ANNUAL REVIEWS

Annual Review of Biomedical Data Science

Computational Approaches for Unraveling the Effects of Variation in the Human Genome and Microbiome

Chengsheng Zhu,[1] Maximilian Miller,[1] Zishuo Zeng,[1] Yanran Wang,[1] Yannick Mahlich,[1] Ariel Aptekmann,[1] and Yana Bromberg[1,2]

[1] Department of Biochemistry and Microbiology, Rutgers University, New Brunswick, New Jersey 08873, USA; email: czhu@bromberglab.org, yana@bromberglab.org

[2] Department of Genetics, Rutgers University, Piscataway, New Jersey 08854, USA

Annu. Rev. Biomed. Data Sci. 2020. 3:411–32

First published as a Review in Advance on
May 12, 2020

The *Annual Review of Biomedical Data Science* is
online at biodatasci.annualreviews.org

https://doi.org/10.1146/annurev-biodatasci-030320-041014

Keywords

computational metagenomics, genome variants, microbiome, microbiome–genome interaction, metagenome, pharmacogenomics, pharmacomicrobiomics, variant effect prediction, precision medicine

Abstract

The past two decades of analytical efforts have highlighted how much more remains to be learned about the human genome and, particularly, its complex involvement in promoting disease development and progression. While numerous computational tools exist for the assessment of the functional and pathogenic effects of genome variants, their precision is far from satisfactory, particularly for clinical use. Accumulating evidence also suggests that the human microbiome's interaction with the human genome plays a critical role in determining health and disease states. While numerous microbial taxonomic groups and molecular functions of the human microbiome have been associated with disease, the reproducibility of these findings is lacking. The human microbiome–genome interaction in healthy individuals is even less well understood. This review summarizes the available computational methods built to analyze the effect of variation in the human genome and microbiome. We address the applicability and precision of these methods across their possible uses. We also briefly discuss the exciting, necessary, and now possible integration of the two types of data to improve the understanding of pathogenicity mechanisms.

> We propose that those with limited knowledge in a domain suffer a dual burden: Not only do they reach mistaken conclusions and make regrettable errors, but their incompetence robs them of the ability to realize it.
>
> —J. Kruger & D. Dunning (1)

THE BEGINNING

With every new and exciting discovery relevant to human health comes the realization that science is still very far away from a broad understanding of how to diagnose, prevent, and treat diseases. By April 2003, the Human Genome Project (HGP) created, at a cost around \$2.7 billion, a reference sequence from a compilation of the genomes of several individuals (2). The first completely sequenced individual-specific genomes of J. Craig Venter (3) and James D. Watson (4) were 1,000-fold less expensive but still cost roughly \$1 million each. Since then, sequencing prices have dropped sufficiently to allow for large-scale studies designed to understand how human genetic variation, initially mainly single-nucleotide variants (SNVs), contribute to human complex traits. [In the scientific literature, SNVs that are frequent in the population (e.g., >1%) are termed single-nucleotide polymorphisms (SNPs; see **Figure 1** for the potential biological effects of SNPs). However, since this threshold is arbitrary and constantly moving, in this review

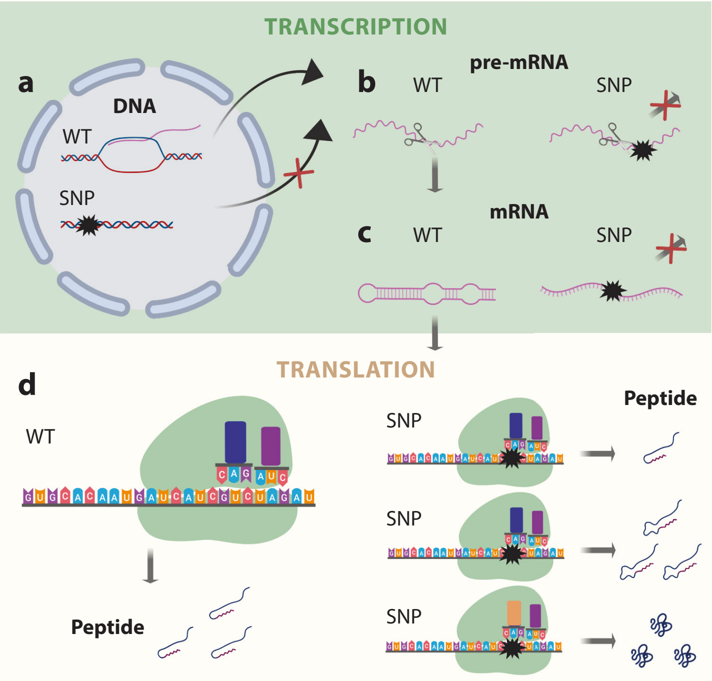

Figure 1

Mechanisms of variant impact on biological function. SNPs may affect (*a*) transcription factor binding, (*b*) pre-mRNA splicing, (*c*) mRNA secondary structure and stability, and (*d*) translational efficiency (i.e., quantity of transcripts), as well as the structure and stability of the protein products. Note that both synonymous and nonsynonymous SNPs can affect the amount, structure, and stability of the resulting gene product (RNA or protein). Also note that other types of variants, e.g., insertions and deletions, are not shown in this image but are responsible for at least as much impact. Abbreviations: mRNA, messenger RNA; pre-mRNA, precursor mRNA; SNP, single-nucleotide polymorphism; WT, wild-type sequence.

we use the term "SNP" without regard for frequency differences.] This price drop was due to advances in both sequencing techniques and analytical methods. According to the National Human Genome Research Institute (NHGRI), the cost of reagents and instrument time necessary for sequencing a complete genome is now around $1,000 (5), with some companies as of October 2019 performing whole-genome sequencing for $600 or even at no cost (Veritas and Nebula, respectively), and whole-exome sequencing can be performed even more cheaply (6). The technical feasibility of using patient genetic data in real clinical settings has thus made obvious the need for fast, accurate, and reliable analytical methods.

Among a multitude of HGP-related efforts, scientists have annotated genome components into ENCODE (Encyclopedia of DNA Elements; 7) and listed their common and not-so-common variants via the International HapMap Project (8), the 1000 Genomes Project (9), ExAC (Exome Aggregation Consortium; 10), and gnomAD (Genome Aggregation Database; 11). They have surveyed the structure and function of the encoded proteins via the Structural Genomics Initiative (12, 13); described gene expression across tissues and conditions via GTEx (Genotype Tissue Expression; 14), GEO (Gene Expression Omnibus; 15), and Allen Brain Atlas (16); and are exploring the genome's three-dimensional organization (17–20) via, e.g., the Roadmap Epigenomics Program (21). Current efforts aim to combine newly gained genomic knowledge with other advances to further understanding of basic biological mechanisms [e.g., the BRAIN (Brain Research through Advancing Innovative Neurotechnologies) Initiative (22)] and pursue better diagnostics and treatments [e.g., Cancer Moonshot (23)]. Progress in genetic counseling [e.g., *CFTR* gene analysis for assessing the incidence of cystic fibrosis among newborns (24) and a cost-effective web-based platform for genetic counseling (25)], diagnostics [e.g., rapid, automated diagnosis of monogenic diseases for newborns (26) and diagnosis of congenital anomalies from peripheral blood (27)], and precision medicine [e.g., predictions of *TP53* variant impact on response to chemotherapy (28), recommendations of medication dosage on the basis of genotype data (29), and genotype-based algorithmic warfarin dosing strategies (30)] are all contributing to improved disease outcomes and increased span and quality of life.

While work on interpreting the genome continues, more recent findings have highlighted the importance of variation in human microbiomes as well. The human microbiome, i.e., the community of microorganisms living in and on the human body, consists of roughly 10 times more cells than the human body (excluding red blood cells) (31) and at least 400 times more unique genes than the human genome (32). The microbiome metagenome, our so-called second genome, is thus a significant additional source of genetic variation, contributing to phenotypes and playing an important role in disease development, progression, and treatment possibilities (33). New treatment strategies involving the microbiome are already being used in the clinic. For example, fecal microbiome transplantation treats recurrent *Clostridium difficile* infection at >90% success rate (34, 35) and has been recommended for other diseases such as inflammatory bowel disease (IBD) and obesity (36). However, our understanding of what defines a healthy microbiome, or how microbiomes can be manipulated to improve health, remains limited.

Historically, and in large part due to the cost of sequencing, 16S ribosomal RNA (rRNA) gene surveys (amplicon sequencing) were used to assess the microbiome composition, i.e., its taxonomic makeup. However, as recent findings have shown, 16S rRNA sequence identity does not precisely identify microbial species (37, 38). Moreover, the microbiome molecular contributions to the functioning of the body are not easily inferred by answering the question "Who is there?" (39–42). One recent study, for example, found that the enrichment/depletion of genes across microbial strains of the same species was associated with host BMI (body mass index) and cholesterol level (43). Shotgun metagenome sequencing (i.e., sequencing all the DNA in a sample) has allowed for deeper exploration of the microbiome. Numerous tools have been developed for the analysis of such data, either by first using microbial genome assembly (44–47) or by making inferences directly from reads (41, 48–51). This new source of data now demands new, and vastly more efficient, methods for the joint analysis of human and microbial genetic data.

Recently there have been many efforts to interpret the human genome (variant association- and effect-based) and metagenome (amplicon- and shotgun-based) data in relation to disease. These efforts have borne major advances in diagnoses and precise stratification of certain diseases, as well as in treatment selection, such as in pharmacogenomics (52) and pharmacomicrobiomics (53). For example, one diagnostic platform that automatically analyzes electronic health record and genome sequencing data was able to successfully (with 97% recall and 99% precision) and rapidly (under 24 h) diagnose genetic diseases for severely ill children in intensive care units (54). A previous study had shown that a dietary intervention induced significant weight loss and concomitant structural changes of the gut microbiota in children affected by Prader-Willi syndrome and simple obesity (55). Our lab revisited these data and revealed further individual-specific responses to the dietary intervention (41). In spite of these advances, however, three limitations of the current state of the art are salient: (*a*) In most cases work remains in the realm of the research labs and is removed from clinical applications; (*b*) easily generalizable methods for performing these types of analyses are missing; and (*c*) there are, to the best of our knowledge, no methods that incorporate both genome and microbiome variation into a single predictive measure. Looking forward to the near future where data availability is no longer a limiting factor for method development, holistic and reproducible approaches that consider both the microbiome and genome factors to reach conclusions about disease are necessary to move science and clinical applications forward.

THE GENOME

Human Genome Variation Drives Functional Changes and Disease Development

Only about 0.1% of human DNA (about 3 million of 3 billion base pairs) is different between two randomly selected human genomes (56). This difference, however, accounts for population diversity, individuality, susceptibility to disease, etc. By definition, heritable diseases are solely due to genomic variation, but in reality the contributions of environmental factors, epigenomics, and other features of specific disease types vary. Some diseases are monogenic; for example, sickle cell anemia is caused by a homozygous SNP resulting in a valine to glutamic acid substitution in the hemoglobin beta-subunit. As of November 2019, there are 5,472 single-gene disorders and traits reported in the Online Mendelian Inheritance in Man (OMIM) (57) database, and we suspect that there are at least as many such rare diseases affecting only a small fraction of the population and thus not yet molecularly specified. Most other known diseases are polygenic and thus display a less clear genetic signal.

In an attempt to understand the genetic architecture of common heritable disease, many genome-wide association studies (GWAS) were carried out in the early years of genome exploration (58). GWAS aim to identify a set of common genomic variants that are associated with a specific phenotypic trait, such as a disease, in a given population. Using SNP arrays (59) (i.e., DNA microarrays used to identify specific SNPs in individual genomes), large-scale GWAS bypass the need to sequence genomes in their entirety, focusing instead on variants common in specific populations. Note that since SNP arrays require the explicit knowledge of the possible SNP at a given position, they are not able to identify new variants. While SNP arrays can be specifically designed to target any variant, they are usually limited to tagging common variants. GWAS take advantage of linkage disequilibrium to tag entire haplotypes with a much smaller set of these common genomic markers. For example, as few as 500,000 common SNPs are estimated to be sufficient to tag more than 10 million variants common to non-African populations (60).

The NHGRI-EBI (European Bioinformatics Institute) GWAS Catalog currently contains 5,687 curated GWAS comprising 71,673 statistically significant (p-value $< 5 \times 10^{-8}$) variant–trait associations from 3,567 studies. The variants identified in GWAS, however, are too common in the population to be causal for the observed traits, hampering the use of GWAS results for biologically meaningful conclusions or clinically relevant diagnoses. A workgroup of clinical laboratory directors and clinicians from the American College of Medical Genetics and Genomics (ACMG), the Association for Molecular Pathology, and the College of American Pathologists recommended guidelines (ACMG guidelines) for the interpretation of sequence variants. The guidelines recommend classifying variants using standardized terminology ("pathogenic," "likely pathogenic," "uncertain significance," "likely benign," and "benign") based on different types of variant evidence, such as population frequency, computational predictions, and functional annotations (61). Incidental findings of common variants (frequency annotated in population-wide databases such as gnomAD, ExAC, and dbSNP) do not, by these guidelines, indicate presence of disease but rather designate the variant as probably benign (61). The term "pathogenic" is not used even when a GWAS-based association with disease exists; rather, these variants are deemed risk alleles (61). In contrast, the frequency of variants in disease-specific databases, such as the Catalogue of Somatic Mutations in Cancer (COSMIC) (62), may indicate disease involvement.

In determining the cause and effect relationships between genetic variation and disease it is important to consider the pathogenicity mechanisms, i.e., variant-caused failures in the normal functioning of molecular pathways. Variants in noncoding regions of the genome may have an effect on overall genome structure, gene regulation, splicing, etc. Some noncoding variants directly mediate Mendelian disease (63), while others play a role in cancer development (64). Noncoding variants mainly affect functional changes by modifying gene expression via mechanisms such as changes to DNA accessibility (65), transcription factor binding (66), and histone modifications (67). A specific coding variant may lead to changes in mRNA stability or speed of translation, and thus protein quantity (68–70), altered protein structure or stability (71), posttranslational modifications (72), subcellular localization (73), ligand binding (74), interaction with other proteins (75), etc. Broadly, a variant may result in enhanced or depleted functionality of the gene that it affects—or produce no change to an assumed wild-type functionality at all. In humans specifically, diploidy also contributes to the complexity: Some genes are haplosufficient, meaning that one nonmutant copy of the gene is enough to carry on normal functioning, while others require the presence of both functional alleles (76). Furthermore, functionality of the nonmutant allele product (protein or RNA) may be additionally disrupted by the presence of a specific mutant allele of the same gene, e.g., via formation of inactive protein multimers (77) or competition for the same ligand (78). Finally, the specific combination of the altered gene functions may lead to disease (79, 80).

Computational Tools Predict SNP Effects, but Often Fail to Define "Effect"

To date, researchers have developed hundreds of computational tools to predict the functional effects of variants (SNPs, as well as structural and insertion/deletion variants). While some methods address effects of all SNPs [e.g., CADD (81), DANN (82), FATHMM-MKL (83), MutationTaster2 (84)], others are more focused on noncoding variants [e.g., GWAVA (85), LINSIGHT (86), ARVIN (87), SIFT Indel (88)] or synonymous variants [e.g., SilVA (89), regSNPs-splicing (90), DDIG-SN (91), IDSV (92)], and most available methods attempt to predict effects of nonsynonymous variants. In addition to the increasing need for appropriate benchmarking data (93), it is a challenge to define what exactly constitutes an effect for a given tool. Some tools aim to find cancer drivers [e.g., FATHMM-cancer (94), VEST (95), CScape (96)], while others look for function, structure, or stability changes [e.g., SNAP (97), PoPMuSiC (98), I-Mutant2.0 (99), I-Mutant3.0 (100)] or variant pathogenicity [e.g., PolyPhen-2 (101), PON-P (102), PON-P2 (103), REVEL (104)]. With the advent of deep mutational scanning (DMS) (105), tools have also been developed to recognize differences in the specific functionality defined by each experiment [e.g., Envision (106)], although their applications to new data may be limited (107). Notably, a method that predicts pathogenicity of variant combinations in gene pairs was recently published (108), suggesting an interesting future direction.

Interestingly, most existing and new tools that do not rely on DMS data fail to explicitly differentiate their target effects among the three basic overlapping but not identical classes: function change, fitness, and pathogenicity. The responsibility for figuring out which of the many methods to use for a particular set of predictions thus falls upon the largely unaware users. To choose correctly, it is important to understand the details of the method training/development data. For example, even something as seemingly well defined as recognizing polymorphisms versus disease variants requires a more in-depth analysis even before the prediction is made: What is a polymorphism? Is it a variant that has been definitively shown not to be disease associated or simply one with high frequency in a population? What variants are designated as disease? Are these only variants causing monogenic diseases or are these GWAS-significant variants associated with, but not causing, disease? Here, it is possible that the method actually differentiates variants by pathogenicity (disease versus no disease), by frequency in a population (e.g., monogenic disease culprits versus common variants), or by functional effect (severe effect versus mild or no effect). However, all of these classifications are equally likely if no detailed information about the development data is provided or discussed. Further complicating the distinction is the significant overlap between classes at the extremes: Observed variants that are lethal at an early age are almost always rare and obviously disadvantageous in terms of fitness. In contrast, common variants (e.g., >1% frequency in a population) may be neutral polymorphisms but also pathogenic in certain genomic contexts, or they may bear functional and phenotypic, but not disease, effects (109). In other words, while stated method goals may vary, their predictions often overlap in extreme cases, but not in intermediate ones (97, 107, 110). Thus, while recognizing method appropriateness for a particular prediction task should be straightforward in principle, in practice, the use boundaries are often vague for both tool builders and users. **Table 1** (with more detail provided in **Supplemental Table 1**) summarizes popular variant effect prediction tools along with their likely uses.

Supplemental Material >

Existing tools also differ in the biological features they use for predicting variant effect. Some, such as SIFT (111) and PROVEAN (112), rely solely on basic biological principles, such as biochemical amino acid similarities and evolutionary conservation information. Others, such as SNAP/SNAP2 (97, 113) and MutPred/MutPred2 (114, 115), use machine learning models trained on biochemical, biophysical, and evolutionary features of large SNP sets with experimentally verified effects. Notably, almost all tools excessively rely on the fact that amino acid substitutions at conserved residues more frequently have an effect than those at nonconserved positions (116). Thus, rather than predicting mutation effect, these tools highlight site conservation, where the threshold for what can be deemed conserved varies. In nature, there are some neutral mutations at conserved sites as well as plenty of moderately non-neutral mutations at nonconserved sites (110, 117). Shared features used by computational predictors, and particularly the use of conservation signal, also make the consensus approach less reliable than desired: If two methods predict the same variant to have an effect, that does not constitute a more reliable outcome if the methods are not independent.

Table 1 Properties of common variant effect prediction methods

Tool	Year	Model	Features	Scope[a]	Impacts[b]	Predicts[c]
IDSV (92)	2019	Random forest	SEQ	sSNP	Protein	Path.
DeFINE (192)	2018	Deep convolutional neural net/gradient boosting	SEQ	SNP	Regulatory	Effect, path.
Envision (106)	2018	Stochastic gradient boosting	SEQ, STR	nsSNP	Protein	Effect
ARVIN (87)	2018	Random forest	SEQ, NET	SNP	Regulatory	Path.
MutPred2 (115)	2017	Neural network	SEQ	nsSNP	Protein	Path.
LINSIGHT (86)	2017	Linear, probabilistic model	SEQ	SNP	Regulatory	Path.
DDIG-SN (91)	2017	Support vector machine	SEQ	sSNP	Protein	Path.
regSNPs-splicing (90)	2017	Random forest	SEQ	sSNP	Protein	Path.
CScape (96)	2017	Multiple kernel learning	SEQ	SNP	Both	Cancer
REVEL (104)	2016	Random forest	ENS	nsSNP	Protein	Path.
PANTHER-PSEP (193)	2016	Phylogenetic analysis	SEQ	nsSNP	Protein	Path.
DANN (82)	2015	Deep neural network	SEQ	SNP	Both	Effect, path.
FATHMM-MKL (83)	2015	Multiple kernel learning	SEQ, KB	SNP	Both	Path.
SNAP2 (113)	2015	Neural network	SEQ	nsSNP	Protein	Effect, path.
PON-P2 (103)	2015	Random forest	SEQ, STR	nsSNP	Protein	Path.
wKinMut2 (194)	2015	Annotation summary	ENS, KB	nsSNP	Protein	Path.
CADD (81)	2014	Support vector machine	SEQ	All	Both	Effect, path.
MutationTaster2 (84)	2014	Naïve Bayes classifier	SEQ, KB	All	Both	Path.
GWAVA (85)	2014	Random forest	SEQ, KB	SNP	Regulatory	Path.
PredictSNP (195)	2014	Consensus scoring	ENS	nsSNP	Protein	Effect, path.
FATHMM-DS (196)	2014	Hidden Markov models	SEQ	nsSNP	Protein	Path.
PolyPhen-2 (101)	2013	Naïve Bayes classifier	SEQ, STR	nsSNP	Protein	Effect
FATHMM (197)	2013	Hidden Markov models	SEQ	nsSNP	Protein	Path.
VEST (95)	2013	Random forest	SEQ	nsSNP	Protein	Path.
FATHMM-cancer (94)	2013	Hidden Markov models	SEQ	nsSNP	Protein	Cancer
Meta-SNP (198)	2013	Random forest	ENS	nsSNP	Protein	Path.
SilVA (89)	2013	Random forest	SEQ	sSNP	Protein	Path.
PROVEAN (112)	2012	Delta alignments scoring	SEQ	All	Both	Path.
SIFTIndel (88)	2012	Decision tree	SEQ, KB	InDel	Protein	Path.
PON-P (102)	2012	Random forest	ENS	nsSNP	Protein	Path.
KinMut (199)	2012	Support vector machine	SEQ	nsSNP	Protein	Path.
MutationAssessor (200)	2011	Functional impact scoring	SEQ, STR	nsSNP	Protein	Effect, path.
MutationTaster (208)	2010	Naïve Bayes classifier	SEQ, KB	All	Protein	Path.
MutPred (114)	2009	Random forest	SEQ	nsSNP	Protein	Path.
PoPMuSiC2.0 (201)	2009	Energy function	SEQ, STR, KB	nsSNP	Protein	Stability
I-Mutant3.0 (100)	2008	Support vector machine	SEQ, STR	nsSNP	Protein	Stability
SNAP (97)	2007	Neural network	SEQ	nsSNP	Protein	Effect
PhD-SNP (202)	2006	Support vector machine	SEQ	nsSNP	Protein	Path.
Align-GVGD (203)	2006	Extended Grantham difference scoring	SEQ, STR	nsSNP	Protein	Effect
FoldX (204)	2005	FoldX force field	STR	nsSNP	Protein	Stability

(*Continued*)

Table 1 ***(Continued)***

Tool	Year	Model	Features	Scope[a]	Impacts[b]	Predicts[c]
I-Mutant2.0 (99)	2005	Support vector machine	SEQ, STR	nsSNP	Protein	Stability
MAPP (205)	2005	Functional impact scoring	SEQ	nsSNP	Protein	Effect
nsSNPAnalyzer (206)	2005	Random forest	SEQ, STR	nsSNP	Protein	Path.
PolyPhen (207)	2002	Rule-based classifier	SEQ, STR	nsSNP	Protein	Effect
SIFT (111)	2001	PSSM-based probabilities	SEQ	nsSNP	Protein	Effect, path.
PoPMuSiC (98)	2000	Energy function	SEQ, STR, KB	nsSNP	Protein	Stability

Abbreviations: ENS, ensemble predictor using output of other predictors; InDel, insertion/deletion; KB, extracted from literature or a knowledge base; NET, extracted from a regulatory network; (n)sSNP, (non)synonymous SNP; path., pathogenicity; PSSM, position-specific scoring matrix; SEQ, sequence-derived; SNP, single-nucleotide polymorphism; STR, structure-derived.

[a] "SNP" means all SNPs and "all" means both SNPs and InDels.

[b] This column indicates whether the tool applies at the protein level, the regulatory level, or both.

[c] This column indicates whether the tool predicts protein structure/function effects (effect); pathogenicity, possibly including cancer (path.); protein stability; or cancer predisposition.

Predicting Disease Risk from Genome Data

Computational effect predictions cannot be directly interpreted as increasing disease risk. Although genetic diseases are usually caused by (combinations of) mutations with severe functional changes, the latter cannot guarantee the former. To identify genes whose altered functionality is responsible for increased risk of disease, studies often rely on the prior experimental/clinical knowledge, such as curated variants from databases such as OMIM (57), ClinVar (118), COSMIC, and HGMD (Human Gene Mutation Database; 119), and expand this knowledge to cover molecular pathways involved in pathogenesis via gene coexpression or protein–protein interaction network analysis (79, 120–124). Statistical analysis of GWAS results also highlights potential disease genes, but does so without evidence for functional changes in the latter (125).

Various data-driven methods have been developed to assess whole human genomes (as opposed to individual variants) to predict whether a person has (a high risk of developing) disease. For example, Wei et al. (126) extracted nearly 179,000 SNPs from a study of 50,000 Crohn's disease (CD) and ulcerative colitis (UC) cases and healthy controls from the International IBD Genetics Consortium's data (127) to build variant-based regression models for accurate association-based identification of CD and UC patients. The PROPS (probability pathway score) (128) method was developed to differentiate between CD and UC patients using variants that affect genes in KEGG pathways (129) and coincidentally identified metabolism-related pathways most discriminative between the two diseases. Our recently published AVA,Dx (analysis of variation for association with disease) support vector machine–based method uses vectors of gene functional changes, as predicted from individual exonic variation, to further predict individual CD status (80). Our method thus identified dozens of previously unreported CD genes by tracing differentially functionally altered genes in diseased patients versus healthy controls. While these human genome–based methods have produced exciting results, adding the human microbiome into the picture may fill in the missing pieces toward the holy grail of precision medicine.

THE MICROBIOME

Taxonomic Annotations Reveal Composition of the Microbial Communities

Shortly after the human genome had been sequenced, two major projects were launched: the European project Metagenomics of the Human Intestinal Tract (130) and the Human

Microbiome Project (HMP) (131), funded by the National Institutes of Health. A major question in microbiome analysis had initially been, "Who are they?", that is, "What is the taxonomic composition, i.e., the list and abundance profiles of member organisms, of the microbial community?" Microbiome composition is often assessed by sequencing 16S rRNA, followed by comparison to reference databases, such as the Ribosomal Database Project (RDP) (132), SILVA (133), and GreenGenes (134). The most widely used computational pipelines for this type of analyses are QIIME (Quantitative Insights into Microbial Ecology) (135) and mothur (136). Benchmark analyses suggest that these tools generate results of comparable accuracy, but QIIME is significantly faster (137). Notably the 2018 QIIME 2 update, which uses a naïve Bayes classifier (138), demonstrated further improved performance, albeit at the cost of increased memory use and CPU time (139).

While 16S rRNA sequencing has been historically widely used, it suffers from limited resolution/precision at lower taxonomic levels (140) and significant annotation disagreements across different reference databases (141). Shotgun whole-metagenome sequencing, although significantly more expensive, targets all genes in the microbiome rather than just the 16S rRNA gene. With shotgun metagenomic data, taxonomic assignment can be done by using either signature genes only [e.g., MetaPhlAn2 (142), mOTUs2 (143)] or all genes [e.g., Centrifuge (144), Bracken (145), Kraken 2 (146), Kaiju (147), CLARK (148)]. While these methods are limited by the lack of complete microbial reference genomes, and thus not as useful for taxonomically placing novel organisms as 16S rRNA–based methods, they offer higher resolution than 16S rRNA analyses. For example, MetaPhlAn2 can accurately assign taxonomy all the way down to the strain level for relatively well-studied microbiome niches. In a recent benchmark study with a variety of test datasets, the all genes methods demonstrated better performance than the signature genes methods, mainly due to the more comprehensive reference databases (149). Recent large-scale efforts to explore the organismal composition of the human gut microbiome human have augmented the reference databases by reconstructing 2,058 (150), 1,952 (151) and 4,930 (152) new/yet-uncultured bacterial species. These results indicate that the human microbiome is far from completely explored.

Functional Annotation of the Microbiome Is Necessary but Difficult

As compared to the question "Who are they?", an arguably more compelling question in microbiome analysis is "What do they do?", that is, "What is the totality of molecular-level activities such as catalysis or binding being carried out by the members of the microbial community?" Here it is important to remember that although functional abilities can be inferred from taxonomic assignments, even taxonomic neighbors can have substantially different functions due to horizontal gene transfer (HGT) (38, 40). Notably, HGT is more frequent in human-associated bacteria than in those from other environments (42). It is estimated that more than half of total genes in human-associated bacterial genomes were obtained via HGT (42). For example, the rapid spread of antibiotic resistance genes via HGT has caused a global medical crisis of multidrug-resistant pathogens (153). Thus, identifying who is present in a particular microbiome, even if possible at a high level of precision, may not be as useful as figuring out what the microbiome is doing as a whole.

In a workflow of metagenome functional annotation, DNA sequences (either reads or assembled contigs) are first subjected to gene finding (154) or simple six-frame translation to predict corresponding peptide sequences, which are then mapped to reference sequence databases. A benchmark analysis using artificial metagenome datasets suggested that assemblers using multiple *k*-mers outperformed single-*k*-mer assemblers (155). However, for complex and highly diverse microbiome samples, assembly is computationally expensive and often plagued by chimeras and

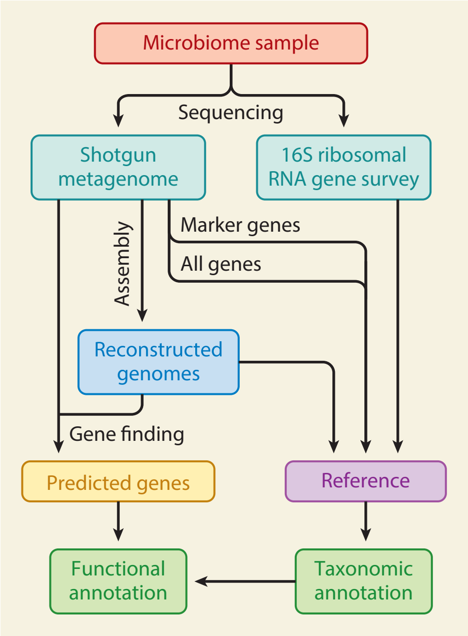

Figure 2

Flowchart of microbiome analysis. Taxonomic annotation can reveal the microbiome composition ("Who are they?"), while functional annotations reveal the molecular functionality that the community members carry out ("What do they do?").

a large fraction of unassembled reads from minor community members (155). Read-based workflows, in contrast, bypass the assembly step and the associated errors, but their annotation is often hampered by short/unreliable alignments (156). Both read- and assembly-based annotation inaccuracies are additionally compounded by the errors in functional annotations of most genes in the reference databases (157).

Various tools, such as MG-RAST (48), HUMAnN/HUMAnN2 (51, 158), ShotMAP (49), and Fun4Me (50), annotate metagenome functions by directly mapping reads to reference sequence databases, such as SEED (159), KEGG (129), MetaCyc (160), and UniRef (161). These methods aim to identify the specific microbial genes present in the metagenome. Our recently developed mi-faser (41) method/database combination was optimized to extract correct functional (as opposed to gene sequence-specific) annotations using a manually curated collection of experimentally verified protein molecular functions. Carnelian (162) followed soon after, using *k*-mer analysis to map to reads to the mi-faser database. Workflows are often database centered, complicating the conversion between annotations for method comparison. For example, MG-RAST uses SEED data as reference, while HUMAnN2 relies on UniRef50; HMP data were mapped by HUMAnN to KEGG pathways. A summary of the microbiome annotation flow can be found in **Figure 2**.

Microbiome Impacts Human Health

It is increasingly accepted that the human microbiome plays a critical role in host health. The gut is by far the most microbially populous niche of the human body (31), harboring different microbial populations across the intestinal microniches (163), from the gut lumen to the intestinal wall mucous layer. The human gut microbiome is critical for human development (164) and has been associated with a variety of diseases, including metabolic disorders such as obesity (165) and type 2 diabetes (166), autoimmune diseases such as inflammatory bowel diseases (167), and mental disorders such as autism (168). Taxonomic surveys of the gastrointestinal microbiome of CD patients have revealed microbial community features that are unique to CD patients, such as loss of microbial diversity (169) and depletion/enrichment of certain bacterial taxa (170). Establishing whether these community shifts contribute to pathogenesis, simply correlate with disease, or result from it requires understanding not only which microbes are involved but also what they do. Studies indicate that in CD, the microbiome molecular functionality is more consistently disturbed than the taxonomic makeup (171). Analysis of CD occurrence in a single family had similarly shown microbial functional differences across patients, as well as between patients and their healthy relatives (41). In type 2 diabetes, dietary-fiber-promoted gut bacteria have been shown to alleviate the symptoms of the disease (164). The steady increase in interest in microbiome shifts associated with a wide range of diseases, from gastrointestinal to neurological, thus suggests the need for exploring joint contributions of the human genome and microbiome to disease.

HUMAN GENOME AND MICROBIOME INTERACTION

Current Knowledge of Healthy Genome–Microbiome Interactions Is Limited

Human genome variation is known to impact the course and severity of infectious disease. As with the sickle cell example described above, individuals heterozygous for the hemoglobin mutation display strongly reduced plasmodium reproduction rates upon infection and thus significantly reduced malaria risks. Incidentally, they also do not suffer from the full range of adverse effects of sickle cell anemia, promoting positive selection for the mutation in malaria-affected regions of the world (172). Associations between human genetic variation and increased susceptibility to infectious diseases such as tuberculosis (173) and leprosy (174) have recently been identified.

It is thus expected that human genome variation would similarly impact the composition of the human-associated microbiome. Microbiome GWAS (mGWAS; not to be confused with the unrelated metabolome GWAS) connect variation across human genomes to microbiome descriptors, such as alpha diversity (the number of species in a microbiome; 175) and beta diversity [pairwise distance, such as Bray–Curtis taxonomic dissimilarity (176), between microbiomes (175)], as well as to the abundance of certain microbial taxa or functions. To date, several mGWAS have been carried out in healthy cohorts to identify hundreds of significant associations, yet only one association, between *Bifidobacterium* and variants in the lactase gene *LCT*, has been validated across different studies and cohorts (177–179). Further increasing the inconsistency in mGWAS, a recent study on 1,046 healthy adults identified no significant associations between host genetics and the microbiome (180). The study's results suggest that the transient environment, as opposed to the genetically defined stable determinants, is the dominant factor in determining microbiome composition: Genetically unrelated individuals who share a household have similar microbiomes, while relatives who have never lived together may differ microbiome-wise.

The inconsistency in mGWAS results can be due to several factors including technical differences (batch effects) and study design differences (host genetics and prior-knowledge-based variant filtering used to increase the statistical power of the study). More specifically, microbiomes contain hundreds of taxa and thousands of encoded functions, which requires stringent multiple testing correction to validate the significance of findings. Researchers thus either select, somewhat arbitrarily, only very common SNPs (181) or limit their studies to candidate genes/SNPs based on prior knowledge (178, 179). For example, although both Bonder et al. (179) and Goodrich et al. (177) aimed to collect a descriptive set of SNPs that guide gut microbiome composition, including involvement in complex diseases, immune traits, metabolic traits, food metabolism, and food preferences, the number of SNPs collected in the former study was twice as high as that in the latter study (76,444 versus 32,378). Furthermore, SNPs evaluated in these two studies were selected on the basis of previous GWAS results and thus may have been subject to the limitations of those GWAS.

Although published mGWAS have not yet been applied to cohorts of sick individuals, studies have reported that variants known to carry higher risks of IBD (for example, affecting *NOD2*, *CARD9*, *ATG16L1*, *IRGM*, and *FUT2* genes) alter the gut microbiome composition in healthy individuals (182).

Exploration of Genome–Microbiome Interactions in Disease Is Only Now Taking Off

The second phase of HMP, iHMP (integrative HMP), carried out both host whole-genome sequencing and microbiome shotgun metagenome sequencing (as well as meta-transcriptomes, meta-metabolomes, etc.) of its participants in three longitudinal cohort studies of pregnancy and preterm birth (vaginal microbiomes of pregnant women), IBD (gut microbiomes), and prediabetes (gut and nasal microbiomes) (183). The data were recently published and made publicly available (184), offering researchers a unique chance to investigate these medical conditions in a combined perspective of both human genome and microbiome. Machine learning models, for example, with additional microbiome information have the potential to improve the prediction precision to a level that can be applied in clinical settings. Technical challenges to the development of such models, however, include the drastically increased feature space [there are over 1,000 bacterial species that could normally live in the human gut, although any one individual may have any combination of these (131)] and significant heterogeneity of input features in terms of type, scale, sparsity, and weight. Advances in deep learning techniques, a class of machine learning algorithms well suited to processing high-dimensional data, provide new means for this type of analysis (185). Deep learning artificial neutral networks can extract features of increasing abstraction progressively via an architecture of consecutive convolution layers. As such, they can be used to effectively encode multidimensional data mapping to the observed signal. Other implementations such as autoencoders [unsupervised artificial neural networks used to learn efficient data encoding (186)] allow researchers to first compress the input dimensionality and train the network in a lower-dimensional space. Since training these networks requires a large training dataset and significant computational resources, deep learning has only very recently become a viable analytical approach. Given the amount of now available and consistently generated genome/metagenome data, deep learning models provide promising a way forward for extracting new insights.

The Future of Pharmacogenomics and -Microbiomics

After millions of years of coevolution, human metabolism has become an amalgamation of both host and microbial attributes (187). Evidence for this abounds; for example, one metabolomics study in germ-free mice illustrated that the gastrointestinal microbiome generates at least 10% of all detectable metabolites in the host serum (188). Specifically, the queuine micronutrient, which is necessary for posttranslational modification of transfer RNAs in all eukaryotes, including humans, can only be produced by bacteria (189). We suspect that disruption of these interactions also drives disease; for example, CD development has been shown to entail both genome-encoded (190) and microbially driven (41) immune system activity. The connection between the genome and the microbiome suggests that the results of pharmacomicrobiomics (53) studies focusing on gut bacterial drug metabolism as related to efficiency and toxicity (191) are likely also picked up

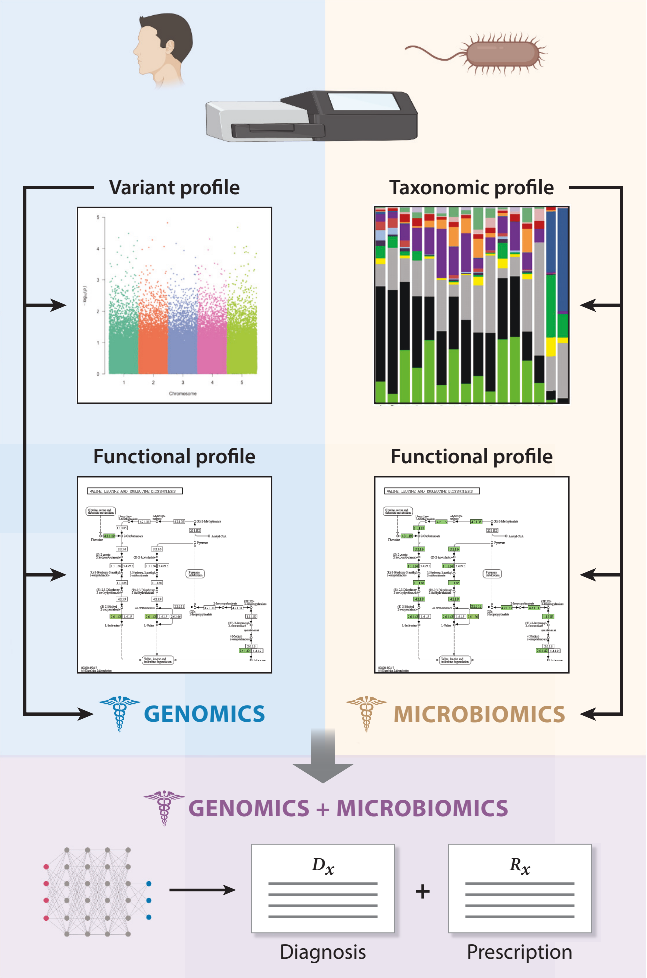

Figure 3

Integration of human genome and microbiome data may improve clinical diagnosis and treatment. In recent years, developments in pharmacogenomics and pharmacomicrobiomics have provided a platform for future joint explorations, e.g., using advances in deep learning. The ability to functionally profile the human genome and microbiome significantly contributes to such efforts, transforming them from statistical analyses to possible cause assessments.

in pharmacogenomics assessments. Moreover, studies that explicitly integrate human genome and microbiome data by looking for human–microbe joint pathways will likely reveal disease mechanisms that have been hidden from one-sided investigations. It seems that the future of personalized medicine lies at the interface between the human genome and microbiome (**Figure 3**). Integrating existing tools and building novel methods to meet the needs of this new type of analysis are thus two of the main challenges that the computational biologists will face in the near future.

DISCLOSURE STATEMENT

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

ACKNOWLEDGMENTS

We would like to thank Dr. Adrienne Hoarfrost, Dr. Loredana Quadro, Dr. Jay Tischfield (all Rutgers), Dr. Burkhard Rost, Christian Dallago (both TU Munich), Dr. Emidio Capriotti, Dr. Rita Casadio (both University of Bologna), Dr. Predrag Radivojac (Northeastern), Dr. Hannah Carter (UCSD), and Dr. Olivier Lichtarge (Baylor College of Medicine) for discussions of the topics covered here. We are grateful to Dr. Russ Altman (Stanford) and Dr. Maricel Kann (University of Maryland) for the invitation to contribute our thoughts to this collection of reviews. Finally, we would also like to express gratitude to all researchers working in microbiome and genome analysis fields and to the people who have made their genomic, microbiome, and medical information available to contribute to a better understanding health and disease. Y.B., M.M., Y.W., and Z.Z. were supported by the NIH (National Institutes of Health) grant U01 GM115486; Y.B., Y.M., and C.Z. were supported by the NSF (National Science Foundation) CAREER award 1553289; and Y.B. was supported by the NIH grant R01 MH115958.

LITERATURE CITED

1. Kruger J, Dunning D. 1999. Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *J. Personal. Soc. Psychol.* 77:1121–34
2. NHGRI (Natl. Human Genome Res. Inst.). 2019. *Human Genome Project FAQ*. Fact Sheet, NHGRI, Bethesda, MD. **https://www.genome.gov/human-genome-project/Completion-FAQ**
3. Levy S, Sutton G, Ng PC, Feuk L, Halpern AL, et al. 2007. The diploid genome sequence of an individual human. *PLOS Biol.* 5:e254
4. Wheeler DA, Srinivasan M, Egholm M, Shen Y, Chen L, et al. 2008. The complete genome of an individual by massively parallel DNA sequencing. *Nature* 452:872–76
5. NHGRI (Natl. Human Genome Res. Inst.). 2019. *The cost of sequencing a human genome*. Fact Sheet, NHGRI, Bethesda, MD. **https://www.genome.gov/about-genomics/fact-sheets/Sequencing-Human-Genome-cost**
6. Schwarze K, Buchanan J, Taylor JC, Wordsworth S. 2018. Are whole-exome and whole-genome sequencing approaches cost-effective? A systematic review of the literature. *Genet. Med.* 20:1122–30
7. ENCODE Proj. Consort. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* 489:57–74
8. Gibbs RA, Belmont JW, Hardenbol P, Willis TD, Yu F, et al. 2003. The International HapMap Project. *Nature* 426:789–96
9. 1000 Genomes Proj. Consort. 2015. A global reference for human genetic variation. *Nature* 526:68–74
10. Karczewski KJ, Weisburd B, Thomas B, Solomonson M, Ruderfer DM, et al. 2017. The ExAC browser: displaying reference data information from over 60 000 exomes. *Nucleic Acids Res.* 45:D840–45
11. Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, et al. 2019. Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes. bioRxiv 531210. **https://doi.org/10.1101/531210**
12. Burley SK, Almo SC, Bonanno JB, Capel M, Chance MR, et al. 1999. Structural genomics: beyond the human genome project. *Nat. Genet.* 23:151–57
13. Stevens RC, Yokoyama S, Wilson IA. 2001. Global efforts in structural genomics. *Science* 294:89–92
14. GTEx Consort. 2013. The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.* 45:580–85
15. Clough E, Barrett T. 2016. The Gene Expression Omnibus database. *Methods Mol. Biol.* 1418:93–110
16. Sunkin SM, Ng L, Lau C, Dolbeare T, Gilbert TL, et al. 2013. Allen Brain Atlas: an integrated spatiotemporal portal for exploring the central nervous system. *Nucleic Acids Res.* 41:D996–1008
17. Marchal C, Sima J, Gilbert DM. 2019. Control of DNA replication timing in the 3D genome. *Nat. Rev. Mol. Cell Biol.* 20:721–37
18. Zheng H, Xie W. 2019. The role of 3D genome organization in development and cell differentiation. *Nat. Rev. Mol. Cell Biol.* 20:535–50

19. van Steensel B, Furlong EEM. 2019. The role of transcription in shaping the spatial organization of the genome. *Nat. Rev. Mol. Cell Biol.* 20:327–37
20. Trieu T, Oluwadare O, Cheng J. 2019. Hierarchical reconstruction of high-resolution 3D models of large chromosomes. *Sci. Rep.* 9:4971
21. Satterlee JS, Chadwick LH, Tyson FL, McAllister K, Beaver J, et al. 2019. The NIH Common Fund/Roadmap Epigenomics Program: successes of a comprehensive consortium. *Sci. Adv.* 5:eaaw6507
22. Mott MC, Gordon JA, Koroshetz WJ. 2018. The NIH BRAIN Initiative: advancing neurotechnologies, integrating disciplines. *PLOS Biol.* 16:e3000066
23. Barlas S. 2016. The White House launches a cancer moonshot: Despite funding questions, the progress appears promising. *P&T* 41:290–95
24. Audrezet MP, Munck A, Scotet V, Claustres M, Roussey M, et al. 2015. Comprehensive *CFTR* gene analysis of the French cystic fibrosis screened newborn cohort: implications for diagnosis, genetic counseling, and mutation-specific therapy. *Genet. Med.* 17:108–16
25. Biesecker BB, Lewis KL, Umstead KL, Johnston JJ, Turbitt E, et al. 2018. Web platform versus in-person genetic counselor for return of carrier results from exome sequencing: a randomized clinical trial. *JAMA Intern. Med.* 178:338–46
26. Saunders CJ, Miller NA, Soden SE, Dinwiddie DL, Noll A, et al. 2012. Rapid whole-genome sequencing for genetic disease diagnosis in neonatal intensive care units. *Sci. Transl. Med.* 4:154ra35
27. Aref-Eshghi E, Bend EG, Colaiacovo S, Caudle M, Chakrabarti R, et al. 2019. Diagnostic utility of genome-wide DNA methylation testing in genetically unsolved individuals with suspected hereditary conditions. *Am. J. Hum. Genet.* 104:685–700
28. Osman AA, Neskey DM, Katsonis P, Patel AA, Ward AM, et al. 2015. Evolutionary action score of *TP53* coding variants is predictive of platinum response in head and neck cancer patients. *Cancer Res.* 75:1205–15
29. Reisberg S, Krebs K, Lepamets M, Kals M, Magi R, et al. 2019. Translating genotype data of 44,000 biobank participants into clinical pharmacogenetic recommendations: challenges and solutions. *Genet. Med.* 21:1345–54
30. Syn NL, Wong AL-A, Lee S-C, Teoh H-L, Yip JWL, et al. 2018. Genotype-guided versus traditional clinical dosing of warfarin in patients of Asian ancestry: a randomized controlled trial. *BMC Med.* 16:104
31. Sender R, Fuchs S, Milo R. 2016. Revised estimates for the number of human and bacteria cells in the body. *PLOS Biol.* 14:e1002533
32. Markowitz VM, Chen I-MA, Palaniappan K, Chu K, Szeto E, et al. 2014. IMG 4 version of the integrated microbial genomes comparative analysis system. *Nucleic Acids Res.* 42:D560–67
33. Grice EA, Segre JA. 2012. The human microbiome: our second genome. *Annu. Rev. Genom. Hum. Genet.* 13:151–70
34. van Nood E, Vrieze A, Nieuwdorp M, Fuentes S, Zoetendal EG, et al. 2013. Duodenal infusion of donor feces for recurrent *Clostridium difficile*. *New Engl. J. Med.* 368:407–15
35. Kelly CR, Khoruts A, Staley C, Sadowsky MJ, Abd M, et al. 2016. Effect of fecal microbiota transplantation on recurrence in multiply recurrent *Clostridium difficile* infection: a randomized trial. *Ann. Intern. Med.* 165:609–16
36. Choi HH, Cho Y-S. 2016. Fecal microbiota transplantation: current applications, effectiveness, and future perspectives. *Clin. Endosc.* 49:257–65
37. Fox GE, Wisotzkey JD, Jurtshuk P Jr. 1992. How close is close: 16S rRNA sequence identity may not be sufficient to guarantee species identity. *Int. J. Syst. Bacteriol.* 42:166–70
38. Zhu C, Delmont TO, Vogel TM, Bromberg Y. 2015. Functional basis of microorganism classification. *PLOS Comput. Biol.* 11:e1004472
39. Smillie CS, Smith MB, Friedman J, Cordero OX, David LA, Alm EJ. 2011. Ecology drives a global network of gene exchange connecting the human microbiome. *Nature* 480:241–44
40. Zhu C, Mahlich Y, Miller M, Bromberg Y. 2018. fusionDB: assessing microbial diversity and environmental preferences via functional similarity networks. *Nucleic Acids Res.* 46:D535–41
41. Zhu C, Miller M, Marpaka S, Vaysberg P, Rühlemann MC, et al. 2018. Functional sequencing read annotation for high precision microbiome analysis. *Nucleic Acids Res.* 46(4):e23

42. Jeong H, Arif B, Caetano-Anollés G, Kim KM, Nasir A. 2019. Horizontal gene transfer in human-associated microorganisms inferred by phylogenetic reconstruction and reconciliation. *Sci. Rep.* 9:5953

43. Zeevi D, Korem T, Godneva A, Bar N, Kurilshikov A, et al. 2019. Structural variation in the gut microbiome associates with host health. *Nature* 568:43–48

44. Peng Y, Leung HC, Yiu SM, Chin FY. 2012. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics* 28:1420–28

45. Luo R, Liu B, Xie Y, Li Z, Huang W, et al. 2012. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *Gigascience* 1:18

46. Li D, Liu CM, Luo R, Sadakane K, Lam TW. 2015. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* 31:1674–76

47. Nurk S, Meleshko D, Korobeynikov A, Pevzner PA. 2017. metaSPAdes: a new versatile metagenomic assembler. *Genome Res.* 27:824–34

48. Meyer F, Paarmann D, D'Souza M, Olson R, Glass E, et al. 2008. The metagenomics RAST server—a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinform.* 9:386

49. Nayfach S, Bradley PH, Wyman SK, Laurent TJ, Williams A, et al. 2015. Automated and accurate estimation of gene family abundance from shotgun metagenomes. *PLOS Comput. Biol.* 11:e1004573

50. Sharifi F, Ye Y. 2017. From gene annotation to function prediction for metagenomics. *Methods Mol. Biol.* 1611:27–34

51. Franzosa EA, McIver LJ, Rahnavard G, Thompson LR, Schirmer M, et al. 2018. Species-level functional profiling of metagenomes and metatranscriptomes. *Nat. Methods* 15:962–68

52. Lavertu A, McInnes G, Daneshjou R, Whirl-Carrillo M, Klein TE, Altman RB. 2018. Pharmacogenomics and big genomic data: from lab to clinic and back again. *Hum. Mol. Genet.* 27:R72–78

53. Rizkallah MR, Saad R, Aziz RK. 2010. The Human Microbiome Project, personalized medicine and the birth of pharmacomicrobiomics. *Curr. Pharmacogenom. Personal. Med.* 8:182–93

54. Clark MM, Hildreth A, Batalov S, Ding Y, Chowdhury S, et al. 2019. Diagnosis of genetic diseases in seriously ill children by rapid whole-genome sequencing and automated phenotyping and interpretation. *Sci. Transl. Med.* 11:eaat6177

55. Zhang C, Yin A, Li H, Wang R, Wu G, et al. 2015. Dietary modulation of gut microbiota contributes to alleviation of both genetic and simple obesity in children. *eBioMedicine* 2:968–84

56. Zhao Z, Jin L, Fu YX, Ramsay M, Jenkins T, et al. 2000. Worldwide DNA sequence variation in a 10-kilobase noncoding region on human chromosome 22. *PNAS* 97:11354–58

57. Hamosh A, Scott AF, Amberger JS, Bocchini CA, McKusick VA. 2005. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.* 33:D514–17

58. Reich DE, Lander ES. 2001. On the allelic spectrum of human disease. *Trends Genet.* 17:502–10

59. LaFramboise T. 2009. Single nucleotide polymorphism arrays: a decade of biological, computational and technological advances. *Nucleic Acids Res.* 37:4181–93

60. Int. HapMap Consort. 2005. A haplotype map of the human genome. *Nature* 437:1299–320

61. Richards S, Aziz N, Bale S, Bick D, Das S, et al. 2015. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet. Med.* 17:405–24

62. Tate JG, Bamford S, Jubb HC, Sondka Z, Beare DM, et al. 2019. COSMIC: the catalogue of somatic mutations in cancer. *Nucleic Acids Res.* 47:D941–47

63. Smedley D, Schubach M, Jacobsen JOB, Köhler S, Zemojtel T, et al. 2016. A whole-genome analysis framework for effective identification of pathogenic regulatory variants in Mendelian disease. *Am. J. Hum. Genet.* 99:595–606

64. Horn S, Figl A, Rachakonda PS, Fischer C, Sucker A, et al. 2013. TERT promoter mutations in familial and sporadic melanoma. *Science* 339:959–61

65. Zhou J, Troyanskaya OG. 2015. Predicting effects of noncoding variants with deep learning–based sequence model. *Nat. Methods* 12:931–34

66. Zhao J, Li D, Seo J, Allen AS, Gordân R. 2017. Quantifying the impact of non-coding variants on transcription factor-DNA binding. In *Proceedings of the 21st Annual International Conference on Research in Computational Molecular Biology (RECOMB 2017)*, ed. SC Sahinalp, pp. 336–52. Cham, Switz.: Springer

67. Hoffman GE, Bendl J, Girdhar K, Schadt EE, Roussos P. 2019. Functional interpretation of genetic variants using deep learning predicts impact on chromatin accessibility and histone modification. *Nucleic Acids Res*. 47(20):10597–611

68. Edwards NC, Hing ZA, Perry A, Blaisdell A, Kopelman DB, et al. 2012. Characterization of coding synonymous and non-synonymous variants in *ADAMTS13* using *ex vivo* and *in silico* approaches. *PLOS ONE* 7:e38864

69. Hunt RC, Simhadri VL, Iandoli M, Sauna ZE, Kimchi-Sarfaty C. 2014. Exposing synonymous mutations. *Trends Genet*. 30:308–21

70. Robert F, Pelletier J. 2018. Exploring the impact of single-nucleotide polymorphisms on translation. *Front. Genet.* 9:507

71. Pechmann S, Frydman J. 2013. Evolutionary conservation of codon optimality reveals hidden signatures of cotranslational folding. *Nat. Struct. Mol. Biol.* 20:237–43

72. Cartegni L, Krainer AR. 2002. Disruption of an SF2/ASF-dependent exonic splicing enhancer in *SMN2* causes spinal muscular atrophy in the absence of *SMN1*. *Nat. Genet*. 30:377–84

73. Nishibori Y, Liu L, Hosoyamada M, Endou H, Kudo A, et al. 2004. Disease-causing missense mutations in *NPHS2* gene alter normal nephrin trafficking to the plasma membrane. *Kidney Int*. 66:1755–65

74. Pires DE, Blundell TL, Ascher DB. 2015. Platinum: a database of experimentally measured effects of mutations on structurally defined protein-ligand complexes. *Nucleic Acids Res*. 43:D387–91

75. Wang Z, Burge CB. 2008. Splicing regulation: from a parts list of regulatory elements to an integrated splicing code. *RNA* 14:802–13

76. Seidman J, Seidman C. 2002. Transcription factor haploinsufficiency: when half a loaf is not enough. *J. Clin. Investig.* 109:451–55

77. Brachmann RK. 2004. p53 mutants: the Achilles' heel of human cancers? *Cell Cycle* 3:1030–34

78. Gong W, Chavez S, Beato M. 1997. Point mutation in the ligand-binding domain of the progesterone receptor generates a transdominant negative phenotype. *Mol. Endocrinol.* 11:1476–85

79. Margolin AA, Nemenman I, Basso K, Wiggins C, Stolovitzky G, et al. 2006. ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinform*. 7:S7

80. Wang Y, Miller M, Astrakhan Y, Petersen B-S, Schreiber S, et al. 2019. Identifying Crohn's disease signal from variome analysis. *Genome Med*. 11:59

81. Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. 2014. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.* 46:310–15

82. Quang D, Chen Y, Xie X. 2015. DANN: a deep learning approach for annotating the pathogenicity of genetic variants. *Bioinformatics* 31:761–63

83. Shihab HA, Rogers MF, Gough J, Mort M, Cooper DN, et al. 2015. An integrative approach to predicting the functional effects of non-coding and coding sequence variation. *Bioinformatics* 31:1536–43

84. Schwarz JM, Cooper DN, Schuelke M, Seelow D. 2014. MutationTaster2: mutation prediction for the deep-sequencing age. *Nat. Methods* 11:361–62

85. Ritchie GRS, Dunham I, Zeggini E, Flicek P. 2014. Functional annotation of noncoding sequence variants. *Nat. Methods* 11:294–96

86. Huang Y-F, Gulko B, Siepel A. 2017. Fast, scalable prediction of deleterious noncoding variants from functional and population genomic data. *Nat. Genet.* 49:618–24

87. Gao L, Uzun Y, Gao P, He B, Ma X, et al. 2018. Identifying noncoding risk variants using disease-relevant gene regulatory networks. *Nat. Commun.* 9:702

88. Hu J, Ng PC. 2012. Predicting the effects of frameshifting indels. *Genome Biol*. 13:R9

89. Buske OJ, Manickaraj A, Mital S, Ray PN, Brudno M. 2013. Identification of deleterious synonymous variants in human genomes. *Bioinformatics* 29:1843–50

90. Zhang X, Li M, Lin H, Rao X, Feng W, et al. 2017. regSNPs-splicing: a tool for prioritizing synonymous single-nucleotide substitution. *Hum. Genet.* 136:1279–89

91. Livingstone M, Folkman L, Yang Y, Zhang P, Mort M, et al. 2017. Investigating DNA-, RNA-, and protein-based features as a means to discriminate pathogenic synonymous variants. *Hum. Mutat.* 38:1336–47

92. Shi F, Yao Y, Bin Y, Zheng C-H, Xia J. 2019. Computational identification of deleterious synonymous variants in human genomes using a feature-based approach. *BMC Med. Genom.* 12:12

93. Peters B, Brenner SE, Wang E, Slonim D, Kann MG. 2018. Putting benchmarks in their rightful place: the heart of computational biology. *PLOS Comput. Biol.* 14:e1006494
94. Shihab HA, Gough J, Cooper DN, Day INM, Gaunt TR. 2013. Predicting the functional consequences of cancer-associated amino acid substitutions. *Bioinformatics* 29:1504–10
95. Carter H, Douville C, Stenson PD, Cooper DN, Karchin R. 2013. Identifying Mendelian disease genes with the Variant Effect Scoring Tool. *BMC Genom*. 14:S3
96. Rogers MF, Shihab HA, Gaunt TR, Campbell C. 2017. CScape: a tool for predicting oncogenic single-point mutations in the cancer genome. *Sci. Rep.* 7:11597
97. Bromberg Y, Rost B. 2007. SNAP: predict effect of non-synonymous polymorphisms on function. *Nucleic Acids Res*. 35:3823–35
98. Gilis D, Rooman M. 2000. PoPMuSiC, an algorithm for predicting protein mutant stability changes: application to prion proteins. *Protein Eng*. 13:849–56
99. Capriotti E, Fariselli P, Casadio R. 2005. I-Mutant2.0: predicting stability changes upon mutation from the protein sequence or structure. *Nucleic Acids Res*. 33:W306–10
100. Capriotti E, Fariselli P, Rossi I, Casadio R. 2008. A three-state prediction of single point mutations on protein stability changes. *BMC Bioinform*. 9:S6
101. Adzhubei I, Jordan DM, Sunyaev SR. 2013. Predicting functional effect of human missense mutations using PolyPhen-2. *Curr. Protoc. Hum. Genet.* 72:7.20.1–7.20.41
102. Olatubosun A, Väliaho J, Härkönen J, Thusberg J, Vihinen M. 2012. PON-P: integrated predictor for pathogenicity of missense variants. *Hum. Mutat.* 33:1166–74
103. Niroula A, Urolagin S, Vihinen M. 2015. PON-P2: prediction method for fast and reliable identification of harmful variants. *PLOS ONE* 10:e0117380
104. Ioannidis NM, Rothstein JH, Pejaver V, Middha S, McDonnell SK, et al. 2016. REVEL: an ensemble method for predicting the pathogenicity of rare missense variants. *Am. J. Hum. Genet.* 99:877–85
105. Fowler DM, Fields S. 2014. Deep mutational scanning: a new style of protein science. *Nat. Methods* 11:801–7
106. Gray VE, Hause RJ, Luebeck J, Shendure J, Fowler DM. 2018. Quantitative missense variant effect prediction using large-scale mutagenesis data. *Cell Syst*. 6:116–24.e3
107. Pejaver V, Babbi G, Casadio R, Folkman L, Katsonis P, et al. 2019. Assessment of methods for predicting the effects of PTEN and TPMT protein variants. *Hum. Mutat.* 40:1495–506
108. Papadimitriou S, Gazzo A, Versbraegen N, Nachtegael C, Aerts J, et al. 2019. Predicting disease-causing variant combinations. *PNAS* 116:11878–87
109. Bromberg Y, Kahn PC, Rost B. 2013. Neutral and weakly nonneutral sequence variants may define individuality. *PNAS* 110:14255–60
110. Miller M, Bromberg Y, Swint-Kruse L. 2017. Computational predictors fail to identify amino acid substitution effects at rheostat positions. *Sci. Rep.* 7:41329
111. Ng PC, Henikoff S. 2001. Predicting deleterious amino acid substitutions. *Genome Res*. 11:863–74
112. Choi Y, Sims GE, Murphy S, Miller JR, Chan AP. 2012. Predicting the functional effect of amino acid substitutions and indels. *PLOS ONE* 7:e46688
113. Hecht M, Bromberg Y, Rost B. 2015. Better prediction of functional effects for sequence variants. *BMC Genom*. 16:S1
114. Li B, Krishnan VG, Mort ME, Xin F, Kamati KK, et al. 2009. Automated inference of molecular mechanisms of disease from amino acid substitutions. *Bioinformatics* 25:2744–50
115. Pejaver V, Urresti J, Lugo-Martinez J, Pagel KA, Lin GN, et al. 2017. MutPred2: inferring the molecular and phenotypic impact of amino acid variants. bioRxiv 134981. **https://doi.org/10.1101/134981**
116. Miller M, Wang Y, Bromberg Y. 2019. What went wrong with variant effect predictor performance for the PCM1 challenge. *Hum. Mutat.* 40:1486–94
117. Miller M, Vitale D, Kahn PC, Rost B, Bromberg Y. 2019. funtrp: identifying protein positions for variation driven functional tuning. *Nucleic Acids Res*. 47(21):e142
118. Landrum MJ, Lee JM, Riley GR, Jang W, Rubinstein WS, et al. 2014. ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res*. 42:D980–85

119. Stenson PD, Mort M, Ball EV, Evans K, Hayden M, et al. 2017. The Human Gene Mutation Database: towards a comprehensive repository of inherited mutation data for medical research, genetic diagnosis and next-generation sequencing studies. *Hum. Genet.* 136:665–77

120. Meyer PE, Lafitte F, Bontempi G. 2008. minet: a R/Bioconductor package for inferring large transcriptional networks using mutual information. *BMC Bioinform.* 9:461

121. Yue Z, Li HT, Yang Y, Hussain S, Zheng CH, et al. 2016. Identification of breast cancer candidate genes using gene co-expression and protein-protein interaction information. *Oncotarget* 7:36092–100

122. Amar D, Vizel A, Levy C, Shamir R. 2018. ADEPTUS: a discovery tool for disease prediction, enrichment and network analysis based on profiles from many diseases. *Bioinformatics* 34:1959–61

123. Cirincione AG, Clark KL, Kann MG. 2018. Pathway networks generated from human disease phenome. *BMC Med. Genom.* 11:75

124. Cowen L, Ideker T, Raphael BJ, Sharan R. 2017. Network propagation: a universal amplifier of genetic associations. *Nat. Rev. Genet.* 18:551–62

125. Lamparter D, Marbach D, Rueedi R, Kutalik Z, Bergmann S. 2016. Fast and rigorous computation of gene and pathway scores from SNP-based summary statistics. *PLOS Comput. Biol.* 12:e1004714

126. Wei Z, Wang W, Bradfield J, Li J, Cardinale C, et al. 2013. Large sample size, wide variant spectrum, and advanced machine-learning technique boost risk prediction for inflammatory bowel disease. *Am. J. Hum. Genet.* 92:1008–12

127. Jostins L, Ripke S, Weersma RK, Duerr RH, McGovern DP, et al. 2012. Host–microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature* 491:119–24

128. Han L, Maciejewski M, Brockel C, Gordon W, Snapper SB, et al. 2017. A probabilistic pathway score (PROPS) for classification with applications to inflammatory bowel disease. *Bioinformatics* 34:985–93

129. Kanehisa M, Sato Y, Kawashima M, Furumichi M, Tanabe M. 2015. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res.* 44:D457–62

130. Qin J, Li R, Raes J, Arumugam M, Burgdorf KS, et al. 2010. A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* 464:59–65

131. Human Microbiome Proj. Consort. 2012. A framework for human microbiome research. *Nature* 486:215–21

132. Cole JR, Wang Q, Fish JA, Chai B, McGarrell DM, et al. 2014. Ribosomal Database Project: data and tools for high throughput rRNA analysis. *Nucleic Acids Res.* 42:D633–42

133. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, et al. 2013. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res.* 41:D590–96

134. McDonald D, Price MN, Goodrich J, Nawrocki EP, DeSantis TZ, et al. 2012. An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. *ISME J.* 6:610–18

135. Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, et al. 2010. QIIME allows analysis of high-throughput community sequencing data. *Nat. Methods* 7:335–36

136. Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, et al. 2009. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl. Environ. Microbiol.* 75:7537–41

137. Plummer E, Twin J, Bulach DM, Garland SM, Tabrizi SN. 2015. A comparison of three bioinformatics pipelines for the analysis of preterm gut microbiota using 16S rRNA gene sequencing data. *J. Proteom. Bioinform.* 8:283–91

138. Bolyen E, Rideout JR, Dillon MR, Bokulich NA, Abnet CC, et al. 2019. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nat. Biotechnol.* 37:852–57

139. Almeida A, Mitchell AL, Tarkowska A, Finn RD. 2018. Benchmarking taxonomic assignments based on 16S rRNA gene profiling of the microbiota from commonly sampled environments. *Gigascience* 7:giy054

140. Janda JM, Abbott SL. 2007. 16S rRNA gene sequencing for bacterial identification in the diagnostic laboratory: pluses, perils, and pitfalls. *J. Clin. Microbiol.* 45:2761–64

141. Delmont TO, Prestat E, Keegan KP, Faubladier M, Robe P, et al. 2012. Structure, fluctuation and magnitude of a natural grassland soil metagenome. *ISME J* 6:1677–87

142. Truong DT, Franzosa EA, Tickle TL, Scholz M, Weingart G, et al. 2015. MetaPhlAn2 for enhanced metagenomic taxonomic profiling. *Nat. Methods* 12:902–3

143. Milanese A, Mende DR, Paoli L, Salazar G, Ruscheweyh H-J, et al. 2019. Microbial abundance, activity and population genomic profiling with mOTUs2. *Nat. Commun.* 10:1014

144. Kim D, Song L, Breitwieser FP, Salzberg SL. 2016. Centrifuge: rapid and sensitive classification of metagenomic sequences. *Genome Res.* 26:1721–29

145. Lu J, Breitwieser FP, Thielen P, Salzberg SL. 2017. Bracken: estimating species abundance in metagenomics data. *PeerJ Comput. Sci.* 3:e104

146. Wood DE, Lu J, Langmead B. 2019. Improved metagenomic analysis with Kraken 2. *Genome Biol.* 20:257

147. Menzel P, Ng KL, Krogh A. 2016. Fast and sensitive taxonomic classification for metagenomics with Kaiju. *Nat. Commun.* 7:11257

148. Ounit R, Wanamaker S, Close TJ, Lonardi S. 2015. CLARK: fast and accurate classification of metagenomic and genomic sequences using discriminative *k*-mers. *BMC Genom.* 16:236

149. Ye SH, Siddle KJ, Park DJ, Sabeti PC. 2019. Benchmarking metagenomics tools for taxonomic classification. *Cell* 178:779–94

150. Nayfach S, Shi ZJ, Seshadri R, Pollard KS, Kyrpides NC. 2019. New insights from uncultivated genomes of the global human gut microbiome. *Nature* 568:505–10

151. Almeida A, Mitchell AL, Boland M, Forster SC, Gloor GB, et al. 2019. A new genomic blueprint of the human gut microbiota. *Nature* 568:499–504

152. Pasolli E, Asnicar F, Manara S, Zolfo M, Karcher N, et al. 2019. Extensive unexplored human microbiome diversity revealed by over 150,000 genomes from metagenomes spanning age, geography, and lifestyle. *Cell* 176:649–62.e20

153. Lerminiaux NA, Cameron ADS. 2019. Horizontal transfer of antibiotic resistance genes in clinical environments. *Can. J. Microbiol.* 65:34–44

154. Zhu W, Lomsadze A, Borodovsky M. 2010. Ab initio gene identification in metagenomic sequences. *Nucleic Acids Res.* 38:e132

155. Sczyrba A, Hofmann P, Belmann P, Koslicki D, Janssen S, et al. 2017. Critical Assessment of Metagenome Interpretation—a benchmark of metagenomics software. *Nat. Methods* 14:1063–71

156. Sinha R, Clarke J, Benson AK. 2015. Alignment behaviors of short peptides provide a roadmap for functional profiling of metagenomic data. *BMC Genom.* 16:1080

157. Schnoes AM, Brown SD, Dodevski I, Babbitt PC. 2009. Annotation error in public databases: misannotation of molecular function in enzyme superfamilies. *PLOS Comput. Biol.* 5:e1000605

158. Abubucker S, Segata N, Goll J, Schubert AM, Izard J, et al. 2012. Metabolic reconstruction for metagenomic data and its application to the human microbiome. *PLOS Comput. Biol.* 8:e1002358

159. Overbeek R, Olson R, Pusch GD, Olsen GJ, Davis JJ, et al. 2014. The SEED and the rapid annotation of microbial genomes using subsystems technology (RAST). *Nucleic Acids Res.* 42:D206–14

160. Caspi R, Altman T, Billington R, Dreher K, Foerster H, et al. 2013. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Res.* 42:D459–71

161. Suzek BE, Wang Y, Huang H, McGarvey PB, Wu CH, UniProt Consort. 2015. UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics* 31:926–32

162. Nazeen S, Yu YW, Berger B. 2020. Carnelian uncovers hidden functional patterns across diverse study populations from whole metagenome sequencing reads. *Genome Biol.* 21(1):47

163. Donaldson GP, Lee SM, Mazmanian SK. 2016. Gut biogeography of the bacterial microbiota. *Nat. Rev. Microbiol.* 14:20–32

164. Dominguez-Bello MG, Godoy-Vitorino F, Knight R, Blaser MJ. 2019. Role of the microbiome in human development. *Gut* 68:1108–14

165. Castaner O, Goday A, Park Y-M, Lee S-H, Magkos F, et al. 2018. The gut microbiome profile in obesity: a systematic review. *Int. J. Endocrinol.* 2018:4095789

166. Karlsson FH, Tremaroli V, Nookaew I, Bergstrom G, Behre CJ, et al. 2013. Gut metagenome in European women with normal, impaired and diabetic glucose control. *Nature* 498:99–103

167. Kostic AD, Xavier RJ, Gevers D. 2014. The microbiome in inflammatory bowel disease: current status and the future ahead. *Gastroenterology* 146:1489–99

168. Li Q, Han Y, Dy ABC, Hagerman RJ. 2017. The gut microbiota and autism spectrum disorders. *Front. Cell. Neurosci.* 11:120

169. Manichanh C, Rigottier-Gois L, Bonnaud E, Gloux K, Pelletier E, et al. 2006. Reduced diversity of faecal microbiota in Crohn's disease revealed by a metagenomic approach. *Gut* 55:205–11

170. Frank DN, St. Amand AL, Feldman RA, Boedeker EC, Harpaz N, Pace NR. 2007. Molecular-phylogenetic characterization of microbial community imbalances in human inflammatory bowel diseases. *PNAS* 104:13780–85

171. Morgan XC, Tickle TL, Sokol H, Gevers D, Devaney KL, et al. 2012. Dysfunction of the intestinal microbiome in inflammatory bowel disease and treatment. *Genome Biol.* 13:R79

172. Ackerman H, Usen S, Jallow M, Sisay-Joof F, Pinder M, Kwiatkowski DP. 2005. A comparison of case-control and family-based association methods: the example of sickle-cell and malaria. *Ann. Hum. Genet.* 69:559–65

173. Thye T, Owusu-Dabo E, Vannberg FO, van Crevel R, Curtis J, et al. 2012. Common variants at 11p13 are associated with susceptibility to tuberculosis. *Nat. Genet.* 44:257–59

174. Wang Z, Sun Y, Xa Fu, Yu G, Wang C, et al. 2016. A large-scale genome-wide association and meta-analysis identified four novel susceptibility loci for leprosy. *Nat. Commun.* 7:13760

175. Whittaker RH. 1960. Vegetation of the Siskiyou Mountains, Oregon and California. *Ecol. Monogr.* 30:279–338

176. Bray JR, Curtis JT. 1957. An ordination of the upland forest communities of southern Wisconsin. *Ecol. Monogr.* 27:325–49

177. Goodrich JK, Davenport ER, Beaumont M, Jackson MA, Knight R, et al. 2016. Genetic determinants of the gut microbiome in UK twins. *Cell Host Microbe* 19:731–43

178. Blekhman R, Goodrich JK, Huang K, Sun Q, Bukowski R, et al. 2015. Host genetic variation impacts microbiome composition across human body sites. *Genome Biol.* 16:191

179. Bonder MJ, Kurilshikov A, Tigchelaar EF, Mujagic Z, Imhann F, et al. 2016. The effect of host genetics on the gut microbiome. *Nat. Genet.* 48:1407–12

180. Rothschild D, Weissbrod O, Barkan E, Kurilshikov A, Korem T, et al. 2018. Environment dominates over host genetics in shaping human gut microbiota. *Nature* 555:210–15

181. Wang J, Thingholm LB, Skieceviciene J, Rausch P, Kummen M, et al. 2016. Genome-wide association analysis identifies variation in vitamin D receptor and other host factors influencing the gut microbiota. *Nat. Genet.* 48:1396–406

182. Imhann F, Vich Vila A, Bonder MJ, Fu J, Gevers D, et al. 2018. Interplay of host genetics and gut microbiota underlying the onset and clinical presentation of inflammatory bowel disease. *Gut* 67:108–19

183. Integr. HMP (iHMP) Res. Netw. Consort. 2014. The Integrative Human Microbiome Project: dynamic analysis of microbiome-host omics profiles during periods of human health and disease. *Cell Host Microbe* 16:276–89

184. Proctor LM, Creasy HH, Fettweis JM, Lloyd-Price J, Mahurkar A, et al. 2019. The Integrative Human Microbiome Project. *Nature* 569:641–48

185. LeCun Y, Bengio Y, Hinton G. 2015. Deep learning. *Nature* 521:436–44

186. Hinton GE, Salakhutdinov RR. 2006. Reducing the dimensionality of data with neural networks. *Science* 313:504–7

187. Gill SR, Pop M, DeBoy RT, Eckburg PB, Turnbaugh PJ, et al. 2006. Metagenomic analysis of the human distal gut microbiome. *Science* 312:1355–59

188. Wikoff WR, Anfora AT, Liu J, Schultz PG, Lesley SA, et al. 2009. Metabolomics analysis reveals large effects of gut microflora on mammalian blood metabolites. *PNAS* 106:3698–703

189. Fergus C, Barnes D, Alqasem MA, Kelly VP. 2015. The queuine micronutrient: charting a course from microbe to man. *Nutrients* 7:2897–929

190. Shaw KA, Cutler DJ, Okou D, Dodd A, Aronow BJ, et al. 2019. Genetic variants and pathways implicated in a pediatric inflammatory bowel disease cohort. *Genes Immun.* 20:131–42

191. Mallory EK, Acharya A, Rensi SE, Turnbaugh PJ, Bright RA, Altman RB. 2018. Chemical reaction vector embeddings: towards predicting drug metabolism in the human gut microbiome. *Pac. Symp. Biocomput.* 23:56–67

192. Wang M, Tai C, Weinan E, Wei L. 2018. DeFine: Deep convolutional neural networks accurately quantify intensities of transcription factor-DNA binding and facilitate evaluation of functional non-coding variants. *Nucleic Acids Res.* 46:e69

193. Tang H, Thomas PD. 2016. PANTHER-PSEP: predicting disease-causing genetic variants using position-specific evolutionary preservation. *Bioinformatics* 32:2230–32

194. Vazquez M, Pons T, Brunak S, Valencia A, Izarzugaza JM. 2016. wKinMut-2: identification and interpretation of pathogenic variants in human protein kinases. *Hum. Mutat.* 37:36–42

195. Bendl J, Stourac J, Salanda O, Pavelka A, Wieben ED, et al. 2014. PredictSNP: robust and accurate consensus classifier for prediction of disease-related mutations. *PLOS Comput. Biol.* 10:e1003440

196. Shihab HA, Gough J, Mort M, Cooper DN, Day INM, Gaunt TR. 2014. Ranking non-synonymous single nucleotide polymorphisms based on disease concepts. *Hum. Genom.* 8:11

197. Shihab HA, Gough J, Cooper DN, Stenson PD, Barker GLA, et al. 2013. Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models. *Hum. Mutat.* 34:57–65

198. Capriotti E, Altman RB, Bromberg Y. 2013. Collective judgment predicts disease-associated single nucleotide variants. *BMC Genom.* 14(Suppl. 3):S2

199. Izarzugaza JMG, del Pozo A, Vazquez M, Valencia A. 2012. Prioritization of pathogenic mutations in the protein kinase superfamily. *BMC Genom.* 13(Suppl. 4):S3

200. Reva B, Antipin Y, Sander C. 2011. Predicting the functional impact of protein mutations: application to cancer genomics. *Nucleic Acids Res.* 39:e118

201. Dehouck Y, Grosfils A, Folch B, Gilis D, Bogaerts P, Rooman M. 2009. Fast and accurate predictions of protein stability changes upon mutations using statistical potentials and neural networks: PoPMuSiC-2.0. *Bioinformatics* 25:2537–43

202. Capriotti E, Calabrese R, Casadio R. 2006. Predicting the insurgence of human genetic diseases associated to single point protein mutations with support vector machines and evolutionary information. *Bioinformatics* 22:2729–34

203. Mathe E, Olivier M, Kato S, Ishioka C, Hainaut P, Tavtigian SV. 2006. Computational approaches for predicting the biological effect of p53 missense mutations: a comparison of three sequence analysis based methods. *Nucleic Acids Res.* 34:1317–25

204. Schymkowitz J, Borg J, Stricher F, Nys R, Rousseau F, Serrano L. 2005. The FoldX web server: an online force field. *Nucleic Acids Res.* 33:W382–88

205. Stone EA, Sidow A. 2005. Physicochemical constraint violation by missense substitutions mediates impairment of protein function and disease severity. *Genome Res.* 15:978–86

206. Bao L, Zhou M, Cui Y. 2005. nsSNPAnalyzer: identifying disease-associated nonsynonymous single nucleotide polymorphisms. *Nucleic Acids Res.* 33:W480–82

207. Ramensky V, Bork P, Sunyaev S. 2002. Human non-synonymous SNPs: server and survey. *Nucleic Acids Res.* 30:3894–900

208. Schwarz JM, Rodelsperger C, Schuelke M, Seelow D. 2010. MutationTaster evaluates disease-causing potential of sequence alterations. *Nat. Methods* 7:575–76

Annual Review of
Biomedical Data
Science

Volume 3, 2020

Contents

Errata

An online log of corrections to *Annual Review of Biomedical Data Science* articles may be found at http://www.annualreviews.org/errata/biodatasci