ANNUAL REVIEWS

Annual Review of Genomics and Human Genetics

Massively Parallel Assays and Quantitative Sequence–Function Relationships

Justin B. Kinney and David M. McCandlish

Simons Center for Quantitative Biology, Cold Spring Harbor Laboratory, Cold Spring Harbor, New York 11724, USA; email: jkinney@cshl.edu, mccandlish@cshl.edu

Annu. Rev. Genom. Hum. Genet. 2019. 20:99–127

First published as a Review in Advance on May 15, 2019

The *Annual Review of Genomics and Human Genetics* is online at genom.annualreviews.org

https://doi.org/10.1146/annurev-genom-083118-014845

Keywords

genotype–phenotype map, epistasis, variants of uncertain significance, biophysical modeling, *cis*-regulatory grammar, deep learning

Abstract

Over the last decade, a rich variety of massively parallel assays have revolutionized our understanding of how biological sequences encode quantitative molecular phenotypes. These assays include deep mutational scanning, high-throughput SELEX, and massively parallel reporter assays. Here, we review these experimental methods and how the data they produce can be used to quantitatively model sequence–function relationships. In doing so, we touch on a diverse range of topics, including the identification of clinically relevant genomic variants, the modeling of transcription factor binding to DNA, the functional and evolutionary landscapes of proteins, and *cis*-regulatory mechanisms in both transcription and mRNA splicing. We further describe a unified conceptual framework and a core set of mathematical modeling strategies that studies in these diverse areas can make use of. Finally, we highlight key aspects of experimental design and mathematical modeling that are important for the results of such studies to be interpretable and reproducible.

1. INTRODUCTION

Deciphering the genetic code was one of the crowning achievements of the early days of molecular biology (64). By cataloging which amino acids are encoded by each of the 64 possible codons, Nirenberg et al. (101) and others were able to complete the mapping from DNA to mRNA to protein sequence. Subsequently, sequence elements that mark the initiation of transcription (117) and translation (145) were identified. By the mid-1970s, the budding field of molecular biology had good reason to hope that a full determination of how genomic sequence encodes molecular function would proceed along similar lines: One could simply catalog all of the discrete sequence elements that have biological function, then locate where in the genome these functional elements occur.

But nature has not been so eager to reveal the genome's secrets. And in retrospect, the genetic code appears aberrantly simple. Unlike the genetic code, which is a discrete mapping from codons to amino acids, the biological codes governing many other aspects of how cells and organisms work are fundamentally continuous, in the sense that variant alleles produce a graded array of molecular phenotypes. Here, we use the term molecular phenotype in a very broad sense to describe any physical, chemical, or biological property affected by a sequence of interest. The need to understand how sequence encodes the quantitative values of molecular phenotypes arises in a diverse set of problems. For example:

- Predicting which variants in the human genome are likely to be pathogenic requires a comprehensive and quantitative understanding of the molecular phenotypes produced by mutation (152).
- Transcription factors (TFs) are proteins that regulate gene expression by binding to specific sites encoded in genomic DNA. Understanding the sequence specificities of TFs—that is, which sites a TF will bind and how strong this binding will be—requires quantitative models that integrate sequence information across the length of each candidate (77).
- Proteins typically have multiple molecular phenotypes, including folding energy, enzymatic activity, and expression level. Understanding how protein sequence governs these molecular phenotypes is complicated by the fact that the peptide chain of a protein typically folds into a specific three-dimensional structure, resulting in interactions between amino acids that are distant in the primary sequence. The study of protein sequence–function relationships (42) thus presents experimental and modeling challenges beyond those encountered in the study of TF–DNA binding.
- *Cis*-regulatory elements (CREs) are genomic sequences that control gene expression. This control can take place at a variety of steps, including transcript initiation, pre-mRNA splicing, and mRNA decay. CREs typically function by binding multiple regulatory proteins or other *trans* factors, such as small RNAs, at multiple distinct binding sites. Quantitative models are needed to describe how interactions between these *trans* factors give rise to CRE activity, but an understanding of the rules that govern these interactions remains largely elusive (79).

Over the last decade, a rich variety of experimental methods based on high-throughput DNA sequencing have been developed for studying these and other quantitative sequence–function relationships. These methods, which we collectively refer to as massively parallel assays, include high-throughput SELEX (HT-SELEX) methods for measuring protein–DNA and protein–RNA binding (77), deep mutational scanning (DMS) methods for determining the molecular phenotypes of protein variants (42), and massively parallel reporter assays (MPRAs) for studying CREs of different types (79). These experimental methods are revolutionizing our understanding

Molecular phenotype: a broad term describing any physical, chemical, or biological property that is affected by a biological sequence of interest

Quantitative model: in the context of this review, a mathematical function that takes a biological sequence as input and returns a numerical value (the score) as output

***Cis*-regulatory element (CRE):** a biological sequence that regulates gene expression in *cis*; examples include bacterial promoters, eukaryotic enhancers, and pre-mRNA sequences that regulate splicing

Quantitative sequence–function relationship: the mapping between all possible sequences that a genetic element can have and the quantitative value of that element's molecular phenotype

Massively parallel assay: an experimental method that, by using high-throughput DNA sequencing as a readout, can simultaneously measure the molecular phenotypes of a large number of sequence variants

of quantitative sequence–function relationships, not only because they generate massive data sets but also because the data they produce can be focused on specific biological contexts (as opposed to genome-wide) both in vitro and inside of cells.

Here, we provide a high-level review of massively parallel assays and the ways in which the data produced by these assays can be used to characterize quantitative sequence–function relationships. We emphasize the shared strategies used in HT-SELEX, DMS, and MPRA experiments, as well as commonalities in the quantitative modeling strategies used to analyze the resulting data. We also discuss the big-picture opportunities that these new approaches present. For human health, massively parallel assays may allow the creation of extensive atlases of genetic variants in human disease genes and key regulatory regions, thus providing a comprehensive solution to the problem of genetic variant interpretation (56, 152, 169). In basic biology, these technologies provide a general strategy for probing the functional architecture of a wide range of genetic elements. Along the way, we highlight impediments to progress in these two areas and suggest potential ways of overcoming these difficulties.

2. MASSIVELY PARALLEL ASSAYS

Massively parallel assays all follow a shared schema (**Figure 1*a***). First, one constructs a library of variants for a genetic element of interest, such as a TF binding site, a protein-coding sequence, or a CRE. This library is then used as input to an experiment that outputs one or more bins of sequences, with the enrichment or depletion of each sequence in each bin being determined by the value of a specific molecular phenotype. The input library (bin 0) and each output bin (bin 1, bin 2, etc.) are sequenced, and the resulting number of times that each variant is observed in each bin is used to quantify that variant's molecular phenotype. The ultrahigh-throughput nature of modern DNA sequencing makes it possible to simultaneously assay thousands to millions of sequence variants in this manner.

Within this shared design, a broad array of different assays have been developed by mixing and matching different binning strategies with different variant libraries. Commonly used binning strategies include in vitro binding, selective cellular growth, fluorescence-activated cell sorting, and mRNA sequencing (**Figure 1*b–e***). Commonly used variant library formats include genomic, element-shuffle, element-swapping, element-scanning, systematic mutation, scattered mutation, randomized window, and fully random libraries (**Figure 2**).

Many different methods for constructing variant libraries have also been described. Classic methods include the digestion and cloning of bulk genomic DNA, standard DNA synthesis with nucleotide mixtures, and error-prone PCR. Commercially available oligo pools, in which approximately 10^3–10^5 individually specified DNA sequences are synthesized, have greatly increased the flexibility with which libraries of sequences less than approximately 150 base pairs in length can be designed (95, 142). New methods for high-throughput directed mutagenesis (37, 170, 174) and high-throughput gene synthesis (115) have also opened up previously inaccessible possibilities for long sequence libraries. And, importantly, high-throughput landing-pad integration methods (90) and template-directed mutagenesis using CRISPR/Cas9 (34, 141) are enabling the investigation of large CRE libraries in realistic chromosomal contexts.

This ability to freely combine libraries of different types with different high-throughput enrichment strategies has led to a proliferation of similar assays developed independently by different groups. This burst of creative work has left in its wake a veritable alphabet soup of assay names. To emphasize the shared concepts and strategies behind these approaches, we have organized these methods into three broad classes: HT-SELEX, DMS, and MPRA.

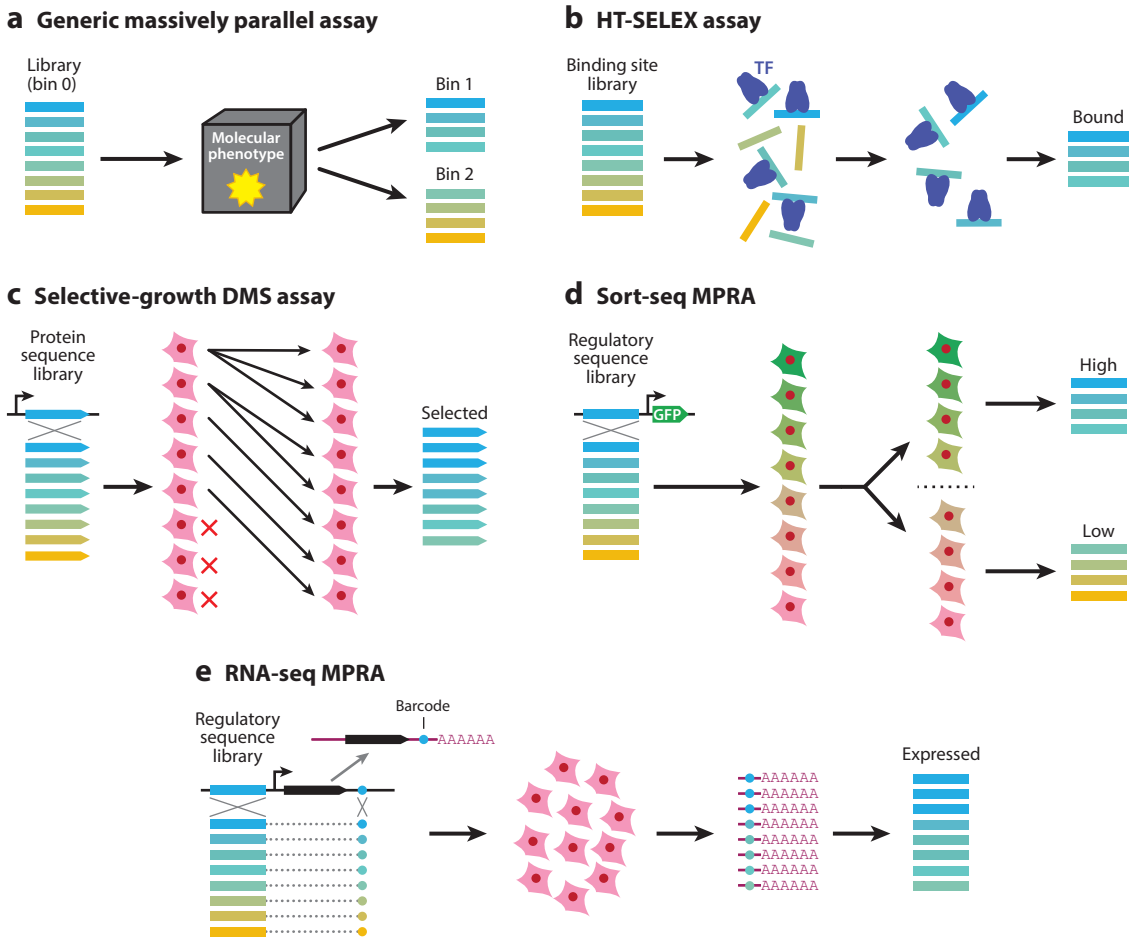

Figure 1

Experimental strategies used in massively parallel assays. (*a*) In the general form of a massively parallel assay, a library of pooled sequences serves as input to an experiment. This experiment then outputs sequences into one or more bins in a manner that depends on each sequence's measured molecular phenotype. The sequences in each bin are then tallied using high-throughput DNA sequencing. (*b*) In an HT-SELEX experiment, a library of candidate DNA binding sites is incubated with a TF of interest, after which TF-bound DNA is isolated and sequenced. (*c*) In a DMS experiment, protein-coding sequences are selected according to a specific molecular phenotype, such as cellular growth rate. (*d*) In a typical sort-seq MPRA, a library of variant CREs is used to drive the expression of a fluorescent reporter gene. Cells expressing this reporter are sorted into bins using fluorescence-activated cell sorting, and the CREs in each bin are then sequenced. (*e*) In a typical RNA-seq MPRA, a library of variant CREs drives the expression of mRNA that contains CRE-specific barcodes, which are then sequenced and tallied. Abbreviations: CRE, *cis*-regulatory element; DMS, deep mutational scanning; HT-SELEX, high-throughput SELEX; MPRA, massively parallel reporter assay; TF, transcription factor.

2.1. High-Throughput SELEX

Much of what is known about protein–DNA and protein–RNA binding specificity has been learned using high-throughput microarray techniques, such as protein-binding microarrays (15) and RNAcompete (120). But over the last decade, HT-SELEX (**Figure 1*b***) has emerged as a simple and highly accessible alternative for assaying binding specificity (160). HT-SELEX is an adaptation of the classic SELEX method (166) and was described by multiple groups circa 2009–2010. Typically, one begins with a library of random DNA sequences, with each sequence

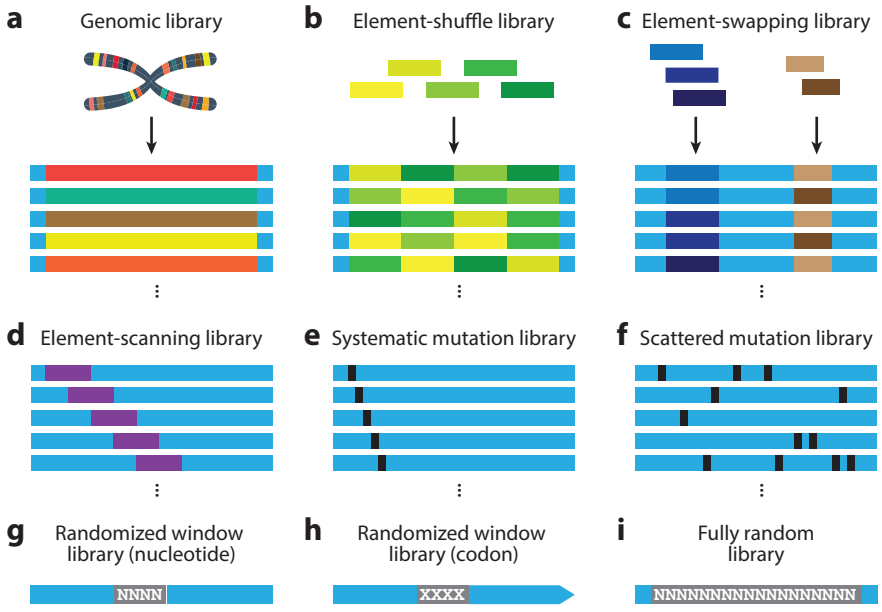

Figure 2

Sequence libraries commonly used in massively parallel assays. Note that, in libraries of all types, variant sequences are usually flanked by fixed DNA that is needed for technical reasons such as PCR amplification. (*a*) Genomic libraries consist of segments of genomic DNA. (*b*) Element-shuffle libraries consist of fixed sequence elements (e.g., TF binding sites) combined in different arrangements. (*c*) Element-swapping libraries consist of a fixed background sequence in which selected elements are varied. (*d*) Element-scanning libraries consist of a fixed background sequence within which sequence elements are placed at systematically varied positions. (*e*) Systematic mutation libraries consist of a background sequence in which all possible single (or even double) mutations are made. (*f*) Scattered mutation libraries consist of a background sequence in which mutations are randomly introduced. (*g*,*h*) Randomized window libraries consist of a fixed sequence context in which a small region is completely randomized. Sequences can be randomized at the level of nucleotides (panel *g*) or codons (panel *h*). (*i*) Fully random libraries consist of completely randomized DNA. Abbreviation: TF, transcription factor.

flanked by constant DNA that enables amplification. This DNA library is then incubated in vitro with a TF of interest, after which TF-bound DNA is isolated and sequenced. As in the standard SELEX procedure, TF-bound DNA can be amplified and used as input for additional rounds of selection if desired.

HT-SELEX can be performed in a variety of ways depending on the system of interest and the goals of the study. The use of random DNA libraries allows one to determine the binding specificities for many TFs in parallel without needing to tailor each library to each individual TF (62). Alternatively, by using DNA libraries in which one or more fixed TF binding sites are partially mutagenized, studies have been able to quantify TF specificity at high precision (187). HT-SELEX methods have also been developed for assaying the specificity of RNA-binding proteins (50). A variety of HT-SELEX-like assays have also been described. *Escherichia coli* one-hybrid is a method that is conceptually similar to HT-SELEX, except that TF–DNA binding is assayed in living *E. coli* cells using selective growth (179). Recent work has also explored the possibility of coupling microfluidic devices with high-throughput DNA sequencing (57, 78), potentially enabling measurements of binding kinetics.

2.2. Deep Mutational Scanning

Epistasis: the phenomenon wherein the effect of a mutation depends on which other mutations are already present in a sequence

We use the term DMS to describe massively parallel mutagenesis studies on proteins and other macromolecules (such as tRNAs) that have complex 3-D structures. Most DMS libraries are constructed to probe the effects of single amino acid substitutions generated either randomly (e.g., using error-prone PCR) (41) or through systematic mutational scanning (174). The study of epistasis between mutations has been a common theme in DMS studies and has been pursued using scattered mutation libraries (136), systematic pairwise mutation libraries (104, 134), and short randomized window libraries (116, 175). An emerging strategy is the use of DMS experiments to assay libraries of protein sequences engineered to have specific properties, e.g., to study functional constraints on unstructured protein domains (151). These experiments are likely to benefit from new methods for synthesizing large libraries of synthetic genes (115).

Many different protein selection procedures have been used in DMS experiments. Protein display methods select for buffer-facing proteins that are able to bind a ligand of interest. Among such methods are phage display (41), yeast display (2, 75, 121), mammalian cell display (40), and RNA display (104). The selection of cells based on their ability to express a fluorescent reporter gene has also proven useful (116, 136, 151). Another common DMS strategy is to select for growth (**Figure 1*c***), e.g., in viral hosts (165, 177), bacteria (94), yeast (53), or mammalian cells (92). Finally, while most DMS studies have focused on proteins, several have also focused on structural RNAs, such as tRNAs (29, 81, 118).

2.3. Massively Parallel Reporter Assays

The term MPRA describes a diverse class of assays used to interrogate many different types of CREs in a diverse set of biological systems. These assays mix and match different experimental strategies (**Figure 1**) with CRE libraries of different types (**Figure 2**). Here, we briefly review the wide range of investigations that have been enabled by MPRAs, highlighting in each case some early relevant work; for a more detailed review of MPRA technology, see Reference 79.

MPRAs have been developed in a wide range of systems, including in vitro expression systems (112), bacteria (69), yeast (142), insect cells (5), mammalian cell culture (95), intact organs (76), and live animals (111). These assays have been used to study many different types of CREs, including promoters (69, 112), enhancers (76, 95, 111), the 5′ and 3′ untranslated regions (UTRs) of mRNAs (31, 103, 140), and pre-mRNA sequences that regulate splicing (9, 66a, 129, 173). Most MPRAs use either a sort-seq strategy or an RNA-seq strategy. Sort-seq MPRAs (**Figure 1*d***) couple gene expression to a fluorescent protein readout (69, 113). Fluorescence-activated cell sorting is then used to sort cells based on expression level, after which the variant CREs in each sorted bin are sequenced. Alternatively, RNA-seq MPRAs (**Figure 1*e***) use the sequencing of expressed mRNA to measure activity. This technique often requires the inclusion of CRE-specific barcodes in expressed transcripts (76, 95, 111, 112), but such barcodes are not always necessary (5, 9, 66a). In some cases, it is useful to couple RNA-seq MPRAs with techniques that provide other information about the mRNA transcript, such as start site location (168) or the presence of an alternatively spliced exon (9, 66a, 129, 173). Less prevalent but no less useful are cell-growth-based methods, which have been used for both MPRAs (26) and MPRA-like studies of DNA replication origins (83).

3. GENOMIC VARIANTS

One of the most potentially impactful applications of massively parallel assays is to address the problem of variant interpretation in whole-genome or whole-exome sequencing (56, 152, 169). The difficulty here is the frequent observation of mutations in human disease genes for which

there is not sufficient evidence under current clinical guidelines (122) to classify them as either likely pathogenic or likely benign. Because these variants of uncertain significance (VUSs) are in aggregate quite common, resolving whether they do or do not affect gene function would be of substantial practical importance for patients and clinicians.

Existing approaches in medical genetics have limited utility for addressing this problem because of the individual rarity of most variants. Statistical approaches such as genome-wide association studies can only identify variants at high enough frequencies to be observed in multiple individuals with the disease phenotype, whereas many VUSs from whole-genome or whole-exome sequencing studies have never been previously observed (56, 152). For similar reasons, because of the rarity of VUSs, classical methods such as pedigree analysis within multiple affected families cannot be employed.

While empirical approaches are hampered due to the rarity of individual variants, current computational approaches (164) also appear to have only limited utility. Such methods typically rely on some subset of population-genetic data, functional-genomic data, signatures of evolutionary conservation, structural data, and existing disease annotations. In practice, however, these models suffer from a lack of precision, wherein many variants identified as deleterious do not display a corresponding phenotype (49, 96, 119, 162). In a recent community-organized prediction challenge, these methods also demonstrated only a moderate ability to predict the quantitative effects of missense mutations (182).

As a result, there is currently an important unmet need for determining the effects of VUSs. Massively parallel assays, which in this context are also referred to as multiplex assays of variant effects (MAVEs), have the potential to help address this need (152, 169). Indeed, there is a strong precedent in medical genetics for the utility of laboratory-based functional assays (125), and established, robust, and reproducible functional assays can already provide strong evidence for or against pathogenicity under current clinical guidelines (122). Provided methods can be developed that match the performance of low-throughput functional assays, comprehensive massively parallel assays of variants in the most clinically relevant and actionable disease genes would go far in addressing the difficulties presented by VUSs.

Variant of uncertain significance (VUS): a mutation that cannot be definitively classified as likely pathogenic or likely benign owing to insufficient evidence

Complementation: the phenomenon wherein the introduction of a variant allele into a loss-of-function genetic background restores the wild-type phenotype

3.1. Massively Parallel Assays of Human Disease Genes

Several different strategies have used massively parallel assays to measure the functional effects of variants in human disease genes. One approach is based on measuring the activity of a reporter gene. For instance, an influential study by Majithia et al. (88) assayed the impact of all possible missense mutations on the protein PPARγ via changes in the expression of CD36, a downstream target of PPARγ. However, like many functional assays, this experimental design is specific to a particular gene of interest, and a different experimental methodology would be needed for each gene assayed in this manner.

A distinct approach that partially overcomes this difficulty is to use complementation assays. Here, one measures cellular growth rate, typically in yeast or a human cell line, using a genetic background where the endogenous locus has been knocked out. The deletion of the endogenous locus results in a measurable fitness defect that is then ameliorated by functional library variants. Sun et al. (162) provided an important proof of concept for this approach by establishing that complementation assays could be used to characterize the effects of previously annotated mutations in 22 human disease genes. In a follow-up paper, Weile et al. (170) performed high-throughput assays on several of the corresponding proteins (UBE2I, TPK1, and CALM1) and used a machine learning approach to predict the effects of all missense variants therein. A completely different strategy was demonstrated by Matreyek et al. (92), who constructed fusions between enhanced

GFP (EGFP) and comprehensive variant libraries for the proteins PTEN and TPMT, then assayed the impact of mutations by measuring steady-state protein abundance. The rationale behind this approach is that destabilizing mutations will decrease abundance, e.g., through the targeted degradation of misfolded protein.

Perhaps the best-studied model system for massively parallel assays of human disease genes is the BRCA1 RING domain. This critically important protein domain has been investigated using several different experimental strategies, including a phage-display-based ubiquitination assay (154), a yeast two-hybrid binding assay (154), a GFP-based reporter assay for homology-directed DNA repair (153), and a growth rate readout in a haploid human cell line in which *BRCA1* has become an essential gene (35). Although all of these assays could likely be further improved, the growth rate assay (35), which uses CRISPR/Cas9 editing of the endogenous *BRCA1* locus, currently appears to have the best performance in distinguishing known pathogenic variants from known benign variants.

3.2. Improvements to Experimental Methods

As shown by the discussion above, current efforts to prospectively assess the effects of mutations in human disease genes appear promising. Nevertheless, substantial improvements in the reproducibility and rigor of these experiments can likely be achieved by incorporating best practices from other fields. In particular, both the area of differential expression analysis in RNA-seq studies (22) and the field of experimental evolution (43) have established standardized methods (described below) for how to reliably measure fold changes in abundance between libraries or time points.

First, studies commonly report enrichment scores or other experiment-specific, semiquantitative measures that are not directly comparable across genes, laboratories, or assays. Often, however, these scores can be replaced by measurements in inherently meaningful units. For instance, complementation assays and growth assays generally measure growth rates (i.e., cell divisions per hour), and there are standard methods from experimental evolution and microbiology for estimating these rates as well as differences between them, i.e., selection coefficients (43). Such methods are already incorporated into the measurement procedures of some reported assays, such as EMPIRIC (53), Enrich2 (132), and FiT-seq (82). Similarly, studies of protein–ligand binding should report absolute dissociation constants, as in Tite-seq (2).

Second, experiments should include both controls and standards. They should also feature an appropriate degree of replication (43) and, when possible, multiple time points (93). Currently, DMS experiments often use the distribution of synonymous and nonsense mutations as internal controls, but the interpretation of these distributions is complicated by the fact that synonymous mutations need not be neutral, e.g., if they affect translational rates or splicing. A better idea is to incorporate an allelic series of variants with a range of known values for the molecular phenotype of interest. For example, a massively parallel assay based on cellular growth might include a small number of variants whose growth rates have already been measured in a low-throughput manner. This inclusion of an allelic series not only enables calibration (e.g., between enrichment scores and growth rates) but can also reveal important features of the experiment that would not otherwise be apparent, such as saturation. Furthermore, because patients care deeply about the uncertainty in what is known for their specific genomic variant, the concordance between replicate experiments should be measured in terms of the error bars on the values assigned to individual variants, rather than collective measures like the rank correlation across all measurements.

Third, improved best practices for both DNA sequencing and the inference of frequency changes from read counts should be followed. These practices include the use of spike-in controls

(e.g., barcodes at known concentrations) (59), unique molecular identifiers (71), and replicate libraries with different sequence-to-barcode associations (173). Variant libraries should also be shared across laboratories, thereby enabling estimates of between-group variation in identical sets of variants. Moreover, it is clear from differential expression studies that DNA sequence counts typically display a strong excess in variance (86) beyond what is expected under the Poisson models used in many analyses. New software is needed for high-throughput phenotyping that appropriately estimates and incorporates this excess uncertainty in the context of phenotypic estimation, rather than just for hypothesis tests of changes in frequency, as in differential expression. Over the next several years, analysis pipelines can also be substantially standardized, and we agree with calls (152, 169) for a database of relevant studies (e.g., 32a) to enable comparisons among studies and the reanalysis of older experiments with more recent computational tools.

Poisson model: a statistical model in which the number of reads observed for a given variant is assumed to be drawn from a Poisson distribution, which is mathematically convenient but exhibits less variance than is often observed in real experiments

We feel that such methodological improvements are critical from the patient's perspective. The overall landscape of genetic disease is complex and filled with uncertainty. Even diseases commonly classified as having a simple Mendelian basis can show substantial variation due to variable expressivity and incomplete penetrance (23). Moreover, the phenotypic variation for individuals harboring the same mutation can be driven by both genetic and environmental factors (23), and while this genetic influence may sometimes arise due to other mutations at the same locus (21), it is likely that most diseases thought to be monogenic are in fact influenced by multiple loci spread across the genome (65). High-quality biochemical assays thus play a critical role, pushing back the boundary of uncertainty and providing patients with simple and clear facts. We may not be able to tell patients whether they (or a loved one) will or will not experience disease symptoms, but at least we can tell them whether they harbor a specific variant, whether this variant disrupts the biochemical function of the gene, and what the effects of this variant are in simple model systems. In keeping with the emphasis under current clinical guidelines of incorporating multiple lines of evidence in coming to a diagnosis (122), we recommend reporting the measured quantities for particular variants to patients and clearly distinguishing them from imputed scores (170) or probabilities of pathogenicity estimated using statistical models (88).

4. PROTEIN–DNA AND PROTEIN–RNA INTERACTIONS

We now move from applications in human health to questions in basic science. Rather than focusing on the effects of individual mutations, the goal here is to use complex libraries of variants—often including double, triple, and higher-order mutations—to interrogate specific genetic elements. One area where this approach is essential is the study of sequence specificity in protein–DNA and protein–RNA interactions. In this section, we review key concepts in the quantitative modeling of sequence–function relationships within this biological context. We concentrate primarily on TF–DNA binding, which historically has been a focal point for efforts to understand how quantitative information is encoded within the genome. We discuss other select problems as well, including nucleosome formation, splice site recognition, and the role of RNA secondary structure in protein–RNA interactions.

Why are quantitative models needed for understanding binding specificity? TFs, like most other DNA- and RNA-binding proteins, recognize a wide range of sequences other than just their strongest binding sites. Consider CRP, an activator in *E. coli* that binds to DNA sites approximately 22 base pairs in length (**Figure 3*a***). CRP recognizes far more sequences than just its strongest binding site (shown in **Figure 3*b***). This binding site degeneracy can be roughly described by an International Union of Pure and Applied Chemistry (IUPAC) motif (**Figure 3*c***), which specifies only the most constrained positions at each position in the binding site. However, such a representation is unable to account for the fact that different binding sites often have functionally important differences in binding affinity (e.g., 61).

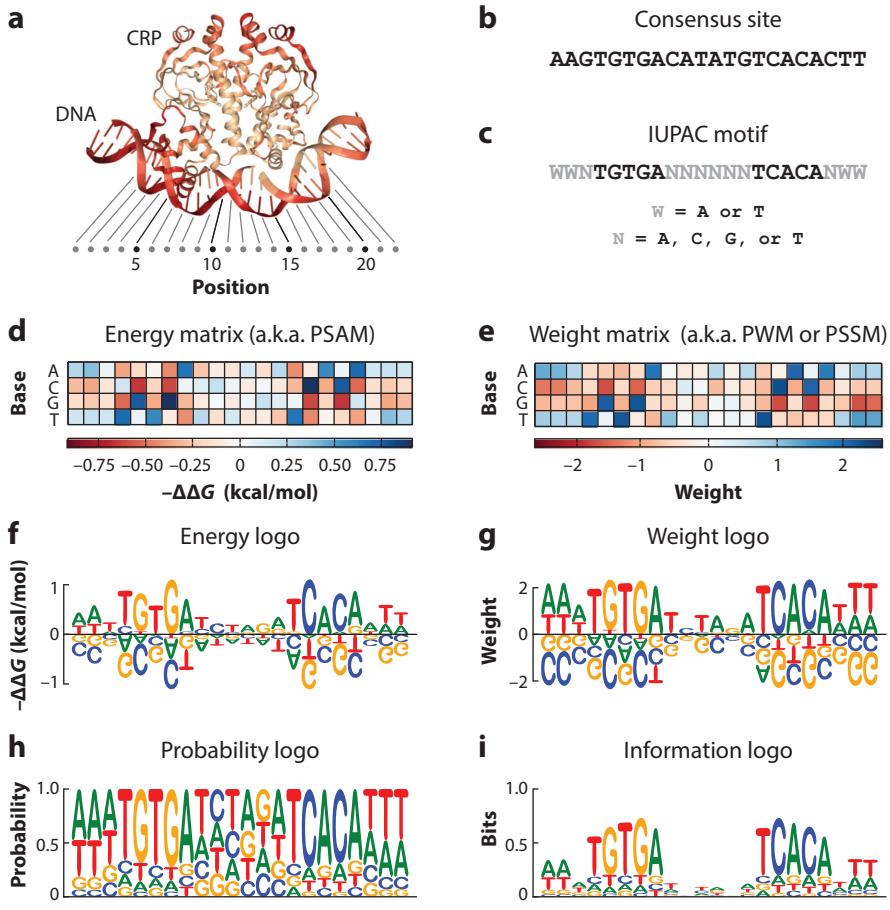

Figure 3

Additive models of TF–DNA binding. (*a*) Structure of the *Escherichia coli* TF CRP bound to its consensus DNA site (Protein Data Bank ID 1CPG) (110). The central 22-base-pair positions of this site are indicated. (*b*) The consensus (i.e., strongest) binding site for CRP. (*c*) An IUPAC motif for CRP. (*d*) An energy matrix for CRP, determined by Kinney et al. (69) and represented as a heat map. (*e*) A weight matrix for CRP, inferred from 358 annotated CRP binding sites in RegulonDB (135). (*f*) An energy logo representation of the energy matrix in panel *d*. The character heights represent the $\Delta\Delta G$ parameters of the energy matrix. (*g–i*) Weight logo (panel *g*), probability logo (panel *h*), and information logo (panel *i*) representations of the weight matrix in panel *e*. In panel *g*, the character heights represent individual nucleotide weights. In panel *h*, the height of each base represents the probability of that base occurring at that position in a binding site. In panel *i*, the total height of each stack of characters quantifies the importance of a nucleotide position using concepts from information theory, while the relative height of each base represents the probability of that base occurring at that position. Logos were created using Logomaker (164a). Abbreviations: IUPAC, International Union of Pure and Applied Chemistry; PSAM, position-specific affinity matrix; PSSM, position-specific scoring matrix; PWM, position weight matrix; TF, transcription factor.

4.1. Structural Predictions and the Need for Quantitative Models

To understand the quantitative determinants of TF–DNA binding, it is helpful to think about the effects that mutations have on the structure of a TF–DNA co-complex. A change from one nucleotide to another at any position in the DNA sequence will result in small changes in the atomic

positions within this structure, a corresponding change in the Gibbs free energy of binding (denoted ΔG), and thus a change in the affinity of that binding site. In principle, a biophysical analysis of co-complex structures should be able to predict these changes and thus provide a quantitative description of TF specificity. And indeed, making such predictions has been the focus of substantial research (63, 77). But this analysis is exceedingly difficult in practice, in large part because the energy scale that determines binding affinity ($k_B T = 0.62$ kcal/mol at 37°C) is very small relative to the scale of the individual chemical interactions involved in complex formation. For example, CRP is estimated to form 31 hydrogen bonds with DNA in the co-complex structure shown in **Figure 3*a*** (110). Failing to account for the presence or absence of just one relatively weak hydrogen bond (say, 1.0 kcal/mol) would throw off a binding affinity estimate by $e^{1.0/0.62}$, or approximately fivefold.

Quantitative modeling provides an alternative means of understanding TF specificity, one that is largely orthogonal to structure-based approaches. A quantitative model is an abstract mathematical function that takes a biological sequence as input and outputs a numerical quantity known as a score. Such models depend on parameters, the values of which must ultimately be inferred from data. Early successful models of TF binding were built from sequence alignments of (often remarkably few) binding sites (159). Today, a variety of high-throughput assays can provide sufficient data for developing quantitative models that are highly complex (77). We now review these different types of models and the strategies used to infer their parameters.

Score: a general term for the output of a quantitative model; scores can represent log odds ratios, Gibbs free energies, or other quantifications of molecular phenotype

Parameter: a numerical value that is inferred from data and then used by a quantitative model when assigning scores to sequences

Functional model: a quantitative model whose score describes the molecular phenotype of an input sequence

Generative model: a quantitative model that describes the probability distribution from which functional sequences are drawn

4.2. Functional Models Versus Generative Models

There are two conceptually distinct types of quantitative models: functional models and generative models. Functional models aim to predict the values of molecular phenotypes; generative models, by contrast, seek to describe the probabilities of observing different sequences within functional genetic elements.

In the context of TF–DNA interactions, the score produced by a functional model usually represents the ΔG of binding. The simplest form of such models is called an energy matrix, also known as a position-specific affinity matrix (PSAM; see Reference 38). Energy matrices assume that each position within a binding site contributes independently to the value of ΔG. The parameters of an energy matrix are the individual energy contributions of each possible base at each nucleotide position and are conventionally denoted by $\Delta\Delta G$. Note that favored bases have lower (i.e., more negative) $\Delta\Delta G$ values, since lower energy corresponds to stronger binding. **Figure 3*d*** illustrates an energy matrix for CRP that was determined by Kinney et al. (69) using data from a sort-seq MPRA.

The score produced by a generative model of TF specificity usually represents the log odds that a functional binding site, as opposed to random DNA under no selective pressure, will have a given sequence. The most common realization of such models is known as a weight matrix, although the terms position weight matrix (PWM) and position-specific scoring matrix (PSSM) are also commonly used (158). Each nucleotide contributes additively to the overall weight matrix score via a parameter called a weight. **Figure 3*e*** shows a weight matrix for CRP computed using an alignment of the 358 annotated CRP binding sites in the *E. coli* genome (135). We refer readers to other reviews (77, 158) for a description of how weight matrices are constructed from alignments such as this.

In two classic papers, Berg & von Hippel (13, 14) proposed an intriguing connection between energy matrices and weight matrices, thus linking functional and generative models of TF binding. They suggested that, if an energy matrix model for a TF is accurate, it should closely resemble the weight matrix model constructed by aligning functional binding sites that have arisen via natural

selection. Stormo and colleagues (52) later presented an alternative rationale for this connection based on an in vitro thought experiment. Empirically, this connection between energy matrices and weight matrices has turned out to be remarkably strong. Indeed, the energy matrix from **Figure 3*d*** is largely similar to the weight matrix from **Figure 3*e***, save for an overall multiplicative factor.

Nevertheless, functional models and generative models do represent very different things, and their equivalence should not be taken for granted. In particular, the in vitro argument relating energy matrices and weight matrices breaks down when TF concentrations are high (130). For this and other reasons, multiple groups have argued that functional models are to be preferred over generative models when practicable (38, 70, 130, 183).

Sequence logo: a graphical representation of an additive model in which the height of each character at each position represents, directly or indirectly, how that character contributes to a model's score

Sequence feature: a mathematical representation of any aspect of a sequence, such as the presence or absence of specific bases at specific positions, or a characterization of DNA geometry, such as minor groove width

4.3. Sequence Logos

Sequence logos provide an evocative way to visualize simple models such as energy matrices and weight matrices (164a). Here, we describe four different kinds of sequence logos that are commonly used in the literature. **Figure 3*f*** shows an energy logo (38) for CRP, where the character heights represent the $\Delta\Delta G$ parameters of the energy matrix from **Figure 3*d***. The weight logo in **Figure 3*g*** similarly illustrates the parameters of the weight matrix model from **Figure 3*e***. Both of these logos contain the same information as their respective heat-map representations but are easier for many readers to interpret. The probability logo in **Figure 3*h*** provides a somewhat more intuitive, though mathematically less direct, way of visualizing a weight matrix: The height of each character is the probability of that base occurring at that position in a binding site. Information logos (e.g., **Figure 3*i***) provide yet another way to graphically represent weight matrices. The total height of each stack of characters quantifies the importance of a nucleotide position using a quantity from information theory called Kullback–Leibler divergence (which has units of bits), while the relative heights of characters within a stack reflects their relative probabilities. Information logos were the first type of sequence logo described in the literature (138) and are still widely used (25, 77). However, they tend to underrepresent the importance of nucleotide positions that are less tightly constrained.

4.4. Modeling Epistatic Interactions

Energy matrices and weight matrices are examples of additive models: They assume that each position within a sequence contributes independently to that sequence's overall score. The potential pitfalls of this independence assumption are well recognized, and substantial effort has gone into developing quantitative models that can express epistatic interactions between positions (77). One way to model epistatic interactions is to make use of sequence features that integrate information across multiple positions (see **Figure 4**). The simplest type of epistatic model is the neighbor model, also known as a dinucleotide model, in which the score is a sum of contributions from sequence features that represent adjacent dinucleotides. One notable example of a neighbor model in the literature was proposed by Segal et al. (139) for describing the positioning of nucleosomes in yeast. Pairwise models are somewhat more complex than neighbor models, as they include contributions from sequence features that represent both adjacent and nonadjacent pairs of positions. Two well-known examples of pairwise models were proposed by Yeo & Burge (180) to describe 3′ and 5′ splice sites in the human genome. Neighbor and pairwise models are naturally generalized by higher-order models, which incorporate contributions from three or more positions at a time. Higher-order models in which coupled positions are contiguous are also referred to as *k*-mer models. In the TF modeling competition organized by Weirauch et al. (172), *k*-mer models performed especially well on TFs that exhibited multiple distinct DNA-binding motifs.

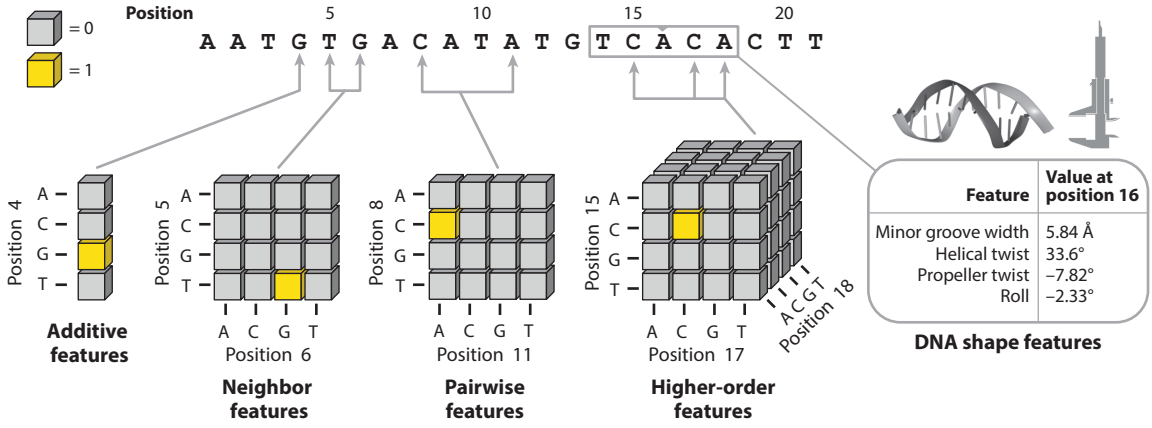

Figure 4

Sequence features commonly used in models of TF–DNA binding. Additive features indicate the presence or absence of individual bases at individual positions. Neighbor features represent adjacent dinucleotides, while pairwise features can represent both adjacent and nonadjacent nucleotide pairs. Higher-order features are constructed analogously by considering three or more nucleotide positions at a time. DNA shape features (186), which characterize the shape of B-form DNA at the center of a small sequence window, are increasingly being incorporated into TF binding models as well. The B-DNA structure shown here is from Protein Data Bank ID 1ILC. Abbreviation: TF, transcription factor.

An alternative means of incorporating epistatic interactions is based on the observation that TFs recognize sequence-dependent aspects of DNA geometry, rather than just specific combinations of nucleotides (127). To facilitate the construction of models that reflect this aspect of TF specificity, Zhou et al. (186) used coarse-grained molecular dynamics simulations to tabulate values for the minor groove width, helical twist, propeller twist, and roll that occur near the center of all 512 possible DNA pentamers within free (i.e., non-TF-bound) B-form DNA (**Figure 4**). This information can be used to supplement additive models of TF specificity (185), yielding what are commonly referred to as DNA shape models. For similar reasons, models of protein–RNA binding often incorporate predictions of what the RNA secondary structure would be in the absence of the RNA-bound protein (66, 91, 105).

One difficulty with epistatic models is that the number of model parameters grows rapidly as interaction order increases. For example, consider a TF that recognizes binding sites 10 base pairs in length. An additive model for this TF will have 40 parameters, a neighbor model will have 144 parameters, a pairwise model will have 720 parameters, a third-order model will have 7,680 parameters, and so on. As the number of parameters increases, so does the risk of overfitting. Overfitting can often be counteracted by using standard regularization methods (124) or sparse models, in which most of the parameters are constrained to be zero (143).

Overfitting: poor model performance resulting from model parameters being fit too specifically to insufficient amounts of data

The above-described models are all examples of linear models because the scores they return are linear combinations of model parameters (see Equation 2 in the sidebar titled Mathematical Forms of Sequence–Function Relationships). Global epistasis models provide an important generalization of the linear model concept: The score of a global epistasis model is a nonlinear transformation of the score of a linear model (Equation 3 in the sidebar). Global epistasis is natural in the study of TF specificity because the fraction of time a DNA site is bound by a TF is a highly nonlinear function of ΔG (158). Evolutionary fitness, which governs the evolution of TF binding sites, has also been observed in some cases to be a nonlinear function of ΔG (100).

Linear model: any model whose score is a linear combination of model parameters; examples include additive, neighbor, pairwise, higher-order, and DNA shape models

MATHEMATICAL FORMS OF SEQUENCE–FUNCTION RELATIONSHIPS

Let S denote an input sequence, L be the length of that sequence, and C be the number of possible characters at each position in S ($C = 4$ for DNA and RNA, $C = 20$ for proteins). An additive model relies on $C \times L$ features, each written as $F_{ci}(S)$, where $F_{ci}(S) = 1$ if character c occurs at position i in sequence S, and $F_{ci}(S) = 0$ otherwise. The score of an additive model is computed as

$$f_{\text{additive}}(S) = \sum_{c=1}^{C} \sum_{i=1}^{L} \theta_{ci} F_{ci}(S), \quad 1.$$

where θ_{ci} denotes the model parameter corresponding to feature F_{ci}. More generally, a linear model of a sequence–function relationship is defined as any model that can be written as

$$f_{\text{linear}}(S) = \sum_{k=1}^{K} \theta_k F_k(S), \quad 2.$$

where K is the number of sequence features in the model, $F_k(S)$ is the kth sequence feature (which is allowed to take an arbitrary value for each sequence, not just 0 or 1), and θ_k is the corresponding model parameter. Global epistasis models include an additional nonlinearity and can be expressed as

$$f_{\text{global}}(S) = g(f_{\text{linear}}(S)), \quad 3.$$

where $f_{\text{linear}}(\cdot)$ is a linear model and $g(\cdot)$ is a nonlinear function.

4.5. Learning Models from Data

Just as important as the mathematical form of a quantitative model is the way in which the values of that model's parameters are learned from data. Parameter inference is a particularly challenging problem because TF-binding experiments almost always measure binding to DNA sequences that are much longer than a single binding site. A large number of motif-finding algorithms, using a wide range of machine learning strategies, have been proposed for this purpose (77). Generative models of TF specificity are often inferred using methods from the field of signal processing, such as the expectation–maximization algorithm (10). Functional models, on the other hand, can be learned using statistical inference methods such as maximum likelihood (38, 183), support vector machines (45), or mutual information maximization (8, 32, 70). The proper way to treat long DNA sequences within functional models, however, is less obvious than it is in generative models. One attractive approach that is becoming increasingly popular (38, 124, 131, 183) is to use thermodynamic models (16, 114, 144), which rely on the equations of statistical physics, to predict the average number of TF molecules that will simultaneously bind to a long DNA sequence.

Thermodynamic model: a biophysical model that describes an ensemble of molecular states governed by the equations of statistical mechanics

4.6. Outlook

The problem of TF–DNA binding has spurred the development of a rich variety of methods for modeling sequence–function relationships. At present, there appear to be sufficiently powerful experimental and computational methods for characterizing the in vitro specificities of individual TFs (77). But our understanding of TF–DNA binding in cells is far from complete, as in vivo binding patterns differ markedly from what one would naively expect from known TF motifs

(55). In eukaryotes, TF–DNA binding is often contingent on binding sites being present in nucleosome-free regions of DNA, and the rules that govern nucleosome positioning remain incompletely understood (161). Epigenetic modifications, such as cytosine methylation, also strongly affect TF binding to DNA (27). Moreover, the interactions of a TF with other DNA-bound proteins (148) or with non-DNA-binding cofactors (147) can affect that TF's specificity in emergent and sometimes surprising ways.

One exciting and increasingly popular strategy that might help to answer some of these lingering questions is the use of deep learning techniques (48) for modeling protein–DNA and protein–RNA specificity. Of particular interest are convolutional neural network models, which can integrate binding signals across long sequences in a highly flexible manner (3). These models have been proposed for characterizing the complex context dependence of TF–DNA binding to chromatin in vivo (68, 184). In the context of protein–RNA binding, these models have shown remarkably good performance, including the ability to capture the effects of RNA secondary structure (73). Convolutional neural networks are less readily interpreted than thermodynamic models, but efforts to improve the interpretability of these and other deep learning models are underway (74, 146).

Deep learning: machine learning methods that use neural networks with multiple hidden layers

Sequence space: the set of all possible sequences for a genetic element of interest

5. PROTEINS

In some ways, the relationship between an amino acid sequence and its biological function is quite different from the relationship between a DNA binding site and its affinity for a TF. First, the amino acid alphabet is much larger than the nucleotide alphabet (20 proteinogenic amino acids versus 4 deoxyribonucleotides), meaning that random protein sequences have only 5% rather than 25% sequence similarity. Moreover, protein sequences tend to be much longer than TF binding sites, typically having tens to thousands of positions rather than $\sim$10 positions. Thus, while it is often experimentally feasible in the context of TF specificity to exhaustively explore sequence space, it is typically impossible to do so for proteins since, e.g., there are 10^{130} possible amino acid sequences of length 100. And within this much larger space of possible sequences, the fraction of functional sequences is far smaller. For example, while a few kilobases of random DNA sequence will typically contain at least one binding site for any given eukaryotic TF, it has been estimated that only approximately 1 in 10^{11} protein sequences will exhibit substantial ATP-binding activity (67). Finally, whereas the molecular phenotype of a TF binding site can often be fully described by a single number—its binding affinity—each protein will typically have many functionally and biophysically distinct molecular phenotypes, including folding energy, enzymatic activity, ligand or cofactor binding affinity, and responsiveness to allosteric regulation.

Despite these differences, current techniques for modeling and understanding sequence–function relationships in proteins are very similar to those used for TF binding sites. These similarities arise for three main reasons. First, because of the size of protein sequence space and the rarity of functional protein sequences, models of protein sequence–function relationships tend to focus on relatively minor perturbations to a sequence known to be functional. Such perturbations tend to maintain the 3-D structure of the protein and the ability to align one protein sequence to another. As a result of this restriction in scope, each position typically has a relatively consistent functional role across sequence backgrounds, and so additive models with one parameter for each possible amino acid at each position often provide good baseline performance. Second, the simplest thermodynamic models of protein folding are very similar to thermodynamic models of TF binding in that the probability of a protein being folded is a nonlinear function of ΔG, with ΔG itself being additive (178). Finally, just as TF sequence–function relationships usually

focus on DNA binding, proteins can bind many different types of ligands, substrates, and cofactors. Studying how amino acid sequence influences the specificity of these interactions has been a major focus of DMS experiments (1, 156), sometimes even enabling measurements of binding specificity in physical units (2).

Two-state model: a thermodynamic model of protein folding that assumes that the protein has only two possible conformations, folded and unfolded

Monotonic function: a mathematical function that is either consistently increasing or consistently decreasing, and thus has no local maxima or minima

5.1. Evidence For and Against the Additive Folding Energy Hypothesis

Starting in the mid-2000s, a synthetic theory emerged from the work of several different groups that attempted to explain a diverse set of observations about the thermodynamics and evolution of natural proteins using simple assumptions about protein sequence–function relationships. These assumptions were as follows: (*a*) For random mutations, most fitness effects are due to defects in protein folding; (*b*) the fraction of time a protein spends properly folded is a logistic function of the free energy of folding ΔG (i.e., the two-state model); and (*c*) the stability effect of any given mutation, $\Delta\Delta G$, is well conserved across sequence backgrounds and can be treated as additive. These parsimonious assumptions were then used to explain and self-consistently describe a wide variety of phenomena, including frequencies of functional sequences in mutagenesis libraries (19), patterns of epistasis (46, 47), the distribution of observed fitness effects of mutations and the marginal thermodynamic stability of proteins (46, 178), and many features of molecular evolution (46, 181), such as the overdispersed molecular clock (i.e., the observation that amino acid substitutions occur in a temporally clustered manner) (18).

This additive folding energy hypothesis makes strong predictions for the types of models that should be able to describe protein sequence–function relationships. In particular, it predicts that protein sequence–function relationships should be well approximated by a global epistasis model (72, 107, 133, 157) in which fitness is a monotonic function of an underlying additive model (Equation 3) and the underlying additive trait is proportional to the free energy of folding ΔG. Moreover, a key qualitative prediction of such a model is that any given mutation should have either a beneficial or deleterious effect across all genetic backgrounds, even if the magnitude of the effect is background dependent (and might include neutrality on, e.g., highly stable backgrounds). This is because, under the additive folding energy hypothesis, a stabilizing mutation will always increase the fraction of time the protein is properly folded, which will always increase fitness (a similar argument holds for destabilizing mutations).

However, the evidence for the additive folding energy hypothesis from DMS experiments has been mixed. Many studies have found a strong correlation between mutational effects on fitness and mutational effects on stability, including in proteins such as TEM-1 β-lactamase (36, 58), the WW domain (4), nucleoprotein (7), and GFP (136). Others, however, have found that mutational effects in the wild-type background are uncorrelated with stability effects (104). And while global epistasis models can sometimes fit DMS data remarkably well (e.g., see **Figure 5**), other studies have revealed very different patterns of mutational effects. For instance, in a recent study, Starr et al. (155) showed that a large proportion of the mutations that have fixed over the evolutionary history of Hsp90 have fitness effects that have changed sign over evolutionary time. Similarly, and also in Hsp90, Bank et al. (11) conducted combinatorial mutagenesis at six sites and observed a pattern where the sign of a mutation's effect depended strongly on the sequence background. These results are incompatible with the additive folding energy hypothesis for Hsp90.

Another important challenge to the additive folding energy hypothesis comes from generative models for homologous proteins (80, 99). While phylogenetic models of molecular evolution typically assume that each site in a protein evolves independently from the others and thus produces a long-term distribution of states described by an additive model (17, 51, 126, 163), generative

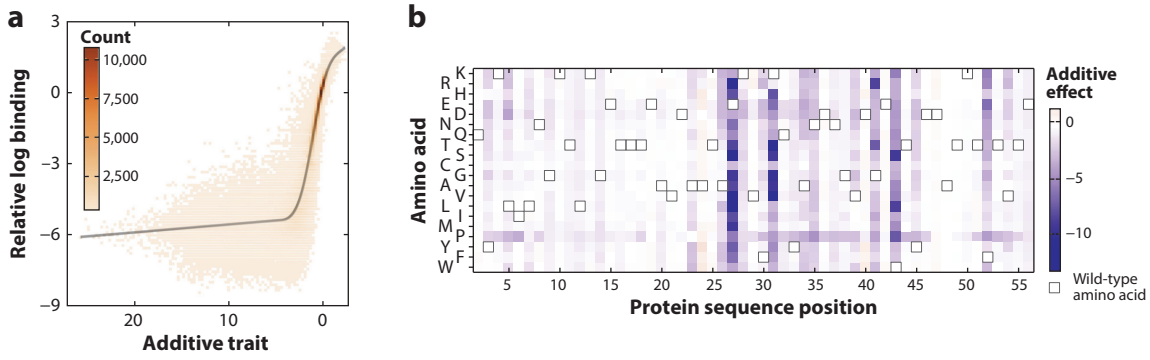

Figure 5

Applying a global epistasis model to DMS data. (*a*) The inferred global nonlinearity (*g* in Equation 3; see the sidebar titled Mathematical Forms of Sequence–Function Relationships) for a DMS data set (104) consisting of all pairs of possible amino acid substitutions in the GB1 domain of protein G. The nonlinear function component of the global epistasis model is shown in gray. The coloring shows the density of observed genotypes (cross-validated $R^2 = .935$). (*b*) The corresponding matrix of context-independent mutational effects on the underlying additive trait. Abbreviation: DMS, deep mutational scanning. Figure adapted from Reference 107.

models of protein alignments typically specify the log-likelihood of a sequence using a pairwise model that captures correlations between amino acids at (nonadjacent) pairs of positions (80). By identifying pairs of sites whose interaction coefficients are unusually large, algorithms that use such pairwise models have shown highly impressive performance at predicting which residues are in direct physical contact in 3-D structures (99). The presence of such pairwise interactions is also plausible from a functional modeling perspective; physicists have long worked with models of folding energy that consider the folding energy to be an additive function of sequence features restricted to sites that contact each other in the 3-D structure (97). Another surprising property of these generative models with pairwise interactions is that they provide state-of-the-art performance in predicting the results of DMS assays; in one study, the performance of a pairwise model was far better than that of the corresponding additive score model (54). Indeed, allowing not only pairwise but also higher-order interactions between sites, via the use of a type of deep learning model called a variational autoencoder (123), can provide even better performance in predicting the results of DMS experiments. Thus, as in the modeling of DNA sequence specificity, generative models of protein sequence appear to provide excellent functional predictions, though the reasons for this remain unclear.

5.2. Case Study: Protein G

Many of the issues discussed above can be seen in the recent literature on protein G, an immunoglobulin-binding protein and model system for protein engineering studies. Olson et al. (104) conducted a DMS experiment based on mRNA display coupled to a pull-down assay using immunoglobulin G bound to beads. By quantifying the change in the frequency of variants before and after selection, they were able to measure enrichment ratios for all single and double mutations within a 56-amino-acid-long domain. [In a follow-up study, the same group also made similar measurements for all $20^4 = 160{,}000$ possible combinations of amino acids at four particularly epistatic sites (175).] Notably, almost no epistasis was observed between many pairs of mutations, particularly for mutations with small fitness effects. This lack of observed epistasis indicates an experiment of extremely high quality, since any noise would tend to produce spurious epistatic interactions.

Enrichment ratio: the ratio between the frequency of a variant in the input library and its frequency after selection for a molecular phenotype of interest

Contrary to the additive folding energy hypothesis, Olson et al. (104) found that the effects of mutations around the wild-type sequence, as quantified by enrichment ratios, were essentially uncorrelated with published measurements of the stability effects of mutations. However, they also somewhat surprisingly found that there were some mutant backgrounds where subsequent mutations had fitness effects that were well correlated with published stability effects (104, 176). Otwinowski et al. (107) reanalyzed these double-mutant data by applying a global epistasis model. They confirmed that, while a global epistasis model provided an excellent fit to the data, the inferred coefficients for the additive part of the model remained uncorrelated with published stability effects. However, in a follow-up paper, Otwinowski (106) fit a more complex biophysical model (89) that treated the probability of a bound and folded complex as a function of both an underlying binding energy and a distinct underlying folding energy, both of which were additive functions of the sequence. In this apparently better-specified model, the inferred energetic effects were well correlated with the published energetic effects of mutations. Moreover, the model provided predictions for a large number of mutations whose energetic effects were unknown. These predictions were then dramatically confirmed by a subsequent high-throughput study that comprehensively measured the effects on free energy of folding for all point mutations in this protein (102).

***Cis*-regulatory grammar:** an umbrella term for rules that govern how multiple *trans*-acting factors that bind a CRE interact with one another and thereby modulate gene expression

While Otwinowski's (106) analysis would suggest that a model with two additive molecular phenotypes (folding energy and binding energy) is largely sufficient to explain the data, two other recent studies (128, 137) suggest that the picture is more complex. Analyzing the same data set, these groups were able to calculate the 3-D structure of the GB1 domain by inferring contacting residue pairs and then using these pairs as constraints for ab initio folding. Such a feat would be impossible if the model described by Otwinowski (106) were complete, since both of the underlying phenotypes in that model were additive and so could not directly contain information about protein contacts. As it stands, the best current explanation is that the model with two underlying additive molecular phenotypes is approximately correct, but deviations from this model are still sufficient to enable the identification of residues that physically interact in the 3-D protein structure.

6. *CIS*-REGULATORY ELEMENTS

Studies of CREs contend with complications beyond those discussed in the previous section. As with proteins, epistatic interactions that straddle large portions of a CRE are often critical for function (20, 150), and quantitative models that can accommodate such interactions are essential. But relative to proteins, much less is known about how CREs actually work. Except in exceedingly well-studied systems, such as the *E. coli lac* promoter (33) and the human interferon-β enhancer (109), the 3-D structures of CREs complexed with the proteins they scaffold have not been determined. Phylogenetic alignments of CREs across species are also less informative than alignments of proteins, as individual binding sites within CREs often appear and disappear on short evolutionary timescales (30, 85).

Given these substantial challenges, MPRAs are proving to be a remarkably powerful technology for studying CRE biology on multiple scales. Using MPRAs, investigators can identify novel CREs and characterize their activities, quantify the effects of genetic variation within CREs, dissect specific CREs of interest in mechanistic detail, and characterize general features of *cis*-regulatory grammar, i.e., the rules that govern how different combinations of binding sites within CREs combine to establish functionality. These different lines of investigation are enabled by mixing and matching different MPRA strategies (**Figure 1**) with different CRE sequence libraries (**Figure 2**) and, when appropriate, using quantitative models to explain the resulting data. We now review a few selected studies in order to illustrate this broad range of applications.

The work of Arnold et al. (5) and Johns et al. (60) illustrates the power of MPRAs for CRE discovery across the phylogenetic tree. Arnold et al. (5) developed an MPRA called self-transcribing active regulatory region sequencing (STARR-seq) to identify novel metazoan enhancers. In this assay, a library of genomic fragments is cloned downstream of a basal promoter. Being positioned in this manner, genomic sequences that have enhancer activity can drive transcription of themselves and thus be identified by sequencing expressed mRNA. Johns et al. (60), on the other hand, developed an MPRA for identifying potentially useful bacterial promoters in large metagenomic databases. Using an oligo pool library comprising 29,249 candidate CREs drawn from 184 prokaryotic genomes, the authors performed a combination sort-seq/RNA-seq MPRA on three industrially important species of bacteria. Based on these data, they were then able to develop synthetic gene circuits that have species-specific activity.

Functional footprint: a visual summary of MPRA data that quantifies how strongly mutations at each nucleotide position affect CRE activity

Complementing the identification of novel genetic elements, Ulirsch et al. (167) and Baeza-Centurion et al. (9) have illustrated two ways in which MPRAs can be used to study genetic variation. Ulirsch et al. (167) used an MPRA to study human genomic loci that had been previously identified in genome-wide association studies of red blood cell traits. Specifically, they used an RNA-seq MPRA to assay a library of human genomic sequences, as well as common variants thereof, for CRE activity. The authors identified 32 candidate loci for follow-up validation using CRISPR/Cas9 genome editing. One validated locus was found to regulate transcription of the gene *RBM38*, which was subsequently shown to encode an important regulator of alternative mRNA splicing in terminal erythropoiesis. Baeza-Centurion et al. (9), by contrast, used an MPRA to study the effects that genomic substitutions across species have on splicing. They used an RNA-seq MPRA to measure exon inclusion rates for 3,071 variants of *FAS* exon 6, representing all combinations of the 12 substitutions that are observed in this exon across the primate lineage. From the resulting data and follow-up studies, the authors identified a remarkably consistent and widespread global epistasis nonlinearity that links sequence variation to the probability of exon inclusion.

Many of the earliest MPRAs were designed to dissect specific CREs of interest at nucleotide resolution (69, 76, 95, 111, 112), thereby providing insight into functional mechanisms. Melnikov et al. (95) used an RNA-seq MPRA to study two enhancers in this manner: a synthetic cAMP-responsive enhancer, and the human interferon-β enhancer. To this end, the authors used a combination of element scanning, systematic mutation, and scattered mutation libraries. For the cAMP-responsive enhancer, functional footprints clearly revealed the locations of binding sites for CREB, the cAMP-responsive TF that drives expression in this context. Functional footprints did not, however, resolve individual TF binding sites within the interferon-β enhancer. These divergent results likely reflect the difference between billboard enhancers and enhanceosomes (6, 150): The cAMP-responsive enhancer is of the billboard type, as it has well-separated TF binding sites that are not strongly coupled; the interferon-β enhancer, on the other hand, forms the canonical example of an enhanceosome (108), a highly structured protein–DNA complex that is easily disrupted by changes to enhancer DNA sequence.

MPRAs can also be used for biophysical studies of in vivo TF–TF interactions that occur at specific CREs of interest. Kinney et al. (69) used a sort-seq MPRA to study a region of the *E. coli lac* promoter that contains binding sites for two proteins: CRP and the σ^{70} RNA polymerase holoenzyme (RNAP). By fitting a thermodynamic model to their MPRA data, they were able to measure a value of -3.3 ± 0.4 kcal/mol for the cooperative interaction between these proteins, which allows CRP to upregulate transcription of the *lac* operon. Belliveau et al. (12) subsequently used this strategy, along with DNA affinity purification and mass spectrometry, to study *E. coli* promoters with little or no prior regulatory annotation. In doing so, they demonstrated a systematic method to identify novel TF binding sites, identify the TFs that bind those sites, and establish biophysical models for how those TFs carry out their regulatory functions. More recently, Forcier et al. (39)

described an alternative MPRA-compatible strategy that substantially increases the precision and clarity with which TF–TF interactions can be measured in vivo.

Rather than focus on a specific CRE of interest, a variety of studies have used MPRAs in attempts to identify general principles that govern *cis*-regulatory grammar. Before the advent of high-throughput DNA sequencing, multiple groups investigated the levels of gene expression produced by artificial bacterial and yeast promoters comprising random arrangements of TF binding sites (24, 44, 84). MPRAs subsequently allowed such studies to be performed in mammalian systems and on orders of magnitude more CREs (98, 149). An alternative approach for studying *cis*-regulatory grammar has been to use systematically varied synthetic CRE libraries. This strategy has been applied to a diverse range of CREs, including yeast promoters (87, 142), yeast 5′ and 3′ UTRs (31, 140), and human promoters (171). MPRAs using fully random libraries have also been used in attempts to characterize *cis*-regulatory grammar relevant for yeast promoters (28), yeast 5′ UTRs (26), and alternative splicing in human cells (129).

General studies of *cis*-regulatory grammar have used a variety of quantitative modeling strategies, including statistical models (149), thermodynamic models (44, 98), and neural network models (26, 28, 129). It remains largely unclear, however, which types of models work best in which situations. One potential way to clarify this issue would be to hold a quantitative modeling competition focused on *cis*-regulatory grammar, akin to the highly influential competition organized by Weirauch et al. (172) to assess methods for modeling TF specificity. Another potential way to validate models of *cis*-regulatory grammar is to perform follow-up studies that apply CRE-dissection MPRAs to a small number of specific CRE variants. For instance, these subsequent experiments might be able to verify model-predicted binding sites for *trans* factors, as well as interactions that are predicted by the model to occur between these *trans* factors. Indeed, combining general studies of *cis*-regulatory grammar with dissection studies on select CREs could prove to be a powerful way of using MPRAs to elucidate the complex sequence–function relationships that govern the regulation of gene expression.

7. CONCLUSION

We have reviewed recent progress in the development, analysis, and application of massively parallel assays. Although these high-throughput experiments have been used to investigate a broad range of biological phenomena, they share many key features, allowing them to be analyzed with a unified set of methods and concepts. Because variants remain pooled at each experimental step, these assays can have enormous throughput and can measure quantitative activities for thousands or even millions of variants in a single experiment. These capabilities suggest new strategies for addressing goals that previously had been barely imaginable, and here we have reviewed progress toward two such visions for contemporary genetics. These goals are quite different, and progress in multiple directions will be necessary to bring these efforts to fruition.

The first vision is to conduct comprehensive measurements of the phenotypic effects of all possible mutations to the most important and actionable human disease genes (56, 152, 169). These prospective measurements would address the problem of genomic variant interpretation by providing patients and genetic counselors with direct evidence for the molecular phenotypes of mutations whose significance would otherwise be uncertain. While the gap between the molecular impact of individual mutations and their consequences at the level of the whole organism remains a substantial challenge (23), comprehensive assays would reduce patient uncertainty by clearly distinguishing worrisome but ultimately benign mutations in disease genes from mutations that substantially affect molecular function and thus are likely to produce a disease state in at least some genetic backgrounds or environmental conditions. For this application, the key areas for progress

primarily revolve around increasing the throughput, precision, replicability, and disease relevance of these high-throughput functional assays, so as to allow for the incorporation of these assays into revised versions of the clinical guidelines for diagnosis and treatment of genetic disease (122).

The second vision is to use massively parallel assays as a general-purpose technology to probe the mechanisms underlying the functionality of any given stretch of genomic DNA. We have reviewed the application of this methodology for understanding TF specificity, protein function, and the architecture of *cis*-regulatory sequences. We have also described some of the major open issues in these applications. Whereas the key areas for progress in variant interpretation are largely experimental, here the primary barrier to progress lies in the limitations of current quantitative modeling capabilities. Simple additive models provide crude but easily interpreted summaries of the sequence–function relationships revealed by high-throughput assays, and such models may in fact be sufficient in some applications, such as identifying likely TF binding sites. However, there is a strong need for models that can capture epistatic interactions of different types within more complex genetic elements, such as proteins and CREs. Most importantly, while it is clear that existing massively parallel assays are providing an unprecedented view of the richness and complexity of sequence–function relationships, better methods are needed to derive mechanistic insights from these observations.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

We would like to thank the participants of the "Measuring and Modeling Quantitative Sequence–Function Relationships" meeting held at the Banbury Center in 2016, discussions with whom played a major role in helping to shape this review. We also apologize to our many colleagues whose work we were unable to cover due to length and citation constraints. This work was supported in part by a Cold Spring Harbor Laboratory/Northwell Health Alliance grant to J.B.K. and by National Institutes of Health Cancer Center Support Grant 5P30CA045508.

LITERATURE CITED

1. Aakre CD, Herrou J, Phung TN, Perchuk BS, Crosson S, Laub MT. 2015. Evolving new protein-protein interaction specificity through promiscuous intermediates. *Cell* 163:594–606
2. Adams RM, Mora T, Walczak AM, Kinney JB. 2016. Measuring the sequence-affinity landscape of antibodies with massively parallel titration curves. *eLife* 5:e23156
3. Alipanahi B, Delong A, Weirauch MT, Frey BJ. 2015. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nat. Biotechnol.* 33:831–38
4. 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. *PNAS* 109:16858–63
5. Arnold CD, Gerlach D, Stelzer C, Boryn LM, Rath M, Stark A. 2013. Genome-wide quantitative enhancer activity maps identified by STARR-seq. *Science* 339:1074–77
6. Arnosti D, Kulkarni M. 2005. Transcriptional enhancers: intelligent enhanceosomes or flexible billboards? *J. Cell. Biochem.* 94:890–98
7. Ashenberg O, Gong LI, Bloom JD. 2013. Mutational effects on stability are largely conserved during protein evolution. *PNAS* 110:21071–76

8. Atwal GS, Kinney JB. 2016. Learning quantitative sequence–function relationships from massively parallel experiments. *J. Stat. Phys.* 162:1203–43
9. Baeza-Centurion P, Miñana B, Schmiedel JM, Valcárcel J, Lehner B. 2019. Combinatorial genetics reveals a scaling law for the effects of mutations on splicing. *Cell* 176:549–63.e23
10. Bailey T, Elkan C. 1995. Unsupervised learning of multiple motifs in biopolymers using expectation maximization. *Mach. Learn.* 21:51–80
11. Bank C, Matuszewski S, Hietpas RT, Jensen JD. 2016. On the (un)predictability of a large intragenic fitness landscape. *PNAS* 113:14085–90
12. Belliveau NM, Barnes SL, Ireland WT, Jones DL, Sweredoski MJ, et al. 2018. Systematic approach for dissecting the molecular mechanisms of transcriptional regulation in bacteria. *PNAS* 115:E4796–805
13. Berg O, von Hippel P. 1987. Selection of DNA binding sites by regulatory proteins. Statistical-mechanical theory and application to operators and promoters. *J. Mol. Biol.* 193:723–50
14. Berg O, von Hippel P. 1988. Selection of DNA binding sites by regulatory proteins. II. The binding specificity of cyclic AMP receptor protein to recognition sites. *J. Mol. Biol.* 200:709–23
15. Berger M, Philippakis A, Qureshi A, He F, Estep P, Bulyk M. 2006. Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. *Nat. Biotechnol.* 24:1429–35
16. Bintu L, Buchler NE, Garcia HG, Gerland U, Hwa T, et al. 2005. Transcriptional regulation by the numbers: models. *Curr. Opin. Genet. Dev.* 15:116–24
17. Bloom JD. 2014. An experimentally determined evolutionary model dramatically improves phylogenetic fit. *Mol. Biol. Evol.* 31:1956–78
18. Bloom JD, Raval A, Wilke CO. 2007. Thermodynamics of neutral protein evolution. *Genetics* 175:255–66
19. Bloom JD, Silberg JJ, Wilke CO, Drummond DA, Adami C, Arnold FH. 2005. Thermodynamic prediction of protein neutrality. *PNAS* 102:606–11
20. Browning DF, Busby SJW. 2016. Local and global regulation of transcription initiation in bacteria. *Nat. Rev. Microbiol.* 14:638–50
21. Castel SE, Cervera A, Mohammadi P, Aguet F, Reverter F, et al. 2018. Modified penetrance of coding variants by cis-regulatory variation contributes to disease risk. *Nat. Genet.* 50:1327–34
22. Conesa A, Madrigal P, Tarazona S, Gomez-Cabrero D, Cervera A, et al. 2016. A survey of best practices for RNA-seq data analysis. *Genome Biol.* 17:13
23. Cooper DN, Krawczak M, Polychronakos C, Tyler-Smith C, Kehrer-Sawatzki H. 2013. Where genotype is not predictive of phenotype: towards an understanding of the molecular basis of reduced penetrance in human inherited disease. *Hum. Genet.* 132:1077–130
24. Cox RS, Surette MG, Elowitz MB. 2007. Programming gene expression with combinatorial promoters. *Mol. Syst. Biol.* 3:145
25. Crooks GE, Hon G, Chandonia JM, Brenner SE. 2004. WebLogo: a sequence logo generator. *Genome Res.* 14:1188–90
26. Cuperus JT, Groves B, Kuchina A, Rosenberg AB, Jojic N, et al. 2017. Deep learning of the regulatory grammar of yeast 5′ untranslated regions from 500,000 random sequences. *Genome Res.* 27:2015–24
27. Dantas Machado AC, Zhou T, Rao S, Goel P, Rastogi C, et al. 2015. Evolving insights on how cytosine methylation affects protein-DNA binding. *Brief. Funct. Genom.* 14:61–73
28. de Boer C, Sadeh R, Friedman N, Regev A. 2018. Deciphering eukaryotic *cis*-regulatory logic with 100 million random promoters. bioRxiv 224907. **https://doi.org/10.1101/224907**
29. Domingo J, Diss G, Lehner B. 2018. Pairwise and higher-order genetic interactions during the evolution of a tRNA. *Nature* 558:117–21
30. Doniger S, Fay J. 2007. Frequent gain and loss of functional transcription factor binding sites. *PLOS Comput. Biol.* 3:e99
31. Dvir S, Velten L, Sharon E, Zeevi D, Carey LB, et al. 2013. Deciphering the rules by which 5′-UTR sequences affect protein expression in yeast. *PNAS* 110:E2792–801

32. Elemento O, Slonim N, Tavazoie S. 2007. A universal framework for regulatory element discovery across all genomes and data types. *Mol. Cell* 28:337–50
32a. Esposito D, Weile J, Shendure J, Starita LM, Papenfuss AT, et al. 2019. An open-source platform to distribute and interpret data from multiplexed assays of variant effect. bioRxiv 555797. **https://doi.org/10.1101/555797**
33. Feng Y, Zhang Y, Ebright RH. 2016. Structural basis of transcription activation. *Science* 352:1330–33
34. Findlay GM, Boyle EA, Hause RJ, Klein JC, Shendure J. 2014. Saturation editing of genomic regions by multiplex homology-directed repair. *Nature* 513:120–23
35. Findlay GM, Daza RM, Martin B, Zhang MD, Leith AP, et al. 2018. Accurate classification of *BRCA1* variants with saturation genome editing. *Nature* 562:217–22
36. Firnberg E, Labonte JW, Gray JJ, Ostermeier M. 2014. A comprehensive, high-resolution map of a gene's fitness landscape. *Mol. Biol. Evol.* 31:1581–92
37. Firnberg E, Ostermeier M. 2012. PFunkel: efficient, expansive, user-defined mutagenesis. *PLOS ONE* 7:e52031
38. Foat B, Morozov A, Bussemaker H. 2006. Statistical mechanical modeling of genome-wide transcription factor occupancy data by MatrixREDUCE. *Bioinformatics* 22:e141–49
39. Forcier TL, Ayaz A, Gill MS, Jones D, Phillips R, Kinney JB. 2018. Measuring cis-regulatory energetics in living cells using allelic manifolds. *eLife* 7:e40618
40. Forsyth CM, Juan V, Akamatsu Y, DuBridge RB, Doan M, et al. 2013. Deep mutational scanning of an antibody against epidermal growth factor receptor using mammalian cell display and massively parallel pyrosequencing. *mAbs* 5:523–32
41. Fowler DM, Araya CL, Fleishman SJ, Kellogg EH, Stephany JJ, et al. 2010. High-resolution mapping of protein sequence-function relationships. *Nat. Methods* 7:741–46
42. Fowler DM, Fields S. 2014. Deep mutational scanning: a new style of protein science. *Nat. Methods* 11:801–7
43. Gallet R, Cooper TF, Elena SF, Lenormand T. 2012. Measuring selection coefficients below 10^{-3}: method, questions, and prospects. *Genetics* 190:175–86
44. Gertz J, Siggia ED, Cohen BA. 2009. Analysis of combinatorial *cis*-regulation in synthetic and genomic promoters. *Nature* 457:215–18
45. Ghandi M, Lee D, Mohammad-Noori M, Beer MA. 2014. Enhanced regulatory sequence prediction using gapped *k*-mer features. *PLOS Comput. Biol.* 10:e1003711
46. Goldstein RA. 2011. The evolution and evolutionary consequences of marginal thermostability in proteins. *Proteins* 79:1396–407
47. Gong LI, Suchard MA, Bloom JD. 2013. Stability-mediated epistasis constrains the evolution of an influenza protein. *eLife* 2:e00631
48. Goodfellow I, Bengio Y, Courville A. 2016. *Deep Learning*. Cambridge, MA: MIT Press
49. Grimm DG, Azencott CA, Aicheler F, Gieraths U, MacArthur DG, et al. 2015. The evaluation of tools used to predict the impact of missense variants is hindered by two types of circularity. *Hum. Mutat.* 36:513–23
50. Guenther UP, Yandek LE, Niland CN, Campbell FE, Anderson D, et al. 2013. Hidden specificity in an apparently nonspecific RNA-binding protein. *Nature* 502:385–88
51. Halpern A, Bruno W. 1998. Evolutionary distances for protein-coding sequences: modeling site-specific residue frequencies. *Mol. Biol. Evol.* 15:910–17
52. Heumann J, Lapedes A, Stormo G. 1994. Neural networks for determining protein specificity and multiple alignment of binding sites. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* 2:188–94
53. Hietpas RT, Jensen JD, Bolon DNA. 2011. Experimental illumination of a fitness landscape. *PNAS* 108:7896–901
54. Hopf TA, Ingraham JB, Poelwijk FJ, Schärfe CPI, Springer M, et al. 2017. Mutation effects predicted from sequence co-variation. *Nat. Biotechnol.* 35:128–35
55. Inukai S, Kock KH, Bulyk ML. 2017. Transcription factor-DNA binding: beyond binding site motifs. *Curr. Opin. Genet. Dev.* 43:110–19

56. Ipe J, Swart M, Burgess KS, Skaar TC. 2017. High-throughput assays to assess the functional impact of genetic variants: a road towards genomic-driven medicine. *Clin. Transl. Sci.* 10:67–77
57. Isakova A, Groux R, Imbeault M, Rainer P, Alpern D, et al. 2017. SMiLE-seq identifies binding motifs of single and dimeric transcription factors. *Nat. Methods* 14:316–22
58. Jacquier H, Birgy A, Le Nagard H, Mechulam Y, Schmitt E, et al. 2013. Capturing the mutational landscape of the beta-lactamase TEM-1. *PNAS* 110:13067–72
59. Jiang L, Schlesinger F, Davis CA, Zhang Y, Li R, et al. 2011. Synthetic spike-in standards for RNA-seq experiments. *Genome Res.* 21:1543–51
60. Johns NI, Gomes ALC, Yim SS, Yang A, Blazejewski T, et al. 2018. Metagenomic mining of regulatory elements enables programmable species-selective gene expression. *Nat. Methods* 15:323–29
61. Johnson A, Meyer BJ, Ptashne M. 1978. Mechanism of action of the *cro* protein of bacteriophage λ. *PNAS* 75:1783–87
62. Jolma A, Yan J, Whitington T, Toivonen J, Nitta KR, et al. 2013. DNA-binding specificities of human transcription factors. *Cell* 152:327–39
63. Joyce AP, Zhang C, Bradley P, Havranek JJ. 2015. Structure-based modeling of protein: DNA specificity. *Brief. Funct. Genom.* 14:39–49
64. Judson HF. 1996. *The Eighth Day of Creation: The Makers of the Revolution in Biology*. Cold Spring Harbor, NY: Cold Spring Harb. Lab. Press
65. Katsanis N. 2016. The continuum of causality in human genetic disorders. *Genome Biol.* 17:233
66. Kazan H, Ray D, Chan ET, Hughes TR, Morris Q. 2010. RNAcontext: a new method for learning the sequence and structure binding preferences of RNA-binding proteins. *PLOS Comput. Biol.* 6:e1000832
66a. Ke S, Shang S, Kalachikov SM, Morozova I, Yu L, et al. 2011. Quantitative evaluation of all hexamers as exonic splicing elements. *Genome Res.* 21:1360–74
67. Keefe AD, Szostak JW. 2001. Functional proteins from a random-sequence library. *Nature* 410:715–18
68. Kelley DR, Snoek J, Rinn JL. 2016. Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. *Genome Res.* 26:990–99
69. Kinney JB, Murugan A, Callan CG, Cox EC. 2010. Using deep sequencing to characterize the biophysical mechanism of a transcriptional regulatory sequence. *PNAS* 107:9158–63
70. Kinney JB, Tkacik G, Callan CG. 2007. Precise physical models of protein-DNA interaction from high-throughput data. *PNAS* 104:501–6
71. Kivioja T, Vähärautio A, Karlsson K, Bonke M, Enge M, et al. 2012. Counting absolute numbers of molecules using unique molecular identifiers. *Nat. Methods* 9:72–74
72. Kondrashov FA, Kondrashov AS. 2001. Multidimensional epistasis and the disadvantage of sex. *PNAS* 98:12089–92
73. Koo PK, Anand P, Paul SB, Eddy SR. 2018. Inferring sequence-structure preferences of RNA-binding proteins with convolutional residual networks. bioRxiv 418459. **https://doi.org/10.1101/418459**
74. Koo PK, Eddy SR. 2018. Representation learning of genomic sequence motifs with convolutional neural networks. bioRxiv 362756. **https://doi.org/10.1101/362756**
75. Kowalsky CA, Faber MS, Nath A, Dann HE, Kelly VW, et al. 2015. Rapid fine conformational epitope mapping using comprehensive mutagenesis and deep sequencing. *J. Biol. Chem.* 290:26457–70
76. Kwasnieski JC, Mogno I, Myers CA, Corbo JC, Cohen BA. 2012. Complex effects of nucleotide variants in a mammalian *cis*-regulatory element. *PNAS* 109:19498–503
77. Lai X, Stigliani A, Vachon G, Carles C, Smaczniak C, et al. 2019. Building transcription factor binding site models to understand gene regulation in plants. *Mol. Plant* 12:P743–63
78. Le DD, Shimko TC, Aditham AK, Keys AM, Longwell SA, et al. 2018. Comprehensive, high-resolution binding energy landscapes reveal context dependencies of transcription factor binding. *PNAS* 115:E3702–11
79. Levo M, Segal E. 2014. In pursuit of design principles of regulatory sequences. *Nat. Rev. Genet.* 15:453–68
80. Levy RM, Haldane A, Flynn WF. 2017. Potts Hamiltonian models of protein co-variation, free energy landscapes, and evolutionary fitness. *Curr. Opin. Struct. Biol.* 43:55–62

81. Li C, Qian W, Maclean CJ, Zhang J. 2016. The fitness landscape of a tRNA gene. *Science* 352:837–40
82. Li F, Salit ML, Levy SF. 2018. Unbiased fitness estimation of pooled barcode or amplicon sequencing studies. *Cell Syst.* 7:521–25.e4
83. Liachko I, Youngblood RA, Keich U, Dunham MJ. 2013. High-resolution mapping, characterization, and optimization of autonomously replicating sequences in yeast. *Genome Res.* 23:698–704
84. Ligr M, Siddharthan R, Cross F, Siggia ED. 2006. Gene expression from random libraries of yeast promoters. *Genetics* 172:2113–22
85. Long HK, Prescott SL, Wysocka J. 2016. Ever-changing landscapes: transcriptional enhancers in development and evolution. *Cell* 167:1170–87
86. Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* 15:550
87. Lubliner S, Regev I, Lotan-Pompan M, Edelheit S, Weinberger A, Segal E. 2015. Core promoter sequence in yeast is a major determinant of expression level. *Genome Res.* 25:1008–17
88. Majithia AR, Tsuda B, Agostini M, Gnanapradeepan K, Rice R, et al. 2016. Prospective functional classification of all possible missense variants in *PPARG*. *Nat. Genet.* 48:1570–75
89. Manhart M, Morozov AV. 2015. Protein folding and binding can emerge as evolutionary spandrels through structural coupling. *PNAS* 112:1797–802
90. Maricque BB, Chaudhari HG, Cohen BA. 2018. A massively parallel reporter assay dissects the influence of chromatin structure on *cis*-regulatory activity. *Nat. Biotechnol.* 37:90–95
91. Maticzka D, Lange SJ, Costa F, Backofen R. 2014. GraphProt: modeling binding preferences of RNA-binding proteins. *Genome Biol.* 15:R17
92. Matreyek KA, Starita LM, Stephany JJ, Martin B, Chiasson MA, et al. 2018. Multiplex assessment of protein variant abundance by massively parallel sequencing. *Nat. Genet.* 50:874–82
93. Matuszewski S, Hildebrandt ME, Ghenu AH, Jensen JD, Bank C. 2016. A statistical guide to the design of deep mutational scanning experiments. *Genetics* 204:77–87
94. McLaughlin RN, Poelwijk FJ, Raman A, Gosal WS, Ranganathan R. 2012. The spatial architecture of protein function and adaptation. *Nature* 491:138–42
95. Melnikov A, Murugan A, Zhang X, Tesileanu T, Wang L, et al. 2012. Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. *Nat. Biotechnol.* 30:271–77
96. Miosge LA, Field MA, Sontani Y, Cho V, Johnson S, et al. 2015. Comparison of predicted and actual consequences of missense mutations. *PNAS* 112:E5189–98
97. Miyazawa S, Jernigan R. 1985. Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation. *J. Am. Chem. Soc.* 18:534–52
98. Mogno I, Kwasnieski JC, Cohen BA. 2013. Massively parallel synthetic promoter assays reveal the in vivo effects of binding site variants. *Genome Res.* 23:1908–15
99. Morcos F, Pagnani A, Lunt B, Bertolino A, Marks DS, et al. 2011. Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *PNAS* 108:E1293–301
100. Mustonen V, Kinney JB, Callan CG, Lässig M. 2008. Energy-dependent fitness: a quantitative model for the evolution of yeast transcription factor binding sites. *PNAS* 105:12376–81
101. Nirenberg M, Leder P, Bernfield M, Brimacombe R, Trupin J, et al. 1965. RNA codewords and protein synthesis, VII. On the general nature of the RNA code. *PNAS* 53:1161–68
102. Nisthal A, Wang CY, Ary ML, Mayo SL. 2018. Protein stability engineering insights revealed by domain-wide comprehensive mutagenesis. bioRxiv 484949. **https://doi.org/10.1101/484949**
103. Oikonomou P, Goodarzi H, Tavazoie S. 2014. Systematic identification of regulatory elements in conserved 3′ UTRs of human transcripts. *Cell Rep.* 7:281–92
104. Olson CA, Wu NC, Sun R. 2014. A comprehensive biophysical description of pairwise epistasis throughout an entire protein domain. *Curr. Biol.* 24:2643–51
105. Orenstein Y, Wang Y, Berger B. 2016. RCK: accurate and efficient inference of sequence- and structure-based protein-RNA binding models from RNAcompete data. *Bioinformatics* 32:i351–59
106. Otwinowski J. 2018. Biophysical inference of epistasis and the effects of mutations on protein stability and function. *Mol. Biol. Evol.* 35:2345–54

107. Otwinowski J, McCandlish DM, Plotkin JB. 2018. Inferring the shape of global epistasis. *PNAS* 115:E7550–58
108. Panne D. 2008. The enhanceosome. *Curr. Opin. Struct. Biol.* 18:236–42
109. Panne D, Maniatis T, Harrison SC. 2007. An atomic model of the interferon-beta enhanceosome. *Cell* 129:1111–23
110. Parkinson G, Wilson C, Gunasekera A, Ebright YW, Ebright RH, et al. 1996. Structure of the CAP-DNA complex at 2.5 Å resolution: a complete picture of the protein-DNA interface. *J. Mol. Biol.* 260:395–408
111. Patwardhan RP, Hiatt JB, Witten DM, Kim MJ, Smith RP, et al. 2012. Massively parallel functional dissection of mammalian enhancers in vivo. *Nat. Biotechnol.* 30:265–70
112. 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.* 27:1173–75
113. Peterman N, Levine E. 2016. Sort-seq under the hood: implications of design choices on large-scale characterization of sequence-function relations. *BMC Genom.* 17:206
114. Phillips R, Kondev J, Theriot J, Garcia HG. 2012. *Physical Biology of the Cell*. New York: Garland Sci. 2nd ed.
115. Plesa C, Sidore AM, Lubock NB, Zhang D, Kosuri S. 2018. Multiplexed gene synthesis in emulsions for exploring protein functional landscapes. *Science* 359:343–47
116. Podgornaia AI, Laub MT. 2015. Pervasive degeneracy and epistasis in a protein-protein interface. *Science* 347:673–77
117. Pribnow D. 1975. Nucleotide sequence of an RNA polymerase binding site at an early T7 promoter. *PNAS* 72:784–88
118. Puchta O, Cseke B, Czaja H, Tollervey D, Sanguinetti G, Kudla G. 2016. Network of epistatic interactions within a yeast snoRNA. *Science* 352:840–44
119. Raraigh KS, Han ST, Davis E, Evans TA, Pellicore MJ, et al. 2018. Functional assays are essential for interpretation of missense variants associated with variable expressivity. *Am. J. Hum. Genet.* 102:1062–77
120. Ray D, Kazan H, Chan ET, Pena-Castillo L, Chaudhry S, et al. 2009. Rapid and systematic analysis of the RNA recognition specificities of RNA-binding proteins. *Nat. Biotechnol.* 27:667–70
121. Reich LL, Dutta S, Keating AE. 2015. SORTCERY—a high-throughput method to affinity rank peptide ligands. *J. Mol. Biol.* 427:2135–50
122. Richards S, Aziz N, Bale S, Bick D, Das S, et al. 2015. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet. Med.* 17:405–14
123. Riesselman AJ, Ingraham JB, Marks DS. 2018. Deep generative models of genetic variation capture the effects of mutations. *Nat. Methods* 15:816–22
124. Riley TR, Lazarovici A, Mann RS, Bussemaker HJ. 2015. Building accurate sequence-to-affinity models from high-throughput in vitro protein-DNA binding data using FeatureREDUCE. *eLife* 4:e06397
125. Rodenburg RJ. 2018. The functional genomics laboratory: functional validation of genetic variants. *J. Inherit. Metab. Dis.* 41:297–307
126. Rodrigue N, Philippe H, Lartillot N. 2010. Mutation-selection models of coding sequence evolution with site-heterogeneous amino acid fitness profiles. *PNAS* 107:4629–34
127. Rohs R, Jin X, West SM, Joshi R, Honig B, Mann RS. 2010. Origins of specificity in protein-DNA recognition. *Annu. Rev. Biochem.* 79:233–69
128. Rollins NJ, Brock KP, Poelwijk FJ, Stiffler MA, et al. 2018. 3D protein structure from genetic epistasis experiments. bioRxiv 320721. **https://doi.org/10.1101/320721**
129. Rosenberg AB, Patwardhan RP, Shendure J, Seelig G. 2015. Learning the sequence determinants of alternative splicing from millions of random sequences. *Cell* 163:698–711
130. Ruan S, Stormo GD. 2017. Inherent limitations of probabilistic models for protein-DNA binding specificity. *PLOS Comput. Biol.* 13:e1005638

131. Ruan S, Swamidass SJ, Stormo GD. 2017. BEESEM: estimation of binding energy models using HT-SELEX data. *Bioinformatics* 33:2288–95
132. Rubin AF, Gelman H, Lucas N, Bajjalieh SM, Papenfuss AT, et al. 2017. A statistical framework for analyzing deep mutational scanning data. *Genome Biol.* 18:150
133. Sailer ZR, Harms MJ. 2017. Detecting high-order epistasis in nonlinear genotype-phenotype maps. *Genetics* 205:1079–88
134. Salinas VH, Ranganathan R. 2018. Coevolution-based inference of amino acid interactions underlying protein function. *eLife* 7:e34300
135. Santos-Zavaleta A, Sánchez-Pérez M, Salgado H, Velázquez-Ramírez DA, Gama-Castro S, et al. 2018. A unified resource for transcriptional regulation in *Escherichia coli* K-12 incorporating high-throughput-generated binding data into RegulonDB version 10.0. *BMC Biol.* 16:91
136. Sarkisyan KS, Bolotin DA, Meer MV, Usmanova DR, Mishin AS, et al. 2016. Local fitness landscape of the green fluorescent protein. *Nature* 533:397–401
137. Schmiedel J, Lehner B. 2018. Determining protein structures using genetics. bioRxiv 303875. **https://doi.org/10.1101/303875**
138. Schneider TD, Stephens RM. 1990. Sequence logos: a new way to display consensus sequences. *Nucl. Acids Res.* 18:6097–100
139. Segal E, Fondufe-Mittendorf Y, Chen L, Thastrom A, Field Y, et al. 2006. A genomic code for nucleosome positioning. *Nature* 442:772–78
140. Shalem O, Sharon E, Lubliner S, Regev I, Lotan-Pompan M, et al. 2015. Systematic dissection of the sequence determinants of gene 3′ end mediated expression control. *PLOS Genet.* 11:e1005147
141. Sharon E, Chen SAA, Khosla NM, Smith JD, Pritchard JK, Fraser HB. 2018. Functional genetic variants revealed by massively parallel precise genome editing. *Cell* 175:544–57.e16
142. Sharon E, Kalma Y, Sharp A, Raveh-Sadka T, Levo M, et al. 2012. Inferring gene regulatory logic from high-throughput measurements of thousands of systematically designed promoters. *Nat. Biotechnol.* 30:521–30
143. Sharon E, Lubliner S, Segal E. 2008. A feature-based approach to modeling protein-DNA interactions. *PLOS Comput. Biol.* 4:e1000154
144. Sherman MS, Cohen BA. 2012. Thermodynamic state ensemble models of *cis*-regulation. *PLOS Comput. Biol.* 8:e1002407
145. Shine J, Dalgarno L. 1975. Determinant of cistron specificity in bacterial ribosomes. *Nature* 254:34–38
146. Shrikumar A, Greenside P, Kundaje A. 2017. Learning important features through propagating activation differences. In *Proceedings of the 34th International Conference on Machine Learning*, pp. 3145–53. Proc. Mach. Learn. Res. Vol. 70. N.p.: PMLR
147. Siggers T, Duyzend MH, Reddy J, Khan S, Bulyk ML. 2011. Non-DNA-binding cofactors enhance DNA-binding specificity of a transcriptional regulatory complex. *Mol. Syst. Biol.* 7:555
148. Slattery M, Riley T, Liu P, Abe N, Gomez-Alcala P, et al. 2011. Cofactor binding evokes latent differences in DNA binding specificity between Hox proteins. *Cell* 147:1270–82
149. Smith RP, Taher L, Patwardhan RP, Kim MJ, Inoue F, et al. 2013. Massively parallel decoding of mammalian regulatory sequences supports a flexible organizational model. *Nat. Genet.* 45:1021–28
150. Spitz F, Furlong EEM. 2012. Transcription factors: from enhancer binding to developmental control. *Nat. Rev. Genet.* 13:613–26
151. Staller MV, Holehouse AS, Swain-Lenz D, Das RK, Pappu RV, Cohen BA. 2018. A high-throughput mutational scan of an intrinsically disordered acidic transcriptional activation domain. *Cell Syst.* 6:444–55.e6
152. Starita LM, Ahituv N, Dunham MJ, Kitzman JO, Roth FP, et al. 2017. Variant interpretation: functional assays to the rescue. *Am. J. Hum. Genet.* 101:315–25
153. Starita LM, Islam MM, Banerjee T, Adamovich AI, Gullingsrud J, et al. 2018. A multiplex homology-directed DNA repair assay reveals the impact of more than 1,000 BRCA1 missense substitution variants on protein function. *Am. J. Hum. Genet.* 103:498–508

154. Starita LM, Young DL, Islam M, Kitzman JO, Gullingsrud J, et al. 2015. Massively parallel functional analysis of BRCA1 RING domain variants. *Genetics* 200:413–22
155. Starr TN, Flynn JM, Mishra P, Bolon DNA, Thornton JW. 2018. Pervasive contingency and entrenchment in a billion years of Hsp90 evolution. *PNAS* 115:4453–58
156. Starr TN, Picton LK, Thornton JW. 2017. Alternative evolutionary histories in the sequence space of an ancient protein. *Nature* 549:409–13
157. Starr TN, Thornton JW. 2016. Epistasis in protein evolution. *Protein Sci.* 25:1204–18
158. Stormo GD. 2013. Modeling the specificity of protein-DNA interactions. *Quant. Biol.* 1:115–30
159. Stormo GD, Fields D. 1998. Specificity, free energy and information content in protein-DNA interactions. *Trends Biochem. Sci.* 23:109–13
160. Stormo GD, Zhao Y. 2010. Determining the specificity of protein-DNA interactions. *Nat. Rev. Genet.* 11:751–60
161. Struhl K, Segal E. 2013. Determinants of nucleosome positioning. *Nat. Rev. Microbiol.* 20:267–73
162. Sun S, Yang F, Tan G, Costanzo M, Oughtred R, et al. 2016. An extended set of yeast-based functional assays accurately identifies human disease mutations. *Genome Res.* 26:670–80
163. Tamuri AU, Goldman N, dos Reis M. 2014. A penalized-likelihood method to estimate the distribution of selection coefficients from phylogenetic data. *Genetics* 197:257–71
164. Tang H, Thomas PD. 2016. Tools for predicting the functional impact of nonsynonymous genetic variation. *Genetics* 203:635–47
164a. Tareen A, Kinney JB. 2019. Logomaker: beautiful sequence logos in Python. bioRxiv 635029. **https://doi.org/10.1101/635029**
165. Thyagarajan B, Bloom JD. 2014. The inherent mutational tolerance and antigenic evolvability of influenza hemagglutinin. *eLife* 3:e03300
166. Tuerk C, Gold L. 1990. Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. *Science* 249:505–10
167. Ulirsch JC, Nandakumar SK, Wang L, Giani FC, Zhang X, et al. 2016. Systematic functional dissection of common genetic variation affecting red blood cell traits. *Cell* 165:1530–45
168. Vvedenskaya IO, Zhang Y, Goldman SR, Valenti A, Visone V, et al. 2015. Massively systematic transcript end readout, "MASTER": transcription start site selection, transcriptional slippage, and transcript yields. *Mol. Cell* 60:953–65
169. Weile J, Roth FP. 2018. Multiplexed assays of variant effects contribute to a growing genotype-phenotype atlas. *Hum. Genet.* 17:241–14
170. Weile J, Sun S, Cote AG, Knapp J, Verby M, et al. 2017. A framework for exhaustively mapping functional missense variants. *Mol. Syst. Biol.* 13:957
171. Weingarten-Gabbay S, Nir R, Lubliner S, Sharon E, Kalma Y, et al. 2019. Systematic interrogation of human promoters. *Genome Res.* 29:171–83
172. Weirauch MT, Cote A, Norel R, Annala M, Zhao Y, et al. 2013. Evaluation of methods for modeling transcription factor sequence specificity. *Nat. Biotechnol.* 31:126–34
173. Wong MS, Kinney JB, Krainer AR. 2018. Quantitative activity profile and context dependence of all human 5′ splice sites. *Mol. Cell* 71:1012–26.e3
174. Wrenbeck EE, Klesmith JR, Stapleton JA, Adeniran A, Tyo KEJ, Whitehead TA. 2016. Plasmid-based one-pot saturation mutagenesis. *Nat. Methods* 13:928–30
175. Wu NC, Dai L, Olson CA, Lloyd-Smith JO, Sun R. 2016. Adaptation in protein fitness landscapes is facilitated by indirect paths. *eLife* 5:e16965
176. Wu NC, Olson CA, Sun R. 2016. High-throughput identification of protein mutant stability computed from a double mutant fitness landscape. *Protein Sci.* 25:530–39
177. Wu NC, Young AP, Al-Mawsawi LQ, Olson CA, Feng J, et al. 2014. High-throughput profiling of influenza A virus hemagglutinin gene at single-nucleotide resolution. *Sci. Rep.* 4:4942
178. Wylie CS, Shakhnovich EI. 2011. A biophysical protein folding model accounts for most mutational fitness effects in viruses. *PNAS* 108:9916–21
179. Xu DJ, Noyes MB. 2015. Understanding DNA-binding specificity by bacteria hybrid selection. *Brief. Funct. Genom.* 14:3–16

180. Yeo G, Burge CB. 2004. Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. *J. Comput. Biol.* 11:377–94
181. Zeldovich KB, Chen P, Shakhnovich EI. 2007. Protein stability imposes limits on organism complexity and speed of molecular evolution. *PNAS* 104:16152–57
182. Zhang J, Kinch LN, Cong Q, Weile J, Sun S, et al. 2017. Assessing predictions of fitness effects of missense mutations in SUMO-conjugating enzyme UBE2I. *Hum. Mutat.* 38:1051–63
183. Zhao Y, Granas D, Stormo GD. 2009. Inferring binding energies from selected binding sites. *PLOS Comput. Biol.* 5:e1000590
184. Zhou J, Troyanskaya OG. 2015. Predicting effects of noncoding variants with deep learning-based sequence model. *Nat. Methods* 12:931–34
185. Zhou T, Shen N, Yang L, Abe N, Horton J, et al. 2015. Quantitative modeling of transcription factor binding specificities using DNA shape. *PNAS* 112:4654–59
186. Zhou T, Yang L, Lu Y, Dror I, Dantas Machado AC, et al. 2013. DNAshape: a method for the high-throughput prediction of DNA structural features on a genomic scale. *Nucl. Acids Res.* 41:W56–62
187. Zuo Z, Stormo GD. 2014. High-resolution specificity from DNA sequencing highlights alternative modes of lac repressor binding. *Genetics* 198:1329–43