Genome **Medicine**

MUSINGS

What if we had whole-genome sequence data for millions of individuals?

Peter M Visscher[1*] and Greg Gibson[2]

The affordable whole-genome sequence is (nearly) here and already costs less than many commercial DNA tests for specific variants or genes. It seems very plausible to us that in the near future millions of people will have their genome sequenced, not just because the cost is coming down but also because individuals (in rich countries) are increasingly becoming 'health consumers'. Here, we muse on what we could infer from having whole-genome sequence data for a million individuals.

In the absence of phenotypic data, there are six types of information that people can obtain from their own genomes. Firstly, their ancestry: how their chromosomes compare with those of typical members of diverse human populations. Secondly, things they can do something about (for example, enzyme deficiencies, *BRCA1*-like mutations, their pharmacogenetic responses). Thirdly, things they are unable to do anything about, most of which will be a source of anxiety [1]. Fourthly, things that explain interesting or medical phenotypes (for example, eye color or rare congenital traits). Fifthly, things they may worry about handing on to their children (and which they might want to check in their (prospective) spouse or prevent their children from being at high risk for). Finally, things they may like to pass on to their children. The first four types of information in this list relate to the person whose genome has been sequenced, the last two are about their children. Of this list, (2) and (5) lead to information that is, from a health perspective, actionable right now. At present, most knowledge on these actionable items are restricted to Mendelian traits.

Actionable Mendelian variants now and in the future

Although recessive Mendelian diseases are rare in the population, the number of carriers is always orders of magnitude larger than the number of homozygotes. For example, for a recessive Mendelian disease with a prevalence of one in a million, we would expect to find a single person who has the disease among our sample of 1 million people whose whole genomes were sequenced, but we would also expect to find about 2,000 people who are carriers. If we were to check their spouses, then on average 1 in 250,000 couples would discover that they are both carriers. These frequencies are just for one very rare disease; in fact, more than 1,000 genes have been identified that cause recessive Mendelian diseases and each person is a carrier for a number of these [2], so that about 1 in 25 couples can expect to discover that they share at least one mutation that has a non-trivial chance of resulting in their offspring having a severe congenital disorder. It seems likely that many individuals or couples will avail themselves of the opportunity to prepare for this possibility by having their genomes sequenced.

Limits of disease susceptibility prediction from sequence data

For common diseases, it will soon be possible to generate genomic risk profiles from the genome sequence, using existing knowledge from genome-wide association studies (GWAS) about multiple risk variants for a disease or complex trait. Prediction of individual risk of disease is not accurate at present because, for most diseases, only a small proportion of genetic variation in risk between people has been accounted for by known genetic variants. There is also a limit to how well a genetic predictor can ever work [3,4] because common disease is caused by the combination of genetic and environmental factors.

For complex traits, the upper limit of the correlation between an as-yet-unobserved phenotype and a genetic predictor for an individual is h, the square root of the proportion of phenotypic variation that is attributable to genetic variation (or the trait's heritability (h^2)). For most complex human traits, h^2 falls in the range of 40 to 80%, so the upper limit of the precision of prediction is about

* Correspondence: peter.visscher@uq.edu.au
[1]Queensland Brain Institute and University of Queensland Diamantina Institute, The University of Queensland, Brisbane, Queensland, Australia
Full list of author information is available at the end of the article

BioMed Central © 2013 BioMed Central Ltd.

60 to 90% (√0.4 to √0.8), and that is only if we know all of the variants in the genome that affect risk and have estimated their effect sizes without error. To put the (in)accuracy of such genetic prediction into perspective, imagine if we had a perfect genetic predictor for height (in which all causal variants are known without error). Then, assuming that $h^2 = 80\%$ and that the standard deviation for height is 7 cm, the prediction error [5] for any individual would be $7 \times \sqrt{(1–0.8)} = 3.1$ cm. So we would predict someone's height with a 95% confidence interval of about ± 6 cm. With discovery data on 1 million genomes, a theoretical prediction of the prediction error is about 3.5 cm [6], not far off the minimum value possible.

What about making predictions about children who have not had their genomes sequenced (for example, because they have not been born yet)? Now the precision of prediction decreases further because we cannot predict from knowing the parental genome sequences which combination of risk alleles the children are going to inherit. The added amount of uncertainty is not trivial: of all genetic variation in the population, 50% occurs between siblings within families.

A useful comparison at this stage is to look at how animal breeders perform genetic predictions. They are mostly interested in item (6) from our list of genomic information types, and use DNA markers to make imperfect predictions that are used in selection decisions. This approach makes better predictions than the pedigree-phenotype-based approach used previously. Their business is about making money from having a progeny generation that is, on average, genetically better than the parental generation [7]. Breeders do not worry about quantifying how well they can predict an individual's phenotype because they are interested in predicting genotypes and thereby the average phenotype of progeny. We think it unlikely that humans will select their partner on the basis of genome-based polygenic predictors with an eye to maximizing the odds that all of their children are superior in some way (although something like this does occur when sperm banks are used).

What might be the utility of risk scores derived from the study of millions of genomes?

To some extent, this comes down to the question of whether it is prediction or classification that matters. For the millions of individuals who are assembling health action plans, and indulging in the self-knowledge provided by software applications that aim to promote personal fitness, it is more about skewing the odds than predicting the future. Given a prediction, whether or not someone ever gets diabetes or coronary disease is less important than having the knowledge that they have gained influence their behaviors in ways that make it less likely that they will get these diseases. In this light, knowing the aspects of your health in which you are genetically at higher risk than the rest of the population might well be just the incentive you need to keep jogging, to strive for more alcohol-free days, or to screen for cancer more regularly.

What if the millions of sequences came with detailed phenotype data? For example, this could comprise disease status for a range of common diseases and measures on quantitative traits that are risk factors for disease. This is not a far-fetched scenario. The Kaiser Permanente and University of California, San Francisco (UCSF) collaboration [8] has obtained detailed phenotyped and genotyped data for over 100,000 people, and earlier this year, it was announced that the UK Biobank sample of 500,000 people will be genotyped using a single-nucleotide polymorphism (SNP) array [9]. Phenotype and sequence data for a million people will allow the discovery of more risk and trait variants and the creation of multiple-variant profiles that can be used for prediction. But is a million sequences enough? For diseases with a prevalence of about 1%, there will be 10,000 cases among the million. Larger GWAS samples already exist for some diseases, such as Crohn's disease and schizophrenia. Although these have identified tens to hundreds of risk variants, polygenic profiles explain only a modest proportion of risk in the population, although they can do better than self-reported 'family history' [10]. Sequence data (instead of solely relying on common variants from GWAS) will improve the prediction of disease by capturing variation resulting from lower-frequency risk variants. With millions of genomes sequenced, the limitation of disease prediction for many traits is likely to result from imperfect information on environmental effects. The question is therefore not 'how well can we predict disease' but 'how can we incorporate probabilistic predictions of disease in personal or clinical decision making'. There are plenty of challenges on the way, including the generation of accurate sequence data, getting all these data together for analysis, and the statistical analysis of millions of genome sequences. Then, there will be the practical challenges of disseminating the results, not to mention encouraging people to act on them for the benefit of their health.

For quantitative traits and disease, we can expect major advances in our ability to explain the genetic component of disease risk and thus to predict disease. What we do with that information is a sociological concern with major public health implications, and now is the time to contemplate the implications.

Abbreviation

GWAS: Genome-wide association study.

Competing interests

The authors declare that they have no competing interests.

Acknowledgments

We thank Bruce Weir and Bill Hill for helpful comments, and acknowledge support from NIH grant GM099568.

Author details

[1]Queensland Brain Institute and University of Queensland Diamantina Institute, The University of Queensland, Brisbane, Queensland, Australia.
[2]School of Biology, Georgia Institute of Technology, Atlanta, GA 30332, USA.

Published: 20 September 2013

References

1. Biesecker LG, Burke W, Kohane I, Plon SE, Zimmern R: **Next-generation sequencing in the clinic: are we ready?** *Nat Rev Genet* 2012, **13:**818–824.
2. Bell CJ, Dinwiddie DL, Miller NA, Hateley SL, Ganusova EE, Mudge J, Langley RJ, Zhang L, Lee CC, Schilkey FD, Sheth V, Woodward JE, Peckham HE, Schroth GP, Kim RW, Kingsmore SF: **Carrier testing for severe childhood recessive diseases by next-generation sequencing.** *Sci Transl Med* 2011, **3:**65ra4.
3. Janssens AC, Aulchenko YS, Elefante S, Borsboom GJ, Steyerberg EW, van Duijn CM: **Predictive testing for complex diseases using multiple genes: fact or fiction?** *Genet Med* 2006, **8:**395–400.
4. Wray NR, Yang J, Goddard ME, Visscher PM: **The genetic interpretation of area under the ROC curve in genomic profiling.** *PLoS Genet* 2010, **6:** e1000864.
5. Henderson CR: **Best linear unbiased estimation and prediction under a selection model.** *Biometrics* 1975, **31:**423–447.
6. Hayes BJ, Visscher PM, Goddard ME: **Increased accuracy of artificial selection by using the realized relationship matrix.** *Genet Res (Camb)* 2009, **91:**47–60.
7. Goddard ME, Hayes BJ: **Mapping genes for complex traits in domestic animals and their use in breeding programmes.** *Nat Rev Genet* 2009, **10:**381–391.
8. Hoffmann TJ, Kvale MN, Hesselson SE, Zhan Y, Aquino C, Cao Y, Cawley S, Chung E, Connell S, Eshragh J, Ewing M, Gollub J, Henderson M, Hubbell E, Iribarren C, Kaufman J, Lao RZ, Lu Y, Ludwig D, Mathauda GK, McGuire W, Mei G, Miles S, Purdy MM, Quesenberry C, Ranatunga D, Rowell S, Sadler M, Shapero MH, Shen L, *et al*: **Next generation genome-wide association tool: design and coverage of a high-throughput European-optimized SNP array.** *Genomics* 2011, **98:**79–89.
9. **The UK Biobank.** [http://www.ukbiobank.ac.uk].
10. Do CB, Hinds DA, Francke U, Eriksson N: **Comparison of family history and SNPs for predicting risk of complex disease.** *PLoS Genet* 2012, **8:**e1002973.

doi:10.1186/gm484
Cite this article as: Visscher and Gibson: **What if we had whole-genome sequence data for millions of individuals?.** *Genome Medicine* 2013 **5**:80.