The role of functional data in interpreting the effects of genetic variation

David L. Young[a] and Stanley Fields[a,b,c]
[a]Department of Genome Sciences, [b]Department of Medicine, and [c]Howard Hughes Medical Institute, University of Washington, Seattle, WA 98195

ABSTRACT **Progress in DNA-sequencing technologies has provided a catalogue of millions of DNA variants in the human population, but characterization of the functional effects of these variants has lagged far behind. For example, sequencing of tumor samples is driving an urgent need to classify whether or not mutations seen in cancers affect disease progression or treatment effectiveness or instead are benign. Furthermore, mutations can interact with genetic background and with environmental effects. A new approach, termed deep mutational scanning, has enabled the quantitative assessment of the effects of thousands of mutations in a protein. However, this type of experiment is carried out in model organisms, tissue culture, or in vitro; typically addresses only a single biochemical function of a protein; and is generally performed under a single condition. The current challenge lies in using these functional data to generate useful models for the phenotypic consequences of genetic variation in humans.**

Monitoring Editor
David G. Drubin
University of California, Berkeley

Received: Jul 23, 2015
Revised: Aug 27, 2015
Accepted: Aug 28, 2015

Genome sequences are available for many individual humans, but these constitute only one level of data. The blueprints for cellular structure and behavior are not stored solely in the linear sequence of the genome, but embedded within the network of interacting molecular components. Via current sequencing technologies, we have become accomplished in determining genotype. One consequence is that, at the protein level, we can readily establish the differences in protein sequences present in any person or during any disease process. Such a difference may lead to a molecular phenotype, whereby a protein is changed in its activity, stability, localization, or other property. This molecular change can lead, in turn, to a cellular phenotype, whereby the cell may alter how it divides, responds to signals, repairs its DNA, or carries out some other process. Finally, the cellular change can lead to an organismal phenotype, such as cancer. However, the complex hierarchical network of interacting components makes it exceedingly difficult to predict with confidence the effect of a mutation on an organism's phenotype, even though a small change in the genome can have profound effects on development, morphology, behavior, and disease.

The many attempts to computationally predict phenotype from genotype have led to remarkable improvements but have not achieved sufficient accuracy for use in a clinical setting. "Prediction" here is meant to include both the ability to accurately label individual variants as either deleterious or benign and the ability to combine individual variant effects into larger models of phenotypic effect, though most efforts to date have focused on the former type of prediction. These approaches use evolutionary conservation-based metrics (Grantham, 1974; Ng and Henikoff, 2001) or apply methods such as machine learning to combine diverse features of a variant, including conservation, amino acid chemical similarity, and surrounding sequence features into statistical models of variant effect (Adzhubei *et al.*, 2010). These models do not predict risk for a specific disease but instead are trained to label variants as either deleterious or benign by using data either from large databases of individually annotated disease genes (e.g., ClinVar; Landrum *et al.*, 2014), which are expected to be deleterious, or from lists of common variants or lists of fixed but recently derived human alleles, which both are expected to be benign (Kircher *et al.*, 2014). The models are therefore limited by the quality and quantity of the labeled variants used during training and will perform poorly for diseases or for types of detrimental variants that are not well represented in the training data (Li *et al.*, 2014). Most of these models

DOI:10.1091/mbc.E15-03-0153
Address correspondence to: Stanley Fields (fields@uw.edu).
Abbreviations used: HDR, homology-directed DNA repair; SNP, single-nucleotide polymorphism.

also rely heavily on measures of evolutionary constraint that might not be constant at a genetic locus across the related sequences used in the alignment. For example, sites that have only recently in evolutionary history either become important or lost their importance for organismal function will lead to false negatives and false positives, respectively, when using distantly related species or sequences to estimate evolutionary constraint. In addition, the relationship between phenotypic effect and evolutionary constraint is complex, as many phenotypes, including some common diseases, might not have a large effect on reproductive success if they cause morbidity and mortality only later in life. Partly for these reasons, false-positive and false-negative rates are too high to use these predictions in the context of human disease (Dong *et al.*, 2015).

Even though the high standard of clinical utility has not been met, these models still perform well, hinting that a more accurate prediction of phenotype from genotype might be possible. To achieve this goal, we will need to experimentally measure the phenotypic effects of genetic variation in a diversity of cellular and environmental contexts. The resulting large data sets of functional measurements will serve as both input variables and as more dense sets of output variables in the training of statistical models, allowing us to predict more accurately the effect of genetic variation on a wide range of both molecular and organismal phenotypes. The accuracy of genotype-based predictions will always be constrained by the heritability of the predicted trait—the amount of phenotypic variation that is attributable to genotypic variation. For this reason, functional data on genetic variation will be most informative for highly heritable traits, and these are the traits that we should focus on first.

CONSTRUCTING A GENOTYPE–PHENOTYPE MAP BY DEEP MUTATIONAL SCANNING

Genotype–phenotype maps were classically derived from forward genetic screens, in which random mutations introduced into a genome yielded interesting phenotypes. The mutations were localized within the genome, and then the corresponding genes were identified, a process that could take months to years. Reverse genetics, in which defined genes are first mutated and the effects of the mutations are then tested, sped up the process of mapping genotype to phenotype. However, we now know that that there are hundreds to thousands of variants in each gene in humans alive today, with more than 97 million single-nucleotide variants catalogued to date (dbSNP build 144, www.ncbi.nlm.nih.gov/news/06-09-2015-dbsnp-build-144). There is no indication yet of a plateau in the discovery of new variants. Most human variation arose recently and is rare, with 99% of variants present in <1% of individuals (Exome Aggregation Consortium, 2015; Fu *et al.*, 2013). Each person has, on average, around 40–90 de novo mutations (Kong *et al.*, 2012; Besenbacher *et al.*, 2015); thus every possible point mutation in the genome—or at least those compatible with life—may have occurred among the seven billion humans on the planet.

Establishing the contribution of rare variation to disease is difficult with most existing methodologies. An examination of the power of whole-genome resequencing-association studies to identify rare variants of modest effect (1% of phenotypic variance explained) showed that both gene-based and single variant–based methods are significantly underpowered even with tens of thousands of individuals (Moutsianas *et al.*, 2015). Worse, purifying selection keeps frequencies of most detrimental variants low, so these rare variants may be enriched for detrimental effects (Gibson, 2012; Tennessen *et al.*, 2012). Whether rare, large-effect variants or common, moderate-effect variants are more likely to explain the heritability of common diseases depends strongly on the relationship between a variant's phenotypic effect size and its effect on reproductive fitness (Henn *et al.*, 2015). When a disease strongly affects reproductive success, rare, large-effect variants are more likely to be the major cause of the disease (Agarwala *et al.*, 2013).

Even among variants identified by their association with disease, our understanding is poor. The Human Gene Mutation Database, which contains only such variants, is rapidly increasing beyond its current 140,000 mutations, and the vast majority have not been functionally characterized (Stenson *et al.*, 2014). Determining the phenotypic consequences of all these variants one at a time using site-directed mutagenesis is obviously infeasible. However, there is a strong incentive to functionally characterize variants that are inaccessible to association studies, as many of them likely affect disease risk.

The direct functional characterization of genetic variation is achieved through mutagenesis, but mutagenesis experiments have typically been low throughput, sometimes analyzing only a few mutations. These studies seek to uncover more about a protein's function or to characterize mutations observed in the context of human disease. However, for both of these goals, the utility of mutagenesis has been limited by its throughput. It is often unknown which substitutions will be most informative, and there are too many potential mutations in every gene and regulatory element to perform another experiment each time a new mutation is observed in humans. Deep mutational scanning (reviewed in Fowler and Fields, 2014) increases the throughput of mutagenesis studies such that tens of thousands of genotype–phenotype associations can be assayed in a single experiment. For most proteins, this approach can analyze all single–amino acid substitutions and a large number of double substitutions at once. Similar strategies have been developed for DNA regulatory elements and for RNA (Patwardhan *et al.*, 2009)

A deep mutational scan begins with the synthesis of a pool of mutants. The mutations can be introduced by "doped" oligonucleotides, in which random mutations are introduced during gene synthesis, by error-prone PCR, or by synthesis of designed oligonucleotide primers that introduce defined mutations (Fowler and Fields, 2014). The pool of mutations must be introduced into an expression system in which each mutant gene and its encoded protein are physically linked. Both cellular and phage-based expression can meet this requirement. Finally, a selective pressure is applied, unique to the protein function being assayed, such that the activity of the mutant proteins can be differentiated. For example, cells expressing the protein can be dependent on it for normal growth. Cells harboring detrimental mutants thus grow more slowly. This difference in growth rate can be measured by high-throughput sequencing.

The steps involved include first sequencing the pool of mutants at some initial time point and determining the frequency of each mutant in the pool (number of reads for mutant X/total number of reads). Second, the pool of mutants is sequenced after a period of selective growth, with the frequency of each mutant again determined by sequencing. Third, the change in frequency of each mutant from the initial to the final time point is calculated. Mutants with higher growth rates will increase in frequency, and those with lower growth rates will decrease in frequency. The data can be analyzed statistically by programs like Enrich (Fowler *et al.*, 2011).

Other methods for differentiating between mutants in the pool include linking protein function to the expression of an essential gene or to a fluorescent readout, in which case flow cytometry can be used to separate functional and nonfunctional mutants. Regardless of the details of the selective pressure, the approach leverages the ability of high-throughput sequencing to provide the frequency of each mutant in a large pool, thus greatly increasing the throughput of mutagenesis studies.

GENOTYPE–PHENOTYPE MAPS IN DISEASE

A compelling reason to create genotype–phenotype maps is that these data will be valuable in diagnosing, treating, and understanding disease. These maps seem especially critical in cancer, a disease characterized by the accumulation of novel changes in genotype. Large projects have coordinated the extensive sequencing and transcriptional analysis of thousands of samples from many of the most common and deadly types of cancer (McLendon *et al.*, 2008; Cancer Genome Atlas Research Network, 2013). The rationale for these studies, which require the coordinated efforts of clinicians, pathologists, sequencing centers, and bioinformaticians, is threefold. First, finding genes frequently mutated in cancers sheds light on the mechanisms of tumorigenesis, with frequently mutated genes often clustering in a small number of pathways (Vogelstein *et al.*, 2013). Second, the molecular profiling of many samples from one type of cancer, previously typed solely by histology, may reveal multiple subtypes with different clinical properties. For example, gene expression profiles are predictive of prognosis in breast cancer (Millar *et al.*, 2009). Finally, molecular characterization may lead to the discovery of predictive biomarkers that can help guide treatment decisions. For example, cancers that express certain proteins respond better to treatment with drugs that inhibit these proteins (Weigel and Dowsett, 2010). In addition, mutations in drug targets may predict response to the drug (Traina *et al.*, 2014).

Despite these few successes, the search for new prognostic and predictive biomarkers, especially those personalized for individual patients, is complicated by the numbers of rare mutations in cancer and the difficulties in interpreting their effects. All cancers are heterogeneous mixes of clones, each with up to hundreds of de novo nonsynonymous mutations (Vogelstein *et al.*, 2013). We currently try to assign causative mutations by finding genes or networks of genes associated with cancer progression or risk and using predictive tools to stratify mutations occurring in those genes. This approach can miss rare variants of large effect as well as variants that are pathogenic only when combined. Even a recurrent mutation may not be clinically important. For example, large numbers of mutations in a panel of genes recurrently mutated in skin cancers were found in histologically normal skin tissue, suggesting that a cancer-inducing environment or some combination of these mutations is necessary for complete transformation (Martincorena *et al.*, 2015). Estimates that it may take more than a decade for tumor expansion to occur after the initiating driver mutation indicate that the multiple changes required for malignancy accumulate slowly (Yachida *et al.*, 2010). However, tumor expansion likely occurs in a quick, punctuated manner without obvious intermediates between normal and bulk tumor cells (Navin *et al.*, 2011), complicating efforts to determine which variants contribute to the tumor phenotype. It would clearly be useful in prioritizing drug targets and assessing risk to know the marginal fitness advantage conferred to a cancer by each genomic change.

In this vein, deep mutational scanning experiments have advantages over association tests. First and foremost, the functional effect of all single-nucleotide variants can be assayed directly without regard to their frequency in the population. An analysis of the breast cancer susceptibility gene, *BRCA1*, measured the effect of thousands of mutations in the BRCA1 RING domain on two biochemical functions: binding to the BARD1 protein and E3 ubiquitin ligase activity (Starita *et al.*, 2015). These functional scores were used as the inputs to model a third function, homology-directed DNA repair (HDR), which is difficult to measure in bulk but correlates best with cancer susceptibility. The model captures roughly 70% of the variation in HDR scores and greatly outperforms commonly used biological effect prediction algorithms. Thus useful, gene-specific information from assaying variants directly is not captured by models based largely on evolutionary data. Many potential variants have the capacity to be damaging in the context of cancer, as the model predicts that roughly 20% of the assayed mutations in the RING domain would show impaired HDR activity.

As more of the variants seen in large-scale cancer sequencing projects are subjected to direct functional assessment, models that integrate these results should provide better predictions and highlight areas of incomplete understanding. Statistical models could be used to integrate these direct measurements with other evolutionary and functional data to predict deleteriousness. Alternatively, for molecular phenotypes that have strong correlations with disease, such as protein instability, functional scores from these assays could be used as the target of prediction. The validated disease variants currently used to train some variant effect prediction models are spread sparsely throughout the genome, and their collection is subject to ascertainment biases, resulting in models that are not likely generalizable. Direct functional measurements would serve as much more dense gold standards for training.

The eventual objective is to advance beyond predicting only the molecular effects of single variants and to attempt to predict individuals' phenotypes from their genome sequences. To this end, direct functional measurements could be used alongside data from genotype–phenotype association studies to predict disease risk, prognosis, or treatment effectiveness. Several groups have recently used mixed linear models on genotype–phenotype association data from large association studies to estimate the total contribution of genetic variation across a large set of genetic variants to phenotypic variation (Yang *et al.*, 2011). The resulting model, which takes into account the marginal effects of all variants simultaneously, explains more of the heritability of complex traits than the much smaller set of individually significantly associated variants and can be used to predict phenotype from the identity of genotyped single-nucleotide polymorphisms (SNPs). However, the prediction accuracy is still low, and the phenotypic variance that is accounted for is still much lower than the actual heritability, as there is sampling error when estimating the effect of each SNP, and this problem will likely require large increases in sample size to reduce, especially when many contributing variants are rare (Chatterjee *et al.*, 2013).

Deep mutational scanning data could potentially improve SNP-based prediction models by providing information as to how each SNP marginally adds to phenotypic variation. An additional set of parameters, one for each assay, could be added, with each parameter representing the contribution of the functional score to the phenotypic variation. These parameters would replace the corresponding parameters for the variance explained by SNP presence. In cancer-resequencing studies, the resulting model could make use of the presence of rare SNPs that would otherwise not contribute to the prediction. Other types of models, like random forests and neural networks, which can implicitly take into account interactions between the molecular function results, might yield even better predictions, providing that the prediction target population is sufficiently similar to the training population (Quang *et al.*, 2015; Stephan *et al.*, 2015). A major challenge in collecting the data will be the choice of phenotypes to assay. It will be essential to determine the types of data (e.g., stability-based, enzymatic, or expression-based) that are most predictive, so as to prioritize experimental efforts and avoid overfitting by adding unhelpful data. This goal could be facilitated by using methods that rank model inputs by their contribution to the model's accuracy (Saeys *et al.*, 2007).

CONTEXT DEPENDENCE

Mutations do not exist in a vacuum, making the prediction of phenotype from genotype even more difficult. Mutations interact with other mutations, and they interact with features of the environment, such that their effects often depend on context. Consider two mutations in the same gene. If the gene is implicated in a phenotype such as growth rate, then we might expect that the combined effect of the two mutations on growth rate could be found by adding their individual effects. When the actual growth rate differs from this expectation, then the two mutations display epistasis, described quantitatively as how much the double mutant's growth rate differs from the expectation. Several deep mutational scanning experiments have demonstrated such intragenic epistasis, both positive and negative (Araya *et al.*, 2012; McLaughlin *et al.*, 2012). However, these same experiments have shown that the general rule for most mutations in a protein is that they do not exhibit epistasis, with the additive model providing a good estimate of log-transformed double-mutant function.

Epistasis can also occur between two separate genes that both affect the same phenotype; for example, the encoded proteins may interact with each other or take part in the same cellular process. Gene-by-gene epistasis is usually studied through gene deletion, as in synthetic lethality screens, but has not been systematically measured at the level of individual single-nucleotide variants. However, a huge number of such interacting variants might exist, given that a screen of 75% of yeast gene pairs found 170,000 genetic interactions affecting cell growth (Costanzo *et al.*, 2010). There remain substantial technical hurdles to analyzing gene-by-gene interactions by deep mutational scanning, but these types of interactions have been probed via short hairpin RNA screens and have proven to be useful for predicting survival and drug response in cancer patients based on the genetic architecture of their tumors (Jerby-Arnon *et al.*, 2014). This result supports the idea that the specific genetic variants unique to a given tumor might render it vulnerable to inhibition of other genes that interact with those variants, providing a first glimpse of the potential utility of predictive models that make use of a mix of functional and association data.

Epistasis is not limited to deviations from the expected interactions of just two genetic changes. Higher-order epistasis can occur after taking into account epistasis at lower levels (Weinreich *et al.*, 2013). The fitness of a triple mutant may deviate from expectation even if all pairwise epistasis scores are considered, and higher-order epistasis coefficients are in many cases as large or larger than pairwise epistatic coefficients (Weinreich *et al.*, 2013). Instances of higher-order intergenic (as opposed to intragenic) epistasis have also been identified (Taylor and Ehrenreich, 2014). With more interacting components, such as those involved in common disease processes, there will be an intractable number of even third-order epistasis terms that contribute significantly to phenotype. In *Drosophila*, whole genetic interaction networks can change substantially based on genetic background (Chari and Dworkin, 2013). This complexity might seem discouraging, but it is unknown to what extent higher-order interactions play a role in observed human phenotypes. The extent to which epistasis affects phenotypic variation among humans or within tumors depends on the relative frequencies of the interacting sites, and there is some evidence that the contribution of epistasis, while significant when assayed directly, may play a smaller role in human phenotypic variation in complex disease (Hill *et al.*, 2008). One estimate from the analysis of expression quantitative trait loci is that pairwise epistasis explains approximately one-tenth the amount of phenotypic variance that additive effects do (Hemani *et al.*, 2014).

Finally, changes in cell type, temperature, nutrients, chemical messengers, and drug treatments, as well as the presence of other organisms, can also mask or unmask a phenotypic change caused by a mutation. Deep mutational scanning experiments performed under more than one condition have shown that these gene-by-environment interactions are common (Guy *et al.*, 2014; Stiffler *et al.*, 2015).

It is still unclear how much gene-by-gene and gene-by-environment interactions contribute to phenotype. For some phenotypes, like the probability that a given splicing event occurs, epistasis between sequence features and cell type dependence are both important enough that accuracy suffers greatly when these interactions are not included in phenotype predictions (Xiong *et al.*, 2015). In other cases, for example, in single-step enzymatic processes, we might expect less dependence on cellular context and less intergenic epistasis, though intragenic epistasis might still be significant. Deep mutational scanning experiments can functionally characterize all single mutants of a given target but only a subset of double and triple mutants, so errors in predictive models due to epistasis will likely not be overcome by brute force. However, it remains to be seen whether the effects of genetic and environmental interactions will have an appreciable impact on accuracy.

CONCLUSIONS

The genomic diversity of humanity is staggering, and the genomic diversity within cancers greater still. Although only a small fraction of this variation may be relevant to any given disease phenotype, identifying these relevant variants will require functional data for all rare variations. High-throughput phenotyping experiments can score many thousands of genetic variants for their effects on a diversity of molecular properties, but it is unclear which of these properties and targets will be most informative when going from molecular to organismal phenotype. Perhaps the most daunting obstacle is the potential contribution of environmental and genetic interaction terms, especially higher-order interaction terms, to phenotype. By combining diverse types of functional data for genetic variants within a structured mathematical model, we can probe the contribution of molecular phenotypes to disease phenotypes directly, for both rare and common variants. Obtaining high-quality, reproducible functional measurements throughout the genome will require a large, coordinated effort across many laboratories, but the effort will be well placed if the result is a more interpretable genome.

ACKNOWLEDGMENTS

We thank Douglas Fowler, Christine Queitsch, William Noble, and Joshua Akey for helpful comments on the manuscript. This work was supported by P41 GM103533 from the National Institutes of Health (NIH). D.L.Y. acknowledges support from National Institute on Drug Abuse Molecular Pharmacology Training Grant (3T32DA007278-21S1) from the NIH. S.F. is an investigator of the Howard Hughes Medical Institute.

REFERENCES

Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR (2010). A method and server for predicting damaging missense mutations. Nat Methods 7, 248–249.

Agarwala V, Flannick J, Sunyaev S, GoT2D Consortium, Altshuler D (2013). Evaluating empirical bounds on complex disease genetic architecture. Nat Genet 45, 1418–1427.

Araya CL, Fowler DM, Chen W, Muniez I, Kelly JW, Fields S (2012). A fundamental protein property, thermodynamic stability, revealed solely from large-scale measurements of protein function. Proc Natl Acad Sci USA 109, 16858–16863.

Besenbacher S, Liu S, Izarzugaza JMG, Grove J, Belling K, Bork-Jensen J, Huang S, Als TD, Li S, Yadav R, *et al.* (2015). Novel variation and de novo mutation rates in population-wide de novo assembled Danish trios. Nat Commun 6, 5969.

Cancer Genome Atlas Research Network (2013). Integrated genomic characterization of endometrial carcinoma. Nature 497, 67–73.

Chari S, Dworkin I (2013). The conditional nature of genetic interactions: the consequences of wild-type backgrounds on mutational interactions in a genome-wide modifier screen. PLoS Genet 9, e1003661.

Chatterjee N, Wheeler B, Sampson J, Hartge P, Chanock SJ, Park J-H (2013). Projecting the performance of risk prediction based on polygenic analyses of genome-wide association studies. Nat Genet 45, 400–405.

Costanzo M, Baryshnikova A, Bellay J, Kim Y, Spear ED, Sevier CS, Ding H, Koh JLY, Toufighi K, Mostafavi S, *et al.* (2010). The genetic landscape of a cell. Science 327, 425–431.

Dong C, Wei P, Jian X, Gibbs R, Boerwinkle E, Wang K, Liu X (2015). Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies. Hum Mol Genet 24, 2125–2137.

Exome Aggregation Consortium (2015). ExAC Browser, Cambridge, MA. http://exac.broadinstitute.org (accessed 25 May 2015).

Fowler DM, Araya CL, Gerard W, Fields S (2011). Enrich: software for analysis of protein function by enrichment and depletion of variants. Bioinformatics 27, 3430–3431.

Fowler DM, Fields S (2014). Deep mutational scanning: a new style of protein science. Nat Methods 8, 801–807.

Fu W, O'Connor TD, Jun G, Kang HM, Abecasis G, Leal SM, Gabriel S, Rieder MJ, Altshuler D, Shendure J, *et al.* (2013). Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. Nature 493, 216–220.

Gibson G (2012). Rare and common variants: twenty arguments. Nat Rev Genet 13, 135–145.

Grantham R (1974). Amino acid difference formula to help explain protein evolution. Science 185, 862–864.

Guy MP, Young DL, Payea MJ, Zhang X, Kon Y, Dean KM, Grayhack EJ, Mathews DH, Fields S, Phizicky EM (2014). Identification of the determinants of tRNA function and susceptibility to rapid tRNA decay by high-throughput in vivo analysis. Genes Dev 28, 1721–1732.

Hemani G, Shakhbazov K, Westra HJ, Esko T, Henders AK, McRae AF, Yang J, Gibson G, Martin NG, Metspalu A, *et al.* (2014). Detection and replication of epistasis influencing transcription in humans. Nature 508, 249–253.

Henn BM, Botigué LR, Bustamante CD, Clark AG, Gravel S (2015). Estimating the mutation load in human genomes. Nat Rev Genet 16, 333–343.

Hill WG, Goddard ME, Visscher PM (2008). Data and theory point to mainly additive genetic variance for complex traits. PLoS Genet 4, e1000008.

Jerby-Arnon L, Pfetzer N, Waldman YY, McGarry L, James D, Shanks E, Seashore-Ludlow B, Weinstock A, Geiger T, Clemons PA, *et al.* (2014). Predicting cancer-specific vulnerability via data-driven detection of synthetic lethality. Cell 158, 1199–1209.

Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J (2014). A general framework for estimating the relative pathogenicity of human genetic variants. Nat Genet 46, 310–315.

Kong A, Frigge ML, Masson G, Besenbacher S, Sulem P, Magnusson G, Gudjonsson SA, Sigurdsson A, Jonasdottir A, Jonasdottir A, *et al.* (2012). Rate of de novo mutations and the importance of father's age to disease risk. Nature 488, 471–475.

Landrum MJ, Lee JM, Riley GR, Jang W, Rubinstein WS, Church DM, Maglott DR (2014). ClinVar: public archive of relationships among sequence variation and human phenotype. Nucleic Acids Res 42, D980–D985.

Li Q, Liu X, Gibbs RA, Boerwinkle E, Polychronakos C, Qu H-Q (2014). Gene-specific function prediction for non-synonymous mutations in monogenic diabetes genes. PLoS One 9, e104452.

Martincorena I, Roshan A, Gerstung M, Ellis P, Van Loo P, McLaren S, Wedge DC, Fullam A, Alexandrov LB, Tubio JM, *et al.* (2015). High burden and pervasive positive selection of somatic mutations in normal human skin. Science 348, 880–886.

McLaughlin RN, Jr., Poelwijk FJ, Raman A, Gosal WS, Ranganathan R (2012). The spatial architecture of protein function and adaptation. Nature 491, 138–142.

McLendon R, Friedman A, Bigner D, Van Meir EG, Brat DJ, Mastrogianakis GM, Olson JJ, Mikkelsen T, Lehman N, Aldape K, *et al.* (2008). Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature 455, 1061–1068.

Millar EK, Graham PH, O'Toole SA, McNeil CM, Browne L, Morey AL, Eggleton S, Beretov J, Theocharous C, Capp A, *et al.* (2009). Prediction of local recurrence, distant metastases, and death after breast-conserving therapy in early-stage invasive breast cancer using a five-biomarker panel. J Clin Oncol 27, 4701–4708.

Moutsianas L, Agarwala V, Fuchsberger C, Flannick J, Rivas MA, Gaulton KJ, Albers PK, McVean G, Boehnke M, Altshuler D, *et al.* (2015). The power of gene-based rare variant methods to detect disease-associated variation and test hypotheses about complex disease. PLoS Genet 11, e1005165.

Navin N, Kendall J, Troge J, Andrews P, Rodgers L, McIndoo J, Cook K, Stepansky A, Levy D, Esposito D, *et al.* (2011). Tumour evolution inferred by single-cell sequencing. Nature 472, 90–94.

Ng PC, Henikoff S (2001). Predicting deleterious amino acid substitutions. Genome Res 11, 863–874.

Patwardhan RP, Lee C, Litvin O, Young DL, Pe'er D, Shendure J (2009). High-resolution analysis of DNA regulatory elements by synthetic saturation mutagenesis. Nat Biotechnol 27, 1173–1175.

Quang D, Chen Y, Xie X (2015). DANN: a deep learning approach for annotating the pathogenicity of genetic variants. Bioinformatics 31, 761–763.

Saeys Y, Inza I, Larrañaga P (2007). A review of feature selection techniques in bioinformatics. Bioinformatics 23, 2507–2517.

Starita LM, Young DL, Islam M, Kitzman JO, Gullingsrud J, Hause RJ, Fowler DM, Parvin JD, Shendure J, Fields S (2015). Massively parallel functional analysis of BRCA1 RING domain variants. Genetics 200, 413–422.

Stenson PD, Mort M, Ball EV, Shaw K, Phillips A, Cooper DN (2014). The Human Gene Mutation Database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine. Hum Genet 133, 1–9.

Stephan J, Stegle O, Beyer A (2015). A random forest approach to capture genetic effects in the presence of population structure. Nat Commun 6.

Stiffler MA, Hekstra DR, Ranganathan R (2015). Evolvability as a function of purifying selection in TEM-1 β-lactamase. Cell 160, 882–892.

Taylor MB, Ehrenreich IM (2014). Genetic interactions involving five or more genes contribute to a complex trait in yeast. PLoS Genet 10, e1004324.

Tennessen JA, Bigham AW, O'Connor TD, Fu W, Kenny EE, Gravel S, McGee S, Do R, Liu X, Jun G, *et al.* (2012). Evolution and functional impact of rare coding variation from deep sequencing of human exomes. Science 337, 64–69.

Traina F, Visconte V, Elson P, Tabarroki A, Jankowska AM, Hasrouni E, Sugimoto Y, Szpurka H, Makishima H, O'Keefe CL, *et al.* (2014). Impact of molecular mutations on treatment response to DNMT inhibitors in myelodysplasia and related neoplasms. Leukemia 28, 78–87.

Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz LA, Kinzler KW (2013). Cancer genome landscapes. Science 339, 1546–1558.

Weigel MT, Dowsett M (2010). Current and emerging biomarkers in breast cancer: prognosis and prediction. Endocr Relat Cancer 17, R245–R262.

Weinreich DM, Lan Y, Wylie CS, Heckendorn RB (2013). Should evolutionary geneticists worry about higher-order epistasis? Curr Opin Genet Dev 23, 700–707.

Xiong HY, Alipanahi B, Lee LJ, Bretschneider H, Merico D, Yuen RKC, Hua Y, Gueroussov S, Najafabadi HS, Hughes TR, *et al.* (2015). The human splicing code reveals new insights into the genetic determinants of disease. Science 347, 1254806.

Yachida S, Jones S, Bozic I, Antal T, Leary R, Fu B, Kamiyama M, Hruban RH, Eshleman JR, Nowak MA, *et al.* (2010). Distant metastasis occurs late during the genetic evolution of pancreatic cancer. Nature 467, 1114–1117.

Yang J, Lee SH, Goddard ME, Visscher PM (2011). GCTA: a tool for genome-wide complex trait analysis. Am J Hum Genet 88, 76–82.