Citation: Clin Transl Sci (2017) 10, 67–77; doi:10.1111/cts.12440

REVIEW

High-Throughput Assays to Assess the Functional Impact of Genetic Variants: A Road Towards Genomic-Driven Medicine

J Ipe[1], M Swart[1], KS Burgess[1,2] and TC Skaar[1,*]

INTRODUCTION

Genome-wide genotyping and DNA sequencing has led to the identification of large numbers of genetic variants that are associated with many clinical phenotypes. The functional impacts of most of the variants are unknown. In this article, we review high-throughput assays that have been developed to assess a variety of the functional impacts of the variants. A better understanding of their functions should facilitate the implementation of many more variants in genomic-driven medicine.

A cornerstone of precision medicine is the incorporation of genetic information into healthcare decisions. This approach relies on understanding the genome complexity, the genetic differences that exist between individuals, and the functional consequences of the genetic variants. In the personal genome era, improvements in sequencing technologies are leading to continuous identification of new variants and further illustrating the complexity of the human genome and the genetic diversity between populations.

Genomic variants across individuals

Large-scale high-throughput sequencing studies, such as the 1000 Genomes and NHLBI GO Exome Sequencing Projects, have already identified millions of genetic variants among individuals from different populations and have established a comprehensive resource on human genetic variation.[1,2] The genetic variants are cataloged in public databases, such as dbSNP (https://www.ncbi.nlm.nih.gov/snp/) and dbVAR (https://www.ncbi.nlm.nih.gov/dbvar/) (**Table 1**). The current build of dbSNP (build 147, updated on 14 April 2016) contains ∼154 million single nucleotide variants (SNVs) of which about 101 million have been validated and nearly 89 million are within genes. dbVAR (updated on 28 September 2016) contains ∼5 million structural variants and ∼2.3 million, 1.3 million, and 1.2 million of these variants are contributed by copy number variants, short tandem repeats, and insertions, respectively.

In the 1000 Genomes Project, sequencing was carried out on 2,504 individuals from 26 populations in Africa, East Asia, Europe, South Asia, and the Americas. More than 88 million variants were identified, of which 84.7 million were single nucleotide polymorphisms (SNPs), 3.6 million were short indels, and 60,000 were structural variants. Only 8 million of the identified autosomal variants were observed in more than 5% of individuals, while 64 million rare variants (frequency of $<0.5\%$) were identified. Substantial differences exist in the distribution of variants between populations, with 762,000 variants being rare (frequency of $<0.5\%$) in the global population, but more common ($>5\%$) in at least one population group. Eighty-six percent of variants were only present in a single continental group. Sequencing of individuals from South Asian and African populations contributed to 24% and 28%, respectively, of novel variants discovered.[1] Sudmant *et al.*[3] reported the identification of 68,818 structural variants when analyzing sequencing data from the 1000 Genomes Project. The majority of these structural variants are deletions (42,279) with a median site size of 2,455 bp and median alleles per individual of 2,788.

The nucleotide substitution rate is an important factor underlying the degree of genetic variation between individuals. Scally[4] reported a present-day germline mutation rate of $0.5 \times 10^{-9}\text{bp}^{-1}\text{year}^{-1}$. This mutation rate translates into ∼30 *de novo* variants in each offspring that are absent in the parents. The introduction of 30 new DNA variants with every meiosis event over a period of 3.7–6.6 million years (evolution of the human species) and rapid expansion of the human population during the last 10,000 years resulted in the observed enormous diversity of the human genome. For most of the genetic variants, the impact on gene function and the effect on disease susceptibility remains unknown.

Genomic variants within individuals

High-throughput sequencing continues to produce a more accurate estimation of how much genetic variation exists within and between genomes of individuals of different ethnicities. Typically, each genome has 4–5 million sites that differ from the reference human genome; the greatest number of variant sites were observed among individuals of African ancestry. Although SNPs and indels account for $>99.9\%$ of variants, the typical genome contains 2,100–2,500 structural variants that affect about 20 million bases of sequence. Deep sequencing allows for the identification of rare variants and an estimated 1–4% of variants (40,000–200,000) observed in a genome are rare (frequency

[1] Indiana University School of Medicine, Department of Medicine, Division of Clinical Pharmacology, Indianapolis, Indiana, USA; [2] Indiana University School of Medicine, Department of Pharmacology and Toxicology, Indianapolis, Indiana, USA. *Correspondence: TC Skaar (tskaar@iu.edu).
Received 16 November 2016; accepted 3 January 2017; published online on 18 February 2017. doi:10.1111/cts.12440

Table 1 Databases that catalog human genetic variation and phenotypic relationships

Databases	Description	Link
1000 Genomes Project	Comprehensive catalog of human genetic variation	http://browser.1000genomes.org/index.html
ClinVar	Information about genomic variation and its relationship to human health	https://www.ncbi.nlm.nih.gov/clinvar/
Catalog of somatic mutations in cancer (COSMIC)	Comprehensive resource for exploring the impact of somatic mutations in human cancer	http://cancer.sanger.ac.uk/cosmic
DBASS	Database of new exon boundaries induced by pathogenic mutations in human disease genes	http://www.dbass.org.uk
dbGaP	NCBI's database of genotypes and phenotypes	https://www.ncbi.nlm.nih.gov/gap
dbSNP	NCBI's database of single nucleotide polymorphisms (SNPs) and multiple small-scale variations (including insertions/deletions, microsatellites and nonpolymorphic variants)	https://www.ncbi.nlm.nih.gov/snp
dbVar	NCBI's database of genomic structural variation	https://www.ncbi.nlm.nih.gov/dbvar
DECIPHER	Web-based database incorporating a suite of tools designed to aid the interpretation of genomic variants	https://decipher.sanger.ac.uk
Exome Aggregation Consortium	Aggregate of exome sequencing data from a variety of large-scale sequencing projects	http://exac.broadinstitute.org
GTEx Portal	Data repository for genotype and tissue-specific gene expression data	http://www.gtexportal.org/home/
miRBase	Database of published microRNA sequences and annotation	http://mirbase.org
MirSNP	Collection of human SNPs in predicted microRNA target sites	http://bioinfo.bjmu.edu.cn/mirsnp/search/
NHGRI-EBI GWAS Catalog	Catalog of published genome-wide association studies	http://www.ebi.ac.uk/gwas/
NHLBI Exome Sequencing Project	Data repository for exome sequence variants related to heart, lung and blood disorders	http://evs.gs.washington.edu/EVS/
Online Mendelian Inheritance in Man	Catalog of human genes and genetic disorders for relationships between phenotype and genotype	http://omim.org/
PharmGKB	Pharmacogenomics knowledge resource encompassing clinical information	https://www.pharmgkb.org
PolymiRTS	Database of naturally occurring DNA variations in microRNA seed regions and microRNA target sites	http://compbio.uthsc.edu/miRSNP/
SPHINX – A resource of the eMERGE Network	A web-based tool to access the pharmacogenetics gene sequence data of the eMERGE-PGx project	https://www.emergesphinx.org
The Human Gene Mutation Database	Collection of published gene lesions responsible for human inherited disease	http://www.hgmd.cf.ac.uk/ac/index.php

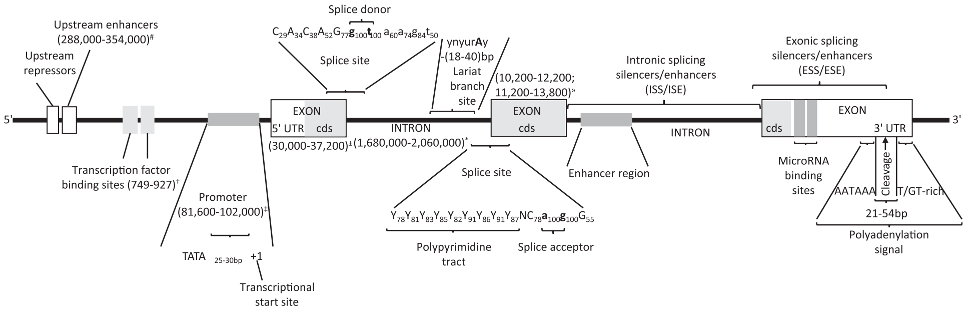

Figure 1 Representation of important gene regions with the number of genetic variants within a typical European or African human genome shown in brackets. Number of genetic variants within upstream enhancer regions#, transcription factor binding sites†, promoter regions‡, 5'- and 3'-untranslated regions±, and intronic regions*. »Number of nonsynonymous;synonymous genetic variants within coding regions.[1]

of <0.5%). A typical genome reportedly contained 149–182 sites with protein truncation variants, 10,000–12,000 sites with nonsynonymous variants and 459,000–565,000 variant sites within regulatory regions (untranslated regions, promoters, insulators, enhancers, and transcription factor binding sites). The number of ClinVar variants (those associated with clinical phenotypes) within a typical genome range from 24–30.[1] Tennessen *et al.*[2] suggested that 2.3% of SNVs per individual exome are thought to disrupt protein function of about 313 of the 23,500 protein-coding genes and nearly 96% of SNPs predicted to affect gene function are rare. (**Figure 1**) is a representation of functionally important regulatory and gene regions with the number of variants within these regions for a typical genome.

Genomic variants associated with phenotypes

Genome-wide association studies (GWAS) have been used to determine which of the identified variants are associated with diseases. To date, more than 3,200 GWA studies have been conducted (http://www.ebi.ac.uk/gwas) and ~10,000 common SNPs have been associated with human traits and diseases through GWA studies.[5] Gusev *et al.*[6] estimated that ~80% of phenotypic heritability of common diseases and traits are explained by variants in noncoding regulatory regions. Approximately 2,000 variants per genome have been associated with complex traits in GWA studies.[1] However, testing of such a large number of SNPs in a GWA study requires correction for multiple testing to decrease the number of false-positive associations by using very stringent significance thresholds. Bonferroni correction for multiple testing (0.05/number of tests) is often used, but it can result in overcorrection and, thus, miss SNPs that really are associated with the phenotype.[7] A large number of study participants are also needed to identify rare causal variants with the use of GWAS.[8]

Genetic variants impact drug metabolism, efficacy, and adverse event risk and are especially relevant to precision medicine. Fujikura *et al.*[9] analyzed sequencing data from the 1000 Genomes and the NHLBI GO Exome Sequencing Projects; they reported a total of 6,165 SNVs in the 57 *cytochrome P450* (*CYP*) genes. Eighty-three percent of the 4,025 SNPs within the coding regions were very rare (frequency of <0.1%) and 65% were nonsynonymous substitutions. The calculated total number of genetic variations in *CYP* genes of 1 million Europeans and Africans was 3.4×10^4 and 4.8×10^4, respectively.[9] Furthermore, every individual of European descent carries on average 94.6 SNVs in *CYP* genes, of which 24.6 are nonsynonymous, within splice sites, or affect stop codons.[9] In the recent PGRN-seq study, 82 genes of pharmacogenomics relevance were sequenced among 5,639 individuals and 40,549 SNVs identified. Of the identified variants, 8,126 were in coding regions (4,858 missense, 3,169 synonymous, and 99 stop gain variants) and 19,923 were in noncoding regions (5,231 intronic, 5,981 upstream, 3,444 downstream, 4,165 3′UTR, 903 5′UTR, and 199 other variants). The majority (~96%) of individuals had one or more Clinical Pharmacogenetics Implementation Consortium Level A actionable variants, while ~23% ($n = 1,273$) of individuals have a single Level A actionable variant.[10]

The Human Gene Mutation Database (HGMD) is a repository of mutations associated with diseases; they are based on published literature, including GWA studies, and as of June 2013, had 141,161 germline mutation entries in 5,700 unique genes. Missense substitutions, nonsense substitutions, splicing substitutions, and substitutions within regulatory elements account for 44%, 11%, 9%, and 2% of the total disease-associated mutations, respectively.[11] Exome sequencing of healthy individuals revealed that each of these individuals carried 40–110 disease-causing mutations as classified by HGMD.[12]

Combining genomic variants for phenotypic prediction

The large amount of genomic variation data now available has created a clear need for the functional characterization of many genetic variants; this should help to distinguish disease-causing variants vs. passenger mutations. In most cases, genotype–phenotype association studies are impractical to assess the role of rare genetic variants, as very large patient cohorts are needed to include enough patients with the rare variants to achieve statistical significance. An alternative is to determine the functional impact of the rare variants and then combine variants with similar functional effects into one group. This has been done with previous studies focusing on *CYP2D6* by assigning an "activity score" to each patient that is derived based on the patients' genotype. Many variants, including rare variants, are classified as functional, partially functional, or nonfunctional. The activity score is then calculated and translated into a predicted CYP2D6 phenotype.[13,14] For any other gene, if the functional impact of the variants are known, this approach could be used to simplify the genotype interpretation and facilitate genotype–phenotype association studies. The remainder of this review will discuss the current status of many high-throughput functional screening assays (**Table 2**). These assays should help distinguish the functional from passenger variants, which will provide valuable information for the successful implementation of genomic-driven medicine.

GENETIC VARIANTS WITHIN CODING REGIONS

The exome is ~1% of the human genome and contains 23,500 protein-coding genes with roughly 180,000 exons. Large-scale sequencing studies have focused on identifying genetic variants within the exome, as these variants could alter protein function.[1,2] The number of identified genetic variants differs significantly between populations, with individuals of African descent being more genetically diverse.[1] African individuals typically carry 12,200 nonsynonymous and 13,800 synonymous variants. The number of nonsynonymous (10,200) and synonymous (11,200) variants identified among individuals of East Asian or European ancestry were fewer (**Figure 1**).[1] Despite the identification of numerous variants in regions coding for proteins, only a portion of these variants disrupt protein function and are likely disease-causing.

The average number of loss-of-function variants per genome ranged from 149 in Europeans to 182 among Africans.[1] Missense or nonsense substitutions in protein-coding genes contribute ~55% of variants implicated in disease.[11] Missense and nonsense nucleotide substitutions and frameshift indels alter the amino acid sequences of the proteins, which can lead to altered secondary, tertiary, and quaternary structures of proteins. These alterations can change many characteristics of the proteins, such as thermodynamic stability and cellular localization and, consequently, cellular functions of the protein, such as enzymatic activity, cell signaling, and ligand binding.[15] Therefore, it is necessary to distinguish loss-of-function variants that could be disease-causing from neutral indels or nucleotide substitutions.

Multiple computational tools have been developed to predict if variants affect protein structure and stability and whether variants in conserved regions are neutral, deleterious, or hyperactivating.[16,17] These currently available tools lack accuracy and, thus, cannot be used in a clinical setting.[18] The models used for these prediction tools are limited by the accuracy of annotated variant effects and evolutionary measures in the training data sets. The complex relationship between evolution and phenotypic effect could also result in high false-positive and false-negative prediction rates by these tools.[17,19,20] The challenges with computational prediction tools can be improved by combining this approach with experimental functional characterization of

Table 2 Summary of high-throughput assays discussed in this review

Genomic region	Name of the assay	Description
Coding	Deep mutational scanning	Mutagenesis method where protein expression and mutant selection are coupled with high-throughput sequencing to determine various functions of variants in the coding region.
Regulatory	Massively parallel reporter assay (MPRA)	Barcoded luciferase plasmids containing variants in *cis*-regulatory elements are inserted into animal model/cell culture and analyzed by RNA-seq.
	CRISPR Cas9-mediated *in situ* saturating mutagenesis	CRISPR-Cas9 mutagenesis to disrupt all sequences within an enhancer region. Cells are sorted and sequenced.
	Multiplexed editing regulatory assay (MERA)	CRISPR-Cas9 based mutational tool to generate variations in the *cis*-regulatory region of genes. Cells are sorted and sequenced.
	Self-transcribing active regulatory region sequencing (STARR-seq)	Ectopic, plasmid-based assay that allows for active enhancers to self-transcribe and analyzed by RNA-seq.
Splicing	High-throughput mini-gene reporter assay	A modified mini-gene assay where a pool of wildtype and variant splice sites are transfected into cells and splice products analyzed by RNA-seq.
	In vitro splicing assay	Radiolabeled RNAs are incubated in nuclear extracts after which the splice products are separated and analyzed by RNA-seq.
	Modified systematic evolution of ligands by exponential enrichment (SELEX)	A method that identifies altered RNA-protein interactions in test splice-sites that contain genetic variations.

variants. Large data sets with experimental measures of the phenotypic effects of variants can also be used to confirm predictions or act as training data sets to improve prediction algorithms.[21]

Mutagenesis studies are commonly used to assess the effect of genetic variants on protein function. A forward genetics approach is time-consuming, because random mutations are created and genes are then identified based on the phenotype that develops. A reverse genetics approach involves the mutagenesis of defined genes, which is followed by functional assays to characterize protein sequence–function relationships; however, these are low-throughput, as the effect of only a small number of variants are assessed. These traditional approaches are also especially laborious as they require the use of a wide variety of techniques and personnel expertise, depending on the function of a specific protein being tested.[22]

High-throughput functional assays

To improve the rate of functional testing, deep mutational scanning has been developed as a high-throughput assay to characterize the function of thousands of variants simultaneously.[21–24] This method can be used to test the functional impact of multiple variant types, including SNPs, indels, and larger structural variants. Design of a deep mutational scanning experiment depends on the type of protein functional assay that is used. For example, for genes that encode enzymes, an enzyme reporter assay may be used. For proteins without known functions or if the measurement of interest is protein stability, then quantifying the protein levels may be desirable. The deep mutational scanning process can be divided into several steps roughly defined as mutagenesis, protein expression, mutant selection, high-throughput sequencing, and statistical analysis. A detailed deep mutational scanning protocol has recently been published by Fowler and Fields.[23] Several studies have used deep mutational scanning with different mutagenesis methods, expression systems, and selection approaches to create sequence-function maps. The first step involves the synthesis of a systematic or random library of mutants that target a specific site in a protein. This step can be performed by creating oligos either designed with defined mutations or mutations introduced randomly through polymerase chain reaction (PCR) amplification. Mingo *et al.*[25] recently developed a one-tube-only standardized site-directed mutagenesis approach. Oligo synthesis is followed by the introduction of the mutant oligo pool into an expression system. Forsyth *et al.*[26] demonstrated the use of mammalian cell-based assays where a protein is expressed from a plasmid or viral vector. Alternatively, M13 or T7 phage-display systems can be used to display up to 10^{12} clones, about 10^{10} clones in bacteria, 10^6 in yeast expression systems, or more than 10^{12} proteins in ribosome display systems.[27–30] The choice of protein expression system to use depends on features of the system, such as how well the expressed protein variant represent the phenotype, ability of the system to do appropriate post-translational modifications, and not only on the number of possible clones. A moderate selection pressure is applied that is appropriate to the protein function assessed. Variant effects have been assessed by testing their impact on protein structure, mechanism of action, catalytic or enzymatic activity (for example, phosphorylation or ubiquitination), thermodynamic stability, protein interaction, peptide binding, DNA or RNA binding, ligand binding, epitope binding, protein aggregation, or expression of a fluorescent protein.[18,21–23] Multiple studies have suggested that additional selection rounds improved accuracy of estimating the fitness for each variant.[31,32]

High-throughput sequencing is then used to identify the variants with the altered phenotypic activity. During the creation of the pool of plasmids a barcoding strategy, where ~20 bp barcodes specific to each mutation is added outside of the open reading frame, is useful for massive parallelization and correction for library amplification biases and the sequencing error rate.[33] A library of 100,000 variants requires about 10^7 sequence reads for adequate coverage (i.e., at least 100 reads per variant).[22,23] It is important to determine the initial frequency of each mutant within the pool before selection. Enrichment of beneficial variants or depletion of deleterious variants are calculated by comparing the frequency (again determined by sequencing) of each mutant after one or multiple rounds of selection to the initial frequency. Statistical analysis is used to identify mutants that are significantly increased or decreased in frequency during selection.[18,21–24,34] Data analysis is straightforward when direct selection is used for a protein property; for example, assessment of thermodynamic stability by thermal denaturation. Fowler *et al.*[35] developed a freely available software package, Enrich, to convert high-throughput sequencing data into a functional score for each variant and create a sequence-function map. Recently, another software package called dms_tools was developed to infer mutation impact from mutation count data by using a likelihood-based analysis.[36] However, standards for deep mutational scanning data analysis have not yet been developed.[22,23]

The context in which mutations occur further complicates the prediction of phenotype from genotype. Deep mutational scanning offers an unbiased technique to test the effect of a combination of mutants at once. Mutants can display epistasis when the observed effect is different from the expected additive effect of the mutants. Mutants might together either cause an unexpected large change in activity or one variant might rescue the destabilizing effect of another.[18] Hemani *et al.*[37] estimated that pairwise epistasis explains approximately one-tenth the amount of phenotypic variance that additive effects do. Wu *et al.*[38] developed a method to calculate the estimated mutational stability effect from double-substitution functional fitness profiles to account for the effect of variant combinations.

Several challenges and potential improvements to the current deep mutational scanning approach have been suggested. Assessment of variant function is difficult for proteins for which the function is unknown; selection in those experiments may be limited to assays assessing thermostability or degradation rate of the gene product.[22,23,34] Selection assays used during deep mutational scanning are often specific to a protein and its function being tested. Designing these assays frequently remains a challenge. For example, coupling of cell-based properties such as protein localization with high-throughput sequencing, and might not reflect the complexity of human disease. Paired-end sequencing and inclusion of replicate samples can be used to correct for the average per-base sequencing error rate of ∼1%. Furthermore, inclusion of known completely nonfunctional variants is useful for estimating error rates and can improve the accuracy of fitness estimates.[21] Kowalsky *et al.*[39] developed a standardized protocol to resolve sequence–function relationships for full-length proteins by using a gene tiling technique to divide long gene sequences into different sequencing libraries to overcome the disadvantages of short sequencing reads.

Deep mutational scanning in precision medicine

Functional characterization of genetic variants with the use of deep mutational scanning, in addition to genotype–phenotype association studies, is valuable in diagnosing, treating, and understanding disease risk or prognosis.[18] For example, variants of unknown significance are continuously identified in *BRCA1* in cancer patients by DNA sequencing. Recently, deep mutational scanning was used in a prospective manner to measure the effects of nearly 2,000 missense substitutions in the BARD1 RING domain of *BRCA1* on its E3 ubiquitin ligase activity and binding to this domain. The resulting variant functional scores were used to create a prediction model of variant effect on homology-directed DNA repair. This model will likely improve the interpretation of variants observed in the clinical sequencing of the *BRCA1* gene.[40]

In addition to understanding the function of variants of uncertain significance, it is also important to be able to discriminate driver mutations from passenger mutations in protein domains or entire cancer-related proteins; furthermore, it is also useful to understand the impact of the mutation on the protein's function, its effect on cellular function, and its drugability. The study by Wagenaar *et al.*[41] is another example of how deep mutational scanning can be used in precision medicine. Mutant selection with vemurafenib exposure in mammalian cells and mouse xenografts was used to identify variants in the kinase active site of *BRAF* that are involved in resistance to treatment of *BRAFV600E*-positive melanomas. The kinase activity of the *BRAFV600E/L505H* mutation combination was higher than that of the well-characterized *V600E* mutation alone. The increased kinase activity of the *BRAFV600E/L505H* mutation combination could result in this mutation combination being moderately resistant to mitogen-activated extracellular signal-regulated kinase (MEK) inhibition. Additional crystal structure comparisons suggest that other BRAF inhibitors will be more effective than a MEK inhibitor in eliciting a response in *BRAFV600E/L505H*-containing melanomas. This method could also be applied to other proteins for evaluating resistance to inhibitors.[41]

A deep mutational scanning method was also developed for antibody complementarity-determining regions to simultaneously determine the effect of every possible single amino acid substitution on antigen binding. This method was then applied to a humanized version of the anti-epidermal growth factor receptor antibody cetuximab. Although the majority of complementarity-determining region substitutions are neutral or deleterious to antibody interaction, 67 of the 1,060 tested point mutations increased its affinity. This approach will likely be useful in the future for the development of additional antibody therapies that target cells with specific genetic mutations or variants.[26]

Cystic fibrosis is an autosomal recessive genetic disorder that affects chloride transport. A genetic variant, G551D, exists in the coding region of the *CF transmembrane conductance regulator (CFTR)* gene. The functional consequence of this variant was discovered with the use of high-throughput assays. Although the variant does not impact transport of the CFTR protein to the cell surface, it impairs the ability of the membrane channel to open. This phenotypic effect is associated with abnormalities in the respiratory, endocrine, gastrointestinal, and reproductive systems. In cystic fibrosis patients, this phenotype can be improved with the therapeutic agent ivacaftor. Ivacaftor is a CFTR potentiator because it can alter the activity of the channel by increasing the opening probability and flow of ions. The development of ivacaftor for use in G551D variant carriers provides a further example of how high-throughput functional assays can facilitate identification of actionable drug targets and development of targeted therapeutics.[42,43]

GENETIC VARIANTS WITHIN REGULATORY REGIONS

Endogenous gene expression is controlled by a variety of regulatory regions in the DNA. These regions serve as binding sites for several activator and repressor proteins and RNAs that alter gene expression. Genetic variations in these enhancer elements, transcription factor binding sites, promoter regions, and untranslated regions (UTRs) can alter the binding of these proteins and RNAs leading to changes in gene expression.[44–46] The ultimate level of gene expression is the combination of the effects of all of these binding sites together. According to the 1000 Genomes Project Consortium, the median number of variants among continental population groupings range from 288,000–354,000 variants in enhancers, 748–927 in transcription factor binding sites, 82,000–102,000 in promoters, and 30,000–37,200 in the UTRs per typical human genome (see **Figure 1**). Estimates indicate that ~500,000 of these variant sites are likely to be functional.[1]

Traditionally, the gold standard for studying this process has been to use individual reporter assays that have enhancer elements inserted upstream of a minimal promoter. The strength and effects of the enhancer elements are measured by determining the expression of a reporter gene (e.g., LacZ, luciferase, and green fluorescent protein (GFP)) that is driven by the minimal promoters and enhancers.[47,48] In addition, these assays are also standard assays for assessing the effect of genetic alterations in miRNA binding sites.[45,49] Given the large number of regulatory SNPs that need to be tested, this has led to advances towards more high-throughput functional testing.

High-throughput functional assays

Massively parallel reporter assays are one of the high-throughput functional assays that have been used to assess genetic variants in regulatory regions. For example, using this technique, enhancer sequences containing variations at many positions in the *cis* core regulatory elements were synthesized using programmable microarrays and inserted into a promoter. Tagging barcodes were also inserted into the expressed sequence. These regulatory variants were then transcribed *in vitro* and the expression of each barcode was measured by RNA-sequencing; the expression of each barcode reflected the relative activity of the promoter variants that each barcode was tagging.[50] This method was then modified by synthesizing over 100,000 enhancer variants that were cloned upstream of a minimal promoter luciferase plasmid with the barcode precloned in the 3'UTR to allow for random barcoding. The library of plasmids was injected into mice via tail vein injection. The plasmids are taken up and expressed in the liver and the reporter expression was measured by RNA-seq.[51] Variations of massively parallel reporter assays have been developed and utilized by several groups involving different cell lines and animal models, as well as adaptations to increase the throughput of the assay.[52–60]

CRISPR-Cas9-mediated *in situ* saturating mutagenesis has been used to assess the effects of genetic variants in the *BCL11A* gene enhancer. This gene is a repressor of fetal hemoglobin levels and a therapeutic target for β-hemoglobin disorders. Using the CRISPR-Cas9 nuclease system, they deleted the 12-kb enhancer of *BCL11A* gene using a pair of guide RNAs (gRNAs) to create paired double-strand breaks.[61] To further assess single nucleotide changes and a complete knockout of the enhancer in *BCL11A*, they then synthesized a saturating gRNA library tiling the enhancer region. They disrupted almost all the sequences within the enhancer with Cas9 cleavage and nonhomologous end joining repair. The library was cloned into a lentiviral vector and transduced into HUDEP-2 cells at low multiplicity to achieve a single gRNA per cell. After expansion and differentiation, cells were sorted by fetal hemoglobin levels, which has been previously validated to be regulated by *BCL11A*. DNA was isolated, sequenced, and mapped back to the genome to assess variations in the enhancer region associated with the high- and low-fetal hemoglobin pools.[61]

Multiplexed editing regulatory assay (MERA) is another high-throughput functional screen for genomic variant effects on gene expression. This technique is a CRISPR-Cas9-based mutational screening tool in which adaptations have allowed for one regulatory element to be targeted per cell. This is performed through integration of a single gRNA expression construct into a universally accessible *ROSA* locus of mouse embryonic stem cells. This gRNA expression construct was driven by a U6 promoter driving the expression of a dummy gRNA inserted into stem cells using CRISPR-Cas9-mediated homologous recombination. A library of over 3,900 gRNAs tiling the *cis*-regulatory region was created for each of the four genes of interest, *Nanog*, *Rpp25*, *Tdgf1*, and *Zfp42*. Homologous recombination was used to replace the dummy gRNA with a gRNA from the library which occurred in ~30% of cells to create a functional gRNA expression construct. GFP knock-in lines that were generated for these four stem cell-specific genes were sorted based on GFP expression and deep-sequencing of gRNA-induced mutations were analyzed to assess which mutations induced loss of GFP expression. A linear regression model was developed in order to detect statistically significant gRNAs that are expressed in the different GFP populations using the GFP targeting gRNAs as the positive control and dummy gRNAs as a negative control.[62]

Another high-throughput functional assay that can be adapted to study SNP effects in enhancers is STARR-seq. Self-transcribing active regulatory region sequencing (STARR-seq) involves cloning enhancer elements downstream of a minimal promoter and into the 3'UTR of reporter genes. This ectopic, plasmid-based assay allows for these active enhancers to self-transcribe and become part of the reporter transcripts when transfected into cells. Expression of the transcripts, which include the inserted enhancer sequences, are measured by RNA-seq. This method was first developed and assessed using the *Drosophila melanogaster* genome and has the capacity to identify and quantify enhancer activity in humans.[63] This method has been applied to several enhancer elements, such as hormone responsive enhancers, as well as a modified capture approach (CapSTARR-seq) in which DNA fragments are captured on a custom-designed microarray and cloned into STARR-seq vector.[64,65]

Collectively, there are multiple high-throughput technologies for assessing the impact of genetic variants on regulatory motifs. The diversity of regulatory mechanisms requires that each type of motif has a specific technology. The variations in regulatory elements that alter gene expression can be detected using assays that sort cells with high and low expression of reporter genes, such as the green fluorescent protein. Then the separate pools can be sequenced to determine which variants resulted in the change in activity. Since regulatory domains are the sites of many clinically important genetic variants, these assays will be critical for identifying the variants with functional implications.

GENETIC VARIANTS THAT ALTER SPLICING

The transcribed regions of most eukaryotic genes are made up of introns (noncoding regions) and exons (coding regions). Following the transcription of the DNA into RNA, a diverse group of trans-acting ribonucleoproteins interact with *cis*-sequences in the pre-mRNA to remove the introns and join the exons to form the mature mRNA. This process of mRNA splicing is not a perfect reaction and the majority of human transcripts (~95%) exist in multiple isoforms due to alternative splicing.[66,67] Alternative splicing can have large impacts on the functions of the proteins by altering the amino acid sequences of the translated proteins, the RNA sequences of regulatory RNAs, or the regulatory domains within the RNAs. Core sequences that are involved in splicing are the exon–intron junction (5' splice site), the intron–exon junction (3' splice site), and a branch point within the intron. These sequences determine how frequently a given splice site is used. Complete conservation of splice site sequences is limited to a GU at the 5' end, an AG at the 3' end, and an A at the branch point of the intron. In contrast, there is 35–80% variation in the other positions around the splice junctions and in the branch points that create variability in the effectiveness of the splice site (**Figure 1**). In addition to the core sequences, auxiliary sequences both within the exon and intron influence the effectiveness of a given splice site. Exonic splicing enhancers (ESEs) within the exons and intronic splicing enhancers (ISEs) within the introns increase the probability of an adjacent splice site being used. Conversely, exonic splicing silencers (ESSs) and intronic splicing silencers (ISSs) suppress splicing.[68,69]

Genetic variations within any of the core or auxiliary sequences can influence splice site selection.[70,71] Since the alternatively spliced transcripts can function differently than the normal transcripts, it is not surprising that variations in splice junctions can impact many phenotypes. About 14% of hereditary diseases are caused by SNPs in noncoding regions, 90% of which are due to splice altering variations.[72] Approximately 30% of disease-causing variants in HGMD are due to variants that perturb normal splicing; around 15% are due to mutations within the conserved dinucleotide motifs.[73] Variants within the highly conserved dinucleotide motifs at the beginning and end of the introns reduce the spliceosome binding and reduce the mRNA splicing.[71] There have also been several reports of diseases attributed to variants around branch points,[74–76] although the identification of disease-causing variants in the branch points has been limited by the low number of known branch point sequences.[77] The majority of the splice-altering variants are located in auxiliary splice sequences. The resulting changes in the levels of aberrant transcripts or the ratio of splice variants have been implicated in a variety of diseases.[78]

Using next-generation sequencing and array-based genome wide screens, large numbers of variants have been associated with clinical phenotypes. Many of these variants are located in noncoding regions of the genome and are predicted to alter splicing. Understanding their role in altering clinical phenotypes requires understanding the function of the variants. However, functionally testing thousands of variants individually is not practical due to the time and resources needed. Typically, computational methods are employed on identified variants to predict functional significance or causality. Examples of widely used computational methods are reviewed by Soemedi *et al*.[79] In addition, newer methods, such as ExonImpact are also continually being developed.[80] Most algorithms that predict the functional significance of SNPs in splice sites scan for disruption of protein-binding motifs in the RNA. RNA-protein binding predictions are similar to DNA-protein predictions, but RNA also has a secondary structure element that may result in altered accessibility and protein binding at sites far away from the site of the variation.[81]

Although the rapidly evolving computational methods help narrow down potential functional variants from large data sets, experimental approaches to test these predictions are important to validate the predictions. While human *in vivo* studies provide the best system in terms of the context of the physiological background, the results may be difficult to interpret amidst the other complex genetic variability and environmental factors in humans. In addition, some of the variants are rare, so it can be difficult to find subjects that carry the variants of interest. Lastly, the analysis of the mRNA splicing often requires tissue samples to evaluate tissue-specific mRNA splicing that are too invasive to be biopsied. Thus, *in vitro* bioassays are most commonly used to determine the impact of variants in splice junctions.

Mini-gene reporter assays are the gold standard for functionally testing sequences and variants predicted to alter splicing.[82] In this assay, a target genomic region containing a *cis*-sequence is inserted between known splicing reporters (usually exons) within an expression plasmid. The ability of the target region to induce splicing is measured by expressing the transcript either *in vitro* (in a nuclear extract) or *in vivo* (transfected into cells) and determining the frequency of splicing in the mature mRNA transcript. The spliced transcripts are usually detected by quantitative PCR (qPCR) assays, which detect specific splice products and lariat fragments. By comparing the splicing of wildtype and variant sequences, one can determine the functional impact of the genetic variants on mRNA splicing. This assay can be used to compare allele-specific splicing by comparing the splice-products of the wildtype and variant version of the test sequence. Although this assay is useful for low-throughput testing and validation of high-throughput assays, it is too resource-intensive to individually test hundreds or thousands of variants.

Since most functional variants in splice-junctions result in gain or loss of RNA–protein interactions, assays that measure RNA–protein interactions can also be used to functionally test variants in splice junctions. There are several low-throughput assays that detect differences in protein–RNA interactions. Some examples include electrophoretic mobility shift assay (EMSA), nitrocellulose membrane binding assay, and immunoprecipitation methods. These assays are commonly performed as *in vitro* assays that involve incubation of nucleic acid constructs with a known protein or a protein pool followed by the separation of protein-bound from free nucleic acids. Although these techniques are useful in low-throughput assays to determine RNA–protein interactions, they will need to be modified to be employed in high-throughput applications. These assays may provide some insights into the mechanism of altered splicing; however, they cannot replace direct measures of mRNA splicing. These assays may be more useful after the functional significance of a variant has been established by assays, such as the mini-gene reporter assay.

High-throughput functional assays

With high volumes of data being generated through genomic technologies and shared through databases, the use of low-throughput assays are a major bottleneck in testing large numbers of variants in splice junctions. There are thousands of variants that have been computationally predicted to alter splicing. Many of those variants are in high linkage disequilibrium with other potentially functional variants, making it necessary to test large numbers of variants. High-throughput bioassays that isolate individual variants and test for intrinsic splicing differences in alleles are critical in identifying the causal variants. Once functional variants are identified, investigation into the mechanism of action of these variants will also require high-throughput methods. Although significant strides have been made in computational methods that predict these functional variants, progress towards developing high-throughput experimental approaches have been lacking until recently.

Pioneering work has recently been done to develop high-throughput functional assays where several low-throughput functional assays were modified into cost-effective high-throughput bioassays. These high-throughput methods use pooled-oligonucleotides containing the target sequences that are synthesized in highly parallel reactions on arrays. Those pools were originally generated by heat-treating custom oligonucleotide arrays.[83] More recently, pools of oligos are now commercially available where thousands of custom oligonucleotides ranging from 10–200 bp in length are electrochemically synthesized on an array and chemically released into a tube to create the custom pool (e.g., OligoMix, LC Biosciences, Houston, TX, USA). Such pools have been used in massively parallel reporter assays, as capture probes, CRISPR-Cas9 applications, and for gene synthesis.[84,85] Pooled-oligonucleotide based assays are ideal for testing the functional significance of SNPs and short indels where the test sequences are of a similar length. In high-throughput studies investigating variants in splice motifs, these pooled oligonucleotides have been used in modified mini-gene splicing assay,[79] *in vitro* splicing assays,[83] and in RNA-binding assays.[86]

In a modified mini-gene reporter assay, a large number of target sequences are inserted into the mini-gene reporter assay to create a pool of reporter plasmids. The targets are synthesized as a pool of single-stranded DNA oligonucleotides that contain the target sequence along with the endogenous flanking region ($\leq$200 nucleotides total length). For each target site, a wildtype and variant version of the oligonucleotide is included in the pool. They are synthesized with universal primer binding sequences on each end, which are used to amplify the targets by PCR. In a single reaction, the oligonucleotides are cloned into the mini-gene vector to create a pool of plasmids that are amplified in bacteria. The pool of plasmids is transfected into cultured cells to be transcribed and spliced. Spliced and unspliced products are quantified by RNA-seq. Following the normalization to the amount of each plasmid in the pool, the ratio of spliced and unspliced product is compared between the wildtype and variant sequences; this determines the effect of the variant on the splicing effectiveness. A limitation to this methodology is the poor representation of the input sequences in the final pool of plasmids.[87] Soemeidi *et al.* found that only $<$35% of the allele pairs in the oligo pool was represented in plasmids pool.[79] An alternate approach is the use of PCR ligation of oligo fragments with overlap to create a PCR library that contains all the intrinsic features required for transcription and splicing.

Another approach to testing splice sites is to use an *in vitro* splicing assay.[88,89] This high-throughput splicing assay uses a pool of pre-mRNAs that have been constructed by synthesizing oligonucleotides with the mini-gene component and target splicing sequence preceded by an upstream T7 promoter. The oligonucleotides are made double stranded by PCR and transcribed in the presence of α^{32}-UTP. The radiolabeled pre-mRNA pool is incubated in a splicing-competent nuclear extract.[90] The products of the splicing reactions are the free 5' exon and lariat intermediate from the first splicing step, and the spliced exons and the shorter lariat product from the second splicing step. To identify and isolate each product, the α^{32}-UTP products are separated by denaturing polyacrylamide gel electrophoresis. The RNA products are purified and analyzed by RNA-seq to confirm the quantity of splicing. This method not only identifies differences in splicing between wildtype and variant sequences, but it also helps identify the steps of splicing that are affected by the variation.

These methods of testing genetic variants for altered splicing also lend themselves well to determining if the mechanism is related to altered protein binding. Since the major reason for altered splicing effectiveness is altered binding to spliceosome proteins, the following approach can be used to identify which proteins may be impacted. This is accomplished using a modified binding motif detection assay, SELEX,[91] to develop a protein binding assay where a pool of RNAs are tested for their ability to bind protein.[83] The pool of double-stranded oligonucleotides, as described above, is transcribed *in vitro* and the resulting RNA pool is incubated with nuclear extracts to facilitate RNA–protein interactions. The RNA fragments that are bound to proteins are isolated by physically separating the bound and unbound fractions. This can be done with a nonspecific method, such as nitrocellulose based protein binding or for specific proteins by immunoprecipitation. The RNA populations in the bound and unbound fractions are reverse-transcribed and analyzed by

RNA-seq. Differences in the binding of wildtype and variant RNAs in the bound fraction indicates altered RNA-protein interactions.

Therapies that target alternative splicing

The impact of aberrant splicing on a variety of human health parameters has stimulated the pursuit of individualized therapies that specifically target the splicing process. Variants in *cis*-motifs that affect nearby splicing events are likely the best candidate targets for designing therapies. Most defects in splicing are due to aberrant splice site selection as a result of altered spliceosome-factor binding. Therapeutic approaches aim to restore normal splice site selection by blocking the interaction between cryptic sites and the spliceosome.[92] Modified oligonucleotides and RNA binding small molecules have been used with some success in animal models and are currently under clinical investigation.[93] For example, the most common form of spinal muscular atrophy is caused by the deficiency of the SMN protein; the deficiency is due to a splice site mutation that results in the skipping of exon 7. 2′-O methyl phosphothioate-modified oligonucleotides that bind to hnRNP A1 (a splicing silencer) binding site, leads to the inclusion of exon 7 and, thus, increase the amount of functional SMN protein.[94] Subcutaneous administration of the modified oligo has been shown to be effective for an extended period of time in animal models.[94] Similarly, partial rescue of aberrant splicing that leads to Duchenne muscular atrophy has been observed using oligonucleotides that target dystrophin pre-mRNA.[93] This approach of using oligonucleotides could also be expanded by the addition of splicing regulatory sequences or factors to antisense targeting.[95]

The therapeutic potential of small molecule regulators of splicing has also been explored with significant success. Using high-throughput drug screens, compounds that promote the inclusion of exon 7 of SMN protein have been identified.[96] One of the compounds stabilizes the interaction between a core spliceosomal small nuclear ribonucleoprotein U1 at the SMN exon 7/intron junction. Other compounds such as kinetin,[97] cardiac glycosides,[98] and RECTAS[99] have been shown to improve the recognition of mutated splice sites in the IKBKAP gene involved in familial dysautonomia. The improved recognition in both of these genes is proposed to be due to stabilization of base paring between the binding site and the RNA component of the ribonucleoprotein.

Spliceosome-mediated RNA trans-splicing (SMaRT) therapies are another approach to treat unwanted splicing aberrations. These are compounds that modify the secondary structure of pre-mRNAs and, thereby, regulate splicing factor accessibility.[100] Trans-splicing is a process that is observed in a variety of organisms ranging from protozoa to mammalian cells. In this process, exons from different pre-mRNAs are spliced together to generate a single transcript. As a therapeutic approach, splicing aberrations that result in a nonfunctional protein are repaired by inducing trans-splicing of endogenous mutated pre-mRNA with an exogenous pre-trans-splicing molecule. The exogenous molecule contains the desirable sequence to replace the aberration, resulting in a chimeric transcript that encodes for a functional protein. Despite poor *in vivo* efficacy of the trans-splicing process, therapies towards treating diseases such as cystic fibrosis,[101] spinal muscular atrophy,[102] Duchenne muscular atrophy,[103] and retinitis pigmentosa[104] are being tested.

CONCLUSIONS AND FUTURE DIRECTIONS

Genome sequencing and genotyping technologies have uncovered enormous genetic variation in the human population. Every person has a large number of variants, but the minor allele frequencies of most individual variants are relatively rare. In addition, many new germline variants are created in every individual. This low frequency of many of the variants makes population genotype–phenotype associations impractical for most variants if their functional impact is not known. Thus, for the majority of the genetic variants, an important first step towards translating this information into clinically usable information is to determine the impact of the variant on the function of the gene product. For genes with known functions and clinical utility, these variants can then be used to guide risk assessment and therapies based on their effect on the host gene. By understanding the functional alterations in genes that have already been associated with clinical phenotypes, this may also help understand the etiology of specific phenotypes and, thus, lead to future curative therapies. For genes that are being tested in clinical association studies, the rare variants within a gene can be grouped together using gene activity scores that can be used for associations with the clinical phenotypes.

In order to functionally classify variants and assign activity scores to the large number of variants that currently have unknown functional consequences, we need more and better high-throughput functional assays. As we have described in this review, there are several relatively high-throughput assays for a variety of functional assays. However, many of them still lack the scalability needed to assess the large number of variants and do it economically. Advances are also needed to improve accuracy and turnaround time for many of them, so as new variants are discovered, they can be clinically implemented. Furthermore, there are many functions that yet still do not have high-throughput assays. For those assays that do exist, and as new ones are developed, centralized databases would be useful to simplify the collection and comparison of data from multiple laboratories and to make the functional data easily accessible by others. These improvements will be critical to maximize the clinical utility of the large amount of existing genomic data.

Acknowledgments. We apologize to fellow researchers whose work could not be directly referenced due to space and reference limitations. This work was supported by R01GM0088076 (to J.I., K.S.B., T.C.S.), R01CA213466 (to T.C.S.), U01HG007762 (to T.C.S.), Vera Bradley Foundation for Breast Cancer (to J.I., M.S., T.C.S.), and F31GM119401 (to K.S.B.).

Author contributions. J.I., M.S., K.S.B., and T.C.S. wrote the article. The first three authors contributed equally to this work.

Conflict of interest. The authors declared no conflicts of interest.

1. The 1000 Genomes Project Consortium. *et al.* A global reference for human genetic variation. *Nature.* **7571**, 68–74 (2015).
2. Tennessen, J.A. *et al.* Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science.* **6090**, 64–69 (2012).
3. Sudmant, P.H. *et al.* An integrated map of structural variation in 2,504 human genomes. *Nature.* **7571**, 75–81 (2015).
4. Scally, A. The mutation rate in human evolution and demographic inference. *Curr. Opin. Genet. Dev.* 36–43 (2016).
5. Welter, D. *et al.* The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res.* **Database issue**, D1001–1006 (2014).
6. Gusev, A. *et al.* Partitioning heritability of regulatory and cell-type-specific variants across 11 common diseases. *Am. J. Hum. Genet.* **5**, 535–552 (2014).
7. Sham, P.C. & Purcell, S.M. Statistical power and significance testing in large-scale genetic studies. *Nat. Rev. Genet.* **5**, 335–346 (2014).
8. Moutsianas, L. *et al.* The power of gene-based rare variant methods to detect disease-associated variation and test hypotheses about complex disease. *PLoS Genet.* **4**, e1005165 (2015).
9. Fujikura, K., Ingelman-Sundberg, M. & Lauschke, V.M. Genetic variation in the human cytochrome P450 supergene family. *Pharmacogenet. Genomics.* **12**, 584–594 (2015).
10. Bush, W.S. *et al.* Genetic variation among 82 pharmacogenes: The PGRNseq data from the eMERGE network. *Clin. Pharmacol. Ther.* **2**, 160–169 (2016).
11. Stenson, P.D. *et al.* The Human Gene Mutation Database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine. *Hum. Genet.* **1**, 1–9 (2014).
12. Xue, Y., Ankala, A., Wilcox, W.R. & Hegde, M.R. Solving the molecular diagnostic testing conundrum for Mendelian disorders in the era of next-generation sequencing: single-gene, gene panel, or exome/genome sequencing. *Genet. Med.* **6**, 444–451 (2015).
13. Borges, S. *et al.* Composite functional genetic and comedication CYP2D6 activity score in predicting tamoxifen drug exposure among breast cancer patients. *J Clin. Pharmacol.* **4**, 450–458 (2010).
14. Gaedigk, A. *et al.* The CYP2D6 activity score: translating genotype information into a qualitative measure of phenotype. *Clin. Pharmacol. Ther.* **2**, 234–242 (2008).
15. Anfinsen, C.B. Principles that govern the folding of protein chains. *Science.* **4096**, 223–230 (1973).
16. Grantham, R. Amino acid difference formula to help explain protein evolution. *Science.* **4154**, 862–864 (1974).
17. Ng, P.C. & Henikoff, S. Predicting deleterious amino acid substitutions. *Genome Res.* **5**, 863–874 (2001).
18. Young, D.L. & Fields, S. The role of functional data in interpreting the effects of genetic variation. *Mol Biol. Cell.* **22**, 3904–3908 (2015).
19. Tang, H. & Thomas, P.D. Tools for predicting the functional impact of nonsynonymous genetic variation. *Genetics.* **2**, 635–647 (2016).
20. Tang, H. & Thomas, P.D. PANTHER-PSEP: predicting disease-causing genetic variants using position-specific evolutionary preservation. *Bioinformatics.* **14**, 2230–2232 (2016).
21. Araya, C.L. *et al.* A fundamental protein property, thermodynamic stability, revealed solely from large-scale measurements of protein function. *Proc. Natl. Acad. Sci. U. S. A.* **42**, 16858–16863 (2012).
22. Fowler, D.M. & Fields, S. Deep mutational scanning: a new style of protein science. *Nat. Methods.* **8**, 801–807 (2014).
23. Fowler, D.M., Stephany, J.J. & Fields, S. Measuring the activity of protein variants on a large scale using deep mutational scanning. *Nat. Protoc.* **9**, 2267–2284 (2014).
24. Araya, C.L. & Fowler, D.M. Deep mutational scanning: assessing protein function on a massive scale. *Trends Biotechnol.* **9**, 435–442 (2011).
25. Mingo, J. *et al.* One-tube-only standardized site-directed mutagenesis: an alternative approach to generate amino acid substitution collections. *PLoS One.* **8**, e0160972 (2016).
26. Forsyth, C.M. *et al.* Deep mutational scanning of an antibody against epidermal growth factor receptor using mammalian cell display and massively parallel pyrosequencing. *MAbs.* **4**, 523–532 (2013).
27. Dai, M. *et al.* Using T7 phage display to select GFP-based binders. *Protein Eng. Des. Sel.* **7**, 413–424 (2008).
28. Levin, A.M. & Weiss, G.A. Optimizing the affinity and specificity of proteins with molecular display. *Mol. Biosyst.* **1**, 49–57 (2006).
29. Mattheakis, L.C., Bhatt, R.R. & Dower, W.J. An in vitro polysome display system for identifying ligands from very large peptide libraries. *Proc. Natl. Acad. Sci. U. S. A.* **19**, 9022–9026 (1994).
30. Smith, G.P. Filamentous fusion phage: novel expression vectors that display cloned antigens on the virion surface. *Science.* **4705**, 1315–1317 (1985).
31. Boucher, J.I. *et al.* Viewing protein fitness landscapes through a next-gen lens. *Genetics.* **2**, 461–471 (2014).
32. Zhao, G. *et al.* Structural and mutational studies on the importance of oligosaccharide binding for the activity of yeast PNGase. *Glycobiology.* **2**, 118–125 (2009).
33. Starita, L.M. *et al.* Activity-enhancing mutations in an E3 ubiquitin ligase identified by high-throughput mutagenesis. *Proc. Natl. Acad. Sci. U. S. A.* **14**, E1263–1272 (2013).
34. Shin, H. & Cho, B.K. Rational Protein Engineering Guided by Deep Mutational Scanning. *Int. J. Mol. Sci.* **9**, 23094–23110 (2015).
35. Fowler, D.M., Araya, C.L., Gerard, W. & Fields, S. Enrich: software for analysis of protein function by enrichment and depletion of variants. *Bioinformatics.* **24**, 3430–3431 (2011).
36. Bloom, J.D. Software for the analysis and visualization of deep mutational scanning data. *BMC Bioinformatics.* **168**. (2015).
37. Hemani, G. *et al.* Detection and replication of epistasis influencing transcription in humans. *Nature.* **7495**, 249–253 (2014).
38. Wu, N.C., Dai, L., Olson, C.A., Lloyd-Smith, J.O. & Sun, R. Adaptation in protein fitness landscapes is facilitated by indirect paths. *Elife.* (2016).
39. Kowalsky, C.A. *et al.* High-resolution sequence-function mapping of full-length proteins. *PLoS One.* **3**, e0118193 (2015).
40. Starita, L.M. *et al.* Massively parallel functional analysis of BRCA1 RING domain variants. *Genetics.* **2**, 413–422 (2015).
41. Wagenaar, T.R. *et al.* Resistance to vemurafenib resulting from a novel mutation in the BRAFV600E kinase domain. *Pigment Cell Melanoma Res.* **1**, 124–133 (2014).
42. Condren, M.E. & Bradshaw, M.D. Ivacaftor: a novel gene-based therapeutic approach for cystic fibrosis. *J. Pediatr. Pharmacol. Ther.* **1**, 8–13 (2013).
43. Ramsey, B.W. *et al.* A CFTR potentiator in patients with cystic fibrosis and the G551D mutation. *N. Engl. J. Med.* **18**, 1663–1672 (2011).
44. Chatterjee, S. *et al.* Enhancer variants synergistically drive dysfunction of a gene regulatory network in Hirschsprung disease. *Cell.* **2**, 355–368 e310 (2016).
45. Ramamoorthy, A. *et al.* In silico and in vitro identification of microRNAs that regulate hepatic nuclear factor 4alpha expression. *Drug Metab. Dispos.* **4**, 726–733 (2012).
46. Jin, Y. *et al.* MicroRNA hsa-miR-25-3p suppresses the expression and drug induction of CYP2B6 in human hepatocytes. *Biochem. Pharmacol.* 88–96. (2016).
47. Davidson, S. *et al.* Differential activity by polymorphic variants of a remote enhancer that supports galanin expression in the hypothalamus and amygdala: implications for obesity, depression and alcoholism. *Neuropsychopharmacology.* **11**, 2211–2221 (2011).
48. Smemo, S. *et al.* Regulatory variation in a TBX5 enhancer leads to isolated congenital heart disease. *Hum. Mol. Genet.* **14**, 3255–3263 (2012).
49. Wu, S. *et al.* MicroRNA-137 inhibits EFNB2 expression affected by a genetic variant and is expressed aberrantly in peripheral blood of schizophrenia patients. *EBioMedicine.* (2016).
50. Patwardhan, R.P. *et al.* High-resolution analysis of DNA regulatory elements by synthetic saturation mutagenesis. *Nat. Biotechnol.* **12**, 1173–1175 (2009).
51. Patwardhan, R.P. *et al.* Massively parallel functional dissection of mammalian enhancers in vivo. *Nat Biotechnol.* **3**, 265–270 (2012).
52. Melnikov, A. *et al.* Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. *Nat. Biotechnol.* **3**, 271–277 (2012).
53. Smith, R.P. *et al.* Massively parallel decoding of mammalian regulatory sequences supports a flexible organizational model. *Nat. Genet.* **9**, 1021–1028 (2013).
54. Kwasnieski, J.C., Mogno, I., Myers, C.A., Corbo, J.C. & Cohen, B.A. Complex effects of nucleotide variants in a mammalian cis-regulatory element. *Proc. Natl. Acad. Sci. U. S. A.* **47**, 19498–19503 (2012).
55. Kheradpour, P. *et al.* Systematic dissection of regulatory motifs in 2000 predicted human enhancers using a massively parallel reporter assay. *Genome Res.* **5**, 800–811 (2013).
56. White, M.A., Myers, C.A., Corbo, J.C. & Cohen, B.A. Massively parallel in vivo enhancer assay reveals that highly local features determine the cis-regulatory function of ChIP-seq peaks. *Proc. Natl. Acad. Sci. U. S. A.* **29**, 11952–11957 (2013).
57. Mogno, I., Kwasnieski, J.C. & Cohen, B.A. Massively parallel synthetic promoter assays reveal the in vivo effects of binding site variants. *Genome Res.* **11**, 1908–1915 (2013).
58. Kamps-Hughes, N., Preston, J.L., Randel, M.A. & Johnson, E.A. Genome-wide identification of hypoxia-induced enhancer regions. *PeerJ.* e1527. (2015).
59. Tewhey, R. *et al.* Direct identification of hundreds of expression-modulating variants using a multiplexed reporter assay. *Cell.* **6**, 1519–1529 (2016).
60. Ulirsch, J.C. *et al.* Systematic functional dissection of common genetic variation affecting red blood cell traits. *Cell.* **6**, 1530–1545 (2016).
61. Canver, M.C. *et al.* BCL11A enhancer dissection by Cas9-mediated in situ saturating mutagenesis. *Nature.* **7577**, 192–197 (2015).
62. Rajagopal, N. *et al.* High-throughput mapping of regulatory DNA. *Nat. Biotechnol.* **2**, 167–174 (2016).
63. Arnold, C.D. *et al.* Genome-wide quantitative enhancer activity maps identified by STARR-seq. *Science.* **6123**, 1074–1077 (2013).
64. Vanhille, L. *et al.* High-throughput and quantitative assessment of enhancer activity in mammals by CapStarr-seq. *Nat. Commun.* 6905. (2015).
65. Shlyueva, D. *et al.* Hormone-responsive enhancer-activity maps reveal predictive motifs, indirect repression, and targeting of closed chromatin. *Mol. Cell.* **1**, 180–192 (2014).
66. Pan, Q., Shai, O., Lee, L.J., Frey, B.J. & Blencowe, B.J. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat. Genet.* **12**, 1413–1415 (2008).
67. Wang, E.T. *et al.* Alternative isoform regulation in human tissue transcriptomes. *Nature.* **7221**, 470–476 (2008).
68. Martinez-Contreras, R. *et al.* Intronic binding sites for hnRNP A/B and hnRNP F/H proteins stimulate pre-mRNA splicing. *PLoS Biol.* **2**, e21 (2006).
69. Kanopka, A., Muhlemann, O. & Akusjarvi, G. Inhibition by SR proteins of splicing of a regulated adenovirus pre-mRNA. *Nature.* **6582**, 535–538 (1996).

70. Sun, H. & Chasin, L.A. Multiple splicing defects in an intronic false exon. *Mol. Cell. Biol.* **17**, 6414–6425 (2000).
71. Roca, X. *et al.* Widespread recognition of 5′ splice sites by noncanonical base-pairing to U1 snRNA involving bulged nucleotides. *Genes Dev.* **10**, 1098–1109 (2012).
72. Stenson, P.D. *et al.* Human Gene Mutation Database (HGMD): 2003 update. *Hum. Mutat.* **6**, 577–581 (2003).
73. Krawczak, M. *et al.* Single base-pair substitutions in exon-intron junctions of human genes: nature, distribution, and consequences for mRNA splicing. *Hum. Mutat.* **2**, 150–158 (2007).
74. Crotti, L. *et al.* A KCNH2 branch point mutation causing aberrant splicing contributes to an explanation of genotype-negative long QT syndrome. *Heart Rhythm.* **2**, 212–218 (2009).
75. Maslen, C., Babcock, D., Raghunath, M. & Steinmann, B. A rare branch-point mutation is associated with missplicing of fibrillin-2 in a large family with congenital contractural arachnodactyly. *Am. J. Hum. Genet.* **6**, 1389–1398 (1997).
76. Webb, J.C., Patel, D.D., Shoulders, C.C., Knight, B.L. & Soutar, A.K. Genetic variation at a splicing branch point in intron 9 of the low density lipoprotein (LDL)-receptor gene: a rare mutation that disrupts mRNA splicing in a patient with familial hypercholesterolaemia and a common polymorphism. *Hum. Mol. Genet.* **9**, 1325–1331 (1996).
77. Taggart, A.J., DeSimone, A.M., Shih, J.S., Filloux, M.E. & Fairbrother, W.G. Large-scale mapping of branchpoints in human pre-mRNA transcripts in vivo. *Nat. Struct. Mol. Biol.* **7**, 719–721 (2012).
78. O'Rourke, J.R. & Swanson, M.S. Mechanisms of RNA-mediated disease. *J. Biol. Chem.* **12**, 7419–7423 (2009).
79. Soemedi, R., Vega, H., Belmont, J.M., Ramachandran, S. & Fairbrother, W.G. Genetic variation and RNA binding proteins: tools and techniques to detect functional polymorphisms. *Adv. Exp. Med. Biol.* 227–266 (2014).
80. Li, M. *et al.* ExonImpact: Prioritizing Pathogenic Alternative Splicing Events. *Hum. Mutat.* (2016).
81. Long, D. *et al.* Potent effect of target structure on microRNA function. *Nat. Struct. Mol. Biol.* **4**, 287–294 (2007).
82. Cooper, T.A. Use of minigene systems to dissect alternative splicing elements. *Methods.* **4**, 331–340 (2005).
83. Reid, D.C. *et al.* Next-generation SELEX identifies sequence and structural determinants of splicing factor binding in human pre-mRNA sequence. *RNA.* **12**, 2385–2397 (2009).
84. Diao, Y. *et al.* A new class of temporarily phenotypic enhancers identified by CRISPR/Cas9-mediated genetic screening. *Genome Res.* **3**, 397–405 (2016).
85. Malina, A. *et al.* Adapting CRISPR/Cas9 for functional genomics screens. *Methods Enzymol.* 193–213. (2014).
86. Watkins, K.H., Stewart, A. & Fairbrother, W. A rapid high-throughput method for mapping ribonucleoproteins (RNPs) on human pre-mRNA. *J. Vis. Exp.* **34**. (2009).
87. Ke, S. *et al.* Quantitative evaluation of all hexamers as exonic splicing elements. *Genome Res.* **8**, 1360–1374 (2011).
88. Ruskin, B., Krainer, A.R., Maniatis, T. & Green, M.R. Excision of an intact intron as a novel lariat structure during pre-mRNA splicing in vitro. *Cell.* **1**, 317–331 (1984).
89. Lin, R.J., Newman, A.J., Cheng, S.C. & Abelson, J. Yeast mRNA splicing in vitro. *J. Biol. Chem.* **27**, 14780–14792 (1985).
90. Krainer, A.R., Maniatis, T., Ruskin, B. & Green, M.R. Normal and mutant human beta-globin pre-mRNAs are faithfully and efficiently spliced in vitro. *Cell.* **4**, 993–1005 (1984).
91. Stoltenburg, R., Reinemann, C. & Strehlitz, B. SELEX–a (r)evolutionary method to generate high-affinity nucleic acid ligands. *Biomol. Eng.* **4**, 381–403 (2007).
92. Svasti, S. *et al.* RNA repair restores hemoglobin expression in IVS2-654 thalassemic mice. *Proc. Natl. Acad. Sci. U. S. A.* **4**, 1205–1210 (2009).
93. Disterer, P. *et al.* Development of therapeutic splice-switching oligonucleotides. *Hum. Gene Ther.* **7**, 587–598 (2014).
94. Hua, Y. *et al.* Peripheral SMN restoration is essential for long-term rescue of a severe spinal muscular atrophy mouse model. *Nature.* **7367**, 123–126 (2011).
95. Owen, N. *et al.* Design principles for bifunctional targeted oligonucleotide enhancers of splicing. *Nucleic Acids Res.* **16**, 7194–7208 (2011).
96. Palacino, J. *et al.* SMN2 splice modulators enhance U1-pre-mRNA association and rescue SMA mice. *Nat. Chem. Biol.* **7**, 511–517 (2015).
97. Axelrod, F.B. *et al.* Kinetin improves IKBKAP mRNA splicing in patients with familial dysautonomia. *Pediatr. Res.* **5**, 480–483 (2011).
98. Liu, L.J. *et al.* [Association of CFTR gene polymorphism with congenital bilateral absence of vas deferens in ethnic Han Chinese patients]. Zhonghua Yi Xue Yi Chuan Xue Za Zhi. **6**, 729–732 (2013).
99. Yoshida, M. *et al.* Rectifier of aberrant mRNA splicing recovers tRNA modification in familial dysautonomia. *Proc. Natl. Acad. Sci. U. S. A.* **9**, 2764–2769 (2015).
100. Zheng, S., Chen, Y., Donahue, C.P., Wolfe, M.S. & Varani, G. Structural basis for stabilization of the tau pre-mRNA splicing regulatory element by novantrone (mitoxantrone). *Chem. Biol.* **5**, 557–566 (2009).
101. Liu, X. *et al.* Partial correction of endogenous DeltaF508 CFTR in human cystic fibrosis airway epithelia by spliceosome-mediated RNA trans-splicing. *Nat. Biotechnol.* **1**, 47–52 (2002).
102. Coady, T.H., Shababi, M., Tullis, G.E. & Lorson, C.L. Restoration of SMN function: delivery of a trans-splicing RNA re-directs SMN2 pre-mRNA splicing. *Mol. Ther.* **8**, 1471–1478 (2007).
103. Lorain, S. *et al.* Dystrophin rescue by trans-splicing: a strategy for DMD genotypes not eligible for exon skipping approaches. *Nucleic Acids Res.* **17**, 8391–8402 (2013).
104. Berger, A. *et al.* mRNA trans-splicing in gene therapy for genetic diseases. *Wiley Interdiscip. Rev. RNA.* **4**, 487–498 (2016).