Massively Parallel Genetics

Jay Shendure*,†,1,2 and Stanley Fields*,‡,1,2

*Department of Genome Sciences and ‡Department of Medicine, University of Washington, and †Howard Hughes Medical Institute, Seattle, Washington 98115

ORCID IDs: 0000-0002-1516-1865 (J.S.); 0000-0001-5504-5925 (S.F.)

ABSTRACT Human genetics has historically depended on the identification of individuals whose natural genetic variation underlies an observable trait or disease risk. Here we argue that new technologies now augment this historical approach by allowing the use of massively parallel assays in model systems to measure the functional effects of genetic variation in many human genes. These studies will help establish the disease risk of both observed and potential genetic variants and to overcome the problem of "variants of uncertain significance."

KEYWORDS DNA; functional assays; genetics; mutations; sequencing

Since genetics began as a field, its rate-limiting step has been the cost and resolution of ascertaining genotypes. However, the recent emergence of massively parallel DNA sequencing has made genotyping both comprehensive and cheap: we can now conduct genetic analyses at a scale scarcely imaginable a decade ago (1000 Genomes Project *et al.* 2015). But the ease of genotyping has exposed the limits of genetic analyses, particularly as they have been applied to human phenotypes. Here, we highlight some of these limitations and argue that massively parallel approaches to experimentally measure the functional consequences of individual variants will advance our ability to interpret human genomes.

As a first example, consider genetic analyses of common diseases. Whereas linkage studies, with some important exceptions, largely proved to be a disappointment in this context, genome-wide association studies have been very successful in identifying thousands of reproducible associations between common variants and common diseases (Price *et al.* 2015). However, the resolution of these associations is inherently limited by linkage disequilibrium in human populations, such that a variant causally underlying an association as well as the gene through which its effects are mediated are rarely known. Although strategies have been developed for fine-mapping, these do not scale or generalize well, and our inability to nail down variants and genes for these associations fundamentally limits this approach. Furthermore, most of the variants discovered by these studies have only a modest effect on phenotype; for example, these variants sum to on the order of 10–20% of narrow-sense heritability for common diseases such as type 2 diabetes, multiple sclerosis, and Crohn's disease (Visscher *et al.* 2012).

As a second example, consider genetic analyses of rare diseases. We are well along the path to the comprehensive elucidation of genes that underlie monogenic syndromes as well as monogenic forms of autism and intellectual disability (Chong *et al.* 2015). However, we remain poorly equipped to predict the consequences of individual variants within implicated genes, such that "variants of uncertain significance" persist as the disappointing outcome in many cases, even for well-studied cancer-predisposing genes such as *BRCA1*. It is unlikely that simply more sequencing will resolve this problem. Although every possible single nucleotide variant (SNV) compatible with life is quite possibly present in a living human somewhere on earth (Shendure and Akey 2015), nearly all such individual variants are exceedingly rare. As such, even with millions of genomes and medical records in hand, and even where there is a specific hypothesis (*e.g.*, that a missense variant resulting in loss of function in *BRCA1* will lead to cancer), we will remain poorly powered to quantify the risk that an individual SNV confers by genotype–phenotype analyses alone.

The shared obstacle of these examples is the fact that we cannot control which variants and haplotypes are present in the individuals that we study: human genetics is necessarily "observational." If we want to confidently interpret variants

doi: 10.1534/genetics.115.180562
[1]These authors contributed equally to this work.
[2]Corresponding authors: Department of Genome Sciences, Box 355065, 3720 15th Ave. NE, Seattle, WA 98195. E-mail: shendure@uw.edu; and E-mail: fields@uw.edu

observed even in only one or a few humans, we propose that "observational" genetics in humans must be supplemented with "perturbational" genetics in model systems: *in vivo* in model organisms, *in vitro* in tissue culture lines, or in cell-free assays of protein function (Fowler *et al.* 2010; Starita *et al.* 2015). This "massively parallel" paradigm requires that we know enough about the function of a gene to establish a DNA sequencing-based assay for its activity. The approach takes advantage of recent methodological developments and involves the following:

1. Generating very large numbers of genetic variants of a sequence of interest, either by directed or random mutagenesis. The toolbox for this task leverages advances in DNA synthesis, including complex pools of microarray-derived oligonucleotides that can be used to program allelic series (Melnikov *et al.* 2014; Kitzman *et al.* 2015).
2. Introducing the allelic series into a model system. This step can use genome editing to insert the alleles at the endogenous locus or at another locus in a relevant human cell line or organoid system (Findlay *et al.* 2014); alternatively, the alleles can be introduced into a nonhuman model organism or into a protein display system (Fowler *et al.* 2014). Because the physiology of model organisms and human cell lines differs from that of the human organism, the choice of functional assay is not always straightforward, and any assay must be validated and calibrated as we describe below.
3. Measuring the functional effects of individual variants within the allelic series in one or several assays. For an assay to be useful, its results must correlate with the organismal phenotype of interest, and assay validation remains an enormous challenge. However, it may be possible to develop assays that generalize to some degree. For example, the impact of an allelic series of regulatory mutations can be studied through their effects on *cis* gene expression, provided that the assay is performed in the appropriate cell type (Patwardhan *et al.* 2009). For genes that play a large role in cellular physiology, global molecular phenotypes such as the transcriptome may represent a generic means of ascertaining functional effects (Hughes *et al.* 2000). As with the construction of the allelic series and its introduction into the model system, a functional assay should be "multiplexed" as well, such that the functional effects of thousands of variants can be studied within a single workflow. For example, an assay that relies on single-cell RNA-seq as a global molecular phenotype requires that this be practical to perform on a very large number of cells (Macosko *et al.* 2015), each of which contains a unique variant in a protein of interest.
4. Partitioning variants based on the outcome of a functional assay. Because of the large number of variants concurrently tested, a distribution of effect sizes will emerge and guide interpretation (Starita *et al.* 2015). For example, subclasses such as nonsense and synonymous mutations will establish the range of the assay from the most to the least detrimental effects that can be observed, thereby informing the interpretation of similarly behaving missense or regulatory mutations.
5. Calibrating experimentally measured effects to human phenotypes. Because *in vitro* functional assays are necessarily imperfect, this step will be essential to validate whether experimental measurements meaningfully capture disease risk in humans and, moreover, to quantify this risk (Majithia *et al.* 2014). We will be better able to carry out this validation because variants that exhibit similar functional profiles from mutational scans can be binned together. From genotype–phenotype data of increasingly larger numbers of humans, we can analyze the odds ratios for disease risk of the variants that fall into each bin. The resulting comparison will establish whether the subclasses defined by functional assay match those defined by patient data. Such validation analyses should be strongly facilitated by large-scale genotype–phenotype databases such as envisioned in the U.S. Precision Medicine Initiative (Collins and Varmus 2015). A functional assay in which the rank order of effects does not match patient data would be confounding and indicate that a different assay is needed to accurately predict disease risk or that our approach is not amenable to this gene.

Functionally testing every possible variant in the human genome in all relevant assays and contexts is of course unrealistic. Furthermore, each of these variants exists within a context of other genetic variation the precise epistatic relationships of which may take decades to decipher. However, by focusing on genes of special value—and assuming that epistasis may not be a significant contributor to the overall burden of disease—we may be able to advance the field of human genetics beyond the bottlenecks that it currently faces. Our proposed approach also benefits from the "multiplexing" of large numbers of independent functional assays within single experiments, as well as on continuing technological improvements, such that the cost in personnel and sequencing reagents for each mutational scan is anticipated to steadily decrease.

Establishing disease risk for all potential variants in cancer-predisposing genes would help to overcome the long-standing problem of variants of uncertain significance. These experiments are currently feasible for many genes the functions of which in signaling, transcription, cell cycle control, and DNA repair are both known and able to be recapitulated in a model system. Defining causal variants and genes within genome-wide association study-implicated haplotypes would dramatically advance our understanding of the genetic basis of common disease. Finally, contemporary computational approaches to variant effect prediction are not as powerful as they might be because they are largely trained on evolutionary metrics and amino acid chemical similarity and they aggregate the effects of variants in many different proteins, leading to poor resolution (Cooper and Shendure 2011). It is likely that by experimentally measuring the functional effects of

thousands to millions of variants in individual genes, we will be in a much better position to apply the sorts of "deep learning" approaches that are proving successful in other fields for complex pattern recognition.

Functional assays have long been used toward the interpretation of individual genetic variants, both for understanding biological mechanisms and for informing clinical decision-making, but largely in a one-off fashion. What we suggest here is simply taking advantage of recent methodological developments to take the next logical step of massively parallelizing such experiments. Whether the resulting data sets of measurements will be predictive of organismal phenotypes remains a hypothesis. However, given the ongoing explosion in human genome sequencing, it seems very much worth testing.

Literature Cited

Chong, J. X., K. J. Buckingham, S. N. Jhangiani, C. Boehm, N. Sobreira *et al.*, 2015 The genetic basis of Mendelian phenotypes: discoveries, challenges, and opportunities. Am. J. Hum. Genet. 97: 199–215.

Collins, F. S., and H. Varmus, 2015 A new initiative on precision medicine. N. Engl. J. Med. 372: 793–795.

Cooper, G. M., and J. Shendure, 2011 Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data. Nat. Rev. Genet. 12: 628–640.

Findlay, G. M., E. A. Boyle, R. J. Hause, J. C. Klein, and J. Shendure, 2014 Saturation editing of genomic regions by multiplex homology-directed repair. Nature 513: 120–123.

Fowler, D. M., C. L. Araya, S. J. Fleishman, E. H. Kellogg, J. J. Stephany *et al.*, 2010 High-resolution mapping of protein sequence-function relationships. Nat. Methods 7: 741–746.

Fowler, D. M., J. J. Stephany, and S. Fields, 2014 Measuring the activity of protein variants on a large scale using deep mutational scanning. Nat. Protoc. 9: 2267–2284.

Hughes, T. R., M. J. Marton, A. R. Jones, C. J. Roberts, R. Stoughton *et al.*, 2000 Functional discovery via a compendium of expression profiles. Cell 102: 109–126.

Kitzman, J. O., L. M. Starita, R. S. Lo, S. Fields, and J. Shendure, 2015 Massively parallel single-amino-acid mutagenesis. Nat. Methods 12: 203–206

Macosko, E.Z., A. Basu, R. Satija, J. Nemesh, K. Shekar *et al.* 2015 Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. Cell 161: 1202–1214.

Majithia, A. R., J. Flannick, P. Shahinian, M. Guo, M. A. Bray *et al.*, 2014 Rare variants in PPARG with decreased activity in adipocyte differentiation are associated with increased risk of type 2 diabetes. Proc. Natl. Acad. Sci. USA 111: 13127–13132.

Melnikov, A., P. Rogov, L. Wang, A. Gnirke, and T. S. Mikkelsen, 2014 Comprehensive mutational scanning of a kinase in vivo reveals substrate-dependent fitness landscapes. Nucleic Acids Res. 42: e112.

1000 Genomes Project ConsortiumAuton, A., L. D. Brooks, R. M. Durbin, E. P. Garrison *et al.*, 2015 A global reference for human genetic variation. Nature 526: 68–74.

Patwardhan, R. P., C. Lee, O. Litvin, D. L. Young, D. Pe'er *et al.*, 2009 High-resolution analysis of DNA regulatory elements by synthetic saturation mutagenesis. Nat. Biotechnol. 27: 1173–1175.

Price, A. L., C. C. Spencer, and P. Donnelly, 2015 Progress and promise in understanding the genetic basis of common diseases. Proc. R. Soc. Lond. B Biol. Soc. 282: 20151684.

Shendure, J., and J. M. Akey, 2015 The origins, determinants, and consequences of human mutations. Science 349: 1478–1483.

Starita, L. M., D. L. Young, M. Islam, J. O. Kitzman, J. Gullingsrud *et al.*, 2015 Massively parallel functional analysis of BRCA1 RING domain variants. Genetics 200: 413–422.

Visscher, P. M., M. A. Brown, M. I. McCarthy, and J. Yang, 2012 Five years of GWAS discovery. Am. J. Hum. Genet. 90 (1): 7–24.

Communicating editor: M. Johnston