Science Translational Medicine
AAAS

The Predictive Capacity of Personal Genome Sequencing
Nicholas J. Roberts *et al.*
Sci Transl Med , (2012);
DOI: 10.1126/scitranslmed.3003380

Science Translational Medicine (print ISSN 1946-6234; online ISSN 1946-6242) is published weekly, except the last week in December, by the American Association for the Advancement of Science, 1200 New York Avenue NW, Washington, DC 20005.

Science Translational Medicine
Rapid Publication
Research Article

The Predictive Capacity of Personal Genome Sequencing

Nicholas J. Roberts,[1*] Joshua T. Vogelstein,[2*] Giovanni Parmigiani,[3] Kenneth W. Kinzler,[1] Bert Vogelstein,[1†] Victor E. Velculescu[1†]

[1]Ludwig Center for Cancer Genetics and Therapeutics and The Howard Hughes Medical Institute at Johns Hopkins Kimmel Cancer Center, Baltimore, MD 21231, USA. [2]Johns Hopkins University, Department of Neuroscience, Baltimore, MD 21205, USA. [3]Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute, and Department of Biostatistics, Harvard School of Public Health, Boston, MA 02115, USA.

*These authors contributed equally to this work.

†To whom correspondence should be addressed. E-mail: vogelbe@jhmi.edu (B.V.); velculescu@jhmi.edu (V.E.V)

New DNA sequencing methods will soon make it possible to identify all germline variants in any individual at a reasonable cost. However, the ability of whole-genome sequencing to predict predisposition to common diseases in the general population is unknown. To estimate this predictive capacity, we use the concept of a "genometype". A specific genometype represents the genomes in the population conferring a specific level of genetic risk for a specified disease. Using this concept, we estimated the capacity of whole-genome sequencing to identify individuals at clinically significant risk for 24 different diseases. Our estimates were derived from the analysis of large numbers of monozygotic twin pairs; twins of a pair share the same genometype and therefore identical genetic risk factors. Our analyses indicate that: (i) for 23 of the 24 diseases, the majority of individuals will receive negative test results, (ii) these negative test results will, in general, not be very informative, as the risk of developing 19 of the 24 diseases in those who test negative will still be, at minimum, 50 - 80% of that in the general population, and (iii) on the positive side, in the best-case scenario more than 90% of tested individuals might be alerted to a clinically significant predisposition to at least one disease. These results have important implications for the valuation of genetic testing by industry, health insurance companies, public policy makers and consumers.

INTRODUCTION

As a result of continuing advances in high-throughput sequencing technologies (*1*–*4*), whole-genome sequencing will soon become an affordable approach to identify all sequence variants in an individual human. Recent evidence suggests that each human genome has more than 3 million sequence variants, some common, some infrequent (*5*). To date, several thousand genomic variants have been associated with human diseases, either as rare variants in Mendelian disorders or as common SNPs in genome-wide association studies (GWAS) (*6*, *7*). Whole-genome or whole-exome sequencing has recently been used to identify new disease predisposing variants in various familial disorders, such as familial pancreatic cancer (*8*) and Miller syndrome (*9*). However, the potential utility of genome-wide sequencing for personalized medicine in the general population is unclear. Suppose, for example, that sequencing becomes sufficiently inexpensive that all individuals, at birth, could have their genomes sequenced at negligible cost. What fraction of the population would benefit from such sequencing? "Benefit" in this context is defined as receiving information indicating that the risk of disease is increased or decreased to a degree that would alter an individual's lifestyle or medical management.

On the surface, it might seem impossible to answer this question at present, as there are millions of genetic variants in every individual and the contribution of nearly all of these variants to any disease is unknown. However, there is one group of individuals in which this question can be immediately addressed: monozygotic twin pairs. If one twin of the pair has a disease, then the probability of the other twin developing that disease is dependent on the genome whenever that disease has some genetic component. We show below that when this logic is applied to a large numbers of twins, estimates of the potential benefits of genome-wide sequencing in the general (non-twin) population can be made.

RESULTS

Conceptual basis

The key to our analysis is the concept of a "genometype". We do not know the genomic sequences of the twin pairs analyzed in the studies described herein, but we do know that each twin pair shares a nearly identical genome (*10*) and that a genome confers a particular genetic risk to every disease.

For each disease, we group genomes that confer identical genetic risks into genometypes. For example, genometypes could be grouped into 20 bins, with genometypes in bin 1 conferring zero genetic risk, genometypes in bin 2 conferring 3% genetic risk, genometypes in bin 3 conferring 10% genetic risk, etc. We can then estimate what distributions of genometypes in the population best reflect the observed monozygotic twin concordancy and discordancy for any given disease.

In twin studies on diseases, heritability (defined in Box 1) is generally based on the difference in the incidence of a disease in monozygotic *versus* dizygotic twins (*11*, *12*). Heritability reflects the average genetic contribution to disease in a twin population. We are interested in the distribution of genetic risks rather than the average. For example, a 30% average risk could reflect a small fraction of twin-pairs with genometypes conferring high genetic risk or a larger fraction of twin-pairs with genometypes conferring a moderate genetic risk. Among all the distributions of genometypes that are compatible with the twin epidemiologic data, we wished to find the distributions that maximized or minimized the potential clinical utility of identifying those genometypes by genomic sequencing.

Whole-genome sequencing-based tests, like any genetic test, can be informative in two ways: negative and positive tests would indicate a substantially lower or higher risk, respectively, than that of the general population. The challenge is to define "substantially" in clinically meaningful and quantitative terms. An example might help put this challenge into perspective. Suppose a woman receives a whole-genome test result indicating that she has a 90% lifetime risk (the total risk over her entire life) of developing breast cancer. She may decide to have a prophylactic double mastectomy to prevent this outcome. Similarly, if the test indicated an 80% or even a 50% lifetime risk of developing breast cancer, she may consider mastectomy. On the other hand, if the test indicated only a 14% risk of developing breast cancer, then mastectomies would be considered by very few women, given that most women today do not opt for prophylactic mastectomies even though the lifetime risk of developing breast cancer in the general population is 12%.

This example illustrates that the risk threshold required for clinical utility represents a balance between the risk reduction afforded by an intervention and its negative consequences. A precedent exists for defining this threshold, in that the decision to implement genetic tests is often based on a positive predictive value (PPV) of at least 10%, implying that more than 1 in 10 patients with a positive test result are expected to develop disease (*13*). While the choice of this threshold will depend on the specific intervention and should ideally be left to the individual, we use this 10% threshold for our population-level analyses of 20 of the 24 diseases analyzed (table S1). In the other four diseases (chronic fatigue syndrome, gastro-esophageal reflux disorder, coronary heart disease-related death and general dystocia), which occur at relatively high frequency in the population, this 10% threshold is inadequate to distinguish individuals with a significantly increased genetic risk from the rest of the population. For these four diseases (table S1), a more appropriate threshold corresponds to one conferring a genetic risk that is at least as great as that of the non-genetic component. Individuals with genometypes conferring this degree of genetic risk would therefore have a total risk at least-twice as large as those without any genetic predisposing factors. This 2x threshold in relative risk is similar to those widely used as clinical benchmarks for common diseases (*14–18*).

For whole-genome testing in healthy individuals, we thereby defined a threshold at which a positive test result would be clinically meaningful as follows. If the non-genetic risk was <5%, then the threshold was set at 10%. If the non-genetic risk was >5%, then the threshold was set at 2x the non-genetic contribution. Though we have used these particular thresholds in most of the examples described below, we also describe how these results varied when other thresholds were considered.

Twin data

We collated monozygotic twin pair data from the Swedish Twin Registry, Danish Twin Registry, Finnish Twin Cohort, Norwegian National Birth Registry and the National Academy of Science – National Research World War II Veteran Twins Registry (*19–31*) (Table 1). From these registries, we selected data representing 24 diseases of diverse etiologies including autoimmune diseases, cancer, cardiovascular diseases, genitourinary diseases, neurological diseases and obesity-associated diseases. Three of these conditions (coronary heart disease, cancer and stroke) represent the leading causes of mortality in the United States, accounted for 54.2% of total deaths in 2007, and are therefore of major public health importance (*32*). The thresholds for a clinically meaningful test result, as defined above, were calculated from disease prevalence and non-genetic risks in the populations from which the twins were drawn (*19–31*) (Materials and Methods, Table 1 and table S2).

Mathematical model

We then developed computational methods to evaluate possible frequency (*f*) and genetic risk (*r*) combinations for a population containing 20 genometypes. Genometype frequency is defined as the proportion of twin pairs in the population that have a given genometype (Box 1). Genometype genetic risk is defined, for each disease, as the absolute increment in risk that an individual with that

genometype will face compared to someone with no genetic risk at all (Box 1). For any combination of genometypes, each with a certain frequency and genetic risk, we obtain an expected distribution of disease-affected individuals among a monozygotic twin cohort. Many different combinations of genometype frequencies and genetic risks match the observed distributions in monozygotic twins; we are interested in those combinations (distributions) that maximize or minimize clinical utility, thus putting bounds on the expectations from whole-genome sequencing. The mathematical framework for our study, and associated statistical and technical issues, are detailed in the Material and Methods.

Clinical implications

These analyses allowed us to address various measures of potential clinical utility. First, for each disease, what is the maximum and minimum fraction of patients with the disease that would receive a positive test, i.e., a result indicating that they have a substantially increased risk of that disease? The answers to this question are graphically shown in Fig. 1 for each of the 24 diseases (for three diseases, we present different answers for males and females, resulting in a total of 27 disease categories). As can be seen from Fig. 1, the fraction of patients that would receive a positive test varies widely from disease to disease. The majority of patients (>50%) who would ultimately develop 13 of the 27 disease categories would *not* test positive, even in the best-case scenario. On the other hand, there were four disease categories - thyroid autoimmunity, type I diabetes, Alzheimer's disease, and coronary heart disease-related deaths in males - for which genetic tests might identify more than 75% of the patients who ultimately develop the disease. Genometype risk and frequency distributions for all diseases are shown in table S3 and graphically for representative diseases in fig. S1.

We could also determine the maximum and minimum fraction of individuals in the population (rather than the fraction of patients with disease) who would receive positive test results for each disease. As shown in Fig. 2, this fraction is generally small, as expected, because the incidence of most diseases is relatively low. Do these negative tests, which would be received by the great majority of individuals for most diseases, have value? Negative tests could be valuable to individual patients if they indicated a considerably lower total risk than would be assumed in the absence of testing. As can be seen from Fig. 3, though, negative tests are generally not very informative in the case of whole-genome sequencing as they are limited by the non-genetic compoment of risk. For 22 of the 27 disease categories studied, a negative test would not indicate a risk that is less than half that in the general population, even in the best-case scenario. This level of risk reduction is probably not sufficient to warrant changes of behavior, lifestyle, or preventative medical practices for these individuals (*33*, *34*). On the other hand, there was one disease category (Alzheimer's disease, Fig. 3) in which a negative test result might indicate as little as a ~12% relative risk of disease compared to the entire twin cohort, at least in the best-case scenario. Knowledge of such a reduced risk might be comforting and relieve anxiety, particularly to those with a family history of Alzheimer's disease.

What is the maximum fraction of individuals that could receive at least one positive test result, i.e., a report indicating that s/he is at risk for at least one of the 24 diseases assessed? From the data depicted in Fig. 2, we estimate that >95% of men and >90% of women could receive at least one positive test result if the risk alleles were actually distributed in the way that produced maximal sensitivity in our model. We assumed that the risk alleles for these 24 diseases were independent in these estimates; if they were not independent, then these figures represent overestimates. On the other hand, these frequencies may represent underestimates as there are a number of additional diseases with hereditary components that have not yet been studied in monozygotic twins or included in our analyses. At the very least, if we consider only distinct disease categories whose pathogenesis is unlikely to be shared, our analyses suggest that, in the best-case scenario, the majority of tested individuals might be alerted to a clinically meaningful risk by whole-genome sequencing.

It was of interest to determine how the results described above varied with the threshold chosen for the analysis. For example, it might be argued that a threshold of 10% was too low for true clinical utility. Our analyses show that the maximum fraction of affected cases testing positive, as well as the maximum fraction of the total population that tests positive, is not changed much when the thresholds are changed to 20% (tables S4 and table S5). With very high thresholds, however, both these measures of sensitivity decrease significantly (table S4 and table S5). Moreover, the maximum predictive value of a negative test drops precipitously at higher thresholds (table S6).

DISCUSSION

The general public does not appear to be aware that, despite their very similar height and appearance, monozygotic twins in general do not always develop or die from the same maladies (*35*, *36*). This basic observation, that monozygotic twins of a pair are not always afflicted by the same maladies, combined with extensive epidemiologic studies of twins and statistical modeling, allows us to estimate upper- and lower-bounds of the predictive value of whole-genome sequencing.

On the negative side, our results show that the majority of tested individuals would receive negative tests for most diseases (Fig. 2). Moreover, the predictive value of these

negative tests would generally be small, as the total risk for acquiring the disease in an individual testing negative would be similar to that of the general population (Fig. 3). On the positive side, our results show that, at least in the best-case scenario, the majority of patients might be alerted to a clinically meaningful risk for at least one disease through whole-genome sequencing.

These conclusions are consistent with what is now known about risk allele loci from genome-wide association studies (GWAS) (*37*). In general, GWAS have shown that many loci can predispose to disease and that each risk allele confers a relatively small effect (*38*, *39*). For example, a recent analysis of large cohorts of individuals with colorectal cancer showed that only ~1.3% of phenotypic variance could be accounted for by the 10 loci discovered through GWAS (*40*). However, it could be argued that the relatively low level of utility that might be inferred from such studies is misleading. In particular, it is possible that a more complete knowledge of disease-associated variants and their epistatic relationships would be able to reliably predict who will and who will not develop disease in the general population. Our results allow us to estimate the *maximum possible* reliability of such tests.

Several of our conclusions are based on the genometype frequency and risk distributions that would *maximize* the clinical utility of genetic testing, i.e., are best-case scenarios. The actual frequency and risk distributions of genometypes in the population are not likely to be distributed in this way. Indeed, other distributions are also consistent with the monozygotic twin data on which our maxima are determined and all other distributions yield less clinical utility than those of the maxima, as shown in Figs. 1 to 3. Moreover, in the real world, it is unlikely that the biomedical correlates of every genetic variant and the epistatic relationships among these variants will ever be completely known, or that the analytic validity of genetic testing will be perfect - as we assume in our ideal scenario. Thus, our conclusions purposely overestimate the value of whole-genome sequencing that will be achieved - they represent an absolute upper bound that cannot be improved by improvements in technology or genetic knowledge. As a practical example of this principle, we estimate that a negative whole-genome sequencing-based test *could* indicate a ~ two-fold decrease in risk for prostate cancer in men and a similar two-fold decrease for urinary incontinence in women. But this two-fold decrease would only apply in a world in which the risk alleles are distributed in a fashion that maximizes the sensitivity of whole genome testing (Fig. 3). In the real world, the risk alleles are not likely to be distributed in this ideal fashion, and omniscience about every variant is not likely to be realized. Thus, the risk of these diseases in patients who test negative will likely be even more similar to that of the general population. For diseases with a lower heritable component, such as most forms of cancer, whole-genome based genetic tests will be even less informative. Thus, our results suggest that genetic testing, at its best, will not be the dominant determinant of patient care and will not be a substitute for preventative medicine strategies incorporating routine checkups and risk management based on the history, physical status and life style of the patient.

It is important to point out that our study focused on testing relatively common diseases in the general population and did not address the utility of whole-genome sequencing to identify the genetic basis of rare monogenic diseases. In such unusual cases, it has already been shown that whole-genome sequencing can prove highly informative (*8*, *9*).

As with any model-based study, our conclusions have a number of caveats. Our analyses are based on data from twin studies and the assumptions made therein (*11*). Specifically, we do not model gene-environment interactions and rely on the prevalence of disease in the twin cohorts; this prevalence, as well as the operative non-genetic contributions, may differ from that in the general population. Though twins are likely to be representative of the general population, the estimates provided by our model could be improved through analyses of larger twin cohorts as these become available, as well as through a more complete phenotypic evaluation of twins of varying ethnicities. Another caveat is that our conclusions about potential utility are based on thresholds that represent a complex balance of personal choices, demographic influences, disease characteristics and the clinical intervention(s) available. We have used a minimum 10% total risk and a minimum relative risk of 2 as the threshold in our analyses. Other thresholds may be more appropriate and meaningful for given situations, though the data in table S4 to table S6 show that our major conclusions are not altered much by the choice of threshold.

In sum, no result, including ours, can or should be used to conclude that whole-genome sequencing will be either useful or useless in an absolute sense. This utility will depend on the results of testing, the individual tested, and the perspectives of individuals and societies. What we hoped to accomplish with this study is to put the debate about the value of such sequencing in a mathematical framework so that the potential merits and limitations of whole-genome sequencing, for any disease, can be quantitatively assessed. Recognition of these merits and limits can be useful to consumers, researchers, and industry, as they can minimize unrealistic expectations and foster fruitful investigations.

MATERIALS AND METHODS

Twin studies used for genometype analyses

We used data from twin studies arising from population-based twin registries to investigate the distribution of disease risk within the population (*19*–*31*). The registries in our study

included the Swedish Twin Registry, Danish Twin Registry, Finnish Twin Cohort, Norwegian National Birth Registry and the National Academy of Science – National Research Council World War II Veteran Twins Registry. Traits were chosen that represented diverse etiologies or were conditions of significant public health importance. We evaluated diseases in the following categories: autoimmune (T1D, thyroid auto-antibodies), neoplastic (breast, colorectal and prostate cancer), cerebrovascular (coronary heart disease-related death and stroke-related death), genitourinary (general dystocia, pelvic organ prolapse, and urinary incontinence), unknown etiology (irritable bowel syndrome, chronic fatigue), neurological (Parkinson disease, Alzheimer's disease and dementia) and obesity-associated (T2D, gallstone disease).

To be included in our analyses, the following data had to be available for each twin study:

1. n_t – total number of monozygotic (MZ) twin pairs where the disease status of each twin was known.
2. n_c – number of disease-concordant MZ twin pairs.
3. n_d – number of disease-discordant MZ twin pairs.
4. n_h – number of healthy-concordant MZ pairs.
5. Heritability (*HER*) – calculated as the proportion of the polygenic liability variation associated with genetic factors.

Using the data from population-based twin studies, we define cohort risk (CR) - the fraction of people in the cohort that had the disease - as follows:

$$CR = (2n_c + n_d) / (2n_t) \quad (1)$$

Model of the predictive capacity of personal genome sequencing

We define the following generative model that characterizes the joint distribution of an individual having a pre-specified disease and a particular genometype. Each individual is characterized by: (i) a binary (Bernoulli) random variable, Z, specifying whether or not s/he has the disease, and (ii) a categorical random variable, G, indicating the genometype of the individual. This means that of the d assumed extant genometypes, each individual can have only one of them. The joint distribution of both the disease and genometype for an individual is given by $P(Z,G)$. This joint distribution decomposes into a product of the likelihood of getting the disease given the genometype, $P(Z \mid G)$, and the prior probability of having the genometype, $P(G)$

$$P(Z,G) = P(Z|G)P(G) \quad (2)$$

Thus, to proceed, we specify both the likelihood function, $P(Z \mid G)$, and the prior, $P(G)$. As mentioned above, G is a categorical random variable taking values $g_1, g_2, \ldots, g_d$, each of which with some probability. Therefore we have:

$$P(G = g_i) = f_i \quad (3)$$

for all $i = 1,2,\ldots,d$. In words, a person can have one of the d assumed extant genometypes, and the probability of having genometype i is given by f_i.

The probability of having the disease given a genometype is $q_i = P(Z = 1 \mid G = g_i)$. Assume that q_i is a sum of a non-genetic risk, e, which is assumed to be constant for the whole population, and genetic risk, r_i, that is, $q_i = e + r_i$ (note that $0 \le q_i \le 1$). Non-genetic risk (e) is the proportion of people in the population that would get the disease if all had the most favorable genometype possible. Non-genetic risk includes all factors that are not inherited, including environmental exposures (e.g., diet, carcinogens), epigenetic alterations and stochastic influences. We estimated it as: $e = CR(1 - HER)$ (see below). This model assumes that all risks are either non-genetic or genetic, i.e., no interactions. We require that the unknown parameters, r_i, must be between 0 and 1 - e, for all i. Therefore, for a given genometype, the likelihood term for genometype i is given by:

$$P(Z|G = g_i) = \begin{cases} e + r_i, & \text{if } z = 1, \\ 1 - e - r_i, & \text{if } z = 0. \end{cases} \quad (4)$$

Thus, the joint distribution of disease and genometype can be written as:

$$P(Z = z, G = g_i) = f_i (e + r_i)^z (1 - e - r_i)^{1-z}, \quad z \in \{0,1\}, g \in \{g_1, \ldots, g_d\}. \quad (5)$$

If the available data included the genometype and disease status of each individual, then inferring estimates of the parameters, $\boldsymbol{r} = (r_1, \ldots, r_d)$, and $\boldsymbol{f} = (f_1, \ldots, f_d)$, would be relatively straightforward. However, the available data include only the disease status of monozygotic twins. These represent observations of disease status in two individuals with identical genometypes. Therefore, we can describe a joint distribution for monozygotic twins having a disease or not. Let $Z_j = Z(X_j)$ be the Bernoulli random variable indicating whether a particular individual has disease and let $Z_k = Z(X_k)$ be the Bernoulli random variable for the co-twin. Similarly, let $G_j = G(X_j)$ and $G_k = G(X_k)$ be categorical random variables indicating whether twin j or k of a pair has some particular genometype. The distribution of disease within monozygotic twins can be divided into three distinct groups, namely: disease concordant, discordant, and healthy concordant pairs.

The probability of disease concordant monozygotic twins is given by:

$$P(Z_j = Z_k = 1 | G_j = G_k) = \sum_i P(Z_j = Z_k = 1 | G_j = G_k = g_i) P(G_j = G_k = g_i), \quad (6a)$$

$$= \sum_i P\left(Z_j = 1 \middle| G_j = g_i\right) P\left(Z_k = 1 \middle| G_k = g_i\right) P\left(G_j = G_k = g_i\right) \tag{6b}$$

$$= \sum_i \left(e + r_i\right)^2 f_i. \tag{6c}$$

Similarly, the probability of healthy concordant monozygotic twin pairs is given by:

$$P\left(Z_j = Z_k = 0 \middle| G_j = G_k\right) = \sum_i P\left(Z_j = Z_k = 0 \middle| G_j = G_k = g_i\right) P\left(G_j = G_k = g_i\right), \tag{7a}$$

$$= \sum_i \left(1 - e - r_i\right)^2 f_i. \tag{7b}$$

And the probability of monozygotic twin pairs discordant for disease is given by:

$$P\left(Z_j \neq Z_k \middle| G_j = G_k\right) = 2\sum_i \left(e + r_i\right)\left(1 - e - r_i\right) f_i. \tag{8}$$

Optimization

For each disease, let n_c, n_h and n_d correspond to the number of concordant disease, healthy and discordant twin pairs. Assuming that there are d genometypes, the expected number of twin pairs of each of the three types is simply the total number of twin pairs times the probability of being each kind of twin pair:

$$\mathrm{E}\left[n_c\right] = n_t \sum_{i \in [d]} \left(e + r_i\right)^2 f_i, \tag{9}$$

$$\mathrm{E}\left[n_h\right] = n_t \sum_{i \in [d]} \left(1 - e - r_i\right)^2 f_i, \tag{10}$$

$$\mathrm{E}\left[n_d\right] = n_t \sum_{i \in [d]} 2\left(e + r_i\right)\left(1 - e - r_i\right) f_i. \tag{11}$$

Because we are interested in the limits of utility of genetic testing, we search for a parameter set that maximizes or minimizes the fraction of patients that will receive a positive test result, given certain constraints. Formally, we define the positive fraction (*PF*) as the proportion, among twin pairs with at least one disease case, that possess a genometype sufficient to change clinical action. In our notation:

$$PF\left(t, e; f, p\right) = \frac{\sum_{i \in [d] | r_i > t} f_i \left[\left(e + r_i\right)^2 + \left(e + r_i\right)\left(1 - e - r_i\right)\right]}{\sum_{i \in [d]} f_i \left[\left(e + r_i\right)^2 + \left(e + r_i\right)\left(1 - e - r_i\right)\right]} \tag{12}$$

where t is the genetic risk required for a person to be at the threshold required for clinical utility and d is the maximum number of genometypes under consideration. The thresholds for each disease are provided in table S2, and for each disease, t is defined as this threshold minus e.

We therefore seek to solve the following optimization problem, for each disease:

$$\underset{f, r}{\text{maximize}} \quad PF\left(t, e; f, p\right), \tag{13}$$

$$\text{subject to} \quad f_i \geq 0, \sum_i f_i = 1, r_i \in (0,1), \sum_{x \in \{c,h,d\}} \left(\hat{n}_x - \mathrm{E}\left[n_x\right]\right)^2 \leq 0.25, \tag{14}$$

where Eq. (14) enforces that none of the residual errors can be larger than 0.5. The parameter n_x is the estimated number of twin pairs of each type obtained by plugging the estimated parameters into Eqs. (9) – (11). This is therefore a quadratically constrained nonlinear optimization problem. We utilize the following algorithm to obtain a local optimum.

For $d' = 2$, i.e., starting with $d' = 2$ genometypes, we implement a grid search over the parameter space and select the parameters that maximize the likelihood over a constrained search space. Let $\theta = (f, r)$ and Θ be the set of all θ's under consideration, as defined by the feasible region specified in Eq. (14). We then discretize this space into nine bins for each element of f and 100 bins for each element of r and denote $P(Z|G)$ by $P_\theta(Z|G)$ to emphasize the dependence of the joint distribution on the parameter. Thus, we aim to solve the following optimization problem:

$$\hat{\theta}^{(2)} = \underset{\theta \in \Theta}{\operatorname{argmax}} \prod_{i,j} P_\theta\left(Z_j, Z_k | G_j = G_k\right) \tag{15}$$

where $\hat{\theta}^{(d')} = \left(\hat{f}^{(d')}, \hat{r}^{(d')}\right)$ is the parameter estimate assuming only d' genometypes. For each $d' = 3, \ldots, 20$, we seek to solve the above optimization problem. To initialize, we pad the previous solution with zeros, yielding $\hat{f}_{(0)}^{(d'+1)} = \left(\hat{f}^{(d')}, 0\right)$ and similarly for $\hat{r}_{(0)}^{(d'+1)}$. Then we use MATLAB's *fmincon* to find a local maximum of *PF* given the constraints. If no improvement in *PF* is obtained for $d'+1$ genometypes using the default "padded" initialization, we try randomly initializing. We stop trying random initializations if any of the following criteria are met: (i) if we find an improvement in *PF* with the constraints satisfied, (ii) if we reach 100% *PF*, or (iii) if we reach 15 random initializations. If criterion (i) is met, we denote the parameters achieving the improvement $\hat{\theta}^{(d'+1)}$ and then increment d' and continue. If criterion (ii) is met, we stop incrementing d', as we have achieved the maximum possible *PF*, so adding additional genometypes cannot possibly maximize it further. If criterion (iii) is met, we let $\hat{\theta}^{(d'+1)} = \hat{\theta}_{(0)}^{(d'+1)}$; that is, we let our final estimate for $d'+1$ simply be our estimate for d' padded with a zero. We then increment d'.

We repeat the above approach for each disease. The parameters that we determined using this approach to maximize *PF* were then used to estimate the percentage of the population testing positive for a given disease, as well as the relative risk of disease for those individuals testing negative, as defined below. We apply this approach separately for each disease, thus assuming independence. To find the minimum PFs compatible with the twin data, we used a simiilar procedure.

Relative risk of disease if testing negative

We determined the relative risk of disease of individuals whose whole-genome sequencing tests were negative after maximizing or minimizing the sensitivity (*PF*) of the test. Disease risk in the population testing negative (DR_{neg}) is the ratio of the number of disease cases testing negative to the number of individuals in the population testing negative:

$$DR_{neg} = \frac{(2n_c + n_d)(1 - PF)}{2n_t \sum_{i \in [d] | r_i < t} f_i} \quad (16)$$

To determine the relative risk of disease if testing negative (RR_{neg}), we calculated the ratio of disease risk of individuals testing negative to the disease risk in the twin cohort (*CR*):

$$RR_{neg} = \frac{DR_{neg}}{CR} \quad (17)$$

Calculation of relative risks

We defined relative risk (*RR*) in table S2 as the minimum total risk of individuals with genometypes carrying a given genetic risk compared to the total risk of individuals with genometypes carrying a genetic risk of 0% (i.e., determined solely by non-genetic factors). The minimum total risk was determined using the standard 10% risk threshold described in the text as well as others (tables S4 to S6). In all cases,

$$RR = \frac{PPV + (CR(1 - HER))}{CR(1 - HER)} \quad (18)$$

Other parameters and models

Equation (14) enforces that none of the residual errors can be larger than 0.5, such that upon rounding we obtain a perfect fit. Changing this parameter from 0.5 to 0.01 did not alter the *PF*'s depicted in Fig. 1 for any disease.

Instead of maximizing *PF*'s, we also determined the distributions of genometype risks (r_i) and frequencies (f_i) that would minimize the relative risk of disease of individuals whose whole-genome sequencing tests were negative. This independent optimization yielded results nearly identical to those reported in Fig. 1, Fig. 2, and Fig. 3.

As noted above, we estimated the non-genetic risk as $e = CR(1 - HER)$. This risk is somewhat higher than that derived from the standard liability threshold (LT) model. However, it has recently been shown that the LT model underestimates the non-genetic contribution to disease because it does not take into account synergistic interactions among genes (*41*). The model described herein does not make any assumptions about the nature of the interactions between genes, such as additivity. However, the LT model can also be used to approximate the maximum capacity of whole genome sequencing to detect individuals at pre-defined risks under certain simplifying assumptions about the distribution of risk alleles in the population. The *PF* predictions from the LT model employing 10% thresholds are provided in table S4 and can be compared to the results of the current model with 10% thresholds (table S4).

Finally, our model can be used to calculate the potential clinical utility of whole-genome sequencing under any assumption about the proportion of non-genetic contributions to disease risk, or estimates thereof. Representative values for each disease, with non-genetic contributions ranging from 10% to 90%, are provided in table S7.

SUPPLEMENTARY MATERIALS

www.sciencetranslationalmedicine.org/cgi/content/full/10.1126/scitranslmed.3003380/DC1

Figure S1. Graphical representation of genometype frequency and risk distributions for A. leukemia, B. Alzheimer's disease, and C. pancreatic cancer.

Table S1. Examples of known risk factors for common human diseases.

Table S2. Thresholds and other parameters used to analyze each disease.

Table S3. The risks and frequencies of each of the 20 genometypes providing maximum sensitivity (PF) for detection of each disease.

Table S4. Percentage of cases (i.e., individuals with disease) testing positive with whole-genome sequencing at varying risk thresholds, or with the liability threshold (LT) model employing a 10% threshold.

Table S5. Percentage of population testing positive with whole-genome sequencing at varying risk thresholds.

Table S6. Relative risk of disease if testing negative with whole-genome sequencing at varying risk thresholds.

Table S7. Percentage of cases testing positive with whole-genome sequencing at varying estimates of non-genetic contributions.

REFERENCES AND NOTES

1. D. R. Bentley, S. Balasubramanian, H. P. Swerdlow, G. P. Smith, J. Milton, C. G. Brown, K. P. Hall, D. J. Evers, C. L. Barnes, H. R. Bignell, J. M. Boutell, J. Bryant, R. J. Carter, R. Keira Cheetham, A. J. Cox, D. J. Ellis, M. R. Flatbush, N. A. Gormley, S. J. Humphray, L. J. Irving, M. S. Karbelashvili, S. M. Kirk, H. Li, X. Liu, K. S. Maisinger, L. J. Murray, B. Obradovic, T. Ost, M. L. Parkinson, M. R. Pratt, I. M. Rasolonjatovo, M. T. Reed, R. Rigatti, C. Rodighiero, M. T. Ross, A. Sabot, S. V. Sankar, A. Scally, G. P. Schroth, M. E. Smith, V. P. Smith, A. Spiridou, P. E. Torrance, S. S. Tzonev, E. H. Vermaas, K. Walter, X. Wu, L. Zhang, M. D. Alam, C. Anastasi, I. C. Aniebo, D. M. Bailey, I. R. Bancarz, S. Banerjee, S. G. Barbour, P. A. Baybayan, V. A. Benoit, K. F. Benson, C. Bevis, P. J. Black, A. Boodhun, J. S. Brennan, J. A. Bridgham, R. C. Brown, A. A. Brown, D.

H. Buermann, A. A. Bundu, J. C. Burrows, N. P. Carter, N. Castillo, E. C. M. Chiara, S. Chang, R. Neil Cooley, N. R. Crake, O. O. Dada, K. D. Diakoumakos, B. Dominguez-Fernandez, D. J. Earnshaw, U. C. Egbujor, D. W. Elmore, S. S. Etchin, M. R. Ewan, M. Fedurco, L. J. Fraser, K. V. Fuentes Fajardo, W. Scott Furey, D. George, K. J. Gietzen, C. P. Goddard, G. S. Golda, P. A. Granieri, D. E. Green, D. L. Gustafson, N. F. Hansen, K. Harnish, C. D. Haudenschild, N. I. Heyer, M. M. Hims, J. T. Ho, A. M. Horgan, K. Hoschler, S. Hurwitz, D. V. Ivanov, M. Q. Johnson, T. James, T. A. Huw Jones, G. D. Kang, T. H. Kerelska, A. D. Kersey, I. Khrebtukova, A. P. Kindwall, Z. Kingsbury, P. I. Kokko-Gonzales, A. Kumar, M. A. Laurent, C. T. Lawley, S. E. Lee, X. Lee, A. K. Liao, J. A. Loch, M. Lok, S. Luo, R. M. Mammen, J. W. Martin, P. G. McCauley, P. McNitt, P. Mehta, K. W. Moon, J. W. Mullens, T. Newington, Z. Ning, B. Ling Ng, S. M. Novo, M. J. O'Neill, M. A. Osborne, A. Osnowski, O. Ostadan, L. L. Paraschos, L. Pickering, A. C. Pike, D. Chris Pinkard, D. P. Pliskin, J. Podhasky, V. J. Quijano, C. Raczy, V. H. Rae, S. R. Rawlings, A. Chiva Rodriguez, P. M. Roe, J. Rogers, M. C. Rogert Bacigalupo, N. Romanov, A. Romieu, R. K. Roth, N. J. Rourke, S. T. Ruediger, E. Rusman, R. M. Sanches-Kuiper, M. R. Schenker, J. M. Seoane, R. J. Shaw, M. K. Shiver, S. W. Short, N. L. Sizto, J. P. Sluis, M. A. Smith, J. Ernest Sohna Sohna, E. J. Spence, K. Stevens, N. Sutton, L. Szajkowski, C. L. Tregidgo, G. Turcatti, S. Vandevondele, Y. Verhovsky, S. M. Virk, S. Wakelin, G. C. Walcott, J. Wang, G. J. Worsley, J. Yan, L. Yau, M. Zuerlein, J. C. Mullikin, M. E. Hurles, N. J. McCooke, J. S. West, F. L. Oaks, P. L. Lundberg, D. Klenerman, R. Durbin, A. J. Smith, Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**, 53-59 (2008).

2. R. Drmanac, A. B. Sparks, M. J. Callow, A. L. Halpern, N. L. Burns, B. G. Kermani, P. Carnevali, I. Nazarenko, G. B. Nilsen, G. Yeung, F. Dahl, A. Fernandez, B. Staker, K. P. Pant, J. Baccash, A. P. Borcherding, A. Brownley, R. Cedeno, L. Chen, D. Chernikoff, A. Cheung, R. Chirita, B. Curson, J. C. Ebert, C. R. Hacker, R. Hartlage, B. Hauser, S. Huang, Y. Jiang, V. Karpinchyk, M. Koenig, C. Kong, T. Landers, C. Le, J. Liu, C. E. McBride, M. Morenzoni, R. E. Morey, K. Mutch, H. Perazich, K. Perry, B. A. Peters, J. Peterson, C. L. Pethiyagoda, K. Pothuraju, C. Richter, A. M. Rosenbaum, S. Roy, J. Shafto, U. Sharanhovich, K. W. Shannon, C. G. Sheppy, M. Sun, J. V. Thakuria, A. Tran, D. Vu, A. W. Zaranek, X. Wu, S. Drmanac, A. R. Oliphant, W. C. Banyai, B. Martin, D. G. Ballinger, G. M. Church, C. A. Reid, Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. *Science* **327**, 78-81 (2010).

3. J. Eid, A. Fehr, J. Gray, K. Luong, J. Lyle, G. Otto, P. Peluso, D. Rank, P. Baybayan, B. Bettman, A. Bibillo, K. Bjornson, B. Chaudhuri, F. Christians, R. Cicero, S. Clark, R. Dalal, A. Dewinter, J. Dixon, M. Foquet, A. Gaertner, P. Hardenbol, C. Heiner, K. Hester, D. Holden, G. Kearns, X. Kong, R. Kuse, Y. Lacroix, S. Lin, P. Lundquist, C. Ma, P. Marks, M. Maxham, D. Murphy, I. Park, T. Pham, M. Phillips, J. Roy, R. Sebra, G. Shen, J. Sorenson, A. Tomaney, K. Travers, M. Trulson, J. Vieceli, J. Wegener, D. Wu, A. Yang, D. Zaccarin, P. Zhao, F. Zhong, J. Korlach, S. Turner, Real-time DNA sequencing from single polymerase molecules. *Science* **323**, 133-138 (2009).

4. M. Margulies, M. Egholm, W. E. Altman, S. Attiya, J. S. Bader, L. A. Bemben, J. Berka, M. S. Braverman, Y. J. Chen, Z. Chen, S. B. Dewell, L. Du, J. M. Fierro, X. V. Gomes, B. C. Godwin, W. He, S. Helgesen, C. H. Ho, G. P. Irzyk, S. C. Jando, M. L. Alenquer, T. P. Jarvie, K. B. Jirage, J. B. Kim, J. R. Knight, J. R. Lanza, J. H. Leamon, S. M. Lefkowitz, M. Lei, J. Li, K. L. Lohman, H. Lu, V. B. Makhijani, K. E. McDade, M. P. McKenna, E. W. Myers, E. Nickerson, J. R. Nobile, R. Plant, B. P. Puc, M. T. Ronan, G. T. Roth, G. J. Sarkis, J. F. Simons, J. W. Simpson, M. Srinivasan, K. R. Tartaro, A. Tomasz, K. A. Vogt, G. A. Volkmer, S. H. Wang, Y. Wang, M. P. Weiner, P. Yu, R. F. Begley, J. M. Rothberg, Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**, 376-380 (2005).

5. K. A. Frazer, S. S. Murray, N. J. Schork, E. J. Topol, Human genetic variation and its contribution to complex traits. *Nat Rev Genet* **10**, 241-251 (2009).

6. E. T. Cirulli, D. B. Goldstein, Uncovering the roles of rare variants in common disease through whole-genome sequencing. *Nat Rev Genet* **11**, 415-425 (2010).

7. T. A. Manolio, Genomewide association studies and assessment of the risk of disease. *N Engl J Med* **363**, 166-176 (2010).

8. S. Jones, R. H. Hruban, M. Kamiyama, M. Borges, X. Zhang, D. W. Parsons, J. C. Lin, E. Palmisano, K. Brune, E. M. Jaffee, C. A. Iacobuzio-Donahue, A. Maitra, G. Parmigiani, S. E. Kern, V. E. Velculescu, K. W. Kinzler, B. Vogelstein, J. R. Eshleman, M. Goggins, A. P. Klein, Exomic sequencing identifies PALB2 as a pancreatic cancer susceptibility gene. *Science* **324**, 217 (2009).

9. S. B. Ng, K. J. Buckingham, C. Lee, A. W. Bigham, H. K. Tabor, K. M. Dent, C. D. Huff, P. T. Shannon, E. W. Jabs, D. A. Nickerson, J. Shendure, M. J. Bamshad, Exome sequencing identifies the cause of a mendelian disorder. *Nat Genet* **42**, 30-35 (2010).

10. C. E. Bruder, A. Piotrowski, A. A. Gijsbers, R. Andersson, S. Erickson, T. Diaz de Stahl, U. Menzel, J. Sandgren, D. von Tell, A. Poplawski, M. Crowley, C.

Crasto, E. C. Partridge, H. Tiwari, D. B. Allison, J. Komorowski, G. J. van Ommen, D. I. Boomsma, N. L. Pedersen, J. T. den Dunnen, K. Wirdefeldt, J. P. Dumanski, Phenotypically concordant and discordant monozygotic twins display different DNA copy-number-variation profiles. *Am J Hum Genet* **82**, 763-771 (2008).
11. F. V. Rijsdijk, P. C. Sham, Analytic approaches to twin data using structural equation models. *Brief Bioinform* **3**, 119-133 (2002).
12. P. M. Visscher, W. G. Hill, N. R. Wray, Heritability in the genomics era--concepts and misconceptions. *Nat Rev Genet* **9**, 255-266 (2008).
13. D. L. Clarke-Pearson, Clinical practice. Screening for ovarian cancer. *N Engl J Med* **361**, 170-177 (2009).
14. P. G. Kopelman, Obesity as a medical problem. *Nature* **404**, 635-643 (2000).
15. W. C. Willett, W. H. Dietz, G. A. Colditz, Guidelines for healthy weight. *N Engl J Med* **341**, 427-434 (1999).
16. A. J. Alberg, J. G. Ford, J. M. Samet, Epidemiology of lung cancer: ACCP evidence-based clinical practice guidelines (2nd edition). *Chest* **132**, 29S-55S (2007).
17. A. Ott, A. J. Slooter, A. Hofman, F. van Harskamp, J. C. Witteman, C. Van Broeckhoven, C. M. van Duijn, M. M. Breteler, Smoking and risk of dementia and Alzheimer's disease in a population-based cohort study: the Rotterdam Study. *Lancet* **351**, 1840-1843 (1998).
18. J. He, L. G. Ogden, L. A. Bazzano, S. Vupputuri, C. Loria, P. K. Whelton, Risk factors for congestive heart failure in US men and women: NHANES I epidemiologic follow-up study. *Arch Intern Med* **161**, 996-1002 (2001).
19. P. Lichtenstein, N. V. Holm, P. K. Verkasalo, A. Iliadou, J. Kaprio, M. Koskenvuo, E. Pukkala, A. Skytthe, K. Hemminki, Environmental and heritable factors in the causation of cancer--analyses of cohorts of twins from Sweden, Denmark, and Finland. *N Engl J Med* **343**, 78-85 (2000).
20. P. S. Hansen, T. H. Brix, I. Iachine, K. O. Kyvik, L. Hegedus, The relative importance of genetic and environmental effects for the early stages of thyroid autoimmunity: a study of healthy Danish twins. *Eur J Endocrinol* **154**, 29-38 (2006).
21. J. Kaprio, J. Tuomilehto, M. Koskenvuo, K. Romanov, A. Reunanen, J. Eriksson, J. Stengard, Y. A. Kesaniemi, Concordance for type 1 (insulin-dependent) and type 2 (non-insulin-dependent) diabetes mellitus in a population-based cohort of twins in Finland. *Diabetologia* **35**, 1060-1067 (1992).
22. D. Katsika, A. Grjibovski, C. Einarsson, F. Lammert, P. Lichtenstein, H. U. Marschall, Genetic and environmental influences on symptomatic gallstone disease: a Swedish study of 43,141 twin pairs. *Hepatology* **41**, 1138-1143 (2005).
23. M. Gatz, N. L. Pedersen, S. Berg, B. Johansson, K. Johansson, J. A. Mortimer, S. F. Posner, M. Viitanen, B. Winblad, A. Ahlbom, Heritability for Alzheimer's disease: the study of dementia in Swedish twins. *J Gerontol A Biol Sci Med Sci* **52**, M117-M125 (1997).
24. C. M. Tanner, R. Ottman, S. M. Goldman, J. Ellenberg, P. Chan, R. Mayeux, J. W. Langston, Parkinson disease in twins: an etiologic study. *JAMA* **281**, 341-346 (1999).
25. P. F. Sullivan, B. Evengard, A. Jacks, N. L. Pedersen, Twin analyses of chronic fatigue in a Swedish national sample. *Psychol Med* **35**, 1327-1336 (2005).
26. A. J. Cameron, J. Lagergren, C. Henriksson, O. Nyren, G. R. Locke, 3rd, N. L. Pedersen, Gastroesophageal reflux disease in monozygotic and dizygotic twins. *Gastroenterology* **122**, 55-59 (2002).
27. M. B. Bengtson, T. Ronning, M. H. Vatn, J. R. Harris, Irritable bowel syndrome in twins: genes and environment. *Gut* **55**, 1754-1759 (2006).
28. S. Zdravkovic, A. Wienke, N. L. Pedersen, M. E. Marenberg, A. I. Yashin, U. De Faire, Heritability of death from coronary heart disease: a 36-year follow-up of 20 966 Swedish twins. *J Intern Med* **252**, 247-254 (2002).
29. S. Bak, D. Gaist, S. H. Sindrup, A. Skytthe, K. Christensen, Genetic liability in stroke: a long-term follow-up study of Danish twins. *Stroke* **33**, 769-774 (2002).
30. M. Algovik, E. Nilsson, S. Cnattingius, P. Lichtenstein, A. Nordenskjold, M. Westgren, Genetic influence on dystocia. *Acta Obstet Gynecol Scand* **83**, 832-837 (2004).
31. D. Altman, M. Forsman, C. Falconer, P. Lichtenstein, Genetic influence on stress urinary incontinence and pelvic organ prolapse. *Eur Urol* **54**, 918-922 (2008).
32. J. Xu, K. D. Kochanek, S. L. Murphy, B. Tejada-Vera, Deaths: final data for 2007. *Natl Vital Stat Rep* **58**, 1-135 (2010).
33. J. Audrain, N. R. Boyd, J. Roth, D. Main, N. F. Caporaso, C. Lerman, Genetic susceptibility testing in smoking-cessation treatment: one-year outcomes of a randomized trial. *Addict Behav* **22**, 741-751 (1997).
34. E. Sabaté, Adherence to long-term therapies: evidence for action, Geneva, *World Health Organiszation*, 288 pages, 2003.
35. A. H. Wong, Gottesman, II, A. Petronis, Phenotypic differences in genetically identical organisms: the epigenetic perspective. *Hum Mol Genet* **14 Spec No 1**, R11-R18 (2005).
36. Identical Twins Not As Identical As Believed. *ScienceDaily*. Accessed on: September 10th 2011. Available from: http://www.sciencedaily.com/releases/2008/02/080215121214.htm.

37. N. R. Wray, M. E. Goddard, P. M. Visscher, Prediction of individual genetic risk to disease from genome-wide association studies. *Genome Res* **17**, 1520-1528 (2007).
38. L. A. Hindorff, P. Sethupathy, H. A. Junkins, E. M. Ramos, J. P. Mehta, F. S. Collins, T. A. Manolio, Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci U S A* **106**, 9362-9367 (2009).
39. N. R. Wray, J. Yang, M. E. Goddard, P. M. Visscher, The genetic interpretation of area under the ROC curve in genomic profiling. *PLoS Genet* **6**, e1000864 (2010).
40. A. Tenesa, M. G. Dunlop, New insights into the aetiology of colorectal cancer from genome-wide association studies. *Nat Rev Genet* **10**, 353-358 (2009).
41. O. Zuk, E. Hechter, S. R. Sunyaev, E. S. Lander, The mystery of missing heritability: Genetic interactions create phantom heritability. *Proc. Natl. Acad. Sci. U.S.A.* **109**, 1193-1198 (2012).

Acknowledgments: We thank Naomi Wray and Donald Geman for critical comments regarding the manuscript, and Katie Kinzler for technical assistance. **Funding:** The project was supported by The Lustgarten Foundation for Pancreatic Cancer Research, The Virginia and D. K. Ludwig Fund for Cancer Research, AACR Stand Up To Cancer-Dream Team Translational Cancer Research Grant, The Dr. Miriam and Sheldon G. Adelson Medical Research Foundation, The European Community's Seventh Framework Programme, NIH grants CA43460, CA57345, CA62924, CA121113, and NCI contract N01-CN-43302. **Author contributions:** N.J.R, J.T.V., G.P., K.W.K, B.V. and V.E.V designed the study; N.J.R, J.T.V. and V.E.V. generated and analyzed data; N.J.R., J.T.V, and B.V. wrote the manuscript. **Competing Interests:** B.V., K.W.K and V.E.V are a co-founders of Inostics and Personal Genome Diagnostics and are members of their Scientific Advisory Boards. K.W.K., B.V., and V.E.V own Inostics and Personal Genome Diagnostics stock, which is subject to certain restrictions under University policy. The terms of these arrangements are managed by the Johns Hopkins University in accordance with its conflict-of-interest policies. G.P. is on the scientific advisory board of Counsyl.

Submitted 25 October 2011
Accepted 2 April 2012
Published 2 April 2012
10.1126/scitranslmed.3003380

Citation: N. J. Roberts, J. T. Vogelstein, G. Parmigiani, K. W. Kinzler, B. Vogelstein, V. E. Velculescu, The Predictive Capacity of Personal Genome Sequencing. *Sci. Transl. Med.* 10.1126/scitranslmed.3003380 (2012).

Fig. 1. The fraction of cases (i.e., patients with disease) who would test positive by whole-genome sequencing. For each disease, the maximum and minimum fraction of cases that would test positive using the thresholds defined in table S2 are plotted.

Fig. 2. Percentage of individuals in the general population who would test positive by whole-genome sequencing. For each disease, the maximum and minimum fraction of individuals in the population that would test positive using the thresholds defined in table S2 are plotted.

Fig. 3. Relative risk of disease in individuals testing negative by whole-genome sequencing. A relative risk of 100% represents the same risk as the general population, i.e., the cohort risk. Relative risks were calculated using the genometype frequencies and genometype genetic risks that maximized or minimized sensitivity for disease detection.

Box 1. Definition of terms

Genometype	A set of genomes that confer a specific genetic risk for a given disease
Genometype genetic risk (r)	The genetic risk conferred by a given genometype
Genometype frequency (f)	The frequency of a given genometype in the general poulation
Threshold	Minimum risk for a given disease considered to be clinically meaningful
Heritability (HER)	Proportion of phenotypic variance associated with genetic factors
Cohort risk (CR)	Risk of disease in the relevant twin cohort
Non-genetic risk (e)	Proportion of cohort risk due to non-genetic factors
Total risk	Sum of genetic risk conferred by a given genometype plus non-genetic risk
Relative risk	Ratio of total risk associated with a given genometype to cohort risk

Table 1. Population-based twin studies used for analysis

Disease/Condition	Sex	Number of MZ Twin Pairs	Number MZ Disease Concordant Pairs	Number MZ Disease Discordant Pairs	Disease Prevalence in Cohort (*CR*)	Reference
Bladder Cancer	Male & Female	15668	5	189	0.6%	Lichtenstein *et al.* [19]
Breast Cancer	Female	8437	42	505	3.5%	Lichtenstein *et al.* [19]
Colorectal Cancer	Male & Female	15668	30	416	1.5%	Lichtenstein *et al.* [19]
Leukemia	Male & Female	15668	2	103	0.3%	Lichtenstein *et al.* [19]
Lung Cancer	Male & Female	15668	18	296	1.1%	Lichtenstein *et al.* [19]
Ovarian Cancer	Female	8437	3	125	0.8%	Lichtenstein *et al.* [19]
Pancreatic Cancer	Male & Female	15668	3	123	0.4%	Lichtenstein *et al.* [19]
Prostate Cancer	Male	7231	40	299	2.6%	Lichtenstein *et al.* [19]
Stomach Cancer	Male & Female	15668	11	223	0.8%	Lichtenstein *et al.* [19]
Thyroid Autoimmunity	Male & Female	284	7	17	5.5%	Hansen *et al.* [20]
Type 1 Diabetes	Male & Female	4307	3	20	0.3%	Kaprio *et al.* [21]
Gallstone Disease	Male & Female	11073	112	956	5.3%	Katsika *et al.* [22]
Type 2 Diabetes	Male & Female	4307	29	113	2.0%	Kaprio *et al.* [21]
Alzheimer's Disease	Male & Female	398	2	8	1.5%	Gatz *et al.* [23]
Dementia	Male & Female	398	3	16	2.8%	Gatz *et al.* [23]
Parkinson Disease	Male & Female	3477	7	60	1.1%	Tanner *et al.* [24]
Chronic Fatigue	Female	1803	133	526	22.0%	Sullivan *et al.* [25]
Chronic Fatigue	Male	1426	48	266	12.7%	Sullivan *et al.* [25]
Gastro Esophageal Reflux Disorder (GERD)	Female	1260	63	284	16.3%	Cameron *et al.* [26]
Gastro Esophageal Reflux Disorder (GERD)	Male	918	32	185	13.6%	Cameron *et al.* [26]
Irritable Bowel Syndrome	Male & Female	1252	14	97	5.0%	Bengtson *et al.* [27]
Coronary heart disease (CHD) Death	Female	2004	97	424	15.4%	Zdravkovic *et al.* [28]
Coronary heart disease (CHD) Death	Male	1640	153	451	23.1%	Zdravkovic *et al.* [28]
Stroke-related Death	Male & Female	3852	35	316	5.0%	Bak *et al.* [29]
General Dystocia	Female	928	40	173	13.6%	Algovik *et al.* [30]
Pelvic Organ Prolapse	Female	3376	34	157	3.3%	Altman *et al.* [31]
Stress Urinary Incontinence	Female	3376	13	87	1.7%	Altman *et al.* [31]

MZ: Monozygotic. Disease prevalence in cohort (cohort risk, *CR*) was determined as described in the Materials and Methods.