Published in final edited form as:

Trends Genet. 2018 October ; 34(10): 746–754. doi:10.1016/j.tig.2018.07.004.

Complex-Trait Prediction in the Era of Big Data

Gustavo de los Campos^{1,2,3,*}, **Ana Ines Vazquez**^{1,2}, **Stephen Hsu**^{4,5}, and **Louis Lello**⁴

¹: Department of Epidemiology and Biostatistics, Michigan State University (MSU), East Lansing, MI, 48824.

²: Department of Statistics and Probability, MSU.

³: Institute for Quantitative Health Science and Engineering, MSU.

⁴: Department of Physics and Astronomy, MSU.

⁵: Cognitive Genomics Laboratory, BGI.

Abstract

Accurate prediction of complex traits requires using large number of DNA-variants. Advances in statistical and machine learning methodology enable the identification of complex patterns in high-dimensional settings. However, training these highly-parameterized methods requires very large data sets. Until recently, such data sets were not available. But the situation is changing rapidly as very large biomedical data sets comprising individual genotype-phenotype data for hundreds of thousands of individuals become available in public and private domains. We argue that the convergence of advances in methodology and the advent of Big Genomic Data will enable unprecedented improvements in complex trait prediction; we review theory and evidence supporting our claim and discuss challenges and opportunities that Big Data will bring to complex-trait prediction.

Keywords

GWAS; SNP; complex traits; disease risk; prediction; Big Data

Many important human traits and diseases are moderately heritable, implying that a sizable fraction of the inter-individual differences in phenotypes (either measurable traits or disease risk) can be attributed to differences in DNA sequence. For such heritable phenotypes, DNA-risk assessments can potentially achieve moderately high prediction accuracy.

In the last fifteen years, Genome-Wide Association (GWA) studies have discovered thousands of variants associated with many important human traits and diseases (e.g., <https://www.ebi.ac.uk/gwas/>); however, for complex phenotypes the accuracy of DNA-risk assessments based on GWAS-significant variants remains low. This problem, referred to as

* Correspondence: gustavoc@msu.edu (G. de los Campos).

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

the “missing heritability” of complex traits, was identified a decade ago [1,2] and remains a concern.

Four main factors affect the accuracy of a genomic predictor [3–5]: (1) the trait/disease heritability provides an upper-bound on prediction accuracy, (2) the genotyping/sequencing technology used determines how well SNPs can capture (via linkage disequilibrium, LD) variation on alleles at causal variants, (3) statistical methods differ in their ability to uncover phenotype-DNA patterns, and (4) the size (and quality) of the data set determines the accuracy of parameter (e.g., SNP-effect) estimates.

Modern SNP arrays offer dense coverage of the genome (and genome coverage can be further increased using imputation). Meanwhile, advances in statistical and machine learning methodology enable identifying complex patterns in high dimensional settings. Examples of this include Bayesian [6–10] and penalized regressions that can handle high-dimensional inputs (e.g., hundreds of thousands of SNPs) in both linear (e.g., Lasso, Elastic Net [11,12]) and non-linear settings (e.g., kernel methods, [13,14] and deep learning [15]). However, training these predictive machines requires using very large data sets. Until recently such data sets were not available, but the situation is changing rapidly as very large biomedical data sets become available in public (e.g., UK-Biobank [16], Million Veteran Program® [17]) and private (e.g., 23andMe®) domains.

We argue that the confluence of advances in methodology and the recent availability of very large biomedical data sets will enable unprecedented improvements in our ability to predict complex human traits and diseases. However, realizing this untapped potential will require a paradigm shift: from a focus on finding variants associated with phenotypes at high statistical confidence to one centered on producing accurate predictions.

Genetic Risk Prediction is an Imperfect Science

The prediction accuracy of a DNA-risk assessment is bounded by the broad-sense heritability [18] (Box 1). However, we have very imperfect knowledge of which loci affect genetic risk and how DNA-sequence translate into genetic risk; therefore, even for highly heritable traits achieving high prediction R-squared (R-sq.) is challenging. Thus, genomic risk prediction is an imperfect task which is concerned with approximating as accurately as possible the underlying phenotype-DNA patterns.

Linear models can be very powerful predictive machines.

At the causal level genes interact in many complex ways and epistasis is pervasive (e.g., [19]). However, theory and empirical evidence suggests that for most complex traits a large proportion of interindividual differences in genetic risk can be captured using linear models [20]. This much-welcomed news implies that we may achieve good prediction accuracy using relatively simple linear models. One of the reasons that explain why linear models often provide a good approximate stem from the nature of the DNA-variation. At any given loci, DNA varies among discrete classes. For SNPs, there are three possible levels (e.g., AA, AB, BB) and when minor allele frequencies are low two classes (one of the homozygous and

the heterozygous) dominate. The discrete and simple nature of genotypes induce patterns of variation in genetic risk that are often well described by a linear model [20].

Imperfect linkage disequilibrium (LD) between SNPs and causal variants impose further limits on prediction accuracy.

Genotyping is typically done using SNP arrays that include anywhere between 500K-1M (K=thousand, M=million) variants. This is only a small fraction of the total number of SNPs in the human genome and there are many forms of genetic variation other than SNPs. Thus, for complex traits most of the causal variants are not genotyped. Signals generated un-genotyped variants can be captured through LD with genotyped SNPs. However, in most cases LD would be imperfect, implying that a fraction of the total genetic signal will be missed. For this reason, the maximum prediction R-sq. that could be achieved using a linear model on SNPs (the SNP or genomic heritability, h_{SNP}^2 , [5,21]) is smaller than the narrow sense heritability (compare levels II and III in Box 1). The gap between h^2 and h_{SNP}^2 depends on the (multi-locus) correlation between the SNP used in the model and the genotypes at causal loci. Increasing marker density improves the probability of having SNPs in high LD with genotypes at causal loci.

Available estimates suggest that for panels of common SNPs the ratio $\frac{h_{SNP}^2}{h^2}$ is about 0.4–0.5 depending on the trait and SNP array used [22]. Estimates based on imputed genotypes indicate a much smaller difference between h^2 and h_{SNP}^2 [23]. However, there has been important controversy about the quality of genomic heritability estimates, especially when the number of SNPs used largely exceeds sample size [24–26]. Nonetheless, given the evidence available, one could establish that a reasonable upper-bound for prediction R-sq. is anywhere between 40–70% of the trait heritability depending on the trait and SNP-panel used. Thus, for moderately heritable traits there is potential to achieve moderately high genomic-prediction accuracy using common SNPs.

Advances in methodology meet Big Data.

Predictions use SNP genotypes and estimated SNP-effects ($\hat{\beta}$ in Box 1). Estimation errors (i.e., differences between true (β) and the estimated ($\hat{\beta}$) effects) make prediction R-sq. smaller than h_{SNP}^2 [5]. Modern statistical/machine learning methods incorporating penalization (e.g., [11,12]) and Bayesian methods (e.g., [27]) allow accurate estimation of very large number of SNP effects. However, training high-dimensional models accurately also requires access to very large data sets. Until recently, these data sets were not available; however, as noted before, this situation is changing rapidly and it is reasonable to expect that in the next decade there will be several data sources with sample size near 1 million records.

Big Data will enable unprecedented improvements in complex trait prediction

Recent studies published using data from the UK-Biobank [28–30] demonstrate for human height (a highly heritable and highly complex trait) that the use Big Data coupled with modern whole-genome regression methods (incorporating tens of thousands of SNPs) can reduce the gap between the narrow-sense heritability and prediction R-sq. The reduction of the missing-heritability-gap has two components. First, the use of large numbers of SNPs (including many that may not reach GWAS-significance) reduces the gap between the narrow-sense heritability and h_{SNP}^2 (levels II and III in Box 1, respectively). Second, the use of Big Data enables accurate estimation of effects, thus reducing the gap between h_{SNP}^2 and prediction R-sq. (levels III and IV in Box 1, respectively).

Figure 1 (adapted from Kim et al., [29]) illustrates how marker density and sample size affect prediction R-sq. The prediction R-sq. achieved with models based on a small number of SNPs (e.g., the top-500, all of them GWA-significant) increased quickly with sample size, reaching a plateau (with an R-sq. ~ 0.1 for the top-500 SNPs). Clearly, a sample size of 80K is enough to achieve high accuracy of estimated effects when only 500 SNPs are used in the model. However, this SNP panel has a small h_{SNP}^2 (~ 0.1). Models including a larger number of SNPs (many of which did not reach GWA-significance) gave much higher prediction accuracy: the authors reported an R-sq. ~ 0.22 (~ 0.24) for the top 50K SNPs (100K SNPs) when SNP effects were estimated using a sample size of 80K. The horizontal lines in Figure 1 represents estimates of the maximum R-sq. that could be achieved by each SNP set, if effects were estimated with a (conceptually) infinite large sample size. For the top-50K SNPs that maximum was estimated to be ~ 0.37 ; however, a sample size of 80K was not enough to reach that maximum.

More recently Lello et al., [30] presented prediction accuracy results based on the full release of the UK-Biobank. Figure 2 (adapted from Lello et al.) shows the prediction accuracy achieved for sex-age adjusted human height as a function of the size of the data set used for model training, for models using 50K GWAS-selected SNPs. With a sample size of $n=400K$ the study reported a prediction R-sq. of 0.38. This value (which close to the maximum forecasted by Kim et al.[29]) represents roughly 50% of the trait heritability. The study also reported results for other traits, including heel-bone mineral density and educational attainment. For those traits prediction R-sq. were lower, in part reflecting the lower heritability of these traits, compared to human height. However, in the case of bone-heel mineral density the prediction R-sq. reported was about 0.2 with a training size of 400K samples, suggesting that for that trait it may also be possible to identify subjects that are expected to have extreme phenotypes with relatively high accuracy.

The results presented above indicate that the use of Big Data can lead to sizable improvements in our ability to predict complex traits. However, capitalizing on the untapped potential of Big Data will require important changes in the way genomic prediction is approached.

A much-needed paradigm shift

Genome-wide association studies were designed to identify variants associated with phenotypes with high statistical confidence. This paradigm is effective at controlling the number of false discoveries but is not optimal for developing accurate risk assessments.

Building an accurate predictor requires using SNPs that are not GWA-significant.

Inference in GWA studies are grounded in the classical Neyman-Pearson hypothesis testing framework [31,32] which is centered on quantifying evidence against the null hypothesis (absence of SNP effect in GWA analyses). Rejection of the null hypothesis happens when evidence against it is overwhelming. This, together with the fact that GWAs test hundreds of thousands of variants simultaneously, has led to the establishment of criteria for determining GWA-significance that are extremely stringent and leave outside of the discovery set large numbers of variants that, collectively, can contribute to prediction accuracy. This is illustrated in Table 1 which shows for five solutions (obtained over the regularization path of a Lasso fit) the number of active SNPs (i.e., with non-zero estimated effect), whether these SNPs were GWAS-significant, and the prediction R-sq. achieved in a testing set. For the solution with ~1K active SNPs, all the active SNPs were GWA-significant. However, the predictive performance of that model was poor compared with that of models that including many more SNPs. Over the regularization path more SNPs become active; many of them were not GWAS-significant (e.g., for the model with ~25K active SNPs there were 15,032 SNPs that were active but did not reach $-\log_{10}(\text{pvalue}) > 5$). Conversely, for the model with ~25K active SNPs, there were 20,262 SNPs that were GWAS-significant but were not active in the Lasso fit.

There are several reasons why GWAS-significance is not an optimal criterion to select the SNPs to be used in a risk assessment. First, as already mentioned, due to the burden of multiple testing GWA-significance is overly conservative. Second, single-marker-regression analyses ignore LD. In a linear model the contribution to variance of a set of SNPs is a function of the size of the SNP-effects, allele frequency, and of the extent of LD between those variants [33], this structure is not considered in single-marker regression.

Exploiting multi-locus LD can increase prediction accuracy.

In regression analyses the use of highly correlated (or co-linear) predictors is often discouraged because high correlation among predictors leads to large standard errors. While the idea may be adequate for inference about individual SNP effects, the principle is not optimal for developing accurate predictors. In prediction, excessive LD-SNP-pruning can negatively affect prediction accuracy.

Kim et al. [29] compared the prediction accuracy of models based on 100K SNPs that were pre-selected using GWAS results with a selection criteria that either ignored LD among the selected SNPs (top-p) versus one that selected only one SNP per LD-block. Selecting SNPs based on GWAS p-values ignoring LD leads to subsets of SNPs in high LD. However, the model based on the top-p SNPs outperformed the predictive performance of the model using only one SNP per LD block. Often no single SNP in a panel is in perfect LD with variants at

the causal loci. Imperfect LD between markers and QTL is one of the main factors hindering prediction accuracy (Box 1). Combining information from multiple SNPs (e.g., using a linear combination of two or more SNPs in an LD block) can capture a larger fraction of the signal generated by an-unobserved causal variant in the same block than any of the SNPs individually. You see a person in a raincoat or a person who is wet-any of these might give the impression it is raining, but noticing a person who is wearing a raincoat and is wet improves your accuracy of guessing it is raining.

Challenges and Opportunities

Risk assessment for diseases with moderately low prevalence will remain challenging.

Very large population data sets may offer reasonably high power for deriving risk assessments for prevalent diseases (e.g., type-2 diabetes, depression). However, population data sets may not provide enough number of cases for diseases with low prevalence (e.g., pancreatic cancer). Several studies have proposed statistical methods to combine data from case-control studies and (mostly) controls from population data [36,37]. However, there are important statistical challenges that remain un-resolved.

Non-random sampling.

Large prospective cohorts such as the UK-Biobank are designed to be representative of a target population. However, due to selection, most observational studies are not strictly random samples. This problem is potentially even more important when data sets originate from electronic medical records of patients admitted to health institutions. Further research is needed to assess how sample selection may affect the accuracy of a genomic predictor when such predictor is applied to randomly chosen subjects.

Increases in sample size may also be accompanied with **increased heterogeneity** (e.g., ethnic and socio-economic differences, heterogeneous exposures and differential access to health care) which in turn may lead to inhomogeneous SNP-effects. In its simplest form, SNP-effects may vary between known disjoint groups (e.g., sex, ethnicity). These forms of effect-heterogeneity can be modeled using SNP-by-group random-effect interactions [38] or with multivariate mixed models [39]. Results from GWAS (e.g., [40]) and from mixed models (e.g., [39]) suggest that the genetic architecture of many complex traits varies between male and female; therefore, advancing complex-trait prediction will require in many cases adequate modelling of sex differences.

Likewise, differences in allele frequencies, as well as cultural and socio-economic differences (e.g., diet, access to health care), can make SNP effects vary between ethnic groups. The clear majority of the available genomic data originates from White Caucasian individuals living in developed countries (predominantly Europe and the US). For non-Caucasian individuals, the available sample sizes are typically small and often originates from subpopulations that are highly heterogeneous (e.g., Latinos in the US). Unfortunately, trans-ethnic prediction leads to a reduction in prediction accuracy (see [29,30]). Thus, further investment in large scale genotyping/phenotyping of under-represented groups and the development of methods that can enable borrowing of information between groups (e.g.,

multi-ethnic prediction methods) would be needed to advance our ability to predict disease risk for non-Caucasian individuals. Otherwise, there will be an increasing disparity induced by un-even advancement of scientific knowledge.

The accuracy of a DNA-risk assessment is bounded by heritability. For low-heritability traits, incorporating information from risk phenotypes (smoking, chronic excessive alcohol consumption) and omic data (e.g., methylation, gene expression) can aid to improve prediction accuracy. **Omic data and risk phenotypes** can be integrated into risk assessments additively (i.e., in models where total risk is the sum of DNA-risk plus other sources of risk) or considering interactions. For risk phenotypes that are defined as presence/absence (e.g., smoking) SNP-by-exposure interactions can be modeled with the same techniques that are used to model sex and ethnic differences. Interactions between SNPs and multi-dimensional exposures (e.g., diet) or multidimensional omics can be modeled using Gaussian process [41,42].

Omic information can also be very valuable for predicting disease progression [42,43]. An important challenge with omic data stems from the fact that the expression of many omics varies in space (e.g., tissue) and time. Assessing omics in the relevant tissue at the relevant time is critical for the success of the integration of omic data into models for prediction of disease risk and progression.

Concluding Remarks and Future Directions

The confluence of Big Data (from biobanks and electronic medical records) and of advances in high-dimensional regression methodology will enable unprecedented improvements in our ability to predict complex traits and diseases. However, developing accurate risk assessments will require a paradigm shift, from one focused on detection of variants associated to risk with high statistical confidence to one focused on maximizing prediction accuracy.

The development of methods and algorithms that can be effectively used to derive risk assessments with big genomic data sets using hundreds of thousands of SNPs should represent an important research priority. Accounting for various forms of heterogeneity (e.g., sex and ethnic differences) and incorporating risk phenotypes into models for prediction of disease risk and progression are two clear avenues to further advance complex-trait prediction.

Acknowledgments

GDLC and AIV received support from NIH awards R01GM09992 and R01GM101219.

Glossary

Heritability

the proportion of phenotypic variance that can be explained by genetic factors. Additive and non-additive (i.e., dominance and epistatic) effects contribute to the *broad-sense heritability*

(H^2). The *narrow-sense heritability* (h^2) quantifies the proportion of phenotypic variance that can be explained by additive effects alone

SNP-heritability

the proportion of phenotypic variance that can be explained by a set of variants (e.g., SNPs). The difference between h^2_{SNP} and h^2 depends on the extent of linkage-disequilibrium between SNPs and causal variants

Prediction R-squared (R-sq.)

the proportion of phenotypic variance that can be explained by a risk-score. It depends on h^2_{SNP} and on the accuracy of the estimated effects

High-dimensional regression

a regression problem involving a very large number of predictors (e.g., hundreds of thousands of SNPs)

Regularized regression

a method for estimating parameters in a regression model that balances model goodness of fit and model complexity

Linkage Disequilibrium

a measure of the correlation of alleles at two or more loci

References

1. Maher B (2008) Personal genomes: the case of the missing heritability. *Nature* 456, 18–21 [PubMed: 18987709]
2. Manolio TA et al. (2009) Finding the missing heritability of complex diseases. *Nature* 461, 747–753 [PubMed: 19812666]
3. Hastie T et al. (2009) *The Elements of Statistical Learning*, Springer.
4. Daetwyler HD et al. (2008) Accuracy of Predicting the Genetic Risk of Disease Using a Genome-Wide Approach. *PLoS One* 3,
5. Goddard ME (2009) Genomic selection: prediction of accuracy and maximisation of long term response. *Genetica* 136, 245–252 [PubMed: 18704696]
6. George EI and McCulloch RE (1993) Variable Selection via Gibbs Sampling. *J. Am. Stat. Assoc* 88, 881–889
7. Ishwaran H et al. Spike and slab variable selection: frequentist and Bayesian strategies. [projecteuclid.org at <https://projecteuclid.org/euclid.aos/1117114335>](https://projecteuclid.org/euclid.aos/1117114335)
8. Meuwissen T and Goddard M (2010) Accurate Prediction of Genetic Values for complex Traits by Whole-Genome Resequencing. *Genetics* 185, 623–631 [PubMed: 20308278]
9. Gianola D et al. (2009) Additive Genetic variability and the {B}ayesian Alphabet. *Genetics* 183, 347–363 [PubMed: 19620397]
10. de los Campos G et al. (2013) Whole Genome Regression and Prediction Methods Applied to Plant and Animal Breeding. *Genetics* 193, 327–345 [PubMed: 22745228]
11. Tibshirani R (1996) Regression shrinkage and selection via the {LASSO}. *J. R. Stat. Soc. Ser. B* 58, 267–288
12. Zou H and Hastie T (2005) Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. {B}* 67, 301–320
13. Cristianini N and Shawe-Taylor J (2000) *An introduction to support vector machines : and other kernel-based learning methods*, Cambridge University Press.

14. de los Campos G et al. (2010) Semi-parametric Genomic-Enabled Prediction of Genetic Values Using Reproducing Kernel Hilbert Spaces Methods. *Genet. Res. (Camb)* 92, 295–308 [PubMed: 20943010]
15. LeCun Y et al. (2015) Deep learning. *Nature* 521, 436–444 [PubMed: 26017442]
16. Sudlow C et al. (2015) UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med.* 12, e1001779 [PubMed: 25826379]
17. Gaziano JM et al. (2016) Million Veteran Program: A mega-biobank to study genetic influences on health and disease. *J. Clin. Epidemiol* 70, 214–223 [PubMed: 26441289]
18. Falconer DS and Mackay TFC (1996) *Introduction to Quantitative Genetics*, Longman.
19. Phillips PC (2008) Epistasis--the essential role of gene interactions in the structure and evolution of genetic systems. *Nat. Rev. Genet* 9, 855–67 [PubMed: 18852697]
20. Hill WG et al. (2008), Data and Theory Point to Mainly Additive Genetic Variance for Complex Traits., in *PLoS Genetics*
21. de los Campos G et al. (2015) Genomic Heritability: What Is It? *PLOS Genet.* 11, e1005048 [PubMed: 25942577]
22. Yang J et al. (2010) Common SNPs explain a large proportion of the heritability for human height. *Nat. Genet* 42, 565–9 [PubMed: 20562875]
23. Yang J et al. (2016) Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index. *Nat. Genet* 47, 1114–1122
24. Speed D et al. (2012) Improved heritability estimation from genome-wide {SNPs}. *Am. J. Hum. Genet* 91, 1011–1021 [PubMed: 23217325]
25. Kumar KK et al. (2015) Limitations of {GCTA} as a solution to the missing heritability problem. *Proc. Natl. Acad. Sci* 113, E61–E70 [PubMed: 26699465]
26. Lehermeier C et al. (2017) Genomic variance estimates: With or without disequilibrium covariances? *J. Anim. Breed. Genet* 134, 232–241 [PubMed: 28508483]
27. Gianola D et al. (2009) Additive genetic variability and the Bayesian alphabet. *Genetics* 183, 347–63 [PubMed: 19620397]
28. Canela-Xandri O et al. (2016) Improved Genetic Profiling of Anthropometric Traits Using a Big Data Approach. *PLoS One* 11, e0166755 [PubMed: 27977676]
29. Kim H et al. (2017) Will Big Data Close the Missing Heritability Gap? *Genetics* 207, 1135–1145 [PubMed: 28893854]
30. Lello L et al. (2017) Accurate Genomic Prediction Of Human Height. *bioRxiv DOI:* 10.1101/190124
31. Neyman J and Pearson ES (1933) On the Problem of the Most Efficient Tests of Statistical Hypothesis. *Philos. Trans. R. Soc. Ser. A* 231, 289–337
32. Lehmann EL (1986) *Testing Statistical Hypotheses*, Springer-Verlag.
33. de Los Campos G et al. (2015) Genomic heritability: what is it? *PLoS Genet.* 11, e1005048 [PubMed: 25942577]
34. Meuwissen THE et al. (2001) Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps. *Genetics* 157, 1819–1829 [PubMed: 11290733]
35. de los Campos G et al. (2010) Predicting genetic predisposition in humans: the promise of whole-genome markers. *Nat. Rev. Genet* 11, 880–886 [PubMed: 21045869]
36. Derkach A et al. (2014) Association analysis using next-generation sequence data from publicly available control groups: the robust variance score statistic. *Bioinformatics* 30, 2179–88 [PubMed: 24733292]
37. Lee S et al. (2017) Improving power for rare-variant tests by integrating external controls. *Genet. Epidemiol* 41, 610–619 [PubMed: 28657150]
38. de Los Campos G et al. (2015) Incorporating Genetic Heterogeneity in Whole-Genome Regressions Using Interactions. *J. Agric. Biol. Environ. Stat* 20, 467–490 [PubMed: 26660276]
39. Rawlik K et al. (2016) Evidence for sex-specific genetic architectures across a spectrum of human complex traits. *Genome Biol.* 17, 166 [PubMed: 27473438]

40. Winkler TW et al. (2015) The Influence of Age and Sex on Genetic Associations with Adult Body Size and Shape: A Large-Scale Genome-Wide Interaction Study. *PLOS Genet.* 11, e1005378 [PubMed: 26426971]
41. Jarquín D et al. (2014) A reaction norm model for genomic selection using highdimensional genomic and environmental data. *Theor. Appl. Genet* 127, 595–607 [PubMed: 24337101]
42. Vazquez AI et al. (2016) Increased Proportion of Variance Explained and Prediction Accuracy of Survival of Breast Cancer Patients with Use of Whole-Genome Multiomic Profiles. *Genetics* 203, 1425–38 [PubMed: 27129736]
43. González-Reymúndez A et al. (2017) Prediction of years of life after diagnosis of breast cancer using omics and omic-by-treatment interactions. *Eur. J. Hum. Genet* 25, 538–544 [PubMed: 28272536]
44. American Psychiatric Association. and American Psychiatric Association. DSM-5 Task Force (2013) Diagnostic and statistical manual of mental disorders : DSM-5, American Psychiatric Association.

Box 1:

Heritability and Prediction Accuracy

Level	Proportion of Variance	Model	Variance Ratio
I	Broad-Sense Heritability	$y_i = g(z_i) + \psi_i$	$H^2 = \frac{Var[g(z_i)]}{Var[y_i]}$
II	Narrow-Sense Heritability	$y_i = z_i' b + \delta_i$	$h^2 = \frac{Var[z_i' b]}{Var[y_i]}$
III	SNP or Genomic-Heritability	$y_i = x_i' \beta + \varepsilon_i$	$h_{SNP}^2 = \frac{Var[x_i' \beta]}{Var[y_i]}$
IV	Prediction R-sq.	$y_i = x_i' \hat{\beta} + \hat{\varepsilon}_i$	$R^2 = \frac{Var[x_i' \hat{\beta}]}{Var[y_i]}$

y_i : phenotype or disease risk;
 z_i : genotypes at causal loci, x : SNPs used for prediction;
 b : allele substitution effects at causal loci, β : SNP-effects, $\hat{\beta}$: estimated SNP-effects.

Highlights

- Genome Wide Association (GWA) studies have discovered thousands of variants associated with many important human traits and diseases.
- However, GWA-significant variants explain only a small fraction of the trait heritability.
- Achieving high genomic prediction accuracy requires using tens or hundreds of thousands of SNPs, including many that do not reach GWA-significance.
- Penalized and Bayesian regressions can be used to fit highdimensional regressions including hundreds of thousands of predictors.
- However, training these high dimensional regressions requires using very large data sets.
- Until recently such data sets were not available, but this situation is changing rapidly.
- We argue that the convergence of advances in methodology and the advent of very large biomedical data sets (comprising hundreds of thousands of genotypes linked to phenotypes) will enable unprecedented improvements in complex-trait prediction.

Outstanding Questions

- For highly heritable complex traits such as human height whole-genome regressions fitted with very large data sets can achieve high prediction accuracy. Will this also hold for other traits?
- For diseases with low prevalence, even very large population data sets won't provide enough number of cases. What methods can be used to combine biobank data with data from case-control studies?
- Differences in allele frequency, in linkage disequilibrium and genetic-by-environment interactions may make SNP effect sex- and ethnic-group-specific. How can we effectively incorporate sex and ethnic difference in complex trait prediction?
- Several studies have demonstrated that rare variants can be responsible by a sizable fraction of the trait heritability. Will the use of rare variants improve complex-trait prediction?
- Several private organizations offer disease risk assessments. Should we develop standardized protocols for evaluating the accuracy of such risk assessments?
- The integration of genomic data with electronic medical records can lead to the creation of very large genomic data sets. However, data acquired through electronic medical records is subject to sampling bias. Will sample selection bias affect the accuracy of risk assessments derived? If so, How should that problem be addressed?

Figure 1.

Prediction R-sq. for (sex-age-adjusted) human height versus the size of the sample used to estimate effects by the number of SNPs used ($K=1,000$) in the model. Adapted from Kim et al. [29] who reported prediction R-sq. results based on data from the interim release of the UK-Biobank. Both training ($N=80K$) and testing set ($N=20K$) included distantly related white Caucasian individuals. Subsets of 500, 2K and 50K SNPs were selected based on a GWAS conducted in the training set. SNP effects were estimated using a Bayesian genomic regression implemented with the BGLR R-package. Dots give the average prediction R-sq. obtained the testing set. The solid lines give a non-linear curve. The dashed horizontal line gives the estimated asymptote for each curve. The authors reported an R-sq. for the top-100K SNPs of 0.24 (results not shown in the figure).

Figure 2.

Prediction R-sq. by size of the data set used to train a LASSO regression using 50K (K=1,000) SNPs. From Lello et al. [30] who reported prediction R-sq. results based on data from the full release of the UK-Biobank. Both training (~400K) and testing set (N=5K) included white Caucasian individuals that were not close relatives. Models used the top 50K GWAS-selected SNPs. SNP effects were estimated using Lasso. Dots give the average prediction R-sq. obtained the testing set. The dashed lines give an estimate of the asymptote for each model.

Table 1.

Prediction R-squared for (sex-age-adjusted) human height by number of SNPs active in a Lasso fit, and number of SNPs by status (adapted from Lello et al. [30]).

Active SNPs		Prediction R-squared	Number of SNPs By Status ^b			
Number	% ^a		Non-Active		Active	
			NS	Sig.	NS	Sig.
1025	1%	0.18	69,570	29,405	0	1,025
2529	3%	0.25	69,565	27,906	5	25,24
5056	5%	0.30	69,309	25,635	261	4,795
10110	10%	0.35	66,901	22,989	2,669	7,441
25200	25%	0.41	54,538	20,262	15,032	10,168

^a % of the 100,000 SNPs offered to the algorithm

^b Active (Non-Active): non-zero (zero) estimated effect in Lasso; NS (Sig.) -log₁₀(p-value) <5 (>5).