Comment on “The Predictive Capacity of Personal Genome Sequencing”

David Golan and Saharon Rosset*

Recalculation of the results in Roberts et al. shows that the true predictive capacity of genomes may be higher than their maximal estimates.

In their recent paper, Roberts et al. (1) take on the formidable task of exploring the limits of utility of genetic information for medical purposes. They define the notion of clinically significant risk threshold and denote by PF (positive fraction) the portion of the general population whose risk exceeds this threshold. They formulate an optimization problem of finding the genetic risk distributions that maximize and minimize PF, subject to being consistent with observed disease patterns among monozygotic twins, and propose a heuristic approach for finding these extreme distributions over genetic risk levels (which they term genometypes). Using these distributions, they quantify the limits of utility of genetic information for 27 conditions (24 diseases, 3 of which are separated by gender) by considering the fraction of cases that could receive positive tests (high genetic risk), the fraction of the general population that could test positive (PF), and the relative risk of people testing negative. They conclude that, although most of the population would test negative for most diseases, the positive tests could cover a significant portion of actual cases. For most diseases, negative tests carry little reduction in risk, but for some, like coronary heart disease (CHD) in males, negative tests could be highly informative. Their analysis emphasizes the upper ends of their results, which they claim “represent an absolute upper bound” on the utility of genetic information.

In this letter, we improve on the results of (1) and critically examine their generality and their utility. We first note that the optimization problem they pose always has a simple optimal solution, which assigns non-zero probability to only three genometypes, and that the risk levels of these genometypes are fully determined by the parameters set in (1) (the minimal risk level, which they denote $e$, and the clinical relevance threshold, which they denote $t$). The actual monozygotic twin data only determine the probability associated with each genometype. This result is a direct consequence of viewing the optimization problem as a semi-infinite linear program and applying known results (2). Furthermore, we can similarly assert that the same three-genometype solutions optimize all three measures considered by (1). Finding the true optimal solution has two pertinent consequences: First, we are able to obtain the exact maximal ranges between minimum and maximum genetic predictive power, which are slightly larger than those of (1). Second, this implies that the extreme genometype distributions should not be considered as indicative of population risk distribution because their nonzero probability locations are determined by the problem formulation.

Another aspect that may contribute to increasing the ranges is data uncertainty. Monozygotic twin data should be considered to be a sample,
and the quantities calculated from it are consequently uncertain. In Fig. 1, A and B, we show the result of taking into account this uncertainty by calculating bootstrap-t 95% confidence intervals (3), which significantly increases the range of possible values for many conditions, in particular, those where the overall number of monozygotic twin pairs, and especially the number of concordant pairs, is small. Note that heritability estimates used in the analysis also carry uncertainty, which we do not know, and accounting for it could cause further increase in the ranges.

A crucial practical question is whether the ranges calculated with the methodology in (1) are indicative of possible true values, even after our corrections, and even under the assumptions that the authors choose to make, such as no gene-environment interaction. One key reason why that might not be the case is the approach they select for accounting for environmental effects, by including a fixed lower bound denoted by \( e \) in the risk distribution, capturing the environmental portion of the risk. It can be shown that this is equivalent to assuming that the environmental risk of monozygotic twins is independent between the twins. That is, if we solve the problem under the assumption that the environmental effect is not fixed but independent of the genetic effect and has expectation \( e \), we obtain the same range as in (1). The independence assumption is of course problematic because we know that monozygotic twins also share part of the environmental effect. In our experiments, fitting variations of the model in (1), which allow for sharing some environment between monozygotic twins, led to a significant increase of the range for some conditions, like dystocia and CHD in males, and small changes in others. Although detailed discussion of the different ways in which this assumption can be relaxed and their implications is beyond the scope of this letter, it is clear that this is a key to improving generality and applicability of the results.

Another view of the generality issue can be obtained by using an alternative approach for calculation and comparing results. As the authors of (1) note, there is standard methodology for estimating the same quantities on the basis of the liability threshold (LT) model (4). They also show in their table S4 the LT-based estimates for the proportion of cases testing positive. We repeat their calculation and include confidence intervals, as shown in Fig. 1. Although they do not comment about these results, the LT-based estimates in Fig. 1 are larger than the maximum of the hypothesized range reported in (1) for 15 of 27 conditions they consider, statistically significantly larger for 13 of these 15. In general, the LT-based estimates have a tendency to imply higher (more predictive) values for the utility of genomes, even compared to the most optimistic estimates of (1). For example, using the LT-based results, we estimate that in expectation 48.9% of the total 8393 cases would test positive for increased risk of their actual condition, compared to an absolute upper bound of 45.5% in (1), when considering their upper bound for all conditions.

In summary, our analyses suggest that the true predictive capacity of whole genomes outside the ranges estimated in (1), specifically higher than their most optimistic estimates.

REFERENCES AND NOTES


Acknowledgments: We thank K. Skorecki and R. Shamir for useful comments. Funding: Israeli Science Foundation grant 1227/09 and IBM Open Collaborative Research grant. D.G. was also supported in part by a fellowship from the Edmond J. Safra Center for Bioinformatics at Tel-Aviv University.

Author contributions: D.G. and S.R. conceived the research, performed the experiments, and wrote the paper together. Competing interests: The authors declare that they have no competing interests.

Submitted 23 April 2012
Accepted 5 May 2012
Published 23 May 2012
10.1126/scitranslmed.3004197