INVITED PAPER

Machine Learning in Genomic Medicine: A Review of Computational Problems and Data Sets

This paper provides an introduction to machine learning tasks that are relevant to important problems in genomic medicine.

By Michael K. K. Leung, Andrew Delong, Babak Alipanahi, and Brendan J. Frey

ABSTRACT | In this paper, we provide an introduction to machine learning tasks that address important problems in genomic medicine. One of the goals of genomic medicine is to determine how variations in the DNA of individuals can affect the risk of different diseases, and to find causal explanations so that targeted therapies can be designed. Here we focus on how machine learning can help to model the relationship between DNA and the quantities of key molecules in the cell, with the premise that these quantities, which we refer to as cell variables, may be associated with disease risks. Modern biology allows high-throughput measurement of many such cell variables, including gene expression, splicing, and proteins binding to nucleic acids, which can all be treated as training targets for predictive models. With the growing availability of large-scale data sets and advanced computational techniques such as deep learning, researchers can help to usher in a new era of effective genomic medicine.

KEYWORDS | Computational biology; deep learning; genetic variants; genome analysis; genome biology; genomic medicine; machine learning; precision medicine

Manuscript received February 17, 2015; revised July 22, 2015; accepted September 3, 2015. Date of publication December 4, 2015; date of current version December 18, 2015.

M. K. K. Leung, **A. Delong**, and **B. Alipanahi** are with the Department of Electrical and Computer Engineering, University of Toronto, Toronto, ON M5S 3G4, Canada (e-mail: mleung@psi.toronto.edu; andrew@psi.toronto.edu; babak@psi.toronto.edu). **B. J. Frey** is with the Department of Electrical and Computer Engineering, University of Toronto, Toronto, ON M5S 3G4, Canada and also with the Program on Genetic Networks and the Program on Neural Computation & Adaptive Perception, Canadian Institute for Advanced Research, Toronto, ON M5G 1Z8, Canada (e-mail: frey@psi.toronto.edu).

Digital Object Identifier: 10.1109/JPROC.2015.2494198

I. INTRODUCTION

Here, we describe how machine learning can be used to solve key problems in genomic medicine. Genomics is the study of the function and information structure encoded in the DNA sequences of living cells, whereas precision medicine is the practice of tailoring treatment based on all relevant information about the patient, including the patient's genome. Each of these disciplines is undergoing an explosion of growth, especially in terms of data [1]–[4]. We see these problem domains as prime opportunities to develop "machine learning that matters" [5], improving the longevity and quality of life for the millions of individuals suffering from a genetic disease or disorder, both now and in years to come.

A genome is an instruction book for building an organism. Since 1953 it has been understood that DNA molecules are the physical medium of genetic information storage [6], and by 2001 the Human Genome Project had drafted the raw information content of a typical human genome [7], [8]. However, the bigger challenge was to interpret the structure, function, and meaning of the genetic information itself. Biologist Eric Lander summarized the situation in seven words: "Genome. Bought the book. Hard to read." Still, much is known about how genetic information is organized into distinct genes. Each gene is a like a chapter in the instruction book, describing how to build a particular family of molecules. So-called protein-coding genes describe how to build large molecules made from amino-acid chains (proteins), whereas noncoding genes describe how to build small molecules made from ribonucleic acid (RNA) chains; see [9] and [10] for introductions to molecular genetics and cell biology.

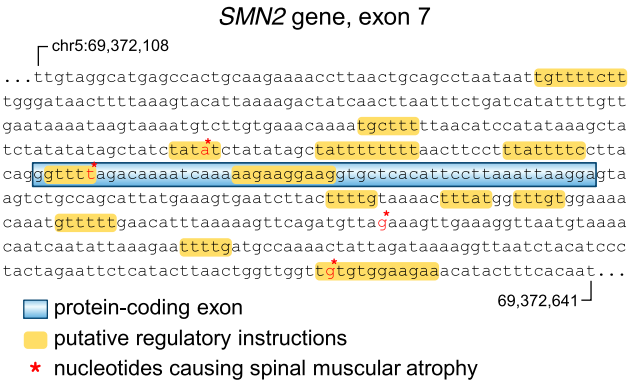


Fig. 1. ***Exon and the regulatory instructions identified using machine learning. If an infant is homozygous in a version of the survival motor neuron gene SMN2, the result is spinal muscular atrophy, a leading cause of infant mortality. Three of the nucleotides lie within genomic instructions that a machine learning technique identified as being important for including this exon when building the protein [44].***

Roughly, the human genome contains 20 000 protein-coding genes [11], and 25 000 noncoding genes [12]. Some genes are crucial for life, some are crucial for health, and some can be deleted in their entirety without apparent harm.

One of the most important information structures within a typical gene is the presence of alternating regions called introns and exons. The boundaries between these regions are determined by patterns in the nucleotide sequence, and many disease-causing mutations act by disrupting these patterns. Spinal muscular atrophy (SMA), which is the leading genetic cause of infant mortality in North America [13], results if a baby's genome is missing the *SMN1* gene, or contains a damaged version of it, resulting in deficient production of the survival motor neuron (SMN) protein. Another version of the gene, called *SMN2*, can compensate for the production of the SMN protein. Fig. 1 shows the nucleotide sequence from the seventh exon of the protein-coding gene *SMN2*. Due to differences in nucleotides at the four positions shown, the cell's machinery fails to recognize the exon, resulting in a protein that does not function properly, thereby unable to compensate for the production of the SMN protein. Researchers are evaluating therapies that restore function of exon 7 in *SMN2* [14], [15]. SMA is well studied and can be diagnosed by outward symptoms, but genetic testing is crucial for confirmation and therapeutic development. In other genetic diseases, the causal mechanisms are more complex. Cancer is a prime example of a heterogeneous disease, i.e., a disease with multiple causal pathways all leading to similar symptoms but requiring different treatments [16]. For cancer, genomic data are becoming essential for providing more detailed diagnoses and targeted treatments [17].

The concept of precision medicine is not entirely new; doctors have been using blood type to tailor blood transfusions for over a century [18]. What is different today is the rapid growth in genomic data that can be quickly and cheaply collected from the patient and the wider community, and the potential for insights from sharing that data. The scale and complexity of genomic data dwarfs the 20–50 measurements that are traditionally used in laboratory tests [17]. In this paper, we focus on machine learning applications in genomic medicine, where one assesses genomic characteristics to find targeted therapies or match existing ones, and to identify disease risks for potential preventative measures.

It is our view that to make genomic medicine a reality, we must develop computer systems that can accurately interpret the text of the genome just as the machinery inside the cell does. While this is a difficult challenge, it will enable the effects of genetic variation and potential therapies to be explored quickly, cheaply, and more accurately than can be achieved using laboratory experiments and model organisms.

What is the current state of the art in genomic medicine? Currently, protein-coding exons are the most understood regions in the genome. The universal genetic code for proteins was experimentally confirmed over 50 years ago [19], and knowing how a coding mutation changes the corresponding amino-acid sequence is a standard feature in genome diagnostic pipelines. For example, if a mutation introduces a "stop codon" into the sequence (called a "nonsense" mutation) then it is known that the protein will be truncated as a general rule. However, predicting whether a mutation will disrupt the stability or structure of the final protein molecule is a long-standing open problem [20]. Furthermore, coding regions make up only ~1.5% of the human genome, even though there is evidence that at least ~5.5% of positions undergo purifying selection [21]. Disease-causing mutations are increasingly being found outside of protein-coding regions [22], indicating that analysis tools for coding regions are not enough. Many of the functional noncoding positions are regulatory sequences, meaning they instruct the cell how to regulate important processes such as gene expression and the reliable identification of exons. This underscores the importance of developing computational models that can automatically identify and understand regulatory instructions in the genome, such as those in Fig. 1. These regulatory elements contribute significantly to the complexity of cell biology, which cannot be accounted for only by the sheer number of genes (e.g., balsam poplar trees have twice as many genes as humans [23]) or the coding regions themselves (e.g., less than 1% of human genes have coding regions that are distinct from those of mice and dogs [24]).

How can we learn to "read the genome"? Unlike familiar cognitive tasks such as visual object detection and speech recognition, humans are not naturally equipped to perceive and interpret genomic sequences nor to understand all the mechanisms, pathways, and interactions that go on inside a living cell. To make headway, a system with

superhuman analytical ability will be required. As described below, a few research groups with sufficient machine learning and genome biology expertise have been developing techniques that can interpret the genome. There are huge ongoing opportunities for machine learning researchers to contribute in this area. In genomics and biology, it is now being acknowledged that resources would be better spent on new computational techniques rather than on pure data collection—something computational biologists have been arguing for years ("The cultural baggage of biology, that privileges data generation over all other forms of science, is holding us back" [2]). For instance, after nearly $1 billion was spent on The Cancer Genome Atlas (TCGA) project, researchers question whether the project should continue to focus on sequencing or shift to analyze function [25].

Computer systems that can read the text of the genome can be used in a variety of ways to support genomic medicine. For example, a recent breakthrough in "gene editing" is allowing scientists to alter the genomes of already living cells, with an efficacy no one thought possible just a few years ago. Gene therapies can now include targeted modifications, such as removing deleterious mutations or even inserting new sequences at predetermined locations in a genome. Genome editing technology [26], [27] opens a door to unprecedented opportunities in genomic medicine, making it more important than ever that we can predict the effects of these edits *in silico*. In other words, knowing how to write is not the same as knowing what to write.

II. USING MACHINE LEARNING TO INTERPRET THE GENOME

Predicting phenotypes (e.g., traits and disease risks) from biomarkers such as the genome is, in principle, a supervised machine learning problem. The inputs are a stretch of DNA sequence (genotype) relevant to the underlying biology, and the outputs are the phenotypes. This approach [Fig. 2(a)] is not ideal for most complex phenotypes and diseases for two reasons. First is the sheer complexity of the relationship between a full genotype and its phenotype. Even within a single cell, the genome directs the state of the cell through many layers of intricate and interconnected biophysical processes and control mechanisms that have been shaped *ad hoc* by evolution. Attempting to infer the outcomes of these complex regulatory processes by observing only genomes and phenotypes is rather like trying to learn how computer chess playing programs work by examining binary code and wins and losses, while ignoring which moves were taken. Second, even if one could infer such models (those that are predictive of disease risks), it is likely that the hidden variables of these models would not correspond to biological mechanisms that can be acted upon. Insight into disease mechanisms is important for the purpose of developing targeted therapies, but can also provide complementary information for phenotypic screens, which traditionally identifies chemicals with desired biological effects without knowledge of the precise targets [28].

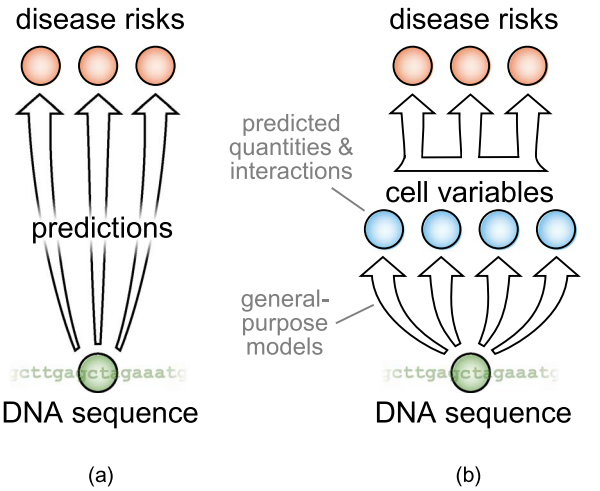


Fig. 2. ***(a) One major goal of genomic medicine is to predict phenotypes, such as disease risks, from a genotype. (b) By training models that predict how genotype defined by as a stretch of DNA sequence influences "cell variables," such as concentrations of proteins, it hugely simplifies and modularizes the machine learning problem, and enables the exploration of therapies that target these crucial variables.***

We follow what we believe is a more powerful approach, where the computational model is trained to predict measurable intermediate cell variables, also known as molecular phenotypes, first, and then these variables can be linked to phenotype [Fig. 2(b)]. For example, in the case of spinal muscular atrophy described above, the cell variable could be the frequency with which the exon is included when the gene is being copied to make a protein. Other examples of cell variables include the locations where a protein binds to a strand of DNA containing a gene, the number of copies of a gene (transcripts) in a cell, the distribution of proteins along the transcript, and concentration of proteins. Examples of cell variables are described in the next section.

This approach addresses the two aforementioned problems. Since these cell variables are more closely related to and more easily determined from genomic sequences than are phenotypes, learning models that map from DNA to cell variables can be more straightforward. High-throughput assay technologies are generating massive amounts of data profiling these cell variables under diverse conditions, and these data sets can be used to train larger and more accurate models. Also, since the cell variables correspond to intermediate biochemically active quantities, such as the concentration of a gene transcript, they are good targets for therapies. If high disease risk is associated with a change in a cell variable compared to a

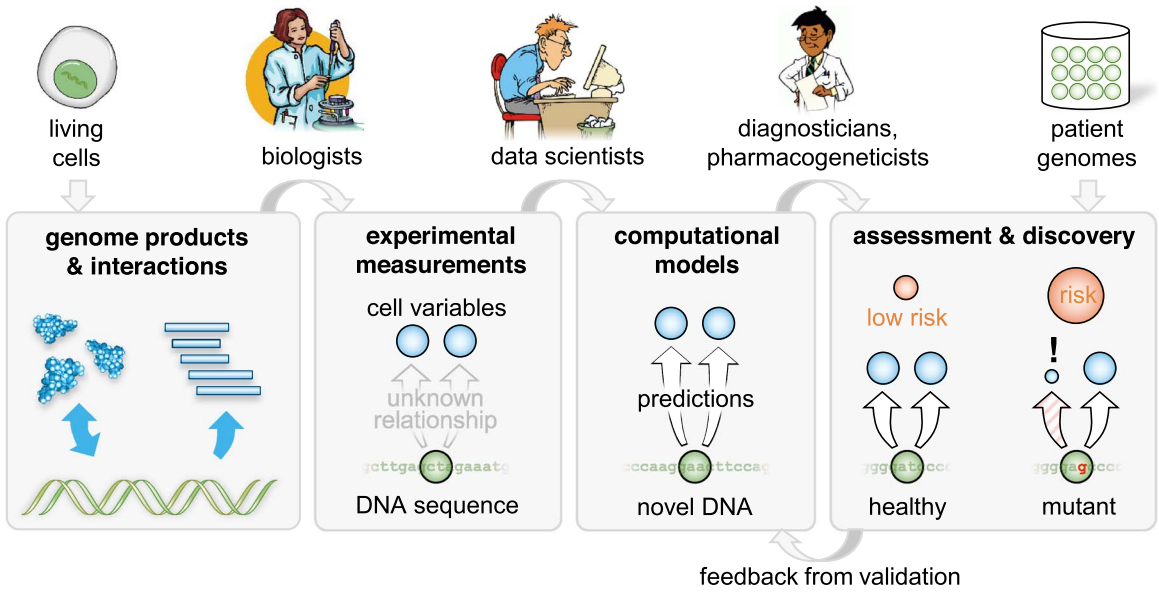


Fig. 3. *Simplified view of how biologists, data scientists, and medical researchers can work toward genomic medicine. Machine learning plays a central role by turning high-throughput measurements into specialized or general-purpose predictive models for what we referred to as "cell variables"—quantities that are relevant to cell function. By knowing how mutations affect disease via cell variables, diagnosticians and pharmacogeneticists can more easily find direct correlates with disease, develop treatments, and plan targeted therapies for individual patients.*

healthy individual, an effective therapy may consist of restoring that cell variable to its normal state. In the above example of spinal muscular atrophy, therapies that modify the genomic instructions so as to increase the frequency with which the exon is included in the protein are currently being tested in clinical trials [14].

The remainder of this paper provides an overview of the role of machine learning in building computational models of cell variables, with an aim to understand the genetic determinants of disease. We will advocate the approach of learning to model cell variables as an intermediate step, and explain how this benefits from the growing availability of diverse types of data. We will describe in detail two types of cell variable that our group has been closely involved in modeling, and briefly summarize how this research has impacted our understanding of spinal muscular atrophy, cancer, and autism spectrum disorder. To place our approach in context, we will review existing techniques that are used for scoring disease risks. Also, we will describe data sets and machine learning formulations of problems that enable data scientists to work in this hugely important area.

III. CELL BIOLOGY, MACHINE LEARNING, AND GENOMIC MEDICINE

In this section, we describe the workflow through which different actors participate toward the goal of genomic medicine, which is summarized in Fig. 3.

To build a computational model of a particular cell variable, an assay to measure the corresponding biological quantity must exist, and training data must be collected under many conditions. Well into the 1990s, biological assays typically required several manual steps and generated small amounts of data. Such techniques are useful for developing and testing hypotheses, but do not provide sufficient data to infer accurate predictive models of complex outcomes. With the commoditization of high-throughput assay technologies, it is now commonplace to acquire hundreds of thousands of measurements for a cell variable in a single low-cost experiment. For example, microarray technology has been used to peer into living cells for decades [29], but new assays and new chemistry are still being developed around this fundamental approach, such as universal protein binding microarrays (PBMs) [30], [31], ChIP-chip [32], [33], and RNAcompete [34], [35]. High-throughput sequencing technologies are likewise being used for a wide range of tasks [36]: identifying protein binding sites, sequencing the genomes of different organisms in evolutionary studies, and profiling the genomes of individuals in medical studies for the purpose of discovering variations, either in regions of interest or across the entire genome.

In addition to measuring genotypes on a large scale, high-throughput technologies can be used to measure cell variables, such as the abundances of different transcripts [37]. Although somatic mutations, which are alteration in the DNA after conception, can occur in cancers and some

neurological diseases [38], [39], the genome of an individual is relatively stable. The "transcriptome," on the other hand, varies from cell to cell, and is affected by the cell's surrounding environment, for example, the tissue type it represents. Previously, microarrays were used to measure transcripts on a large scale, but now high-throughput sequencing is the method of choice. Another application of high-throughput sequencing is to profile how proteins interact with specific regions of DNA [40]. Binding of proteins can influence how the instructions in the genome are utilized, offering a layer of complexity that can be exploited for regulation of cell biology. Data such as these, which measure particular cell variables of interest, allow us to peer into the underlying workings of the cell, at the most fundamental level of the instructions that define an organism. High-throughput assay technologies have made it feasible to measure cell variables of interest covering vast portions of the genome at various cell states, including disease conditions, and a wealth of data is now publicly available. This presents an exceptional opportunity for data scientists to infer predictive models of cell variables using machine learning techniques.

The inputs to the computational model include sequence characteristics from a stretch of DNA, such as the frequency of particular nucleotides or presence of certain motifs, some of which can be learned from the sequence themselves [41]. To account for instructions encoded in the DNA that impact cell variables through biochemical processes and structures, additional features can be derived, for example, the binding of proteins to DNA and RNA, nucleosome positioning and occupancy profiles [42], and RNA secondary structures [43]. Generally, it is beneficial to require that the model's inputs be extractable from DNA sequences. For a computational model to be useful in making predictions in the context of genomic medicine, it is desirable for the inputs to be easily obtainable. Given that the cost of whole genome sequencing continues to rapidly decrease, a growing number of genomes will be available for training purposes, and within the context of genomic medicine, it will likely become standard for a patient's genome to be available.

An important aspect of the approach illustrated in Fig. 3 is the use of machine learning to infer models that are capable of generalizing to new genetic contexts. For example, we may infer a model using the publicly available reference genome and data profiling transcripts in healthy tissues, but then apply it to the genome of a diseased cell and ascertain how the distribution of transcripts changes in the diseased cell. This notion of generalization is a crucial aspect of the models that need to be inferred. From a modeling perspective, we expect a greater ability to generalize to new genetic context for those cell states that were observed during training. Consequently, an important aspect of model development is validation using DNA sequences that the model has never seen before and using data for cell states that are different from those used during training. We cannot expect the models to be accurate for any DNA sequences and cell states that are extremely different from those used during training, so the validation procedure should also attempt to characterize the inputs for which the model is reliable.

If a model is good at generalization, it can analyze mutated DNA sequences that lead to changes in cell variables that may be indicative of disease state, without needing experimental measurements from diseased cells. In practice, this kind of "zero-shot" learning has been successfully used to identify mutations that cause a variety of diseases, using a model that was trained using the reference genome and healthy tissues [44].

While the model that predicts cell variables does not directly take into account information pertaining to disease, if the model accurately reflects how the instructions in the genome are processed, then it should be able to detect diseases that are caused by mutations that change cell variables. This approach has been shown to work very well for a large number of mutations and disease [44], but of course it makes errors. If mutations are scored by how much they cause a change in the cell variables, then false positives will arise when a mutation causes a large change in cell variables that have no impact on disease, for example, a mutation that changes a cell variable, which leads to a change in hair color. False negatives will arise for mutations that act through cell variables that are not being modeled. Both kinds of error will also arise due to inaccuracies in the computational models. When investigating specific diseases, scores for mutations can be combined with disease-specific data, such as population data. In this way, sets of candidate mutations can be filtered to identify the ones that are most likely to have a causal effect on a cell variable. More generally, these scores can be used as input features for models that are specific to certain diseases, where the models may utilize many such scores across multiple regions in the genome.

After reviewing the processes involved in gene expression, we provide concrete examples of this approach. The focus is on splicing and protein nucleic acid binding, but there are a variety of cell variables that are relevant to disease, such as transcription rate [45], DNA methylation [46], [47], polyadenylation [48], chromatin structure [49], [50], RNA folding [51], and protein folding [52].

A. Gene Expression

During gene expression, the gene is first copied (transcribed) to make a messenger RNA (mRNA) and then the mRNA is translated to make a protein. The DNA sequence containing both exons and introns is first transcribed into RNA, which is referred to as the precursor mRNA (pre-mRNA), as shown in Fig. 4. The term "precursor" refers to the fact that the pre-mRNA needs to be further processed within the nucleus to make a mature mRNA. Various modifications take place during RNA processing, one of which is splicing [53]. Splicing

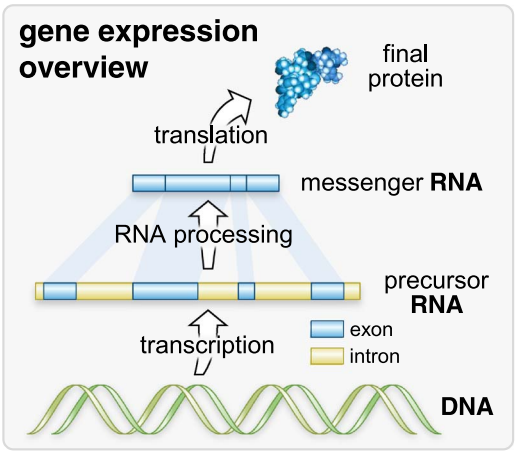


Fig. 4. *Gene expression consists of three high-level steps. Transcription creates an RNA molecule that is essentially a copy of the DNA in the gene being transcribed; at this stage, the RNA molecule is called precursor messenger RNA (pre-mRNA). RNA processing then modifies the pre-mRNA, which includes splicing out long stretches of sequence called introns and connecting the flanking regions called exons; at this stage, the RNA molecule is called messenger RNA (mRNA). Translation creates a protein molecule (an amino-acid chain) by reading the three-letter "codes" in the mRNA sequence. Other processes include polyadenylation, wherein adenine bases are appended to the end of the mRNA; mRNA stabilization, wherein the mRNA molecule is processed so as to make it less likely to degrade; mRNA localization, wherein the mRNA is moved to a location suitable for translation; and protein localization, wherein the protein is moved to a specific type of location in the cell.*

removes the introns from the pre-mRNA and connects the exons together. Another step is polyadenylation, which appends a sequence of adenine bases to the end of the mRNA [48]. In the standard model, splicing removes introns and retains all exons, as illustrated in Fig. 4, but most genes may be spliced in different ways, so that exons are sometimes removed and/or introns are retained, which increases the variety of proteins. Splicing is a critical cellular process that is deeply integrated in the gene regulatory networks [54]. Splicing may occur after transcription is complete, but it frequently occurs concurrently with transcription, so that the processes of transcription and splicing interact [55]. Finally, the mRNA is transported out of the nucleus to a ribosome, which translates the mRNA into protein. These are the main processes involved in gene expression, but others include stabilization of the mRNA and localization of the protein.

B. A Computational Model of Splicing

In spite of being able to generate highly complex tissues and organs, the human genome contains only ~20 000 genes [11], in addition to other functional molecules, such as microRNAs. One way that it achieves complexity is through the use of instructions to direct how single genes can perform different functions by splicing the same gene in different ways. In alternative splicing (AS), some exons may be removed during splicing, depending on the cellular context, such as the tissue the cell is in, and the genomic instructions that are nearby. Control of this process is called "splicing regulation" and it depends on complex interactions between numerous genomic elements within the DNA and the pre-mRNA, which are features near the site of regulation, and transfactors, which are proteins and other molecules that interact with the genomic elements [56]. Given that there are multiple exons within a gene, AS can connect exons in different ways and produce different protein isoforms (Fig. 5, top), thus "expanding" the protein repertoire of the cell [57]. The latest statistics indicate that the average number of transcripts per protein coding gene is approximately four [11]. In a notable instance, alternative splicing of exon 6 of Interleukin 7 receptor, IL7R, can render the translated protein water soluble or water insoluble and membrane localized [58]. The importance of AS is supported by evidence that at least 95% of human multiple-exon genes are alternatively spliced [59]. Many human diseases have been found to be affected by mutations in the instructions that regulate splicing [60].

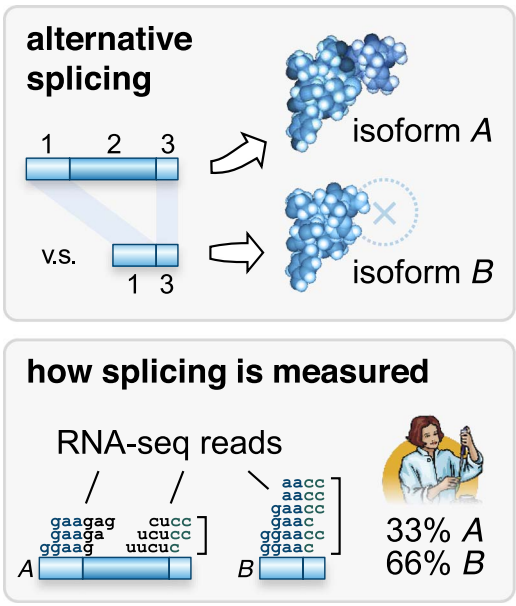


Fig. 5. *Gene can generate different proteins by the process of alternative splicing, where genomic instructions cause an exon to sometimes be included or excluded, depending on cellular conditions, such as the cell type. Using RNA-seq, the frequency with which each of tens of thousands of exons is included in a specific cell type can be measured, and these data can be used to train a computational model that discovers the instructions that control splicing and combines them to predict splicing.*

AS is observed more frequently in more complex cell types, and among all cell types found in different vertebrate species, the human brain exhibits the most complex alternative splicing patterns [61]. It is not surprising, then, that aberrant splicing is indicated as a major contributor in several psychiatric disorders. For example, transcriptomic analyses have found consistent deviations in AS patterns in the cortical regions of autism spectrum disorder (ASD) cases [62], [63], pointing to splicing misregulation as an ASD mechanism.

López-Bigas *et al.* estimate that up to 60% of genetic diseases caused by mutations are related to defects in the splicing process [64]. In particular, nearly one third of all disease-causing mutations alter a splice site [65], mostly resulting in abnormal exon skipping [66], [67]. Moreover, nearly 45% of disease/trait-associated variants reside in introns, and most of them are believed to modulate splicing patterns [22]. Aberrant splicing causes abnormalities in two major ways: 1) it results in inactive or less effective protein isoforms; or 2) it disrupts the balance of protein isoforms. Major splicing-related diseases include neurological and psychiatric disorders, cystic fibrosis, Parkinsonism, spinal muscular atrophy, myotonic dystrophy, amyotrophic lateral sclerosis, premature aging, and dozens of cancers [56], [66], [68].

Over the past decade, our group has worked to develop accurate computational models of splicing, discover new insights into the biological mechanisms of splicing, and determine the genetic causes of diverse diseases and neurological disorders. By accurately modeling splicing and AS computationally, we have been able to predict how it is affected by variations in the genome, and then to assess whether a mutation in the genome affects disease risk. The computational model of splicing uses input features that are extracted from the genome near regions where splicing occurs and then predicts the frequency with which the corresponding exons are kept or excluded from the mRNA [44], [69]–[71]. To train the model, measurements from RNA sequencing (RNA-seq) technologies are used to determine which isoforms are present in cells of different tissue types. A brief description of RNA-seq is provided here, details of which can be found in [72]. Sequencing involves first fragmenting mRNAs present in a population of cells under investigation, and then sequencing these fragments. The sequenced fragments, or "reads," are then mapped to the reference genome of the organism, allowing for a small number of mismatches. The number of reads is then used to determine the relative abundance of isoforms in the sample [73]. For splicing, reads mapped to the junction between the boundaries of exons are analyzed [44], [74]. This gives a measure of how often an exon is included or excluded in the mRNA (Fig. 5, bottom).

Recently, we showed how this approach can be used to discover new ways that mutations can affect splicing and lead to disease [44]. To analyze a specific genetic variant, the normal DNA sequence and the mutated DNA sequence are fed into the computational model and the predicted change in the splicing level is used as an indication of whether the mutation may be deleterious. Our approach led to specific predictions of the effects of clinical variants and synthetic variants in a variety of disorders, including spinal muscular atrophy, hereditary nonpolyposis colorectal cancer, and autism spectrum disorder. Validation experiments showed that the predictions matched experimental data quite well [44]. By analyzing all rare genomic variants in ASD patients, Xiong *et al.* identified 19 aberrantly spliced genes involved in central nervous system development, synaptic transmission, and neuron projection [44], significantly expanding the set of putative ASD-related genes.

This example illustrates the general workflow outlined in Fig. 3, where experimental data are first gathered in the lab to quantify cell variables of interest, and then passed to data scientists, who create models of data that are used for analysis of genetic diseases.

C. A Computational Model of Protein-DNA and Protein-RNA Binding

A "protein-sequence interaction" is a chemical attraction between a protein molecule and a DNA or RNA strand, either in a living cell (*in vivo*) or in a controlled medium (*in vitro*). The strength of protein-sequence interactions is a low-level category of cell variable for which we have much training data. These interactions are important to model accurately, because they influence so many key processes in the cell. DNA- and RNA-binding proteins bind selectively to DNA or RNA strands, respectively, and are sensitive to mutations at the binding site. A wide spectrum of ailments are known to be caused by mutations that alter binding sites, including cholesterol overproduction [75], melanoma [76], and prostate cancer [77] [cf., Fig. 5(h)]. Accurate models of protein-sequence binding are essential for interpreting the genome and for predicting the effects of mutations—this is where machine learning can play a key role.

The human genome encodes for at least ~1400 DNA-binding proteins [78] and at least ~1500 RNA-binding proteins (RBPs) [79], making these the largest categories of proteins. Many DNA-binding proteins are called transcription factors (TFs) because, when they bind, they influence the rate at which specific genes are copied (transcribed) to RNA. RNA-binding proteins can influence the subsequent processing of RNA, for example, by folding the RNA [80] or by dictating how certain exons within mRNA are spliced [81]. In fact, computational models of RBPs provide essential input features to the computational model of splicing we presented earlier. Fig. 6 (top) depicts just two important ways that these protein-DNA and protein-RNA interactions can influence cell state.

Biologists have developed high-throughput experiments that measure the sequence specificity of individual proteins. These measurements can be used to train

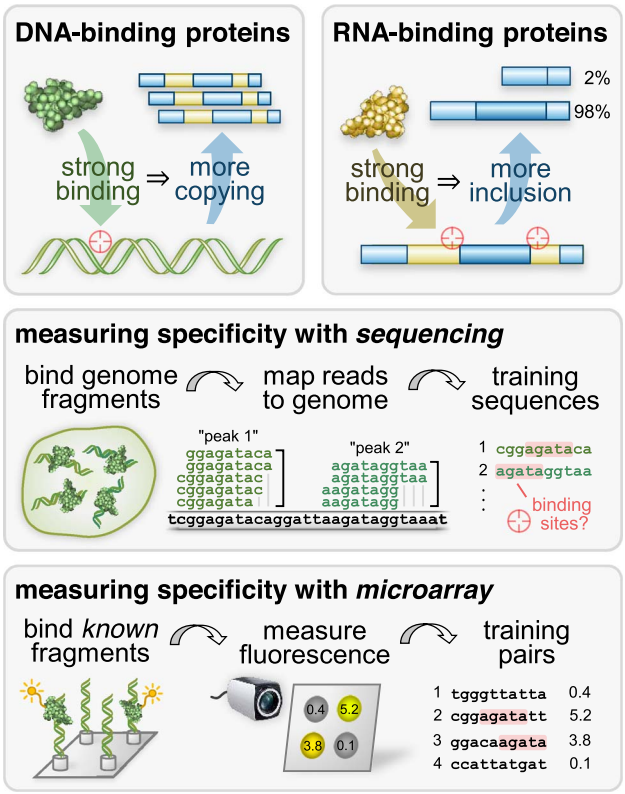


Fig. 6. ***DNA- and RNA-binding proteins regulate many processes in the cell, including gene transcription rates (top left) and splicing rates (top right). To determine the motifs(s) that a particular protein binds to (e.g.,*** `agata`***), biologists collect "specificity" data either by sequencing protein-bound fragments from a living cell (middle), or by exposing tagged proteins to synthetic fragments on a microarray (bottom).***

computational models of protein-sequence interaction: the input is a genomic sequence, and the output is a binding score [82]. However, before endeavoring to train a predictive model, it is helpful to understand what these measurements represent, how they were acquired, and what biases may lurk within the data. Two popular approaches to high-throughput experiments are sequencing-based methods and microarray-based methods. Large data sets of both kinds are publicly available. We first review sequencing and microarray procedures, and then describe popular computational models. For an overview of other protein-DNA measurement technologies available, see the review by Levo and Segal [83].

1) Sequencing: Sequencing methods work by isolating short DNA or RNA fragments (length ∼100) that are bound to the protein of interest, then sequencing those fragments, and finally mapping them to a reference genome. The idea is that, wherever mapped reads overlap and pile up to form a "peak," we can infer that the protein prefers binding in the vicinity of that peak (Fig. 6, middle). A common experimental protocol for DNA-binding proteins is ChIP-seq [40], and for RNA-binding proteins there are for example RIP- and CLIP-seq [84]. Each sequencing experiment results in a list of 100–100 000 short sequences (one for each distinct peak) that have all been excerpted from the reference genome. Machine learning researchers would consider these peaks "weakly labeled" because the appearance and locations of the binding sites are not known, much like the ImageNet classification challenge [85]. From this list of sequences, we can learn to discriminate between sequences that the protein binds to versus sequences that the protein presumably ignores. Computational biologists have developed several tools for finding recurring patterns (motifs) among a list of peaks. A popular software suite is MEME-ChIP [86], which uses expectation–maximization to find one or more motifs that discriminate peaks (positives) from a statistical model of background sequences (negatives); the MEME suite has been in development since the 1990s.

One difficulty with sequencing is that the resulting peaks often contain motifs of several proteins that bind in the same vicinity as the protein targeted by the experiment. For computer vision researchers, this would be akin to training an unsupervised model on images of the character "*q*" that were coarsely cropped from real text, only to find that the final model also responds strongly to the character "*u*." What we have, then, is not a model for recognizing a particular handwritten character, but rather a model for recognizing the context in which a particular character tends to appear. For biologists, it is important that the right patterns be associated with the right protein, and so the interpretability of the computational model can be a valid concern here. Currently, biologists examine the output of tools like MEME-ChIP by hand, visualizing each motif that was discovered and then reasoning about which motif likely belongs to the target protein based on prior knowledge of the existing literature.

2) Microarrays: High-throughput protein binding microarray (PBM) experiments expose a protein of interest to a library of ∼40 000–250 000 distinct "probe" sequences [87]. It is not known at the outset to which probes, or to where within those probes, the protein will bind. The experiment relies on a specially manufactured slide containing a matrix of tiny wells on its surface—one well for each type of probe in the library. Copies of the protein are then introduced over the slide so that they may bind selectively to the probes, after which the unbound proteins are washed away. The remaining proteins are made to emit visible light by laser excitation of an attached fluorophore. The slide is photographed and the intensity of each well is measured, revealing the concentration of protein in each well (Fig. 6, bottom). The final result of a PBM experiment is a list of (sequence, score) pairs that can be used to train a computational model of the protein's sequence specificity.

The probe sequences are explicitly designed to contain good coverage of all possible subsequences of some length *k*.

The k-mer length is usually approximately 10 for high-throughput array designs. The set of probes are generated starting with a de Brujin sequence of order k, and then partitioned with overlap to generate the desired probe sequences, each of length ∼40 [35]. Therefore, unlike sequencing data, microarray data do not have the problem of unrelated motifs appearing frequently together in the positive set. The raw intensity measurements from each microarray experiment can still have noise, artifacts, and biases. That is why intensities are often postprocessed, including outlier removal, spatial detrending, and quantile normalization (see [35], supplementary information).

3) Basic Computational Models: An early way to model binding sites was by "consensus sequence," where the protein is assumed to bind preferably to one designated pattern, sometimes with one or two mismatched positions allowed. Today, the workhorse of binding site modeling is the position-frequency matrix (PFM), along with several variations of the idea [88]. Parameter (i, j) of a protein's PFM represents the frequency at which base i appears at position j when the PFM is aligned to a binding site for that protein [89]. The frequencies in each column are often normalized to become probabilities. For example, Fig. 7 (left) shows a published PFM for modeling the human RBFOX1 protein [35]. RBFOX1 is an RBP that influences splicing [Fig. 6, top] during neuronal development, and is associated with autism [90]. The PFM parameters shown are from [35], and were inferred by identifying recurring motifs in the highest intensity probes of a microarray-based experiment. The consensus sequence for RBFOX1 is widely recognized to be `gcaug`, yet the PFM shown suggests that it binds even more preferably to the longer motif `ugcaug`—a secondary preference sometimes written as `(u)gcaug` in consensus sequence notation.

Once a PFM has been ascertained for a protein, it can act as a soft pattern for template-matching along the genome. Given a sequence $s = (s_1, \ldots, s_n)$ of length n where each $s_i \in \{1, 2, 3, 4\}$ (representing `a`, `c`, `g`, `t`, resp.), the binding score z_i for a normalized PFM with $4 \times m$ parameters M is

$$z_i = \prod_{j=1}^{m} M_{s_{(i+j-1)}, j}.$$

After scoring each individual position i, the entire sequence can be scored by taking the maximum, the average, or the total sum of scores across positions. For researchers familiar with deep neural networks, PFMs play a similar role as the filters do in a convolutional neural network [91], where here the input is a 1-D "image" with four input channels representing the occurrence of {`a`, `c`, `g`, `t`}, as depicted in Fig. 7 (right).

On the spectrum of binding complexity, our RBFOX1 example from Fig. 7 is known as one of the proteins most easily modeled by PFMs. The sequence specificity of many proteins is more accurately modeled by a combination of two PFMs, where the secondary PFM captures an extra "mode" of binding [92]. Using default settings, the MEME-ChIP software suite [86] actually outputs a list of up to three distinct PFMs, each found to be statistically significant with regard to MEME's modeling assumptions. The MatrixREDUCE software suite [93] outputs up to 20 PFM-like matrices called position-specific affinity matrices (PSAMs). One difficulty with PFMs and PSAMs is that the length of the pattern (number of columns in the matrix) must be guessed beforehand, and this guess can influence the quality of results; common practice is to overestimate the length and then trim high-entropy columns ("don't care" positions) from the matrix in a postprocessing step.

4) Beyond PFM-Based Models: The most obvious drawback to PFMs is that they make the strong assumption that each position contributes independently to binding strength. For some proteins this assumption seems to hold, but for many it does not. Computational biologists are actively developing models that capture the sequence specificity of more challenging proteins. Examples include models that capture global features of the binding site [94], models specialized for specific protein families such as C2H2 zinc fingers [95], [96], and models of "structure-selective" RBPs that are sensitive to the way RNA folds in on itself [97]. Proteins can also bind at adjacent locations on the strand, leading to protein–protein interactions that complicate binding specificity [83]. Advanced models often rely on biological insights such as the protein's crystal structure or knowledge of the "binding domains" on the protein's surface. Stormo and Zhao [98] provide a more in-depth review of DNA-binding protein experiments and PFM-based computational models. See [84] for a recent review of RBP experiments and models. Slattery *et al.* [99] provide an excellent review of more advanced issues in TF-DNA binding.

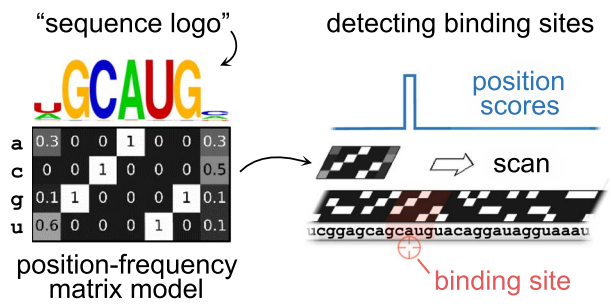


Fig. 7. ***Position-frequency matrix (PFM) model of the RNA-binding protein RBFOX1 as reported by Ray et al. [35]. A PFM can be used to score each position along a target sequence. Biologists commonly visualize PFMs as a sequence logo [89].***

For machine learning researchers interested in the sequence specificity problem, a good starting point is the DREAM5 TF-DNA Motif Recognition Challenge [100]. In that work, Weirauch *et al.* provided PBM and ChIP-seq data and evaluated the performance of 26 specialized algorithms from the computational biology community, including MEME-ChIP and MatrixREDUCE. Alipanahi *et al.* [41] recently found that convolutional neural networks can be adapted to this challenge, and that they outperform the algorithms participating in the original challenge on both PBM and ChIP-seq data. The DREAM5 challenge is part of a large series of computational biology challenges; see [101] and [102] for an overview.

IV. MACHINE LEARNING IN COMPUTATIONAL BIOLOGY

In recent years, machine learning researchers have focused their most high-profile efforts on speech recognition [103] and computer vision [104]. Computer vision in particular has a long history in machine learning due to its intuitive, accessible nature. Human beings are exceedingly good at computer vision tasks, and so when our learning algorithms do not satisfy our own expectations we often find new insights. In fact, the handwritten digit recognition data set known as MNIST [105] has been called "the drosophila of machine learning"—a reference to the fruit fly model organism in biology—owing to MNIST's widespread use as a test bed for new learning algorithms.

One of our goals in this paper is to explain computational biology problems in a way that is accessible to machine learning researchers. The situation in biology is fundamentally different from the situation in computer vision. The visual world is directly accessible to us, and researchers exploit our knowledge of how images are generated through light, occlusion, and projection. The nanoscale world of a cell's machinery is not directly accessible and, despite decades of painstaking effort, our knowledge of the mechanisms at play is woefully incomplete [106], [107]; this is true even for single-cell organisms like yeast [108], [109]. The genotype-to-phenotype relationship [Fig. 2(a)] is arguably orders of magnitude more complex than the pixels-to-labels relationship in high-profile vision challenges such as ImageNet [85]. The details of many interactions, quantities, and processes in the cell are "hidden" from us because we do not have the technology to systematically measure them. In other words, the few cell variables that we can observe [Fig. 2(b)] are the outcome of many layers of interacting cell variables that we cannot. That is why we believe advanced machine learning, and deep learning in particular, will play an important role as biology moves toward high-throughput experiments.

In this section, we first describe some approaches to map genetic variants with disease risks by association and through the use of comparative genomics. These methods directly model the genotype-to-phenotype relationship and are in common use. We then give a brief outline of additional cell variables that are of interest to the research community, some of which have predictive models that have leveraged the use of deep learning to improve performance. These cell variables may be used alone or in combination for disease risk modeling. We also provide additional publicly available sources of data for machine learning researchers to contribute to the field.

A. Current Approaches to Model the Genetic Basis of Disease Risks

1) Genome-Wide Association Studies: The goal of genome-wide association studies (GWASs) is to detect how traits within a population can be related to variants in particular genomic locations, or loci. Early GWAS experiments used microarrays that were designed by the most easily determined variants in the human population: single-nucleotide polymorphisms (SNPs), which are variations that are relatively frequent across humans (frequency greater than 1%). Modern GWAS analysis use more comprehensive sets of variants and even whole genome sequencing data, which is not restricted to a subset of variants. Here, we focus on some of the challenges of GWAS; for an excellent review of the current state of GWAS, see [110].

From a data analysis point of view, one of the main difficulties with GWAS is establishing statistical significance between a potentially causal variant with a change in risk for particular diseases between the affected group of individuals compared to a control. The main problem in GWAS and any association-based technique [e.g., expression quantitative trait loci (eQTLs)] is that they indicate correlation, not causation. Due to confounding hidden variables, such as correlations between nearby variants caused by cross-over (linkage disequilibrium) or differences in subpopulations caused by factors such as migration, two or more genomic loci might be correlated, and an SNP could be picked up by GWAS, simply because some other genomic locus is causal [111]. The causal variant is often not even observed in the GWAS study. GWAS furthermore provides a huge number of putative causal mutations, and researchers may be biased toward candidates that have greater "narrative potential" [112].

Some of the bigger GWAS studies involve tens of millions of SNPs that are conducted on thousands of individuals. Assessing the statistical significance of an immense number of SNPs is challenging and requires careful multiple-hypothesis correction or false discovery rate analysis [113]. The problem is compounded by the fact that many common variants have weak effects, and those that have strong effects tend to be rare [114]. To improve significance, some studies limit the profiling of SNPs to the coding region of the genome [115], with the assumption that mutations in these regions are more likely to impact

risks as they can alter the function of the proteins [116]. Another way to address this problem is by increasing the sample size. Significant resources have been channeled into large population study initiatives such as TCGA and HapMap, which have raised debates within the research community on the cost-benefit ratio of these projects [25], [117], [118]. Another major hurdle is population structure and its stratification. A recent paper [119] that used a genetic classifier based on SNPs for ASD detection raised some controversy [120] and it was alleged that most of the observed signal was due to "potential population stratification" (genetic differences due to ancestry).

One approach to make better use of GWAS data beyond statistical associations is to use computational models that take as input the SNP profiles of individuals to predict disease risks. These SNP profiles tend to be high dimensional, and typically have a large proportions of SNPs that are not relevant to the disease at hand (and therefore noisy). Several tools are available for prioritizing causal variants (e.g., PolyPhen [121], SIFT [122], SPANR [44]), and machine learning algorithms have been used for learning predictive models of disease risks [123], [124].

2) Evolutionary Conservation: Comparative genomics is a powerful way to identify genomic sequences that have function. The most well-known resource that comparative genomics provides is sequence conservation. The rationale behind sequence conservation is as follows. First, consider evolution as being driven by two forces: the slow accumulation of random mutations, and selective pressures against mutations that damage reproductive fitness within a population [125]. Now consider the genomes of several species that diverged from a common ancestor long ago; long enough that random mutations have had plenty of time to occur. When we compare the genomes, we find many long distinct sequences that are nearly identical, or "conserved," across species. When a sequence is conserved, it is strong evidence that evolution is exerting selective pressure on the positions within those sequences. Studies estimate at least 5%–6% of the human genome is conserved with mammals [21], [126].

Detection of conserved sequences has been instrumental in annotating functional elements in the human genome [21], such as exons. Conservation scores are available for multiple organisms from the software tool phastCons [127]. For each position in a genome, phastCons provides a number between 0 and 1, where 0 indicates no discernible conservation, and 1 indicates 100% conservation across all species considered. Other methods for quantifying conservation include GERP [128] and phyloP [129]. Conservation scores for each position in the human genome can be viewed online as a "track" within the UCSC Genome Browser [130].

A mutation that lowers reproductive fitness is called deleterious, whereas a mutation that causes a disease is called pathogenic [112]. Many mutations are, of course, both deleterious and pathogenic, such as mutations causing Tay Sachs disease, but it is important to understand that conservation only provides information about deleteriousness. Even so, conservation-based techniques have been an extremely useful input feature for predictive models of disease. One recent example is the combined annotation dependent depletion (CADD) method [131]. Kircher *et al.* first developed a "mutation simulator" that generates realistic synthetic mutations without regard to selective pressure. They then trained an ensemble of ten linear support vector machines to discriminate between synthetic mutations (assumed deleterious) and the ~16 million actual human mutations that have survived selective pressure (nondeleterious) since the human–chimpanzee common ancestor.

B. Current Directions

We believe the cell variable approach will be important for making reliable *in silico* predictions in genomic medicine and provide insight to disease mechanisms. We have described two examples of cell variables (splicing and protein-sequence binding) that may be useful to ascertain disease risk from genetic variants. For splicing, we have shown that our model is quite different from existing techniques, so it can be used to complement them and significantly improve their sensitivity [44]. In addition, the cell variable approach can provide putative explanations as to how variants affect disease risks. Referring to the example in Fig. 1, four differences in nucleotides between the *SMN1* and *SMN2* gene cause spinal muscular atrophy. Our splicing model is able to suggest that the synonymous mutation in the exon reduces the binding affinity of SF2/ASF (a splicing regulator protein) and thereby causes the exon to be skipped [44], [132]. Approaches like GWAS do not offer such information.

For both these problems, there have been recent works that utilized deep learning to improve predictive performance, such as feedforward neural networks for alternative splicing patterns by Leung *et al.* [71] and convolutional neural networks for binding specificity by Alipanahi *et al.* [41]. In another example, Quang *et al.* used the previously described CADD data set to train a deep neural network to lower error rate [133].

Current research includes improved modeling of other cell variables related to regulation at the DNA level. Some of these are listed in Table 1. Many of these variables are coregulated. For example, transcription is tightly coupled with splicing [55], and therefore having a good model for one cell variable can often improve the prediction of another. For the studies cited, the authors do not necessarily present their models for use in disease risk analysis. However, the goal here is to provide to the reader a larger exposure to measurable quantities in the cell that researchers have developed predictive models for, many of which can potentially act as intermediate cell variables for disease risk prediction [134]. Even though deep learning is

Table 1 Sample of Cell Variables Related to Genomic Regulatory Mechanisms

Cell Variable	Brief Description	Relevance to Disease	Reviews	Related Works
Identification of structural and functional regions of the genome	Attaching meaning to, or annotating, different regions of the DNA, such as marking the boundary of introns and exons, and identifying parts that have regulatory functions.	Changes in genomic sequences can cause a region which previously served a particular function to become non-functional and vice-versa, or changing its intended function, thereby affecting regulation.	[208]–[210]	[211], [212]
Binding sites for transcription regulation	Binding of proteins to specific sequence elements of the DNA controls whether transcription can occur, as well as the rate at which it happens.	Sequence variations to sequence patterns that proteins, such as transcription factors and complexes that 'unwind' the DNA, bind to can alter whether a gene is transcribed.	[45], [213]	[41], [97]
Splicing patterns	Splicing modifies the pre-mRNA by removing introns and selecting which exons are retained.	Changes to the regulatory elements that control splicing can change the characteristics of the gene products, and in some cases, cause them to be non-functional.	[53]	[44], [69]–[71]
Cleavage site selection and polyadenylation	The ends of transcripts are cleaved and a stretch of adenine bases are attached before they are ready for translation. Cleavage can occur in one of multiple sites within a transcript.	Modifications to sequence elements can alter where cleavage occurs, which determines whether binding sites for regulatory proteins are present or absent on the transcript. This alters its stability and translation efficiency.	[48], [214]	[215], [216]
RNA structure	The RNA folds into three-dimensional (3D) structures, which influence how it interacts with other molecules in the cell.	The mRNA, beyond the information it contains for encoding protein, has 3D structure. This structure can affect processes that it is involved with, such as transcription, splicing, and translation.	[51], [217]	[43]
Protein structure	The outcome of translation is a sequence of amino acids that folds into a protein. The protein's 3D structure is crucial for its function, as it interacts with DNA, RNA, and other proteins.	Structure affects function. The ability to predict protein structure from sequences can help in understanding the biological function of a gene, and how mis-folding of proteins contribute to disease.	[218]	[219], [220]

not currently common practice for most of these models, it is utilized by some and in certain instances there has been significant improvement.

C. Large Data Initiatives for Modeling the Genotype to Phenotype Relationship

Beyond data sets of single cell variables, large international efforts are now in place to coordinate and acquire measurements from different levels of the biological system from genotype to particular phenotypes, such as cancer, for a large number of individuals. Some of these initiatives are listed in Table 2. A good variety of "omic" data are publicly available from these initiatives, including genomic, transcriptomic, epigenomic, and proteomic information [106]. Another source of information is the dbGaP database maintained by the National Center of Biotechnology Information, which archives the results of studies that have investigated the interaction between genotype and phenotype [135]. Using multiple sources of data can compensate for missing information from any single data source, and can help bridge the gap between the many layers of interactions between genotype and phenotype toward a more complete biological model of regulation.

V. DISCUSSION AND FUTURE DIRECTIONS

Based on our experience, we expect the role of machine learning in genome biology, genome medicine, and precision medicine to grow rapidly in the coming years. Particularly relevant is the dramatic progress that has been made in the deep learning community, including the discovery of techniques that can more effectively learn from very large and much more complex data sets than ever before. Since genotype and phenotype are related through many layers of biophysical processes and interactions, most of which are not fully understood, we anticipate that only through the development and application of even more powerful computational techniques can we hope to model the outcomes of these complex processes and interactions.

Machine learning and most recently deep learning have achieved human-level performance in domains such as image recognition, speech recognition, and natural language processing. However, from a machine learning perspective, genome biology differs from these domains in a very important way. Humans are very good at the former tasks, which involve human perception (e.g., seeing images, hearing speech) and human action (e.g., grabbing an object, responding to words). In stark contrast, we can think of no reason why the genome should be interpretable by humans. Whereas there has been evolutionary pressure for humans to perceive, interpret, and respond to patterns of light, such as that produced by an advancing tiger, there has been no pressure for the genome to be interpretable by humans or for humans to develop the capacity to interpret the genome. Consequently, it is important to incorporate the latest biological knowledge and data into learning

Table 2 Overview of Large Initiatives and the Available Sources of "omic" Data Set

Source	Types of data	References
Genotype Tissue Expression (GTEx)	**Genotype:** SNP arrays, exome and whole-genome sequencing **Transcriptome:** RNA-seq **Phenotype:** comprehensive profiles of subjects	[221]
National Cancer Institute Anticancer Drug Screen (NCI-60)	**Genotype:** exome sequencing **Transcriptome:** m/miRNA microarray **Proteome:** SWATH profiles **Phenotype:** cancer cell lines (drug tested)	[222]
Encyclopedia of DNA Elements (ENCODE)	**Genotype:** whole genomes for subset of cell lines **Transcriptome:** RNA-seq **Epigenome:** ChIP-seq, DNASE, 5C	[166]
International Cancer Genome Consortium (ICGC)	**Genotype:** cancer whole-genomes **Phenotype:** pathology reports	[223]
The Cancer Genome Atlas (TCGA) and The Cancer Proteome Atlas (TCPA)	**Genotype:** cancer whole-genome and exome sequencing **Transcriptome:** RNA-seq (m/miRNA) **Epigenome:** methylation **Proteome:** expression levels for signaling pathways (reverse-phase protein arrays); some samples matched with TCGA **Phenotype:** pathology reports (baseline and drug-tested)	[224], [225]
The 1000 Genomes Project	**Genotype:** whole-genome sequencing, high-quality variant calls **Transcriptome:** RNA-seq for a large fraction of cell lines (through project Geuvadis) **Phenotype:** different populations, trios (parents and offsprings)	[226]
NIH Roadmap Epigenomics Project	**Genotype:** whole-genome (a subset of cell lines) **Transcriptome:** RNA-seq & smRNA-seq **Epigenome:** comprehensive ChIP-seq **Phenotype:** tens of cell lines and ex vivo differentiated cells	[227], [228]
Genetic Investigation of Anthropometric Traits (GIANT)	**Genotype:** SNP arrays **Phenotype:** body size and measures of obesity	[229]–[231]

algorithms and to carefully and rapidly validate models in different ways, since the models cannot be "checked by eye."

It is possible that the association of the genome to some diseases might simply be too complex to be modeled from a practical number of "inputs." This is in contrast to image or speech recognition, where we know what the prediction ought to be given the input. Furthermore, it should be noted that due to the inherent stochasticity of cellular processes, environmental factors that differ from person to person (even for identical twins), and uninherited variants from the parent that can affect offsprings, the genotype of an individual may not be sufficient to completely determine their phenotype [136]. Therefore, we do not expect computational methods to be able to entirely replace laboratory and clinical diagnosis, but they should greatly shorten the time required for these methods of analysis by reducing the search space of hypotheses that need to be validated.

With respect to machine learning, there are some similarities between genome biology and other domains. The genotype–phenotype relationship can be thought of as a landscape. This landscape has extremely steep valleys, where small perturbations in genotype give rise to vastly different phenotypes [137]. It also has large plateaus, where seemingly unrelated genotypes yield an equivalent phenotype [138], [139]. To some extent, the same observation applies to other application domains, such as speech recognition, where a perturbed vocalization can dramatically alter semantics, or computer vision, where image interpretation must be invariant to a huge space of possible image transformations. Deep learning has facilitated great progress in speech and vision. Perhaps most exciting are the recent successes in "end-to-end" learning, where systems with many layers can learn from extremely low-level (unprocessed, uninterpreted) data [140], [141]. If comparable progress can be made on the computational challenges facing biology and genomics, the potential impact on genomic medicine is very real.

Cell variables are more difficult to measure than phenotypic observations such as whether a patient is sick. However, consider measuring few variables per patient for a large number of individuals, versus taking hundreds of thousands of cell variable measurements per patient for a smaller group of people. We believe that the latter approach gives us a better chance at deciphering the genomic instructions of the cell, where there is much more information available about the biological mechanisms at play, and therefore more data overall for a model to learn from. In a sense, we are making the "genomic invariance" assumption; that is, we assume that regulatory processes

act the same across the entire genome and so we can learn the DNA-to-cell variable relationship by treating different locations in the genome as independent measurements.

The immense growth of genomic data poses storage, privacy, and computing challenges that make it difficult for small research groups to participate in at scale. With regard to computation, there exist parallel distributed training algorithms that handle extremely large data sets [142], [143]. These algorithms are effective on compute clusters with ~1000 nodes [~16 000 central processing unit (CPU) cores]. However, it was recently shown that large-scale machine learning can be done fast and much more economically on "consumer off-the-shelf" clusters that are accelerated by graphics processing units (GPUs) [141], [144].

Cloud-based computing can also facilitate reproducible research through the use of virtual machine images. Dudley and Butte [145] call this the whole system snapshot exchange (WSSE) approach, where data sets, code libraries, processing pipelines, and experimental results are all packaged and made available for inspection and for follow-up research. Similarly, Stein *et al.* make a point that reducing sequencing costs are outpacing information technology and storage support in laboratories around the world, challenging the traditional use case where practitioners are accustomed to downloading genomic data to local computer for analysis [146], [147]. Configuring a large-scale machine learning experiment requires significant software engineering expertise, and the resulting pipelines often have complex, brittle, soon-outdated dependencies that are extremely burdensome to reproduce [148]. To participate in the era of "Big Data" genomics, machine learning research projects should be designed with scalability, portability, and reproducibility in mind, even at the earliest stages.

In addition to genomic and disease risk data, we believe the best way forward is to also leverage cell-level data and explicitly model how they are influenced by the genome. Given a new patient in need of diagnosis and treatment, the "computational models of cell variables" approach [Fig. 2(b)] has economic advantages. The idea is to measure variables for which the measurement technology is inexpensive/fast/noninvasive (e.g., genome sequencing) and to then predict variables that would be expensive/slow/invasive to measure on per-patient basis (e.g., splicing in the brain of an autism patient). Still, biologists are developing new high-throughput technologies for measuring cell variables, such as DNA methylation [149], novel mRNA isoforms via long-read technology [150]–[152], and mRNA levels in a single cell [153]–[155]. Integrating the many emerging sources of data into analysis will be a challenge for some time to come, and no gold standard has emerged [106]. For example, much of the raw data generated by the experiments in Table 2 require a background in the relevant biology to preprocess them before they can be used to train models. For machine learning researchers, this hurdle can be high, although perhaps not very high in comparison to the potential value of transforming medicine and saving lives. In the remaining text, we discuss how to engage machine learning researchers, as well as future research directions.

A. Engaging the Machine Learning Community

Today, researchers in machine learning tend to develop their algorithmic ideas around applications in speech recognition, natural language processing, and computer vision. In doing so, they have transformed these fields and supplanted many specialized systems developed within those communities. In our experience, machine learning researchers are eager to prove the efficacy of their algorithms on important applications. Genomics and computational biology could become areas that are actively targeted, just as computer vision is, if only the problems and data were prepared and communicated in a more accessible manner.

A successful example of machine learning community engagement is the Merck Molecular Activity Challenge hosted on the Kaggle platform. Deep learning researchers were able to significantly improve a state-of-the-art drug discovery pipeline, despite no knowledge of what biochemical properties the training features represented [156]. Another example of community engagement is protein side-chain prediction, an important step in protein folding and protein design [157], [158]. Side-chain prediction has received attention from researchers working on inference algorithms [159]–[162] because the problem was formulated in terms of graphical models—a familiar and well-studied abstraction. As for the more challenging problem of protein folding itself, the biannual critical assessment of protein structure prediction (CASP) challenge [20] provides the data and proving ground for folding-related tasks. The critical assessment of protein function annotation (CAFA) challenge is another initiative where the community is asked to predict protein function from its sequence [163]. However, both the CASP and CAFA challenge data and problem descriptions are designed for experts in that field. Many problems in computational biology, such as splicing, genome annotation, protein folding, and protein-sequence binding, would be well served by being packaged as accessible Kaggle-like competitions. These community challenges can foster knowledge sharing and attract members from different communities, such as those from machine learning, to bring new perspectives to otherwise specialized computational biology problems, and can often lead to improvements to the state of the art [101], [164].

Several organizations have been successful in machine learning engagement. DREAM, the dialog for reverse engineering assessments and methods (dreamchallenges.org), poses fundamental questions about biology and medicine, uses rigorous practices to assess the performance of different methods (e.g., holding data out for testing), and fosters collaborations. CAGI, the critical assessment of genome annotation (genomeinterpretation.org), is

a community experiment to objectively assess computational methods for predicting phenotypic impacts of genomic variation and to inform future research directions. CASP (predictioncenter.org) is another organization that aims to establish the current state of the art in protein structure prediction, identifying what progress has been made, and highlighting where future effort may be most productively focused.

B. Less Reliance on Evolutionary Conservation

As noted in Section IV-A, the current assortment of variant analysis tools (e.g., CADD) rely heavily on conservation features. One drawback to relying on conservation is that the resulting models can seem to perform well even if they trivially rely on conservation and do not discover any meaningful patterns in the genome itself. For example, the computational model of splicing in Section III-B is more accurate when given access to conservation features [44], [69], presumably because there are subtle aspects of splicing that the model did not ascertain from the raw messenger RNA sequences in the training set. Also, since the relevant biochemical processes do not get to directly examine information about evolutionary conservation, it seems inappropriate for a model that is meant to mimic the biochemical processes to have access to it.

Another drawback to relying on conservation is that it only represents sequence conservation, not functional conservation. Not all conserved sequences are functional, and not all functional sequences are conserved [165], [166]. We would expect a lack of functional conservation to be more associated with disease but, at this time, it is not known how to measure or to estimate the function of many gene products. Also, even when conservation can tell us that a mutation likely hurts reproductive fitness, it cannot tell us why the mutation is damaging nor what therapies may help the patient. Furthermore, conservation can only tell us what has survived millions of years of natural selection. It is indifferent to variations in DNA that cause pain or that—like Alzheimer's, Parkinson's, heart disease, and most cancers—tend to manifest long after reproductive age. In some instances, disease-associated mutations can also appear as wild-type in another species [167]. Finally, conservation is blind to variants that alter nucleotides in recently evolved parts of the genome. For example, methods that aim to classify disease-causing genetic variants from databases such as Human Gene Mutation Database (HGMD) are known to be heavily biased toward the variants in the conserved genomic regions [168]. In short, conservation can be a source of information that is predictive, but it will never be able to tell the whole story.

C. Recurrent Neural Networks (RNNs)

Other research communities that deal with sequence data have seen dramatic gains in accuracy by moving from Markov models toward deep recurrent models. In text-based natural language processing, the classic *n*-gram model is fast being replaced with RNNs [169] and related long short-term memory models (LSTMs) [170]. In speech recognition, hidden Markov models (HMMs) have been supplanted by LSTMs due to the power of learning better representations of the data [103]. We believe that similar gains may be expected for sequential prediction applications in computational biology. For example, genome annotation is a problem for which HMM-based models are still prevalent [171]–[173], treating genomic position as the "time" axis and using position-specific genomic measurements as the observed variables. The official genome annotations of the recent Encyclopedia of DNA Elements (ENCODE) project were determined by an HMM-based model [174]. HMM have also been used to predict whether single nucleotide variants are potentially pathogenic [175]. Another example is the modeling of cell variable dynamics through time. The work of Karr *et al.* [176], [177] is widely regarded as the first whole-cell computational model with any reasonable degree of predictive accuracy. They simulate the human parasitic bacterium *Mycoplasma genitalium* by training a model to update 16 cell variables at one minute intervals—a time-series task for which RNNs may be highly suitable. As another example where RNNs may be powerful, we observe that DNA-binding proteins (Section III-C) are known to arrive at binding sites through a dynamic process wherein the protein migrates along the DNA backbone [178]. During migration, the protein may be influenced by intervening sequence patterns or (in a real cell) by the state of chromatin. This dynamical view of TF-DNA interaction seems to call for a sequential state model of binding based on RNNs or LSTMs. Another potential application of RNNs is for the imputation of epigenomic tracks. Regression trees were utilized for this problem in the work of Ernst and Kellis [179], but since the problem can be viewed as a "sequence-to-sequence mapping" problem, it may benefit from an RNN architecture [180].

D. Interpretability

Interpretability is not a well-defined concept. Despite calls in the 1990s to define a clear measure of interpretability for machine learning models [181], there is still no universally agreed-upon definition. The validity of an interpretation is dependent on the framework used to formalize and communicate concepts. Also, a "simple" interpretation may in fact appear complicated when presented in a different framework.

Within some application domains, interpretability is deemed to be quite important [182]. Following the movement to rely more on data-driven explanations rather than incorrect conceptual "explanations," we advocate the development of systems that can be queried by human experts so as to make predictions that can be experimentally tested. Instead of examining the parameters of a neural network and coming up with an "interpretation," a more useful exercise would be to ask the system about

relationships between inputs and outputs, for instance, whether a cell variable will increase or decrease if a particular nucleotide is changed, or, whether changing a pair of nucleotides leads to a change in the cell variable that cannot be accounted for by independent, additive contributions. This question-and-answer interaction between the expert and the machine learning model provides a quantitative, data-driven interpretation.

Traditional methods can be used to identify important input features. For example, given a machine learning model that has been trained on well-understood data, a domain expert can inspect the model and recognize familiar features, patterns, or hidden variables that are known to be relevant to the prediction task. The input features can be ranked by importance, such as done with linear models, decision trees, and random forests.

The community's ability to derive interpretations from machine learning models is likely to improve as these models become more effective in practice. Some of the world's best researchers are setting their sights on the problem of interpretability, and progress is being made rapidly. However, consider for a moment what it would mean to wait for interpretability challenges to be "solved"; to forego the benefit of more accurate models in genomic and precision medicine. Throughout history, many advances were made by noticing a pattern without understanding the precise causal mechanisms involved. For example, in 1847, Ignaz Semmelweis found that washing hands before delivering babies was correlated with fewer maternal deaths. He achieved a two third reduction in mortality rate a full 25 years before Louis Pasteur established the relationship between germs and disease. Viewed from a different perspective, if machine learning can be used to identify the genetic cause of a disease and an effective therapy, it is unlikely that the patient who benefits will care about interpretability.

Regardless of the progress in machine learning, machine learning researchers working with biologists should be prepared for a strong bias toward interpretability and historical models. For example, it is widely understood that PFMs (Section III-C) make unrealistic simplifying assumptions with regard to protein-sequence binding (e.g., positional independence), but they persist as a popular model of TF and RBP sequence specificity precisely because they are so simple. Furthermore, their specificity rules can be visually depicted as sequence logos (Fig. 7), which are intuitive and can be understood at a glance by domain experts. Some successors to PFMs attempt to provide logo-like visualizations, e.g., the feature motif model of Sharon *et al.* [94], but they are inevitably more complex and thus much less popular among biologists today, despite their better accuracy.

There have been many efforts to improve the interpretability of machine learning models, and in particular deep architectures like deep neural networks. Erhan *et al.* introduced the idea of tuning the input to maximize the activation of a hidden unit. This enables one to see what kind of inputs a hidden unit is sensitive to [183]. The method has been applied to deep architectures trained on millions of images, where neurons that correspond to face, cat, and human body detectors have been found [184]; they do so by designing a norm-constrained input that maximizes the activity of a neuron deep inside the network. Zeiler and Fergus [185] aim to visualize the input variations that high-level features respond to in a convolutional neural network; they do so by generating several diverse inputs that each cause high activations in a feature map deep within the network. Several compelling visualization approaches use back-propagation to efficiently visualize how deep architectures respond to input perturbations [186], [187] and to understand invariances in deeper layers [188]. Another approach in deep learning has been to simplify computations within the actual models themselves. Examples include the optimal brain damage [189] method, filter decomposition techniques [190]–[193], and model compression [194]. However, the motivation for model simplification techniques has been to provide faster predictions, not to gain insights into the characteristics of the learned function. We propose that model compression, in particular, can be repurposed to provide interpretable approximations to large black-box models such as deep neural networks.

There are examples in machine learning where interpretability has led to important biological insights. The computational pathologist (C-Path) project [195] is an example at the intersection of computer vision and cancer diagnostics. The authors of C-Path discovered that their machine learning model relied much more heavily on features of the stromal cells (connective tissue) than expected, providing human pathologists with new insights into the development of breast cancer. The continued popularity of decision trees in computational biology is in part due to the potential for insight from the rules that they learn [196]. In their HMM model of T-cell chromatin states, Ernst and Kellis [172] explicitly simplify the state space (from 79 to 51) so as to assign meaningful biological function to each state; the insight gained from these interpretable states is an important aspect of their contribution. Going forward, techniques such as deep neural networks must be able to provide similar insights "out of the box" for the sake of nonmachine learning experts.

E. Adversarial Data for Genomics

Recent work in computer vision has highlighted the fact that neural networks make wildly inaccurate predictions when presented with adversarial inputs. These inputs are called adversarial because they are explicitly designed to "fool" a model into making mistakes. The inputs are often designed with a specific model instance in mind, such as a particular neural network that has already been trained, or is in the midst of being trained.

Szegedy *et al.* [197] showed that adversarial inputs need not appear unusual or pathological. As just one example, a state-of-the-art neural network that correctly classified a particular image as car was found to misclassify that same image as ostrich after introducing an imperceptibly small perturbation of the image pixels. For genomic data, in many circumstances, a small variation (e.g., a single mutation) really should have a drastic effect [137]. Szegedy *et al.* showed that this troubling behavior was exhibited for a wide variety of images and classes, even when the images being perturbed were taken directly from the training set. Less surprisingly, adversarial inputs are also abundant when we move far from the training data [198]. In either case, adversarial inputs designed to fool one specific model were also good at fooling a different model trained on the same data, indicating that ensemble predictions are highly susceptible as well.

A key insight into why adversarial examples exist is that, given any training example, we can always perturb it in a direction that aligns well with the weights of a neural network, thereby amplifying its effect on the output [199]. It is important to note that these adversarial examples may not occur naturally and so the system may work fine for naturally occurring inputs. However, one of our objectives is to use our computational models to predict the effects of therapies, such as making small changes to the genome, for example, using genome editing technologies. The resulting genome sequences may be unnatural and so the question of testing for adversarial input arises. To address this, it is possible to synthesize adversarial genomic variants and compare predictions to real experiments, thereby validating and then improving the computational models. We believe adversarial examples will play an important role in shaping and validating the invariances learned by data-driven models in biology and genomics.

F. Genome Editing

One of the most promising directions in genomic medicine is genome editing that is performed by using RNA-guided DNA endonuclease Cas9 (CRISPR-associated endonuclease 9) enzyme from the type II bacterial adaptive immune system—called clustered regularly interspaced short palindromic repeats (CRISPR) [26], [27]. The elegance of CRISPR-Cas9 systems is that they only need the Cas9 enzyme and a single guide RNA (sgRNA) [200]. Given its sgRNA as the template, Cas9 can target specific locations in the genome and cut the genome at those locations. The CRISPR-Cas9 system can be used for the modification, insertion, or deletion of genomic instructions. The system has been used to control gene expression by deactivating the endonuclease domain [201], study gene functions in neural cells [202], develop synthetic biology applications [203], fix pathogenic mutations causing β-thalassemia [204], and to repair cystic fibrosis transmembrane conductor receptor mutations [205]. There are still challenges to practical genome editing of individual tissues in the human body, but these challenges seem surmountable in the near term [206]. We believe computational models not only can improve the efficacy of CRISPR-Cas9 system [207], but also can predict the phenotypic effects of genome edits that in turn will maximize the potential of this exciting technology. As we stated above, knowing how to write to the genome is not the same as knowing what to write. ■

Acknowledgment

The authors would like to acknowledge members of the Frey Lab, especially H. Y. Xiong, for helpful discussions and comments. The authors would also like to thank the reviewers for their contributions to improve the paper.

REFERENCES

[1] M. C. Schatz and B. Langmead, "The DNA data deluge," *IEEE Spectrum*, vol. 50, no. 7, pp. 28–33, Jul. 2013.

[2] V. Marx, "Biology: The big challenges of big data," *Nature*, vol. 498, no. 7453, pp. 255–260, 2013.

[3] E. A. Ashley, "The precision medicine initiative: A new national effort," *J. Amer. Med. Assoc.*, vol. 313, no. 21, pp. 2119–2120, 2015.

[4] Z. D. Stephens *et al.*, "Big data: Astronomical or genomical?" *PLoS Biol.*, vol. 13, no. 7, 2015, Art. ID e1002195.

[5] K. Wagstaff, "Machine learning that matters," in *Proc. 29th Int. Conf. Mach. Learn.*, 2012, pp. 529–536.

[6] J. D. Watson and F. H. C. Crick, "Molecular structure of nucleic aids: A structure for deoxyribose nucleic acid," *Nature*, vol. 171, no. 4356, pp. 737–738, 1953.

[7] E. S. Lander *et al.*, "Initial sequencing and analysis of the human genome," *Nature*, vol. 409, no. 6822, pp. 860–921, 2001.

[8] E. S. Lander, "Initial impact of the sequencing of the human genome," *Nature*, vol. 470, no. 7333, pp. 187–197, 2011.

[9] T. Strachan and A. Read, *Human Molecular Genetics*. New York, NY, USA: Garland Science, 2010.

[10] B. Alberts *et al.*, *Molecular Biology of the Cell*. New York, NY, USA: Garland Science, 2002.

[11] E. de Klerk and P. A. C. 't Hoen, "Alternative mRNA transcription, processing, and translation: Insights from RNA sequencing," *Trends Gen.*, vol. 31, no. 3, pp. 128–139, 2015.

[12] J. Harrow *et al.*, "GENCODE: The reference human genome annotation for the ENCODE project," *Genome Res.*, vol. 22, no. 9, pp. 1760–1774, 2012.

[13] L. Cartegni and A. R. Krainer, "Disruption of an SF2/ASF-dependent exonic splicing enhancer in SMN2 causes spinal muscular atrophy in the absence of SMN1," *Nature Gen.*, vol. 30, no. 4, pp. 377–384, 2002.

[14] Y. Hua *et al.*, "Peripheral SMN restoration is essential for long-term rescue of a severe spinal muscular atrophy mouse model," *Nature*, vol. 478, no. 7367, pp. 123–126, 2011.

[15] N. A. Naryshkin *et al.*, "SMN2 splicing modifiers improve motor function and longevity in mice with spinal muscular atrophy," *Science*, vol. 345, no. 6197, pp. 688–693, 2014.

[16] D. Hanahan and R. A. Weinberg, "Hallmarks of cancer: The next generation," *Cell*, vol. 144, no. 5, pp. 646–674, 2011.

[17] M. A. Rubin, "Make precision medicine work for cancer care," *Nature*, vol. 520, no. 7547, pp. 290–291, 2015.

[18] F. S. Collins and H. Varmus, "A new initiative on precision medicine," *New England J. Med.*, vol. 372, no. 9, pp. 793–795, 2015.

[19] F. H. Crick, L. Barnett, S. Brenner, and R. J. Watts-Tobin, "General nature of the genetic code for proteins," *Nature*, vol. 192, pp. 1227–1232, 1961.

[20] J. Moult, K. Fidelis, A. Kryshtafovych, T. Schwede, and A. Tramontano, "Critical assessment of methods of protein structure prediction (CASP)-round x," *Proteins*, vol. 82, no. S2, pp. 1–6, 2014.

[21] K. Lindblad-Toh *et al.*, "A high-resolution map of human evolutionary constraint using 29 mammals," *Nature*, vol. 478, no. 7370, pp. 476–482, 2011.

[22] L. A. Hindorff *et al.*, "Potential etiologic and functional implications of genome-wide association loci for human diseases and traits," *Proc. Nat. Acad. Sci. USA*, vol. 106, no. 23, pp. 9362–9367, 2009.

[23] G. A. Tuskan *et al.*, "The genome of black cottonwood, Populus trichocarpa (Torr. & Gray)," *Science*, vol. 313, no. 5793, pp. 1596–1604, 2006.

[24] M. Clamp *et al.*, "Distinguishing protein-coding and noncoding genes in the human genome," *Proc. Nat. Acad. Sci. USA*, vol. 104, no. 49, pp. 19428–19433, 2007.

[25] H. Ledford, "End of cancer-genome project prompts rethink," *Nature*, vol. 517, no. 7533, pp. 128–129, 2015.

[26] L. Cong *et al.*, "Multiplex genome engineering using CRISPR/Cas systems," *Science*, vol. 339, no. 6121, pp. 819–823, 2013.

[27] P. Mali *et al.*, "RNA-guided human genome engineering via Cas9," *Science*, vol. 339, no. 6121, pp. 823–826, 2013.

[28] J. G. Moffat, J. Rudolph, and D. Bailey, "Phenotypic screening in cancer drug discovery-past, present and future," *Nature Rev. Drug Discov.*, vol. 13, no. 8, pp. 588–602, 2014.

[29] M. Schena, D. Shalon, R. W. Davis, and P. O. Brown, "Quantitative monitoring of gene expression patterns with a complementary DNA microarray," *Science*, vol. 270, no. 5235, pp. 467–470, 1995.

[30] M. F. Berger *et al.*, "Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities," *Nature Biotechnol.*, vol. 24, no. 11, pp. 1429–1435, 2006.

[31] A. A. Philippakis, A. M. Qureshi, M. F. Berger, and M. L. Bulyk, "Design of compact, universal DNA microarrays for protein binding microarray experiments," *J. Comput. Biol.*, vol. 15, no. 7, pp. 655–665, 2008.

[32] M. J. Buck and J. D. Lieb, "ChIP-chip: Considerations for the design, analysis, and application of genome-wide chromatin immunoprecipitation experiments," *Genomics*, vol. 83, no. 3, pp. 349–360, 2004.

[33] J. W. K. Ho *et al.*, "ChIP-chip versus ChIP-seq: Lessons for experimental design and data analysis," *BMC Genomics*, vol. 12, p. 134, 2011.

[34] D. Ray *et al.*, "Rapid and systematic analysis of the RNA recognition specificities of RNA-binding proteins," *Nature Biotechnol.*, vol. 27, no. 7, pp. 667–670, 2009.

[35] D. Ray *et al.*, "A compendium of RNA-binding motifs for decoding gene regulation," *Nature*, vol. 499, no. 7457, pp. 172–177, 2013.

[36] M. L. Metzker, "Sequencing technologies-the next generation," *Nature Rev. Genetics*, vol. 11, no. 1, pp. 31–46, 2010.

[37] A. Mortazavi, B. A. Williams, K. McCue, L. Schaeffer, and B. Wold, "Mapping and quantifying mammalian transcriptomes by RNA-Seq," *Nature Methods*, vol. 5, pp. 621–628, 2008.

[38] I. R. Watson, K. Takahashi, P. A. Futreal, and L. Chin, "Emerging patterns of somatic mutations in cancer," *Nature Rev. Genetics*, vol. 14, no. 10, pp. 703–718, 2013.

[39] A. Poduri, G. D. Evrony, X. Cai, and C. A. Walsh, "Somatic mutation, genomic variation, and neurological disease," *Science*, vol. 341, no. 6141, pp. 43–51, 2013.

[40] S. G. Landt *et al.*, "ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia," *Genome Res.*, vol. 22, no. 9, pp. 1813–1831, 2012.

[41] B. Alipanahi, A. Delong, M. T. Weirauch, and B. J. Frey, "Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning," *Nature Biotechnol.*, vol. 33, no. 8, pp. 831–838, 2015.

[42] T. van der Heijden, J. J. F. A. van Vugt, C. Logie, and J. van Noort, "Sequence-based prediction of single nucleosome positioning and genome-wide nucleosome occupancy," *Proc. Nat. Acad. Sci. USA*, vol. 109, no. 38, pp. E2514–E2522, 2012.

[43] R. Lorenz *et al.*, "ViennaRNA package 2.0," *Algorithms Mol. Biol.*, vol. 6, no. 1, p. 26, 2011.

[44] H. Y. Xiong *et al.*, "The human splicing code reveals new insights into the genetic determinants of disease," *Science*, vol. 347, no. 6218, 2014, DOI: 10.1126/science.1254806.

[45] T. I. Lee and R. Young, "Transcriptional regulation and its misregulation in disease," *Cell*, vol. 152, no. 6, pp. 1237–1251, 2013.

[46] K. D. Robertson, "DNA methylation and human disease," *Nature Rev. Genetics*, vol. 6, no. 8, pp. 597–610, 2005.

[47] H. Heyn and M. Esteller, "DNA methylation profiling in the clinic: Applications and challenges," *Nature Rev. Genetics*, vol. 13, no. 10, pp. 679–692, 2012.

[48] R. Elkon, A. P. Ugalde, and R. Agami, "Alternative cleavage and polyadenylation: Extent, regulation and function," *Nature Rev. Genetics*, vol. 14, no. 7, pp. 496–506, 2013.

[49] B. Hendrich and W. Bickmore, "Human diseases with underlying defects in chromatin structure and modification," *Human Mol. Gen.*, vol. 10, no. 20, pp. 2233–2242, 2001.

[50] N. Kaplan *et al.*, "The DNA-encoded nucleosome organization of a eukaryotic genome," *Nature*, vol. 458, no. 7236, pp. 362–366, 2009.

[51] C. Laing and T. Schlick, "Computational approaches to RNA structure prediction, analysis, and design," *Current Opinion Struct. Biol.*, vol. 21, no. 3, pp. 306–318, 2011.

[52] T. P. J. Knowles, M. Vendruscolo, and C. M. Dobson, "The amyloid state and its association with protein misfolding diseases," *Nature Rev. Mol. Cell Biol.*, vol. 15, no. 6, pp. 384–396, 2014.

[53] Z. Wang and C. B. Burge, "Splicing regulation: From a parts list of regulatory elements to an integrated splicing code," *RNA*, vol. 14, no. 5, pp. 802–813, 2008.

[54] U. Braunschweig, S. Gueroussov, A. M. Plocik, B. R. Graveley, and B. J. Blencowe, "Dynamic integration of splicing within gene regulatory pathways," *Cell*, vol. 152, no. 6, pp. 1252–1269, 2013.

[55] D. L. Bentley, "Coupling mRNA processing with transcription in time and space," *Nature Rev. Genetics*, vol. 15, no. 3, pp. 163–175, 2014.

[56] R. K. Singh and T. A. Cooper, "Pre-mRNA splicing in disease and therapeutics," *Trends Mol. Med.*, vol. 18, no. 8, pp. 472–482, 2012.

[57] T. W. Nilsen and B. R. Graveley, "Expansion of the eukaryotic proteome by alternative splicing," *Nature*, vol. 463, no. 7280, pp. 457–463, 2010.

[58] I. Evsyukova, S. S. Bradrick, S. G. Gregory, and M. Garcia-Blanco, "Cleavage and polyadenylation specificity factor 1 (CPSF1) regulates alternative splicing of interleukin 7 receptor (IL7R) exon 6," *RNA*, vol. 19, no. 1, pp. 103–115, 2013.

[59] Q. Pan, O. Shai, L. J. Lee, B. J. Frey, and B. J. Blencowe, "Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing," *Nature Genetics*, vol. 40, no. 12, pp. 1413–1415, 2008.

[60] J. Tazi, N. Bakkour, and S. Stamm, "Alternative splicing and disease," *Biochim. Biophys. Acta-Mol. Basis Dis.*, vol. 1792, no. 1, pp. 14–26, 2009.

[61] N. L. Barbosa-Morais *et al.*, "The evolutionary landscape of alternative splicing in vertebrate species," *Science*, vol. 338, no. 6114, pp. 1587–1593, 2012.

[62] I. Voineagu *et al.*, "Transcriptomic analysis of autistic brain reveals convergent molecular pathology," *Nature*, vol. 474, no. 7351, pp. 380–384, 2011.

[63] M. Irimia *et al.*, "A highly conserved program of neuronal microexons is misregulated in autistic brains," *Cell*, vol. 159, no. 7, pp. 1511–1523, 2014.

[64] N. López-Bigas, B. Audit, C. Ouzounis, G. Parra, and R. Guigó, "Are splicing mutations the most frequent cause of hereditary disease?" *FEBS Lett.*, vol. 579, no. 9, pp. 1900–1903, 2005.

[65] K. H. Lim, L. Ferraris, M. E. Filloux, B. J. Raphael, and W. G. Fairbrother, "Using positional distribution to identify splicing elements and predict pre-mRNA processing defects in human genes," *Proc. Nat. Acad. Sci. USA*, vol. 108, no. 27, pp. 11093–11098, 2011.

[66] G.-S. Wang and T. A. Cooper, "Splicing in disease: Disruption of the splicing code and the decoding machinery," *Nature Rev. Genetics*, vol. 8, no. 10, pp. 749–761, 2007.

[67] S. Stamm, *Encyclopedia of Life Sciences*. Chichester, U.K.: Wiley, 2011.

[68] R. F. Luco, M. Allo, I. E. Schor, A. R. Kornblihtt, and T. Misteli, "Epigenetics in alternative pre-mRNA splicing," *Cell*, vol. 144, no. 1, pp. 16–26, 2011.

[69] Y. Barash *et al.*, "Deciphering the splicing code," *Nature*, vol. 465, no. 7294, pp. 53–59, 2010.

[70] H. Xiong, Y. Barash, and B. Frey, "Bayesian prediction of tissue-regulated splicing using RNA sequence and cellular context," *Bioinformatics*, vol. 27, no. 18, pp. 2554–2562, 2011.

[71] M. K. K. Leung, H. Y. Xiong, L. J. Lee, and B. J. Frey, "Deep learning of the tissue-regulated splicing code," *Bioinformatics*, vol. 30, no. 12, pp. i121–i129, 2014.

[72] Z. Wang, M. Gerstein, and M. Snyder, "RNA-Seq: A revolutionary tool for transcriptomics," *Nature Rev. Genetics*, vol. 10, no. 1, pp. 57–63, 2009.

[73] M. Garber, M. G. Grabherr, M. Guttman, and C. Trapnell, "Computational methods for transcriptome annotation and quantification using RNA-seq," *Nature Methods*, vol. 8, no. 6, pp. 469–477, 2011.

[74] B. Kakaradov, H. Y. Xiong, L. J. Lee, N. Jojic, and B. J. Frey, "Challenges in estimating percent inclusion of alternatively spliced junctions from RNA-seq data," *BMC Bioinf.*, vol. 13, no. Suppl. 6, p. S11, 2012.

[75] I. De Castro-Orós *et al.*, "Functional analysis of LDLR promoter and 5′ UTR mutations in subjects with clinical diagnosis of familial hypercholesterolemia," *Human Mutat.*, vol. 32, no. 8, pp. 868–872, 2011.

[76] F. W. Huang *et al.*, "Highly recurrent TERT promoter mutations in human melanoma," *Science*, vol. 339, no. 6122, pp. 957–959, 2013.

[77] Q. Huang *et al.*, "A prostate cancer susceptibility allele at 6q22 increases RFX6 expression by modulating HOXB13 chromatin binding," *Nature Genetics* , vol. 46, no. 2, pp. 126–135, 2014.

[78] J. M. Vaquerizas, S. K. Kummerfeld, S. Teichmann, and N. M. Luscombe, "A census of human transcription factors: Function, expression and evolution," *Nature Rev. Genetics*, vol. 10, no. 4, pp. 252–263, 2009.

[79] S. Gerstberger, M. Hafner, and T. Tuschl, "A census of human RNA-binding proteins," *Nature Rev. Genetics*, vol. 15, no. 12, pp. 829–845, 2014.

[80] D. Herschlag, "RNA chaperones and the RNA folding problem," *J. Biol. Chem.*, vol. 270, no. 36, pp. 20871–20874, 1995.

[81] B. J. Blencowe, "Alternative splicing: New insights from global analyses," *Cell*, vol. 126, no. 1, pp. 37–47, 2006.

[82] M. T. Weirauch *et al.*, "Determination and inference of eukaryotic transcription factor sequence specificity," *Cell*, vol. 158, no. 6, pp. 1431–1443, 2014.

[83] M. Levo and E. Segal, "In pursuit of design principles of regulatory sequences," *Nature Rev. Genetics*, vol. 15, no. 7, pp. 453–468, 2014.

[84] K. B. Cook, T. R. Hughes, and Q. D. Morris, "High-throughput characterization of protein-RNA interactions," *Brief. Funct. Genomics*, vol. 14, no. 1, pp. 74–89, 2014.

[85] J. D. J. Deng *et al.*, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 248–255.

[86] P. Machanick and T. L. Bailey, "MEME-ChIP: Motif analysis of large DNA datasets," *Bioinformatics*, vol. 27, no. 12, pp. 1696–1697, 2011.

[87] M. F. Berger and M. L. Bulyk, "Universal protein-binding microarrays for the comprehensive characterization of the DNA-binding specificities of transcription factors," *Nature Protoc.*, vol. 4, no. 3, pp. 393–411, 2009.

[88] G. D. Stormo, "DNA binding sites: Representation and discovery," *Bioinformatics*, vol. 16, no. 1, pp. 16–23, 2000.

[89] T. D. Schneider and R. M. Stephens, "Sequence logos: A new way to display consensus sequences," *Nucleic Acids Res.*, vol. 18, no. 1990, pp. 6097–6100, 1990.

[90] B. L. Fogel *et al.*, "RBFOX1 regulates both splicing and transcriptional networks in human neuronal development," *Human Mol. Genetics*, vol. 21, no. 19, pp. 4171–4186, 2012.

[91] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2323, Nov. 1998.

[92] G. Badis *et al.*, "Diversity and complexity in DNA recognition by transcription factors," *Science*, vol. 324, no. 5935, pp. 1720–1723, 2009.

[93] B. C. Foat, A. V. Morozov, and H. J. Bussemaker, "Statistical mechanical modeling of genome-wide transcription factor occupancy data by MatrixREDUCE," *Bioinformatics*, vol. 22, no. 14, pp. 141–149, 2006.

[94] E. Sharon, S. Lubliner, and E. Segal, "A feature-based approach to modeling protein-DNA interactions," *PLoS Comput. Biol.*, vol. 4, no. 8, 2008, DOI: 10.1371/journal.pcbi.1000154.

[95] K. N. Lam, H. Van Bakel, A. G. Cote, A. Van Der Ven, and T. R. Hughes, "Sequence specificity is obtained from the majority of modular C2H2 zinc-finger arrays," *Nucleic Acids Res.*, vol. 39, no. 11, pp. 4680–4690, 2011.

[96] A. Gupta *et al.*, "An improved predictive recognition model for Cys2-His2 zinc finger proteins," *Nucleic Acids Res.*, vol. 42, no. 8, pp. 4800–4812, 2014.

[97] X. Li, G. Quon, H. D. Lipshitz, and Q. Morris, "Predicting *in vivo* binding sites of RNA-binding proteins using mRNA secondary structure," *RNA*, vol. 16, no. 6, pp. 1096–1107, 2010.

[98] G. D. Stormo and Y. Zhao, "Determining the specificity of protein-DNA interactions," *Nature Rev. Genetics*, vol. 11, no. 11, pp. 751–760, 2010.

[99] M. Slattery *et al.*, "Absence of a simple code: How transcription factors read the genome," *Trends Biochem. Sci.*, vol. 39, no. 9, pp. 381–399, 2014.

[100] M. T. Weirauch *et al.*, "Evaluation of methods for modeling transcription factor sequence specificity," *Nature Biotechnol.*, vol. 31, no. 2, pp. 126–134, 2013.

[101] I. Jarchum and S. Jones, "DREAMing of benchmarks," *Nature Biotechnol.*, vol. 33, no. 1, pp. 49–50, 2015.

[102] Q. Zhou and J. S. Liu, "Extracting sequence features to predict protein-DNA interactions: A comparative study," *Nucleic Acids Res.*, vol. 36, no. 12, pp. 4137–4148, 2008.

[103] A. Graves, A. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2013, pp. 6645–6649.

[104] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[105] Y. LeCun, C. Cortes, and C. J. C. Burges, *The MNIST database of handwritten digits*, 1998 . [Online]. Available: http://yann.lecun.com/exdb/mnist/.

[106] M. D. Ritchie, E. R. Holzinger, R. Li, S. A. Pendergrass, and D. Kim, "Methods of integrating data to uncover genotype-phenotype interactions," *Nature Rev. Genetics*, vol. 16, no. 2, pp. 85–97, 2015.

[107] F. W. Albert and L. Kruglyak, "The role of regulatory variation in complex traits and disease," *Nature Rev. Genetics*, vol. 16, no. 4, pp. 197–212, 2015.

[108] M. Costanzo *et al.*, "The genetic landscape of a cell," *Science*, vol. 327, no. 2010, pp. 425–431, 2010.

[109] B. Lehner, "Genotype to phenotype: Lessons from model organisms for human genetics," *Nature Rev. Genetics*, vol. 14, no. 3, pp. 168–178, 2013.

[110] P. M. Visscher, M. A. Brown, M. I. McCarthy, and J. Yang, "Five years of GWAS discovery," *Amer. J. Human Genetics*, vol. 90, no. 1, pp. 7–24, 2012.

[111] J. K. Pritchard and M. Przeworski, "Linkage disequilibrium in humans: Models and data," *Amer. J. Human Genetics*, vol. 69, pp. 1–14, 2001.

[112] D. G. MacArthur *et al.*, "Guidelines for investigating causality of sequence variants in human disease," *Nature*, vol. 508, no. 7497, pp. 469–476, 2014.

[113] R. C. Johnson *et al.*, "Accounting for multiple comparisons in a genome-wide association study (GWAS)," *BMC Genomics*, vol. 11, p. 724, 2010.

[114] G. Gibson, "Rare and common variants: Twenty arguments," *Nature Rev. Genetics*, vol. 13, no. 2, pp. 135–145, 2012.

[115] M. J. Bamshad *et al.*, "Exome sequencing as a tool for Mendelian disease gene discovery," *Nature Rev. Genetics*, vol. 12, no. 11, pp. 745–755, 2011.

[116] S. B. Ng *et al.*, "Targeted capture and massively parallel sequencing of 12 human exomes," *Nature*, vol. 461, no. 7261, pp. 272–276, 2009.

[117] "A haplotype map of the human genome," *Nature*, vol. 437, no. 7063, pp. 1299–1320, 2005.

[118] D. M. Altshuler *et al.*, "Integrating common and rare genetic variation in diverse human populations," *Nature*, vol. 467, no. 7311, pp. 52–58, 2010.

[119] E. Skafidas *et al.*, "Predicting the diagnosis of autism spectrum disorder using gene pathway analysis," *Mol. Psychiatry*, vol. 19, pp. 504–510, 2014.

[120] E. B. Robinson *et al.*, "Response to 'Predicting the diagnosis of autism spectrum disorder using gene pathway analysis,'" *Mol. Psychiatry*, vol. 19, no. 8, pp. 859–861, 2014.

[121] I. A. Adzhubei *et al.*, "A method and server for predicting damaging missense mutations," *Nature Methods*, vol. 7, no. 4, pp. 248–249, 2010.

[122] P. Kumar, S. Henikoff, and P. C. Ng, "Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm," *Nature Protoc.*, vol. 4, no. 8, pp. 1073–1081, 2009.

[123] C. Kooperberg, M. LeBlanc, and V. Obenchain, "Risk prediction using genome-wide association studies," *Genetics Epidemiol.*, vol. 34, no. 7, pp. 643–652, 2010.

[124] J. Kruppa, A. Ziegler, and I. R. König, "Risk estimation and risk prediction using machine-learning methods," *Human Genetics*, vol. 131, no. 10, pp. 1639–1654, 2012.

[125] A. Ureta-Vidal, L. Ettwiller, and E. Birney, "Comparative genomics: Genome-wide analysis in metazoan eukaryotes," *Nature Rev. Genetics*, vol. 4, no. 4, pp. 251–262, 2003.

[126] E. T. Dermitzakis *et al.*, "Evolutionary discrimination of mammalian conserved non-genic sequences (CNGs)," *Science*, vol. 302, no. 5647, pp. 1033–1035, 2003.

[127] A. Siepel *et al.*, "Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes," *Genome Res.*, vol. 15, no. 8, pp. 1034–1050, 2005.

[128] G. M. Cooper *et al.*, "Distribution and intensity of constraint in mammalian genomic sequence," *Genome Res.*, vol. 15, no. 7, pp. 901–913, 2005.

[129] K. S. Pollard, M. J. Hubisz, K. R. Rosenbloom, and A. Siepel, "Detection of nonneutral substitution rates on mammalian phylogenies," *Genome Res.*, vol. 20, no. 1, pp. 110–121, 2010.

[130] W. James Kent *et al.*, "The human genome browser at UCSC," *Genome Res.*, vol. 12, no. 6, pp. 996–1006, 2002.

[131] M. Kircher *et al.*, "A general framework for estimating the relative pathogenicity of human genetic variants," *Nature Genetics*, vol. 46, no. 3, pp. 310–315, 2014.

[132] L. Cartegni, M. L. Hastings, J. A. Calarco, E. de Stanchina, and A. R. Krainer, "Determinants of exon 7 splicing in the spinal muscular atrophy genes, SMN1 and SMN2," *Amer. J. Human Genetics*, vol. 78, no. 1, pp. 63–77, 2006.

[133] D. Quang, Y. Chen, and X. Xie, "DANN: A deep learning approach for annotating the pathogenicity of genetic variants," *Bioinformatics*, vol. 31, no. 5, pp. 761–763, 2014.

[134] L. W. Barrett, S. Fletcher, and S. D. Wilton, "Regulation of eukaryotic gene expression by the untranslated gene regions and other non-coding elements," *Cell. Mol. Life Sci.*, vol. 69, pp. 3613–3634, 2012.

[135] M. D. Mailman *et al.*, "The NCBI dbGaP database of genotypes and phenotypes," *Nature Genetics*, vol. 39, no. 10, pp. 1181–1186, 2007.

[136] A. Burga and B. Lehner, "Beyond genotype to phenotype: Why the phenotype of an individual cannot always be predicted from their genome sequence and the environment that they experience," *FEBS J.*, vol. 279, no. 20, pp. 3765–3775, 2012.

[137] J. R. Hart *et al.*, "The butterfly effect in cancer: A single base mutation can remodel the cell," *Proc. Nat. Acad. Sci.*, vol. 112, no. 4, pp. 1131–1136, 2015.

[138] D. Devos and A. Valencia, "Practical limits of function prediction," *Proteins Struct. Funct. Genetics*, vol. 41, no. 1, pp. 98–107, 2000.

[139] D. M. Fowler *et al.*, "High-resolution mapping of protein sequence-function relationships," *Nature Methods*, vol. 7, no. 9, pp. 741–746, 2010.

[140] A. Hannun *et al.*, "Deep speech: Scaling up end-to-end speech recognition," arXiv1412.5567, 2014, pp. 1–12.

[141] R. Wu, S. Yan, Y. Shan, Q. Dang, and G. Sun, "Deep image: Scaling up image recognition," arXiv1501.02876, 2015.

[142] J. Dean *et al.*, "Large scale distributed deep networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1223–1231.

[143] A. Agarwal, O. Chapelle, M. Dudík, and J. Langford, "A reliable effective terascale linear learning system," *J. Mach. Learn. Res.*, vol. 15, pp. 1111–1133, 2014.

[144] A. Coates, B. Huval, T. Wang, D. J. Wu, and A. Y. Ng, "Deep learning with COTS HPC systems," in *Proc. 30th Int. Conf. Mach. Learn.*, 2013, vol. 28, pp. 1337–1345.

[145] J. T. Dudley and A. J. Butte, "In silico research in the era of cloud computing," *Nature Biotechnol.*, vol. 28, no. 11, pp. 1181–1185, 2010.

[146] L. D. Stein, "The case for cloud computing in genome informatics," *Genome Biol.*, vol. 11, no. 5, p. 207, 2010.

[147] L. D. Stein, B. M. Knoppers, P. Campbell, G. Getz, and J. O. Korbel, "Data analysis: Create a cloud commons," *Nature*, vol. 523, no. 7559, pp. 149–151, 2015.

[148] D. Sculley *et al.*, "Machine learning: The high interest credit card of technical debt," in *Proc. SE4ML Softw. Eng. Mach. Learn.*, 2014, pp. 1–9.

[149] J. Shim *et al.*, "Nanopore-based assay for detection of methylation in double-stranded DNA fragments," *ACS Nano*, vol. 9, no. 1, pp. 290–300, 2015.

[150] H. Tilgner *et al.*, "Accurate identification and analysis of human mRNA isoforms using deep long read sequencing," *G3 (Bethesda)*, vol. 3, no. 3, pp. 387–397, 2013.

[151] D. Sharon, H. Tilgner, F. Grubert, and M. Snyder, "A single-molecule long-read survey of the human transcriptome," *Nature Biotechnol.*, vol. 31, no. 11, pp. 1009–1014, 2013.

[152] B. Treutlein, O. Gokce, S. R. Quake, and T. C. Südhof, "Cartography of neurexin alternative splicing mapped by single-molecule long-read mRNA sequencing," *Proc. Nat. Acad. Sci. USA*, vol. 111, no. 13, pp. E1291–E1299, 2014.

[153] O. Stegle, S. A. Teichmann, and J. C. Marioni, "Computational and analytical challenges in single-cell transcriptomics," *Nature Rev. Genetics*, vol. 16, no. 1, pp. 133–145, 2015.

[154] G. K. Marinov *et al.*, "From single-cell to cell-pool transcriptomes: Stochasticity in gene expression and RNA splicing," *Genome Res.*, vol. 24, no. 3, pp. 496–510, 2014.

[155] S. S. Dey, L. Kester, B. Spanjaard, and A. Van, "Integrated genome and transcriptome sequencing from the same cell," *Nature Biotechnol.*, vol. 33, no. 3, pp. 285–289, 2015.

[156] J. Ma, R. P. Sheridan, A. Liaw, G. Dahl, and V. Svetnik, "Deep neural nets as a method for quantitative structure-activity relationships," *J. Chem. Inf. Model.*, vol. 55, no. 2, pp. 263–274, 2015.

[157] C. N. Pace, B. A. Shirley, M. McNutt, and K. Gajiwala, "Forces contributing to the conformational stability of proteins," *FASEB J.*, vol. 10, no. 1, pp. 75–83, 1996.

[158] C. L. Kingsford, B. Chazelle, and M. Singh, "Solving and analyzing side-chain positioning problems using linear and integer programming," *Bioinformatics*, vol. 21, no. 7, pp. 1028–1036, 2005.

[159] C. Yanover and Y. Weiss, "Approximate inference and protein-folding," in *Proc. Adv. Neural Inf. Process. Syst.*, 2002, pp. 1457–1464.

[160] Y. Weiss, C. Yanover, and T. Meltzer, "MAP estimation, linear programming and belief propagation with convex free energies," in *Proc. 23rd Conf. Uncertain. Artif. Intell.*, 2007, pp. 416–425.

[161] D. Sontag, T. Meltzer, A. Globerson, T. Jaakkola, and Y. Weiss, "Tightening LP relaxations for MAP using message passing," arXiv1206.3288, 2008.

[162] V. Jojic, S. Gould, and D. Koller, "Accelerated dual decomposition for MAP inference," in *Proc. 27th Int. Conf. Mach. Learn.*, 2010, pp. 503–510.

[163] P. Radivojac *et al.*, "A large-scale evaluation of computational protein function prediction," *Nature Methods*, vol. 10, no. 3, pp. 221–227, 2013.

[164] K. R. Lakhani *et al.*, "Prize-based contests can provide solutions to computational biology problems," *Nature Biotechnol.*, vol. 31, no. 2, pp. 108–111, 2013.

[165] M. C. Frith *et al.*, "Evolutionary turnover of mammalian transcription start sites," *Genome Res.*, vol. 16, pp. 713–722, 2006.

[166] M. Kellis *et al.*, "Defining functional DNA elements in the human genome," *Proc. Nat. Acad. Sci. USA*, vol. 111, no. 17, pp. 6131–6138, 2014.

[167] J. Xu and J. Zhang, "Why human disease-associated residues appear as the wild-type in other species: Genome-scale structural evidence for the compensation hypothesis," *Mol. Biol. Evol.*, vol. 31, no. 7, pp. 1787–1792, 2014.

[168] P. D. Stenson *et al.*, "The human gene mutation database (HGMD) and its exploitation in the fields of personalized genomics and molecular evolution," *Current Protoc. Bioinf.*, pp. 1–13, 2012.

[169] T. Mikolov, "Statistical language models based on neural networks," Ph.D. dissertation, Brno Univ. Technol., Brno, Czech Republic, 2012.

[170] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," arXiv1411.4555, 2014.

[171] A. Krogn, "Two methods for improving performance of an HMM and their application for gene finding," in *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, 1997, vol. 5, pp. 179–186.

[172] J. Ernst and M. Kellis, "Discovery and characterization of chromatin states for systematic annotation of the human genome," *Nature Biotechnol.*, vol. 28, no. 8, pp. 817–825, 2010.

[173] J. Ernst and M. Kellis, "ChromHMM: Automating chromatin-state discovery and characterization," *Nature Methods*, vol. 9, no. 3, pp. 215–216, 2012.

[174] M. M. Hoffman *et al.*, "Integrative annotation of chromatin elements from ENCODE data," *Nucleic Acids Res.*, vol. 41, no. 2, pp. 827–841, 2013.

[175] H. A. Shihab *et al.*, "Predicting the functional, molecular, phenotypic consequences of amino acid substitutions using hidden Markov models," *Human Mutat.*, vol. 34, no. 1, pp. 57–65, 2013.

[176] J. R. Karr *et al.*, "A whole-cell computational model predicts phenotype from genotype," *Cell*, vol. 150, no. 2, pp. 389–401, 2012.

[177] J. R. Karr, N. C. Phillips, and M. W. Covert, "WholeCellSimDB: A hybrid relational/HDF database for whole-cell model predictions," *Database*, vol. 2014, pp. bau095–bau095, 2014.

[178] O. G. Berg, R. B. Winter, and P. H. von Hippel, "Diffusion-driven mechanisms of protein translocation on nucleic acids. 1. Models and theory," *Biochemistry*, vol. 20, no. 24, pp. 6929–6948, 1981.

[179] J. Ernst and M. Kellis, "Large-scale imputation of epigenomic datasets for systematic annotation of diverse human tissues," *Nature Biotechnol.*, vol. 33, no. 4, pp. 364–376, 2015.

[180] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," arXiv1409.3215, 2014.

[181] Y. Kodratoff, "The comprehensibility manifesto," *KDD Nugget Newslett.*, vol. 94, no. 9, 1994.

[182] C. Rudin and K. L. Wagstaff, "Machine learning for science and society," *Mach. Learn.*, vol. 95, no. 1, pp. 1–9, 2013.

[183] D. Erhan, Y. Bengio, A. Courville, and P. Vincent, "Visualizing higher-layer features of a deep network," Dept. IRO, Univ. Montréal, Montréal, QC, Canada, Tech. Rep., 2009, vol. 1341.

[184] Q. V. Le *et al.*, "Building high-level features using large scale unsupervised learning," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2013, pp. 8595–8598.

[185] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Proc. Comput. Vis.*, 2013, pp. 818–833.

[186] K. Simonyan and A. Vedaldi, "Deep inside convolutional networks: Visualising image classification models and saliency maps," in *Proc. Int. Conf. Learn. Represent.* arXiv1312.6034, 2014.

[187] A. Mahendran and A. Vedaldi, "Understanding deep image representations by inverting them," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 5188–5196.

[188] K. Lenc and A. Vedaldi, "Understanding image representations by measuring their equivariance and equivalence," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 991–999.

[189] Y. Le Cun, J. S. Denker, and S. a Solla, "Optimal brain damage," in *Proc. Adv. Neural Inf. Process. Syst.*, 1990, vol. 2, pp. 598–605.

[190] R. Rigamonti, A. Sironi, V. Lepetit, and P. Fua, "Learning separable filters," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2013, pp. 2754–2761.

[191] E. Denton, W. Zaremba, and J. Bruna, "Exploiting linear structure within

convolutional networks for efficient evaluation," arXiv1404.0736, 2014.

[192] M. Jaderberg, A. Vedaldi, and A. Zisserman, "Speeding up convolutional neural networks with low rank expansions," arXiv1405.3866, 2014.

[193] V. Lebedev, Y. Ganin, M. Rakhuba, I. Oseledets, and V. Lempitsky, "Speeding-up convolutional neural networks using fine-tuned CP-decomposition," in *Int. Conf. Learn. Represent.* arXiv1412.6553, 2014.

[194] C. Bucilua, R. Caruana, and A. Niculescu-Mizil, "Model compression," in *Proc. 12th ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2006, p. 535.

[195] A. H. Beck *et al.*, "Systematic analysis of breast cancer morphology uncovers stromal features associated with survival," *Sci. Transl. Med.*, vol. 3, no. 108, pp. 108ra113–108ra113, 2011.

[196] C. Kingsford and S. L. Salzberg, "What are decision trees?," *Nature Biotechnol.*, vol. 26, no. 9, pp. 1011–1013, 2008.

[197] C. Szegedy, W. Zaremba, and I. Sutskever, "Intriguing properties of neural networks," arXiv1312.6199, 2013.

[198] A. Nguyen, J. Yosinski, and J. Clune, "Deep neural networks are easily fooled: High confidence predictions for unrecognizable images," arXiv1412.1897, 2014.

[199] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *Proc. Int. Conf. Learn. Represent.* arXiv1412.6572, 2015.

[200] J. E. Garneau *et al.*, "The CRISPR/Cas bacterial immune system cleaves bacteriophage and plasmid DNA," *Nature*, vol. 468, no. 7320, pp. 67–71, 2010.

[201] L. S. Qi *et al.*, "Repurposing CRISPR as an RNA-guided platform for sequence-specific control of gene expression," *Cell*, vol. 152, no. 5, pp. 1173–1183, 2013.

[202] L. Swiech *et al.*, "In vivo interrogation of gene function in the mammalian brain using CRISPR-Cas9," *Nature Biotechnol.*, vol. 33, no. 1, pp. 102–106, 2014.

[203] L. Nissim, S. D. Perli, A. Fridkin, P. Perez-Pinera, and T. K. Lu, "Multiplexed and programmable regulation of gene networks with an integrated RNA and CRISPR/Cas toolkit in human cells," *Mol. Cell*, vol. 54, no. 4, pp. 698–710, 2014.

[204] Y. Wu *et al.*, "Correction of a genetic disease in mouse via use of CRISPR-Cas9," *Cell Stem Cell*, vol. 13, no. 6, pp. 659–662, 2013.

[205] G. Schwank *et al.*, "Functional repair of CFTR by CRISPR/Cas9 in intestinal stem cell organoids of cystic fibrosis patients," *Cell Stem Cell*, vol. 13, no. 6, pp. 653–658, 2013.

[206] D. B. T. Cox, R. J. Platt, and F. Zhang, "Therapeutic genome editing: Prospects and challenges," *Nature Med.*, vol. 21, no. 2, pp. 121–131, 2015.

[207] H. Xu *et al.*, "Sequence determinants of improved CRISPR sgRNA design," *Genome Res.*, 2015, DOI: 10.1101/gr.191452.115.

[208] M. Yandell and D. Ence, "A beginner's guide to eukaryotic genome annotation," *Nature Rev. Genetics*, vol. 13, no. 5, pp. 329–342, 2012.

[209] R. P. Alexander, G. Fang, J. Rozowsky, M. Snyder, and M. B. Gerstein, "Annotating non-coding regions of the genome," *Nature Rev. Genetics*, vol. 11, no. 8, pp. 559–571, 2010.

[210] K. Y. Yip, C. Cheng, and M. Gerstein, "Machine learning and genome annotation: A match meant to be?" *Genome Biol.*, vol. 14, no. 5, p. 205, 2013.

[211] S. Sonnenburg, G. Schweikert, P. Philips, J. Behr, and G. Rätsch, "Accurate splice site prediction using support vector machines," *BMC Bioinf.*, vol. 8, no. Suppl. 10, p. S7, 2007.

[212] Y. Saeys, T. Abeel, S. Degroeve, and Y. Van de Peer, "Translation initiation site prediction on a genomic scale: Beauty in simplicity," *Bioinformatics*, vol. 23, no. 1987, pp. 418–423, 2007.

[213] G. A. Maston, S. K. Evans, and M. R. Green, "Transcriptional regulatory elements in the human genome," *Annu. Rev. Genomics Human Genetics*, vol. 7, pp. 29–59, 2006.

[214] S. Danckwardt, M. W. Hentze, and A. E. Kulozik, "3′ end mRNA processing: Molecular mechanisms and implications for health and disease," *EMBO J.*, vol. 27, no. 3, pp. 482–498, Mar. 2008.

[215] M. N. Akhtar, S. A. Bukhari, Z. Fazal, R. Qamar, and I. A. Shahmuradov, "POLYAR, a new computer program for prediction of poly(A) sites in human sequences," *BMC Genomics*, vol. 11, no. 1, p. 646, 2010.

[216] T.-H. Chang *et al.*, "Characterization and prediction of mRNA polyadenylation sites in human genes," *Med. Biol. Eng. Comput.*, vol. 49, no. 4, pp. 463–472, 2011.

[217] Y. Wan, M. Kertesz, R. C. Spitale, E. Segal, and H. Y. Chang, "Understanding the transcriptome through RNA structure," *Nature Rev. Genetics*, vol. 12, no. 9, pp. 641–655, 2011.

[218] C. A. Floudas, "Computational methods in protein structure prediction," *Biotechnol. Bioeng.*, vol. 97, no. 2, pp. 207–213, 2007.

[219] O. G. Troyanskaya, "Deep supervised and convolutional generative stochastic network for protein secondary structure prediction," in *Proc. 31st Int. Conf. Mach. Learn.*, 2014, vol. 32, pp. 745–753.

[220] P. Di lena, K. Nagata, and P. Baldi, "Deep architectures for protein contact map prediction," *Bioinformatics*, vol. 28, no. 19, pp. 2449–2457, 2012.

[221] T. G. Consortium, "The genotype-tissue expression (GTEx) project," *Nature Genetics*, vol. 45, no. 6, pp. 580–585, 2013.

[222] R. H. Shoemaker, "The NCI60 human tumour cell line anticancer drug screen," *Nature Rev. Cancer*, vol. 6, no. 10, pp. 813–823, 2006.

[223] T. J. Hudson *et al.*, "International network of cancer genome projects," *Nature*, vol. 464, no. 7291, pp. 993–998, 2010.

[224] K. Chang *et al.*, "The cancer genome atlas pan-cancer analysis project," *Nature Genetics*, vol. 45, no. 10, pp. 1113–1120, 2013.

[225] J. Li *et al.*, "TCPA: A resource for cancer functional proteomics data," *Nature Methods*, vol. 10, no. 11, pp. 1046–1047, 2013.

[226] G. Project *et al.*, "An integrated map of genetic variation from 1,092 human genomes," *Nature*, vol. 491, no. 7422, pp. 556–665, 2012.

[227] B. E. Bernstein *et al.*, "The NIH roadmap epigenomics mapping consortium," *Nature Biotechnol.*, vol. 28, no. 10, pp. 1045–1048, 2010.

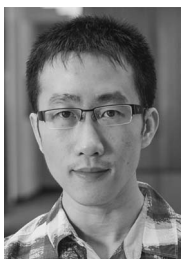
[228] R. E. Consortium *et al.*, "Integrative analysis of 111 reference human epigenomes," *Nature*, vol. 518, no. 7539, pp. 317–330, 2015.

[229] A. R. Wood *et al.*, "Defining the role of common variation in the genomic and biological architecture of adult human height.," *Nature Genetics*, vol. 46, no. 11, pp. 1173–1186, 2014.

[230] D. Shungin *et al.*, "New genetic loci link adipose and insulin biology to body fat distribution," *Nature*, vol. 518, no. 7538, pp. 187–196, 2015.

[231] A. E. Locke *et al.*, "Genetic studies of body mass index yield new insights for obesity biology," *Nature*, vol. 518, no. 7538, pp. 197–206, 2015.

ABOUT THE AUTHORS



Michael K. K. Leung received the B.A.Sc. degree in engineering science and the M.Sc. degree in medical biophysics from the University of Toronto, Toronto, ON, Canada, in 2007 and 2010, respectively, where he is currently working toward the Ph.D. degree in the Department of Electrical and Computer Engineering.

His research interests span various domains, having coauthored peer-reviewed papers in the fields of machine learning, computational biology, medical imaging, and radiation oncology.

Mr. Leung was awarded a Natural Sciences and Engineering Research Council of Canada (NSERC) Graduate Scholarship and a Canadian Institutes of Health Research (CIHR) Graduate Scholarship for his works at the University of Toronto.

Andrew Delong received the B.Math. degree in computer science from the University of Waterloo, Waterloo, ON, Canada, in 2003 and the M.Sc. and Ph.D. degrees in computer science from Western University, London, ON, Canada, in 2006 and 2011, respectively.

He is a Postdoctoral Fellow in the Department of Electrical and Computer Engineering, University of Toronto, Toronto, ON, Canada. Prior to entering graduate school he worked in the computer graphics industry. His research interests include machine learning, computational biology, computer vision, and combinatorial optimization.

Dr. Delong was awarded a Natural Sciences and Engineering Research Council of Canada (NSERC) Postgraduate Scholarship, an NSERC Postdoctoral Fellowship, and a Heffernan Commercialization Fellowship from the University of Toronto.

Babak Alipanahi received the B.Sc. degree in electrical engineering from Amirkabir University of Technology, Tehran, Iran, in 2003, the M.Sc. degree in electrical engineering from the University of Tehran, Tehran, Iran, in 2005, and the Ph.D. degree from the Bioinformatics Lab, University of Waterloo, Waterloo, ON, Canada, in 2011, working on protein NMR structure determination.

He did his Postdoctoral Fellowship at the University of Toronto, Toronto, ON, Canada, from 2011 to 2015, working on gene regulatory models and disease genomics. The focus of his research is on designing deep learning algorithms for modeling cellular processes and using them to analyze noncoding variants. In particular, his research interests include modeling alternative splicing, protein/nucleic acid binding, and discovery of genes involved in different complex polygenic disorders, such as cancer and autism.

Brendan J. Frey received the B.Sc. degree in computer engineering and physics from the University of Calgary, Calgary, AB, Canada, in 1990, the M.Sc. degree from the University of Manitoba, Winnipeg, MB, Canada, in 1993, and the Ph.D. degree from the University of Toronto, Toronto, ON, Canada, in 1997.

He is a Professor at the University of Toronto, Toronto, ON, Canada, with appointments in Engineering and Medicine, and he is the CEO and President of Deep Genomics Inc. He has consulted for several industrial research and development laboratories in Canada, the United States, and the United Kingdom, and he has served on the Technical Advisory Board of Microsoft Research. His former students and postdoctoral fellows include professors, entrepreneurs, and industrial researchers at centers across Canada, the United States, and Europe. He conducts research in the fields of genome biology and machine learning, and is best known for his work on using machine learning to understand the genome and how natural and therapeutic variations impact disease, the splicing code, affinity propagation, and factor graphs.

Dr. Frey holds the Canada Research Chair in Biological Computation, and is a Fellow of the Royal Society of Canada, the Canadian Institute for Advanced Research, the American Institute for the Advancement of Science, and the Institute of Electrical and Electronic Engineers. He has received several distinctions, including the John C Polanyi Award, the EWR Steacie Fellowship, and Canada's Top 40 Leaders Under 40 Award.