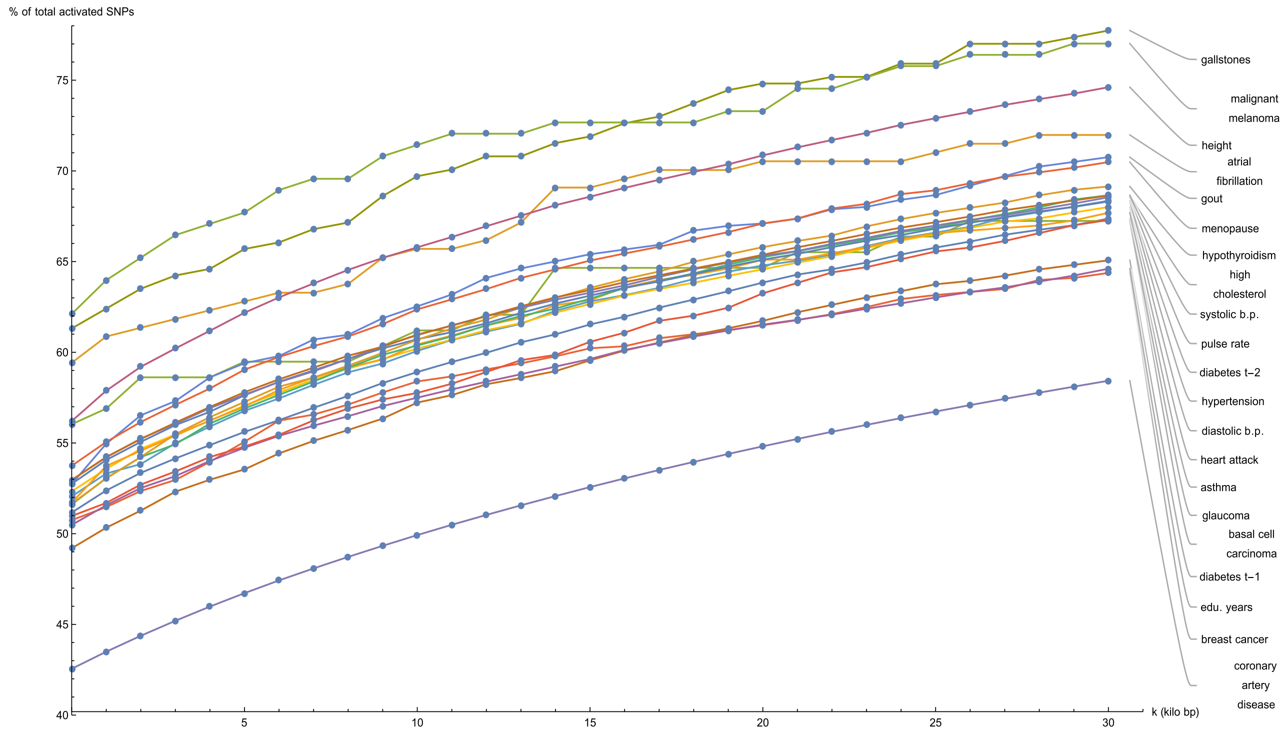

Figure 1. Plots of the number of predictor SNPs located within genic regions, expressed as a percentage of the total number of predictor SNPs for that disease condition, against expansion of GENCODE Release 19 gene boundaries by k kilo base pairs.

Because LDPred takes LD into account, it is quite likely that this predictor involves a disproportionate amount of intergenic SNPs.

For contrast, the situation where all activated predictor SNPs are replaced with SNPs randomly selected from the 800,000+ variants measured by the UK Biobank Axiom Array is examined. Supplementary Figure S1 in Appendix A displays the change in the number of randomly selected SNPs located within genic regions as the reference gene boundaries are expanded, expressed as a percentage of the total number of SNPs in randomly-selected sets, where each set is of size equal to the activated set in the predictor for the disease condition labelling the plot. The CAD predictor was excluded from this comparison since its active set included more than 6.6 million SNPs, most of which do not overlap with the UK Biobank array. It can be seen that Supplementary Figure S1 is mostly similar to Fig. 1. Once again, the growth of the number of genic SNPs with increasing distance from the reference gene boundaries is roughly linear, and the rate of growth does not appear to depend on the size of the set of random SNPs used. At ($k = 0$), the percentage of predictor SNPs which are genic ranges from about 50% to about 55%; while at $k = 30$, this percentage rises to between 60 and 70%—meaning a small overall decrease of genic SNPs for a few individual plots, by about 5%, when compared to Fig. 1.

Figure 2 shows for each disease condition how the variance accounted for by the predictor SNPs located in genic regions—expressed as a percentage of the total variance accounted for by all predictor SNPs for that condition—changes as the boundaries of every gene (according to GENCODE Release 19) are expanded by k kilo base pairs at both ends. At the reference gene boundaries ($k = 0$), the percentage of predictor variance accounted for by SNPs in genic regions ranges from about 40% (breast cancer, type-1 diabetes) to 90% (gallstones, gout, malignant melanoma), with notable outliers at 25% (atrial fibrillation) and 20% (coronary artery disease). For the majority of disease conditions, the percentage of variance accounted for by the genic predictor SNPs remains approximately flat as k is increased, meaning that practically all the variance accounted for by genic predictor SNPs is due to SNPs contained within the reference gene boundaries. Noticeable exceptions occur for glaucoma at $k = 6.5$, breast cancer at $k = 17.5$, and menopause at $k = 27.5$, where the observed large jumps in variance indicate the presence of some SNP(s) at that genomic location with a significant effect on that specific disease condition.

Now that the extent of the variance accounted for by SNPs located within genic regions has been established for each disease condition, it is natural to investigate next how this genic variance is distributed between individual (protein-coding) genes. When considering a specific disease condition, for each gene, the variance accounted for by all predictor SNPs located within the (effective) gene boundary coordinates is summed and expressed as a percentage of the total variance accounted for by all predictor SNPs for that condition. For the purposes of this calculation, the boundaries of the genic regions were chosen to be at $k = 30$.

Supplementary Figures S2 through S22 in Appendix A show the percentage of predictor variance accounted for by single genes for asthma, atrial fibrillation, basal cell carcinoma, breast cancer, coronary artery disease, type-1 diabetes, type-2 diabetes, diastolic blood pressure, educational years, gallstones, glaucoma, gout, heart attack, height, high cholesterol, hypertension, hypothyroidism, malignant melanoma, menopause, pulse rate, and

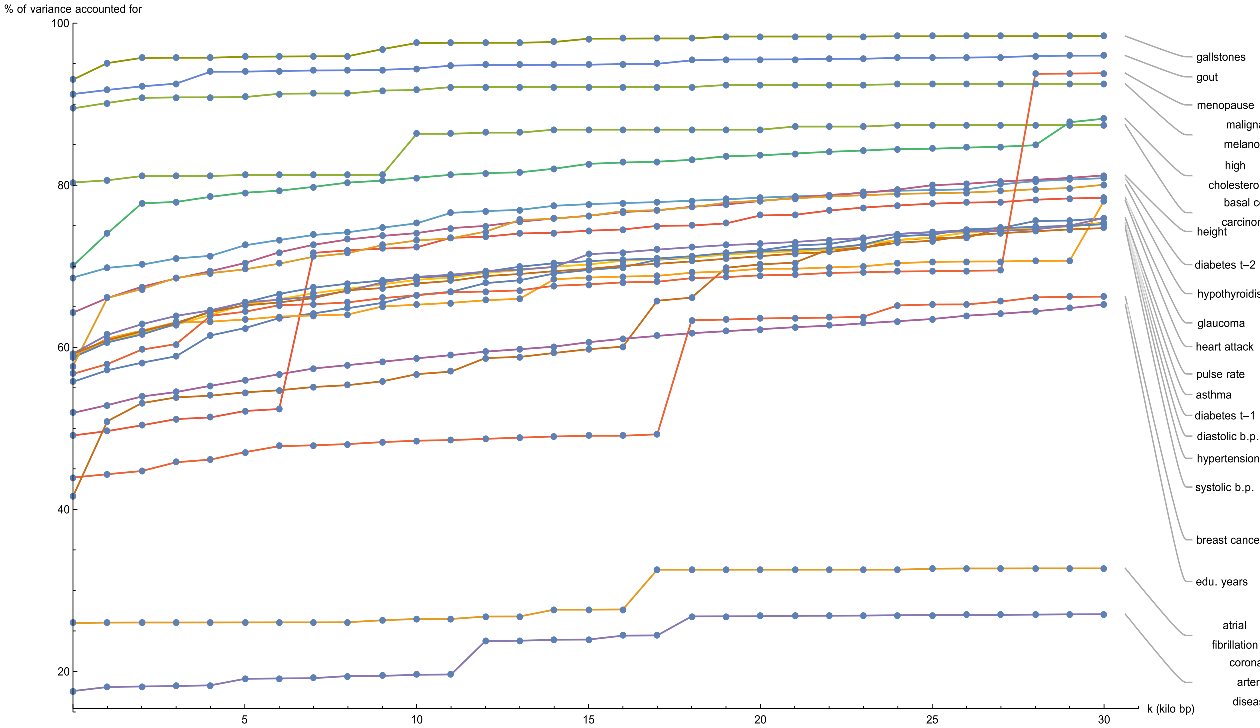

Figure 2. Plots of the variance accounted for by predictor SNPs located within genic regions, expressed as a percentage of the total variance accounted for by all predictor SNPs for that disease condition, against expansion of GENCODE Release 19 gene boundaries by *k* kilo base pairs.

systolic blood pressure. Only the fifteen largest values of variance accounted for by a single gene are displayed for each condition. As each genic predictor SNP may lie within the boundaries of more than one gene due to the expanded gene boundaries overlapping, multiple genes may share the exact same set of predictor SNPs and therefore the same value of total variance accounted for by single genes. For every disease condition considered here, the full list of genes responsible for the top fifteen values of variance displayed here can be found in Supplementary Tables S23 through S43 in Appendix B.

It is evident that certain disease conditions have one (or perhaps two or three) dominating value(s) of variance accounted for by a single gene. The most striking results are found with basal cell carcinoma (one gene, IRF4, supplying 28% of the predictor variance out of the 87% total genic variance at $k = 30$ and a second, TGM3, supplying 12% of the predictor variance), breast cancer (two genes, FGFR2 and TOX3, each supplying 14–15% of the predictor variance out of 66% total genic variance), type-2 diabetes (one gene, TCF7L2, supplying 25% of the predictor variance out of 81% total genic variance), gallstones (three genes, ABCG8, ABCG5 and DYNC2LI1, each supplying 80–83% of the predictor variance out of 98% total genic variance), glaucoma (two genes, ALDH9A1 and TMCO1, each supplying 19% of the predictor variance out of 78% total genic variance), gout (one gene, ABCG2, supplying 57% of the predictor variance out of 96% total genic variance and another, SLC2A9, supplying 12% of the predictor variance), heart attack (one gene, LPA, supplying 21% of the predictor variance out of 78% total genic variance), malignant melanoma (five genes, AC092143.1, TUBB3, TCF25, MC1R and DEF8, each supplying 37–43% of the predictor variance out of 92% total genic variance), and menopause (one gene, UTY, supplying 24% of the predictor variance out of 94% total genic variance).

The more novel among these results are the strong links observed between glaucoma and ALDH9A1, between gallstones and DYNC2LI1, between menopause and UTY, and between malignant melanoma and the four genes AC092143.1, TUBB3, TCF25, and DEF8. These relationships are either not well-established up until now, or have not yet been suggested to be possible—therefore we take a closer look at the specific genetic variants involved. Supplementary Tables S44 through S47 list the predictor SNPs located within these novel genes, together with the amount of variance accounted for by each SNP. Clearly, for each disease condition, very few (< 10) SNPs are involved and just one SNP accounts for most of the variance: Affx-20090007 (on DYNC2LI1) with 77% of the predictor variance, rs4656461 (on ALDH9A1) with 19%, Affx-35293625 (on AC092143.1, TUBB3, TCF25, and DEF8) with 34%, and rs1236440 (on UTY) with 24%.

The rest of our findings confirm previous work by other groups. The association of IRF4[55] and TGM3[56] with basal cell carcinoma is well-known, as is the relationship of FGFR2[57] and TOX3[58] to breast cancer, the link between TCF7L2[59] and type-2 diabetes, the association of ABCG8[60] and ABCG5[61] with gallstones, the role of TMCO1[62] in glaucoma, the relationship of ABCG2[63,64] and SLC2A9[65] to gout, the connection between LPA[66] and heart attack/coronary artery disease, and the role of MC1R[67,68] in malignant melanoma.

For reference, Supplementary Tables S48 through S56 in Appendix B name all predictor SNPs located within the high-variance genes named above, together with the values of variance accounted for by each SNP, and the

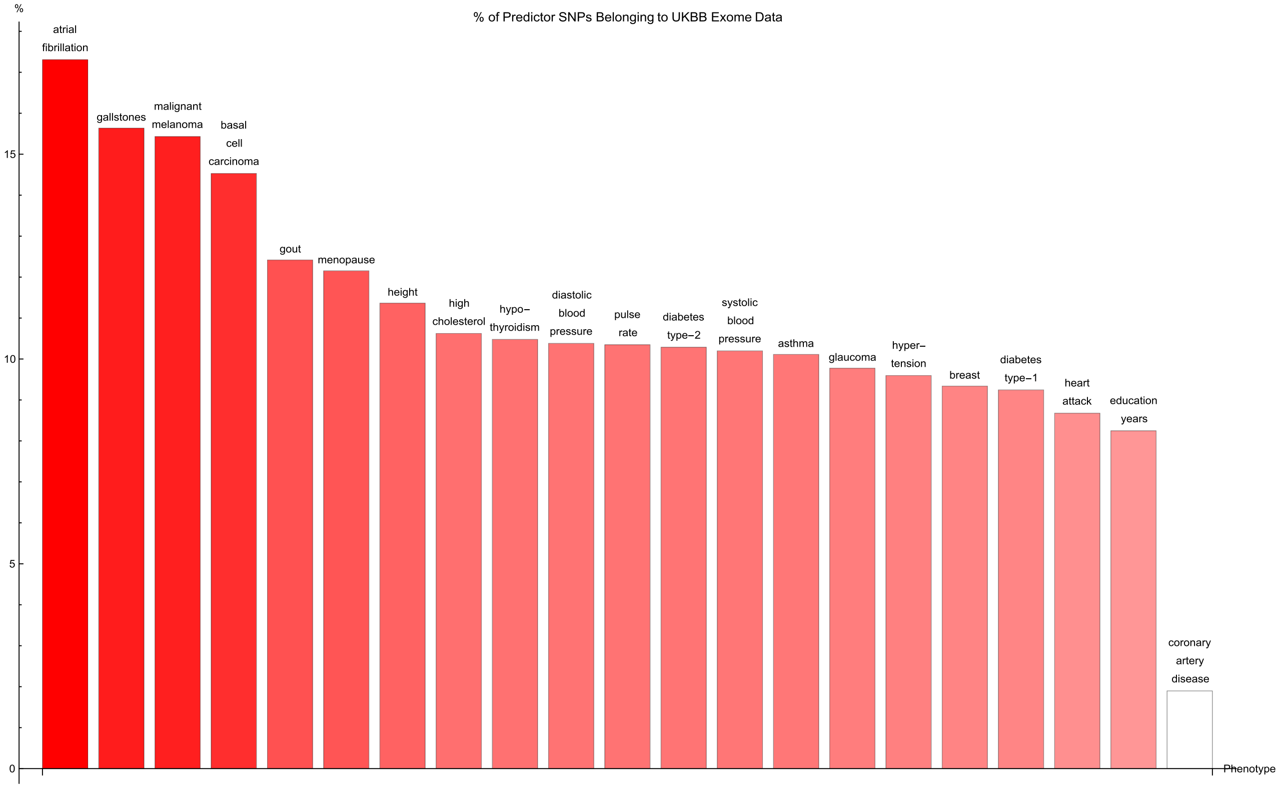

Figure 3. The percentage of predictor SNPs which are found both in genic regions and the UK Biobank exome data, for each disease condition. The disease conditions are listed from left to right on the horizontal axis in order of decreasing percentage. Each vertical bar is colored red with a depth of shade proportional to the height of the bar. Here, "genic" SNPs are contained within the GENCODE Release 19 gene boundaries plus 30 kilo base pairs at both ends.

particular gene(s) with which the SNP is associated. Once again, for each disease condition, we see that just one or two SNPs is responsible for the lion's share of the variance carried by each high-variance gene. To be specific: the basal cell carcinoma predictor has rs12203592 (on IRF4) carrying 28% of the predictor variance and rs214803 (on TGM3) with 12%, the breast cancer predictor has rs2981575 (on FGFR2) and rs4784227 (on TOX3) each with 14% of the predictor variance, the type-2 diabetes predictor has rs7903146 (on TCF7L2) with 25% of the predictor variance, the gallstones predictor has Affx-20090007 (on ABCG8, ABCG5, and DYNC2LI1 simultaneously) with 77% of the predictor variance, the glaucoma predictor has rs4656461 (on ALDH9A1 and TMCO1) with 19% of the predictor variance, the gout predictor has rs2231142 (on ABCG2) with 55% of the predictor variance and rs3775948 (on SLC2A9) with 8%, the heart attack predictor has rs10455872 (on LPA) with 14% of the predictor variance and rs117733303 (on LPA) with 7%, the malignant melanoma predictor has Affx-35293625 (on AC092143.1, TUBB3, TCF25, MC1R and DEF8) with 34% of the predictor variance and rs11538871 (on AC092143.1, TUBB3, TCF25, and MC1R) with 6%, and the menopause predictor has rs1236440 (on UTY) with 24% of the predictor variance.

Overlap between predictor SNPs and whole-exome sequencing data. Modern whole-exome sequencing techniques are expected to be able to access about 85% of known disease-related variants[50,51]. Assuming this is correct, we would expect about the same fraction of genic SNPs belonging to our predictors for disease conditions to be identifiable via exome-sequencing data. To verify this, we compared our sets of predictor SNPs against the whole-exome sequencing data released by the UK Biobank in March 2019[52] (to be specific, the version of the whole-exome sequencing data set generated using a Functionally Equivalent (FE) pipeline[53]). In this section, once again, genic SNPs are defined as those SNPs located within the GENCODE Release 19 protein-coding gene boundaries extended by 30 kilo base pairs at both ends (or $k = 30$).

Figure 3 shows for each disease condition the percentage of predictor SNPs located in genic regions which are also found in the UK Biobank exome data, where the disease conditions are listed from left to right on the horizontal axis in order of decreasing percentage. For about two-thirds of the conditions surveyed, 10% or so of the genic SNPs for each predictor also formed part of the set of SNPs identified via exome-sequencing. The remaining disease conditions have up to about 17% (in terms of numbers) of their genic predictor SNPs detected by the exome-sequencing data set—with one exception, coronary artery disease, which displays an extraordinarily low (2%) value of this overlap.

Figure 4 shows for each disease condition the variance accounted for by the genic predictor SNPs from Fig. 3, expressed as a percentage of the total variance accounted for by all the SNPs in the predictor, where the disease

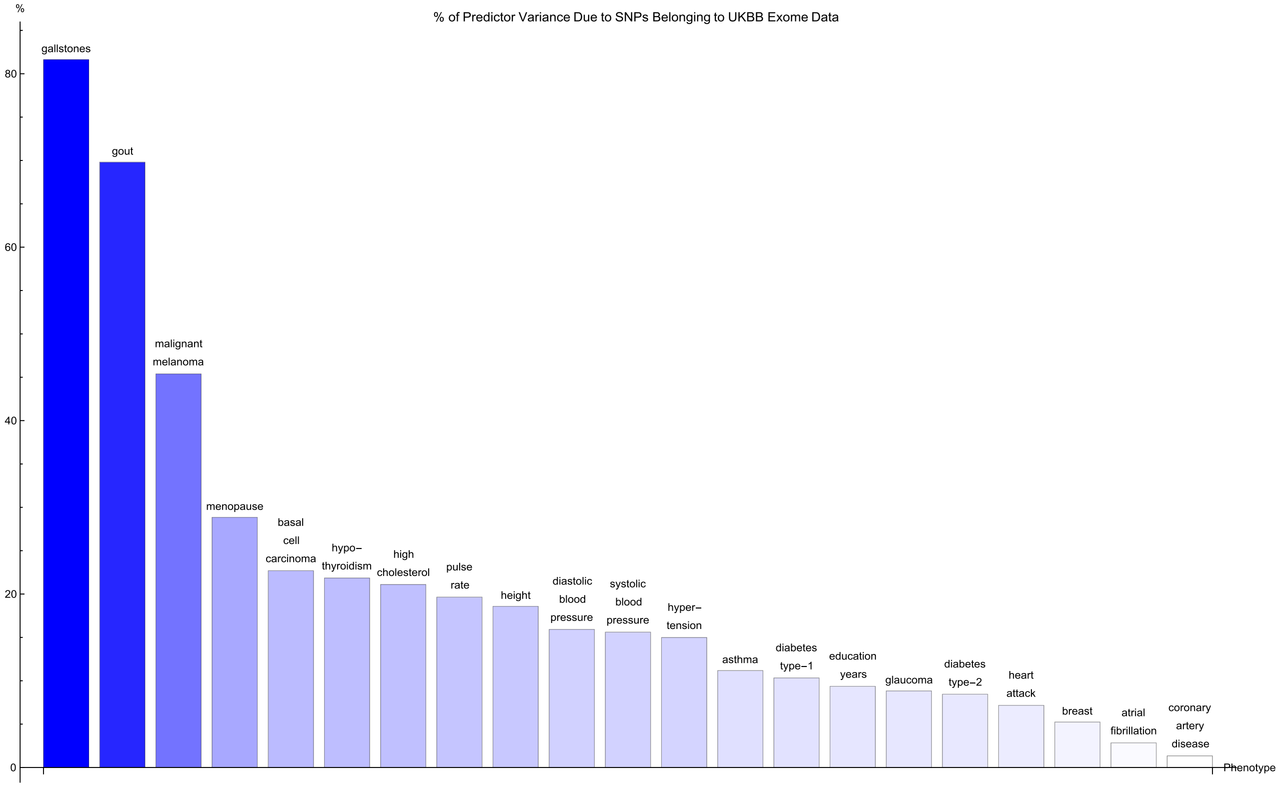

Figure 4. The variance accounted for by predictor SNPs which are both in genic regions and detected by the UK Biobank exome data, as a percentage of the total variance accounted for by all predictor SNPs, for each disease condition. The disease conditions are listed from left to right on the horizontal axis in order of decreasing percentage. Each vertical bar is colored blue with a depth of shade proportional to the height of the bar. Here, "genic" SNPs are contained within the GENCODE Release 19 gene boundaries plus 30 kilo base pairs at both ends.

conditions are listed from left to right on the horizontal axis in order of decreasing percentage. For the majority of conditions, 20% or less of the total variance accounted for by the predictor SNPs comes from genic SNPs which show up in the UK Biobank exome data. In fact, less than 5% of the variance accounted for by the the breast cancer, atrial fibrillation, and coronary artery disease predictor SNPs is detected by the exome data. Exceptions to this are the predictors for gallstones (80% of predictor variance detected by the exome data), gout (70% of predictor variance detected by the exome data), malignant melanoma (45% of predictor variance identified by the exome data), and menopause (30% of predictor variance identified by the exome data).

It may be worth noting that the ordering of the predictors by percentage of predictor SNPs which are genic (in Fig. 1) and the ordering according to percentage of predictor SNPs which are both genic and accessible via the exome data in (Fig. 3) is about the same. A similar observation can be made regarding Figs. 2 and 4 (here in terms of variance accounted for).

In Fig. 5, an overall view of the extent to which predictor SNPs in genic and non-genic regions are identifiable based on exome-sequencing data is given. Figure 5 shows for each disease condition the breakdown of the percentage of predictor SNPs according to whether their location is genic and whether the exome data serves to probe them. This breakdown does not vary much between predictors: in general about 10–17% of predictor SNPs are both in genic regions and found in the exome data ('Genic/Exonic', colored blue), 0% (as might be expected) are not in genic regions but are present in the exome data ('Non-genic/Exonic', colored yellow), 55–65% are located in genic regions but not found in the exome data ('Genic/Non-exonic', colored green), and 20–35% are neither in genic regions nor in the exome data ('Non-genic/Non-exonic', colored red). The only deviation comes from the coronary artery disease predictor SNPs—less than 5% are 'Genic/Exonic', while nearly 45% are 'Non-genic/Non-exonic'.

Pairwise comparison of predictors.

We now focus on finding connections between disease conditions based upon similarities in their predictors. In this section, the analysis involving the coronary artery disease predictor was restricted to the top twenty thousand predictor SNPs as ranked by value of variance accounted for.

Table 1 shows the pairwise overlap between disease conditions in terms of the number of predictor SNPs that each pair of conditions has in common. Here, pairs of SNPs less than 4,000 base pairs apart were considered to be essentially the same SNP (where this separation was chosen so that most or all SNP pairs with high linkage disequilibrium levels are expected to be identified with one another[54]). Each row of the table corresponds to a particular disease condition, and reading the row from left to right (going from one column to the next) gives

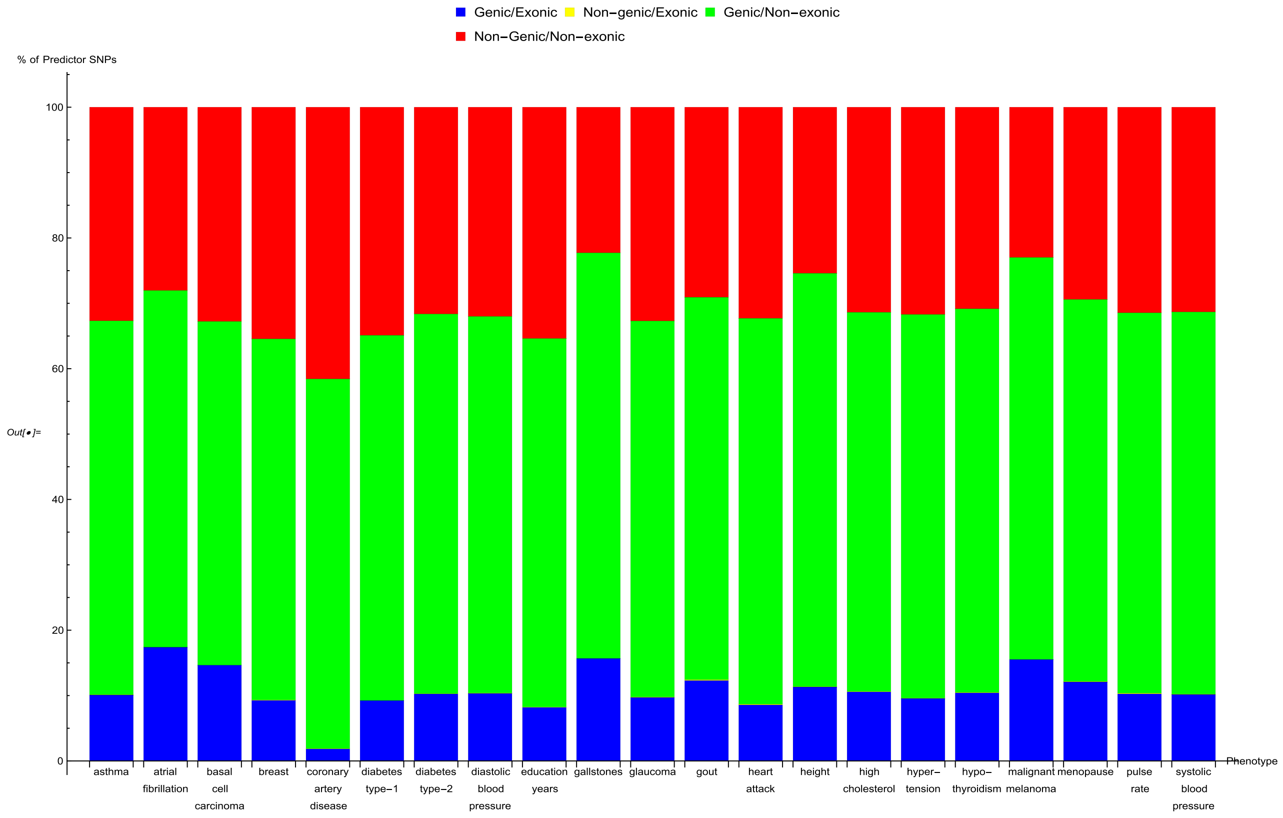

Figure 5. The breakdown of the percentage of predictor SNPs according to whether their location is genic and whether the exome data serves to probe them, for each predictor. The bar sections representing predictor SNPs in genic regions and in the exome data are labelled 'Genic/Exonic' and colored blue, those representing predictor SNPs not in genic regions but present in the exome data are 'Non-genic/Exonic' and colored yellow, those representing predictor SNPs which are located in genic regions but not found in the exome data are 'Genic/Non-exonic' and colored green, and those representing predictor SNPs neither in genic regions nor in the exome data are 'Non-genic/Non-exonic' and colored red. As expected, the yellow 'Non-genic/Exonic' bar sections are too small to be discernible. Here, "genic" SNPs are contained within the GENCODE Release 19 gene boundaries plus 30 kilo base pairs at both ends.

the number of predictor SNPs (expressed as a percentage of the total number of SNPs in the row-label predictor) that the conditions labelling each column share with the condition that the row corresponds to.

A pair of conditions may be considered to have a significant connection if the percentage of SNPs in common is substantial when read off the table both ways. Analyzing the table in this manner produces two notable groupings, with all the conditions in each group having large pairwise overlap.

1. Asthma–diastolic blood pressure–hypertension–systolic blood pressure–education years–height: The first four disease conditions form a combination which is not too surprising, but the same cannot be said about the last two traits. The table shows that all possible pairs taken from these six conditions overlap by about 10% or so, with the following exceptions which exceed this level by some way: 38% of the systolic blood pressure SNPs also belong to the diastolic blood pressure predictor, while 35% of the diastolic blood pressure SNPs also belong to the systolic blood pressure predictor; 17% of the diastolic blood pressure SNPs belong to the hypertension predictor, while 18% of the hypertension SNPs belong to the diastolic blood pressure predictor; and 18% of the hypertension SNPs belong to the systolic blood pressure predictor, while 19% of the systolic blood pressure SNPs belong to the hypertension predictor.

2. Basal cell carcinoma–malignant melanoma: 7.8% of the basal cell carcinoma SNPs belong to the malignant melanoma predictor, while 8.1% of the malignant melanoma SNPs belong to the basal cell carcinoma predictor.

Next, variance accounted for is used to measure the overlap according to Eq. (6), and can be seen in Table 2. This results in the discovery of more relations between different predictors. Table 2 displays this overlap as a percentage, and is basically Table 1 expressed in terms of variance (weighted according to correlation sign) accounted for by the overlapping predictor SNPs.

Once again, significant connections are observed in the case of diastolic blood pressure–hypertension–pulse rate–systolic blood pressure and basal cell carcinoma–malignant melanoma. In the case of the first group of conditions, the overlap of weighted variance ranges from 8% (systolic blood pressure–pulse rate) to 51% (systolic blood pressure–diastolic blood pressure). In the second case, 10% of the basal cell carcinoma predictor variance is shared with the malignant melanoma predictor, while 58% of the malignant melanoma predictor variance belongs to the basal cell carcinoma predictor.

	asthma	atrial fibrillation	basal cell carcinoma	breast cancer	CAD 20 *k*	type 1 diabetes	type 2 diabetes	diastolic blood pressure	education years	gallstones	glaucoma	gout	heart attack	height	high cholesterol	hypertension	hypothyroidi˙. sm	malignant melanoma	menopause	pulse rate	systolic blood pressure
asthma	100	0.11	0.09	1.1	3.3	2	2.2	10	11	0.19	0.8	0.31	1.1	14	3.9	10	4.3	0.097	3.5	10	9.6
atrial fibrillation	7.2	100	0	0.97	2.9	0	1.4	14	12	0	0.97	0.97	0.48	14	3.4	13	3.4	0	3.4	14	11
basal cell carcinoma	10	0	100	2.6	1.7	2.6	2.6	14	13	0	4.3	0	0.86	21	3.4	15	6.9	7.8	3.4	8.6	13
breast cancer	7.6	0.11	0.17	100	3.6	1.9	2.3	9.8	9.8	0.11	0.89	0.39	0.67	13	3.8	9.7	3.6	0.11	3.4	8.7	8.7
CAD 20 *k*	5.9	0.1	0.055	1	100	1.6	2.6	10	6.8	0.25	0.54	0.57	3.6	13	5.2	9.3	2.5	0.065	2.7	8.5	9.7
type 1 diabetes	8.3	0	0.13	1.1	3.6	100	3.8	10	9.6	0.19	1	0.41	0.82	13	4.1	9.8	4.3	0.16	3.4	9.6	9.6
type 2 diabetes	8.2	0.086	0.086	1.3	4.5	3.4	100	12	13	0.35	0.98	0.69	1.3	16	6.1	15	3.9	0.029	3.6	12	12
diastolic blood pressure	7.7	0.18	0.11	1	4	1.9	2.5	100	11	0.14	0.88	0.44	1	15	4.1	17	3.4	0.13	3.7	14	35
education years	7.5	0.12	0.082	0.96	2.8	1.7	2.4	10	100	0.12	0.74	0.42	0.91	13	3.7	10	3.2	0.092	3.5	9.5	9.9
gallstones	8.4	0	0	0.73	6.6	2.2	4.4	8	8	100	1.1	1.1	1.5	13	6.6	11	2.9	0	4	11	9.1
glaucoma	7.5	0.14	0.36	1.2	3.2	2.2	2.4	11	10	0.22	100	0.65	0.72	15	3.1	10	3.4	0.072	3.9	10	11
gout	5.4	0.26	0	0.92	3.5	1.7	2.9	10	10	0.39	1.3	100	0.65	16	3.3	12	4.7	0.13	5.8	9.9	11
heart attack	9.5	0.07	0.07	0.77	10	1.9	3.2	13	12	0.28	0.7	0.35	100	15	8.2	14	4	0.07	3.6	10	12
height	8.1	0.14	0.13	1.1	3.9	1.9	2.5	12	11	0.17	0.99	0.57	0.98	100	4.3	12	4.1	0.15	4.2	11	11
high cholesterol	7.9	0.14	0.063	1.1	5.2	2	3.3	11	11	0.31	0.71	0.42	2	14	100	17	4.1	0.094	3.4	10	11
hypertension	7.7	0.16	0.1	1.1	3.9	1.8	3.1	18	11	0.18	0.84	0.52	1.2	15	6.3	100	3.5	0.079	3.5	11	18
hypothyroidi˙. sm	9.4	0.12	0.17	1.2	3.7	2.4	2.3	10	9.8	0.13	0.79	0.61	1.1	15	4.4	10	100	0.067	3.7	9.9	9.2
malignant melanoma	7.5	0	8.1	1.2	5	3.1	0.62	16	13	0	0.62	0.62	0.62	20	3.7	9.9	3.1	100	6.8	9.9	13
menopause	7.3	0.13	0.081	1.1	3.3	1.8	2	10	11	0.18	0.9	0.68	0.86	15	3.5	9.6	3.6	0.13	100	10	9.8
pulse rate	7.9	0.21	0.059	0.98	3.4	1.9	2.6	15	11	0.17	0.88	0.47	0.87	15	4	11	3.6	0.11	3.8	100	12
systolic blood pressure	7.6	0.17	0.11	1	4	1.9	2.6	38	11	0.16	0.89	0.49	1.1	15	4.4	19	3.3	0.13	3.7	12	100

Table 1. Table containing the pairwise overlap between predictors in terms of the number of predictor SNPs in common, expressed as a percentage of the total number of SNPs in the predictor for the condition labelling each row. Pairs of SNPs less than 4,000 base pairs apart were considered to be identified with one another. For a particular row, the number in each column represents the percentage of SNPs from the row predictor that can be identified with the SNPs from the column predictor.

We now also have groups of conditions which appear to be strongly positively correlated in terms of variance overlap, but were unremarkable in terms of number of SNPs in common. The most obvious is coronary artery disease–heart attack–high cholesterol–hypertension, where the magnitude of the pairwise overlap in terms of weighted variance ranges from 4% (hypertension–coronary artery disease) to 35% (high cholesterol–heart attack). Other single pairs of conditions with large overlapping variances are gout—high cholesterol (where 62% of the gout predictor variance also belongs to the high cholesterol predictor, while 35% of the high cholesterol predictor variance also belongs to the gout predictor), type-1 diabetes—type-2 diabetes (where 23% of the type-1 diabetes predictor variance also belongs to the type-2 diabetes predictor, while 30% of the type-2 diabetes predictor variance also belongs to the type-1 diabetes predictor), coronary artery disease—type-2 diabetes (4% and 25%), and gallstones—high cholesterol (85% and 4%). Also worth mentioning is the overlap of height with

	asthma	atrial fibrillation	basal cell carcinoma	breast cancer	CAD 20 *k*	type 1 diabetes	type 2 diabetes	diastolic blood pressure	education years	gallstones	glaucoma	gout	heart attack	height	high cholesterol	hypertension	hypothyroidism	malignant melanoma	menopause	pulse rate	systolic blood pressure
asthma	92	0.074	0.52	1.1	3.2	2	2.3	3.9	4.7	0.00024	0.53	0.18	2	11	3.3	5.5	6.7	−0.062	2.7	4.8	2.4
atrial fibrillation	1.7	100	0	0.11	2.4	0	−0.14	11	1.6	0	0.21	0.018	0.25	9.3	0.11	0.045	0.27	0	0.41	60	5
basal cell carcinoma	3.3	0	100	6.6	−0.08	−0.0053	−0.18	0.94	31	0	3.9	0	−2	43	1.7	3.1	31	10	−0.53	1.9	2.7
breast cancer	1	0.0021	0.048	100	0.48	0.62	−1.1	17	2.8	0.0022	0.43	−0.054	2.9	−3	−0.25	2.9	0.31	$-7. \times 10^{-6}$	1.5	2.7	16
CAD 20 *k*	2.4	0.0036	−0.04	−6.4	100	−0.063	3.6	−8.5	−0.096	−0.014	−0.099	0.16	15	−5.9	6.4	8.4	0.033	0.00028	0.15	0.18	6.7
type 1 diabetes	0.13	0	−0.36	−0.34	2.4	97	23	−0.3	−5.8	−0.032	0.087	0.12	−0.24	4.9	3.9	7.6	5.8	0.063	0.61	3.5	4.2
type 2 diabetes	2.2	−0.0046	−0.18	−0.18	25	30	99	3	1.7	0.091	0.086	1.1	0.59	7.7	32	−14	1.4	0.0002	1.9	31	4.4
diastolic blood pressure	3.7	−0.041	−0.27	0.56	2.2	0.95	1.7	93	5.6	0.24	0.42	0.77	3.8	13	5.6	32	2.7	0.076	1.7	11	50
education years	2.6	0.038	0.17	0.42	−0.15	0.66	1.1	4.8	94	0.032	0.053	0.14	0.35	8.1	1.6	4.6	1.7	0.028	1.6	5	5
gallstones	0.22	0	0	0.0049	78	−0.54	3.4	2.2	1.8	100	0.023	1.5	4.5	6.2	85	−80	−76	0	0.16	80	−0.051
glaucoma	−1.1	0.013	0.2	0.41	−0.64	0.93	0.4	2.1	5.1	−0.3	99	0.0014	2.6	13	1.4	20	1.3	0.0019	1.5	3.9	5.4
gout	58	−0.089	0	−0.29	1.1	5.4	0.72	5.4	3.6	3.2	0.072	100	0.39	−0.35	62	5.9	3	−0.0035	7.1	7.7	65
heart attack	3.7	0.00019	−0.044	0.12	23	3.5	−0.23	6.9	2.2	−0.1	0.85	0.91	99	13	32	25	2.4	−0.0058	1	12	12
height	4.1	0.026	0.022	0.85	1.5	0.35	2	7.8	6.7	0.3	0.65	0.92	0.45	83	2.2	7.8	2.5	−0.026	2.1	6.6	11
high cholesterol	3	−0.22	−0.012	−0.088	10	12	−0.32	−3.1	1.6	4.2	−0.2	7.4	35	15	89	14	2	−0.0021	3.5	20	5.6
hypertension	4.5	−0.041	−0.65	1.4	4	1.7	3.4	39	4.7	0.067	0.67	1.6	4.7	12	13	93	3	0.051	1.3	7.6	40
hypothyroidism	9.8	0.00089	1.3	−0.86	8.5	9.9	8.4	16	4.1	0.058	0.21	6.7	6.5	8.5	1.6	9.6	96	0.016	0.1	11	9.8
malignant melanoma	5.3	0	58	0.43	0.35	0.14	0.00059	−29	37	0	0.56	−0.017	−0.039	−1.1	−0.26	−34	0.83	100	34	0.65	1.9
menopause	3	0.014	−0.092	2.8	−0.09	0.4	0.66	2.8	3	0.014	0.1	−0.43	3	7	0.45	0.089	0.69	−0.01	94	2.6	14
pulse rate	3.1	0.6	0.064	0.4	0.24	1.2	1.9	29	4.7	0.34	−0.23	0.64	0.72	12	2.5	5.8	1.2	0.18	2.3	88	18
systolic blood pressure	3.5	0.0031	−0.35	0.7	2.9	1.3	0.26	51	5.3	0.31	0.77	0.77	2.9	11	8	37	2.4	0.014	1.8	7.7	91

Table 2. As described by Eq. (6), this table contains the pairwise overlap between predictors in terms of the variance accounted for by predictor SNPs in common, expressed as a percentage of the total variance accounted for by the SNPs belonging to the predictor for the condition labelling each row. The overlapping variance is weighted according to the sign of the correlation between each pair of associated SNPs (4,000 base pairs or less apart). This can cause diagonal elements to be less than one hundred percent, as anti-correlated SNPs may be included in the overlap calculation.

respectively: education years, asthma, diastolic blood pressure, hypertension, pulse rate, systolic blood pressure—which all are about 10%.

It may be asked whether similar results are obtained when pairwise genetic correlations are used to compare disease conditions instead. Supplementary Figure S62 in Appendix D shows previously published genetic correlations estimated using LD Score regression[69] for some of the UK Biobank phenotypes that we consider. Among the groupings of phenotypes with largest positive genetic correlations in this set of results, heart attack–high cholesterol–hypertension and diastolic blood pressure–hypertension–pulse rate–systolic blood pressure very obviously stand out, i.e., groupings that we also obtained. This lends support to our results for the overlap between disease conditions described above.

Conclusions

This paper explores the genetic architectures of a number of common disease conditions and complex traits, as revealed by the most important SNPs used in genomic predictors.

The results are complex—primarily summarized in the many figures and tables in the main text and Supplementary Information. However, we can make some general statements:

I. The fraction of SNPs in or near genic regions varies widely by phenotype. For example, in the case of Coronary Artery Disease and Atrial Fibrillation, less than 20-30 percent of the total risk variance is due to SNPs near genic regions.

II. For the majority of disease conditions studied, *most* of the variance is accounted for by SNPs whose state cannot be determined from exome-sequencing data. This suggests that exome data alone will miss much of the heritability for these traits. Stated somewhat differently: exome sequencing data for a specific individual misses much of the information necessary to compute their PRS score!

III. The DNA regions used in disease risk predictors so far constructed seem to be largely disjoint (with a few interesting exceptions), suggesting that individual genetic disease risks could be largely uncorrelated.

Observation III has interesting implications for pleiotropy[70–72]. We found that genetic risks are largely uncorrelated for different conditions. This suggests that there can exist individuals with, e.g., low risk simultaneously in each of multiple conditions, for any essentially any combination of conditions. There is no trade-off required between different disease risks (at least, not among the ones studied here). One could speculate that a lucky individual with exceptionally low risk across multiple conditions might have an unusually long life expectancy. Of course, our conclusions can only be preliminary because we only have access to a subset of risk alleles and in fact only common variants. Rare variants and as yet undiscovered variants may have qualitatively different behavior.

Of course, the same applies for high risk: some unlucky individuals have high risk for multiple conditions simultaneously. In fact, there appear to be combinations of SNPs that could make a specific individual an outlier in each of the conditions studied, simultaneously.

Note. Recently, it was pointed out[73] that the processing of the whole-exome sequencing data via the FE pipeline had been carried out in a manner that failed to take into account the presence of alternative contigs in the GRCh38 reference genome. This is expected to have led to fewer variants being called than there should be in the resultant data set. Out of the total of 204,829 genomic regions comprising 39.20 MBp of the human genome targeted by the whole-exome sequencing process, data from 7554 regions extending across 1.53 MBp were potentially affected by this error.

In an analysis of this issue[74], Jia et al. compared the number of exome variants per gene identified when using whole-exome sequencing data from the UK Biobank versus using data from gnomAD. They found 641 genes for which the UK Biobank exome data contains no variants whatsoever. In contrast, they calculated that it is highly probable for the UK Biobank exome data to identify at least one variant per gene in the case of the vast majority (93%) of these 641 genes.

With the aim of gauging the extent to which our results may have been impacted by this discrepancy, we examined the overlap between our lists of top genes ranked by variance accounted for by predictor SNPs (Supplementary Tables S23–S43) and the 641 potentially problematic genes. We found that for most (17 out of 21) conditions, the genes responsible for the top fifteen values of variance accounted for do not include any of the 641 potentially problematic genes.

The exceptions to this are asthma, basal cell carcinoma, type-1 diabetes, and hypothyroidism. To be specific: Asthma has 2 genes (HLA-DQB1 and HLA-DQA1 with variance 2–3% each) out of its top 18 genes included among the 641 potentially problematic genes. For comparison, asthma has 3% as the highest percentage of predictor variance accounted for by a single gene. Basal cell carcinoma has 1 gene (HLA-DQA1 with variance 2%) out of its top 20 genes included among the 641 potentially problematic genes, while 28% is the highest fraction of predictor variance accounted for by a single gene. Type-1 diabetes has 12 genes (HSPA1L, HLA-DRB1, BTNL2, HLA-DQB1, HLA-DQB2, NEU1, HLA-DOA, HLA-DQA1, HLA-DRA, HSPA1B, C6orf48, LSM2) out of its top 25 genes among the 641 potentially problematic genes. These 12 genes include the top four genes ranked by variance accounted for, with 14%, 9%, 7%, and 4% of predictor variance respectively. Hypothyroidism has 4 genes (HLA-DPA1, HLA-DQB1, HLA-DQA1, and HLA-DPB1, where two have 5% of predictor variance each and two have 1% each) out of its top 17 genes included among the 641 potentially problematic genes, while 8% is the highest value of predictor variance accounted for by a single gene.

Clearly, for all conditions except type-1 diabetes, the 641 potentially problematic genes play very little part in determining the variance accounted for by the predictor SNPs. We feel that this justifies our opinion that the upcoming corrections to the UK Biobank exome data will not qualitatively change our findings as regards these conditions. There is a possibility, of course, that there could be a significant shift in our results for the type-1 diabetes predictor.

Received: 25 February 2020; Accepted: 30 June 2020
Published online: 21 July 2020

References

1. Vattikuti, S., Lee, J. J., Chang, C. C., Hsu, S. D. & Chow, C. C. Applying compressed sensing to genome-wide association studies. *GigaScience* **3**, 10 (2014).
2. Ho, C. M. & Hsu, S. D. Determination of nonlinear genetic architecture using compressed sensing. *GigaScience* **4**, 44 (2015).

3. Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. GCTA: a tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.* **88**, 76–82 (2011).
4. Vilhjálmsson, B. J. *et al.* Modeling linkage disequilibrium increases accuracy of polygenic risk scores. *Am. J. Hum. Genet.* **97**, 576–592 (2015).
5. Lello, L., Raben, T. G., Yong, S. Y., Tellier, L. C. & Hsu, S. D. H. Genomic prediction of 16 complex disease risks including heart attack, diabetes, breast and prostate cancer. *Sci. Rep.* **9**, 2019 (2019).
6. Lello, L. *et al.* Accurate genomic prediction of human height. *Genetics* **210**, 477–497 (2018).
7. Khera, A. V. *et al.* Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nat. Genet.* **50**, 1219 (2018).
8. Marigorta, U. M., Rodriguez, J. A., Gibson, G. & Navarro, A. Replicability and prediction: lessons and challenges from GWAS. *Trends Genet.* **34**, 504–517 (2018).
9. Tam, V. *et al.* Benefits and limitations of genome-wide association studies. *Nat. Rev. Genet.* **20**, 467–484 (2019).
10. Chatterjee, N., Shi, J. & García-Closas, M. Developing and evaluating polygenic risk prediction models for stratified disease prevention. *Nat. Rev. Genet.* **17**, 392–406 (2016).
11. Euesden, J., Lewis, C. M. & O'Reily, P. F. PRSice: Polygenic Risk Score software. *Bioinformatics* **31**, 1466–1468 (2015).
12. Torkamani, A., Wineinger, N. E. & Topol, E. J. The personal and clinical utility of polygenic risk scores. *Nat. Rev. Genet.* **19**, 581–590 (2018).
13. Shieh, Y. *et al.* Breast cancer risk prediction using a clinical risk model and polygenic risk score. *Breast Cancer Res. Treat.* **159**, 513–525 (2016).
14. Lewis, C. M. & Vassos, E. Prospects for using risk scores in polygenic medicine. *Genome Med.* **9**, 96 (2017).
15. Abraham, G. & Inouye, M. Genomic risk prediction of complex human disease and its clinical application. *Curr. Opin. Genet. Dev.* **33**, 10–16 (2015).
16. Priest, J. R. & Ashley, E. A. Genomics in clinical practice. *BMJ Heart* **100**, 1569–1570 (2014).
17. Jacob, H. J. *et al.* Genomics in clinical practice: lessons from the front lines. *Sci. Transl. Med.*https://doi.org/10.1126/scitranslmed.3006468 *(2013).*
18. Veenstra, D. L., Roth, J. A., Garrison, L. P., Ramsey, S. D. & Burke, W. A formal risk-benefit framework for genomic tests: facilitating the appropriate translation of genomics into clinical practice. *Genet. Med.* **12**, 686–693 (2010).
19. Bowdin, S. *et al.* Recommendations for the integration of genomics into clinical practice. *Genet. Med.* **18**, 1075–1084 (2016).
20. Francisco, M. & Bustamante, C. D. Polygenic risk scores: a biased prediction?. *Genome Med.* **10**, 1–3 (2018).
21. Martin, A. R. *et al.* Clinical use of current polygenic risk scores may exacerbate health disparities. *Nat. Genet.* **51**, 584–591 (2019).
22. Nelson, H. D., Pappas, M., Cantor, A., Haney, E. & Holmes, R. Risk assessment, genetic counseling, and genetic testing for BRCA-related cancer in women: updated evidence report and systematic review for the US Preventive Services Task Force. *JAMA* **322**, 666–685 (2019).
23. Amir, E., Freedman, O. C., Seruga, B. & Evans, D. G. Assessing women at high risk of breast cancer: a review of risk assessment models. *J. Natl. Cancer Inst.* **102**, 680–691 (2010).
24. Offit, K. BRCA mutation frequency and penetrance: new data, old debate. *J. Natl. Cancer Inst.* **98**, 23 (2006).
25. Ford, D., Easton, D. F. & Peto, J. Estimates of the gene frequency of BRCA1 and its contribution to breast and ovarian cancer incidence. *Am. J. Hum. Genet.* **57**, 1457–62 (1995).
26. Whittemore, A. S. *et al.* Prevalence of BRCA1 mutation carriers among U.S. non-Hispanic Whites. *Cancer Epidemoiol. Biomark. Prev.* **13**, 2078–83 (2004).
27. Kuchenbaecker, K. *et al.* Evaluation of polygenic risk scores for breast and ovarian cancer risk prediction in BRCA1 and BRCA2 mutation carriers. *JNCI J. Natl. Cancer Inst.* **109**, 7 (2017).
28. Mavaddat, N. *et al.* Polygenic risk scores for prediction of breast cancer and breast cancer subtypes. *Am. J. Hum. Genet.* **104**, 21–34 (2019).
29. Kakushadze, Z., Raghubanshi, R. & Yu, W. Estimating cost savings from early cancer diagnosis. *Data* **2**, 30 (2017).
30. Cohen, L. E. Idiopathic short stature: a clinical review. *JAMA* **311**, 1787–1796 (2014).
31. Bryant, J., Baxter, L., Cave, C. B. & Milne, R. Recombinant growth hormone for idiopathic short stature in children and adolescents. *Cochrane Database Syst. Rev.* **3**, 004440 (2007).
32. Finkelstein, B. S. *et al.* Effect of growth hormone therapy on height in children with idiopathic short stature: a meta-analysis. *Arch. Pediatr. Adolesc. Med.* **156**, 230–240 (2002).
33. Cohen, P. *et al.* ISS Consensus Workshop participants, 2008. Consensus statement on the diagnosis and treatment of children with idiopathic short stature: a summary of the Growth Hormone Research Society, the Lawson Wilkins Pediatric Endocrine Society, and the European Society for Paediatric Endocrinology Workshop. *J. Clin. Endocrinol. Metab.* **93**, 4210–4217 (2007) .
34. Wit, J. M. *et al.* Idiopathic short stature: definition, epidemiology, and diagnostic evaluation. *Growth Horm. IGF Res.* **18**, 89–110 (2008).
35. Sudlow, C. *et al.* UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med.* **12**, 3 (2015).
36. Bycroft, C., Freeman, C. & Petkova, D. The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).
37. Azodi, C. .B. *et al.* Benchmarking parametric and machine learning models for genomic prediction of complex traits. *G3: Genes Genomes Genet.* **9**, 3691–3702 (2019).
38. Cunningham, F. *et al.* Ensembl 2019. *Nucleic Acids Res.* **47**, D745–D751 (2018).
39. UK Biobank. https://www.ukbiobank.ac.uk/. Accessed: 1 Aug 2018.
40. Bycroft, C. *et al.* Genome-wide genetic data on˜ 500,000 UK Biobank participants. *BioRxiv* **166298** (2017).
41. Donoho, D. & Tanner, J. Observed universality of phase transitions in high-dimensional geometry, with implications for modern data analysis and signal processing. *Philos. Trans. R. Soc. A Math. Phys. Eng. Sci.* **367**, 4273–4293 (2009).
42. Donoho, D. L. & Tanner, J. Precise undersampling theorems. *Proc. IEEE* **98**, 913–924 (2010).
43. Donoho, D. L. & Tanner, J. Sparse nonnegative solution of underdetermined linear equations by linear programming. *Proc. Natl. Acad. Sci.* **102**, 9446–9451 (2005).
44. Vattikuti, S., Lee, J. J., Chang, C. C., Hsu, S. D. & Chow, C. C. Applying compressed sensing to genome-wide association studies. *GigaScience* **3**, 2047–217X (2014).
45. Harrow, J. *et al.* GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res.* **22**, 1760–1774 (2012).
46. Church, D. M. *et al.* Modernizing reference genome assemblies. *PLoS Biol.* **9**, e1001091 (2011).
47. Gerstein, M. B. *et al.* What is a gene, post-ENCODE? History and updated definition. *Genome Res.* **17**, 669–681 (2007).
48. Gingeras, T. R. Origin of phenotypes: genes and transcripts. *Genome Res.* **17**, 682–690 (2007).
49. Portin, P. & Wilkins, A. The evolving definition of the term "gene". *Genetics* **205**, 1353–1364 (2017).
50. https://www.illumina.com/techniques/sequencing/dna-sequencing/targeted-resequencing/exome-sequencing.html.
51. Van Dijk, E. L., Auger, H., Jaszczyszyn, Y. & Thermes, C. Ten. years of next-generation sequencing technology. *Trends Genet.* **30**, 418–426 (2014).

52. Van Hout, C. V. *et al.* Whole exome sequencing and characterization of coding variation in 49,960 individuals in the UK Biobank. *bioRxiv* (2019) .
53. Regier, A. A. *et al.* Functional equivalence of genome sequencing analysis pipelines enables harmonized variant calling across human genetics projects. *Nat. Commun.* **9**, 4038 (2018).
54. Abecasis, G. R. *et al.* Extent and distribution of linkage disequilibrium in three genomic regions. *Am. J. Hum. Genet.* **68**, 191–197 (2001).
55. Stacey, S. N. *et al.* New basal cellcarcinoma susceptibility loci. *Nat. Commun.* **6**, 6825 (2015).
56. Stacey, S. N. *et al.* Germline sequence variants in TGM3 and RGS22 confer risk of basal cell carcinoma. *Hum. Mol. Genet.* **23**, 3045–3053 (2014).
57. Hunter, D. J. *et al.* A genome-wide association study identifies alleles in FGFR2 associated with risk of sporadic postmenopausal breast cancer. *Nat. Genet.* **39**, 870 (2007).
58. Easton, D. F. *et al.* Genome-wide association study identifies novel breast cancer susceptibility loci. *Nature* **447**, 1087 (2007).
59. Grant, S. F. *et al.* Variant of transcription factor 7-like 2 (TCF7L2) gene confers risk of type 2 diabetes. *Nat. Genet.* **38**, 320 (2006).
60. Buch, S. *et al.* A genome-wide association scan identifies the hepatic cholesterol transporter ABCG8 as a susceptibility factor for human gallstone disease. *Nat. Genet.* **39**, 995 (2007).
61. Jiang, Z. Y. *et al.* Increased expression of LXRα, ABCG5, ABCG8, and SR-BI in the liver from normolipidemic, nonobese Chinese gallstone patients. *J. Lipid Res.* **49**, 464–472 (2008).
62. Burdon, K. P. *et al.* Genome-wide association study identifies susceptibility loci for open angle glaucoma at TMCO1 and CDKN2B-AS1. *Nat. Genet.* **43**, 574 (2011).
63. Woodward, O. M. *et al.* Identification of a urate transporter, ABCG2, with a common functional polymorphism causing gout. *Proc. Natl. Acad. Sci.* **106**, 10338–10342 (2009).
64. Matsuo, H. *et al.* Common defects of ABCG2, a high-capacity urate exporter, cause gout: a function-based genetic analysis in a Japanese population. *Sci. Transl. Med.* **1**, 5–11 (2009).
65. Vitart, V. *et al.* SLC2A9 is a newly identified urate transporter influencing serum urate concentration, urate excretion and gout. *Nat. Genet.* **40**, 437 (2008).
66. Trégouët, D. A. *et al.* Genome-wide haplotype association study identifies the SLC22A3-LPAL2-LPA gene cluster as a risk locus for coronary artery disease. *Nat. Genet.* **41**, 283 (2009).
67. Valverde, P. *et al.* The Asp84Glu variant of the melanocortin 1 receptor (MC1R) is associated with melanoma. *Hum. Mol. Genet.* **5**, 1663–1666 (1996).
68. Kennedy, C. *et al.* Melanocortin 1 receptor (MC1R) gene variants are associated with an increased risk for cutaneous melanoma which is largely independent of skin type and hair color. *J. Investig. Dermatol.* **117**, 294–300 (2001).
69. *MS Windows NT Kernel Description.* http://www.nealelab.is/uk-biobank/. Accessed: 23 May 2020.
70. Hackinger, S. *Pleiotropy in complex traits, Diss* (University of Cambridge, Cambridge, 2019).
71. Hackinger, S. & Zeggini, E. Statistical methods to detect pleiotropy in human complex traits. *Open Biol.* **7**, 170125 (2017).
72. Socrates, A. *et al.* Polygenic risk scores applied to a single cohort reveal pleiotropy among hundreds of human phenotypes. *bioRxiv* **203257** (2017).
73. https://www.ukbiobank.ac.uk/wp-content/uploads/2019/12/UK-Biobank-50k-Exome-ReleaseFAQ-December-2019.pdf.
74. Jia, T., Munson, B., Allen, H. L., Ideker, T. & Majithia, A. R. Thousands of missing variants in the UK BioBank are recoverable by genome realignment. *bioRxiv* (2019).

Acknowledgements

S.Y.Y., T.G.R., L.L., and S.D.H.H. acknowledge support from the Office of the Senior Vice President for Research and Innovation at Michigan State University. L.L. was supported by Genomic Prediction, Inc. during part of this work. Computational resources were provided by the Institute for Cyber-Enabled Research at Michigan State University. This research was conducted using the UK Biobank Resource under UK Biobank Main Application 15326.

Author contributions

S.Y.Y. analyzed data and prepared figures. S.Y.Y., T.G.R., and L.L. developed code used in the analysis. S.Y.Y., T.G.R., and S.D.H.H. wrote the text of the manuscript. S.D.H.H. managed and designed the project. All authors reviewed the manuscript.

Competing interests

Stephen Hsu is a shareholder and serves on the board of directors of Genomic Prediction, Inc.. Louis Lello joined the company, becoming an employee and shareholder, during the writing and submission of this paper. Soke Yuen Yong and Timothy Raben declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41598-020-68881-8.

Correspondence and requests for materials should be addressed to S.Y.Y.