EPFL

Thèse n° 8464

The genetic architecture of complex human traits at the dawn of genomic medicine

Présentée le 19 avril 2021

Faculté des sciences de la vie
Groupe Fellay
Programme doctoral en biotechnologie et génie biologique

pour l'obtention du grade de Docteur ès Sciences

par

Olivier Noël Marie NARET

Acceptée sur proposition du jury

Prof. A. L. A. Persat, président du jury
Prof. J. Fellay, directeur de thèse
Prof. O. Delaneau, rapporteur
Prof. Z. Kutalik, rapporteur
Prof. B. Deplancke, rapporteur

École
polytechnique
fédérale
de Lausanne

2021

Heard melodies are sweet,
but those unheard, are sweeter.
— John Keats, Ode On A Grecian Urn And Other Poems

To my brother...

Acknowledgements

I would like to thank first of all Jacques Fellay who gave me this wonderful opportunity to pursue a PhD in the exceptional environment that is the Fellay Lab. Professor Jacques Fellay offered me a very flexible mentorship with excellent guidance and framework when needed while leaving me a lot of freedom. By valuing out-of-the-box thinking and the exploration of fields outside of the core line of the lab, he has made these 4 years of PhD fascinating, diverse and eventful.
I also warmly thank Bruno Lemaitre who accepted to take on the role of mentor, Bart Deplancke, Zoltan Kutalik, Olivier Delaneau and Alexandre Persat who all accepted to be part of the jury where they gave some of their precious time to apply their expertise to my work.
I would also like to thank Jonathan K Pritchard, who offered me a wonderful opportunity by hosting me in his lab at Stanford for a period of 6 months and Yuval Simons with whom I had the pleasure to collaborate.
Special thanks to all my former colleagues who passed through the Fellay Lab. First of all Nimisha Chaturvedi who was the first person I could have an exciting collaboration with, but also Samira Asgari, Alessandro Borghesi, Petar Scepanovic, Christian Hammer, Christian Thorball and Sina Rueger. Similarly, I thank my current colleagues whom I would have liked to see more of during the past year impacted by covid, in particular Flavia Hodel who, in addition to being an excellent scientist with an exceptional work ethic, was a wonderful neighbour always full of good humour, as well as Dylan Lawless, Thomas Junier, Zhi Ming Xu (aka Mack) and Konstantin Popadin. All of them have been excellent comrades in the daily life of the lab as well as in more extraordinary contexts such as conferences on the other side of the world and epic lab outings in the Swiss mountains on skis and on foot.
Finally, I thank my family and friends, and especially Shahrzad, my wonderful partner in life, the nicest person in the world and a brilliant scientist with whom I have lived 4 wonderful years of PhD and many more to come.

Lausanne, March 2021 O. N.

Abstract

The focus of the work presented in this thesis is the exploration of the genetic architecture of complex human traits - at the dawn of genomic medicine.

The underlying mechanisms explaining the enormously polygenic nature of most human complex traits are still unknown. The first chapter explores a possible explanatory model in which variant effects are due to an indirect mechanism, namely competition among genes for shared intracellular resources such as ribosomes. Our findings show that under most reasonable assumptions, resource competition should not be expected to have much impact on either protein expression levels of individual genes or on complex trait outcomes.

The prediction accuracy of polygenic scores (PGS) remains relatively modest compared to what is expected given the estimated heritability of traits. Traditionally, the construction of PGS uses a large number of genetic variations, most of which have weak additive effects. Recent machine learning methods could improve PGS by also aggregating epistatic effects. To evaluate these different methods, we conducted an experiment based on an innovative concept of crowdsourcing, detailed in the second chapter. We collaborated with opensnp.org, an open repository where people share their genotyping data and phenotypic information, and with crowdai.org, a platform that allowed us to create a public competition for the genomic prediction of height. The challenge lasted three months and attracted 138 participants. This was the first crowd-sourcing challenge based on publicly available genome-wide genotyping data.

Due to the enormous number of potential combinations of variants, it is difficult to integrate epistatic effects into PGS. In the third chapter, we present a method where we limit the possible combinations to the boundaries of each topologically associated domain (TAD) independently. With the UK Biobank, for the height phenotype, we included 17,560 variants in an artificial neural network (ANN) and compared the variance explained (R^2) by the PGS with or without the knowledge of the TADs. We found that it brings a significant improvement with an average R^2 going from 0.287 to 0.293 (with a p-value = $10E-5$ for n=20). We concluded that it should be possible to build better PGS using ANNs and epistasis in TADs.

The effect of genetic ancestry on phenotypes is not taken into account in PGS-based risk estimates. Doing so could accelerate the adoption of genomic medicine for underrepresented populations and mixed-race individuals. The fourth chapter presents a method for its integration through a secondary score derived from genome-wide genotyping data, the PC score

(PCS). We compared two models, one using only the PGS and the other using both the PGS and the PCS. Using the UK Biobank, we found an improvement in genetic prediction for all phenotypes tested: <10% for blood pressure, BMI and baldness, 16% for menarche age, 38% for height, 71% for menopausal age, 138% for bone mineral density, 350% for education and 2800% for skin color. These results were reproduced when the trained models were applied to an external cohort (Cohort Lausannoise).
Each advance in the understanding of complex traits and the calculation of PGS has the potential to improve genomic medicine when used routinely in clinical practice. During these four years, I have had the opportunity to act at different levels to participate in this long-awaited evolution.

Key words: complex traits, omnigenic trait, polygenic score, precision medicine, topologically associated domains, artificial neural network, crowdsourcing

Résumé

Le travail présenté dans cette thèse est axé sur l'exploration de l'architecture génétique des traits humains complexes - à l'aube de la médecine génomique.

Les mécanismes sous-jacents expliquant la nature extrêmement polygénique de la plupart des traits humains complexes sont encore inconnus. Le premier chapitre explore un modèle explicatif possible dans lequel les effets de variantes sont dus à un mécanisme indirect, à savoir la compétition entre les gènes pour des ressources intracellulaires partagées telles que les ribosomes. Nos résultats montrent que, selon les hypothèses les plus raisonnables, la compétition pour les ressources ne devrait pas avoir beaucoup d'impact sur les niveaux d'expression protéique des gènes individuels ou sur les résultats des traits complexes.

La précision des prédictions des scores polygéniques (PGS) reste relativement modeste par rapport à ce que l'on attend compte tenu de l'héritabilité estimée des traits. Traditionnellement, la construction de PGS utilise un grand nombre de variations génétiques, dont la plupart ont de faibles effets additifs. Les méthodes récentes d'apprentissage machine pourraient améliorer le PGS en agrégeant également les effets épistatiques. Pour évaluer ces différentes méthodes, nous avons mené une expérience basée sur un concept novateur de crowdsourcing, détaillé dans le deuxième chapitre. Nous avons collaboré avec opensnp.org, un dépôt ouvert où les gens partagent leurs données de génotypage et leurs informations phénotypiques, et avec crowdai.org, une plateforme qui nous a permis de créer un concours public pour la prédiction génomique de la taille. Le défi a duré trois mois et a attiré 138 participants. C'était le premier défi d'approvisionnement par la foule basé sur des données de génotypage génomique accessibles au public.

En raison du nombre énorme de combinaisons potentielles de variantes, il est difficile d'intégrer des effets épistatiques dans les PGS. Dans le troisième chapitre, nous présentons une méthode où nous limitons les combinaisons possibles aux limites de chaque domaine topologiquement associé (TAD) indépendamment. Avec la Biobank britannique, pour le phénotype de hauteur, nous avons inclus 17 560 variantes dans un réseau neuronal artificiel (ANN) et comparé la variance expliquée (R^2) par le PGS avec ou sans la connaissance des TAD. Nous avons constaté qu'elle apporte une amélioration significative avec une moyenne de R^2 passant de 0,287 à 0,293 (avec une p-value = $10E-5$ pour n=20). Nous avons conclu qu'il devrait être possible de construire de meilleurs SGP en utilisant les ANN et l'épistasie dans les TAD.

L'effet de l'ascendance génétique sur les phénotypes n'est pas pris en compte dans les estimations de risque basées sur les SGP. Cela pourrait accélérer l'adoption de la médecine génomique pour les populations sous-représentées et les individus de race mixte. Le quatrième chapitre présente une méthode pour son intégration par le biais d'un score secondaire dérivé des données de génotypage à l'échelle du génome, le score PC (PCS). Nous avons comparé deux modèles, l'un utilisant uniquement le PGS et l'autre utilisant à la fois le PGS et le PCS. En utilisant la Biobanque du Royaume-Uni, nous avons constaté une amélioration de la prédiction génétique pour tous les phénotypes testés : <10% pour la pression sanguine, l'IMC et la calvitie, 16% pour l'âge de la ménarche, 38% pour la taille, 71% pour l'âge de la ménopause, 138% pour la densité minérale osseuse, 350% pour l'éducation et 2800% pour la couleur de la peau. Ces résultats ont été reproduits lorsque les modèles formés ont été appliqués à une cohorte externe (Cohorte Lausannoise).
Chaque avancée dans la compréhension des traits complexes et le calcul des SGP peut améliorer la médecine génomique lorsqu'elle est utilisée en routine dans la pratique clinique. Au cours de ces quatre années, j'ai eu l'occasion d'agir à différents niveaux pour participer à cette évolution tant attendue.

Mots clefs : traits complexes, trait omnigenic, score polygénique, médecine de précision, domaines topologiquement associés, réseau neural artificiel, crowdsourcing

Contents

1 Introduction

Following the completion of the Human Genome Project in 2001, there were high hopes that genomic approaches would transform clinical care by favoring the emergence of tailored treatments, diagnostic and prognostic tools that take into account the individual genetic risk. Since then, however, the promises have not yet been fulfilled, due to the many hurdles that have emerged along the way revealing a complexity beyond what was expected. These challenges are both fundamental, with the persistence of questions about the architecture of complex diseases, and applied, with the need to translate the knowledge gained from genomic studies into clinical care that benefits everyone.

A constant in the history of the study of complex traits and diseases is the discrepancy between the degree of complexity observed and expected. Fisher's first predictions in 1918 stated that the more complex a trait is - that is, the more genes it is under the influence of - the smaller the individual contribution of each gene would be[1]. Although this statement is still valid, the number of associated loci goes far beyond previous expectations and still leaves us puzzled about the architecture of complex traits. For example, the degree of complexity for human standing height is such that one variant is expected to be causal every 100 kbp[2]. Recently, a new model called "omnigenic" from J.K. Pritchard's laboratory has attempted to frame this observation by classifying genes into two categories. On the one hand, core genes, corresponding to the orthodox belief in genes directly involved in the biological pathway responsible for a trait. On the other hand, peripheral genes, described as a new category of causative genes that have an indirect impact on a trait through their random interaction with the core genes. The mechanistic link between core and peripheral genes remained unexplained, which led us to hypothesize that it could be due to a competition between peripheral and core genes for the use of cellular resources. This work is presented in Chapter 3 of the thesis. We have been able to establish, through a mathematical model, that while under certain specific conditions such competition is indeed possible, it is unlikely to be the main driving force of the omnigenic architecture.

Genome-wide association studies (GWAS) for complex traits and diseases have identified multiple associations with genetic variants. Their cumulative impact can be included in a

point estimate polygenic score (PGS) which can be used in genomic prediction models. To date, the predictive power of PGS is still relatively modest compared to the genetic component estimated from heritability methods [3] or twin studies [4]. Several avenues should be explored to reduce this gap, from an increase in the size of the cohorts and the number of variants tested, to the improvement of methods that address the genotype-to-phenotype equation to construct PGS. Recently, new machine learning methods have been successfully applied to large datasets in many fields. These techniques are numerous and their ability to accurately predict complex features needs to be evaluated. Crowdsourced contests for data scientists have succeeded in bringing together data science experts and enthusiasts from many fields to solve real-world problems through online challenges. However, because human genomics is based on highly sensitive data, such an approach has not yet be applied to the field. Thanks to OpenSNP, a community of citizen scientists who share their direct-to-consumer genotyping data, we were able to organize a genomic prediction challenge of the height phenotype that attracted 138 participants. This project is described in Chapter 4.

Complex traits result from the combined effects of individual genes, but could also be partly determined by gene x gene interactions. These potential interactions represent an unknown part of the phenotypic variance [5] and are not integrated in current PGS, which typically rely solely on the additive effects of genetic variants. As described in Chapter 5, we decided to build an artificial neural network that could uncover potential gene x gene interactions by exploiting topological association domains (TADs). TADs are functionally delimited genomic regions of ~ 800kb with self-interacting DNA, meaning that DNA sequences inside a TAD physically interact with each other more frequently than with DNA sequences outside the TAD. We were able to demonstrate a significant difference for a PGS of height, which was improved by 3.5% when trained on a sequence where the integrity of the TADs was preserved.

Over the past three decades, the bulk of human genomic research has been performed in North American and European research institutions. As a consequence, it mainly focused on populations of European ancestry. This has become an acute problem at the dawn of genomic medicine, as existing resources on common genetic variation and associations with complex traits are heavily skewed in favour of Europeans, who represent about 80% of the individuals included in genomic studies. To counterbalance, we have developed a method, described in Chapter 6, that allows the inclusion of the genetic ancestry component with PGS in predictive models. We then demonstrated its ability to improve phenotypic prediction over a wide range of phenotypes.

Over the past years, the field of genomic research has undergone major changes. The broad availability of data from very large cohorts such as the UK Biobank[6] has made it possible to ask new questions about the extreme polygenicity of traits, which has led to the development of the omnigenic model[2]. There has been an explosion of interest in PGS, motivated by claims of their maturity and their potentially imminent use in clinical practice[7]. This has been accompanied by growing concerns about equity in genomic research: certain populations are under-represented, potentially exacerbating the risk of health inequalities[8]. The use of

artificial neural networks has grown rapidly, culminating recently in a closely related field with the remarkable success of the DeepMind protein folding algorithm[9].

Now is a fascinating time to do research - and indeed a PhD - in the field of human genomics. The journey from understanding of complex diseases to concrete implementation of genomic medicine has only begun.

2 Background Research

2.1 Genome-wide association studies

2.1.1 Key concepts

History

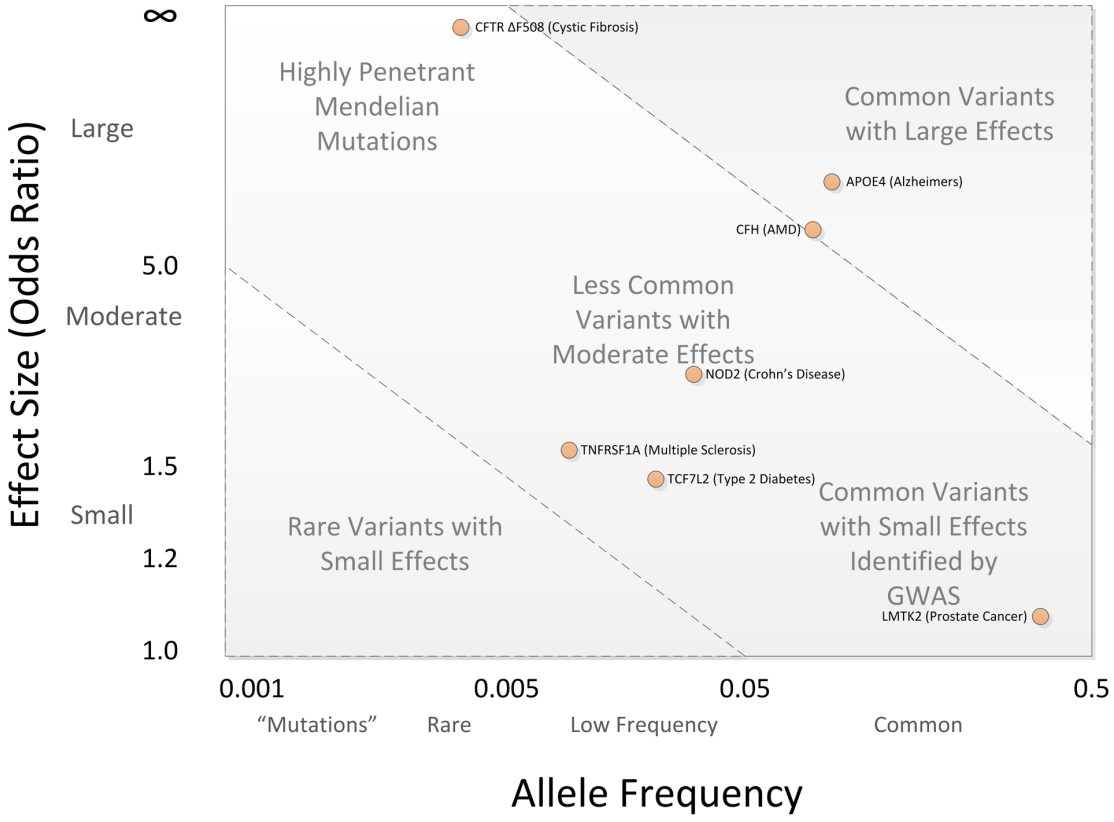
At the dawn of the 21st century, the completion of the Human Genome Project carried the hopes that genomics would profoundly change medicine by making it more precise and more personal [10, 11]. Soon after, the first Phase of the HapMap project was completed in 2005, which provided crucial information about human genomic diversity. The mapping of Single Nucleotide Polymorphisms (SNPs) across the genome was the key to drive population-based studies and ushered in the era of Genome-Wide Association Studies (GWAS) [12]. GWAS aim to identify common genetic variants that vary systematically between individuals with different disease states. SNPs are by far the most abundant form of genetic variation in the human genome [13]. More than 100 million have been validated so far: SNP density has been estimated for around once in every 1000 nucleotides, totaling roughly 3.5 to 4 million SNPs in a person's genome. GWAS have identified countless associations between SNPs and complex traits or diseases over the past 15 years.

Common disease common variant

We distinguish between rare variants that are directly linked to a rare disease (e.g.: DMD gene variants causing Duchenne muscular dystrophy, inherited in an X-linked recessive manner) through linkage analysis, and common variants - referred to as SNPs - that can statistically co-occur with a disease more often than it would be expected by chance, linked to common diseases through association analysis. A linkage analysis (base on known models of inheritance) which works for rare diseases has no power to identify genetic factors involved in complex disease. This reveals a different underlying genetic architecture for common diseases, which led to the "common disease common variant" hypothesis [14]. It relies on the fact that a

common variant will have a small effect size, and that a common polygenic disease will thus be influenced by multiple SNPs [15]. To decipher complex diseases, family-based genetic studies are not likely to be successful, which prompted a shift toward population-based studies [16]. An alternate hypothesis, the "common disease rare variant" hypothesis, argues that multiple rare DNA sequence variations, each with relatively large effect size, are the major contributors to genetic susceptibility to common diseases. GWAS studies initially relied mostly on the "common disease common variant" hypothesis, however the recent availability of very large cohorts has shown that rare variants also contribute to the heritability of common diseases. A SNP is characterized by both its frequency (minor allele frequency, MAF) and its effect size (beta for continuous traits or odd ratio for case-control phenotypes). The Fig.2.1 shows the various combinations of MAF and effect sizes observed for human SNPs, and their respective (theoretical) effects on diseases and traits.

Figure 2.1 – Spectrum of Disease Allele Effects.



Genetic variants categorisation and the corresponding association studies can be represented in two dimensions: allele frequency and effect size. Highly penetrant alleles for Mendelian disorders are extremely rare with large effect sizes (upper left), while most GWAS findings are associations of common SNPs with small effect sizes (lower right). The bulk of discovered genetic associations lie on the diagonal denoted by the dashed lines. Figure from [16]

Linkage disequilibrium & tagging SNPs

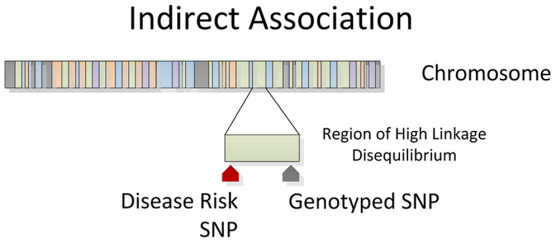
Much of the human genome is b̈lock like, with little recombination within block and block boundaries that are hotspots of intense recombination. [17, 18] The human genome size is about 3,100 mega basepairs. The amount of data generated by full genome sequencing is therefore massive, resulting in a practical need to reduce the genotyping target. It is possible to maximize the information brought by a single genotyped SNP by leveraging the linkage disequilibrium (LD) property of genetic data.

LD is measured by the coefficient of determination D'[19]. It is scaled between 0 and 1 (with 0 for a total equilibrium) such as with $D_{AB} = P_{AB} - P_A P_B$ with $D' = \frac{D}{D_{max}}$. Alternatively, r can be used as the coefficient of correlation with $r = \frac{D}{P_A(1-P_A)P_B(1-P_B)}$. Because these metrics are sensitive to any recombination, they will decay over time.

Genotyping a carefully selected set of SNPs is sufficient to determine untyped SNP through imputation using haplotype reference panels. Although SNPs are genotyped one by one, methods allow inference of phase [20]. Of the 100 million discovered SNPs, ẗags SNPsäre selected which allow the capture of variation at nearby sites thought stretch of LD [21, 22, 23]. 80% of the commonly occurring SNPs in a European population can be captured using 500,000 to 1 million SNPs scattered across the genome [24]. In the downstream analysis of GWAS results, it is important to distinguish between functional SNPs that are directly the cause of the trait, and indirectly associated SNPs [25] that are in LD with the variant causing the trait. See Fig.2.2.

Some missing genotype information can be recovered by doing statistical inference based on the observed genotypes at neighboring SNPs. It is based on a prior haplotype estimation or phasing, relying on external panels built from whole-genome sequencing projects such as the International HapMap[26], the 1000 Genome Project[27], or the Haplotype Reference Consortium [28]. Initially, imputation was done using multinomial models in which each possible haplotype was given an unknown frequency parameter estimated by the expectation-maximization (EM) algorithm. Nowadays, the most accurate methods use Hidden Markov models (HMM). Imputation allows to increase the SNP density and therefore the number of tested variants. More than 10 million genetic variants are nowadays commonly tested in GWASs[29].

Figure 2.2 – Indirect association.



Genotyped SNPs often lie in a region of high linkage disequilibrium with an causal allele. The genotyped SNP will be statistically associated with disease as a surrogate. Figure from [16]

2.1.2 Statistical power and significance

Multiple testing burden

Because many SNPs are tested simultaneously in a GWAS, there is a need to define a strategy to deal with multiple testing issues. Family-wise error rates like Bonferroni correction are very stringent, assuming that tests are independent, it necessitates a p-value of 5.10^{-8} for 5% type-I error with a 1 million SNPs GWAS. The premise however of independent tests is not correct due to the LD properties of the human genome. Therefore, modifications have been proposed through the use of an *effective number* of the independent test (Me), based on counting the number of LD blocks and SNP singleton [30], or through the eigenvalue of the correlation matrix of the SNP allele counts [31]. Applied to the latest Illumina SNP array, containing 2.45 million SNPs, it estimated the number of 1.37 million Me raising a corresponding *genome-wide significant threshold* of $3.63.10^{-8}$.

Type I error can also be approximated by a Monte Carlo permutation procedure where phenotype labels are randomized over conserved genotype data. All M tests are then calculated and the smallest p-value is pulled out. The process is repeated to construct an empirical frequency distribution of the smallest p-values. The p-values calculated from the real data are then compared to this distribution to determine an empirical adjusted p-value. $P_adj = \frac{(r+1)}{(n+1)}$, with n the number of permutation and r the number of p-value smaller than the one obtained from test on real data [32]. Finally, the false discovery rate (FDR) is an estimate of the proportion of the significant results that are false positive. It has been used in some GWAS studies. [33, 34]

Statistical power estimation

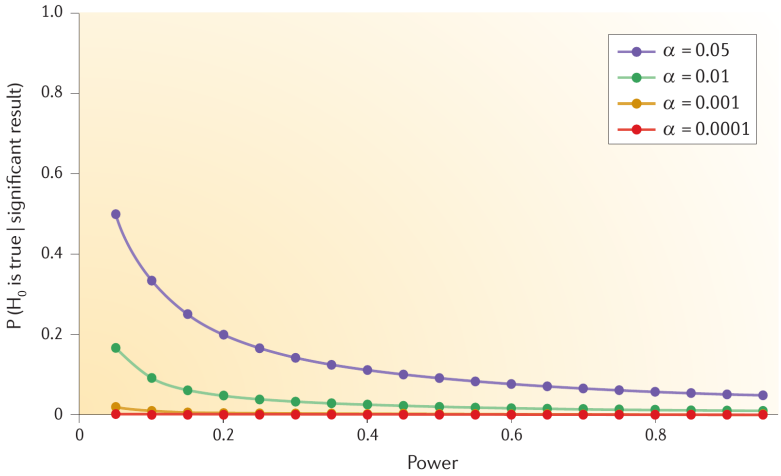
In a landmark paper published almost 20 years ago, it was found that only 6 of the 166 initial association findings were reliably replicated in subsequent studies [35]. The false-positive report probability [36] (FPRP), which is $P(H_0 \mid P \leq \alpha)$, the probability of no true association between a genetic variant and a disease given a statistically significant finding, depends not only on the observed p-value but also on both the prior probability that the association between the genetic variant and the disease is real and the statistical power of the test.

$$\begin{aligned} P(H_0 \mid P \leq \alpha) &= \frac{P(P \leq \alpha \mid H_0)P(H_0)}{P(P \leq \alpha \mid H_0)P(H_0) + P(P \leq \alpha \mid H_1)P(H_1)} \\ &= \frac{\alpha\pi_0}{\alpha\pi_0 + (1-\beta)(1-\pi_0)} \end{aligned} \tag{2.1}$$

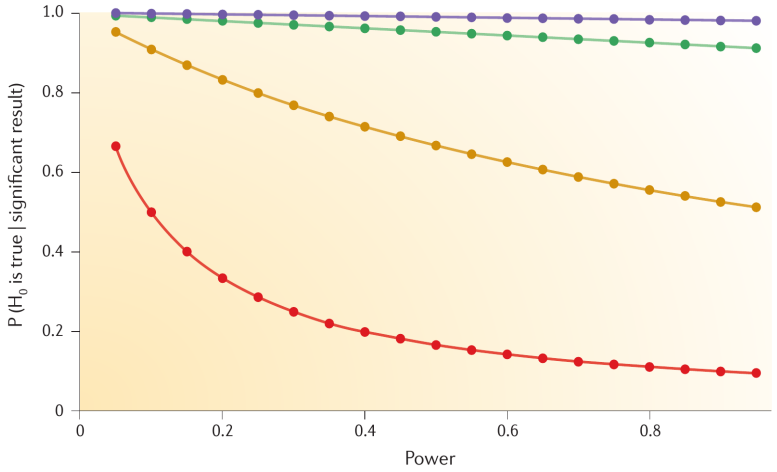
Where π_0 is the prior probability for H_0 to be true. For complex traits in the context of GWAS, multiple SNPs have a true association with the trait, and $P(P \leq \alpha \mid H_1)$ could be considered as the average statistical power of all SNPs for which H_1 is true. See Fig. 2.3A, 2.3B

Figure 2.3 – Prior probability of H_0

(A) Prior probability that H0 is true = 0.5



(B) Prior probability that H0 is true = 0.999



Posterior probability of H_0 given the critical significance level and the statistical power of a study, for different prior probabilities of H_0. We see that the probability of false-positive association decreases with increasing power, decreasing significance level, and decreasing the prior probability of the null hypothesis (H_0). Figure from [37]

We could set alpha to control FPRP

$$\alpha = \frac{P(H_0|P \leq \alpha)}{1 - P(H_0|P \leq \alpha)} \frac{1-\pi_0}{\pi_0}(1-\beta) \tag{2.2}$$

The difficult part remains in evaluating the power, which requires making assumptions on the underlying disease model.

Most common diseases have a complex genetic architecture that involves multiple risk loci combined with environmental factors. Rare alleles or alleles with very small effect sizes will be especially hard to detect because they account for little variance in the global liability (predisposition to cause disease).

If a causal SNP is untyped nor imputed, a proxy SNP correlated with r to the causal SNP would require a sample increased by a factor $\frac{1}{r^2}$ to be detected. Plus, due to MAF difference, a rare variant cannot be in strong LD with any of the common variants usually present on commercial SNP arrays. Rare variants could be assigned more weight as they are more likely to be subject to negative selection. For detecting them, a gene or pathway-based approach could be used.

2.1.3 Biological and Statistical models binding

GWAS Outcome

In a case-control study, the predictor (SNPs) will be linked to a binary - case or control status - outcome. Whereas in quantitative studies, the predictor will be linked to quantitative outcomes (e.g. plasma concentration of a biomarker).

Single SNP test

The ideal test will depend on the unknown underlying biological model (e.g., recessive, dominant, over-dominant) of the disease-predisposing variant. The additive model is commonly used as a standard choice, it assumes that the heterozygous risk is intermediate between the two homozygous risks.

Traits are analyzed using a linear model, including linear regression for quantitative traits and logistic regression for case-control studies. Case-control uses generalized linear model because of the non-normality of the residuals. In such a case, the outcome is transformed using a logistic function that predicts the probability of having case status given a genotype class. Here we fits $Y = s(\alpha + \beta X_1 + \gamma X_2)$, where s is the logit function. The logistic regression model assumes that the log-odds of an observation y can be expressed as a linear function of the K input variables x:

$$log\frac{P(x)}{1-P(x)} = \sum_{j=0}^{K} b_j x_j \tag{2.3}$$

Here the disease risk of individuals is equated to $\beta_0, \beta_1, \beta_2$ according to the genotype, and a likelihood-ratio test of this model will be computed against the null hypothesis of $\beta_0 = \beta_1 = \beta_2$. If we fix the beta to $\beta_1 = \frac{\beta_0+\beta_2}{2}$ to stick to additive model a 1_{df} tests is obtained, equivalent

to the Armitage test. For single SNP analysis, there is not a lot of interest in these more complicated test models, however, when it comes to multiple SNP testing with complex epistatic and environmental interactions or covariates such as population stratification, it offers much more flexibility. The logistic regression can provide adjusted odds ratios as a measure of effect size, which is the chances of being a case given the presence of a SNP, proportional to the chances of being a case without the SNP.

Collapsing tests, rare variant analysis

Collapsing the signal from multiple SNPs into one is both a test of association for multiple SNPs and a way to leverage rare variants [38]. With high-throughput sequencing, the whole genome can now be sequenced quickly and at reasonable price, making methods relying on LD to tagged untyped causal SNPs less relevant.

Running single marker association analyses that include rare variants, at the scale of the genome, would dramatically increase the multiple testing burden. For this reason, region-based analysis is preferred (genes, moving windows, networks/pathways). A burden test will aggregate variant information in a region into a summary dose variable and will be more powerful if all variants are causal with the same effect sizes and association directions. A burden test can be weighted to give more importance to rare variants, which theoretically have a bigger effect size, and/or to variants with predicted functional effects.

Multi-locus analysis, epistatic effects : gene-gene interaction

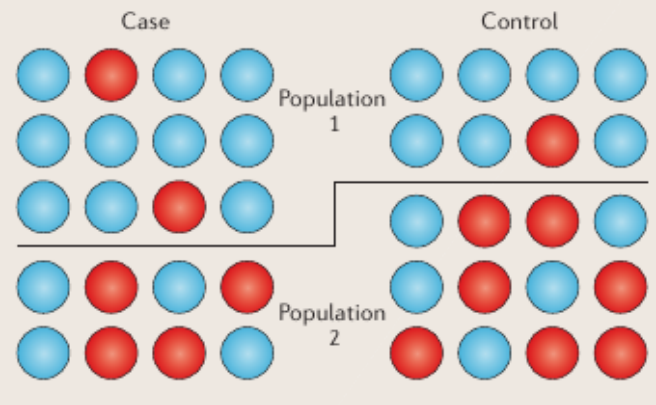
It is commonly admitted that genes interact with each other and with environmental factor to cause complex diseases[39]. Because of the number of SNPs across the genome, well-powered global pairwise study are challenging [40]. One solution is to only select the SNPs that have a marginal effect for pairwise analysis, searching for potential interactions with any other SNP in standard regression models[41, 42]. However, a lot of multi-locus interactions will be missed, which have purely epistatic effect[40, 43]. Another approach is to select SNPs based on biological relevance for the studied trait, as for example a biological pathway, structure, or function. Epistatic effects can be evaluated by a deviation from a model of additive effects, but it could be either on a logarithmic or linear scale, which implies a different definition.

Confounding factors and population stratification

Some non-genetic factors such as age, sex, and location can impact statistical tests and produce spurious associations. Covariate adjustment will minimize these effects but use additional degrees of freedom. Population stratification can also lead to spurious associations in GWAS. It happens when within the case or control group, a population is overrepresented in comparison to the other group (Fig. 2.4). There are several possible reasons for a subgroup to be overrepresented among cases: a higher proportion of causal alleles in a subgroup; different

environmental factors between subgroups, which either modify the penetrance of an allele or directly impact the disease outcome; a sample bias. All of these can lead to ascertainment bias. Population structure is a misnomer: it actually represents a pattern of distant kinship. Spurious associations arise when cases or controls are more related to one population than the other group.

Figure 2.4 – Population Stratification.



We can see that case and control contains in different proportions, two different subpopulations. This could lead to spurious associations. Figure from [44]

The most used techniques to control for population stratification are based on principal component (PC) analysis of the genotyping matrix followed by the inclusion of the eigen values of the significant PC axes as covariates in the regression analyses[45].

Nowadays, the use of mixed linear model is the most popular method to correct for multiple levels of relatedness simultaneously, including the presence of cryptic relatedness and population stratification [46, 47].

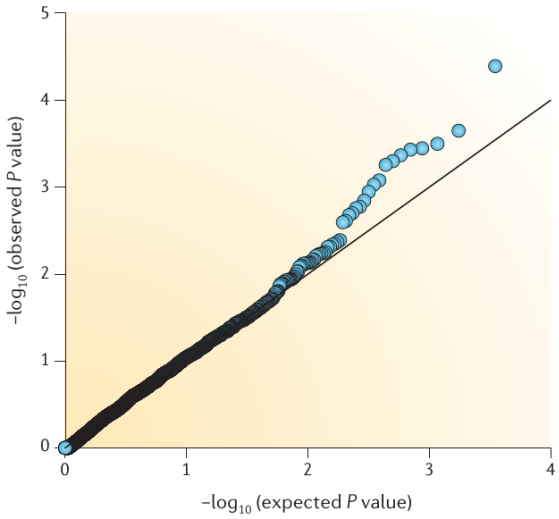
A common representation of association of multiple SNPs is the quantile-quantile (QQ-)plot of p-values (Fig. 2.5. Negatively ranked log p-values are plot against their null expectations. Early departure of negatively ranked log p-values from their expected values is a strong indication of the presence of population stratification. Nevertheless, it has been shown that highly polygenic trait shows inflation due not to spurious factor but the true underlying genetic architecture [48].

2.1.4 Results and Beyond

Polygenic score

: One use of GWAS results is the computation polygenic score (PGS) sometimes referred to as a polygenic risk score (PRS) in the context of diseases. It aims as aggregating the effect of the many variants associated with a trait or disease - with a lenient significant threshold - to an individual score. This score can then be used, ideally with other existing clinical predictors, to create a higher-level composite score, i.e., to stratify individuals by risk and potentially individualize medical care [49].

Figure 2.5 – Quantile-quantile plot.



QQ-plot representation summaries the global association profile on the study. Inflation at the end (lowest p-value) correspond to causal associations. Inflation that begins too early corresponds to population stratification. Figure from [44]

Ancestry issues

The GWAS field suffers from structural inequality that commonly hit societies, including the representativity of the samples that have been included in studies. In 2020, the GWAS diversity monitor shows that there is a huge imbalance in terms of representativity with 84% of samples from individuals of European ancestry [50]. This could have a dramatic impact on the practical implementation of genomic medicine, making it much harder for people of non-European ancestry to benefit from recent genomic advances. For these reasons, there is an urgent need of expanding GWAS studies to multiple ancestries, as done for example by consortia such as PAGE, H3Africa, the African Genome Variation Project, and GenomeAsia 100k. In addition, techniques should be developed to allow for a quick transfer of existing results to admixed populations.

Summary

As of 2020-12-02, the GWAS Catalog contains 4795 publications and 222,000 for 55,000 unique loci and 5000 traits. The first GWAS in 2005[51] analyzed 146 samples, while some studies now include more than a million participants. The number of variants tested as also increased from <500,000 in the early days to >10 millions today.

2.2 Heritability

2.2.1 Heritability definition

Heritability is a dimensionless population parameter summarizing the proportion of variance that is attributable to genetic variation in a given environment, distinguishing what

is attributable to nature vs. what is attributable to nurture. It is a ratio of variances: the numerator contains the genetic component; the denominator contains the observed phenotypic component. Broad-sense heritability (H^2) describes the proportion of phenotypic variance attributable to the total genetic variance, $H^2 = frac\sigma_G^2\sigma_P^2$. σ_P^2 can be broken down in $\sigma_P^2 = \sigma_G^2 + \sigma_E^2 + \sigma_{G\times E}^2$ where σ_E^2 is the phenotypic variance caused by environmental factors and $\sigma_{G\times E}^2$ the interaction between environmental and genetic factors. σ_G^2 can be broken down in $\sigma_A^2 + \sigma_{G\times G}^2 + \sigma_D^2$ where σ_A^2 is the additive, $\sigma_{G\times G}^2$ the epistatic (interaction between different loci alleles), and σ_D^2 the dominant (interaction between same locus alleles) genetic component. Narrow-sense heritability (h^2) describes the proportion of the additive genetic component in the observed phenotypic variance, $h^2 = \frac{\sigma_A^2}{\sigma_P^2}$.

2.2.2 Heritability of height

In 2003, a twin study estimated height H_{twin}^2 at 0.8 [52]. In 2006, a comparable result was obtained through genome-wide markers using the correlation between the amount of identical-by-descent (IBD) segments and the phenotype for 4,401 pairs of siblings estimating H_{fam}^2 to be 0.8 [53]. We consider here that those two studies estimated the broad-sense heritability. Indeed, twins or siblings share enough genomic regions to inherit entirely (twins) or partially (siblings) the genetic interaction component of the phenotypic variance [5]. In other words, the gene-gene interaction component is (at least partially) captured by the pedigree. Even if we would expect a lower heritability estimate from the study on siblings, sharing half of their genome, than from the study on twins, many parameters such as different populations, environmental factors, and technical biases have to be considered.

In 2008, a meta-analysis of three different GWAS [54, 55, 56] including 63,000 samples in total uncovered 40 loci associated with height, with 57 genome-wide significant markers. Surprisingly, these variants only explained 5% of the heritability of height ($h_{GWS}^2 = 0.05$), resulting in a call for the creation of a big consortium [57] to capture what was called the 'missing heritability of height'. The missing heritability was described as the dark matter of GWAS, detectable through family studies but invisible in the results of association studies [58]. Different hypotheses were proposed, such as a large number of variants with small effect size, the existence of untagged rare variants with larger effect size, poorly tagged structural variants, or the contribution of a gene-gene interaction component. In 2010, a GWAS with 183,727 samples estimated height h_{GWS}^2 at 0.12 with 180 loci [59]. It showed that most loci were in biologically relevant pathways such as skeletal growth; most associated variants were near likely causal genes; variants were enriched for likely functional effect on genes and; there was allelic heterogeneity. No convincing interaction was found by testing pairwise associations between genome-wide significant markers.

In 2010 the 'Restricted Maximum Likelihood' (REML) technique was proposed, which takes

advantage of dense genotyping data to exploit small differences in the proportion of genome shared between apparently unrelated individuals to estimate narrow-sense heritability [60]. On a cohort of 3,925 individuals, the single nucleotide polymorphism (SNP) based heritability h^2_{SNP} was estimated to be 0.45. REML estimates the narrow-sense heritability attributable to common SNPs because non-related individuals are unlikely to share most potential gene-gene interactions and therefore only expose the additive genetic component to linear models. It was speculated in the paper that the remaining missing heritability would be mostly due to incomplete linkage disequilibrium (LD) between markers and causative alleles. In 2011 the REML technique was included in a new tool dedicated to the characterization of complex traits, GCTA as GREML-SC (single component) [3]. In 2014, the GIANT consortium study powered by 253,288 samples estimated h^2_{GWS} at 0.16 with 697 genome-wide significant markers (associated with 423 loci) and h^2_{GWAS} with a significant threshold of 5×10^-3 at 0.29 with 9,500 markers, accounting for 60% of the variance attributable to all common SNPs [61].

In 2015, the new techniques GREML-LDMS (LD and MAF stratified) estimated h^2_{SNP} at 0.56 [62]. Because of the maximum possible LD correlation between two genetic variants declines as their difference in minor allele frequency (MAF) increases, the imputation quality can decrease from 97% for common variants to 68% for rare variants. Leveraging this parameter, it was estimated that h^2 could be as big as 0.6 - 0.7. Most of the 'missing heritability' described in 2010 was thus coming from many common variants with small effect size [60], while rare and low-frequency variants were estimated to account for 1.7% of heritability [63]. In addition, CNVs were reported to be associated with short stature [64] and recently reported to be associated with height with effect size as big as 2.4cm [65].

2.2.3 Heritability estimators

Over the past decade, our ability to estimate and interpret heritability of complex traits from population-level, genome-wide data has made great progress, yet many issues remain [66] [67]. In 2012, the GCTA GREML-SC heritability estimator was outperformed by LDAK which takes into account LD [68]. Indeed, GREML-SC overestimated heritability for causal variants in high LD regions and underestimated heritability for those in low LD regions. In 2015, LDAK was criticized as giving too much weight to rare variants, and GCTA GREML-LDMS, which conditions on LD and MAF, was introduced [62]. In 2017, LDAK was improved and heritability estimation now includes LD, MAF, and SNP genotyping quality [69]. Nonetheless, those techniques are assessed by simulation where the assumptions of the simulations follow the assumptions used to evaluate the quality of the models. Therefore, those results have to be taken carefully. Following extensive comparison based on whole-genome sequencing data, GREML-LDMS seems to be the least biased. Despite this, most of the literature report h^2_{SNP} from GREML-SC estimates [67].

In 2015, a technique called 'stratified LD score regression' was proposed, which allows heritability to be partitioned depending on categories made of subsets of SNPs [70]. Leveraging data from FANTOM [71], ENCODE [72], and relying on GWAS summary statistics, the authors found that conserved regions (which account for 2.6% of all SNPs) explained 35% of the heritability across all traits, and that the heritability per chromosome was proportional to its length. In September 2017, the method was updated from partitioning the heritability of SNPs based on binary annotations to continuous annotations. The authors investigated LD-dependent architecture and suggested considering the age of variant apparition rather than LD [73]. The rationale is that more recent variants have less LD and are more likely to have a high effect sizes because negative selection hasn't happened yet.

2.3 Evolution of complex traits understanding

2.3.1 History

In the early 1900s, a debate raged between Mendelians and Biometricians to explain the underlying rules explaining the distribution of continuous traits, as it was understood for Mendelians traits. This was largely solved by R.A. Fisher [1], who demonstrated that, if a quantitative trait results from the expression of multiple genes, the random drawing of alleles would result in a normally distributed phenotype within the population. As the number of genes would grow, their individual effect size would decrease. This laid the foundation of what would later become the infinitesimal model. Until year 2000, the number of genes implicated in any complex trait was unknown, but it was expected to be in the 10s or 100s.

2.3.2 An extended polygenicity

The first surprise came with the GWAS era, which revealed that significantly associated variants only explain a small fraction of heritability, implying that many more gene variants are implicated in complex traits [58]. It was estimated that 62% of all common SNPs are associated with a non-zero effect on height [2]. The second surprise came from the partition of heritability [74]. It revealed that the variance explained by each chromosome is proportional to its length, hinting to a uniform distribution of causal variants across the genome. It was estimated that 71 to 100% of 1Mb windows contribute to the risk of schizophrenia [75]. The third surprise came from fine mapping. While Mendelian diseases are mostly caused by variants resulting in protein-coding changes, complex traits seem to be mainly driven by variants mapping to regulatory regions [76, 77, 78]. Finally, associated SNPs are enriched in genes active in relevant cell type. There is little enrichment of genes specifically active in relevant cell type versus genes broadly active in every cell type, but there is a deprivation of genes inactive in relevant cell type [2]. While the missing heritability was partially solved when it was shown that common variants below the significant threshold contribute to complex traits [60], little

was understood regarding the presence of so many causally associated genes in seemingly irrelevant functions.

2.3.3 Toward the omnigenic model

Those observations led to the recent introduction of the so-called 'omnigenic model' as a new perspective to understand extreme polygenic traits [2]. Under this assumption, causal genes can be dichotomized in two categories. On the one hand, the core genes. They are causal because they are involved directly in the etiology of a trait. They are enriched in the most significant markers. Because they are a minority, they explain a modest part of heritability. On the other hand, the peripheral genes. They are causal because they are expressed in relevant cell types. In comparison to core genes, they are enriched in the lower tier of the significant markers. Because they constitute the majority, they explain the biggest part of heritability. The rationale is that because peripheral genes are expressed in a highly interconnected network of pathways, the 'small-world property' of network applies, and a change in the expression of a peripheral gene will have an impact on a core gene through a limited number of steps. This model also correlates with recent findings showing a high degree of pleiotropy between complex traits [79].

The omnigenic model was criticised in 2018[80]. Namely, it argues that the robustness of biological systems based on multiple backup cues that core genes can be much more, but lacks due to the current imperfect annotation of gene function. Furthermore, in the context of common diseases, there is no evidence that the type of genes enriched for rare variants merits special attention, as the effect size has little relationship to clinical relevance. They argue that the study design should not be redirected to whole-exome sequencing with the objective of identifying core genes.

3 Resource competition between genes is unlikely a driving force of complex trait architecture?

Olivier Naret[1], Yuval Simons[2], Jacques Fellay[1], Jonathan K Pritchard[2,3]

[1] School of Life Sciences, École Polytechnique Fédérale de Lausanne, Switzerland
[2] Department of Genetics, Stanford University, Stanford, CA 94305, USA
[3] Department of Biology, Stanford University, Stanford, CA 94305, USA

Most human complex traits are enormously polygenic, with tiny effects contributed by thousands of variants spread nearly uniformly across the genome. These observations raise a question about why so many variants–and so many genes–impact any given phenotype. Here we explore a possible explanatory model in which variant effects are due to an indirect mechanism, namely competition among genes for pools of shared intracellular resources such as ribosomes. To this end, we describe a simple theoretical model of resource competition during translation. We investigate whether the combined effects of thousands of eQTLs genome-wide can be expected to exert a meaningful influence on complex trait variation by altering the abundances of these shared resources. Our findings show that under most reasonable assumptions, resource competition should not be expected to have much impact on either the protein expression levels of individual genes, nor on complex trait outcomes. We conclude that, for most complex traits, resource competition is not a viable explanation for high polygenicity.

3.1 Introduction

Since the advent of genome-wide association studies around 15 years ago, there has been huge progress in determining the genetic basis of many human complex traits [51, 81, 82]. However, early studies found something perplexing: namely that the lead variants for any given trait typically explained only a small fraction of the heritability that had been predicted by family studies [56, 58]. This gap between the heritability accounted for by top variants, and the heritability observed in family studies, caused so much consternation that in 2008 it was referred to as the "mystery of missing heritability" [83].

This mystery was largely resolved when it was shown that most of the trait heritability comes from large numbers of common variants with very small effect sizes, whose signals fall far below genome-wide significance [84, 60]. Further work since then has shown that for many complex traits, there are on the order of 10^4 or even 10^5 variants across the genome that affect trait variance [85, 86, 87, 88]. Although there is heritability enrichment for coding variants, most of the heritability comes from non-coding variants impacting gene regulation [89, 76, 70]. These variants are spread surprisingly uniformly across the genome rather than being strongly concentrated near important genes or in particular chromosomal regions [75, 90]. Indeed the overall genetic architecture of most complex traits bears a striking resemblance to the classic infinitesimal model of quantitative genetics, developed starting from Fisher's 1919 paper [1].

However, such analyses also indicate another curious feature of the data. While the strongest GWAS signals are usually enriched near trait-relevant genes, in most cases these trait-relevant genes seem to contribute only a small fraction of the heritability [91, 61, 2, 92, 93]. For most diseases and other phenotypes, this conclusion is slightly tenuous as we generally have quite incomplete knowledge of the most relevant pathways; however the observation that the SNPs contributing heritability are spread relatively uniformly across the genome [75] implies that a large fraction of genes must be contributing to the trait variance [2].

Furthermore, in certain traits we do know much more about the molecular pathways that are directly involved in the trait biology, and detailed analysis of such traits is consistent with what was inferred from more complicated traits. For example, a recent paper from our group examined GWAS data for three molecular traits–urate, IGF-1, and testosterone–where a great deal is known about the relevant biological pathways [88]. Aside from one major effect locus for urate, that paper concluded that in aggregate the lead biological pathways for each trait only explain about 10% of the total SNP-based heritability. Instead, for all three traits, the bulk of the heritability comes from a large number of SNPs spread relatively uniformly across the genome: we estimated around 4,000-12,000 causal variants for the three molecular traits and 80,000 causal variants for height. Hence, paradoxically, for typical traits most of the heritability appears to act mainly through seemingly trait-irrelevant genes.

Why do so many genes affect trait variance? Thus, the resolution of the missing-heritability question leads to a second, and more mechanistic question: *Why are complex traits so enormously polygenic, and why do so many different genes affect trait variance?*

In two recent papers, our group proposed a simple quantitative model that we referred to as the "omnigenic" or sometimes "core gene" model, to explain this [2, 94]. Summarized very briefly, this model proposes that a modest fraction of all genes have direct effects on a phenotype of interest; these are referred to as "core genes". Meanwhile, all of the other genes expressed in trait-relevant cell types are referred to as "peripheral genes". While the peripheral genes do not exert direct effects on the trait, by definition, the expression levels of peripheral genes can have indirect effects on the trait via gene regulatory networks. Indeed, we propose that the large majority of the heritability actually flows through indirect trans-regulatory effects from periperhal genes.

This model is currently difficult to test directly due to our limited knowledge of gene networks and causal variants. However, our work with molecular traits strongly supports the conclusion that core genes typically contribute only small fractions of the heritability [88]. Recent work on correlations between polygenic scores for various traits and whole blood gene expression of likely core genes also supports the network component of the model [95].

Furthermore, we showed that there is a natural connection between our model and estimates of cis- and trans-heritability of gene expression [94]. Surveying work that measures gene expression heritability in a variety of cell types and species, we estimated that typically ~ 70% of gene expression heritability is due to trans regulation [94]. Since trans-eQTLs have very small effect sizes compared to cis-eQTLs, this implies that a typical gene must be regulated by very large numbers of trans-eQTLs–most of which lie far below the detection threshold for current studies. Based on the 70% estimate for trans heritability of expression, our model implies that peripheral genes can be expected to contribute ~ 70% to nearly 100% of the heritability for any given trait, depending on the number of core genes and their relative positions within the regulatory network.

It's worth noting that other types of effects also contribute to the observed architectures of complex traits but do not resolve the paradox of extreme polygenicity, and will not be considered in detail in this paper. First, many disease endpoints are impacted by multiple separate intermediate processes, each of which is polygenic in its own right. For example, diabetes risk is affected by adiposity, lipid levels and distribution, and liver function, each of which has a polygenic basis in its own right [96]. Thus, any variants that affect the intermediate processes can potentially be detected in GWAS of the endpoint trait [97, 98, 99, 78, 96]. While this hierarchical nature of traits certainly contributes to the high polygenicity of some disease endpoints, it seems unlikely to be a complete and general explanation given that virtually all complex traits show high polygenicity. To give just one counter-example, urate, which is controlled mainly by solute channels in the kidneys was estimated to have ~ 12,000 causal variants [88].

A second relevant effect is that selective constraint can play a "flattening" role on signals by lowering the allele frequencies of the large-effect variants [87]. This effect may contribute to the modest contributions of core genes, but does not help to explain the very high polygenicity of traits.

The role of resource competition in trans-regulation and heritability. In this paper we consider the role of a mechanism for trans regulation that is distinct from the network-based model that we considered previously [2, 94]. In the original phrasing of our model we assumed implicitly that the effects from peripheral genes are transmitted via specific molecular interactions in cell regulatory networks. One obvious regulatory interaction of this type is found in the relationship between transcription factors and their target genes, but any type of specific molecular interaction between genes that affects their expression would fit within that framework. However, in the present paper, we consider a completely different form of regulation, namely via resource competition.

The fundamental notion of the resource competition model is that each cell contains finite pools of shared molecules that are important for gene expression and regulation, including for example RNA PolII, nucleotides, tRNAs and ribosomes. Consider a cis-eQTL for a particular gene. If an individual carries the high-expressing genotype then we can expect this to very slightly reduce the number of ribosomes and other shared resources available to all other genes. Hence, resource limitation could mean that a cis-eQTL acts as a (very weak) trans-eQTL for every other gene.

We should clearly expect the net effect of any single cis-eQTL to be tiny, but what about in aggregate? We know that a large fraction of genes have cis-eQTLs [100]. **If there are** 10^4 **eQTLs in a cell type of interest, then could these in aggregate drive a meaningful effect on the variance of any given gene, or on the heritability of a trait controlled by that cell type?**

3.2 A model for intracellular resource competition

We study this question using a simple model of resource competition in a scenario of extreme resource scarcity. To make the model specific, we describe it in terms of competition for ribosomes, but competition for other types of resources would be modeled very similarly. We first examine the effect of the resource competition on the variation in protein level of a single gene. We then apply this model to a complex trait under the omnigenic model of Liu et al. Although we assume a relatively simple competition model for illustrative purposes, our results are generalizable. We discuss the effects of some extensions of our basic model and explain why our results remain robust.

Alternative to the last two lines: Although we assume a relatively simple competition model for illustrative purposes, our model's simplicity arises from assuming an extreme form of resource competition with no free ribosomes. Therefore, our results should hold for more

complex models with less stringent resource competition. Furthermore, our results stem not from model specifics but from order of magnitude arguments and should therefore hold quite generally, see discussion below.

The two main steps of gene expression are transcription and translation, and resource competition has previously been described for the translation step [101, 102]. In addition, competition between gene products can also occur at the level of downstream molecular mechanisms, such as the access of different protein species to a transport mechanism. Mathematically, all forms of resource competition are similar. However, to have a concrete example in mind, we focus on translation for two reasons. Firstly, it has been shown in Escherichia coli that during high growth, only 30% of the RNA polymerases are active against 80% of the ribosomes [101, 102]. Second, translation is the most energy-consuming intracellular process, which we consider as an approximation of resource consumption [103]. Competition during translation can come from different factors including ribosomes, tRNAs or translation factors. It has been shown that in Saccharomyces cerevisiae ribosomes are the main limiting factor [104].

A basic model of competition for ribosomes. Previous work has developed sophisticated quantitative models of translation [105, 106]. However, to focus more clearly on the parameters that are directly relevant here, we use a simpler model that gives the protein level of a particular gene relative to two random components: the basal mRNA level of that gene, and the mRNA level of the rest of the transcriptome.

If we assume that translation is limited by ribosomes and that ribosomes have equal affinity for each mRNA species, then the rate at which ribosomes bind to the mRNAs of a given gene is:

$$R_i = \frac{N_i}{\sum_k^m N_k} = \frac{N_i}{N_{tot}} \tag{3.1}$$

with R_i being the rate at which ribosomes bind to mRNAs of the i gene, N_i being the number of mRNA copies of the i gene, m being the number of genes, and $N_{tot} = \sum_k^m N_k$ defining the total number of mRNA copies. We assume that $m >> 1$. For simplicity, this expression ignores variation in translational efficiency across mRNAs; this would be incorporated as additional constant scaling factors for each term in Eq. 3.1.

At equilibrium, the protein level for the ith gene is then proportional to R_i:

$$P_i \propto R_i = \frac{N_i}{\sum_k^m N_k} = \frac{N_i}{N_{tot}} \tag{3.2}$$

with P_i being the protein level of the i gene. (Here we simplify the model by ignoring differences in decay rates between genes.) Without loss of generality, we set the constant proportionality to 1, which gives $P_i = R_i$. This is equivalent to changing the units of P_i.

We assume that N_i results from the cis-regulatory elements of the i gene. This assumption

means that $Cov[N_i, N_j] = 0$ for $i \neq j$. We will discuss the effects of the trans-regulatory elements later. In addition to the cis-regulatory effects, the transcriptome will have trans-pQTL effects on P_i through resource competition, which we define as the **competitive effect**. Therefore, the cis-eQTL of each gene is a trans-pQTL for all other genes by the competitive effect. This model is shown in Figure 3.1.

Alternative paragraph: We assume low levels of gene co-regulation, i.e. that $Cov[N_i, N_j] \approx 0$ for $i \neq j$. While this is true in general for gene pairs, core genes that affect a specific trait might be co-regulated [94]. We discuss this possibility later and explain why it only serves to further diminish the relative effect of resource competition.
In addition to regulatory effects, the transcriptome will have trans-pQTL effects on P_i through resource competition, which we define as the **competitive effect**. Therefore, an eQTL of a gene is a trans-pQTL for all other genes by the competitive effect. This model is shown in Figure 3.1.

We initially assume that the mean and variance of the level of expression are the same for all genes, and define $\mu_N \equiv E(N_i)$ and $V[N] \equiv V(N_i)$. We will discuss the implications of this hypothesis later.

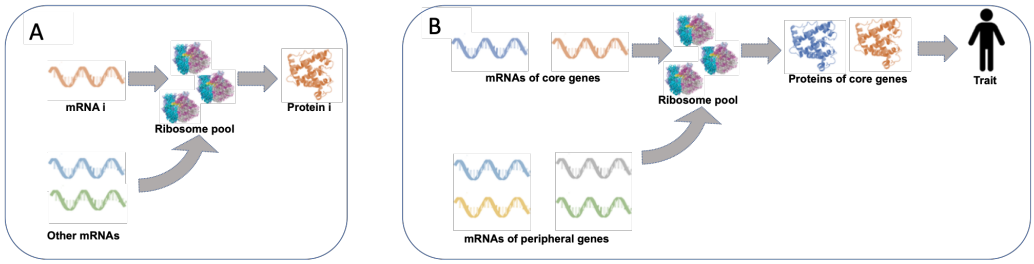


Figure 3.1 – **Illustration of the competition model for ribosomes. (A)** *Translation of a single gene may be affected by ribosome depletion due to translation of other genes.* **(B)** *This same model extends to complex traits by distinguishing between translation of core and peripheral genes.*

3.3 Effect of resource competition on the variance of a single gene

Before getting to the more-complicated model of a complex trait, we focus first on the effect of resource competition on protein expression of a single gene. Here we ask: **What proportion of the variance of a protein level comes from resource competition versus non-competitive effects?**

According to our model (3.2), the variance of protein level i is :

$$V[P_i] = V\left[\frac{N_i}{\sum_k^m N_k}\right] = V\left[\frac{N_i}{N_{tot}}\right]$$

The variance of a ratio can be approximated by a first-order Taylor expansion [107, 108] as:

$$V\left[\frac{f}{g}\right] \approx \frac{1}{E[g]^2}\left(V[f] - 2\frac{E[f]}{E[g]}Cov[f,g] + \frac{E[f]^2}{E[g]^2}V[g]\right). \tag{3.3}$$

Using this approximation, we can write:

$$V[P_i] \propto \underbrace{V[N_i]}_{basal} \underbrace{-2\frac{E[N_i]}{E[N_{tot}]}Cov[N_i,N_{tot}] + \frac{E[N_i]^2}{E[N_{tot}]^2}V[N_{tot}]}_{competition}. \tag{3.4}$$

Without loss of generality, we set the proportionality constant of the Equation 3.4 to 1, which is equivalent to measuring the levels of gene expression in units of the expected total expression level, $E[N_{tot}]$.

The first term on the right side of the Equation 3.4 reflects the variation in protein levels due to gene regulation and the second and third terms reflect the effects of resource competition. These two resource competition terms in this equation represent two opposite effects: (1) An increase in the expression level of gene i also leads to an increase in overall expression levels, leaving fewer ribosomes available for translation of gene i; and conversely for a decrease in expression level. This leads to a *reduction* in the variance of the protein as represented by the second term of the equation 3.4. (2) A fluctuation in the expression level of any gene leads to a fluctuation in the number of ribosomes available for the translation of the i gene and thus to an *increase* in the variance of the protein level as represented by the third term of the equation 3.4. As can be seen, the two effects of resource competition depend on the ratio of the mean mRNA level of the gene of interest to the mean sum of all genes. This suggests that if the gene of interest constitutes only a small fraction of the total mRNA, as is expected to be the case for almost all genes in almost all tissues for almost all organisms, resource competition has only a small effect on the variation in protein levels.

We can easily quantify this effect under the simplifying assumption that the expression levels of all genes are identically and independently distributed. In this case, Equation 3.4 is simplified to:

$$\begin{aligned} V[P_i] &= V[N] - 2\frac{1}{m}V[N] + \frac{1}{m^2}mV[N] \\ &= \underbrace{V[N]}_{basal} \underbrace{-\frac{1}{m}V[N]}_{competition} \end{aligned} \tag{3.5}$$

with $V[N]$ being the variance of the protein level of a gene. When all the gene expression patterns are identical, resource competition *actually leads to a reduction in protein variance*. However, this reduction is m times smaller than the baseline variance. This is because the relative change in the denominator of P_i (equation 3.2) is m times smaller than the relative change in the numerator. Therefore, resource competition will have only a negligible effect on variation in protein levels.

If we relax our hypothesis that genes are identically distributed, we find that the direction of

the overall effect of resource competition may differ between genes, although its magnitude remains small. We denote the mean and variance of gene expression in i as $E[N_i]$ and $V[N_i]$ and their mean values as $\bar{\mu} = \frac{1}{m}\sum_i E[N_i]$ and $\bar{V} = \frac{1}{m}\sum_i V[N_i]$. Equation 3.5 then becomes

$$\begin{aligned} V[P_i] &= V[N_i] - 2\frac{1}{m}\frac{E[N_i]}{\bar{\mu}}V[N_i] + \frac{1}{m}\frac{E[N_i]^2}{\bar{\mu}^2}\bar{V} \\ &= V[N_i]\left(1 \quad \underbrace{-2\frac{1}{m}\frac{E[N_i]}{\bar{\mu}} + \frac{1}{m}\frac{\bar{V}}{\bar{\mu}^2}\frac{E[N_i]^2}{V[N_i]}}_{competition}\right). \end{aligned} \tag{3.6}$$

The first resource competition term, representing the suppression of protein variation by the limited pool of ribosomes, remains small as long as the mean expression level of gene i is low relatively to overall gene expression, i.e. as long as $\frac{E[N_i]}{\bar{\mu}} << m$. The second term of resource competition, representing the increase in protein variation due to fluctuations in ribosome availability for gene i, remains relatively small as long as there is variability in the level of gene expression. However, if there is almost no variability in the expression level of gene i, i.e. if $\frac{V[N_i]}{E[N_i]} << \frac{\bar{V}}{2\bar{\mu}}$, then this term will be important since almost all variation in protein level will be due to resource competition. Interestingly, since the two terms of competition have opposite directions, the direction of the overall effect on protein level variation will vary between genes and will depend on the mean and variance of the gene expression level.

In Figures 3.2A and 3.2B, we show by simulations (see supplementary materials section 3.8.1) that the relative effect of resource competition is small and at a scale of $1/m$.

Because we have seen that, first, competition has little effect on the variance of protein level and, second, competition acts entirely in trans, if 70% of the heritability of gene expression comes from trans-regulation as estimated in Liu et al. [94], then it cannot come from competition alone.

Although the expression of most gene pairs is not highly correlated, if genes are subject to trans-regulation such that a single transcription factor binds to many cis-regulatory elements, and thus control the expression of many genes, they can co-vary. We estimated that the inclusion of such a behavior would have little impact on the effect of competition (see complementary documents, fig.3.6A and 3.6B). Though competition acts entirely in trans, our results suggest it cannot account for much of this trans-regulated variation in gene expression. This suggests that other mechanisms are responsible for the bulk of trans-regulated variation.

3.4 Effect of resource competition on the variance of complex traits

The second question we ask is: **what is the impact of competition on the phenotypic variance of a complex trait?**

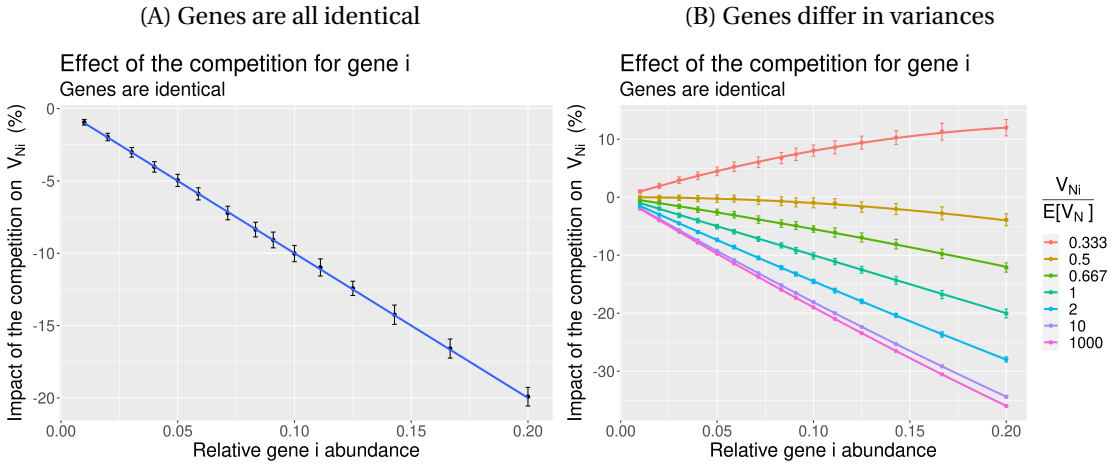


Figure 3.2 – **Impact of competition on protein level variance.** *Both figures represent that impact of the competition on the protein level variance.* **(A)** *Case where all genes are identical and in variable number m from 5 to 100.* **(B)** *Case where genes have heterogeneous expression patterns and in variable number m from 5 to 100. We overlay the simulation data (dots with an error bar) and the formula (line). Each point represents the average result of 100 simulations based on 10,000 samples with a 90% confidence interval.*

Phenotype expression. Following the omnigenic model laid out in Boyle et al. and Liu et al. [2, 94], we consider a number of c core proteins species that affect a trait, Y. We define the effect size per core protein level on the phenotype as γ, with γ_i being the effect size for the core protein species i. According to the Liu et al model, the phenotype is given by:

$$Y_i = \bar{Y} + \sum_{j=1}^{c} \gamma_j (P_{i,j} - \bar{P_j}) + \epsilon_{Y_i} \tag{3.7}$$

Next, incorporating variance from competition we have:

$$\begin{aligned} Y_i &= \bar{Y} + \sum_{j=1}^{c} \gamma_j \left(\frac{N_{i,j}}{N_{tot,i}} - \frac{\bar{N_j}}{\bar{N_{tot}}}\right) + \epsilon_{Y_i} \\ &= \bar{Y} + \frac{N_{\gamma,i}}{N_{tot,i}} - \frac{\bar{N_\gamma}}{\bar{N_{tot}}} + \epsilon_{Y_i} \end{aligned} \tag{3.8}$$

where we define, for convenience, $N_{\gamma,i} = \sum_{j=1}^{c} \gamma_j N_{j,i}$. The random error term ϵ_{Y_i} represents environmental and stochastic effects.

We now use our approximation (equation 3.3) to obtain an expression for the phenotypic variance:

$$V[Y] = \underbrace{V[N_\gamma]}_{basal} \underbrace{-2\frac{E[N_\gamma]}{E[N_{tot}]} Cov[N_\gamma, N_{tot}] + \frac{E[N_\gamma]^2}{E[N_{tot}]^2} V[N_{tot}]}_{competition} \tag{3.9}$$

This expression is identical in form to the equation 3.4 except that N_γ replaces N_i. As in the previous section (eq. 3.4), we can see that the effect of resource competition will depend crucially on the ratio between the averages of N_γ and N_{tot}. This ratio is different from the ratio for the variance of a single gene seen in the previous section in two ways: (1) it concerns the c number of core genes and not just one, and (2) each gene is associated to a γ values that can be either positive or negative.

To understand the effects of these qualitative differences, we turn, once again, to the very simple model where the expression levels of all genes are independent and have identical distributions with a mean μ_N and a variance $V[N]$. The equation 3.9 is then simplified to

$$\begin{aligned} V[Y] &= \overline{\gamma^2} c V[N] - 2\frac{\overline{\gamma} c}{m}\overline{\gamma} c V[N] + \frac{\overline{\gamma}^2 c^2}{m^2} m V[N] \\ &= \underbrace{\overline{\gamma^2} c V[N]}_{basal} \underbrace{- \overline{\gamma}^2 \frac{c}{m} c V[N]}_{competition} \end{aligned} \tag{3.10}$$

and we immediately see that, in comparison to what we saw in the previous section (eq. 3.5), there are now two distinct reasons for the competition term to be relatively small. The effect of resource competition will be small if either of these reasons is present:

(1) $\frac{c}{m} << 1$, i.e. the number of core genes is much smaller than the total number of genes. As with a single gene, this is because the core genes contribute only a small fraction of the overall mRNA, whose translation is limited by resource competition.

or

(2) $\overline{\gamma}^2 << \overline{\gamma^2}$, i.e. the mean effect size squared is much smaller than the mean squared effect size. If the effect of a gene on the trait is not correlated with the level of expression of the gene, then the mean effect will be very small because γ is equally likely to be positive or negative, making the effect of resource competition on the trait very small.

These two conditions are shown on Figures 3.3A and 3.3B, by simulations (see Supplement Section 3.8.2). The relative effect of resource competition is expected to be small, it scales with $\frac{c}{m}$ and is conditioned on $\frac{\overline{\gamma}^2}{\overline{\gamma^2}} \neq 0$. We consider it unlikely that both conditions are met.

This result should be valid even if the genes have different modes of expression and are trans-regulated. The ratio $\frac{E[N_\gamma]}{E[N_{tot}]}$ should range as $\frac{c}{m}$ even if the genes differ in their average expression, with small effects for resource competition if $\frac{c}{m} << 1$. The effects of heterogeneity in the variance of gene expression should only be significant if the core genes differ systematically in their variation from the peripheral genes. Since this is unlikely, the negative term of resource competition will generally be larger than the positive term, and thus resource competition will tend to (slightly) reduce the variance of the traits.

(A) Phenotype variance in function of core genes ratio for different $\frac{\overline{\gamma}^2}{\overline{\gamma^2}}$

(B) Phenotype variance in function of $\frac{\overline{\gamma}^2}{\overline{\gamma^2}}$ for different core genes ratio

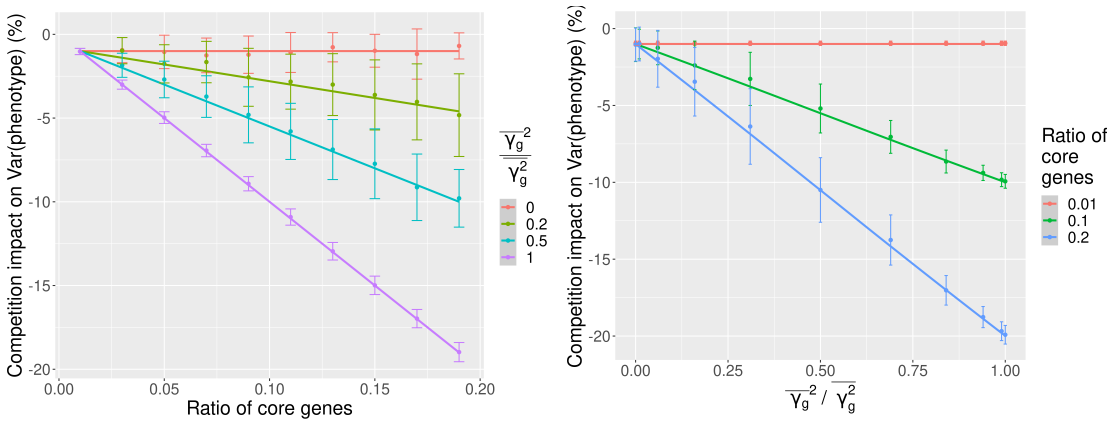


Figure 3.3 – **Impact of competition on phenotypic variance.** *The two figures represent the impact of resource competition on the variance of characteristics.* **(A)** *The ratio of core genes varies for different* $\frac{\overline{\gamma}^2}{\overline{\gamma^2}}$. **(B)** *the* $\frac{\overline{\gamma}^2}{\overline{\gamma^2}}$ *varies for different core gene ratios. Simulation data (dots with error bars) and formula (line) are superimposed. Each point represents the mean result within its 90% confidence interval for 100 simulations with 10,000 samples.*

Trans-regulation, which induces correlations between the expression of different genes, may have a significant effect here. As Liu et al [94] indicate, if the core genes tend to be co-regulated, trans effects can dominate the variance of a trait, thus inflating $V[N_\gamma]$ considerably compared to $Cov[N_\gamma, N_{tot}]$ and $V[N_{tot}]$. Therefore, in such a case, the relative effect of resource competition will be even smaller (see supplementary materials section 3.8.3, fig.3.6A and 3.6B).

Interestingly, though the overall effect of resource competition is to reduce trait variance, resource competition increases the relative contribution of trans-effect to variance (fig. TBA). Resource competition decreases cis-effects on trait variance because it dampens fluctuations in gene protein levels (middle term in eq. 3.9). On the other hand, resource competition increases trans-effects on trait variance since a fluctuation in the level of any gene affects the protein level of all other genes (last term in eq. 3.9). For most conditions, the decrease in cis-variance is large than the increase in trans-variance resulting in an overall reduction in trait variation.

3.5 Discussion

In this paper, we explored the possible contribution of resource competition to complex trait variance. Since different genes compete for the same cellular resources during transcription and translation, a variant affecting a single gene may affect the availability of cellular resources to all other genes. Therefore, resource competition is expected to increase the relative contri-

bution of trans-effects to variance in protein and trait levels. However, it is unclear, ab initio, if this could be a substantial effect.

We have presented a simple model of resource competition at translation between genes. We have shown that resource competition should only have a minor effect on variation in the protein level of any given gene, as long as that gene contributes only a small fraction of the overall pool of mRNAs. Similarly, if the core genes directly affecting a given complex trait only contribute a small fraction of the overall pool of mRNAs, resource competition would only have a minor effect on trait variation. Moreover, even if core genes do contribute most of the mRNA pool, resource competition would remain a small effect on trait variance unless trait values are strongly correlated (or anti-correlated) with the overall expression level or core genes.

Resource competition would be a large effect only taits whose core genes contribute a large fraction of the overall pool of mRNAs and the core genes's expression is correlated with trait values. While some traits may meet one of these two conditions, only a few traits should meet both. We don't know much about the expected number of core genes but it could be large for some traits. Even a trait that has a small number of core genes may locally have a large proportion of core genes in a specific tissue if expression is compartmentalized, or during specific periods of development. Despite this, most traits probably have a small number of core genes. For the second condition, we do not know of any category of traits that fulfills it, although it may exist. The general conclusion is that for the vast majority of traits, resource competition should have a negligible effect.

The work presented here is based on a very simple competition model, but we believe that the conclusions would hold with a more sophisticated model as it reproduces the expected trend but with competition having exacerbated effect. First because our model assumes extreme resource competition with no free ribosomes. Second, because our model assumes extreme affinity between mRNA and ribosome. In both cases, a more realistic model would have meant weaker competition and, ultimately, weaker effects of resource competition on the trait variance. We could also have modeled a differential binding affinity for the different mRNA species, which would have resulted in a weighting of the mRNA counts but would have left the conclusions intact.

Other forms of resource competition are expected to produce near-identical results. Strong competition between mRNAs during translation should give the same equation 3.2, regardless of whether the competition is for ribosomes, tRNAs, or other translation factors. Strong competition at the transcription level would result in a similar model, except that instead of mRNAs (the N_is), we would have a binding affinity to polymerase or a similar proxy for transcription rate. The models may also mix a few different forms of competition. However, the conclusions are not based on these details and would remain the same.

We have explored here resource competition as a possible contributing mechanism for trans-effects in complex trait heritability. Our model suggests that, for most traits, this is a small

effect. However, as sample sizes of QTL and eQTL mapping increases, even such a small effect may become meaningful. We have laid out a foundation for understanding the effects of resource competition on the architecture of expression, protein and trait level variance.

3.6 Acknowledgements

This work supported by NIH grants HG011432 and HG008140 to JKP, and HG011202 to YS. Much of this project was conducted during an extended visit by ON to the Pritchard lab at Stanford in 2019. We thank several people for suggesting the resource limitation model to us.

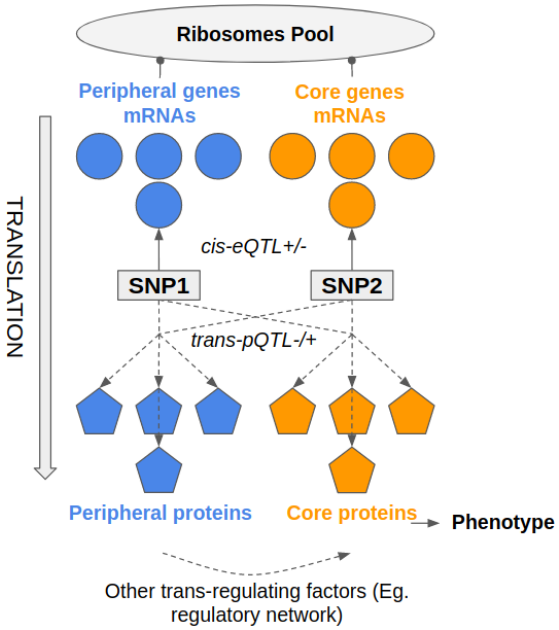
3.7 A partition of the phenotypic variance between core and peripheral genes expression

The last question we asked is, *how is the phenotypic variance partitioned between the expression of the peripheral and core genes?*

3.7.1 Competitive effect on and from the peripheral and core genes

In our model, the phenotype results entirely on the level of core proteins. But the expression of any gene - core or peripheral - is both generating and under the pressure of a **competitive effect**. For that reason, the heritability will be partitioned between the genetic determinants driving the level of expression of core and peripheral genes. Each of these genetic determinants are both cis-eQTL and trans-pQTL through the **competitive effect** (fig.3.4).

Figure 3.4 – Illustration of the competitive effect acting on core and peripheral genes.



SNP1 is a peripheral gene eQTL, SNP2 is a core gene eQTL. Both SNP1 and SNP2 are also pQTLs for all the other genes through competition. The quantity of the different core proteins - which determines the phenotype - species results from both peripheral and core genes eQTLs. Other mechanisms - like the densely connected regulatory network - can add up to the competitive effect as other sources of trans regulating mechanisms of the peripheral genes expression on core genes.

3.7.2 The effect size of the competitive effect

We define δ_n, the effect size of the **competitive effect** per unit of peripheral mRNA expressed in sample n. While within an individual δ_n is the same for all peripheral genes, it will differ between individuals. Therefore, δ it is a random variable of the individual population. See details in the supplementary materials section 3.8.7. We can describe multiple cases of the competitive effect with respect to δ

Omnigenic non-competitive case: if core genes effect sizes are not correlated ($E[\gamma] = 0$) and P and γ are distributed independently, the expected value of core genes overall effect is zero :

$$E[\delta] \propto E[\sum_{i}^{c} \gamma_i P_i] = 0$$

In such a case, the competitive effect is null.

Omnigenic competitive case: if core genes effect sizes are correlated ($E[\gamma] \neq 0$) and P and γ are distributed independently, the expected value of core genes overall effect different from zero :

$$E[\delta] \propto E[\sum_{i}^{c} \gamma_i P_i] \neq 0$$

In such a case, there is a competitive effect.

Another singular case is described in the supplementary materials section 3.8.7.

3.7.3 Partition of the phenotypic variance

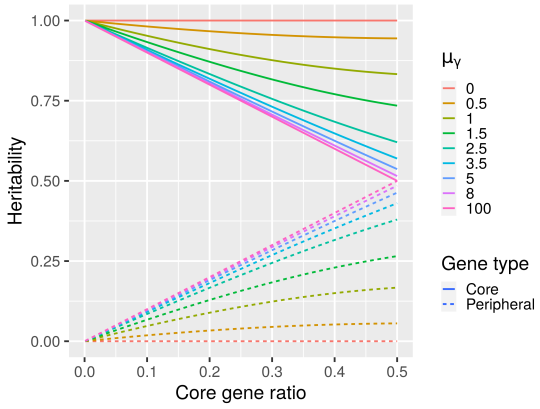
The resulting effect of the peripheral genes on the phenotype for sample n is $\delta_n \sum^p N_{p,n}$. Its variance, the genetic variance coming from peripheral genes V_{periph} is (details in the supplementary materials section 3.2):

$$V_{periph} = \mu_\gamma^2 \frac{r_c^2 r_p}{m\bar{\mu}^2} V[N] \tag{3.11}$$

By definition, $V_P = V_{periph} + V_{core}$. The fig. 3.5. represents the partition of the heritability in function of μ_γ and r_c or r_p between core and peripheral genes expression level.

In conclusion, a significant portion of the heritability can come from peripheral gene expressions due to the competitive effect at two conditions. First, there is a need for a relatively big quantity of core proteins being expressed. Second, $E[\gamma]$ must be not equal to 0.

Figure 3.5 – Partition of the heritability between the core genes and peripheral genes expression level, with respect to the ratio of core genes, and for different values of μ_γ.



When μ_γ is equal to zero there is no heritability coming from peripheral genes. When μ_γ increases there is a growing portion of the heritability that is transferred from core genes to peripheral genes until it reaches a maximum where all core genes share the same direction. While here μ_γ grows positively, we would observe the same thing with μ_γ growing negatively. When r_c increases, the heritability transferred to peripheral genes increases. We see that there is a maximum of heritability transferred increasing with both μ_γ and r_c

3.8 Supplementary materials

3.8.1 Simulation of protein level

To validate Eq.3.6 we simulated a dataset for 10,000 samples and a total of 100 n-genes. The N gene expression data were derived from a $\mathcal{N}$ normal distribution with $\mu_N = 1000$. For the homogeneous case (Fig. 3.2A) all genes have a standard deviation of $\sigma_N = 1$. For the heterogeneous case (Fig.3.2B) the gene of interest has different standard deviations such that $\sigma_{N_i} \in 0.001, 0.1, 0.5, 1, 1.5, 2, 3$ while the other genes have a standard deviation of $\sigma_N = 1$. The protein level P_n is obtained by applying our model (Eq. 3.2) such that $\forall i \; P_i = \frac{N_i}{\sum_j^m N_j}$ with an adjustment to make protein and gene expression at the same magnitude by $P_{i,adj} = P_i \sum_i^m E[N_i]$. $dV[P_i]$, the percentage of difference in variance induced by the competition is calculated as $dV[P_i] = \frac{V_{Pi,adj} - V_{Ni}}{V_{Ni}}$. For each set of parameters, the simulations were repeated 100 times in order to estimate the corresponding $\overline{dV[P_i]}$ and $\sigma_{dV[P_i]}$. The same parameters were introduced in the formula (Eq. 3.6) to compare the simulations to the theoretical results and validate our approach.

3.8.2 Simulation of phenotypes

A similar strategy was followed to validate Eq.3.10. On the same basis, we simulated 10,000 samples and a total of 100 n-genes. The N gene expression data were taken from a normal distribution $\mathcal{N}$ with $\mu_N = 1000$ and $\sigma_N = 1$ for all genes. As in the previous section, the protein level data P_n were obtained by applying 3.2 and fitted.

Here, in a second step, a base gene effect size is generated for a variable number of base genes c $c[1.20]$. We use different $\mu_\gamma \in 0, 0.25, 0.5, 1, 2, 4, 100$ for Fig. 3.3B and $\mu_\gamma \in 0, 0.11, 0.25, 0.43, 0.67, 1, 1.5, 2.33, 4, 9, 100$ for the figure 3.3A with a constant $\sigma_\gamma = 1$.

A phenotype is generated such that $Y_1 = \sum_i^c \gamma_i P_{i,adj}$ and the equivalent without competition $Y_0 = \sum_i^c \gamma_i N_i$. $dV[Y]$, the percentage of variance of the phenotype due to competition is calculated such that $dV[Y] = \frac{V_{Y_1} - V_{Y_0}}{V_{Y_0}}$. For each set of parameters, the simulations were repeated 100 times to estimate the corresponding $\overline{dV[Y]}$ and $\sigma_{dV[Y]}$. The same parameters were introduced into the formula (Eq. 3.10) in parallel to produce the theoretical results and validate our approach.

3.8.3 Trans-regulation

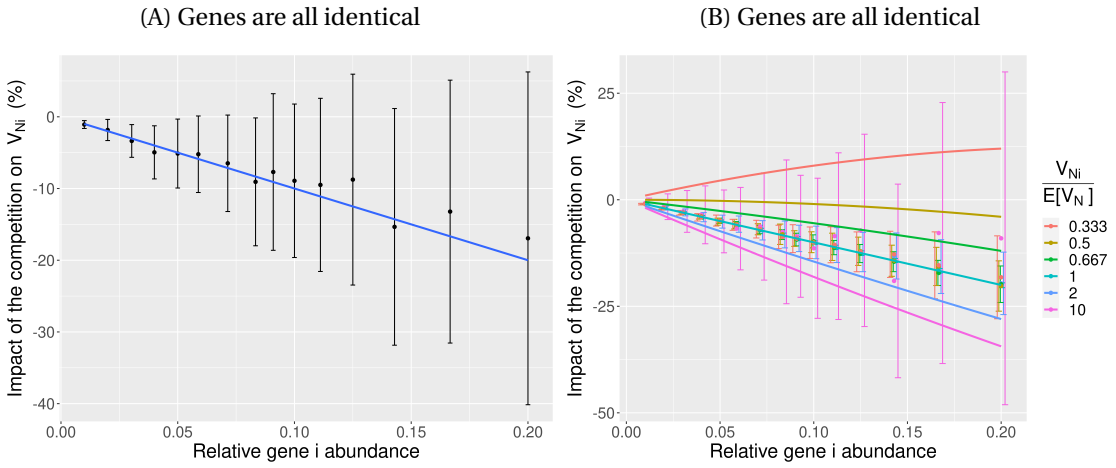
Considering that all possible $n \times n$ covariance matrices Σ can be expressed as

$$\Sigma = P' Diag(\sigma_1, \sigma_2, ..., \sigma_n) P$$

with P an orthogonal matrix and $\sigma_1 \geq \sigma_2 \geq ... \geq \sigma_n$. In this case, σ are main components pointing in the direction of the row of of P.

We generate a $n \times n$ matrix from a normal distribution $\mathcal{N}(0,1)$ on which we compute the QR decomposition to obtain P. We then compute the cross product to obtain Σ such that $\Sigma = P \times (\sigma_1, \sigma_2, ..., \sigma_n)$.

Figure 3.6 – Impact of competition on protein level variance with underlying widespread trans-regulation

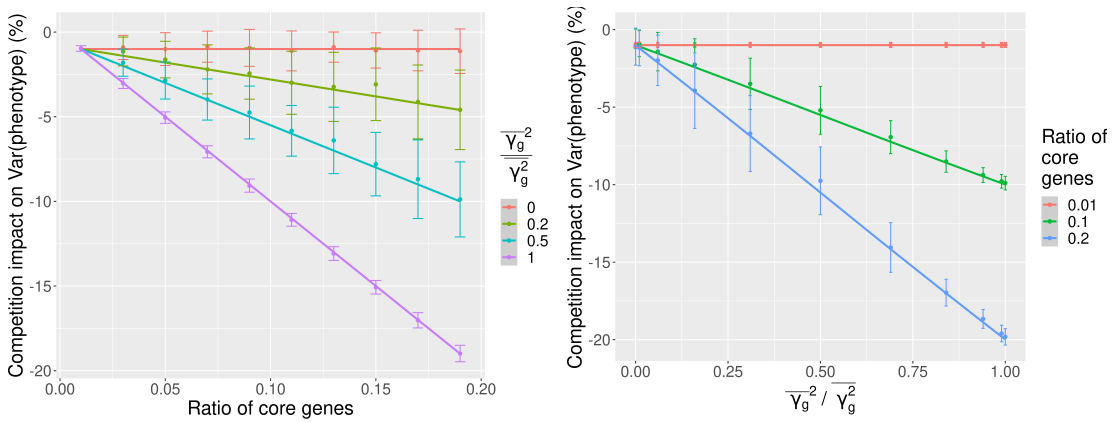


Both figures represent that impact of the competition on the protein level variance in a case with underlying widespread trans-regulation. The fig.3.6A shows a case where all genes are identical and in variable number m from 5 to 100. The fig.3.6B shows a case where genes have heterogeneous expression patterns and in variable number m from 5 to 100. We overlay the simulation data (dots with an error bar) and the formula (line). Each point represents the average result of 100 simulations based on 10,000 samples with a 90% confidence interval.

Figure 3.7 – Impact of competition on phenotypic variance with underlying widespread trans-regulation.

(A) Impact of competition on phenotypic variance in function of core genes ratio for different $\frac{\overline{\gamma}^2}{\gamma^2}$

(B) Impact of competition on phenotypic variance in function of $\frac{\overline{\gamma}^2}{\gamma^2}$ for different core genes ratio



The two figures represent the impact of resource competition on the variance of characteristics in a case with underlying widespread trans-regulation. On fig.3.7B, the ratio of core genes varies for different $\frac{\overline{\gamma}^2}{\gamma^2}$. On fig.3.7A, the $\frac{\overline{\gamma}^2}{\gamma^2}$ varies for different core gene ratios. Simulation data (dots with error bars) and formula (line) are superimposed. Each point represents the mean result within its 90% confidence interval for 100 simulations with 10,000 samples.

3.8.4 Variance of a single gene protein level value

Expression of the protein variance $V[P]$

Introduction

Considering our model 3.2, the variance of the protein level is as:

$$V[P_i] = V[\frac{N_i}{\sum_k^m N_k}]$$

The variance of a ratio of random variables can be approximated with Taylor expansion series. With $\theta = (\mu_1, ..., \mu_m)$ for expansion point we get:

$$\begin{aligned} P_i &= P(\theta;i) + P'_{N_i}(\theta;i)(N_i - \theta_i) + \sum_k^m P'_{N_k;i}(\theta;i)(N_k - \theta_k) + R \\ &\approx P(\theta;i) + P'_{N_i}(\theta;i)(N_i - \theta_i) + \sum_k^m P'_{N_k;i}(\theta;i)(N_k - \theta_k) \end{aligned} \tag{3.12}$$

The general expression of the variance :

$$V[P_i] = E[(P_i - E[P_i])^2] \tag{3.13}$$

Approximation of $E[P_i]$

$$E[P_i] = E[P(\theta;i) + P'_{N_i}(\theta;i)(N_i - \theta_i) + \sum_{k\neq i}^{m} P'_{N_k;i}(\theta;i)(N_k - \theta_k)]$$

With ∀ i, $E[P'_{N_i}(\theta;i)(N_i - E[N_i])] = 0$

$$E[P_i] = E[P(\theta;i)] = P(\theta;i) \tag{3.14}$$

With

$$P(\theta;i) = \frac{E[N_i]}{\sum_k^m E[N_k]}$$

Approximation of $V[P_i]$

First-order Taylor series formulation: If we plug (3.12) and (3.14) in (3.13):

$$\begin{aligned}
V[P_i] &= E[(P(\theta;i) + (P'_{N_i}(\theta;i)(N_i - \theta_i) + \sum_{k\neq i}^{m} P'_{N_k;i}(\theta;i)(N_k - \theta_k))^2 - P(\theta;i))^2] \\
&= E[(P'_{N_i}(\theta;i)(N_i - \theta_i) + \sum_{k\neq i}^{m} P'_{N_k;i}(\theta;i)(N_k - \theta_k))^2] \\
&= E[(P'^2_{N_i}(\theta;i)(N_i - E[N_i])^2 + \sum_{k\neq i}^{m} P'^2_{N_k;i}(\theta;i)(N_k - E[N_k])^2 + \\
&\sum_{k1\neq i}^{m} \sum_{k2=k1+1;\neq i}^{m} 2P'_{N_{k1}}(\theta;i)(N_{k1} - \mu_{k1})P'_{N_{k2}}(\theta;i)(N_{k2} - \mu_{k2}))^2]
\end{aligned}$$

Substituting the expression of Variance and Covariance:

$$\begin{aligned}
&= P'^2_{N_i}(\theta;i)V[N_i] + \sum_{k\neq i}^{m} P'^2_{N_k;i}(\theta;i)V[N_k] \\
&+ \sum_{k1\neq i}^{m} \sum_{k2=k1+1;\neq i}^{m} 2P'_{N_{k1}}(\theta;i)P'_{N_{k2}}(\theta;i)Cov(N_{k1}, N_{k2})
\end{aligned}$$

If mRNAs levels are uncorrelated, we get:

$$V[P_i] = P'^2_{N_i}(\theta;i)V[N_i] + \sum_{k\neq i}^{m} P'^2_{N_k;i}(\theta;i)V[N_k] \tag{3.15}$$

For the transcriptome, $E[N_T]$ is the sum of the m mRNA species means such as $E[N_T] = \sum_i^m E[N_i]$, and $\sum_k^m V[N_k]$ is the sum of the m mRNA variances such as $\sum_k^m V[N_k] = \sum_i^m V[N_i]$.

$$P'_{N_i}(\theta;i) = \frac{\sum_{k\neq i}^m E[N_k]}{(\sum_k^m E[N_k])^2} = \frac{E[N_T] - E[N_i]}{E[N_T]^2} \tag{3.16}$$

$$P'_{N_k;i}(\theta;i) = -\frac{RE[N_i]}{T(\sum_k^m E[N_k])^2} = -\frac{E[N_i]}{E[N_T]^2} \tag{3.17}$$

And the squared versions:

$$P'^2_{N_i}(\theta;i) = \frac{(\sum_{k\neq i}^m E[N_k])^2}{(\sum_k^m E[N_k])^4} = \frac{(E[N_T] - E[N_i])^2}{(E[N_T])^4}$$
$$P'^2_{N_k;i}(\theta;i) = \frac{E[N_i]^2}{(\sum_k^m E[N_k])^4} = \frac{E[N_i]^2}{(E[N_T])^4}$$

Developed formulation:

Substituting those in (3.15) we get:

$$\begin{aligned} V[P_i] &= \frac{(E[N_T] - E[N_i])^2}{E[N_T]^4} V[N_i] + \sum_{k\neq i}^{m} \frac{E[N_i]^2}{(E[N_T])^4} V[N_k] \\ &= \frac{1}{E[N_T]^4}[V[N_i](E[N_T] - E[N_i])^2 + E[N_i]^2(\sum_k^m V[N_k] - V[N_i])] \end{aligned}$$

Finally we have:

$$V[P_i] = (\frac{1}{E[N_T]})^2[V[N_i]\frac{(E[N_T] - E[N_i])^2}{E[N_T]^2} + E[N_i]^2\frac{\sum_k^m V[N_k] - V[N_i]}{E[N_T]^2}]$$

If we express both protein and mRNA in unit of phenotype change it become proportional

$$\begin{aligned} &\propto = V[N_i]\frac{(E[N_T] - E[N_i])^2}{E[N_T]^2} + E[N_i]^2\frac{\sum_k^m V[N_k] - V[N_i]}{E[N_T]^2} \\ &\propto V[N_i] - 2\frac{E[N_i]}{E[N_T]}V[N_i] + \frac{E[N_i]^2}{E[N_T]^2}\sum_k^m V[N_k] \end{aligned}$$

We mark the mean and variance of the gene expression of gene i as $E[N_i]$ and $V[N_i]$ and their mean values as $\bar{\mu} = \frac{1}{m}\sum_i E[N_i]$ and $\bar{V} = \frac{1}{m}\sum_i V[N_i]$

With $\frac{\sum_k^m V[N_k]}{E[N_T]} = \frac{\overline{V}}{\overline{\mu}}$

$$\boxed{V[P_i] \propto \underbrace{V[N_i]}_{basal} \underbrace{-2\frac{1}{m}\frac{E[N_i]}{\overline{\mu}}V[N_i] + \frac{1}{m}\frac{E[N_i]^2}{\overline{\mu}^2}\overline{V}}_{competition}} \tag{3.18}$$

If all the genes expression are i.i.d:

To compare the relative importance of each of these two components we consider that all the genes are identically distributed and independent.

$$\begin{aligned} E[N_i] &= \overline{\mu} \\ V[N_i] &= V[N] \\ E[N_T] &= \sum_k^m E[N_k] = m\overline{\mu} \\ \sum_k^m V[N_k] &= mV[N] \end{aligned} \tag{3.19}$$

Then in substituting these in (3.18).

$$\begin{aligned} V[P] &= V[N] - 2\frac{\overline{\mu}}{m\overline{\mu}}V[N] + \frac{\overline{\mu}^2}{(m\overline{\mu})^2}mV[N] \\ &= V[N] - 2\frac{1}{m}V[N] + \frac{1}{m^2}mV[N] \end{aligned}$$

$$\boxed{V[P] = \underbrace{V[N]}_{basal} - \underbrace{\frac{1}{m}V[N]}_{competition}} \tag{3.20}$$

3.8.5 Covariance between protein level value

The phenotype is then given by

$$Y = \bar{Y} + \sum_{i=1}^{c} \gamma_i P_i \tag{3.21}$$

From the variance of one protein level, we know investigate the phenotypic variance.

$$\begin{aligned} V[Y] &= V[\sum_{i}^{c} \gamma_i P_i] \\ &= \sum_{i}^{c} V[\gamma_i P_i] + \sum_{i}^{c} \sum_{j \neq i}^{c} C[\gamma_i P_i, \gamma_j P_j] \\ &\text{Because } \gamma \text{ is fixed per gene, it behaves as a constant} \\ &= \sum_{i}^{c} \gamma_i^2 V[P_i] + \sum_{i}^{c} \sum_{j \neq i}^{c} \gamma_i \gamma_j C[P_i, P_j] \end{aligned} \tag{3.22}$$

Because of the competition during translation, the protein levels are not independent variables.

We investigate the covariance term.

$$\begin{aligned} C[P_i, P_j] &= E[(P_i - E[P_i])(P_j - E[P_j])] \\ &= E[P_i P_j] - E[P_i]E[P_j] \end{aligned}$$

Approximation to the first order

Approximation of $E[P_i P_j]$

Similarly to the earlier case, we are dealing with a ratio of random variables and have to use Taylor series expansion. With $\theta = (E[N_i], E[N_j], ..., E[N_T])$, the first-order Taylor series is:

$$\begin{aligned} P_i P_j &= g(N; i, j) \\ &= \frac{N_i N_j}{(\sum_k^m N_k)^2} \end{aligned}$$

$$\begin{aligned} g(N; i, j) = g(\theta; i, j) &+ g'_{N_i}(\theta; i, j)(N_i - E[N_i]) + g'_{N_j}(\theta; i, j)(N_j - E[N_j]) + \\ &\sum_{k}^{m} g'_{N_k}(\theta; i, j)(N_k - E[N_k]) \end{aligned}$$

We know from the first part that the first-order Taylor series approximation for a function is

such as $E[g(N;i,j)] = g(\theta;i,j)$

$$g(\theta;i,j) = \frac{E[N_i]E[N_j]}{(\sum_k^m E[N_k])^2} = \frac{E[N_i]E[N_j]}{E[N_T]^2} \tag{3.23}$$

$$\begin{aligned} C[P_i,P_j] &= E[g(N;i,j)] - E[P_i]E[P_j] \\ &= \frac{E[N_i]E[N_j]}{(E[N_T])^2} - \frac{E[N_i]E[N_j]}{(E[N_T])^2} \\ &= 0 \end{aligned}$$

Henceforth, we need to go one step further and extend our formulation to the second-order.

Approximation to the second order

Approximation of $E[P_i]$

With P' the first order terms.

$$\begin{aligned} P_i - P' = P(\theta;i) + \frac{1}{2}(P''_{N_iN_i}(\theta;i)(N_i - E[N_i])^2 + \sum_{k\neq i}^{m} P''_{N_kN_k;i}(\theta;i)(N_k - E[N_k])^2 + \\ \sum_{i}^{m}\sum_{k=i+1}^{m} 2P''_{N_iN_k}(\theta;i,j)(N_i - E[N_i])(N_k - E[N_k])) \end{aligned}$$

$$\begin{aligned} E[P_i] = P(\theta;i) + \frac{1}{2}(P''_{N_iN_i}(\theta;i)V[N_i] + \sum_{k\neq i}^{m} P''_{N_kN_k;i}(\theta;i)V[N_k] + \\ \sum_{i}^{m}\sum_{k=i+1}^{m} 2P''_{N_iN_k}(\theta;i,j)C[N_i,N_k]) \end{aligned} \tag{3.24}$$

With the covariance terms for genes expression being 0 it becomes:

$$= P(\theta;i) + \frac{1}{2}(P''_{N_iN_i}(\theta;i)V[N_i] + \sum_{k\neq i}^{m} P''_{N_kN_k;i}(\theta;i)V[N_k])$$

Approximation of $E[P_i]E[P_j]$

Using (3.24)

$$\begin{aligned}E[P_i]E[P_j] &= (P(\theta;i)+\frac{1}{2}(P''_{N_iN_i}(\theta;i)V[N_i]+\sum_{k\neq i}^{m}P''_{N_kN_k;i}(\theta;i)V[N_k]))\\ &\quad (P(\theta;j)+\frac{1}{2}(P''_{N_jN_j}(\theta;j)V[N_j]+\sum_{k\neq j}^{m}P''_{N_kN_k;j}(\theta;j)V[N_k]))\\ &= P(\theta;i)P(\theta;j)+\frac{1}{2}(P(\theta;i)P''_{N_jN_j}(\theta;j)V[N_j]+P(\theta;j)P''_{N_iN_i}(\theta;i)V[N_i]+\\ &\quad P(\theta;i)P''_{N_kN_k;j}(\theta;j)\sum_{k\neq j}^{m}V[N_k]+P(\theta;j)P''_{N_kN_k;i}(\theta;i)\sum_{k\neq i}^{m}V[N_k])\end{aligned} \tag{3.25}$$

Approximation of $E[P_iP_j]$

$$\begin{aligned}E[P_iP_j] &= E[(P(\theta;i)+P'_{N_i}(\theta;i)(N_i-\theta_i)+\sum_{k\neq i}^{m}P'_{N_k;i}(\theta;i)(N_k-\theta_k)+\\ &\quad \frac{1}{2}(P''_{N_iN_i}(\theta;i)(N_i-\theta_i)^2+\sum_{k\neq i}^{m}P''_{N_kN_k;i}(\theta;i)(N_k-\theta_k)^2))\\ &\quad (P(\theta;j)+P'_{N_j}(\theta;j)(N_j-\theta_j)+\sum_{k\neq j}^{m}P'_{N_k;j}(\theta;j)(N_k-\theta_k)+\\ &\quad \frac{1}{2}(P''_{N_jN_j}(\theta;j)(N_j-\theta_j)^2+\sum_{k\neq j}^{m}P''_{N_kN_k;j}(\theta;j)(N_k-\theta_k)^2))]\end{aligned}$$

... Considering only max second-order terms:

$$\begin{aligned}&= E[P(\theta;i)[P(\theta;j)+P'_{N_j}(\theta;j)(N_j-\theta_j)+\sum_{k\neq i}^{m}P'_{N_k;i}(\theta;i)(N_k-\theta_k)+\\ &\frac{1}{2}(P''_{N_jN_j}(\theta;j)(N_i-\theta_j)^2+\sum_{k\neq j}^{m}P''_{N_kN_k;j}(\theta;j)(N_k-\theta_k)^2)]+\\ &P'_{N_i}(\theta;i)(N_i-\theta_i)[P(\theta;j)+P'_{N_j}(\theta;j)(N_j-\theta_j)+\sum_{k\neq j}^{m}P'_{N_k;j}(\theta;j)(N_k-\theta_k)]]+\\ &\sum_{k\neq i}^{m}P'_{N_k;i}(\theta;i)(N_k-\theta_k)[P(\theta;j)+P'_{N_j}(\theta;j)(N_j-\theta_j)+\sum_{k\neq j}^{m}P'_{N_k;j}(\theta;j)(N_k-\theta_k)]+\\ &\frac{1}{2}(P''_{N_iN_i}(\theta;i)(N_i-\theta_i)^2+\sum_{k\neq i}^{m}P''_{N_kN_k;i}(\theta;i)(N_k-\theta_k)^2)[P(\theta;j)]]\end{aligned}$$

... Removing the first order which will be subtracted as seen at 3.23:

$$
\begin{aligned}
&= E[P(\theta;i)[P(\theta;j) + \frac{1}{2}(P''_{N_jN_j}(\theta;j)(N_j-\theta_j)^2 + \sum_{k\neq j}^{m} P''_{N_kN_k;j}(\theta;j)(N_k-\theta_k)^2)]+ \\
&P'_{N_i}(\theta;i)(N_i-\theta_i)[P'_{N_j}(\theta;j)(N_j-\theta_j) + \sum_{k\neq j}^{m} P'_{N_k;j}(\theta;j)(N_k-\theta_k)]]+ \\
&\sum_{k\neq i}^{m} P'_{N_k;i}(\theta;i)(N_k-\theta_k)[P'_{N_j}(\theta;j)(N_j-\theta_j) + \sum_{k\neq j}^{m} P'_{N_k;j}(\theta;j)(N_k-\theta_k)]+ \\
&\frac{1}{2}(P''_{N_iN_i}(\theta;i)(N_i-\theta_i)^2 + \sum_{k\neq i}^{m} P''_{N_kN_k;i}(\theta;i)(N_k-\theta_k)^2)[P(\theta;j)]]
\end{aligned}
$$

... Developing:

$$
\begin{aligned}
&= P(\theta;i)P(\theta;j) + \frac{1}{2}(P(\theta;i)P''_{N_jN_j}(\theta;j)V[N_j] + P(\theta;i)P''_{N_kN_k;j}(\theta;j)\sum_{k\neq j}^{m} V[N_k]+ \\
&P(\theta;j)P''_{N_iN_i}(\theta;i)V[N_i] + P(\theta;j)P''_{N_kN_k;i}(\theta;i)\sum_{k\neq i}^{m} V[N_k])+ \\
&P'_{N_i}(\theta;i)P'_{N_j}(\theta;j)C[N_i,N_j] + P'_{N_i}(\theta;i)P'_{N_k;j}(\theta;j)\sum_{k\neq j}^{m} C[N_i,N_k]+ \\
&P'_{N_j}(\theta;j)P'_{N_k;i}(\theta;i)\sum_{k\neq i}^{m} C[N_j,N_k] + P'_{N_k;j}(\theta;j)P'_{N_k;i}(\theta;i)\sum_{k\neq j}^{m}(N_k-\theta_k)\sum_{k\neq i}^{m}(N_k-\theta_k)
\end{aligned}
\tag{3.26}
$$

General formulation

Using (3.25) and (3.26)

$$
\begin{aligned}
C[P_i,P_j] &= E[P_iP_j] - E[P_i]E[P_j] \\
&= P'_{N_i}(\theta;i)P'_{N_j}(\theta;j)C[N_i,N_j]+ \\
&P'_{N_i}(\theta;i)P'_{N_k;j}(\theta;j)\sum_{k\neq j}^{m} C[N_i,N_k] + P'_{N_j}(\theta;j)P'_{N_k;i}(\theta;i)\sum_{k\neq i}^{m} C[N_j,N_k]+ \\
&P'_{N_k;j}(\theta;j)P'_{N_k;i}(\theta;i)\sum_{k\neq j}^{m}(N_k-\theta_k)\sum_{k\neq i}^{m}(N_k-\theta_k)
\end{aligned}
$$

$$
\begin{aligned}
C[P_i,P_j] = {} & P'_{N_i}(\theta;i)P'_{N_j}(\theta;j)C[N_i,N_j]+ \\
& \Big| P'_{N_i}(\theta;i)P'_{N_k;j}(\theta;j)V[N_i] + P'_{N_j}(\theta;j)P'_{N_k;i}(\theta;i)V[N_j] + \Big| \\
& \Big| P'_{N_k;j}(\theta;j)P'_{N_k;i}(\theta;i)\sum_{k\neq i\&j}^{m} V[N_k] + \Big| \\
& P'_{N_i}(\theta;i)P'_{N_k;j}(\theta;j)\sum_{k\neq i\&j}^{m} C[N_i,N_k] + P'_{N_j}(\theta;j)P'_{N_k;i}(\theta;i)\sum_{k\neq i\&j}^{m} C[N_j,N_k]+ \\
& P'_{N_k;j}(\theta;j)P'_{N_k;i}(\theta;i)(\sum_{k1\neq i\&j}^{m}\sum_{k2=k1+1;\neq i\&j}^{m} 2C[N_{k1},N_{k2}]+ \\
& \sum_{k\neq i}^{m} C[N_i,N_k] + \sum_{k\neq j}^{m} C[N_j,N_k] + C[N_i,N_j])
\end{aligned}
\tag{3.27}
$$

If no covariance/co-regulation (the outside of the dashed box is zeroed):

$$
\begin{aligned}
C[P_i,P_j] = {} & P'_{N_i}(\theta;i)P'_{N_k;j}(\theta;j)V[N_i] + P'_{N_j}(\theta;j)P'_{N_k;i}(\theta;i)V[N_j]+ \\
& P'_{N_k;i}(\theta;i)P'_{N_k;j}(\theta;j)\sum_{k\neq i\&j}^{m} V[N_k]
\end{aligned}
$$

Using (3.16) and (3.17)

$$
\begin{aligned}
C[P_i,P_j] = {} & \frac{1}{E[N_T]^4}(E[N_i]E[N_j]\sum_{k\neq i\&j}^{m} V[N_k]- \\
& (E[N_T]-E[N_i])E[N_j]V[N_i] - (E[N_T]-E[N_j])E[N_i]V[N_j]) \\
& \propto E[N_i]E[N_j](\sum_{k}^{m} V[N_k] - V[N_i] - V[N_j]) \\
& - (E[N_T]-E[N_i])E[N_j]V[N_i] - (E[N_T]-E[N_j])E[N_i]V[N_j] \\
& \propto E[N_i]E[N_j]\sum_{k}^{m} V[N_k] - E[N_T]E[N_j]V[N_i] - E[N_T]E[N_i]V[N_j] \\
& \propto \frac{\sum_k^m V[N_k]}{E[N_T]} - \frac{V[N_i]}{E[N_i]} - \frac{V[N_j]}{E[N_j]}
\end{aligned}
$$

With coefficient of proportionality similar to 3.18:

$$
\propto \frac{E[N_i]E[N_j]}{E[N_T]}(\frac{\sum_k^m V[N_k]}{E[N_T]} - \frac{V[N_i]}{E[N_i]} - \frac{V[N_j]}{E[N_j]})
$$

$$\boxed{C[P_i,P_j] \propto \frac{1}{m}\frac{E[N_i]E[N_j]}{\overline{\mu}}(\frac{\overline{V}}{\overline{\mu}} - \frac{V[N_i]}{E[N_i]} - \frac{V[N_j]}{E[N_j]})} \tag{3.28}$$

If all the genes expression are i.i.d.

$$C[P_i,P_j] \quad \propto \frac{\overline{\mu}^2}{m\overline{\mu}}(\frac{mV[N]}{m\overline{\mu}} - 2\frac{V[N]}{\overline{\mu}})$$

$$\boxed{C[P_i,P_j] \propto -\frac{1}{m}V[N]} \tag{3.29}$$

3.8.6 Phenotypic variance

General formulation:

$$\begin{aligned}
V[Y] &= V[\sum_i^c \gamma_i P_i] \\
&= \sum_i^c V[\gamma_i P_i] + \sum_i^c \sum_{j\neq i}^c C[\gamma_i P_i, \gamma_j P_j] \\
&\text{Because } \gamma \text{ is fixed per gene, it behaves as a constant} \\
&= \sum_i^c \gamma_i^2 V[P_i] + \sum_i^c \sum_{j\neq i}^c \gamma_i\gamma_j C[P_i,P_j] \\
&= \sum_i^c \gamma_i^2 (V[N_i] - 2\frac{1}{m}\frac{E[N_i]}{\overline{\mu}}V[N_i] + \frac{1}{m}\frac{E[N_i]^2}{\overline{\mu}^2}\overline{V}) + \\
&\sum_i^c \sum_{j\neq i}^c \gamma_i\gamma_j(\frac{1}{m}\frac{E[N_i]E[N_j]}{\overline{\mu}}(\frac{\overline{V}}{\overline{\mu}} - \frac{V[N_i]}{E[N_i]} - \frac{V[N_j]}{E[N_j]}))
\end{aligned}$$

$$\boxed{\begin{aligned}
V[Y] = &\underbrace{\sum_i^c \gamma_i^2 V[N_i]}_{basal} + \underbrace{\sum_i^c \gamma_i^2 (-2\frac{1}{m}\frac{E[N_i]}{\overline{\mu}}V[N_i] + \frac{1}{m}\frac{E[N_i]^2}{\overline{\mu}^2}\overline{V}) + }_{competition} \\
&\underbrace{\sum_i^c \sum_{j\neq i}^c \gamma_i\gamma_j(\frac{1}{m}\frac{E[N_i]E[N_j]}{\overline{\mu}}(\frac{\overline{V}}{\overline{\mu}} - \frac{V[N_i]}{E[N_i]} - \frac{V[N_j]}{E[N_j]}))}_{competition}
\end{aligned}} \tag{3.30}$$

If all the genes expression are i.i.d:

Reusing (3.20) and (3.29)

$$
\begin{aligned}
E[\gamma]^2 &= \mu_\gamma^2 \\
E[\gamma^2] &= \overline{\gamma^2} = E[\gamma]^2 + V[\gamma] \\
E[\gamma^2] &= \mu_\gamma^2 + \sigma_\gamma^2 \frac{(c-1)}{c} \\
E[\gamma_i \gamma_j] &= \overline{\gamma^2} - \sigma_\gamma^2 \\
E[\gamma_i \gamma_j] &= \mu_\gamma^2 - \frac{\sigma_\gamma^2}{c}
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
\sum_i^c \gamma_i^2 V[P_i] &= \sum_i^c E[\gamma^2] E[V[P]] \\
&= c\overline{\gamma^2}(1 - \frac{1}{m}) V[N] \\
&= [c\mu_\gamma^2 + \sigma_\gamma^2(c-1)](1 - \frac{1}{m}) V[N]
\end{aligned}
$$

$$
\begin{aligned}
\sum_i^c \sum_{j \neq i}^c \gamma_i \gamma_j C[P_i, P_j] &= \sum_i^c \sum_{j \neq i}^c E[\gamma_i \gamma_j] E[C[P_i, P_j]] \\
&= -c(c-1)(\overline{\gamma^2} - \sigma_\gamma^2) \frac{1}{m} V[N] \\
&= -c(c-1)(\mu_\gamma^2 - \frac{\sigma_\gamma^2}{c}) \frac{1}{m} V[N]
\end{aligned}
$$

$$
\begin{aligned}
V[Y] &= E[\sum_i^c \gamma_i^2 V[P_i]] + E[\sum_i^c \sum_{j \neq i}^c \gamma_i \gamma_j C[P_i, P_j]] \\
&= c\overline{\gamma^2}(1 - \frac{1}{m}) V[N] - c(c-1)(\overline{\gamma^2} - \sigma_\gamma^2) \frac{1}{m} V[N] \\
&= c(\sigma_\gamma^2 + \mu_\gamma^2) V[N] - c(c\mu_\gamma^2 + \sigma_\gamma^2) \frac{1}{m} V[N]
\end{aligned}
$$

Finally:

$$\begin{aligned} V[Y] &= \underbrace{c\overline{\gamma^2}V[N]}_{basal} - \underbrace{c^2[\overline{\gamma^2} - \sigma_\gamma^2\frac{(c-1)}{c}]\frac{1}{m}V[N]}_{competition} \\ &= \underbrace{c(\sigma_\gamma^2+\mu_\gamma^2)V[N]}_{basal} - \underbrace{c^2(\mu_\gamma^2+\frac{\sigma_\gamma^2}{c})\frac{1}{m}V[N]}_{competition} \end{aligned} \tag{3.31}$$

3.8.7 Partitioning the heritability between core and peripheral genes

Peripheral effect

We define γ'_i, an adjustment of γ_i corresponding to the effect size per unit of core gene mRNA expressed. With $V[N_i] = V[N_i]$. Considering that $V[\delta] \approx 0$.

$$V_Y \approx \underbrace{\sum_i^c \gamma_i'^2 V[N_i, core]}_{c\ core\ terms} + \underbrace{E[\delta]^2 \sum_j^{m-c} V[N_j, peripheral]}_{(m-c)\ peripheral\ terms} \tag{3.32}$$

Both the core and peripheral genes produce a competitive effect. Each gene decreases the level of all the core proteins proportionally to its own level of expression. Therefore, the competitive effect is a factor of the **core genes overall effect** that correspond to $\sum^c \gamma_c P_c$. The core gene overall effect can be negative, positive or null. The competitive effect is a ratio of the core genes overall effect and $\sum N$, the total gene expression level:

$$\delta_n = -\frac{\sum^c \gamma_c P_{c,n}}{\sum_i^m N_{i,n}}$$

Case 3: alternate omnigenic competitive

While further on we will only consider that P and γ are distributed independently (case 1 or 2). It is to note that if P and γ are jointly distributed, positively or negatively, there will be a competitive effect even if core genes effects are not correlated. There is one singular case where the P_i and γ distributions cancels each other such as $E[\sum_i^c \gamma_i P_i]] = 0$.

The effect size of the peripheral effect, δ

When P and γ are distributed independently (case 1 or 2):

$$\begin{aligned}\delta &= -\frac{\sum^c \gamma_c P_c}{\sum N} \\ &= -\frac{c\overline{\gamma_c}\overline{P_c}}{\sum N} \\ &= -\frac{c(\frac{\sum^c \gamma_c}{c}\frac{\sum^c P_c}{c})}{\sum N} \\ &= -\frac{\sum^c \gamma_c}{c}\frac{\sum^c P_c}{\sum N} \\ &= -E[\gamma]\frac{\sum^c P_c}{\sum N}\end{aligned}$$

If we want to express δ in function of gene expression only, with N and γ are distributed independently:

$$\begin{aligned}\delta &= -\frac{\sum^c \gamma_c P_c}{\sum N} \\ &= -\frac{\sum^c \gamma_c \frac{N_c}{\sum N}}{\sum N} \\ &= -\frac{\sum^c \gamma_c N_c}{(\sum N)^2} \\ &= -\frac{\sum^c \gamma_c N_c}{(\sum^c N_c + \sum^p N_p)^2} \\ &= -E[\gamma]\frac{\sum^c N_c}{(\sum^c N_c + \sum^p N_p)^2}\end{aligned}$$

The variance of the peripheral effect, δ

$$V[\delta] = V[-\frac{\sum^c \gamma_c P_c}{\sum N}]$$

If γ and P are independent (case 1 and 2)

$$= V[\frac{\sum^c \gamma_c}{c}\frac{\sum^c P_c}{\sum N}]$$

Because γ is a constant across individuals:

$$= \frac{(\sum^c \gamma_c)^2}{c^2} V[\frac{\sum^c P_c}{\sum N}]$$

$$= E[\gamma^2] V[\frac{\sum^c P_c}{\sum N}]$$

In function of gene expression:

$$= (\mu_\gamma^2 + \sigma_\gamma^2) V[\frac{\sum^c N_c}{(\sum N)^2}]$$

The phenotypic variance coming from peripheral genes

$$\begin{aligned} V[Y, periph] &= V[\zeta] \\ &= V[\delta \sum^p N_p] \\ &= \delta^2 V[\sum^p N_p] \\ &= (-\mu_\gamma \frac{\sum^c N_c}{(\sum N)^2})^2 V[\sum^p N_p] \end{aligned}$$

If genes are i.i.d:

$$\begin{aligned} V[Y, periph] &= V[\zeta] \\ &= \mu_\gamma^2 \frac{c^2 p}{m^4} \frac{V[N]}{\overline{\mu}^2} \\ V[periph] &= \frac{\mu_\gamma^2 r_c^2 r_p}{m\overline{\mu}^2} V[N] \end{aligned}$$

4 Phenotype prediction from genome-wide genotyping data: a crowdsourcing experiment

Olivier Naret[1,*], David AA Baranger[2], Sharada Prasanna Mohanty[1], Bastian Greshake Tzovaras[3,4,5], Marcel Salathé[1‡], Jacques Fellay[1‡], with the openSNP and crowdAI community

[1] School of Life Sciences, Ecole Polytechnique Fédérale de Lausanne, Switzerland
[2] Department of Psychological and Brain Sciences, Washington University, St. Louis, MO, USA
[3] Lawrence Berkeley National Laboratory, Berkeley, CA, USA
[4] Department for Applied Bioinformatics, Goethe University, Frankfurt am Main, Germany
[5] Center for Research and Interdisciplinarity (CRI), Université de Paris, INSERM U1284, Paris, France

Background: The increasing statistical power of genome-wide association studies is fostering the development of precision medicine through genomic predictions of complex traits. Nevertheless, it has been shown that the results remain relatively modest. A reason might be the nature of the methods typically used to construct genomic predictions. Recent machine learning techniques have properties that could help to capture the architecture of complex traits better and improve genomic prediction accuracy.
Methods: We relied on crowd-sourcing to efficiently compare multiple genomic prediction methods. This represents an innovative approach in the genomic field because of the privacy concerns linked to human genetic data. There are two crowd-sourcing elements building our study. First, we constructed a dataset from openSNP (opensnp.org), an open repository where people voluntarily share their genotyping data and phenotypic information in an effort to participate in open science. To leverage this resource we release the 'openSNP Cohort Maker', a tool that builds a homogeneous and up-to-date cohort based on the data available on opensnp.org. Second, we organized an open online challenge on the CrowdAI platform (crowdai.org) aiming at predicting height from genome-wide genotyping data.
Results: The 'openSNP Height Prediction' challenge lasted for three months. A total of 138 challengers contributed to 1275 submissions. The winner computed a polygenic risk score using the publicly available summary statistics of the GIANT study to achieve the best result ($r^2 = 0.53$ versus $r^2 = 0.49$ for the second-best).

Conclusion: We report here the first crowd-sourced challenge on publicly available genome-wide genotyping data. We also deliver the 'openSNP Cohort Maker' that will allow people to make use of the data available on opensnp.org.

4.1 Background

As costs for genetic analyses keep dropping, genetic testing is becoming more available and affordable for increasing numbers of people - a trend that can be seen in the rising number of customers that use Direct-To-Consumer (DTC) genetic testing services like 23andMe and AncestryDNA [109]. Decreasing costs and increased availability have lead to the creation of a number of genomic data resources, such as the Personal Genome Project [110], DNA.land [111] and openSNP [112]. Amongst these data resources, openSNP is unique, in that it offers open participation and open access to the data: Participants of openSNP can use the platform to openly share their existing DTC genetic test data, putting their data in the public domain. In addition, participants can share phenotypic traits, such as eye color, hair color or height. Since its start in 2011, over 5,000 people have used the platform to make their genetic data available.

Crowd-sourced competitions of data analysis have become more and more popular in the past few years allowing data science experts and enthusiasts to collaboratively solve real-world problems, through online challenges. This approach allows the broad exploration of the model space on a specific dataset by people with data analysis skills coming from very different backgrounds. In the context of genomic prediction of complex diseases, it is unprecedented. While the most widely used platform, kaggle.com, offers monetary rewards, crowdai.org is more academic-centered and offered the winner the opportunity to present her work at a scientific conference.

We hereby present a crowd-sourcing experiment where participants could compete on crowdai.org to produce the best possible prediction of the height phenotype using data from opensnp.org.

4.2 Materials and methods

4.2.1 openSNP Cohort Maker

Because on opensnp.org, no restrictions are enforced on what users can upload, after downloading the data dump of the whole community, there is a need for in-depth data curation to produce a clean cohort of genome-wide genotyped individuals. To make these data accessible to anyone, we developed the openSNP Cohort Maker tool that through a systematic approach produced a clean and up-to-date openSNP cohort of genome-wide genotyped individuals.

When running the openSNP Cohort Maker, the data processing starts by downloading the archive containing all data that were uploaded on opensnp.org by the community. Then, files are removed if: they are not text or compressed text; they correspond to exome sequencing; they are genotyping data from decodeme; they are set on the GRCh38 reference; they are corrupted. For individuals who submitted multiple genome-wide genotyping data, either as duplicates or from different DTC companies, only the largest file is kept. A set of tools are

integrated into the pipeline: genotyping data with coordinates based on NCBI36 are upgraded to match the GRCh37 reference [113] with liftOver [114]; PLINK [115, 116] is used to convert file formats. VCFtools [117] is used to sort variants; BCFtools [118] is used to normalize reference and alternate alleles on the GRCh37 reference genome, rename samples, index files, and finally merge all individuals into one file. The output file can be directly imputed on the Sanger Imputation Service [28]. The openSNP Cohort Maker is available on GitHub Supporting information. Leveraging parallel computing, with 28 CPUs it takes 16 hours to produce a single file containing the curated openSNP cohort. From the initial archive containing 2487 different individuals, 2341 remain after filtering. From those, 2034 are from 23andme, 186 are from ancestry.com, and 121 are from ftdna-illumina.

4.2.2 CrowdAI Challenge

The dataset that we used for the challenge was produced by the openSNP Cohort Maker and imputed on Sanger Imputation Service with HRC (r1.1). We sent to the opensnp.org community a survey asking for their height, allowing us to create a dataset regrouping 921 individuals with both height phenotype and genotyping data.

Challenge participants could use two versions of the genotyping data. One version was a sub-dataset containing 9,894 genetic variants, including the top 9,207 variants ($p < 5E-3$) associated with height in the GIANT study [61], and 687 Y chromosome variants. The second versions was a full dataset containing 7,252,636 variants which passed a quality threshold, defined as an imputation score $INFO > 0.8$, genotyping missingness frequency $F_m < 0.1$, and a Hardy-Weinberg equilibrium exact test p-value $< 1E-50$. Both versions of the data were given in the VCF format, as well as in an additive format where each genetic variant is represented by 0 (homozygous for reference), 1 (heterozygous), 2 (homozygous for the alternate allele) or NA (missing data or variants of allosomes), easier to handle for participants unfamiliar with genetics.

The data were separated into two sets, a training set with 784 samples and a test set of 137 samples (an 85/15 split). The challengers were provided the training set with the genotyping data and the height phenotype and the test set with the genotyping data only. The challengers needed to train their model on the training set and produce predictions for the samples of the test set. The test set predictions were then submitted to the CrowdAI platform for evaluation and scoring. The score was produced based on the Pearson's correlation (r^2) between the predicted and true height. The challengers could submit as many prediction models as they wanted in an attempt to improve their method and beat their best score. The scoring method was protected from known exploits [119]. The data are available online on the zenodo platform Supporting information, and the webpages presenting the challenge Supporting information and the leaderboard Supporting information have been saved to PDF from the CrowdAI platform.

4.3 Results

A total of 138 challengers participated, contributing a total of 1275 submissions. The winner computed a polygenic risk score (PRS) using the publicly available summary statistics of the GIANT study to achieve the best result ($r^2 = 0.53$ versus $r^2 = 0.49$ for the second-best).

The training set and testing set were combined for quality control and data preparation. As self-reported sex was not provided, participant's chromosomal sex (i.e. XX vs XY) was imputed using PLINK, which uses the X chromosome inbreeding coefficient (F) to impute sex. Standard cutoffs were used, whereby $F < 0.2$ yielded an XX call, while $F > 0.8$ yielded an XY call. One participant yielded an F of exactly 0.2, and was removed from subsequent analyses (they were in the training data). Of the remaining 920 individuals, 396 (43%) were XX, and 524 (57%) were XY.

The openSNP platform contains genomic data of relatives. The presence of relatives has the potential to bias results, as closely-related individuals will dominate the estimation of principal components and will inflate prediction accuracy statistics [120] . The genetic relationship between participants was calculated using the PLINK computation of identity-by-descent (IBD), which is an estimate of the percent of the genome (excluding sex chromosomes) shared between two individuals. The IBD analysis identified seven pairs of strongly-related individuals (first-cousin or greater), including two pairs of monozygotic twins. The analysis also identified a surprising cluster of 18 individuals estimated to be 3^{rd} cousins, or equivalent. All but one member of each family-group was removed from analyses (N=24), all from the training data. It is well-established that the frequency of genetic variants and correlational structure of the genome differs across ancestral populations [120, 121, 122]. These differences are the major barrier to combining genomic data across ancestries in genome-wide association studies [123, 124]. Genome-wide principal components were computed using PLINK. A scree plot of eigenvalues indicated an elbow at three components. The large eigenvalue of the first principal component, and the shape of all three components, clearly showed that both the training and testing data contained participants of multiple ancestries (i.e. participants of European, African, and Asian ancestry were present in both data sets), though the majority of participants were of European descent.

Genomic data were further processed in PLINK, following the steps outlined by PRSice [125] for the computation of PRS. This included removal of variants that were missing for more than 2% of participants, removal of variants with a minor-allele-frequency less than 0.02, removal of variants with a Hardy-Weinberg equilibrium exact test $p_{value} < 10^{-6}$, removal of variants within the major histocompatibility complex on chromosome 6, and removal of non-synonymous variants. Because neighboring genetic variants can be correlated due to linkage disequilibrium, genetic variants were clumped in PLINK, wherein groups of variants correlated at $r^2 > 0.1$ were identified across a sliding window of 250 kilobases. Within each group, only the variant with the lowest p-value in the GIANT genome-wide association study of height [61] was retained.

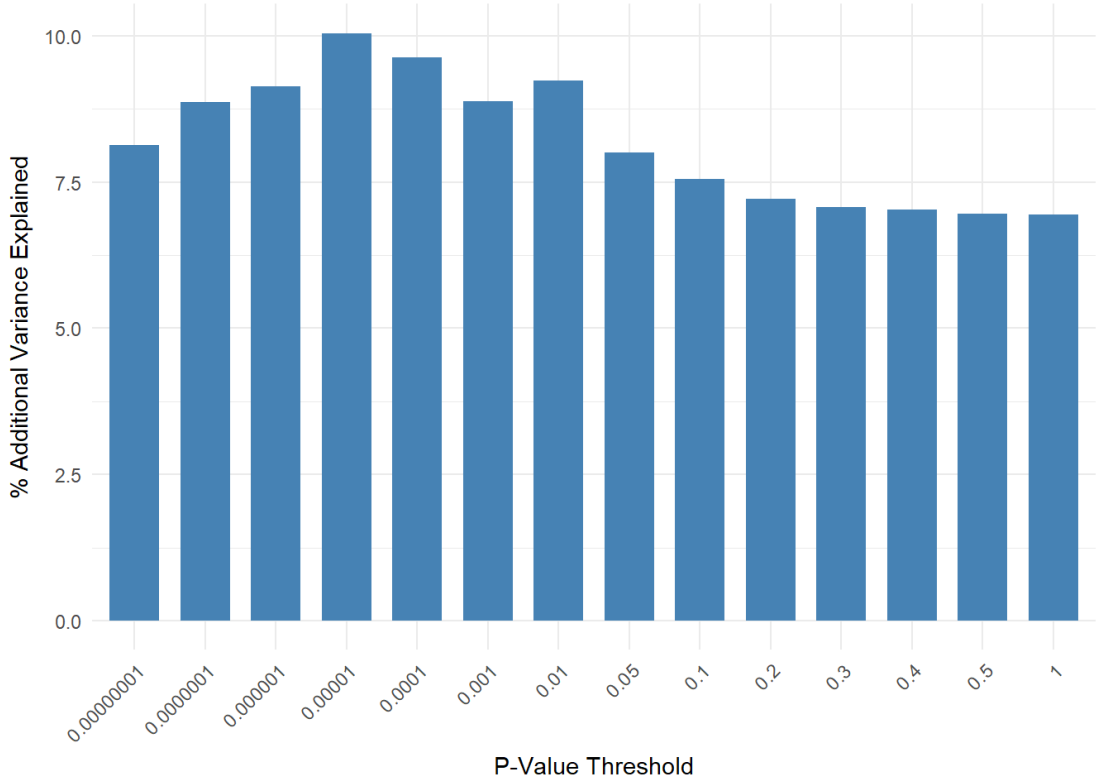
A PRS is a metric reflecting an individual's genetic burden for a disease or trait of interest. [126, 84]. Prior work on the genetic basis of height has found that a PRS for height captures over 20% of the variance in independent samples [61]. PRS are calculated by averaging the number of disease-associated alleles, weighted by their effect size, from an independent study [127]. Put differently, a linear regression predicting the outcome trait is modeled at each individual variant, using the effect size from an independent study. These predictions are then averaged across all models. The one free parameter is the decision of which variants to include in the calculation of the PRS. Typically, the significance of the association of each variant in the independent study is used. Thus, multiple scores are calculated, including only variants that are associated below different p-value thresholds (e.g. $p < 5E-1$, $p < 5E-2$, $p < 5E-3$, etc.). While a PRS including all variants (i.e. $p < 1.0$) typically does not perform the best, neither does a PRS including only variants which surpass family-wise error rate correction for multiple comparison (i.e. $p < 5E-8$). Finally, the confounding effects of ancestral populations apply to PRS analyses as well. PRS work best when the independent study (e.g. the GIANT study used here) was conducted within a homogeneous sample of participants all of whom are from the same ancestral background, and when the sample the PRS is being calculated for is of the same background. For instance, PRS computed from studies of individuals of European descent are well-known to produce biased results in samples of African or Asian descent [128, 129].

To win, PRSice and PLINK were used to compute PRS for height in the openSNP sample, using the results from the GIANT study of height. PRS were computed at 14 different p-value thresholds ($p < 10^{-8}$ to $p < 1.0$), shown in Fig 4.1. Linear regressions predicting height in the training data were fit in R. Chromosomal sex was the first variable included in the model, followed by the top three genome-wide principal components, which help to control for differences in ancestral background [45] . Chromosomal sex predicted 46.81% of the variance in the training data, and the addition of the three principal components subsequently explained 0.91% of variance. Finally, each of the 14 PRS were added to the model and compared. The PRS at $p < 1E10^{-5}$ was observed to perform best, and captured an additional 10.08% of variance. Thus, the final linear regression model explained a total of 57.80% of the variation of height in the training data set, shown in Fig 4.2. Predictions for height in the test data set were then generated from this regression model, predicting 53.45% of variance (MSE = 47.32).

4.4 Discussion

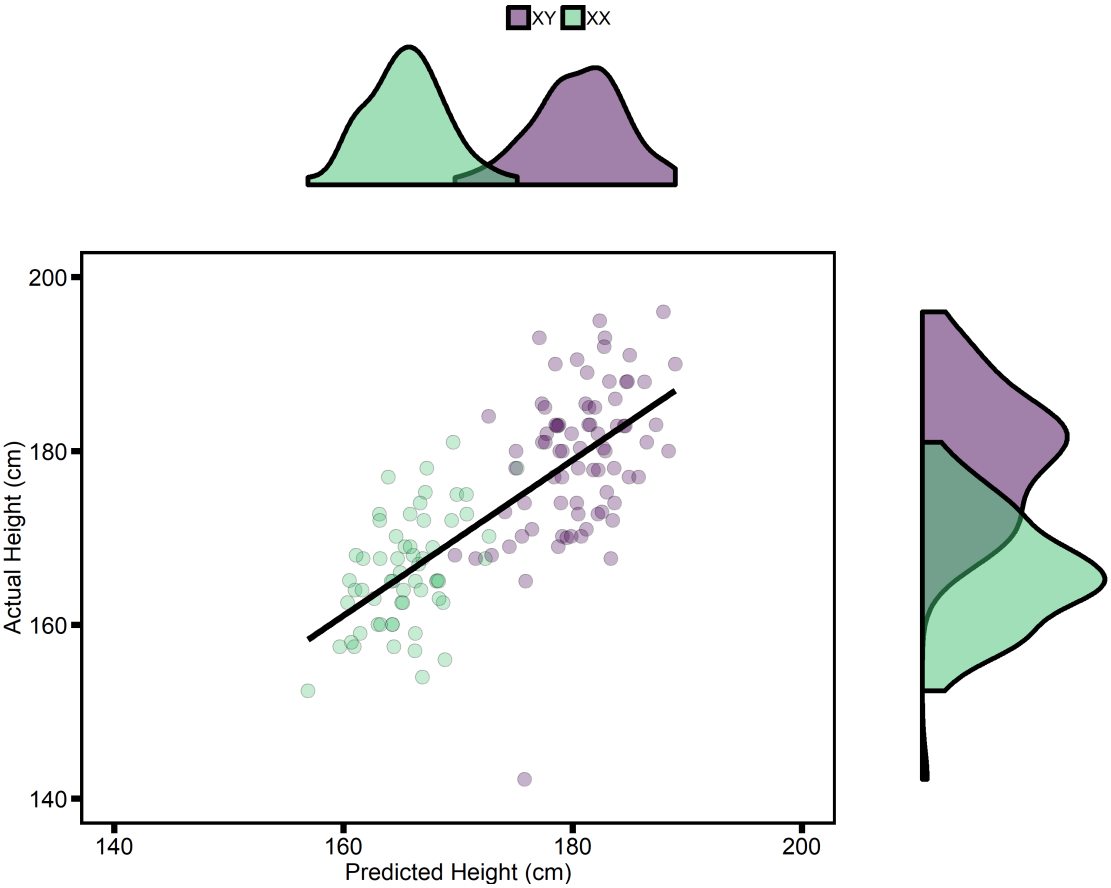
Height is an extremely polygenic trait where even the hundreds of genome-wide significant variants contribute all together for only a small portion of heritability [130]. Because of the modest size of the OpenSNP cohort, the lack of statistical power was the main difficulty for the challengers to capture the association signals coming from the genetic variants. The winning model of the challenge incorporated the GWAS summary statistics from the GIANT study to compute a PRS, in addition to deriving each participant's sex. It should be noted that PRS is a standard and widely-used technique in the field of statistical genetics. While

Figure 4.1 – Variance explained as a function of p_{value} threshold.



PRS were produced at a range of p-value thresholds (x-axis). Y-axis represents Nagelkerke's r-squared from training-sample linear regressions. The model with the best performance in the training data ($p < 5 \times 10^{-4}$) was then used to predict height in the test-sample.

Figure 4.2 – Predicted height distribution versus real height distribution.



Predicted height in the training-sample (x-axis) is displayed relative to true height (y-axis). Points are colored by chromosomal sex. X and y-axis density plots show the predicted and true overlap of height between the sexes.

cross-population PRS have been shown to be unreliable in multiple cases, such as Type II Diabetes [131], coronary artery disease [132], and height [133], the similarities between the GIANT and openSNP cohorts were sufficient to provide a winning strategy. This is likely because only a small portion of samples were of non-European ancestry (7%).

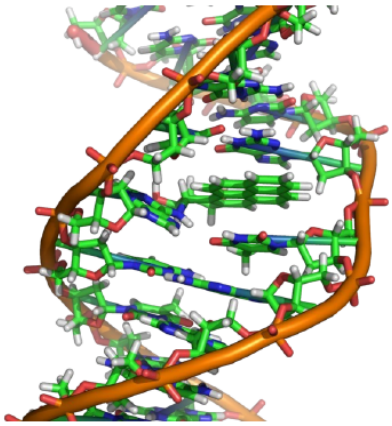
So far, PRS are classically limited to additive models which might not represent the whole complexity of the genetic architecture of some traits. Indeed, the phenotypic variance explained by PRS remains modest in comparison to the heritability of the traits [4] (the so-called 'missing heritability' problem). The inability to consider gene-gene interactions is one of the many factors potentially explaining for this PRS weakness. In this case, the effect of a variant depends on the presence or absence of another variant, a mechanism that is not captured by additive models and accounts for an unknown part of the phenotypic variance [5]. Eventually, more advanced statistical approaches relying on machine learning could improve on the prediction accuracy provided by purely additive risk scores. Because of the diversity in available methods and the world-wide distribution of excellent data scientists, we believe that crowd-sourcing approaches represent a promising strategy to help improve phenotypic prediction from large-scale genomic data.

4.5 Conclusion

Because of privacy concerns, studies relying on crowd-sourcing are almost impossible to set up in the field of human genomics. A first experiment was carried out in 2016 to predict anti-TNF treatment response in rheumatoid arthritis [134], but participants had to apply to participate in the challenge. Here - thanks to the OpenSNP community - we released the first crowd-sourced and fully open challenge based on publicly available genome-wide genotyping data. The competition attracted 138 challengers, with diverse backgrounds, from the vibrant machine learning community. It resulted in the assessment of a broad variety of methods for genotype-based phenotypic prediction through a total of 1275 submissions. We hereby also report a tool to create an up-to-date and curated OpenSNP cohort, making this open genomic resource much more user-friendly.

4.6 Supporting information

S1. Appendix



OpenSNP Height Prediction

OpenSNP

By EPFL

Completed	16070 Views	138 Participants	1275 Submissions

Overview

This challenge aims at predicting height based on genetics (DNA variation).

Background

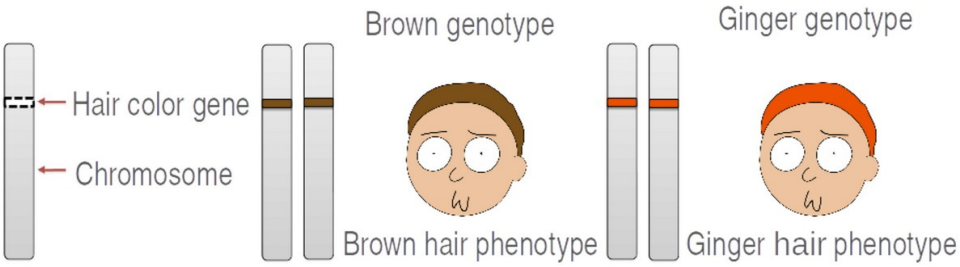
Genetics is the study of genes, genetic variation, and heredity in living organisms. DNA is the support that allows most living organism to pass information from generation to generation. It consists in long strands of nucleotides that build a higher order structures, the chromosomes. There are four different nucleotides represented by the letters A, T, C and G that together make up the genetic code.

The human genome is made of > 3 billion nucleotides, and each individual harbors about 4 million genetic variants (mostly single nucleotide polymorphisms, or SNPs). A specic position on a chromosome is called a genetic locus, and different versions of the same genetic locus are called alleles. Humans being diploid organisms, they have two genome copies - one inherited from each parent - and thus two alleles at each genetic locus. For one particular genetic locus, an individual is homozygous if the two alleles are identical, and heterozygous if the two alleles are different.

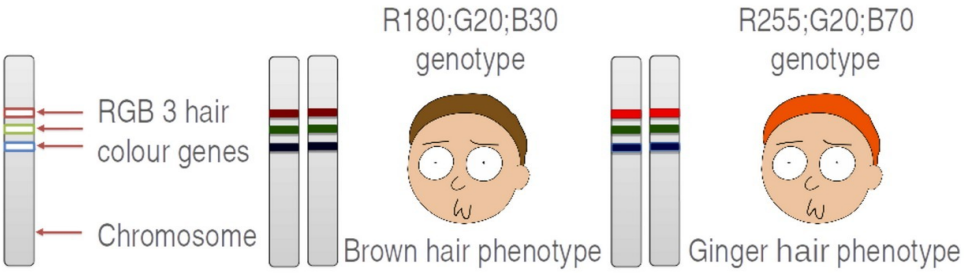
We call genotype the DNA sequence of an individual that determines a specic observable characteristic. That characteristic is called phenotype.

Released challenge presentation to the participant on the CrowdAI website.

Monogenic phenotypes are under the control of a single gene. For example, if hair color was a monogenic phenotype, inheriting two brown alleles of a hypothetical hair color gene would result in the brown hair phenotype. Conversely, inheriting two ginger hair alleles would result in the ginger hair phenotype.



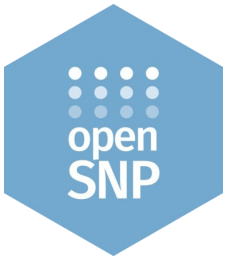
Polygenic phenotypes, on the contrary, are under the control of multiple genetic variants across the genome. If hair color was polygenic, it might for example work like the RGB (Red, Green, Blue) color model. In this case, three different genes would add their effects and interact to control hair color.



Heritability of a phenotype measures how much of the observed variance of the phenotype in the population is due to genetic factors. Missing heritability represents the difference between the estimated heritability of a given phenotype, and the heritability that is explained by known genetic factors. Heritability of human height is estimated to be as high as 80%, but large genomic studies have so far only been able to explain about 25% of the observed variance. Height is a model phenotype to study complex traits, and here we want to test whether part of the missing heritability can be explained using innovative approaches to genetic datasets, including deep learning.

Data

The data comes from OpenSNP, which allows customers of direct-to-customer genetic tests to publicly share their genome-wide genotyping data.



We provide two datasets for a total of 921 samples divided into a training set of 784 sample `subset_cm_train.npy` and a test set of 137 samples in `subset_cm_test.npy`.

It contains a set of 9,894 genetic variants known to be associated with height[1] (9207 variants) and the one on Y chromosome (687 variants). This numpy le has shape `(784, 9894)` for the training set and `(137, 9894)` for the test set. Each genetic variant is represented by 0 (homozygous for reference) , 1 (heterozygous), 2 (homozygous for the genetic variant) or NA (missing information or absence of the position in the case of Y chromosome in women). The rst 9207 rows are the genetic variants known to be associated with height, the last 687 correspond to the Y chromosome.

Finally, height is provided in a separate numpy le of shape `(784, 1)` named `openSNP_heights.npy` for the training set only.

While we recommend to start with this simplied dataset, more advanced user might try to analyze an extended version of OpenSNP data which description is available here.

Submission

```
import crowdai
challenge = crowdai.Challenge("OpenSNPChallenge2017", "YOUR_CROWDAI_API_KEY_HER

data = ... #a list of 137 predicted heights for all the 137 corresponding data
challenge.submit(data)
challenge.disconnect()
```

More instructions to make submissions, and starter code is available at :

STARTER KIT https://github.com/crowdAI/opensnp-challenge-starter-kit

Evaluation

The evaluation will be done based on two scores :

- Co-ecient of Determination (R^2) (Primary Score)
- Mean Squared Error (Secondary Score)

between the actual heights of the individuals in the test set and the submitted predictions.

NOTE : During the challenge, the scores will be computed only on 20% of the test dataset. The nal standings on the leaderboard will be decided computing the same scores on the 100% of the dataset after the challenge.

Prizes

The winner will be invited to the 2nd Applied Machine Learning Days at EPFL in Switzerland on January 29 & 30, 2018, with travel and accommodation covered.

Resources

MIT Open Course Ware can help you go further in understanding biological concepts related to this challenge.

- Lesson 19 on Discovering Quantitative Trait Loci (QTLs)
- Lesson 20 on Human Genetics, SNPs, and Genome Wide Associate Studies

The most important publications describing associations between genetic factors and human height :

- based on common SNPs [1]
- on rare SNPs [2].

To transform the VCF les it can be convenient to use plink .

References

1 Wood, Andrew R, Tonu Esko, Jian Yang, Sailaja Vedantam, Tune H Pers, Stefan Gustafsson, Audrey Y Chu, et al. “Dening the Role of Common Variation in the Genomic and Biological Architecture of Adult Human Height.” Nature Genetics 2014. doi:10.1038/ng.3097.

2 Marouli, Eirini, Mariaelisa Graff, Carolina Medina-Gomez, Ken Sin Lo, Andrew R. Wood, Troels R. Kjaer, Rebecca S. Fine, et al. “Rare and Low-Frequency Coding Variants Alter Human Adult Height.” Nature 2017. doi:10.1038/nature21039.

S2. Appendix

Figure 4.3 – Challenge leaderboard.

Δ	#	Participant	Coefficient of Determination (R^2)	Mean Squared Error	Entries	Last Submission (UTC)
•	01.	 David_B...	0.566	44.069	20	Tue, 6 Mar 2018 22:09
•	02.	 spMohanty Admin	0.492	50.937	42	Wed, 19 Jul 2017 23:15
•	03.	Nurislam	0.486	51.569	17	Fri, 14 Jul 2017 04:49
•	04.	 Muhamma...	0.479	53.012	26	Tue, 26 Sep 2017 23:11
•	05.	SergeKrier	0.476	53.297	11	Thu, 28 Sep 2017 16:23
•	06.	mmi333	0.475	53.402	15	Wed, 30 Aug 2017 17:41
•	07.	NB	0.474	53.455	42	Fri, 9 Mar 2018 14:07

First page of the leaderboard.

S1. Software
OpenSNP cohort maker: https://github.com/onaret/opensnp-cohort-maker

S1. Dataset
Challenge dataset: https://zenodo.org/record/1442755#.XlTwyHVKh1M

4.7 Declaration

4.7.1 Ethics approval and consent to participate

The research proposal was submitted for evaluation by the Ethics Commission of Canton Vaud (CER-VD), under number Req-2016-00421. It was exempted from detailed ethics review due to the nature of the project and the focus on an anthropometric trait without any direct impact on health. All participants shared their genotyping data on an online platform and explicitly allowed anyone to use it without additional consent.

4.7.2 Availability of data and material

The tool 'OpenSNP cohort maker' is available on GitHub repository [135]. The datasets used during the challenge are available on the Zenodo repository [136].

4.7.3 Competing interests

the authors declare no conflict of interest.

4.7.4 Funding

The project was supported by in-house funding from EPFL to JF and MS. DB was supported by the NIH (T32-GM008151) and NSF (DGE-1143954).

4.7.5 Authors' contributions

ON built the 'OpenSNP cohort maker', the OpenSNP cohort dataset, and designed the challenge. DB won the challenge and detailed his methods. BG supervised the collaboration with the OpenSNP community. SPM administrated the challenge on the CrowdAI platform. JF and MS have substantively revised the work. All authors read and approved the final manuscript.

4.7.6 Acknowledgements

We thank the OpenSNP community who took part in the experiment and kindly provided us their height phenotype through our survey. We thank all the challengers who provided their time and expertise to compete in the CrowdAI challenge. We thank the people who contributed to the CrowdAI platform and provided the support to have this challenge running seamlessly.

5 Using the epistasis within topologically associated domains to improve polygenic score

Olivier Naret[1], Jacques Fellay[1]

[1] School of Life Sciences, Ecole Polytechnique Fédérale de Lausanne, Switzerland

Traditionally, the construction of polygenic score (PGS) uses a large number of genetic variants, the bulk of which have weak additive effects and are inevitably entangled with noise. Here, we designed a method to aggregate additionally multiple weak epistatic effects. Due to the enormous number of potential combinations of variants, it is difficult to integrate epistatic effects into PGS. In order to reduce the possible combinations, we limit the possible combinations to the boundaries of each topologically associated domain (TAD). Using UK Biobank data and focusing on the height phenotype, we included 17,560 variants in an artificial neural network and compared the variance explained (R^2) by the PGS with or without knowledge of the TADs. We found that allowing for potential epistatic interactions within TAD borders brings a significant improvement with an average R^2 going from 0.287 to 0.293 (with a p-value $= 10E-5$ for n=20. We concluded that, for a highly polygenic trait such as height, PGS can be improved when epistasis restricted to TAD borders is considered, and that artificial neural networks have the potential to uncover weak epistatic signals.

5.1 Background

5.1.1 Epistasis

Definition

Gene-gene interactions (G x G) or epistasis have various definitions depending on biological or statistical considerations. Initially, it was defined by Bateson as a deviation of Mendelian inheritance where an allele at a given locus masks the expression of another allele at another locus[137]. From a statistical standpoint, Fisher defined it as any non-linear effect combining multiple allelic effects[1]. From a biological standpoint, Moore added that it must be associated with a physical interaction between biomolecules[39, 138]. Here, we tested whether topologically associated domains (TADs), which are domains of the genome with a higher density of physical interaction, are associated with detectable epistatic effects.

In topologically associated domains

Topologically associated domains (TADs) are functionally delimited genomic regions of ~ 800kb with self-interacting DNA, meaning that DNA sequences inside a TAD physically interact with each other more frequently than with DNA sequences outside the TAD[139]. TADs strongly influence contacts between gene promoters and their associated enhancers. Therefore, a simple theoretical model of interaction would be such that an enhancer would exist in two allelic forms - weaker or stronger - acting on a promoter also existing in two allelic forms - weaker or stronger - which impact on the phenotype is a function of the regulated gene level of expression.

$$\text{phenotype} \sim \beta 1.\text{enhancer} + \beta 2.\text{promoter} + \beta 3.(\text{enhancer} \times \text{promoter})$$

As a fraction of heritability

The importance of epistasis in the missing heritability of traits has been debated. Some believe that the totality of heritability can be captured by purely additive models[66, 140]. But others pointed out that depending on the phenotype, additivity is not always sufficient to explain the missing heritability[141]. If it is established that the lion's share of heritability for complex phenotypes can be modeled additively, we seek here to evaluate the potential contribution that can be made by epistasis. The challenge in general when studying epistasis is to identify truly interacting variants and to understand the underlying biological mechanisms. Therefore, it has similarities with the task of interpreting GWAS result, but with an additional degree of complexity. However, the models used for the construction of polygenic scores (PGS) from GWAS results do not rely on the understanding of the biological mechanisms. In such a case, the models aggregate the additive effects of a set of variants, many of which are well below the genome-wide significance threshold, meaning that they are in fact a mixture of truly

associated variants and - to some degree - false positives. With a similar mindset, we are not trying here to identify precise epistasis signals but to design a model aggregating weak epistatic effects to improve the PGS.

The curse of high dimensionality

The main difficulty in studying epistasis is the curse of high dimensionality caused by the many possible combinations of interacting variants. The interaction between two variants yields $3^2 = 9$ two locus genotype cells {(0,0), (0,1), (1,0), ..., (2,2)}, three variants yield already $3^3 = 27$ three locus genotype and so on and so forth. These numerous possibilities are to the detriment of models whose number of parameters increases proportionally, leading to decreased performances, over-fitting, and a high computational burden. Two remedies for this are feature selection and feature extraction that we will jointly use based on the prior knowledge of standard GWAS summary statistics, i.e., for each variant, its marginal effect size and the corresponding p-value (see method section for more details). Finally, by limiting the possible interactions to intra-TAD space, we considerably reduce the number of parameters.

5.1.2 Deep learning incentives

Deep learning has become an important tool in the field of genomics. It has surpassed standard methods for predicting alternative splicing events[142], the specificity of DNA- and RNA-binding protein sequences[143], to predict the effects of non-coding variants[144] and to annotate presumptive pathogenic variants[145]. More recently, AlphaFold, which predicts protein folding, has largely solved a decades-old problem[9]. Nevertheless, the performance of deep learning approaches for constructing PGS is mixed. The first experiment was limited to a simple model of a disease based on two loci and have demonstrated the potential of artificial neural network (ANN) to model epistasis[146]. However, on larger datasets, convolutional neural networks (CNN) and MultiLayer Perceptron (MLP) either failed to outperform the linear model[147, 148] or, it was estimated by simulation that it could bring an improvement if there were enough loci involved, and if the non-additive variance was large enough[149].

5.1.3 Phenotype selection

We study the standing height phenotype, which is highly hereditary (80%)[52, 53], highly polygenic[54, 55, 56, 61] and widely available. It is noteworthy that height is a phenotype that additive models summarize very well, but we test here if better can be achieved by allowing epistatic effect[140, 141].

5.2 Material and methods

5.2.1 UK Biobank

We use the UK Biobank (UKB)[6], a cohort containing a total of 488,377 samples (262,933 females and 221,113 males) of which 4675 samples are excluded for various reasons (lack of size phenotype, poor quality genotypes, non-compatible sex, aneuploidy, extreme outlier) and 77546 for non-white British ancestry resulting in a remaining set of 407,500 samples. We study the phenotype of height, which we normalized through residualization by sex and rank-based inverse normal transformation.

We use as reported in the ENCODE project [150] with GEO:GSE105988

5.2.2 Genome-wide association study

Rare variants with an effect sizes of up to 2cm per allele have already been reported[63]. We estimated the minimum minor allele frequency (MAF) threshold to use for a power of 0.8 and an effect size of 2.5cm at $\geq 6e^{-4}$ with a power analysis[151]. Of the initial 92 million SNVs, 84 million were excluded on the basis of the MAF threshold, the high missingness rate (> 0.01), the rejection of the Hardy-Weinberg equilibrium test with ($p_{val} < 5e^{-50}$), or an insufficient INFO imputation score (< 0.8).

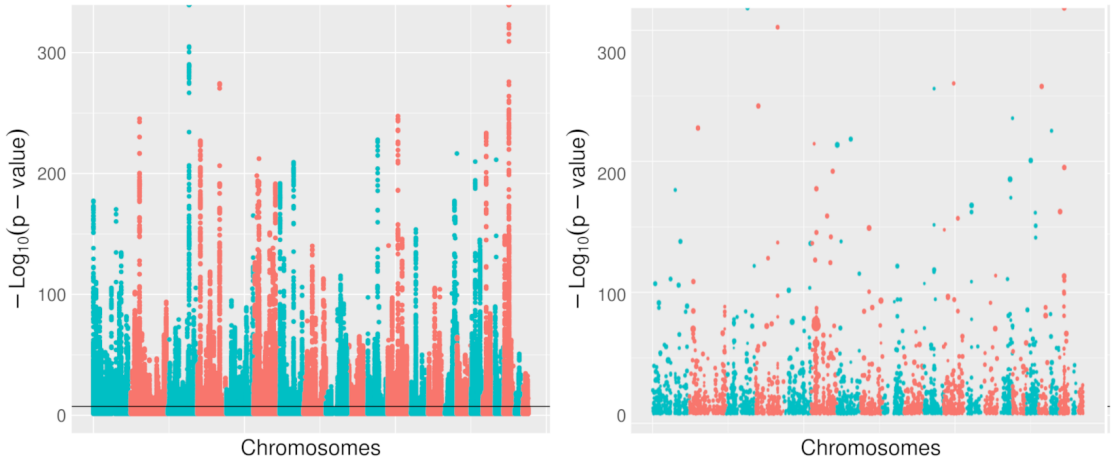
Of the 407,500 samples, 5,000 were randomly selected and removed to create an independent *test set*. The remaining 402,500 samples form the *basic set* on which a GWAS was performed for height using BOLT-LMM. We included as covariates the genotyping batch, the array type, the first 40 PCs and age. The heritability estimated by GREML-SC was $h^2_{SNP} = 0.56$ which is consistent with the literature. The results for height are represented by the Manhattan plot in fig.5.1A with 249,787 genome-wide significant SNVs.

Feature extraction was performed by clumping using plink. Briefly, the clumping step leverages the LD properties of the genome to construct groups of SNPs (or clumps) bellow a given maximal p-value threshold. The LD properties were calculated from the 1000 genomes project[152, 27], with a maximum window of 250kb for each group and a $r^2 < 0.01$. For each group, the most significant SNP was retained. Fig.5.1B shows the results of the clumping step, with a total of 1,924,915 clumps. Furthermore, because imperfect LD could lead to phantom epistasis, it is necessary to have SNPs in linkage equilibrium [153].

Feature selection was performed with PRSice[125]. After the clumping step, PRSice determined by cross-validation on the independent *test set* that the optimal filtering p-value threshold to build PGS was p-value < 0.01. These SNPs were extracted per TAD with their phase previously calculated by the UK Biobank with Shapeit[6]. In the end, few additional SNPs are removed because they could not be included in a TAD.

Figure 5.1 – Manhattan plot of GWAS on height and the clumped results

(A) Manhattan plot of height GWAS (B) Manhattan plot of clumped height GWAS



The fig.5.1A is a Manhattan plot of the GWAS of height. The fig.5.1B shows the result of the GWAS clumped with the size of each clump proportional to the number of variants clumped for $r^2 = 0.01$ *and a window of 500 kb*

5.2.3 Artificial neural network

We designed the ANN shown on fig.5.2B. The model is based on a multi-layer perceptron (MLP) with a hidden layer of fully connected nodes and constrained by TAD, as for n TAD, there are n input layers $\in \mathbb{R}^m$ each with its own m number of variants, followed at a depth of 2 by n hidden layers $\in \mathbb{R}^m$.

Each phase propagates separately in the MLP part to merge into a final hidden layer at depth 3 through a simple addition. From this last hidden layer, a final score is produced on the output node. The activation function used is the Scaled Exponential Linear Unit (SELU)[154]

$$\text{SELU}(x) = \lambda \begin{cases} x & \text{if } x > 0 \\ \alpha e^x - \alpha & \text{if } x \leq 0 \end{cases} \tag{5.1}$$

From the *base set*, 5,000 additional samples were randomly selected to create a *validation set*, and a *training set* with the remaining 397,377 samples. For the training session, we implemented an early stopping procedure to avoid overfitting. The model was trained on the *training set* and at the end of each epoch a validation score was calculated on the *validation set*. The best model was selected as the one after which 5 consecutive epochs have not brought any improvement in the validation score. At the very end when the training is over, a R^2 score was computed on the *test set*. The batch size was set to 16.384 with a learning rate of 5E-5. The model had a total of 216,795 parameters.

5.3 Results

To test if prior knowledge of the TADs can be used to boost PGS, we trained the same model design on two variations of the datasets as shown in fig.5.2A. For H1, the SNPs of the same TAD can propagate together in the MLP section. In this case, the potential interactions can be captured. For H0, the SNPs are randomly swapped genome-wide and therefore, SNPs initially of the same TAD do not end propagating together in the MLP section. In this case, the potential interaction within the TADs cannot be captured. We trained 10 models with each case and calculated the mean phenotypic variance explained on the *test set*. The result is presented in fig.5.3 where we observe a significant difference in the means for height, blood pressure, and bone density heel but not for BMI.

It is interesting to note that in terms of the number of SNPs included (supplementary fig.5.5) and the SNP-based heritability estimate, BMI is at an intermediate level. The h^2_{SNP} of BMI is 0.30, between the max for height at 0.57 and the min for blood pressure at 0.19. Similarly, the number of SNPs for BMI is 12,545, between the max for height at 17,690 and the min for bone density heel at 3,420. Therefore, it naturally questions if within-TAD epistasis is trait-specific. We consider these results very preliminary and further exploration needs to be conducted

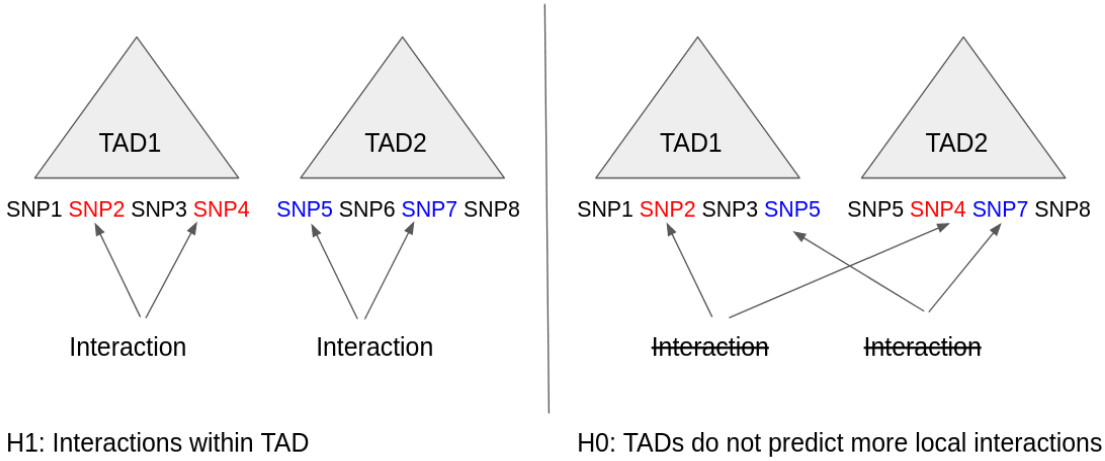
5.4 Conclusion & Next steps

To conclude, we found that building a model with a prior knowledge of TADs significantly improves the predictive power of the PGS for the height phenotype.

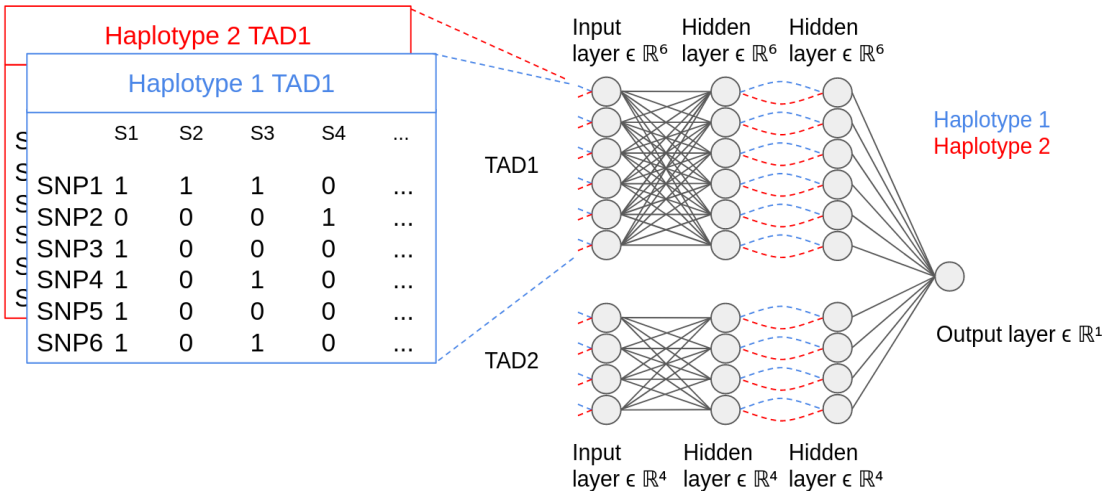
Further tools should be developed to allow the interpretation of the estimated parameters within a neural network to extend its usage to functional work. A potential direction could be to try to identify TADs where the gain is particularly higher compared to a linear combination. For example, based on our model on Fig.5.2B, $\forall$ n TADs with m SNPs, the corresponding second hidden layer $\in \mathbb{R}^m$ could converge to a single node of an additional hidden layer $\in \mathbb{R}^n$. In such a case, the weight of each node at the output of each TAD can be compared with the weight of the equivalent linear combination. If some TADs show a larger statistical gain signal, a second step of research with functional data could help identifying new biological insights.

Traditionally, PGS are built based on variants in simple linear models with additive effects. We have shown here that improvements can be obtained with a model allowing interactions. In the future, we expect methods producing PGS to become more comprehensive and closer to the true complexity of the human genome.

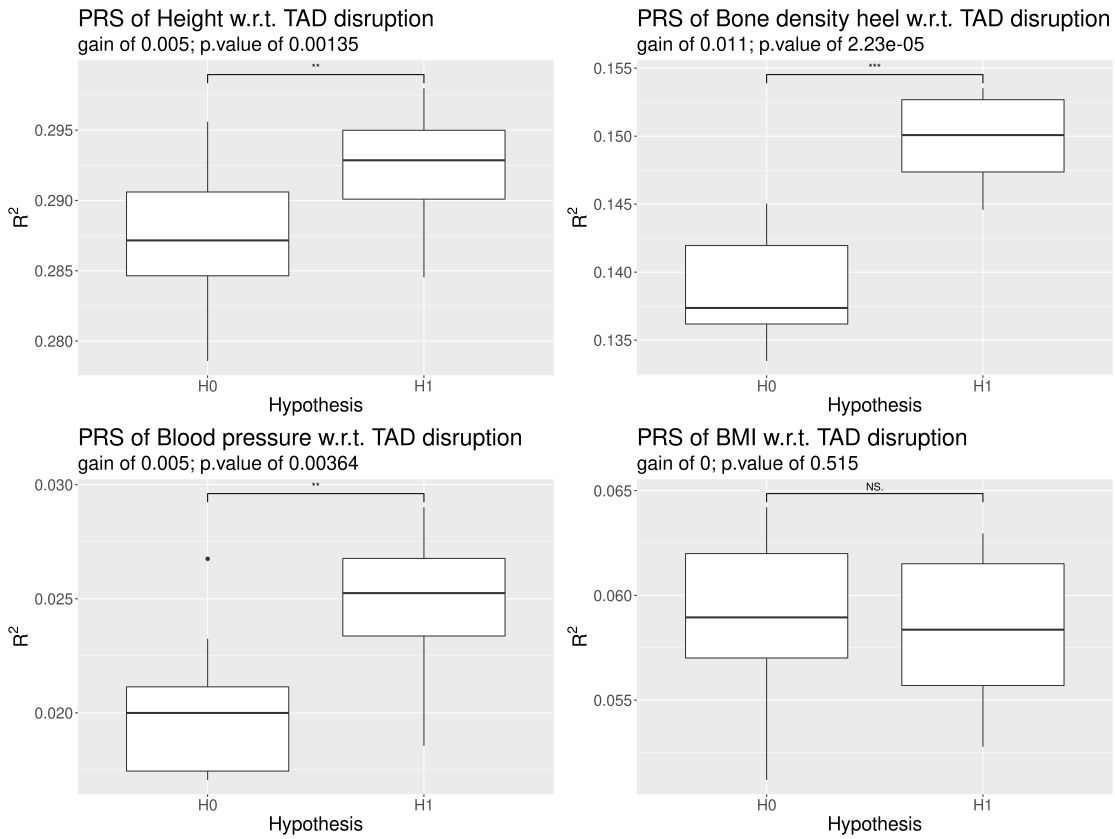
(A) Experimental design



(B) Artificial Neural Network architecture

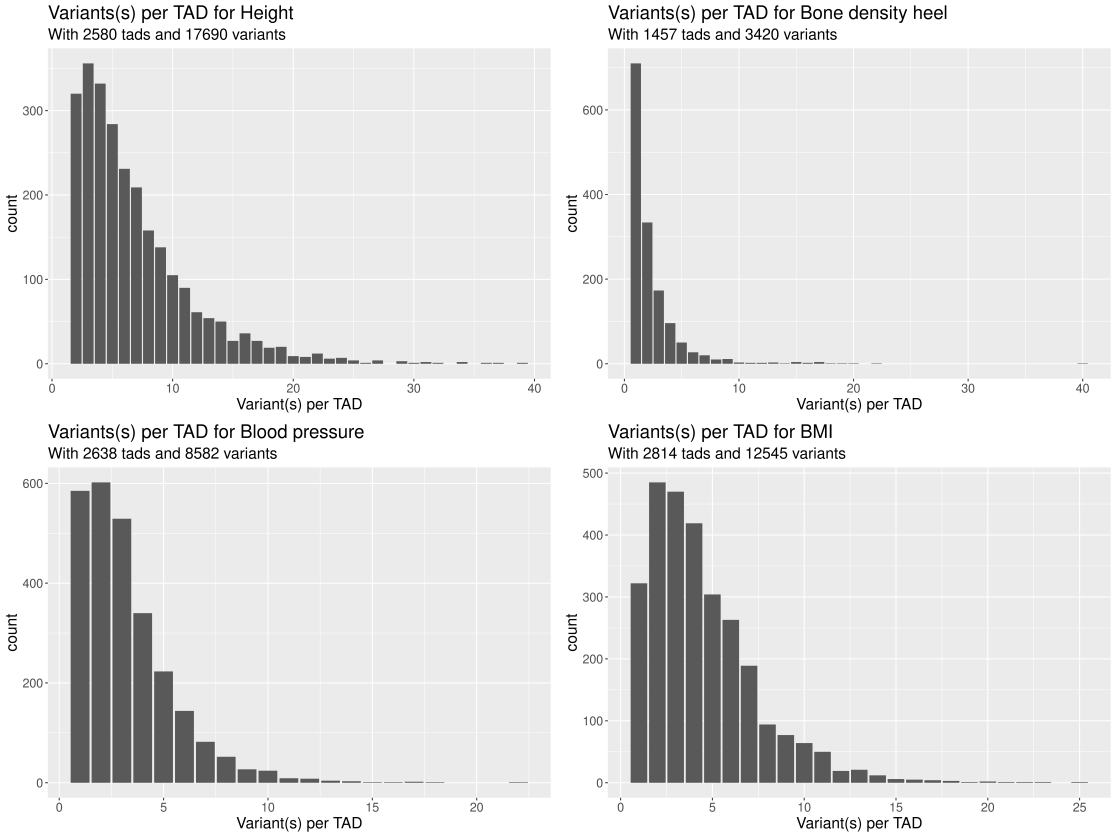


The fig. 5.2B shows the Artificial Neural Network architecture. The two colours represent the haplotype 1 and 2 which propagate in the neural network one after the other and are added before the last hidden layer. Here, only two TADs are represented but there are 2580 in total with 17,560 SNPs.



The fig. 5.3 Shows the difference in variance explained by the PGS with the TAD-ordered SNPs (H1) compared to the swapped SNPs (H0).

5.5 Supplementary materials



Distribution of the number of SNPs per TAD.

6 Improving polygenic score with genetically inferred ancestry

Olivier Naret[1], Jacques Fellay[1]

[1] School of Life Sciences, École Polytechnique Fédérale de Lausanne, Switzerland

Genome-wide association studies (GWAS) have identified multiple genetic variants associated with traits or diseases. They confirmed that most common diseases have a strong genetic component coming from many genetic variants, each having a small effect size. Using summary statistics from highly powered GWAS, it is now possible to build polygenic scores (PGS) that can help predict the individual risk of common diseases. We here propose to improve PGS-based risk estimate by leveraging genetic ancestry, derived from genome-wide genotyping data. We provide a tool that 1) maps a discovery cohort to the principal component (PC) space of a map cohort; 2) runs an association analysis between the phenotype and the PCs in the map cohort space; 3) maps a target individual to the same PC space; 4) uses the association summary statistics to compute their PC score. We compared two models, one using PGS only and one using PGS and PC score jointly. Using large population-based cohorts from the UK (UK Biobank) and Switzerland (CoLaus), we show that this method improves genetic prediction for all tested phenotypes: <10% for blood pressure, BMI, and baldness, 16% for menarche age, 38% for height, 71% for menopause age, 138% for bone mineral density, 350% for educational attainment and 2800% for skin color. This approach could allow diverse populations to benefit from genomic medicine.

6.1 Background

6.1.1 Polygenic scores for clinical applications

Most common diseases of major public health importance have a complex genetic architecture [155, 156, 157, 158, 159, 160, 161, 162]. A polygenic score (PGS) (or polygenic risk score) is the weighted sum of the number of risk alleles carried by an individual. By predicting part of the individual risk of a disease, a PGS allows the stratification of people in different risk categories. Defining such categories has a potentially broad range of clinical applications. As an example, individuals in the top 0.5% polygenic score rank have a five-fold risk increase for CAD [7] that could be counterbalanced by encouraging healthy lifestyle or by pharmacological interventions [163]. PGS alone are already equal or superior to the best clinical risk model to predict prostate cancer, breast cancer, and type-1 diabetes[164, 165, 160]. Because PGS are likely to be integrated in clinical practice in the coming years, their inherent limitations should be urgently addressed [126, 49].

6.1.2 Polygenic scores portability between populations

The computation of PGS relies on the availability of genome-wide association study (GWAS) summary statistics[166]. The GWAS is run on a *base cohort* (or study/discovery cohort) and the PGS is calculated for a *target cohort* or target individual. For optimal PGS performance, the base and target cohort must have matching ancestries[167, 133]. Ancestry mismatches have been shown to result in weaker PGS predictive power for schizophrenia[131], T2D, CAD[132] and height[133]. This is due to both genetic and non-genetic factors. Genetic factors include genetic drift[133], different allele frequencies[168] and linkage disequilibrium patterns[168], presence of ancestry-specific variants[169] and variable effect sizes[170, 171]. Non-genetic factors include gene-environment interactions, family environment[172], age[173], socioeconomic status[174], and technical artefacts including different genotyping arrays[175]. Because of the systematic differences of PGS distribution between populations, there is a need to adjust statistical/clinical models with an ancestry-based risk parameter.

6.1.3 Disambiguation: ancestry and race

The *race* is a social construct which classifies people regardless of the biological component. Its meaning changed over time and varies between societies. The *genetic ancestry* is defined by the systematic associations between genetic variation and the history of populations and it can be useful for biomedical applications [176]. For example, some variants associated with greater risks of prostate cancer are more likely to be present in people with a higher African ancestry.

To develop a method using ancestry without being confounded by the notion of race, continuous variables accounting for the degrees in various ancestries are to be preferred to grouping

by categories. While clusters of individuals could still be observed, it must be label-free to prevent dubious interpretation with potential overlapping racial categories. We propose to use as a variable the genetically inferred ancestry (GIA).

6.1.4 Breaking down the phenotypic variance

Considering the phenotypic variance:

$$V_P = V_G + V_E$$

With V_G, the genetic variance, and V_E, the environmental variance.

For a cohort including diverse populations, we define $V_{G,Individual}$, the fraction of V_G coming from variants shared between ancestries and $V_{G,Ancestry}$, the fraction of V_G coming from variants that are ancestry-specific. The fraction $V_{G,Ancestry}$ is driven by the joint effect of ancestry-specific causal alleles.

Because causal ancestry-specific alleles will co-vary with non-causal ancestry-specific alleles, it is not possible in a GWAS to distinguish between a causal and non-causal locus in presence of population stratification. To avoid false-positive associations, it is necessary to correct for population structure using GIA. GIA can be calculated by running a principal component analysis (PCA) on the genome-wide genotyping data. The top eigenvectors (PC) capture the sample ancestry and can be added to the regression model [45]. Other risk parameters are typically included as covariates in a GWAS statistical model such as age or sex [177]. A typical GWAS model to estimate the effect size of each variant m on the phenotype with the p top PCs as covariates is:

$$phenotype \sim Variant_m + \sum_{i}^{p} PC_i + covariates \tag{6.1}$$

Because the PGS is computed from the GWAS variants summary statistics, it is a composite predictor of the $V_{G,Individual}$ fraction of V_G only.

Similarly to GWAS, we define the PC association study (PCAS) statistical model to estimate the effect size of each of the p top PCs on the phenotype as:

$$phenotype \sim \sum_{i}^{p} PC_i + covariates \tag{6.2}$$

We define the PC score (PCS) which is computed from the PCAS summary statistics, it is a

composite predictor of mainly $V_{G,Ancestry}$ but potentially of $V_{E,cultural}$ a component of the environmental variance.

We define $V_{E,cultural}$, the fraction of V_E that stem from the cultural habits of individuals who modify their environment accordingly such as: $Cultural \rightarrow Environment$. We define $V_{E,race}$, the fraction of V_E driven by the consequences of being assigned to a racial group in a specific society, such as: $Environment \rightarrow Individuals$. Both are optionally associated with ancestry. While $V_{E,cultural}$ can be transferred from one study in a given environment to another, $V_{E,race}$ will depend solely on the society. Because they can be correlated, disentangling both can be necessary when mixing data coming from different societies necessitating inputs from social scientists.

Altogether, the phenotypic variance can be broken down as:

$$V_P = \underbrace{V_{G,Individual} + V_{G,ancestry} + V_{G,epistasis}}_{Genetic factor} + \underbrace{V_{E,cultural} + V_{E,race} + V_{E,other}}_{Environmental factor} \tag{6.3}$$

With $V_{G,epistasis}$ doing the closure with the broad sense heritability taking into account non-additive variance and $V_{E,other}$ accounting for other environmental factors.

While the association between variables such as age and sex can easily be estimated in a discovery cohort and implemented as risk parameters in a clinical model, the GIA is more complex. The PCs order resulting from the PCA on the base cohort is specific to its composition in diverse populations and their relative proportion. Because the base cohort is normally not accessible, the meaning of each PCs cannot be deciphered and therefore it cannot be transferred. Here, we offer a method to circumvent that issue and use the association between the ancestry and the phenotype estimated by the PCAS in a potential risk parameter for a clinical model applied to any target.

Our method relies on the use of an intermediate *map cohort* which needs to have two main properties. First, it must be available for *the geneticists* (owning the base cohort, producing summary statistics) and *the clinicians* (owning the target cohort, calculating scores). Second, it must have enough diversity to separate the populations present in the base and target cohorts.

The method presented here allows 1) to capture trait differences associated with ancestry in a label-free manner along a continuous axis. As an example, for CAD the actual best integrative risk model is QRISK2 and already includes ethnicity as a categorical risk parameter. While PRS-CAD is likely to be the first PGS that will be implemented in clinical care, we argue that ancestry inferred from genetic data would make medical investigation more reliable and that continuous measurement should improve the quality of the score; 2) to publicly share as

metadata of genetic studies the precise ethnic components of the individuals.

6.2 Materials and methods

6.2.1 Map cohort: One Thousand Genome Project

We use the 1000 Genome (1KG) phase 3 dataset. It contains 2,404 samples of diverse ancestries classified in 5 super populations, Europeans (EUR, n=503), Africans (AFR, n=661), Native Americans (AMR, n=347), East Asian (EAS, n=504), and South Asians (SAS, n=489) and is publicly available[152, 27].

6.2.2 Base and target cohort: UK Biobank

We derive our base and target cohorts from the UK Biobank (UKB). The UKB is a fairly mixed cohort with 488,000 individuals including 81,000 of non-white British ancestry. The recruitment process has been described previously [178]. Briefly, the participants attended one of 22 UKB assessment centres located throughout England, Scotland and Wales between 2006 and 2010. All participants completed a touchscreen questionnaire and a verbal interview and had a range of physical measurements and blood, urine and saliva samples taken for long-term storage. Participants were 40-69 years old enrolment (mean age ± SD: 56.5 ± 8.1) with 54.2% of female.

Genotyping and imputation of UKB participants have been fully described by Bycroft et al. [179, 6]. Briefly, samples were genotyped using the UK BiLEVE Axiom array (Affymetrix) (10.2%), or the UK Biobank Axiom array (Applied Biosystems). Genotypes were phased using SHAPEIT3 with the 1KG phase 3 dataset as a reference, then imputed with IMPUTE4 using the Haplotype Reference Consortium data, 1KG phase 3, and UK10K data as references. People were removed if they had non-matching sex between the one submitted and the one determined using genotypes calling on the Y chromosome and the sex-specific region of the X chromosome; if they had non-XX or XY sex chromosome karyotype; if they were flagged as outliers in terms of heterozygosity or missing rates.

Two base and two target cohorts were generated. One set of base and target cohorts with samples of white British ancestry only - *base/target-WBO* – another set with samples of all ancestries - *base/target-UKB-all*. The split between base and target cohort left on average 10,000 samples in the target cohort for an overall sample size varying with respect to the phenotype. We avoided over-representation of individuals of white British ancestry in the *target-UKB-all* (see supplementary section). The phenotypes were selected based on their high degree of stratification between the world super populations as characterized by the Global Distribution of Genetic Traits (GADGET)[180]. If needed, phenotypes were normalized using residualization by sex and/or rank-based inverse normal transformation. When the degree of a phenotype is defined categorically it is transformed into a discrete variable. The

details of the phenotypes is given on Tab.6.1.

Table 6.1 – Phenotype details

Phenotype	F_{STAT}	**Type**	**Transformation**	**Sample size**	**White only**
Skin color	774	Cat(6)	Cont	478,929	403,189
Menopause age	499	Cont	INV	149,435	127,370
HBMD*	404	Cont	INV/Sex Res	274,000	237,166
Blood pressure	319	Cont	INV/Sex Res	455,457	381,383
Menarche age	172	Cont	INV	255,616	149,435
Baldness	162	Cat(4)	Cont	220,192	186,127
BMI	64	Cont	INV/Sex Res	484,587	406,956
Height	55	Cont	INV/Sex Res	485,043	407,318
Educational	NA	Cont	INV	418,573	350,305

*Distribution of the different phenotypes including the samples size; the ancestry (all samples vs white only); the type (continuous or categorical); the transformation procedure (INV: inverse normal transformation; Sex Res: residualised on sex; Cont: transformed from categorical phenotype to continuous); FSTAT (degree of the phenotype difference between super populations). *Heel bone mineral density*

6.2.3 External target cohort: CoLaus

We used the *cohort Lausannoise* (CoLaus), a population-based research study initiated in 2003 from Lausanne, Switzerland, as an independent target cohort. It enrolled 6,188 individuals of Swiss origin, 35 to 75 years old at enrolment (mean ± SD: 51.1 ± 10.9) with 52.5% of female [181]. The institutional review boards of the University of Lausanne approved this study, and written consent was obtained from all participants.

DNA samples from 5'399 CoLaus participants were genotyped using the BB2 GSK-customized Affymetrix Axiom Biobank array. Quality control procedures described by Anderson et al. were applied[182]. Participants whose genetic sex did not match their self-reported gender were removed, and we confirmed that none were related, of non-European ancestry or with missing data rates > 5%. A total of 4,781 individuals were included for further analyses. Genotype imputation was performed using two independent reference panels: the HRC reference panel and the merged 1000 Genomes Phase 3 and UK10K reference panel[183, 184, 185]. Both phasing and imputation were performed on Sanger Imputation Service (https://imputation.sanger.ac.uk). The imputed data set consisted of 93,091,315 SNPs.

We used standing height, body mass index (BMI), and blood pressure as phenotypic outcomes.

6.2.4 Method

The four steps of the method can be visualized in Fig.6.1. The separation between the two groups of scientists are materialized by the dashed line, with the group owning the base cohort and running associations studies on the left-hand side and the group owning the target cohort and computing phenotypic scores on the right-hand side. Only the map cohort in the middle is accessible to both.

1. Mapping of the base and target cohorts on the map cohort. First, the base and target cohorts must be mapped on the principal component (PC) space of the map cohort. To do that, the PC space has to be based on an optimal SNP set selected with respect to both the base and the map cohort such as: 1) only the SNPs present in both cohorts are retained; 2) on the base cohort side, SNPs are filtered for imputation infoscore (INFO) > 0.9; on the map cohort side, SNPs are filtered for minor allele frequency (MAF) > 0.01 and pruned with $r = 0.5$ and $l = 500kb$[116]; 3) The intersection of these two subsets produces the *map SNP set*. On this optimal set of SNP references, a PCA is run on the map cohort to produce the set map-PC-loadings. On the base cohort side, this set of map-PC-loadings is used to project the base samples in the PC space of the map cohort, producing a new set of PCs that we call *map-PCs*. On the target cohort side, to project the samples using the PC-loading, some SNPs are likely to be missing. The linkage disequilibrium (LD) properties of the map cohort can be used to find the best tagging SNP. In our case, we used the ones which max the score given by $Log_{10}(INFO \times r^2)$ in a 500kb windows. Once the missing SNPs substituted, similarly the set of map-PC-loadings is used to project the target samples in the PC space of the map cohort to produce the corresponding *map-PCs*.

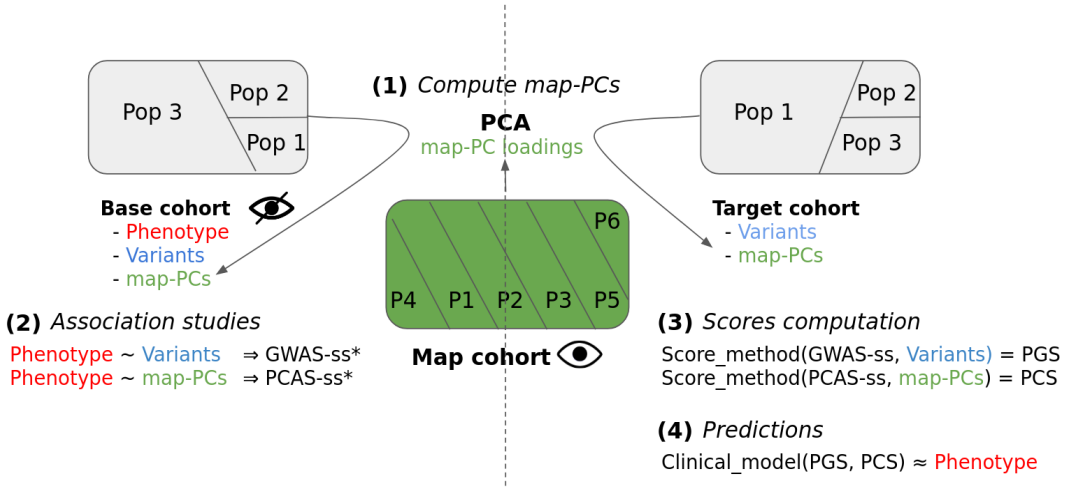
2. Associations analyses on the base cohort. In this second step, we run two association studies on the base cohort in parallel: the PCAS (Principal Components Association Study, as eq. 6.2) and the GWAS (eq. 6.1). For the PCAS, a multiple linear regression model based on scikit-learn implementation[186] is trained on the *map-PCs* to produce the PCAS summary statistics. For the GWAS, a linear mixed model based on BOLT-LMM[187] produces the GWAS summary statistics. The imputed variant are filtered for MAF> $1e^{-4}$ and INFO> 0.8. For both the GWAS and PCAS, the models are also adjusted for the age at recruitment, genetic sex and genetic array depending on the phenotype. For the GWAS, the PCs corresponding to the result of a PCA on UKB cohort - that we call *base-PCs* - are used to correct for ancestry. These base-PCs are part of the UK Biobank dataset[6]) and were calculated with the fast-PCA method[188]. The summary of the series of GWAS is available on supplementary materials Tab.6.2.

3. Computation of PCS and PGS. The PCAS and GWAS summary statistics obtained in the base cohort are used to compute respectively the PCS and the PGS point estimate of the samples in the target cohort. The PGS are computed with PRSice which uses a strategy based on variants clumping and p-value thresholding to determine the optimal set of SNPs to construct the PGS[125]. Briefly, the clumping step leverages the LD properties of the genome

to construct groups of SNPs (or clumps) bellow a given maximal p-value threshold. These LD properties have to be either computed from the target cohort or drawn from an external reference panel. In our case, we chose to use 1KG as a reference panel, with a maximal windows for each clump of 250kb and a r^2 limit of 0.01. From each clump, one SNP is retained to be included in the final set of SNPs used to build the polygenic score. PRSice then determine based on cross-validation which set of SNPs is the more efficient to build polygenic score for different p-value thresholds. The summary of the PGS is described on supplementary materials Tab.6.3. The PCS are computed from the PCAS summary statistics (or trained model) with scikit-learn.

4. Phenotypic predictions. PGS and PCS are two composite variables that stem from the same genome-wide genotyping data. Therefore, in order to estimate the predictive value of PCS, it is necessary to evaluate it jointly with PGS. We expect the combination of both within the same predictive model to increases the phenotypic variance explained (PVE).

Figure 6.1 – General workflow of the method



Representation of the different components and steps of the method. The map cohort is used to defined the PC space of both the base and target cohorts. The summary statistics of the two kinds of association studies performed in the base cohort (GWAS-ss and PCAS-ss) are used to compute the PGS and PCS on the target cohort. Finally, the two scores (PGS and PCS) are jointly used as predictive model risk parameters.

6.3 Results

6.3.1 Cohorts projection on the 1KG PC space

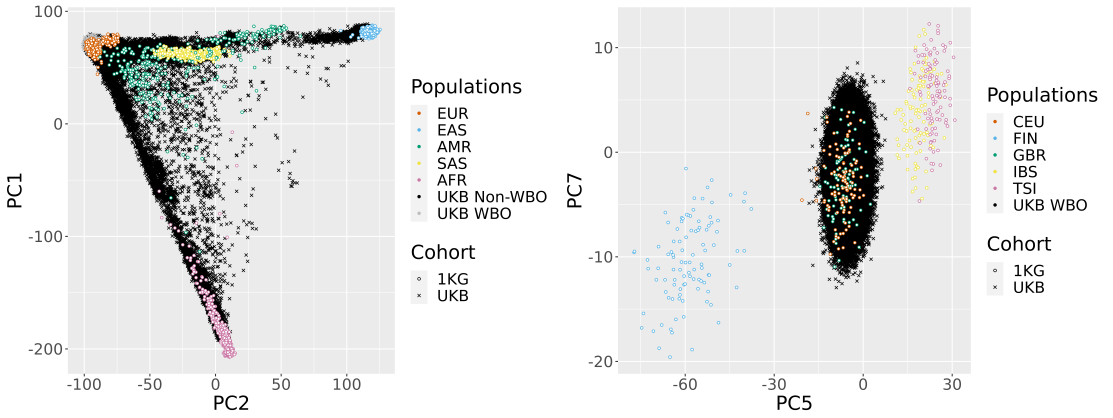
In the following section, we describe the PCs values corresponding to the map-PCs of the base and target cohorts plotted jointly with the corresponding PCs for the samples of 1KG. The Fig.6.2A shows the overlay for PC1 and PC2. The broad overlap between *UKB-all* and 1KG demonstrates the high diversity presents in UKB. The Fig.6.2B shows the equivalent

but for *UKB-WBO* and the subset of European samples in 1KG for PC5 and PC7, the axes discriminating the most samples from European populations (see supplementary section Tab.6.5). As expected, there is a sharp overlap between *UKB-WBO* and the British (GBR) cluster only. Similarly but with CoLaus, the Fig.6.2C shows the overlay with 1KG for PC1 and PC2 and validate that CoLaus is exclusively European. The Fig .6.2D shows the equivalent but for PC5 and PC7 and the subset of European samples in 1KG. Because we observe what is expected from the Swiss population - a central European population - it hints that substituting the missing variants in the target by the best-tagging ones did not introduce any seemingly excessive bias.

Figure 6.2 – Cohorts projection on the 1KG PC space

(A) Projection of UKB on 1KG: PC1-PC2

(B) Projection of UKB on 1KG: PC5-PC7



(C) Projection of the CoLaus on 1KG: PC1-PC2

(D) Projection of the CoLaus on 1KG: PC5-PC7

PC plot of base and target cohorts projected on the 1KG PC space. For UKB-all, map-PC1 and map-PC2 are plotted jointly with PC1 and PC2 of 1KG samples (A). For the UKB-WBO fraction, map-PC5 and map-PC7 are plotted jointly with the PC5 and PC7 of the 1KG samples of European ancestry (B). Similarly for CoLaus, map-PC1 and map-PC2 are plotted jointly with PC1 and PC2 of 1KG samples (C) and map-PC5 and map-PC7 are plotted jointly with the PC5 and PC7 of the 1KG samples of European ancestry(B)

We characterized the populations in UKB by clustering using the k-nearest neighbor method

on the first 7 map-PCs to form 4 clusters based on the super populations given by 1KG which are Europeans (EUR), Africans (AFR), South Asians (SAS) and East Asian (EAS). The center of each cluster along the 7 axes is given by the corresponding median on each axes of each super population in 1KG. UKB can be roughly broken down as, 464,031 EUR including 407,377 of White British ancestry; 9,109 AFR, 11,629 SAS, 2,626 EAS (see supplementary section Fig.6.6A). One limitation of using 1KG as a map cohort for UKB is the lack of samples from the Middle-East and North Africa. We expect these samples to correspond to the bulk observed between the European and African groups. It is interesting to note that while these samples are absent from UKB, they are seemingly mapped at a relevant position.

The same method was followed to classify the different samples of CoLaus based on the PC-mapped 5,7 and 8 into the corresponding closest European population catergory. The CoLaus samples can be classified as follow: 1500 British in England and Scotland (GBR); 1634 Northern and Western European (CEU); 552 Toscani in Italia (TSI); 694 Iberian Population in Spain (IBS); 411 Finnish in Finland (FIN) (see supplementary section Fig.6.6A).

6.3.2 Method evaluation in an optimal setup with UK Biobank

In the following section, the performances of our genetic predictors are assessed through linear regression models with either PGS alone or the combination of PGS (PolyGenic Score) and PCS (Principal Components Score). The models performances are measured in a 10 times 10-fold cross-validation on *target-UKB* with the mean phenotypic variance explained (PVE, or coefficient of determination (R^2)). The design of this statistical analysis is summarized in 6.3A.

We first studied the impact of the base cohort - *base-UKB-WBO* versus *base-UKB-all* - on a model with PGS alone for *target-UKB-WBO*. The Fig.6.3B shows similar performances for the different phenotypes with both base cohorts for $\alpha = 0.05$ (p-value = 0.51; H0: the increase/decrease with a given base cohort is equally likely to occur (p=0.5)).

We then assessed the predictive power of PGS and PCS together with *base-UKB-all* and *target-UKB-all* cohorts. We estimated the narrow-sense heritability in *target-UKB-all* with the genome-based restricted maximum likelihood method (GREML)[60, 3] to display the theoretical upper limit of the PVE by linear models. The Fig.6.3C shows that adding the PCS parameter increases the PVE for all the phenotypes with varying magnitude. There is a slight increase for blood pressure (4.3%), BMI (6.7%) and baldness (9.1%); a greater one for menarche age (15.6%), height (37.5%), and menopause age (70.8%); it does more than double for mineral bone density (137.5%) and educational attainment (350%); it is exacerbated for skin color (2800%) where most of the PVE comes from PCS. Skin pigmentation is more polygenic in populations of African ancestry, therefore it is expected that well-powered studies are needed to capture its many underlying genetic variations. Because of the lack of African samples, the PGS alone performs poorly in comparison to the PCS[189]. We applied the same method based on *base-UKB-WBO* and *target-UKB-WBO* cohorts (see supplementary section Fig.6.7). We do not observe any gain by adding PCS except a very weak one for skin color. When the

base and target cohorts are both very similar and homogeneous there is little to gain by adding PCS.

We then estimated the impact of the PCs - map-PCs versus the base-PCs - used to construct the PCS on its resulting predictive power. The Fig.6.3D shows the relative increase/decrease rate when the PCS added to PGS is from base-PC versus map-PCs. In theory, the base-PCs represent the upper limit of the gain and therefore should do at least equally well overall. As expected, we observe a slight mean gain increase with the base-PCs of 3.57%, while suggestive it is not significant for $\alpha = 0.05$ (p-value = 0.089; H0: there is no increase when base-PCs is used for PCS instead map-PCs). The slightly greater gain for educational attainment with base-PCs (from 0.042 to 0.056) could come from the lack of relevant populations in 1KG having a different mean educational attainment. On the other hand, the slight decrease for menopause age (from 0.015 to 0.012) can only be explained by random fluctuations, a hint to dispel over-interpretation.

6.3.3 Generalization of the method using an independent cohort

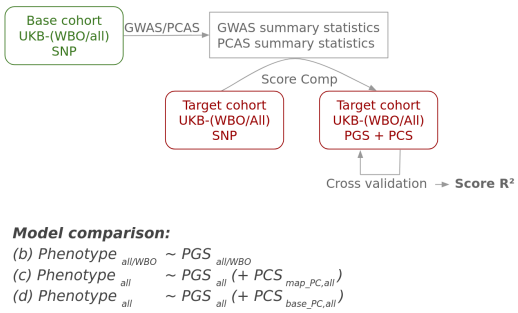
In this section, the models are assessed on CoLaus in order to estimate the potential of generalization of our method with independent samples in a study design similar to a clinical setup (Fig.6.4A). The set of SNPs used to build the PGS in CoLaus is selected by PRSice through cross-validation on *target-UKB-all/WBO*. An intermediate step of calibration of the predictive model is done on *target-UKB-all/WBO* where the coefficients for the PGS and PCS risk parameters are estimated. Once calibrated, the predictive model is used to generate the phenotypic scores on *CoLaus* which are assessed against the true phenotype through R^2 the coefficient of determination.

The results based on *UKB-all* for GWAS and calibration are shown in Fig.6.4B, we observe a gain of PVE for height of 38.4% when the PCS is added in the predictive model, a similar magnitude to what was observed in the previous section (37.5%). As expected from the previous results, blood pressure and BMI are weakly associated with PCS and do not improve the predictions in CoLaus. We note that even if they end up having a negative impact, it is very negligible. As expected with *UKB-WBO*, because the PCS is not significantly associated at the calibration step, the PCS is legitimately discarded from the predictive model (see supplementary section Fig.6.4C).

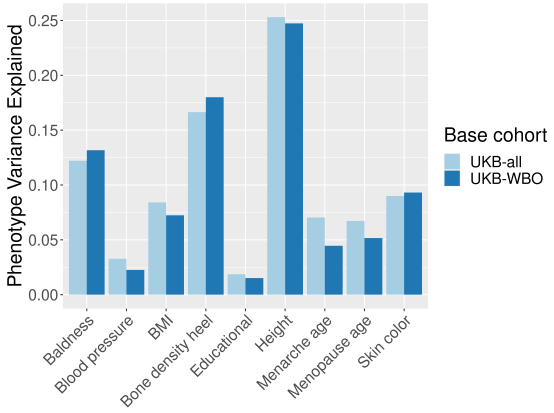
It is interesting to note that while the PGS alone for height shows a higher PVE in *UKB-WBO* of 0.216 (see supplementary section Fig.6.4C) than in *UKB-all* with a PVE of 0.185 (see Fig.6.4B)), the best-fit between are roughly equivalent with a PVE of 0.226 for *UKB-all* and a PVE of 0.219 for *UKB-WBO* (see supplementary section Tab.6.4). The difference that we observe is introduced by the greater phenotypic variance and population diversity in *target-UKB-all* than *target-UKB-WBO* at the calibration step. Because *target-UKB-all* is more distant than *target-UKB-all* on average to *CoLaus*, it produces a shift in the estimated intercept (see Fig.6.5B vs Fig.6.5D). We can see on Fig.6.5C that this difference is compensated when the

Figure 6.3 – Method evaluation in an optimal setup with UK Biobank

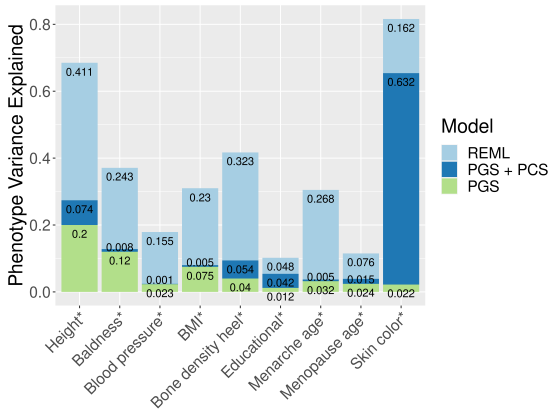
(A) Study design with UKB



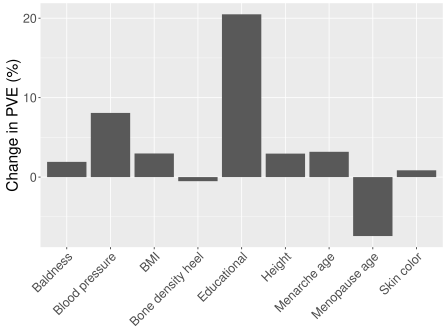
(B) Impact of base UKB



(C) PVE with PGS and PCS in UKB-all



(D) Evolution of PVE with PCS from base-PCs



*Following the study design described on (A) including 1) the GWAS and PCAS; 2) the score computation for target-UKB-all/WBO and CoLaus; 3) the 10 times 10-fold cross-validation on target-UKB-All/WBO cohorts. The barplot (B) shows the difference of PVE by PGS w.r.t the base cohort used for the GWAS. The barplot (C)) show the portion of PVE with PGS only (green) PGS combined with PCS (dark blue) in comparison to the trait heritability (light blue) based on UKB-all. *The PCS was significantly associated at the calibration step.*

PCS is included in the model as it allows to adjust for the diversity in *UKB-all*. The score is maximal for height in CoLaus when both PCS and PGS are used jointly with GWAS and PCAS conducted on *UKB-all* for a maximal PVE of 0,238. This score can be contrasted with the PVE of 0.216 following a standard strategy where only PGS is used from a GWAS restricted on the samples of *UKB-WBO* to reduce the population diversity. PCS allows the phenotypic score improvement in two ways. First, by providing the association with the degree of ancestry. Second, by providing the baseline value for the phenotype w.r.t the ancestry to correct the intercept. In that sense, the PCS is similar to a map between a given ancestry and a mean phenotypic value. Providing the model of the fitted PCS with the corresponding intercept is potentially a great gain as intercepts are usually not shared with GWAS summary statistics.

We conclude positively on the generalization of our method. The best results are obtained when the GWAS and PCAS steps were done on *UKB-all*, a cohort yet structurally more divergent from CoLaus than *UKB-WBO*.

To characterize further the predictive gain stemming from map-PCs, we investigated their association with the height phenotype. First, we see that the map-PCs separating the most the CoLaus samples is the PC5 which is the axis showing the highest standard deviation (see supplementary section Tab.6.5). We see with 1KG populations that PC5 also corresponds to the axis separating the most the populations of European ancestry. Second, we compared the association coefficients of the different PCs with the height phenotype in CoLaus and found out that PC5 is the one with the biggest coefficient. Indeed, PC5 alone explains up to 10% of the height variation in CoLaus. Finally, we investigated the association between PCS and height. The Fig.6.5A shows the comparison of the best fit between PCS and height versus the predictive model calibrated on *target-UKB-all*. Besides the matching profiles, we see that PCS discriminates between the different labelled samples of CoLaus, previously defined from European sub-populations. 6.5A.

6.4 Discussion

Because of the lack of diversity in the populations included in GWAS studies, the implementation of genomic medicine in clinical care risk to exacerbate health disparities[190]. According to the GWAS catalog, despite the fact that people of European ancestry make up only 16% of the world population, they represent 79% of the population in GWAS[191, 190]. The representativity has improved - in 2009 as much as 96% of GWAS people's ancestry was European[192] - but in an unbalanced way, mainly to the benefit of East Asian people. Only 3.8% of all cohorts include Africans, Hispanics, or indigenous peoples. Consequently, the predictive power of PGS is lower in these underrepresented populations[193]. There is a clear need to encourage research on non-European populations to balance out the current lack of diversity, for example through targeted grant allocation.

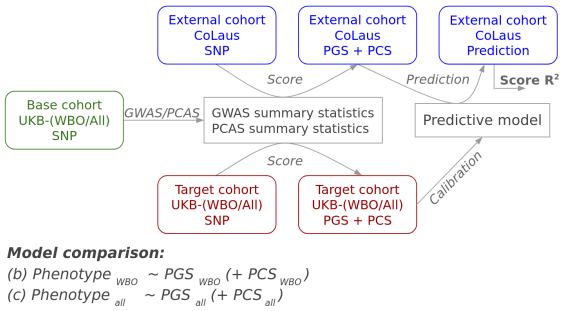
We show in this paper that it is feasible to already increase the predictivity of PGS in diverse populations, using today's existing genomic resources.

Nevertheless, it has been reported that the choice of the base cohort can impact the PGS quality in different ways. On the one hand, the statistical power increase resulting from broader inclusion criteria can improve the effect size estimate for variants widely shared across populations[129]. On the other hand, due to the limitations of linear models, the heterogeneous genetic effects between populations can produce an epistatic signal that can weaken the estimation of the effect sizes of variants. Nevertheless, previous works suggested that the best performing PGS comes from meta GWAS integrating populations of different origins[194].

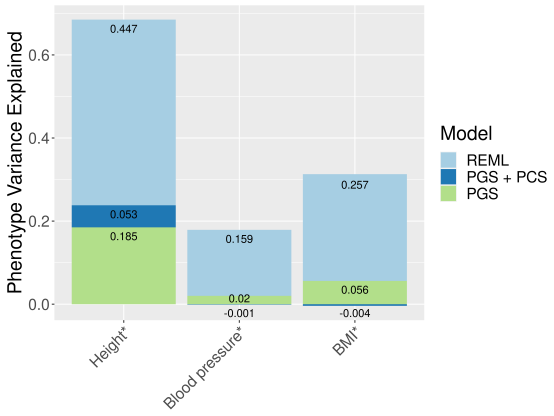
The ClinGen Complex Disease Working Group has defined a standard method to report risk

Figure 6.4 – Method evaluation in a clinical setup with CoLaus

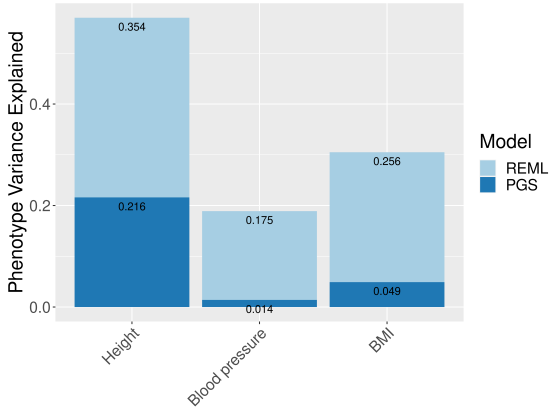
(A) Study design with CoLaus



(B) PVE in CoLaus by PGS and PCS from base-UKB-all



(C) PVE in CoLaus by PGS and PCS from base-UKB-WBO



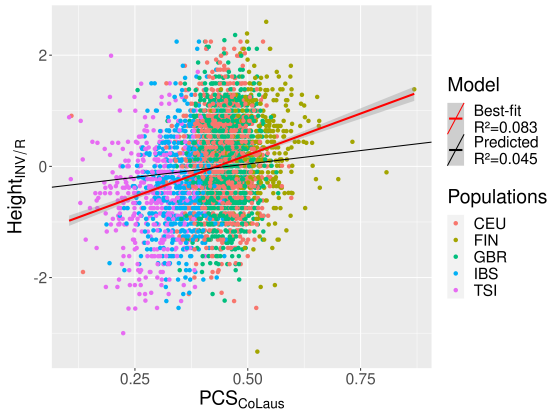
*Following the study design described on (A) including 1) the GWAS and PCAS; 2) the score computation for target-UKB-all/WBO and CoLaus; 3) the calibration of the predictive model on target-UKB-All/WBO cohorts; 4) the computation of the phenotypic score for CoLaus with the predictive model; 5) the evaluation of the phenotypic scores. The barplots show the portion of PVE with PGS only (green) PGS combined with PCS (dark blue) in comparison to the trait heritability (light blue) based on UKB-all (B) or UKB-WBO (C). *The PCS was significantly associated at the calibration step.*

models based on PGS[195] in collaboration with the Polygenic Score Catalog. The Polygenic Score Catalog is a rapidly growing repository for GWAS summary statistics (PGS file)[196]. Such a repository could also host supplementary data useful to compute risk parameters such as PCAS summary statistics (PCS file), PC-loadings, or metadata such as map-PCs. Today, researchers are strongly encouraged to share GWAS summary statistics to allow meta-analyses and accelerate research. Similarly, we encourage researchers to share the data to allows the computation of PCS. Doing it would increase the portability of GWAS summary statistics.

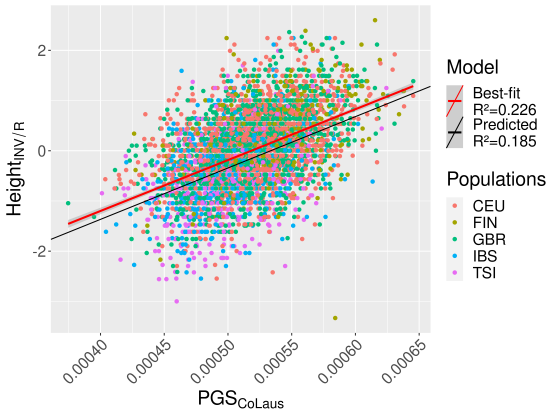
The assessment of an individual's ancestry has long been done using methods based on Self-Identified Race/Ethnicity (SIRE), which have important limitations[197]: they can be confounded by the variable definition of the ancestry categories; they do not offer a clear solution for admixed individuals who are predicted to become a greater fraction of the world population[198]; and they do not solve the problem of people ignoring their ancestry. Our method allows each sample to be characterized accurately with respect to its ancestry.

Figure 6.5 – Association between risk parameters and height in CoLaus

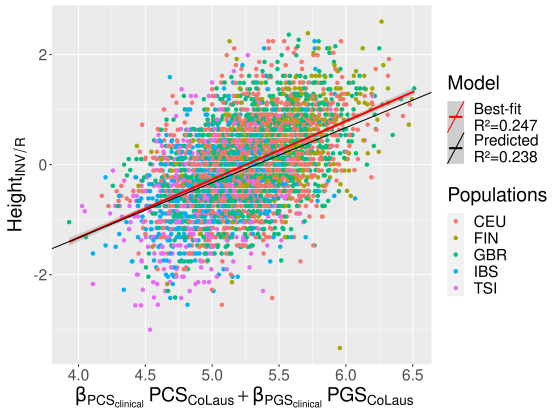
(A) Association between PCS and height in CoLaus, based on UKB-all



(B) Association between PGS and height in CoLaus, based on UKB-all



(C) Association between PCS + PGS and height in CoLaus, based on UKB-all



(D) Association between PGS and height in CoLaus, based on UKB-WBO

Representation of the association between height and PCS (A), PGS (B), PCS + PGS (C) based on UKB-all and PGS (D) based on UKB-WBO. The best-fit given by the regression line (red) has to be compared with the one corresponding to the predictive model calibrated on target-UKB-all(A,B,C)/WBO(D) (black)

Besides association studies, map-PCs can be used to share an accurate description of the diversity present in a discovery cohort as a new kind of study metadata. These data could be useful to assess the compatibility between available GWAS summary statistics of a discovery cohort and a targeted individual/cohort.

It is likely that PCs are not linearly associated with the phenotype. Further development relying on more sophisticated statistical methods encompassing putative interactions between gene variations could achieve higher predictive power with map-PCs and trained model could be shared.

We expect our method to be more sensitive if there is an excessive divergence between the base cohort and the target environments. In general, because the socioeconomic status correlates strongly with $V_{E,race}$[199] the phenotypes which are associated with socioeconomic status are going to be more impacted by a mismatch between the base and target cohort. Therefore, environmental factors should be carefully considered before including PCS in a predictive model depending on the phenotype.

The method we introduced is also a quick and efficient way to produce an equivalent to PCA for big cohorts. During the PCA computation, the computation of the covariance matrix scales as $O(min(N^2p, Np^2)$, with N for the samples size, and p for the number of SNPs used; and the eigendecomposition scales as $O(min(N^3, p^3)$. The fast-PCA tool[188] allows to speed up the process through approximations producing a number of m top PCs. This fast algorithm is mandatory to produce the top PCs for cohorts as big as UKB. Here, once the PCA on the map cohort has been done, mapping a cohort only scales as $O(Npm)$ with m for the number of PCs needed.

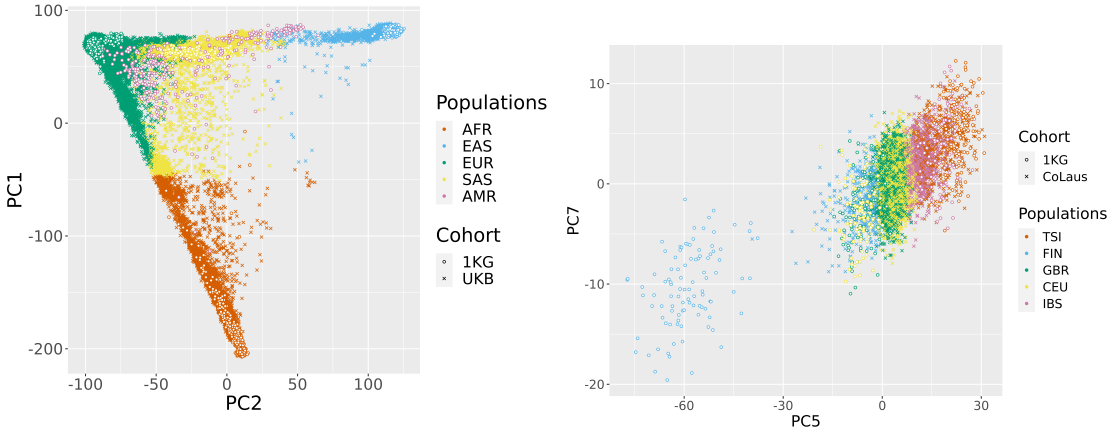
6.5 Conclusion

We introduced a method that fosters predictive models based on PGS to promote their clinical application in diverse populations. We are at a pivotal moment for medicine: large-scale personal data can start being efficiently used to develop more individualized approaches to disease prevention and treatment. Ensuring equitable access to the advances promised by genomic medicine is a major responsibility for the biomedical research community.

6.6 Supplementary materials

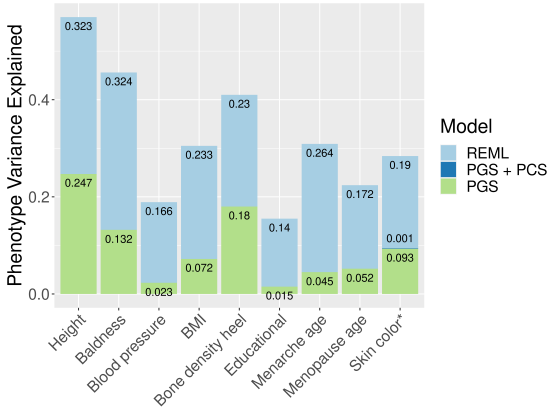
Figure 6.6 – Projection of labeled cohorts

(A) Map-PC1,map-PC2 of UKB and PC1,PC2 of 1KG (B) Map-PC5,map-PC7 of CoLaus and PC5,PC7 of 1KG



PC plot of cohorts projected through our method on the 1KG PC space. The samples have been labeled following our k-means based method on 1KG populations. For UKB, map-PC1 and map-PC2 are plotted jointly with PC1 and PC2 of 1KG samples (A). Similarly for CoLaus, map-PC5 and map-PC7 are plotted jointly with PC5 and PC7 of 1KG samples (B).

Figure 6.7 – PVE in UKB with PGS and PCS based on UKB-WBO



Following the study design described on (A) including 1) the GWAS and PCAS; 2) the score computation for target-UKB-all/WBO and CoLaus; 3) the 10 times 10-fold cross-validation on target-UKB-all/WBO cohorts; The bar plots show the portion of PVE with PGS only (green) PGS combined with PCS (dark blue) in comparison to the trait heritability (light blue) based on UKB-WBO

Table 6.2 – PGS details from UKB to UKB

	Base all		**Base WBO**	
Phenotype	h^2_{GREML}	λ	h^2_{GREML}	λ
Skin color	0.221	1.184	0.284	1.216
Menopause age	0.228	1.130	0.224	1.127
HBMD*	0.409	1.372	0.410	1.385
Blood pressure	0.187	1.393	0.189	1.398
Menarche age	0.310	1.345	0.309	1.343
Baldness	0.440	1.280	0.456	1.292
BMI	0.301	1.682	0.305	1.727
Height	0.558	2.087	0.570	2.223
Educational	0.147	1.288	0.155	1.286

Details of the GWAS run with BOLT-LMM. The narrow-sense heritability is estimated with GREML[60, 3]. The lambda indicates inflation. Elevated inflation has already been reported for a highly polygenic trait on UKB [130]

Table 6.3 – PGS details from UKB to UKB

	Target all			**Target WBO**					
	Base all						**Base WBO**		
Phenotype	**Thres**	**#SNPs**	R^2	**Thres**	**#SNPs**	R^2	**Thres**	**#SNPs**	R^2
Skin color	3.5e-4	34,127	0.0229	5e-5	4,235	0.0921	5e-5	2,151	0.0958
Menopause age	5e-5	1,050	0.0251	1e-4	1,296	0.0721	1e-4	1,133	0.0566
HBMD*	5e-5	5,002	0.0411	1.5e-4	5,404	0.1711	1e-4	4,209	0.1847
Blood pressure	2e-4	5,153	0.0231	0.003	20,508	0.0344	8.1e-3	37,888	0.0245
Menarche age	5e-5	3,247	0.0322	2e-4	4,258	0.0729	5e-5	1,914	0.0481
Baldness	5e-5	3,694	0.1210	5e-5	3,357	0.1246	1e-4	3,697	0.1347
BMI	2e-4	9,089	0.0751	0.0035	31,489	0.0861	4.5e-4	10,676	0.0742
Height	5e-5	14,745	0.2007	7e-4	26,629	0.2596	0.0018	35,561	0.2550
Educational	1.5e-4	2,739	0.0127	0.0051	28,096	0.0202	0.01	46,566	0.0167

Details of the PGS parameters produced by PRSice for the combination of base and target cohorts (base-UKB-all/target-UKB-all, base-UKB-all/target-UKB-WBO, or base-UKB-WBO/target-UKB-WBO). The p-value threshold for each PGS is determined by PRSice through cross-validation on the target-UKB cohort, limiting the PGS to a restricted number of SNPs. The R^2 reported correspond to the best-fit between the PGS and the phenotypes of the target-UKB cohort.

Table 6.4 – PGS details from UKB to CoLaus

	With base UKB all			**With base UKB WBO**		
Phenotype	**Thres**	R^2	**#SNPs**	**Thres**	R^2	**#SNPs**
Height	5e-5	0.2260	14,430	0.00015	0.2190	17,834
Blood pressure	2e-4	0.0206	5,052	0.0057	0.0144	31,254
BMI	2e-4	0.0646	8,897	0.0009	0.0526	15,131

Details of the PGS parameters produced by PRSice for the combination of base and target cohorts (base-UKB-all/target-UKB-all or base-UKB-WBO/target-UKB-WBO). The p-value threshold for each PGS is determined by PRSice through cross-validation on the target-UKB cohort, limiting the PGS to a restricted number of SNPs selected to be also present in CoLaus. The R^2 reported correspond to the best-fit between the PGS and the phenotypes of the CoLaus cohort.

Table 6.5 – CoLaus map-PC characteristics.

PC	**SD**	β	**p-val**
PC1	3.83	0.00413	3.16e-8
PC2	2.81	0.00465	0.353
PC3	3.37	0.00440	0.311
PC4	2.28	0.00534	0.611
PC5	6.77	0.00283	9.69e-13
PC6	2.16	0.00802	0.0322
PC7	2.84	0.00570	0.126
PC8	6.33	0.00609	0.0441
PC9	6.00	0.00624	0.129
PC10	5.22	0.00321	3.21e-3
PC11	3.21	0.00486	0.540
PC12	5.17	0.00327	0.0565
PC13	5.55	0.00253	2.13e-6

Details of all the top map-PCs significantly associated with height after Bonferroni threshold adjustment for $\alpha = 0.05$.

7 Discussion

In my PhD, I have analyzed the genetic architecture of complex human traits from a variety of angles: I tested the hypothesis that competition for cellular resources could explain the omnigenic model of complex trait variation; I organized the first crowdsourcing challenge of genomic prediction; I used deep learning to assess the potential impact of epistatic interactions on complex phenotypes; finally, I developed a method to improve polygenic score calculation in diverse populations and make genomic medicine more inclusive.

7.1 Going further to characterize the omnigenic model

Following my interest in the omnigenic model, I had the opportunity - through the SNSF Doc mobility program - to work for 6 months in Jonathan Pritchard's lab at Stanford, where the idea was born. Although we have found no mathematical evidence to support that our initial hypothesis – that a competition for intracellular resources explains the indirect impact of peripheral genes – is a key element of omnigenic architecture, by answering the question we have narrowed the field of possibilities. The main hypothesis remains that most peripheral genes have a transregulatory effect on the core genes due to the dense connectivity gene regulatory network.

7.2 New horizons to interrogate the genome

The recent advances in large-scale data science has given rise to new study designs, such as crowdsourcing. Because of the nature of the data that geneticists work with, it is not straightforward to come up with projects based on crowdsourcing. Nevertheless, by taking advantage of the explosion of direct-to-consumer genetic testing, we were able to open up a challenge that gave us the opportunity to get hundreds of people with diverse skills to work on a genetic prediction problem. From this completely free entry challenge, it is not excluded that in the future an intermediate form of crowdsourcing challenge may occur again. For example, biobanks could create relatively large-scale challenges that are accessible to all those who have

gained access to their data and could encourage participants to work on specific genomic question. This could provide researchers with new opportunities to gain both awareness and money for their research, as data scientists already do on kaggle.com and aicrowd.com. The visibility we have received will hopefully be part of a broader movement to renew the study design paradigm in genomics research.

7.3 Integrating polygenic scores in genomic medicine

As we are at the dawn of the deployment of genomics in the clinical world, there are many challenges ahead, which I have decided to make the core item of this discussion.

7.3.1 PGS methods overview

New PGS computation methods

Nowadays, different flavors of Bayesian mixed-model are implemented in tools like BOLT-LMM or BayesR[200] and are becoming the standard to run GWASs. Those models rely on the prior estimation of the SNP heritability (h^2_{SNP}), which currently is by default following the GCTA-GREML model. This model assumes a uniform distribution of the heritability across the genome. It has been shown that more realistic heritability models such as the BLD-LDAK which includes both properties from the MAF and LD following LDAK[68, 69], and from functional annotations following LDSC[70, 73] improves the resulting GWAS summary statistics[201]. In turn, the quality of the GWAS summary statistics have a positive impact on the resulting PGS[202]. When constructing PGS from summary statistics, the classical method is to combine p-value thresholding and marker pruning on a validation set to select a set of optimal variants. New methods such as LDPred2[203] uses variants jointly with a correlation matrix rather than SNPs individually. As a result, it does not need the intermediate validation set to estimate the hyper-parameters and can incorporate long-range LD existing in the HLA region - particularly important for the prediction of risk in a series of psychiatric disorders. Other methods such as SBayesR[204] follow a similar idea by expressing a multiple linear regression likelihood as a function of GWAS summary statistics and a LD reference correlation matrix. SBayesR is also coupled with a mixture of normal distributions prior on the genetic effects that incorporate sparsity.

Clinical utility of PGS

From a simple blood or saliva sample, using DNA extraction and genotyping technologies, PGS can now be calculated for a wide range of traits and diseases. These technologies have become relatively inexpensive ($\sim 50-100$\$) and need to be used only once per patient to study all kind of traits. Their clinical value still needs to be demonstrated, but is likely to become more and more obvious as our knowledge of the associations between genetic variants and

human diseases keeps improving.

Complex diseases

While many Mendelian / monogenic diseases can be diagnosed based on the typing of a single or a few known high-penetrance variant(s), the determination of PGS for complex diseases will not have such strong predictive power. Firstly, because genetics is responsible for only part of the risk of common disease, the rest being made up of multiple environmental factors. Second, because the GWAS capture only a fraction of this genetic factor that is commensurate with the overall quality of the GWAS study in terms of statistical power, method used, data quality and control of statistical bias. Nevertheless, many medical decisions implemented in public health for prevention purposes are already motivated by existing clinical risk prediction models. For example, the individual risk of coronary artery disease depends on various factors such as biochemical, clinical, lifestyle or historical risk. Together, these known factors can be used to predict the 10-year risk of coronary artery disease and achieve good accuracy with an area under the curve (AUC) of 80-85%[205, 206, 207, 208]. Although the proposed PGS for coronary artery disease alone are less efficient, these models could benefit from the inclusion of PGS as an additional risk parameter.

There continues to be a strong interest in expanding cohorts that now exceed one million participants, as in the Million Veterans Programs or the 23andme database (more than 10 millions). Even if it will take a long time to functionally characterize all associations, they can already be used clinically in the form of PRS, which don't require understanding of the underlying biological mechanisms. Increasing the sample size also allows for the inclusion of more loci and a better estimate of their effect size, which improves the predictive power of the PGS. This trend will continue to reduce the remaining gap between the narrow heritability estimate and the variance explained by the PGS.

PGS which are not predictive enough today could in the near future become usable. Another factor that will improve PGS comes from the increasingly big reference panels based on whole-genome sequencing which allow the imputation of up to 150 million testable variants in GWAS. Because complex diseases are the result also of multiple non-genetic factors, the quality of their prediction scores will also be dependent on the following two factors: first, the statistical methods, including machine learning, which will be used to extract the most meaningful information from multiple prognostic factors including PGS; second, the capacity of new technologies to accurately measure a growing number of these environmental factors.

To optimize the clinical utility of PGS, it could initially be applied only to individuals who are already at a higher risk of disease due to other factors. In their case, PGS is more likely to provide additional information that can be helpful to make a clinical decision.

For rare disease and cancer

PGS may also be useful for monogenic disorders like Mendelian diseases, neurodevelopmental disorders and familial cancers, where the aggregate of common variants has been shown to be a modifier of the penetrance of the rare deleterious variants causing the diseases. For example, the risk of developing ovarian and breast cancers for carriers of BRCA1 mutations has been shown to be modulated by common variants[209, 210]. A model based solely on 303 variants obtained an AUC of 0.63 which, although modest, is still interesting considering that only a few breast cancers are caused by rare mutations of BRCA1 and BRCA2[159]. Here, PGS could be used to improve screening programs and define for each woman at what age screening should begin and at which intervals.

For noninvasive prenatal testing

Non-invasive prenatal tests (NPTs) based on cell-free DNA have been extremely rapidly adopted for the detection of Down syndrome[211] and other human trisomies (13, 18). Since NPTs may also be able to detect adult-onset diseases and non-disease traits, there will be many ethical challenges, including the possibility of selecting embryos based on the likelihood of complex diseases or non-disease traits. sub-optimal[212].

Pharmacogenetics

Pharmacogenetics studies how variants can impact an individual's response to treatment, as demonstrated by several known associations, e.g. between hypersensitivity to abacavir and HLA-B*57:01, or between severe cutaneous reaction to carbamazepine and HLA-B*15:02. Another avenue would be to immediately identify patients who will not respond to certain drugs in order to save valuable time to start effective treatment while avoiding wasting costly resources for the patient and the healthcare system.

The need for clinical translation

There is a constant generation of data from new GWAS studies involving larger cohorts, newly defined phenotypes, populations whose ancestry had not been studied or involving new statistical methods, as well as the definition of new clinical models integrating various non-genetic risk factor. For this reason, there is a need for clinicians to be regularly updated to use state-of-the-art science in their clinical practice. In years to come, intermediary services are likely to be developed that will ensure an efficient flow of information from researchers to clinicians. From the researcher's perspective, it will be necessary to constantly integrate and contextualize the data coming from new studies, i.e.: [A] To characterize the results of newly produced GWAS by validating their quality and comparing it with existing results, [B] To combine new GWAS output data with existing ones, in a similar way to meta-GWAS, in order to increase their predictive power, [C] To adapt the different GWAS output which typically is

the GWAS summary statistics but in the future is expected to diversify with alternative trained models to produce enhanced polygenic scores (e.g., trained artificial neural networks which integrated the Topologically Associated Domain (TAD) feature); [D] Finally, to integrate the PGS into existing clinical models and calibrate them. From the clinician's perspective, the required steps will be: [A] To characterize the patient and deduce the optimal model to use to produce a PGS (for example, based on the patient's ancestry); [B] To provide the clinician with different calibrated models to choose from based on the risk factors available to him/her; [C] And to calculate the corresponding PGS for the patient and most likely the clinical score and provide it on a visual interface with a detailed report to guide the clinician in his decision making.

Initiatives such as the PGS catalog aim to standardize the GWAS summary statistics reporting procedure and are a useful resource[197, 196] and a start in the right direction.

Inclusivity in genomic medicine

A major ethical requirement and practical challenge is to avoid exacerbating health disparities while deploying genomic medicine [8]. The future of genomic medicine will be defined by the ability to offer it to populations of all ancestries and socio-economic backgrounds. Contemporary genomic research has been mostly focused on populations of European ancestry. There is therefore an urgent need to adapt by taking up two challenges in parallel: first, the lack of research on non-European populations, which can be achieved by prioritizing research on under-studied populations through targeted funding; second, the difficulty of providing interpreted genomic information to the growing number of people with mixed origins. Mixed individuals cannot simply be attributed to one "genetic ancestry" group. Methods must therefore be developed that have the ability to provide every individual with tailored genomic medicine, irrespective of his or her ancestry.

The notion of ancestry group is a sensitive subject because it is closely linked to the idea of race and has already been the source of heated debates[213, 214]. The notion of race is a social construct that classifies people into different categories, regardless of their genetic make-up. The meaning and the number of categories that derive from it are specific to each society and evolve over time. In some countries, such as France, race is not even officially recognized. Because of the history of racism in medicine, it should be kept away from biological science and genetics[213]. But it has also been said that scientists, out of fear, should not leave a vacuum around this notion and make room for pseudo-science that could fill political agendas[214]. The most responsible and practical way forward for genetic researchers is probably to use genetically-inferred ancestry measurements, which place human beings in the same space along continuous axes of diversity. Not only does this avoid the creation of subcategories, but because the axes correlate with certain phenotypes, they can be used to improve the models and accelerate discoveries, also for under-studied and mixed people[176].

Another reason for moving away from European-centric studies is that the marginal value of

ever-growing cohort in the same population for gene discovery is lower. Moving towards understudied populations makes it more likely to discover variants with relatively high minor allele frequency and effect size, thereby identifying potential targets for therapies. In addition, populations of African ancestry have smaller blocks of LD, which should facilitate the fine mapping process of association signals that can be more difficult in European populations[215].

Conclusion

PGS have started to move from research to clinical implementation.

To date, there is a set of 59 genes that have been designated by the American College of Medical Genetics as clinically useful and that must be reported (in the USA) when genetic testing is performed for other purposes[216]. Potentially deleterious variants in these genes are typically called "secondary findings" and are supposed to be reported preemptively, when the patient has not yet developed the related phenotype. Along the same lines, the calculation of PGS, integrated with clinical scores, could become part of routine health follow-up to identify at-risk individuals who should be screened frequently for specific complex diseases.

A distant goal is to have genotyping or sequencing data available along with the rest of the relevant health data in the individual electronic health records (EHRs), and to use them on an ongoing basis to monitor the health of individuals. The clinical importance of PGS will vary throughout the lifespan depending on other factors such as age and exposure to various environmental factors. As such, its prognostic value will have to be assessed several times over a lifetime.

Because great knowledge comes with great responsibility, it is also necessary to anticipate certain ethical and philosophical questions that revolve around genomics. Several problems need to be anticipated, for example if genome sequencing becomes ubiquitous in coming years, through sequencing at birth or routine genome analysis in the context of healthcare, the public's understanding of genetic determinism must be realistic to avoid overestimations of the impact of genetic factors on many aspects of life, health and personality, which would be highly problematic and create undue stress. While we have discussed the interests and challenges of deploying PGS in clinical practice, it should be duly noted that there is also a need to limit their use to whenever they could be applied for discriminatory purposes. Banks and insurance companies have already shown an interest in the subject, the jurisdiction may consider that their use should be limited to the common good of improving public health in general [217]. This potential risk of discrimination could extend to the world of work, or even to dating platforms, thus deepening inequalities and sowing discord. It will be necessary to redefine what it means to be sick in relation to being at risk, when both could have similar consequences, for example by prescribing lifestyle changes. It is likely that some people will explicitly ask for the right not to know their risks, questioning the doctor's need to go or not go beyond the patient's will in the case of an immediate critical risk.

The results of genomic research have created an enormous amount of data that now needs to be translated to fulfill the promise of genome-based medicine through the development of a new range of tools for the clinician.

Bibliography

1. Fisher RA. XV.—The Correlation between Relatives on the Supposition of Mendelian Inheritance. en. Transactions of the Royal Society of Edinburgh 1919; 52:399–433. DOI: 10.1017/S0080456800012163. Available from: http://www.journals.cambridge.org/abstract_S0080456800012163 [Accessed on: 2017 Feb 3]

2. Boyle EA, Li YI, and Pritchard JK. An Expanded View of Complex Traits: From Polygenic to Omnigenic. en. Cell 2017 Jun; 169:1177–86. DOI: 10.1016/j.cell.2017.05.038. Available from: http://linkinghub.elsevier.com/retrieve/pii/S0092867417306293 [Accessed on: 2017 Sep 26]

3. Yang J, Lee SH, Goddard ME, and Visscher PM. GCTA: A Tool for Genome-wide Complex Trait Analysis. American Journal of Human Genetics 2011 Jan; 88:76–82. DOI: 10.1016/j.ajhg.2010.11.011. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3014363/ [Accessed on: 2016 Nov 15]

4. Lewis CM and Vassos E. Prospects for using risk scores in polygenic medicine. en. Genome Medicine 2017 Dec; 9. DOI: 10.1186/s13073-017-0489-y. Available from: https://genomemedicine.biomedcentral.com/articles/10.1186/s13073-017-0489-y [Accessed on: 2017 Dec 22]

5. Zuk O, Hechter E, Sunyaev SR, and Lander ES. The mystery of missing heritability: Genetic interactions create phantom heritability. en. Proceedings of the National Academy of Sciences 2012 Jan; 109:1193–8. DOI: 10.1073/pnas.1119675109. Available from: http://www.pnas.org/cgi/doi/10.1073/pnas.1119675109 [Accessed on: 2017 Feb 6]

6. Bycroft C, Freeman C, Petkova D, Band G, Elliott LT, Sharp K, Motyer A, Vukcevic D, Delaneau O, O'Connell J, Cortes A, Welsh S, McVean G, Leslie S, Donnelly P, and Marchini J. Genome-wide genetic data on 500,000 UK Biobank participants. bioRxiv 2017. DOI: 10.1101/166298. Available from: https://www.biorxiv.org/content/early/2017/07/20/166298

7. Khera AV, Chaffin M, Aragam KG, Haas ME, Roselli C, Choi SH, Natarajan P, Lander ES, Lubitz SA, Ellinor PT, and Kathiresan S. Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. en. Nature Genetics 2018 Aug. DOI: 10.1038/s41588-018-0183-z. Available from: http://www.nature.com/articles/s41588-018-0183-z [Accessed on: 2018 Aug 14]

8. Martin AR, Kanai M, Kamatani Y, Okada Y, Neale BM, and Daly MJ. Hidden 'risk' in polygenic scores: clinical use today could exacerbate health disparities. en. bioRxiv 2018 Oct :441261. DOI: 10.1101/441261. Available from: https://www.biorxiv.org/content/early/2018/10/11/441261 [Accessed on: 2018 Oct 29]

9. AlphaFold: Using AI for scientific discovery. ALL. Available from: /blog/article/AlphaFold-Using-AI-for-scientific-discovery [Accessed on: 2020 Dec 13]

10. Lander ES et al. Initial sequencing and analysis of the human genome. Nature 2001; 409:860–921. DOI: 10.1038/35057062

11. Venter JC et al. The sequence of the human genome. Science (New York, N.Y.) 2001; 291:1304–51. DOI: 10.1126/science.1058040

12. The International HapMap Consortium. A haplotype map of the human genome. Nature 2005; 437:1299–320. DOI: 10.1038/nature04226. Available from: http://www.nature.com/nature/journal/v437/n7063/full/nature04226.html [Accessed on: 2016 May 21]

13. Consortium T1GP. A map of human genome variation from population-scale sequencing. Nature 2010; 467:1061–73. DOI: 10.1038/nature09534. Available from: http://www.nature.com/nature/journal/v467/n7319/full/nature09534.html [Accessed on: 2016 May 21]

14. Reich DE and Lander ES. On the allelic spectrum of human disease. Trends in genetics: TIG 2001; 17:502–10

15. Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, and Manolio TA. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proceedings of the National Academy of Sciences of the United States of America 2009; 106:9362–7. DOI: 10.1073/pnas.0903103106

16. Bush WS and Moore JH. Chapter 11: Genome-Wide Association Studies. en. PLOS Computational Biology 2012; 8. Publisher: Public Library of Science:e1002822. DOI: 10.1371/journal.pcbi.1002822. Available from: https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1002822 [Accessed on: 2020 Dec 4]

17. Ardlie KG, Kruglyak L, and Seielstad M. Patterns of linkage disequilibrium in the human genome. Nature Reviews. Genetics 2002; 3:299–309. DOI: 10.1038/nrg777

18. Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, Blumenstiel B, Higgins J, DeFelice M, Lochner A, Faggart M, Liu-Cordero SN, Rotimi C, Adeyemo A, Cooper R, Ward R, Lander ES, Daly MJ, and Altshuler D. The structure of haplotype blocks in the human genome. Science (New York, N.Y.) 2002; 296:2225–9. DOI: 10.1126/science.1069424

19. Devlin B and Risch N. A comparison of linkage disequilibrium measures for fine-scale mapping. Genomics 1995; 29:311–22. DOI: 10.1006/geno.1995.9003

20. Fallin D and Schork NJ. Accuracy of Haplotype Frequency Estimation for Biallelic Loci, via the Expectation-Maximization Algorithm for Unphased Diploid Genotype Data. American Journal of Human Genetics 2000; 67:947–59. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1287896/ [Accessed on: 2016 May 21]

21. Chapman JM, Cooper JD, Todd JA, and Clayton DG. Detecting disease associations due to linkage disequilibrium using haplotype tags: a class of tests and the determinants of statistical power. Human Heredity 2003; 56:18–31. DOI: 73729

22. Stram DO. Tag SNP selection for association studies. Genetic Epidemiology 2004; 27:365–74. DOI: 10.1002/gepi.20028

23. Carlson CS, Eberle MA, Rieder MJ, Yi Q, Kruglyak L, and Nickerson DA. Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. American Journal of Human Genetics 2004; 74:106–20. DOI: 10.1086/381000

24. Li M, Li C, and Guan W. Evaluation of coverage variation of SNP chips for genome-wide association studies. European journal of human genetics: EJHG 2008; 16:635–43. DOI: 10.1038/sj.ejhg.5202007

25. Hirschhorn JN and Daly MJ. Genome-wide association studies for common diseases and complex traits. Nature Reviews Genetics 2005; 6:95–108. DOI: 10.1038/nrg1521. Available from: http://www.nature.com/nrg/journal/v6/n2/abs/nrg1521.html [Accessed on: 2016 May 21]

26. Gibbs RA, Belmont JW, Hardenbol P, Willis TD, Yu F, Yang H, Ch'ang LY, Huang W, Liu B, and Shen Y. The international HapMap project. Nature 2003; 426:789–96

27. Siva N. 1000 Genomes project. en. Nature Biotechnology 2008 Mar; 26. Number: 3 Publisher: Nature Publishing Group:256–6. DOI: 10.1038/nbt0308-256b. Available from: https://www.nature.com/articles/nbt0308-256b [Accessed on: 2020 Aug 2]

28. McCarthy S et al. A reference panel of 64,976 haplotypes for genotype imputation. Nature Genetics 2016 Aug; 48:1279–83. DOI: 10.1038/ng.3643. Available from: http://www.nature.com/doifinder/10.1038/ng.3643 [Accessed on: 2017 Dec 1]

29. Loos RJF. 15 years of genome-wide association studies and no signs of slowing down. en. Nature Communications 2020 Nov; 11. Number: 1 Publisher: Nature Publishing Group:5900. DOI: 10.1038/s41467-020-19653-5. Available from: https://www.nature.com/articles/s41467-020-19653-5 [Accessed on: 2020 Dec 12]

30. Duggal P, Gillanders EM, Holmes TN, and Bailey-Wilson JE. Establishing an adjusted p-value threshold to control the family-wide type 1 error in genome wide association studies. BMC Genomics 2008; 9:516. DOI: 10.1186/1471-2164-9-516. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2621212/ [Accessed on: 2016 May 22]

31. Nyholt DR. A Simple Correction for Multiple Testing for Single-Nucleotide Polymorphisms in Linkage Disequilibrium with Each Other. American Journal of Human Genetics 2004; 74:765–9. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1181954/ [Accessed on: 2016 May 22]

32. North BV, Curtis D, and Sham PC. A Note on the Calculation of Empirical P Values from Monte Carlo Procedures. American Journal of Human Genetics 2002; 71:439–41. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC379178/ [Accessed on: 2016 May 22]

33. Hochberg Y and Benjamini Y. More powerful procedures for multiple significance testing. Statistics in Medicine 1990; 9:811–8. DOI: 10.1002/sim.4780090710. Available from: http://onlinelibrary.wiley.com/doi/10.1002/sim.4780090710/abstract [Accessed on: 2016 May 22]

34. Oord EJCG van den. Controlling false discoveries in genetic studies. American Journal of Medical Genetics. Part B, Neuropsychiatric Genetics: The Official Publication of the International Society of Psychiatric Genetics 2008 Jul 5; 147B:637–44. DOI: 10.1002/ajmg.b.30650

35. Hirschhorn JN, Lohmueller K, Byrne E, and Hirschhorn K. A comprehensive review of genetic association studies. Genetics in Medicine 2002 Mar; 4:45–61. DOI: 10.1097/00125817-200203000-00002. Available from: http://www.nature.com/gim/journal/v4/n2/full/gim200210a.html [Accessed on: 2016 May 22]

36. Wacholder S, Chanock S, Garcia-Closas M, El Ghormli L, and Rothman N. Assessing the probability that a positive report is false: an approach for molecular epidemiology studies. Journal of the National Cancer Institute 2004 Mar 17; 96:434–42

37. Sham PC and Purcell SM. Statistical power and significance testing in large-scale genetic studies. Nature Reviews Genetics 2014 Apr; 15:335–46. DOI: 10.1038/nrg3706. Available from: http://www.nature.com/doifinder/10.1038/nrg3706 [Accessed on: 2016 May 9]

38. Lee S, Abecasis GR, Boehnke M, and Lin X. Rare-Variant Association Analysis: Study Designs and Statistical Tests. American Journal of Human Genetics 2014; 95:5–23. DOI: 10.1016/j.ajhg.2014.06.009. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4085641/ [Accessed on: 2016 Jun 1]

39. Moore JH. The Ubiquitous Nature of Epistasis in Determining Susceptibility to Common Human Diseases. Human Heredity 2003; 56. Publisher: Karger Publishers:73–82. DOI: 10.1159/000073735. Available from: https://www.karger.com/Article/FullText/73735 [Accessed on: 2020 Oct 5]

40. Moore JH and Ritchie MD. THe challenges of whole-genome approaches to common diseases. JAMA 2004 Apr 7; 291:1642–3. DOI: 10.1001/jama.291.13.1642. Available from: http://dx.doi.org/10.1001/jama.291.13.1642 [Accessed on: 2016 May 22]

41. Todd JA. Statistical false positive or true disease pathway? Nature Genetics 2006 Jul; 38:731–3. DOI: 10.1038/ng0706-731

42. Lake SL, Lyon H, Tantisira K, Silverman EK, Weiss ST, Laird NM, and Schaid DJ. Estimation and tests of haplotype-environment interaction when linkage phase is ambiguous. Human Heredity 2003; 55:56–65. DOI: 71811

43. Millstein J, Conti DV, Gilliland FD, and Gauderman WJ. A Testing Framework for Identifying Susceptibility Genes in the Presence of Epistasis. American Journal of Human Genetics 2006 Jan; 78:15–27. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1380213/ [Accessed on: 2016 May 22]

44. Balding DJ. A tutorial on statistical methods for population association studies. Nature Reviews Genetics 2006 Oct; 7:781–91. DOI: 10.1038/nrg1916. Available from: http://www.nature.com/doifinder/10.1038/nrg1916 [Accessed on: 2016 May 9]

45. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, and Reich D. Principal components analysis corrects for stratification in genome-wide association studies. en. Nature Genetics 2006 Aug; 38:904–9. DOI: 10.1038/ng1847. Available from: http://www.nature.com/ng/journal/v38/n8/full/ng1847.html [Accessed on: 2016 May 22]

46. Loh PR, Kichaev G, Gazal S, Schoech AP, and Price AL. Mixed-model association for biobank-scale datasets. en. Nature Genetics 2018 Jun :1. DOI: 10.1038/s41588-018-0144-6. Available from: https://www.nature.com/articles/s41588-018-0144-6 [Accessed on: 2018 Jun 23]

47. Yu J, Pressoir G, Briggs WH, Vroh Bi I, Yamasaki M, Doebley JF, McMullen MD, Gaut BS, Nielsen DM, Holland JB, Kresovich S, and Buckler ES. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. en. Nature Genetics 2006 Feb; 38. Number: 2 Publisher: Nature Publishing Group:203–8. DOI: 10.1038/ng1702. Available from: https://www.nature.com/articles/ng1702 [Accessed on: 2020 Dec 12]

48. Yang J, Weedon MN, Purcell S, Lettre G, Estrada K, Willer CJ, Smith AV, Ingelsson E, O'Connell JR, Mangino M, Mägi R, Madden PA, Heath AC, Nyholt DR, Martin NG, Montgomery GW, Frayling TM, Hirschhorn JN, McCarthy MI, Goddard ME, and Visscher PM. Genomic inflation factors under polygenic inheritance. European Journal of Human Genetics 2011 Jul; 19:807–12. DOI: 10.1038/ejhg.2011.39. Available from: http://www.nature.com/doifinder/10.1038/ejhg.2011.39 [Accessed on: 2016 May 10]

49. Lambert SA, Abraham G, and Inouye M. Towards clinical utility of polygenic risk scores. en. Human Molecular Genetics 2019 Nov; 28:R133–R142. DOI: 10.1093/hmg/ddz187. Available from: https://academic.oup.com/hmg/article/28/R2/R133/5540980 [Accessed on: 2020 Jul 10]

50. Mills MC and Rahal C. The GWAS Diversity Monitor tracks diversity by disease in real time. en. Nature Genetics 2020 Mar; 52. Number: 3 Publisher: Nature Publishing Group:242–3. DOI: 10.1038/s41588-020-0580-y. Available from: https://www.nature.com/articles/s41588-020-0580-y [Accessed on: 2020 Dec 12]

51. Klein RJ, Zeiss C, Chew EY, Tsai JY, Sackler RS, Haynes C, Henning AK, SanGiovanni JP, Mane SM, Mayne ST, Bracken MB, Ferris FL, Ott J, Barnstable C, and Hoh J. Complement factor H polymorphism in age-related macular degeneration. eng. Science (New York, N.Y.) 2005 Apr; 308:385–9. DOI: 10.1126/science.1109557

52. Silventoinen K, Sammalisto S, Perola M, Boomsma DI, Cornes BK, Davis C, Dunkel L, De Lange M, Harris JR, Hjelmborg JV, et al. Heritability of adult body height: a comparative study of twin cohorts in eight countries. Twin research 2003; 6:399–408. Available from: http://journals.cambridge.org/abstract_S1369052300004001 [Accessed on: 2017 Feb 27]

53. Visscher PM, Medland SE, Ferreira MA, Morley KI, Zhu G, Cornes BK, Montgomery GW, and Martin NG. Assumption-free estimation of heritability from genome-wide identity-by-descent sharing between full siblings. PLoS Genet 2006; 2:e41. Available from: http://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.0020041 [Accessed on: 2017 Feb 3]

54. Sanna S, Jackson AU, Nagaraja R, Willer CJ, Chen WM, Bonnycastle LL, Shen H, Timpson N, Lettre G, Usala G, Chines PS, Stringham HM, Scott LJ, Dei M, Lai S, Albai G, Crisponi L, Naitza S, Doheny KF, Pugh EW, Ben-Shlomo Y, Ebrahim S, Lawlor DA, Bergman RN, Watanabe RM, Uda M, Tuomilehto J, Coresh J, Hirschhorn JN, Shuldiner AR, Schlessinger D, Collins FS, Smith GD, Boerwinkle E, Cao A, Boehnke M, Abecasis GR, and Mohlke KL. Common variants in the *GDF5-UQCC* region are associated with variation in human height. En. Nature Genetics 2008 Feb; 40:198. DOI: 10.1038/ng.74. Available from: https://www.nature.com/articles/ng.74 [Accessed on: 2017 Dec 21]

55. Gudbjartsson DF, Walters GB, Thorleifsson G, Stefansson H, Halldorsson BV, Zusmanovich P, Sulem P, Thorlacius S, Gylfason A, Steinberg S, Helgadottir A, Ingason A, Steinthorsdottir V, Olafsdottir EJ, Olafsdottir GH, Jonsson T, Borch-Johnsen K, Hansen T, Andersen G, Jorgensen T, Pedersen O, Aben KK, Witjes JA, Swinkels DW, Heijer Md, Franke B, Verbeek ALM, Becker DM, Yanek LR, Becker LC, Tryggvadottir L, Rafnar T, Gulcher J, Kiemeney LA, Kong A, Thorsteinsdottir U, and Stefansson K. Many sequence variants affecting diversity of adult human height. En. Nature Genetics 2008 May; 40:609. DOI: 10.1038/ng.122. Available from: https://www.nature.com/articles/ng.122 [Accessed on: 2017 Dec 21]

56. Weedon MN, Lango H, Lindgren CM, Wallace C, Evans DM, Mangino M, Freathy RM, Perry JRB, Stevens S, Hall AS, Samani NJ, Shields B, Prokopenko I, Farrall M, Dominiczak A, Johnson T, Bergmann S, Beckmann JS, Vollenweider P, Waterworth DM, Mooser V, Palmer CNA, Morris AD, Ouwehand WH, Caulfield M, Munroe PB, Hattersley AT, McCarthy MI, and Frayling TM. Genome-wide association analysis identifies 20 loci that influence adult height. en. Nature Genetics 2008 May; 40. Number: 5 Publisher: Nature Publishing Group:575–83. DOI: 10.1038/ng.121. Available from: https://www.nature.com/articles/ng.121 [Accessed on: 2021 Jan 7]

57. Visscher PM. Sizing up human height variation. Nature genetics 2008; 40:489–90

58. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A, Cho JH, Guttmacher AE, Kong A, Kruglyak L, Mardis E, Rotimi CN, Slatkin M, Valle D, Whittemore AS, Boehnke M, Clark AG, Eichler EE, Gibson G, Haines JL, Mackay TFC, McCarroll SA, and Visscher PM. Finding the

missing heritability of complex diseases. en. Nature 2009 Oct; 461:747–53. DOI: 10.1038/nature08494. Available from: http://www.nature.com/nature/journal/v461/n7265/abs/nature08494.html [Accessed on: 2016 Jun 28]

59. Lango Allen H et al. Hundreds of variants clustered in genomic loci and biological pathways affect human height. Nature 2010 Oct; 467:832–8. DOI: 10.1038/nature09410. Available from: http://www.nature.com/doifinder/10.1038/nature09410 [Accessed on: 2017 Dec 21]

60. Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR, Madden PA, Heath AC, Martin NG, Montgomery GW, Goddard ME, and Visscher PM. Common SNPs explain a large proportion of the heritability for human height. Nature Genetics 2010 Jul; 42:565–9. DOI: 10.1038/ng.608. Available from: http://www.nature.com/doifinder/10.1038/ng.608 [Accessed on: 2016 Sep 26]

61. Wood AR et al. Defining the role of common variation in the genomic and biological architecture of adult human height. Nature Genetics 2014 Oct; 46:1173–86. DOI: 10.1038/ng.3097. Available from: http://www.nature.com/doifinder/10.1038/ng.3097 [Accessed on: 2016 Jun 14]

62. Yang J, Bakshi A, Zhu Z, Hemani G, Vinkhuyzen AAE, Lee SH, Robinson MR, Perry JRB, Nolte IM, Vliet-Ostaptchouk JV van, Snieder H, Esko T, Milani L, Mägi R, Metspalu A, Hamsten A, Magnusson PKE, Pedersen NL, Ingelsson E, Soranzo N, Keller MC, Wray NR, Goddard ME, and Visscher PM. Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index. Nature Genetics 2015 Aug; 47:1114–20. DOI: 10.1038/ng.3390. Available from: http://www.nature.com/doifinder/10.1038/ng.3390 [Accessed on: 2017 Nov 26]

63. Marouli E et al. Rare and low-frequency coding variants alter human adult height. Nature 2017 Feb. DOI: 10.1038/nature21039. Available from: http://www.nature.com/doifinder/10.1038/nature21039 [Accessed on: 2017 Feb 3]

64. Dauber A, Yu Y, Turchin MC, Chiang CW, Meng YA, Demerath EW, Patel SR, Rich SS, Rotter JI, Schreiner PJ, Wilson JG, Shen Y, Wu BL, and Hirschhorn JN. Genome-wide Association of Copy-Number Variation Reveals an Association between Short Stature and the Presence of Low-Frequency Genomic Deletions. The American Journal of Human Genetics 2011 Dec; 89:751–9. DOI: 10.1016/j.ajhg.2011.10.014. Available from: http://www.sciencedirect.com/science/article/pii/S0002929711004484 [Accessed on: 2017 Feb 23]

65. Macé A et al. CNV-association meta-analysis in 191,161 European adults reveals new loci associated with anthropometric traits. en. Nature Communications 2017 Dec; 8. DOI: 10.1038/s41467-017-00556-x. Available from: http://www.nature.com/articles/s41467-017-00556-x [Accessed on: 2017 Nov 22]

66. Visscher PM, Hill WG, and Wray NR. Heritability in the genomics era — concepts and misconceptions. Nature Reviews Genetics 2008 Apr; 9:255–66. DOI: 10.1038/nrg2322. Available from: http://www.nature.com/doifinder/10.1038/nrg2322 [Accessed on: 2017 Nov 27]

67. Evans L, Tahmasbi R, Vrieze S, Abecasis G, Das S, Bjelland D, deCandia T, Goddard M, Nealie B, Yang J, Visscher P, and Keller M. Comparison of methods that use whole genome data to estimate the heritability and genetic architecture of complex traits. bioRxiv 2017. DOI: 10.1101/115527. Available from: https://www.biorxiv.org/content/early/2017/03/09/115527

68. Speed D, Hemani G, Johnson MR, and Balding DJ. Improved Heritability Estimation from Genome-wide SNPs. American Journal of Human Genetics 2012 Dec; 91:1011–21. DOI: 10.1016/j.ajhg.2012.10.010. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3516604/ [Accessed on: 2017 Nov 26]

69. Speed D, Cai N, Johnson MR, Nejentsev S, and Balding DJ. Reevaluation of SNP heritability in complex human traits. Nature Genetics 2017 May; 49:986–92. DOI: 10.1038/ng.3865. Available from: http://www.nature.com/doifinder/10.1038/ng.3865 [Accessed on: 2017 Nov 23]

70. Finucane HK, Bulik-Sullivan B, Gusev A, Trynka G, Reshef Y, Loh PR, Anttila V, Xu H, Zang C, Farh K, Ripke S, Day FR, Purcell S, Stahl E, Lindstrom S, Perry JRB, Okada Y, Raychaudhuri S, Daly MJ, Patterson N, Neale BM, and Price AL. Partitioning heritability by functional annotation using genome-wide association summary statistics. Nature Genetics 2015 Sep; 47:1228–35. DOI: 10.1038/ng.3404. Available from: http://www.nature.com/doifinder/10.1038/ng.3404 [Accessed on: 2016 Nov 2]

71. Clst (dgt) The FANTOM Consortium tRP et al. A promoter-level mammalian expression atlas. Nature 2014 Mar; 507:462. DOI: 10.1038/nature13182. Available from: https://www.nature.com/articles/nature13182

72. Consortium TEP. The ENCODE (ENCyclopedia Of DNA Elements) Project. en. Science 2004 Oct; 306:636–40. DOI: 10.1126/science.1105136. Available from: http://science.sciencemag.org/content/306/5696/636 [Accessed on: 2017 Dec 27]

73. Gazal S, Finucane HK, Furlotte NA, Loh PR, Palamara PF, Liu X, Schoech A, Bulik-Sullivan B, Neale BM, Gusev A, and Price AL. Linkage disequilibrium–dependent architecture of human complex traits shows action of negative selection. Nature Genetics 2017 Sep; 49:1421–7. DOI: 10.1038/ng.3954. Available from: http://www.nature.com/doifinder/10.1038/ng.3954 [Accessed on: 2017 Nov 23]

74. Yang J, Manolio TA, Pasquale LR, Boerwinkle E, Caporaso N, Cunningham JM, Andrade M de, Feenstra B, Feingold E, Hayes MG, Hill WG, Landi MT, Alonso A, Lettre G, Lin P, Ling H, Lowe W, Mathias RA, Melbye M, Pugh E, Cornelis MC, Weir BS, Goddard ME, and Visscher PM. Genome partitioning of genetic variation for complex traits using common SNPs. Nature Genetics 2011 Jun; 43:519–25. DOI: 10.1038/ng.823. Available from: http://www.nature.com/doifinder/10.1038/ng.823 [Accessed on: 2017 Nov 23]

75. Loh PR, Bhatia G, Gusev A, Finucane HK, Bulik-Sullivan BK, Pollack SJ, Schizophrenia Working Group of the Psychiatric Genomics Consortium, Candia TR de, Lee SH, Wray NR, Kendler KS, O'Donovan MC, Neale BM, Patterson N, and Price AL. Contrasting genetic architectures of schizophrenia and other complex diseases using fast variance-components analysis. en. Nature Genetics 2015 Dec; 47:1385–92. DOI: 10.1038/ng.3431. Available from: https://www.nature.com/articles/ng.3431 [Accessed on: 2018 Nov 17]

76. Pickrell JK. Joint analysis of functional genomic data and genome-wide association studies of 18 human traits. eng. American Journal of Human Genetics 2014 Apr; 94:559–73. DOI: 10.1016/j.ajhg.2014.03.004

77. Welter D, MacArthur J, Morales J, Burdett T, Hall P, Junkins H, Klemm A, Flicek P, Manolio T, Hindorff L, and Parkinson H. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. eng. Nucleic Acids Research 2014 Jan; 42:D1001–1006. DOI: 10.1093/nar/gkt1229

78. Pickrell JK, Berisa T, Liu JZ, Ségurel L, Tung JY, and Hinds DA. Detection and interpretation of shared genetic influences on 42 human traits. en. Nature Genetics 2016 Jul; 48:709–17. DOI: 10.1038/ng.3570. Available from: https://www.nature.com/articles/ng.3570 [Accessed on: 2020 Feb 20]

79. Visscher PM and Yang J. A plethora of pleiotropy across complex traits. en. Nature Genetics 2016 Jun; 48:ng.3604. DOI: 10.1038/ng.3604. Available from: https://www.nature.com/articles/ng.3604 [Accessed on: 2017 Nov 16]

80. Wray NR, Wijmenga C, Sullivan PF, Yang J, and Visscher PM. Common Disease Is More Complex Than Implied by the Core Gene Omnigenic Model. en. Cell 2018 Jun; 173:1573–80. DOI: 10.1016/j.cell.2018.05.051. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0092867418307141 [Accessed on: 2018 Jun 15]

81. Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. eng. Nature 2007 Jun; 447:661–78. DOI: 10.1038/nature05911

82. Claussnitzer M, Cho JH, Collins R, Cox NJ, Dermitzakis ET, Hurles ME, Kathiresan S, Kenny EE, Lindgren CM, MacArthur DG, North KN, Plon SE, Rehm HL, Risch N, Rotimi CN, Shendure J, Soranzo N, and McCarthy MI. A brief history of human disease genetics. en. Nature 2020 Jan; 577. Number: 7789 Publisher: Nature Publishing Group:179–89. DOI: 10.1038/s41586-019-1879-7. Available from: https://www.nature.com/articles/s41586-019-1879-7 [Accessed on: 2021 Jan 7]

83. Maher B. Personal genomes: The case of the missing heritability. eng. Nature 2008 Nov; 456:18–21. DOI: 10.1038/456018a

84. International Schizophrenia Consortium, Purcell SM, Wray NR, Stone JL, Visscher PM, O'Donovan MC, Sullivan PF, and Sklar P. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. eng. Nature 2009 Aug; 460:748–52. DOI: 10.1038/nature08185

85. Zhang Y, Qi G, Park JH, and Chatterjee N. Estimation of complex effect-size distributions using summary-level statistics from genome-wide association studies across 32 complex traits. en. Nature Genetics 2018 Aug :1. DOI: 10.1038/s41588-018-0193-x. Available from: https://www.nature.com/articles/s41588-018-0193-x [Accessed on: 2018 Aug 13]

86. Frei O, Holland D, Smeland OB, Shadrin AA, Fan CC, Maeland S, O'Connell KS, Wang Y, Djurovic S, Thompson WK, Andreassen OA, and Dale AM. Bivariate causal mixture model quantifies polygenic overlap between complex traits beyond genetic correlation. en. Nature Communications 2019 Jun; 10. Number: 1 Publisher: Nature Publishing Group:2417. DOI: 10.1038/s41467-019-10310-0. Available from: https://www.nature.com/articles/s41467-019-10310-0 [Accessed on: 2021 Jan 8]

87. O'Connor LJ, Schoech AP, Hormozdiari F, Gazal S, Patterson N, and Price AL. Extreme Polygenicity of Complex Traits Is Explained by Negative Selection. The American Journal of Human Genetics 2019 Aug. DOI: 10.1016/j.ajhg.2019.07.003. Available from: http://www.sciencedirect.com/science/article/pii/S0002929719302666 [Accessed on: 2019 Aug 22]

88. Sinnott-Armstrong N, Naqvi S, Rivas M, and Pritchard JK. GWAS of three molecular traits highlights core genes and pathways alongside a highly polygenic background. en. bioRxiv 2020 Apr. Publisher: Cold Spring Harbor Laboratory Section: New Results:2020.04.20.051631. DOI: 10.1101/2020.04.20.051631. Available from: https://www.biorxiv.org/content/10.1101/2020.04.20.051631v1 [Accessed on: 2021 Jan 8]

89. Trynka G, Sandor C, Han B, Xu H, Stranger BE, Liu XS, and Raychaudhuri S. Chromatin marks identify critical cell types for fine mapping complex trait variants. Nature Genetics 2012 Dec; 45:124–30. DOI: 10.1038/ng.2504. Available from: http://www.nature.com/doifinder/10.1038/ng.2504 [Accessed on: 2017 Nov 23]

90. Shi H, Kichaev G, and Pasaniuc B. Contrasting the Genetic Architecture of 30 Complex Traits from Summary Association Data. en. The American Journal of Human Genetics 2016 Jul; 99:139–53. DOI: 10.1016/j.ajhg.2016.05.013. Available from: http://linkinghub.elsevier.com/retrieve/pii/S0002929716301483 [Accessed on: 2016 Nov 16]

91. Jostins L et al. Host–microbe interactions have shaped the genetic architecture of inflammatory bowel disease. en. Nature 2012 Nov; 491:119–24. DOI: 10.1038/nature11582. Available from: https://www.nature.com/articles/nature11582 [Accessed on: 2020 Feb 9]

92. Zhu X and Stephens M. Large-scale genome-wide enrichment analyses identify new trait-associated genes and pathways across 31 human phenotypes. en. Nature Communications 2018 Oct; 9:1–14. DOI: 10.1038/s41467-018-06805-x. Available from: https://www.nature.com/articles/s41467-018-06805-x [Accessed on: 2020 Feb 9]

93. Fernández-Tajes J, Gaulton KJ, Bunt M van de, Torres J, Thurner M, Mahajan A, Gloyn AL, Lage K, and McCarthy MI. Developing a network view of type 2 diabetes risk pathways through integration of genetic, genomic and functional data. Genome Medicine 2019 Mar; 11:19. DOI: 10.1186/s13073-019-0628-8. Available from: https://doi.org/10.1186/s13073-019-0628-8 [Accessed on: 2020 Feb 9]

94. Liu X, Li YI, and Pritchard JK. Trans Effects on Gene Expression Can Drive Omnigenic Inheritance. English. Cell 2019 May; 177:1022–1034.e6. DOI: 10.1016/j.cell.2019.04.014. Available from: https://www.cell.com/cell/abstract/S0092-8674(19)30400-3 [Accessed on: 2019 May 18]

95. Võsa U et al. Unraveling the polygenic architecture of complex traits using blood eQTL meta-analysis. en. preprint. Genomics, 2018 Oct. DOI: 10.1101/447367. Available from: http://biorxiv.org/lookup/doi/10.1101/447367 [Accessed on: 2019 Jul 8]

96. Udler MS. Type 2 Diabetes: Multiple Genes, Multiple Diseases. en. Current Diabetes Reports 2019 Jul; 19:55. DOI: 10.1007/s11892-019-1169-7. Available from: https://doi.org/10.1007/s11892-019-1169-7 [Accessed on: 2021 Jan 8]

97. Turkheimer E. Three Laws of Behavior Genetics and What They Mean. en. Current Directions in Psychological Science 2000 Oct; 9. Publisher: SAGE Publications Inc:160–4. DOI: 10.1111/1467-8721.00084. Available from: https://doi.org/10.1111/1467-8721.00084 [Accessed on: 2021 Jan 8]

98. Gottesman II and Gould TD. The Endophenotype Concept in Psychiatry: Etymology and Strategic Intentions. American Journal of Psychiatry 2003 Apr; 160. Publisher: American Psychiatric Publishing:636–45. DOI: 10.1176/appi.ajp.160.4.636. Available from: https://ajp.psychiatryonline.org/doi/full/10.1176/appi.ajp.160.4.636 [Accessed on: 2021 Jan 8]

99. Bittante G, Penasa M, and Cecchinato A. Invited review: Genetics and modeling of milk coagulation properties. en. Journal of Dairy Science 2012 Dec; 95:6843–70. DOI: 10.3168/jds.2012-5507. Available from: http://www.sciencedirect.com/science/article/pii/S0022030212007072 [Accessed on: 2021 Jan 8]

100. GTEx Consortium, Gamazon ER, Segrè AV, Bunt M van de, Wen X, Xi HS, Hormozdiari F, Ongen H, Konkashbaev A, Derks EM, Aguet F, Quan J, Nicolae DL, Eskin E, Kellis M, Getz G, McCarthy MI, Dermitzakis ET, Cox NJ, and Ardlie KG. Using an atlas of gene regulation across 44 human tissues to inform complex disease- and trait-associated variation. en. Nature Genetics 2018 Jun. DOI: 10.1038/s41588-018-0154-4. Available from: http://www.nature.com/articles/s41588-018-0154-4 [Accessed on: 2018 Jul 2]

101. Bremer H, Baracchini E, Little R, and Ryals J. Control of RNA synthesis in bacteria. Genetics of translation: new approaches. NATO ASI Series, Series H, Cell biology 1987; 14:63–74

102. Laffend L and Shuler ML. Ribosomal protein limitations in Escherichia coli under conditions of high translational activity. en. Biotechnology and Bioengineering 1994; 43:388–98. DOI: 10.1002/bit.260430507. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1002/bit.260430507 [Accessed on: 2019 Sep 5]

103. Lane N and Martin W. The energetics of genome complexity. en. Nature 2010 Oct; 467:929–34. DOI: 10.1038/nature09486. Available from: https://www.nature.com/articles/nature09486 [Accessed on: 2019 Sep 12]

104. Chu D, Barnes DJ, and Haar T von der. The role of tRNA and ribosome competition in coupling the expression of different mRNAs in Saccharomyces cerevisiae. Nucleic Acids Research 2011 Aug; 39:6705–14. DOI: 10.1093/nar/gkr300. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3159466/ [Accessed on: 2019 Sep 10]

105. Haar T von der. MATHEMATICAL AND COMPUTATIONAL MODELLING OF RIBOSOMAL MOVEMENT AND PROTEIN SYNTHESIS: AN OVERVIEW. Computational and Structural Biotechnology Journal 2012 Apr; 1:e201204002. DOI: 10.5936/csbj.201204002. Available from: http://www.sciencedirect.com/science/article/pii/S2001037014601045 [Accessed on: 2019 Sep 4]

106. Zur H and Tuller T. Predictive biophysical modeling and understanding of the dynamics of mRNA translation and its evolution. en. Nucleic Acids Research 2016 Sep :gkw764. DOI: 10.1093/nar/gkw764. Available from: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkw764 [Accessed on: 2019 Sep 11]

107. Elandt-Johnson RC and Johnson NL. Survival Models and Data Analysis. en. *Survival Models and Data Analysis*. John Wiley & Sons, 1980 Sep :69

108. Kendall MG, Stuart A, and Ord JK. Kendall's Advanced Theory of Statistics. en. *The advanced theory of statistics in 3 volumes. 1 1*. OCLC: 1071028235. London: Griffin, 1994 :351

109. x. Exponential Growth of the AncestryDNA Database. Ed. by x. 2018 Oct. Available from: %7Bhttps://wiki.uiowa.edu/display/2360159/2017/09/15/Exponential+Growth+of+the+AncestryDNA+Database%7D

110. Ball MP, Bobe JR, Chou MF, Clegg T, Estep PW, Lunshof JE, Vandewege W, Zaranek AW, and Church GM. Harvard Personal Genome Project: lessons from participatory public research. Genome Medicine 2014 Feb; 6:10. DOI: 10.1186/gm527. Available from: https://doi.org/10.1186/gm527 [Accessed on: 2018 Sep 27]

111. Yuan J, Gordon A, Speyer D, Aufrichtig R, Zielinski D, Pickrell J, and Erlich Y. DNA.Land is a framework to collect genomes and phenomes in the era of abundant genetic information. en. Nature Genetics 2018 Feb; 50:160–5. DOI: 10.1038/s41588-017-0021-8. Available from: https://www.nature.com/articles/s41588-017-0021-8 [Accessed on: 2018 Sep 27]

112. Greshake B, Bayer PE, Rausch H, and Reda J. OpenSNP–a crowdsourced web resource for personal genomics. PLoS One 2014; 9:e89204. Available from: http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0089204 [Accessed on: 2016 May 9]

113. Lander ES. Initial impact of the sequencing of the human genome. Nature 2011 Feb; 470:187–97. DOI: 10.1038/nature09792. Available from: http://www.nature.com/doifinder/10.1038/nature09792 [Accessed on: 2017 Dec 21]

114. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, and Haussler aD. The Human Genome Browser at UCSC. en. Genome Research 2002 Jun; 12:996–1006. DOI: 10.1101/gr.229102. Available from: http://genome.cshlp.org/content/12/6/996 [Accessed on: 2018 Sep 27]

115. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, and Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. GigaScience 2015 Feb; 4. DOI: 10.1186/s13742-015-0047-8. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4342193/ [Accessed on: 2018 Sep 27]

116. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, Bakker PIW de, Daly MJ, and Sham PC. PLINK: a tool set for whole-genome association and population-based linkage analyses. eng. American Journal of Human Genetics 2007 Sep; 81:559–75. DOI: 10.1086/519795

117. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, McVean G, Durbin R, and 1000 Genomes Project Analysis Group. The variant call format and VCFtools. en. Bioinformatics 2011 Aug; 27:2156–8. DOI: 10.1093/bioinformatics/btr330. Available from: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btr330 [Accessed on: 2018 Sep 27]

118. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, and 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map format and SAMtools. eng. Bioinformatics (Oxford, England) 2009 Aug; 25:2078–9. DOI: 10.1093/bioinformatics/btp352

119. Pirvu B and Wilms J. How to get the exact y-values of all data points used for the computation of the public leaderboard score in the "Mercedes-Benz Greener Manufacturing" competition on Kaggle. en. https://crowdstats.eu/ 2017 :7

120. Price AL, Zaitlen NA, Reich D, and Patterson N. New approaches to population stratification in genome-wide association studies. Nature Reviews Genetics 2010; 11:459–63

121. Bouaziz M, Ambroise C, and Guedj M. Accounting for population stratification in practice: a comparison of the main strategies dedicated to genome-wide association studies. eng. PloS One 2011; 6:e28845. DOI: 10.1371/journal.pone.0028845

122. Kittles RA, Chen W, Panguluri RK, Ahaghotu C, Jackson A, Adebamowo CA, Griffin R, Williams T, Ukoli F, Adams-Campbell L, Kwagyan J, Isaacs W, Freeman V, and Dunston GM. CYP3A4-V and prostate cancer in African Americans: causal or confounding association because of population stratification? en. Human Genetics 2002 Jun; 110:553–60. DOI: 10.1007/s00439-002-0731-5. Available from: https://doi.org/10.1007/s00439-002-0731-5 [Accessed on: 2018 Sep 27]

123. Rosenberg NA, Huang L, Jewett EM, Szpiech ZA, Jankovic I, and Boehnke M. Genome-wide association studies in diverse populations. Nature reviews. Genetics 2010 May; 11:356–66. DOI: 10.1038/nrg2760. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3079573/ [Accessed on: 2018 Sep 27]

124. Zaitlen N, Paşaniuc B, Gur T, Ziv E, and Halperin E. Leveraging Genetic Variability across Populations for the Identification of Causal Variants. American Journal of Human Genetics 2010 Jan; 86:23–33. DOI: 10.1016/j.ajhg.2009.11.016. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2801753/ [Accessed on: 2018 Sep 27]

125. Euesden J, Lewis CM, and O'Reilly PF. PRSice: Polygenic Risk Score software. eng. Bioinformatics (Oxford, England) 2015 May; 31:1466–8. DOI: 10.1093/bioinformatics/btu848

126. Torkamani A, Wineinger NE, and Topol EJ. The personal and clinical utility of polygenic risk scores. en. Nature Reviews Genetics 2018 May :1. DOI: 10.1038/s41576-018-0018-x. Available from: https://www.nature.com/articles/s41576-018-0018-x [Accessed on: 2018 Jul 4]

127. Bogdan R, Baranger DA, and Agrawal A. Polygenic Risk Scores in Clinical Psychology: Bridging Genomic Risk to Individual Differences. Annual Review of Clinical Psychology 2018; 14:119–57. DOI: 10.1146/annurev-clinpsy-050817-084847. Available from: https://doi.org/10.1146/annurev-clinpsy-050817-084847 [Accessed on: 2018 Sep 27]

128. Vilhjálmsson BJ, Yang J, Finucane HK, Gusev A, Lindström S, Ripke S, Genovese G, Loh PR, Bhatia G, Do R, Hayeck T, Won HH, Kathiresan S, Pato M, Pato C, Tamimi R, Stahl E, Zaitlen N, Pasaniuc B, Belbin G, Kenny EE, Schierup MH, De Jager P, Patsopoulos NA, McCarroll S, Daly M, Purcell S, Chasman D, Neale B, Goddard M, Visscher PM, Kraft P, Patterson N, and Price AL. Modeling Linkage Disequilibrium Increases Accuracy of Polygenic Risk Scores. en. The American Journal of Human Genetics 2015 Oct; 97:576–92. DOI: 10.1016/j.ajhg.2015.09.001. Available from: http://linkinghub.elsevier.com/retrieve/pii/S0002929715003651 [Accessed on: 2016 Nov 15]

129. Márquez-Luna C, Loh PR, and Price AL. Multi-ethnic polygenic risk scores improve risk prediction in diverse populations. Genetic epidemiology 2017 Dec; 41:811–23. DOI: 10.1002/gepi.22083. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5726434/ [Accessed on: 2020 Jul 16]

130. Yengo L, Sidorenko J, Kemper KE, Zheng Z, Wood AR, Weedon MN, Frayling TM, Hirschhorn J, Yang J, and Visscher PM. Meta-analysis of genome-wide association studies for height and body mass index in ~700000 individuals of European ancestry. en. Human Molecular Genetics 2018 Oct; 27:3641–9. DOI: 10.1093/hmg/ddy271. Available from: https://academic.oup.com/hmg/article/27/20/3641/5067845 [Accessed on: 2019 Feb 22]

131. Curtis D. Polygenic risk score for schizophrenia is more strongly associated with ancestry than with schizophrenia. en-US. Psychiatric Genetics 2018 Oct; 28:85–9. DOI: 10.1097/YPG.0000000000000206. Available from: https://journals.lww.com/psychgenetics/Fulltext/2018/10000/Polygenic_risk_score_for_schizophrenia_is_more.2.aspx [Accessed on: 2020 Jul 13]

132. Reisberg S, Iljasenko T, Läll K, Fischer K, and Vilo J. Comparing distributions of polygenic risk scores of type 2 diabetes and coronary heart disease within different populations. en. PLOS ONE 2017 Jul; 12:e0179238. DOI: 10.1371/journal.pone.0179238. Available from: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0179238 [Accessed on: 2018 Oct 29]

133. Martin AR, Gignoux CR, Walters RK, Wojcik GL, Neale BM, Gravel S, Daly MJ, Bustamante CD, and Kenny EE. Human Demographic History Impacts Genetic Risk Prediction across Diverse Populations. en. The American Journal of Human Genetics 2017 Apr; 100:635–49. DOI: 10.1016/j.ajhg.2017.03.004. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0002929717301076 [Accessed on: 2018 Oct 2]

134. Sieberts SK et al. Crowdsourced assessment of common genetic contribution to predicting anti-TNF treatment response in rheumatoid arthritis. en. Nature Communications 2016 Aug; 7:12460. DOI: 10.1038/ncomms12460. Available from: https://www.nature.com/articles/ncomms12460 [Accessed on: 2019 Mar 5]

135. Olivier Naret. OpenSNP Cohort Maker. 2018 Oct. Available from: https://github.com/onaret/opensnp-cohort-maker [Accessed on: 2019 Mar 1]

136. Olivier Naret. CrowdAI / OpenSNP - height prediction challenge - leaderboard/overview. 2020 Jan. DOI: 10.5281/zenodo.3604246. Available from: https://zenodo.org/record/3604246#.XhiZS8ZKhhE [Accessed on: 2020 Jan 10]

137. Bateson W. Mendel's principles of heredity. 1909. Mendel's principles of heredity 1909

138. Moore JH and Williams SM. Traversing the conceptual divide between biological and statistical epistasis: systems biology and a more modern synthesis. en. BioEssays 2005; 27. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/bies.20236:637–46. DOI: 10.1002/bies.20236. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1002/bies.20236 [Accessed on: 2020 Oct 5]

139. Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, Shen Y, Hu M, Liu JS, and Ren B. Topological domains in mammalian genomes identified by analysis of chromatin interactions. en. Nature 2012 May; 485:376–80. DOI: 10.1038/nature11082. Available from: https://www.nature.com/articles/nature11082 [Accessed on: 2018 May 22]

140. Hivert V, Sidorenko J, Rohart F, Goddard ME, Yang J, Wray NR, Yengo L, and Visscher PM. Estimation of non-additive genetic variance in human complex traits from a large sample of unrelated individuals. en. bioRxiv 2020 Nov. Publisher: Cold Spring Harbor Laboratory Section: New Results:2020.11.09.375501. DOI: 10.1101/2020.11.09.375501. Available from: https://www.biorxiv.org/content/10.1101/2020.11.09.375501v1 [Accessed on: 2020 Dec 30]

141. López-Cortegano E and Caballero A. Inferring the Nature of Missing Heritability in Human Traits Using Data from the GWAS Catalog. en. Genetics 2019 Jul; 212. Publisher: Genetics Section: Investigations:891–904. DOI: 10.1534/genetics.119.302077. Available from: https://www.genetics.org/content/212/3/891 [Accessed on: 2020 Dec 30]

142. Leung MKK, Xiong HY, Lee LJ, and Frey BJ. Deep learning of the tissue-regulated splicing code. en. Bioinformatics 2014 Jun; 30:i121–i129. DOI: 10.1093/bioinformatics/btu277. Available from: http://bioinformatics.oxfordjournals.org/content/30/12/i121 [Accessed on: 2016 Jun 17]

143. Alipanahi B, Delong A, Weirauch MT, and Frey BJ. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. Nature Biotechnology 2015 Jul; 33:831–8. DOI: 10.1038/nbt.3300. Available from: http://www.nature.com/doifinder/10.1038/nbt.3300 [Accessed on: 2016 Jun 16]

144. Zhou J and Troyanskaya OG. Predicting effects of noncoding variants with deep learning–based sequence model. Nature Methods 2015 Aug; 12:931–4. DOI: 10.1038/nmeth.3547. Available from: http://www.nature.com/doifinder/10.1038/nmeth.3547 [Accessed on: 2016 Jun 16]

145. Quang D, Chen Y, and Xie X. DANN: a deep learning approach for annotating the pathogenicity of genetic variants. Bioinformatics 2014 :btu703. Available from: http://bioinformatics.oxfordjournals.org/content/early/2014/10/22/bioinformatics.btu703.short [Accessed on: 2016 May 9]

146. Günther F, Wawro N, and Bammann K. Neural networks for modeling gene-gene interactions in association studies. en. BMC Genetics 2009; 10:87. DOI: 10.1186/1471-2156-10-87. Available from: http://bmcgenet.biomedcentral.com/articles/10.1186/1471-2156-10-87 [Accessed on: 2017 Feb 6]

147. Bellot P, Campos Gdl, and Pérez-Enciso M. Can Deep Learning Improve Genomic Prediction of Complex Human Traits? en. Genetics 2018 Nov; 210:809–19. DOI: 10.1534/genetics.118.301298. Available from: http://www.genetics.org/content/210/3/809 [Accessed on: 2019 Feb 15]

148. Xu Y, Vuckovic D, Ritchie SC, Akbari P, Jiang T, Grealey J, Butterworth AS, Ouwehand WH, Roberts DJ, Di Angelantonio E, Danesh J, Soranzo N, and Inouye M. Learning polygenic scores for human blood cell traits. en. preprint. Genetics, 2020 Feb. DOI: 10.1101/2020.02.17.952788. Available from: http://biorxiv.org/lookup/doi/10.1101/2020.02.17.952788 [Accessed on: 2020 Apr 13]

149. Abdollahi-Arpanahi R, Gianola D, and Peñagaricano F. Deep learning versus parametric and ensemble methods for genomic prediction of complex phenotypes. eng. Genetics, selection, evolution: GSE 2020 Feb; 52:12. DOI: 10.1186/s12711-020-00531-z

150. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. eng. Nature 2012 Sep; 489:57–74. DOI: 10.1038/nature11247

151. Southern California U of. Quanto. 2009 May. Available from: http://biostats.usc.edu/Quanto.html [Accessed on: 2018 Jan 2]

152. Auton A et al. A global reference for human genetic variation. en. Nature 2015 Sep; 526:68–74. DOI: 10.1038/nature15393. Available from: http://www.nature.com/doifinder/10.1038/nature15393 [Accessed on: 2018 Apr 15]

153. Campos Gdl, Sorensen DA, and Toro MA. Imperfect Linkage Disequilibrium Generates Phantom Epistasis (& Perils of Big Data). en. G3: Genes, Genomes, Genetics 2019 Mar :g3.400101.2019. DOI: 10.1534/g3.119.400101. Available from: http://www.g3journal.org/content/early/2019/03/15/g3.119.400101 [Accessed on: 2019 Apr 2]

154. Klambauer G, Unterthiner T, Mayr A, and Hochreiter S. Self-Normalizing Neural Networks. arXiv:1706.02515 [cs, stat] 2017 Sep. arXiv: 1706.02515. Available from: http://arxiv.org/abs/1706.02515 [Accessed on: 2020 Dec 29]

155. Läll K, Mägi R, Morris A, Metspalu A, and Fischer K. Personalized risk prediction for type 2 diabetes: the potential of genetic risk scores. en. Genetics in Medicine 2017 Mar; 19:322–9. DOI: 10.1038/gim.2016.103. Available from: https://www.nature.com/articles/gim2016103 [Accessed on: 2018 Dec 20]

156. Gatz M, Reynolds CA, Fratiglioni L, Johansson B, Mortimer JA, Berg S, Fiske A, and Pedersen NL. Role of Genes and Environments for Explaining Alzheimer Disease. en. Archives of General Psychiatry 2006 Feb; 63:168–74. DOI: 10.1001/archpsyc.63.2.168. Available from: https://jamanetwork.com/journals/jamapsychiatry/fullarticle/209307 [Accessed on: 2019 Feb 24]

157. Zheutlin AB, Dennis J, Restrepo N, Straub P, Ruderfer D, Castro VM, Chen CY, Kirchner HL, Chabris CF, Davis LK, and Smoller JW. Penetrance and pleiotropy of polygenic risk scores for schizophrenia in 90,000 patients across three healthcare systems. en. bioRxiv 2018 Sep :421164. DOI: 10.1101/421164. Available from: https://www.biorxiv.org/content/early/2018/09/18/421164 [Accessed on: 2019 Jan 9]

158. Howard DM, Adams MJ, Clarke TK, Hafferty JD, Gibson J, Shirali M, Coleman JRI, Hagenaars SP, Ward J, Wigmore EM, Alloza C, Shen X, Barbu MC, Xu EY, Whalley HC, Marioni RE, Porteous DJ, Davies G, Deary IJ, Hemani G, Berger K, Teismann H, Rawal R, Arolt V, Baune BT, Dannlowski U, Domschke K, Tian C, Hinds DA, Trzaskowski M, Byrne EM, Ripke S, Smith DJ, Sullivan PF, Wray NR, Breen G, Lewis CM, and McIntosh AM. Genome-wide meta-analysis of depression identifies 102 independent variants and highlights the importance of the prefrontal brain regions. En. Nature Neuroscience 2019 Mar; 22:343. DOI: 10.1038/s41593-018-0326-7. Available from: https://www.nature.com/articles/s41593-018-0326-7 [Accessed on: 2019 Feb 24]

159. Mavaddat N et al. Polygenic Risk Scores for Prediction of Breast Cancer and Breast Cancer Subtypes. The American Journal of Human Genetics 2018 Dec. DOI: 10.1016/j.ajhg.2018.11.002. Available from: http://www.sciencedirect.com/science/article/pii/S0002929718304051 [Accessed on: 2019 Jan 3]

160. Schumacher FR et al. Association analyses of more than 140,000 men identify 63 new prostate cancer susceptibility loci. eng. Nature Genetics 2018; 50:928–36. DOI: 10.1038/s41588-018-0142-8

161. McPherson Ruth and Tybjaerg-Hansen Anne. Genetics of Coronary Artery Disease. Circulation Research 2016 Feb; 118:564–78. DOI: 10.1161/CIRCRESAHA.115.306566. Available from: https://www.ahajournals.org/doi/full/10.1161/CIRCRESAHA.115.306566 [Accessed on: 2018 Dec 13]

162. Clarke SL and Assimes TL. Genome-Wide Association Studies of Coronary Artery Disease: Recent Progress and Challenges Ahead. en. Current Atherosclerosis Reports 2018 Jul; 20:47. DOI: 10.1007/s11883-018-0748-4. Available from: https://doi.org/10.1007/s11883-018-0748-4 [Accessed on: 2018 Nov 27]

163. Khera AV, Emdin CA, Drake I, Natarajan P, Bick AG, Cook NR, Chasman DI, Baber U, Mehran R, Rader DJ, Fuster V, Boerwinkle E, Melander O, Orho-Melander M, Ridker PM, and Kathiresan S. Genetic Risk, Adherence to a Healthy Lifestyle, and Coronary Disease. eng. The New England Journal of Medicine 2016; 375:2349–58. DOI: 10.1056/NEJMoa1605086

164. Sharp SA, Rich SS, Wood AR, Jones SE, Beaumont RN, Harrison JW, Schneider DA, Locke JM, Tyrrell J, Weedon MN, Hagopian WA, and Oram RA. Development and Standardization of an Improved Type 1 Diabetes Genetic Risk Score for Use in Newborn Screening and Incident Diagnosis. eng. Diabetes Care 2019 Feb; 42:200–7. DOI: 10.2337/dc18-1785

165. Maas P, Barrdahl M, Joshi AD, Auer PL, Gaudet MM, Milne RL, Schumacher FR, Anderson WF, Check D, Chattopadhyay S, Baglietto L, Berg CD, Chanock SJ, Cox DG, Figueroa JD, Gail MH, Graubard BI, Haiman CA, Hankinson SE, Hoover RN, Isaacs C, Kolonel LN, Le Marchand L, Lee IM, Lindström S, Overvad K, Romieu I, Sanchez MJ, Southey MC, Stram DO, Tumino R, VanderWeele TJ, Willett WC, Zhang S, Buring JE, Canzian F, Gapstur SM, Henderson BE, Hunter DJ, Giles GG, Prentice RL, Ziegler RG, Kraft P, Garcia-Closas M, and Chatterjee N. Breast Cancer Risk From Modifiable and Nonmodifiable Risk Factors Among White Women in the United States. eng. JAMA oncology 2016 Oct; 2:1295–302. DOI: 10.1001/jamaoncol.2016.1025

166. Wray NR, Goddard ME, and Visscher PM. Prediction of individual genetic risk to disease from genome-wide association studies. Genome Research 2007 Oct; 17:1520–8. DOI: 10.1101/gr.6665407. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1987352/ [Accessed on: 2020 Jul 13]

167. Bustamante CD, De La Vega FM, and Burchard EG. Genomics for the world. en. Nature 2011 Jul; 475:163–5. DOI: 10.1038/475163a. Available from: https://www.nature.com/articles/475163a [Accessed on: 2018 Dec 11]

168. Wang Y, Guo J, Ni G, Yang J, Visscher PM, and Yengo L. Theoretical and empirical quantification of the accuracy of polygenic scores in ancestry divergent populations. en. preprint. Genetics, 2020 Jan. DOI: 10.1101/2020.01.14.905927. Available from: http://biorxiv.org/lookup/doi/10.1101/2020.01.14.905927 [Accessed on: 2020 Apr 27]

169. Durvasula A and Lohmueller KE. Negative selection on complex traits limits genetic risk prediction accuracy between populations. en. bioRxiv 2019 Aug. Publisher: Cold Spring Harbor Laboratory Section: New Results:721936. DOI: 10.1101/721936. Available from: https://www.biorxiv.org/content/10.1101/721936v1 [Accessed on: 2020 May 11]

170. Brown BC, Ye CJ, Price AL, and Zaitlen N. Transethnic Genetic-Correlation Estimates from Summary Statistics. The American Journal of Human Genetics 2016 Jul; 99:76–88. DOI: 10.1016/j.ajhg.2016.05.001. Available from: http://www.sciencedirect.com/science/article/pii/S0002929716301355 [Accessed on: 2019 Jan 9]

171. Wojcik GL, Graff M, Nishimura KK, Tao R, Haessler J, Gignoux CR, Highland HM, Patel YM, Sorokin EP, Avery CL, Belbin GM, Bien SA, Cheng I, Cullina S, Hodonsky CJ, Hu Y, Huckins LM, Jeff J, Justice AE, Kocarnik JM, Lim U, Lin BM, Lu Y, Nelson SC, Park SSL, Poisner H, Preuss MH, Richard MA, Schurmann C, Setiawan VW, Sockell A, Vahi K, Verbanck M, Vishnu A, Walker RW, Young KL, Zubair N, Acuña-Alonso V, Ambite JL, Barnes KC, Boerwinkle E, Bottinger EP, Bustamante CD, Caberto C, Canizales-Quinteros S, Conomos MP, Deelman E, Do R, Doheny K, Fernández-Rhodes L, Fornage M, Hailu B, Heiss G, Henn BM, Hindorff LA, Jackson RD, Laurie CA, Laurie CC, Li Y, Lin DY, Moreno-Estrada A, Nadkarni G, Norman PJ, Pooler LC, Reiner AP, Romm J, Sabatti C, Sandoval K, Sheng X, Stahl EA, Stram DO, Thornton TA, Wassel CL, Wilkens LR, Winkler CA, Yoneyama S, Buyske S, Haiman CA, Kooperberg C, Le Marchand L, Loos RJF, Matise TC, North KE, Peters U, Kenny EE, and Carlson CS. Genetic analyses of diverse populations improves discovery for complex traits. en. Nature 2019 Jun; 570. Number: 7762 Publisher: Nature Publishing Group:514–8. DOI: 10.1038/s41586-019-1310-4. Available from: https://www.nature.com/articles/s41586-019-1310-4 [Accessed on: 2020 May 11]

172. Mostafavi H, Harpak A, Agarwal I, Conley D, Pritchard JK, and Przeworski M. Variable prediction accuracy of polygenic scores within an ancestry group. eLife 2020 Jan; 9. Ed. by Loos R, Eisen MB, and O'Reilly P. Publisher: eLife Sciences Publications, Ltd:e48376. DOI: 10.7554/eLife.48376. Available from: https://doi.org/10.7554/eLife.48376 [Accessed on: 2020 May 9]

173. Khera AV, Chaffin M, Wade KH, Zahid S, Brancale J, Xia R, Distefano M, Senol-Cosar O, Haas ME, Bick A, Aragam KG, Lander ES, Smith GD, Mason-Suares H, Fornage M, Lebo M, Timpson NJ, Kaplan LM, and Kathiresan S. Polygenic Prediction of Weight and Obesity Trajectories from Birth to Adulthood. English. Cell 2019 Apr; 177:587–596.e9. DOI: 10.1016/j.cell.2019.03.028. Available from: https://www.cell.com/cell/abstract/S0092-8674(19)30290-9 [Accessed on: 2019 Jul 5]

174. Ge T, Chen CY, Neale BM, Sabuncu MR, and Smoller JW. Phenome-wide heritability analysis of the UK Biobank. en. PLOS Genetics 2017; 13. Publisher: Public Library of Science:e1006711. DOI: 10.1371/journal.pgen.1006711. Available from: https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1006711 [Accessed on: 2020 Jul 18]

175. De La Vega FM and Bustamante CD. Polygenic risk scores: a biased prediction? Genome Medicine 2018 Dec; 10:100. DOI: 10.1186/s13073-018-0610-x. Available from: https://doi.org/10.1186/s13073-018-0610-x [Accessed on: 2019 Jan 3]

176. Vyas DA, Eisenstein LG, and Jones DS. Hidden in Plain Sight — Reconsidering the Use of Race Correction in Clinical Algorithms. New England Journal of Medicine 2020 Jun; 0. Publisher: Massachusetts Medical Society _eprint: https://doi.org/10.1056/NEJMms2004740:null. DOI: 10.1056/NEJMms2004740. Available from: https://doi.org/10.1056/NEJMms2004740 [Accessed on: 2020 Jun 22]

177. Chatterjee N, Shi J, and García-Closas M. Developing and evaluating polygenic risk prediction models for stratified disease prevention. en. Nature Reviews Genetics 2016 Jul; 17:392–406. DOI: 10.1038/nrg.2016.27. Available from: https://www.nature.com/articles/nrg.2016.27 [Accessed on: 2018 Dec 19]

178. Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J, Downey P, Elliott P, Green J, Landray M, Liu B, Matthews P, Ong G, Pell J, Silman A, Young A, Sprosen T, Peakman T, and Collins R. UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. en. PLOS Medicine 2015 Mar; 12. Publisher: Public Library of Science:e1001779. DOI: 10.1371/journal.pmed.1001779. Available from: https://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.1001779 [Accessed on: 2020 Oct 22]

179. Bycroft C, Freeman C, Petkova D, Band G, Elliott LT, Sharp K, Motyer A, Vukcevic D, Delaneau O, O'Connell J, Cortes A, Welsh S, Young A, Effingham M, McVean G, Leslie S, Allen N, Donnelly P, and Marchini J. The UK Biobank resource with deep phenotyping and genomic data. En. Nature 2018 Oct; 562:203. DOI: 10.1038/s41586-018-0579-z. Available from: https://www.nature.com/articles/s41586-018-0579-z [Accessed on: 2018 Nov 27]

180. Chande AT, Wang L, Rishishwar L, Conley AB, Norris ET, Valderrama-Aguirre A, and Jordan IK. GlobAl Distribution of GEnetic Traits (GADGET) web server: polygenic trait scores worldwide. en. Nucleic Acids Research 2018 Jul; 46:W121–W126. DOI: 10.1093/nar/gky415. Available from: https://academic.oup.com/nar/article/46/W1/W121/4999244 [Accessed on: 2019 Mar 14]

181. Firmann M, Mayor V, Vidal PM, Bochud M, Pécoud A, Hayoz D, Paccaud F, Preisig M, Song KS, Yuan X, Danoff TM, Stirnadel HA, Waterworth D, Mooser V, Waeber G, and Vollenweider P. The CoLaus study: a population-based study to investigate the epidemiology and genetic determinants of cardiovascular risk factors and metabolic

syndrome. en. BMC Cardiovascular Disorders 2008 Mar; 8:6. DOI: 10.1186/1471-2261-8-6. Available from: https://doi.org/10.1186/1471-2261-8-6 [Accessed on: 2020 Sep 1]

182. Ca A, Fh P, Gm C, Lr C, Ap M, and Kt Z. Data quality control in genetic case-control association studies. English. Nature Protocols 2010 Aug; 5:1564–73. DOI: 10.1038/nprot.2010.116. Available from: https://europepmc.org/article/PMC/3025522 [Accessed on: 2020 Oct 22]

183. Loh PR, Danecek P, Palamara PF, Fuchsberger C, A Reshef Y, K Finucane H, Schoenherr S, Forer L, McCarthy S, Abecasis GR, Durbin R, and L Price A. Reference-based phasing using the Haplotype Reference Consortium panel. Nature Genetics 2016 Oct; 48:1443–8. DOI: 10.1038/ng.3679. Available from: http://www.nature.com/doifinder/10.1038/ng.3679 [Accessed on: 2017 Dec 11]

184. Birney E and Soranzo N. Human genomics: The end of the start for population sequencing. eng. Nature 2015 Oct; 526:52–3. DOI: 10.1038/526052a

185. Walter K et al. The UK10K project identifies rare variants in health and disease. en. Nature 2015 Oct; 526. Number: 7571 Publisher: Nature Publishing Group:82–90. DOI: 10.1038/nature14962. Available from: https://www.nature.com/articles/nature14962 [Accessed on: 2020 Oct 22]

186. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, et al. Scikit-learn: Machine learning in Python. Journal of machine learning research 2011; 12:2825–30

187. Loh PR, Tucker G, Bulik-Sullivan BK, Vilhjálmsson BJ, Finucane HK, Salem RM, Chasman DI, Ridker PM, Neale BM, Berger B, Patterson N, and Price AL. Efficient Bayesian mixed-model analysis increases association power in large cohorts. Nature Genetics 2015 Feb; 47:284–90. DOI: 10.1038/ng.3190. Available from: http://www.nature.com/doifinder/10.1038/ng.3190 [Accessed on: 2017 Jul 24]

188. Abraham G and Inouye M. Fast principal component analysis of large-scale genome-wide data. eng. PloS One 2014; 9:e93766. DOI: 10.1371/journal.pone.0093766

189. Martin AR, Lin M, Granka JM, Myrick JW, Liu X, Sockell A, Atkinson EG, Werely CJ, Möller M, Sandhu MS, Kingsley DM, Hoal EG, Liu X, Daly MJ, Feldman MW, Gignoux CR, Bustamante CD, and Henn BM. An Unexpectedly Complex Architecture for Skin Pigmentation in Africans. en. Cell 2017 Nov; 171:1340–1353.e14. DOI: 10.1016/j.cell.2017.11.015. Available from: http://www.sciencedirect.com/science/article/pii/S0092867417313247 [Accessed on: 2020 Jul 15]

190. Martin AR, Kanai M, Kamatani Y, Okada Y, Neale BM, and Daly MJ. Clinical use of current polygenic risk scores may exacerbate health disparities. En. Nature Genetics 2019 Apr; 51:584. DOI: 10.1038/s41588-019-0379-x. Available from: https://www.nature.com/articles/s41588-019-0379-x [Accessed on: 2019 Apr 2]

191. Petrovski S and Goldstein DB. Unequal representation of genetic variation across ancestry groups creates healthcare inequality in the application of precision medicine. Genome Biology 2016 Jul; 17:157. DOI: 10.1186/s13059-016-1016-y. Available from: https://doi.org/10.1186/s13059-016-1016-y [Accessed on: 2018 Dec 11]

192. Popejoy AB and Fullerton SM. Genomics is failing on diversity. en. Nature 2016 Oct; 538:161–4. DOI: 10.1038/538161a. Available from: http://www.nature.com/doifinder/10.1038/538161a [Accessed on: 2018 Dec 11]

193. Duncan L, Shen H, Gelaye B, Meijsen J, Ressler K, Feldman M, Peterson R, and Domingue B. Analysis of polygenic risk score usage and performance in diverse human populations. en. Nature Communications 2019 Jul; 10. Number: 1 Publisher: Nature Publishing Group:1–9. DOI: 10.1038/s41467-019-11112-0. Available from: https://www.nature.com/articles/s41467-019-11112-0 [Accessed on: 2020 May 10]

194. Dikilitas O, Schaid DJ, Kosel ML, Carroll RJ, Chute CG, Denny JA, Fedotov A, Feng Q, Hakonarson H, Jarvik GP, Lee MTM, Pacheco JA, Rowley R, Sleiman PM, Stein CM, Sturm AC, Wei WQ, Wiesner GL, Williams MS, Zhang Y, Manolio TA, and Kullo IJ. Predictive Utility of Polygenic Risk Scores for Coronary Heart Disease in Three Major Racial and Ethnic Groups. eng. American Journal of Human Genetics 2020; 106:707–16. DOI: 10.1016/j.ajhg.2020.04.002

195. Wand H, Lambert SA, Tamburro C, Iacocca MA, O'Sullivan JW, Sillari C, Kullo IJ, Rowley R, Brockman D, Venner E, McCarthy MI, Antoniou AC, Easton DF, Hegele RA, Khera AV, Chatterjee N, Kooperberg C, Edwards K, Vlessis KR, Kinnear K, Danesh JN, Parkinson H, Ramos EM, Roberts MC, Ormond KE, Khoury MJ, Janssens ACJ, Goddard KA, Kraft P, MacArthur JA, Inouye M, and Wojcik GL. Improving reporting standards for polygenic scores in risk prediction studies. en. medRxiv 2020 May. Publisher: Cold Spring Harbor Laboratory Press:2020.04.23.20077099. DOI: 10.1101/2020.04.23.20077099. Available from: https://www.medrxiv.org/content/10.1101/2020.04.23.20077099v1 [Accessed on: 2020 Jul 3]

196. Lambert SA, Gil L, Jupp S, Ritchie SC, Xu Y, Buniello A, Abraham G, Chapman M, Parkinson H, Danesh J, MacArthur JA, and Inouye M. The Polygenic Score Catalog: an open database for reproducibility and systematic evaluation. en. medRxiv 2020 May. Publisher: Cold Spring Harbor Laboratory Press:2020.05.20.20108217. DOI: 10.1101/2020.05.20.20108217. Available from: https://www.medrxiv.org/content/10.1101/2020.05.20.20108217v1 [Accessed on: 2020 Jul 3]

197. Morales J, Welter D, Bowler EH, Cerezo M, Harris LW, McMahon AC, Hall P, Junkins HA, Milano A, Hastings E, Malangone C, Buniello A, Burdett T, Flicek P, Parkinson H, Cunningham F, Hindorff LA, and MacArthur JAL. A standardized framework for representation of ancestry data in genomics studies, with application to the NHGRI-EBI GWAS Catalog. Genome Biology 2018 Feb; 19. DOI: 10.1186/s13059-018-1396-2. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5815218/ [Accessed on: 2020 Jul 9]

198. Sandra L. Colby JMO. Projections of the Size and Composition of the U.S: 2014-2060. EN-US. The United States Census Bureau 2015. Ed. by Bureau UC. Library Catalog: www.census.gov Section: Government. Available from: https://www.census.gov/library/publications/2015/demo/p25-1143.html [Accessed on: 2020 Jul 14]

199. Williams DR, Priest N, and Anderson N. Understanding Associations between Race, Socioeconomic Status and Health: Patterns and Prospects. Health psychology : official journal of the Division of Health Psychology, American Psychological Association 2016 Apr; 35:407–11. DOI: 10.1037/hea0000242. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4817358/ [Accessed on: 2020 Jul 15]

200. Moser G, Lee SH, Hayes BJ, Goddard ME, Wray NR, and Visscher PM. Simultaneous Discovery, Estimation and Prediction Analysis of Complex Traits Using a Bayesian Mixture Model. en. PLOS Genetics 2015 Apr; 11. Publisher: Public Library of Science:e1004969. DOI: 10.1371/journal.pgen.1004969. Available from: https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1004969 [Accessed on: 2021 Mar 16]

201. Speed D, Holmes J, and Balding DJ. Evaluating and improving heritability models using summary statistics. en. Nature Genetics 2020 Apr; 52. Number: 4 Publisher: Nature Publishing Group:458–62. DOI: 10.1038/s41588-020-0600-y. Available from: https://www.nature.com/articles/s41588-020-0600-y [Accessed on: 2021 Mar 10]

202. Zhang Q, Privé F, Vilhjálmsson B, and Speed D. Improved genetic prediction of complex traits from individual-level data or summary statistics. en. preprint. Genetics, 2020 Aug. DOI: 10.1101/2020.08.24.265280. Available from: http://biorxiv.org/lookup/doi/10.1101/2020.08.24.265280 [Accessed on: 2021 Mar 16]

203. Prive F and Arbel J. LDpred2: better, faster, stronger. en :8

204. Lloyd-Jones LR, Zeng J, Sidorenko J, Yengo L, Moser G, Kemper KE, Wang H, Zheng Z, Magi R, Esko T, Metspalu A, Wray NR, Goddard ME, Yang J, and Visscher PM. Improved polygenic prediction by Bayesian multiple regression on summary statistics. en. Nature Communications 2019 Nov; 10. Number: 1 Publisher: Nature Publishing Group:5086. DOI: 10.1038/s41467-019-12653-0. Available from: https://www.nature.com/articles/s41467-019-12653-0 [Accessed on: 2021 Mar 5]

205. Wilson PW, D'Agostino RB, Levy D, Belanger AM, Silbershatz H, and Kannel WB. Prediction of coronary heart disease using risk factor categories. eng. Circulation 1998 May; 97:1837–47. DOI: 10.1161/01.cir.97.18.1837

206. Assmann Gerd, Cullen Paul, and Schulte Helmut. Simple Scoring Scheme for Calculating the Risk of Acute Coronary Events Based on the 10-Year Follow-Up of the Prospective Cardiovascular Münster (PROCAM) Study. Circulation 2002 Jan; 105. Publisher: American Heart Association:310–5. DOI: 10.1161/hc0302.102575. Available from: https://www.ahajournals.org/doi/10.1161/hc0302.102575 [Accessed on: 2020 Dec 20]

207. Wilson PWF, Meigs JB, Sullivan L, Fox CS, Nathan DM, and D'Agostino RB. Prediction of incident diabetes mellitus in middle-aged adults: the Framingham Offspring Study. eng. Archives of Internal Medicine 2007 May; 167:1068–74. DOI: 10.1001/archinte.167.10.1068

208. Hippisley-Cox J, Coupland C, Vinogradova Y, Robson J, Minhas R, Sheikh A, and Brindle P. Predicting cardiovascular risk in England and Wales: prospective derivation and validation of QRISK2. en. BMJ 2008 Jun; 336. Publisher: British Medical Journal Publishing Group Section: Research:1475–82. DOI: 10.1136/bmj.39609.449676.25. Available from: https://www.bmj.com/content/336/7659/1475 [Accessed on: 2020 Dec 20]

209. Niemi MEK, Martin HC, Rice DL, Gallone G, Gordon S, Kelemen M, McAloney K, McRae J, Radford EJ, Yu S, Gecz J, Martin NG, Wright CF, Fitzpatrick DR, Firth HV, Hurles ME, and Barrett JC. Common genetic variants contribute to risk of rare severe neurodevelopmental disorders. en. Nature 2018 Sep :1. DOI: 10.1038/s41586-018-0566-4. Available from: https://www.nature.com/articles/s41586-018-0566-4 [Accessed on: 2018 Sep 27]

210. Weiner DJ, Wigdor EM, Ripke S, Walters RK, Kosmicki JA, Grove J, Samocha KE, Goldstein JI, Okbay A, Bybjerg-Grauholm J, Werge T, Hougaard DM, Taylor J, Skuse D, Devlin B, Anney R, Sanders SJ, Bishop S, Mortensen PB, Børglum AD, Smith GD, Daly MJ, and Robinson EB. Polygenic transmission disequilibrium confirms that common and rare variation act additively to create risk for autism spectrum disorders. en. Nature Genetics 2017 Jul; 49. Number: 7 Publisher: Nature Publishing Group:978–85. DOI: 10.1038/ng.3863. Available from: https://www.nature.com/articles/ng.3863 [Accessed on: 2020 Dec 19]

211. Sparks AB, Wang ET, Struble CA, Barrett W, Stokowski R, McBride C, Zahn J, Lee K, Shen N, Doshi J, Sun M, Garrison J, Sandler J, Hollemon D, Pattee P, Tomita-Mitchell A, Mitchell M, Stuelpnagel J, Song K, and Oliphant A. Selective analysis of cell-free DNA in maternal blood for evaluation of fetal trisomy. en. Prenatal Diagnosis 2012; 32. _eprint: https://obgyn.onlinelibrary.wiley.com/doi/pdf/10.1002/pd.2922:3–9. DOI: https://doi.org/10.1002/pd.2922. Available from: https://obgyn.onlinelibrary.wiley.com/doi/abs/10.1002/pd.2922 [Accessed on: 2020 Dec 20]

212. Wilson C. Exclusive: A new test can predict IVF embryos' risk of having a low IQ. en-US. Available from: https://www.newscientist.com/article/mg24032041-900-exclusive-a-new-test-can-predict-ivf-embryos-risk-of-having-a-low-iq/ [Accessed on: 2020 Dec 19]

213. Letter O. Opinion: How Not To Talk About Race And Genetics. en-US. BuzzFeed News 2018 Mar. Available from: https://www.buzzfeednews.com/article/bfopinion/race-genetics-david-reich [Accessed on: 2020 Jul 25]

214. Reich D. Opinion | How Genetics Is Changing Our Understanding of 'Race'. en-US. The New York Times 2018 Mar. Available from: https://www.nytimes.com/2018/03/23/opinion/sunday/genetics-race.html [Accessed on: 2020 Jul 25]

215. Willer CJ et al. Discovery and refinement of loci associated with lipid levels. eng. Nature Genetics 2013 Nov; 45:1274–83. DOI: 10.1038/ng.2797

216. Kalia SS, Adelman K, Bale SJ, Chung WK, Eng C, Evans JP, Herman GE, Hufnagel SB, Klein TE, Korf BR, McKelvey KD, Ormond KE, Richards CS, Vlangos CN, Watson M, Martin CL, and Miller DT. Recommendations for reporting of secondary findings in clinical exome and genome sequencing, 2016 update (ACMG SF v2.0): a policy statement of the American College of Medical Genetics and Genomics. en. Genetics in Medicine 2017 Feb; 19. Number: 2 Publisher: Nature Publishing Group:249–55. DOI: 10.1038/gim.2016.190. Available from: https://www.nature.com/articles/gim2016190 [Accessed on: 2020 Dec 20]

217. Polygenic Risk Scores: Combining Thousands of Genetic Variants to Predict Disease. Available from: https://www.rgare.com/knowledge-center/media/articles/polygenic-risk-scores-combining-thousands-of-genetic-variants-to-predict-disease [Accessed on: 2020 Dec 19]

Olivier Naret

PhD student

Education

2017–present **Doctoral school EDBB (Biotechnologies & Bioingénierie)**, École polytechnique fédérale de Lausanne, Lausanne.
(2018) Deep learning (EE-559)
(2017) Economics of innovation in the biomedical industry (MGT-403)

2015–2016 **Master degree (M.Sc II) in System biology**, Université Paul Sabatier, Toulouse.
Bioinformatic techniques, NGS, machine learning.

2013–2014 **Master degree (M.Sc II) in Computer science**, Université Joseph Fourier, Grenoble.

2009–2012 **Bachelor degree and M.Sc I in Biology**, Université Pierre et Marie Curie, Paris.

Experience

Biologist

Jul2019-Dec2020 **Visiting Student**, Stanford, Pritchard Lab, Stanford.
Studying the validity of a competition based omnigenic model

2017–present **PhD Student**, Ecole Polytechnique Federale de Lausanne, Fellay Lab, Lausanne.
Developing methods for the genomic prediction of complex human traits to promote the application of personalized medicine.
Setting up a challenge on the CrowdAI platform in collaboration with OpenSNP. CrowdAI is an online platform developed by Profesor Marcel Salathe (EPFL) which provide various thematic challenges that can be freely undertaken at the condition of using deep learning techniques. Thanks to our collaboration with OpenSNP platform, we were able to set up a challenge, which aims to predict height using genotypic data of the OpenSNP community.

2016 **Reasearch Assistant**, Ecole Polytechnique Federale de Lausanne, Fellay Lab, Lausanne.
Development of a R framework to simulate genome-to-genome (G2G) studies. G2G is a recently developed method that allow the joint analysis of genotypic data from a host organism and its pathogen. The framework allowed us to elaborate the best statistical model to correct for both host and pathogen stratification caused by systematic ancestry difference. The results first predicted by simulations were observed on a cohort of patient infected by HIV-1.

2013 **Research Internship**, UPMC Oceanological Observatory, Banyuls.
Characterization of the quorum sensing activity within the marine environment thanks to metagenomic data from the Global Ocean Sampling campaign.

École Polytechnique Fédérale de Lausanne
SV GHI GR-FE, AAB 2.19, CH-1015 Lausanne, Switzerland
(041) 767 798 507 • onaret@gmail.com

Software Engineer

2014–2015 **Software Engineer**, AXWAY SOFTWARE, Annecy.
Within an international Software Vendor Company.
Architect and developer of a test automation tools server (Java, Maven) and web client (HTML/CSS, PHP, AJAX, Javascript (JQuery)), relying on a PostgreSQL database for the continuous integration platform (Jenkins) of an Axway product (AISuite Datastore).
Developer (Java) in Agile environment (Scrum) on an Axway Product (Track & Trace).
On the main Axway Product (Central Governance). Implemented an interface for the common shared toolbox with JBehave and spread the BDD (Behavior Driven Development) best practices to QA Engineers (how to write acceptance tests with natural language).

Publications, Conferences and Grant

2020 **O. Naret, J. Fellay**, *Submitted*.
Improving clinical risk models with ancestry.

2020 **O. Naret, Yuval Simons, JK. Pritchard**, *Submitted*.
Is competition for resource between genes a driving force of the omnigenic architecture of complex traits?

2020 **O. Naret, J Fellay**, *European Society of Human Genomics talk*.
Improving clinical risk models with ancestry.

2019 **O. Naret**, *Doc. Mobility*.
Grant to work 6 months in JK. Pritchard Lab in Stanford, CA, USA.

2019 **O. Naret, David AA Baranger et al., ..., J. Fellay**, *biorXiv*.
Phenotype prediction from genome-wide genotyping data: a crowdsourcing experiment.

2019 **OpenSNP-Maker**, *Software*.
Tool to create a clean up to date cohort from the publicly available OpenSNP database

2018 **O. Naret, N Chaturverdi et al., ..., J. Fellay**, *Front. Genet., 2018;10.3389*.
Correcting for Population Stratification Reduces False Positive and False Negative Results in Joint Analyses of Host and Pathogen Genomes.

2017 **G2G-Simulator**, *Software*.
Framework to generate genome-to-genome dataset with complex population stratification

2017 **O. Naret, J Fellay**, *Rapid-fire talks: Applied Machine Learning Days*.
Predicting complex phenotypes using Neural Networks.

Teaching

2018 **Genomics and bioinformatics**, *EPFL*, Teaching assistant for Prof. Jacques Rougemont.

2018 **Genetics and genomics**, *EPFL*, Teaching assistant for Prof. Jacques Fellay and Bart DePlancke.

2018 **Student supervision**, *EPFL*, Assessing machine learning techniques to predict complex phenotypes..

2018–2019 **Probability and Statistics I and II**, *EPFL*, Teaching assistant for Prof. Darlene Goldstein.

École Polytechnique Fédérale de Lausanne
SV GHI GR-FE, AAB 2.19, CH-1015 Lausanne, Switzerland
(041) 767 798 507 • onaret@gmail.com