BioEssays

www.bioessays-journal.com

# The Protein-Coding Human Genome: Annotating High-Hanging Fruits

*Klas Hatje, Stefanie Mühlhausen, Dominic Simm, and Martin Kollmar**

**The major transcript variants of human protein-coding genes are annotated to a certain degree of accuracy combining manual curation, transcript data, and proteomics evidence. However, there is considerable disagreement on the annotation of about 2000 genes—they can be protein-coding, noncoding, or pseudogenes—and on the annotation of most of the predicted alternative transcripts. Pure transcriptome mapping approaches seem to be limited in discriminating functional expression from noise. These limitations have partially been overcome by dedicated algorithms to detect alternative spliced micro-exons and wobble splice variants. Recently, knowledge about splice mechanism and protein structure are incorporated into an algorithm to predict neighboring homologous exons, often spliced in a mutually exclusive manner. Predicted exons are evaluated by transcript data, structural compatibility, and evolutionary conservation, revealing hundreds of novel coding exons and splice mechanism re-assignments. The emerging human pan-genome is necessitating distinctive annotations incorporating differences between individuals and between populations.**

## 1. Introduction

Accuracy of human genome annotation significantly impacts whole genome, transcriptome, and exome sequencing studies and is therefore of high impact for biological and medical interpretation of experimental results. Usually, the more complete the gene annotation is, the more accurate are downstream analyses, such as short read mapping for spliced reads[1] or identification of genetic variations from whole genome sequencing studies.[2] In contrast, overestimated numbers of annotated exons and transcripts lead to less robust gene abundance and differential gene expression estimates.[3,4] Gene annotation impacts the choice of genetic regions included in exome sequencing, targeted sequencing approaches, and precise genome editing utilizing CRISPR/Cas9. Genomic experiments are more and more applied in clinical practice and, therefore, the accuracy of these methods is highly relevant for diagnosis and treatment of patients. Accurate annotation of genes also has an enormous influence on drug discovery. Gene therapies already allow direct manipulation of the genome, transcripts can be targeted with oligonucleotides, and exon splicing events can be adjusted by small molecules.[5,6]

In contrast to the centralized worldwide effort to assemble the human reference genome led by the Genome Reference Consortium (GRC), efforts to annotate the human genome are split between several major consortia and innumerable smaller groups. Consequently, after decades of identifying human genes a consistent and standardized catalogue of human protein-coding genes is still missing. In addition, many human multi-exon genes undergo alternative splicing and differences between annotated isoforms are even larger between databases. In this review, we will report on recent studies that performed human gene, transcript, and protein identification, while focusing on efforts to investigate those exonic regions that are difficult to detect.

## 2. Do Current Approaches Annotate Functional Regions, Junk, or Both?

Annotating a genome comprises all efforts to assign biological functions, mechanistic and structural roles, and observations linked to genomic positions to every nucleotide in the genome. Thus, a set of publications reporting on the final data from the Encyclopedia of DNA Elements project (ENCODE) and claiming that most of the human genome consists of functional regions[7] initiated a lively, highly controversial and ongoing debate about "function" versus "junk."[8–16] In short, ENCODE generated

Dr. K. Hatje
Roche Pharmaceutical Research and Early Development
Pharmaceutical Sciences
Roche Innovation Center Basel
F. Hoffmann-La Roche Ltd.
Grenzacherstr. 124, 4070 Basel, Switzerland

Dr. S. Mühlhausen, D. Simm, Dr. M. Kollmar
Group Systems Biology of Motor Proteins
Department of NMR-based Structural Biology
Max-Planck-Institute for Biophysical Chemistry
Am Fassberg 11, 37077 Göttingen, Germany
E-mail: mako@nmr.mpibpc.mpg.de

D. Simm
Theoretical Computer Science and Algorithmic Methods
Institute of Computer Science
Georg-August-University Göttingen
Goldschmidtstr. 7, 37077 Göttingen, Germany

The ORCID identification number(s) for the author(s) of this article can be found under https://doi.org/10.1002/bies.201900066



**DOI: 10.1002/bies.201900066**

**Box 1**
**What is a gene?**

At the time when the word "gene" was coined, the term was looked at from the phenotype perspective as a distinct region, a "locus," on a chromosome explaining mechanisms of heredity, development, and physiological function. Later, with the discovery of DNA and the publication of the "Central Dogma" of molecular biology, a gene became a physical entity that is transcribed and finally translated into protein. While the phenotype-based view (also called functionalist approach) did not substantially change over time, the genocentric view and molecular understanding was significantly refined including RNA-genes, "genes in pieces," and regulatory aspects. When the human genome sequence became finished, the gene still was a genomic region with clear structural boundaries. However, alternative splicing already challenged the genotype-phenotype relationship, because it generates different protein isoforms implying different physiological functions derived from the same gene. Subsequent further discoveries seem to have completely dissolved this relationship. While "gene" is one of the major terms in biology, there is no unifying definition for different purposes and disciplines.[121] The "instrumental gene" comprises the concept in which a phenotype (a disease, a Mendelian trait or another observable characteristic) is related to a locus that is not necessarily a molecular gene but could be any other functional DNA element. While being clear on the phenotype, this concept remains vague on the genotype. In contrast, the "nominal gene" describes the molecular entity encoded by DNA and transcribed into RNA. This concept constitutes the definition used by the Encyclopedia of DNA Elements project (ENCODE) project: "A gene is a union of genomic sequences encoding a coherent set of potentially overlapping functional products."[19] While this clearly defines a genotype, it is completely oblivious of physiological consequence. The ENCODE definition is more restricted than a previous definition by the Sequence Ontology consortium: "A locatable region of genomic sequence, corresponding to a unit of inheritance, which is associated with regulatory regions, transcribed regions, and/or other functional sequence regions."[122] Although both concepts are highly used and especially the molecular gene is a critical practical tool for communication between bioscientists, both concepts did not keep pace with the data generated and knowledge acquired in the past 15 years. Given the transcription of almost the entire genome, genes hardly have defined boundaries,[123] exons from different genes can be part of the same transcript,[124] some microbial genomes contain thousands of scrambled genes that need to be decrypted during development,[125] the functional status of a gene can be passed down to a daughter cell, and at least mammals and plants can rewrite their DNA based on RNA inherited from past generations.[126] Accordingly, the traditional reductionist way of thinking is continuously changing toward a more systemic concept, which is summarized by a so-called "relational or systemic gene" or "postgenomics gene."[127–129] Advancing current human genome annotation to such a holistic view will be an enormous step forward allowing the readout of DNA/RNA/protein sequences based on biological questions and not predefined computational categories. In such a system, pieces of DNA, for example, could be assembled independent of their (current) status (exon, regulatory, protein-binding, etc.) and their (current) assignment to a dedicated gene.

massive data from high-throughput RNA-sequencing, transcription-factor-binding, chromatin structure, and histone modification experiments and concluded that 80.4% of the genome shows some sort of biochemical activity. Their conclusion that this activity is completely functional was strongly criticized for ignoring the "affirming the consequent"[11] and for showing only "causal role" but no "selected-effect" functionality of the respective genomic regions. According to the critique, only those regions showing "selected-effect" functionality should be termed functional, and regions having a "causal role" should better be described to have an "effect," "role," "consequence," or "activity." This and most other points of critique such as the size of detected regions and the missing explanation of the C-value paradox[17] boiled down to the question what function is and what it is not. It has been argued that this is not a semantic question because the answer would extend deeply into many biological disciplines, such as biochemistry, molecular and evolutionary biology, and genetics.[10,18] While ENCODE provided a very detailed definition of the term "gene"[19]—which is another fundamental concept with changing meaning across disciplines and time (**Box 1**)—ENCODE missed to explain their use of the term "function."[14,18,20]

Unfortunately, currently available guidelines such as those outlined in the "evolutionary classification of genomic function"[15] are unsuitable when it comes to such complex problems as human genome annotations. For example, many genes are nonessential, about 75% of yeast genes,[21,22] and while the exact number is impossible to be determined for humans, a good estimate might be that at least 6% of genes are nonfunctional in a small part of the Iceland population.[23] Should the nonessential genes be annotated, therefore, as "indifferent DNA [...] whose main function is being there, but whose exact sequence is not important?"[15] Should the same loci, that are nonfunctional in the Iceland population, be termed "junk DNA" (the suggested term for pseudogenes) in Icelander genomes, but functional, "literal DNA" in genomes of other populations?

To resolve the complexity of the concept "function" with respect to genome annotation we suggest using the term in combination with the respective annotation layer. For example, according to the latest Ensembl genome annotation, 41.56% of the human genome is made up of protein-coding genes (40.78% according to NCBI). Therefore, at the layer of "genes," 41.56% of the human genome is functional. However, only 1.15% of the genome represents protein-coding sequence (sum of all constitutive and

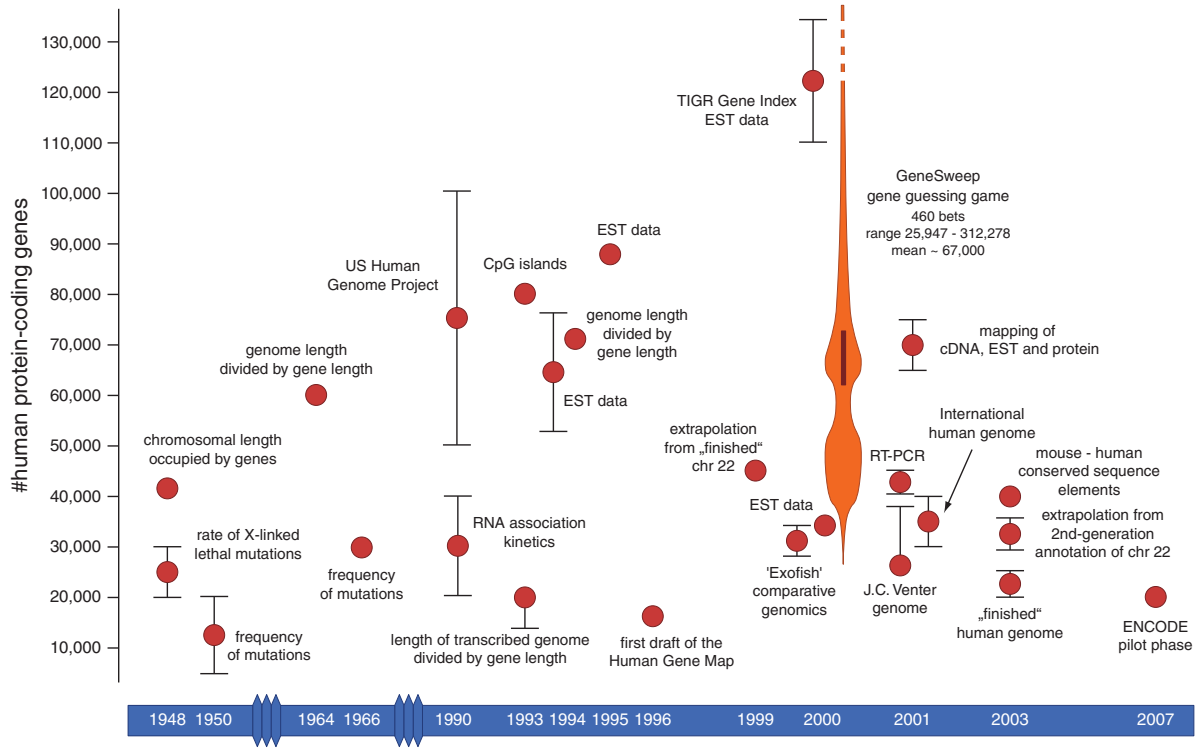



**Figure 1.** History of the prediction of the number of human genes. Early estimates of the number of human genes in the 1940s to 1960s were based on extrapolations of the few known genetic markers and sequence lengths. In 1990, the U. S. Human Genome Project proposed a range of 50 000 to 100 000 human genes without specifying the data underlying this estimation. Subsequent numbers were derived by various experimental methods and the numbers reported in the publications of the draft genome assemblies were derived from ab initio gene predictions. The latter numbers were subsequently refined by mapping extensive experimental data.

alternative exons; 1.17% according to NCBI) and thus determines the functional genome at the layer of protein-coding nucleotides. Similarly, at the resolution of the assays employed by ENCODE, 80.4% of the human genome might be considered "biochemically functional," while at the resolution of nucleotides, the number is much lower. In case of the chromatin immunoprecipitation assay, for example, the resolution of the assay with about 600 nucleotides results in 8.5% of the human genome representing transcription factor binding sites, while at the layer of nucleotides (transcription factor binding sites are usually not longer than 3 to 15 nucleotides), in fact only about 0.14% might bind transcription factors.[11] To be consistent with the "selected-effect" functionality concept, only annotations at the nucleotide level could be termed functional (still, there is the population problem), while annotations at all other layers could be termed "role" or "activity." This would, however, not be consistent with the use of the term "function" in genetics, cell biology, and most other biological disciplines.[24] The various functional annotations of the human genome are usually available as separate data tracks in genome browsers and gene reference entries.

## 3. The Number of Human Protein-Coding Genes Went Up and Down in the Pre-Genome Era

Estimating the number of human genes dates back to the 1940s when the genetic code and even the structure of DNA were unknown. In 1948, James N. Spuhler estimated the number of genes a) to 42 000 by assuming human genes occupying the same mean chromosomal length than fruit fly genes and b) to 19 890–30 420 by extrapolating the number of loci in the nonhomologous segment of the sex chromosome derived from X-linked lethal mutations (**Figure 1** and **Box 2**).[25] At about the same time, Hermann J. Muller estimated the frequency of mutations in humans resulting in 5000 to 20 000 human genes,[26] which is, on the upper limit, very close to the numbers discussed today. Later, Friedrich Vogel for the first time used physical entities estimating the number of genes based on the weight of genes of average length extrapolated to the weight of one human haploid chromosome set.[27] He calculated two numbers: one number he based on the length of haemoglobin genes resulting in 6.7 Mio genes, which he already dismissed as disturbingly high. Unfortunately, this number was nevertheless presented by Pertea and Salzberg as Vogel's number of genes,[28] which likely caused others to later cite this wrong number as well.[11,29] The number Friedrich Vogel considered more reliable, 60 000 human genes, he derived by dividing the length of the human genome by the gene-length of 50 000 nucleotides inferred from the length of genes in Dipteran giant chromosomes.[27] In 1966, Muller revised his earlier estimate to "not much more than 30 000" genes.[30] These early estimates of 20 000–40 000 human genes based on genetic load arguments became the reference in textbooks and publications for the next 25 years, although the respective primary publications did not receive the citations they would have deserved.

**Box 2**
**Early estimates of the number of human protein-coding genes**

At the time when Spuhler and Muller first estimated the number of human genes,[25,26] the entity "gene" was thought to be a certain part, a "locus", on the physical chromosome that harbors a trait that is affected by mutations. Muller estimated not only the number of human genes close to today's values (5000–20 000 genes), but, interestingly, he also estimated the number of *Drosophila* genes to a minimum of 5000–10 000 in the same article, which is also very close to the number of genes known nowadays from genome sequence analysis. The discovery of the genetic code and the amino acid sequences of the $\alpha$- and $\beta$-chains of human hemoglobin allowed Vogel in 1964 to first estimate the number of human genes based on physical entities.[27] However, the number of genes based on the lengths of these proteins was disturbingly high (6.7 Mio genes). Vogel thus calculated another number based on gene lengths in Diptera, which he regarded more reliable (60 000 human genes). Although introns and repetitive regions were not known at that time, it was already well established that the amount of DNA in closely related species, with likely very similar numbers of genes, can differ by two orders of magnitude,[130] and that, at least in bacteria, parts of gene material work as regulators of other gene material[131] providing reasonable explanations for genes being longer than their actual coding sequences. Shortly thereafter, Muller revised his earlier estimate to "not much more than 30 000" genes based on newer data on spontaneous mutations and frequencies of X-ray induced mutations.[30] Geneticists referred to these numbers, 30 000 genes and an upper limit of 40 000, in the following 25 years albeit without properly citing the original studies. In 1990, the U. S. Human Genome Project proclaimed to sequence the human genome and to locate the suspected 50 000–100 000 human genes without providing any data or reference for this estimate.[31] Fast progress in sequencing genomic DNA led to human gene lengths bridging more than three orders of magnitude, and, depending on the methods applied to generating an average, the extrapolated number of human genes ranged from 20 000[35] to 71 000.[33] Application of high-throughput techniques provided further numbers, including 20 000–40 000 genes implied by the measurement of RNA re-association kinetics,[132] 80 000 genes implied by determining and extrapolating CpG island coverage,[32] and 64 000 genes implied by expressed sequence tag (EST) sequencing followed by clustering and extrapolation.[33] From today's perspective, it seems weird that the authors of the CpG island-based estimate (80 000 genes) strongly insisted against reduction of their estimate to 67 000 genes by others.[33,133] In 1996, a first human gene map was constructed to complement the human genetic map by mapping gene-based sequence tagged site markers resulting in 16 354 distinct loci.[134] The authors of this study did not present an own estimate of the total number of human genes, as referenced and repeated by others (e.g., [28,29]), but repeated the expectation from the announcement of the Human Genome Project. The publication of the human draft genomes in early 2001 did not stop speculations on higher gene numbers. Extrapolation of RT-PCR data of chromosome 22 predicted 41 000–45 000 genes[135] and mapping of available cDNA, EST, and protein data combined with gene predictions suggested 65 000–75 000 genes.[136] Still in 2003, when the "completion" of the human genome sequence was announced, researchers predicted 29 000–36 000 genes based on the extrapolation of a refined annotation of chromosome 22[137] and up to 40 000 protein-coding genes based on analysis of conserved sequence elements between human and mouse.[138]

In 1990, the U. S. Human Genome Project proclaimed 50 000–100 000 human genes,[31] a number without any basis but which subsequently became the cited reference (Figure 1). The disruptive success of sequencing and high-throughput technologies soon superimposed the older genetic load-based estimates and it became more popular to estimate higher numbers. Accordingly, research groups generating large-scale CpG island coverage and expressed sequence tag (EST) data predicted 64 000–87 983 human genes.[32–34] All these estimates were, however, extrapolations from nonexhaustive data or data from single chromosomes and there were always groups favoring more conservative assumptions. Different assumptions about average gene length and subsequent extrapolation of gene density, for example, resulted in 14 000–20 000[35] and 71 000 human genes[33] (Figure 1).

While approaching the release of the first draft of the human genome, researchers from The Institute for Genomic Research (TIGR) predicted 110 000 to 134 000 genes made available in the TIGR Gene Index based on massive EST data[36] (erroneously, this estimate has been referenced as "57 000" genes in later reviews[28,29]). In the same journal issue, other researchers predicted 33 630 and 34 700 genes based on similar EST data[37] and 28 000–34 000 genes by comparison with pufferfish[38] (Figure 1). The latter estimates were close to the ranges predicted from the draft genome assemblies, 26 588–38 588 genes[39] and 30 000–40 000 genes.[40] Still, in the following months and years, many scientists betted for considerably higher numbers of up to 312 278 genes with a mean of 67 006 in the GeneSweep gene guessing game.[41,42] Accordingly, the number of predicted genes became cited as "surprisingly low" in the introduction of almost every subsequent paper, despite the many previous studies presenting similar numbers and multiple further arguments that the genome-based numbers were not surprising at all.[43] Today, the exact number of human protein-coding genes is still unknown and given as 17 694 in neXtProt release 01/2019, 19 033 in Consensus Coding Sequence [CCDS] database release 22 (06/2018), 19 975 in GENCODE release 31 (06/2019), 20 203 in NCBI Homo sapiens Annotation Release 109 (03/2018), and 20 465 in Ensembl release 96 (04/2019).

All efforts differ by the criteria used for gene counting and annotation.

## 4. Evidence for Protein-Coding Genes Comes from Multiple Sources

A genome annotation is usually performed in two steps termed as structural and functional annotations. Structural annotation comprises the identification of protein-coding genes, RNA-coding genes, regulatory regions, protein-binding sequences, pseudogenes, noncoding RNA, transposons, and other repeats, while functional annotation assigns specific functions to each of the identified regions. In many eukaryotes, especially in humans, another level of complexity is reached through alternative combination of genomic regions ("alternative splicing") leading to different and overlapping transcripts that encode for proteins with often at least slightly different functions. In essence, a genome annotation does not prove the function of a genomic region, but is a structured collection of predictions and observations that can be used as reference.

There are multiple layers of evidence for genes, transcripts, and exons and support can be inferred from computational and experimental data. In a most basic approach an initial gene annotation is derived from genomic sequence alone using (generalized) hidden Markov models, conditional random fields, or support vector machines.[44] This approach is based on simple assumptions such as "protein-coding genes should contain start and stop codons" and "protein-coding genes should consist of concatenated exons without internal stop codons." Coding regions are then identified based on sequence patterns distinctive for exons, introns, and intergenic sequences. These statistical features are enhanced by specific, probabilistic patterns for, e.g., transcript splice sites and polyadenylation sequences.[45] Pure ab initio gene predictions are of little use for the identification of novel human genes and transcripts, because human genome annotation is one of the best genome annotations available. Still, unguided gene predictions have a use in evaluating the predictive power of sequence motifs for gene encoding. Used as such, unguided transcript reconstructions yield valid isoforms for about 41% of expressed genes including both completely and partially overlapping transcripts.[46]

Considerably better gene annotations are obtained if gene prediction software is supported by transcript data for feature training and gene model evaluation. Extensive evaluation of available gene prediction software and protocols on human gene annotation within RGASP (RNA-seq Genome Annotation Assessment Project) demonstrated that there is, at least currently, an upper limit of about 20% to 40% completely recovered transcripts plus an additional 20% to 30% transcripts recovered missing one exon.[46] Unfortunately, the fraction of false positive exon and transcript predictions (i.e., the number of predicted exons that are highly likely to be noncoding) has not extensively been evaluated yet. In a small-scale attempt, only 3.2% of a selection of 221 computationally predicted exons could experimentally be validated.[47]

Information from transcript data alone, from ESTs in earlier times to high-throughput RNA-sequencing (RNA-seq) today (**Box 3**), is used to both identify and evaluate transcribed regions and to refine annotation models.[48–50] More sequencing data also means more transcriptional noise and, therefore, various types of filters (e.g., presence of reads from multiple individuals and different tissues) are applied to collect transcripts of strong evidence only. Although protein-coding genes have, in general, specific nucleotide patterns and transcripts of these genes have polyadenylation tails allowing their physical enrichment for EST/cDNA data generation, there is also a low level of pseudo-gene transcription.[51,52] For many candidate genes their status as protein-coding gene or non-coding RNA gene is not resolved yet.[53–55] At this level, proteogenomics provides protein-level evidence combining genomics data and mass spectrometry-based proteomics. In these experiments, gene predictions and translations of transcript data are combined into protein sequence databases, which are used to compute theoretical peptide mass spectra, which in turn are compared to the experimental mass spectra for peptide identification.[54,56–60] The proteomics approach unfolds its full potential when tissue-specific transcript data are available allowing to prepare tissue-specific sequence databases to avoid mis-assignment of spectra.[60–63] Complementing direct evidence from transcriptional and translational data, comparative genomics allows finding functional genomic elements that are conserved between species. Genes, exons, and other DNA elements under natural selection are often highly conserved between close relatives and the core set of human genes are conserved in all bilateria.[64–67]

## 5. The Genome Is Annotated at Multiple Overlapping Layers

Functional annotations can be derived by in silico predictions from sequence alone or through comparison with knowledge bases. The latter means transferring functional assignments such as a protein domain or a tRNA isoacceptor type from homologous regions of one protein/RNA to another, which is essentially a prediction by homology. Functional annotations are done at multiple layers. Single nucleotides varying in a population are annotated as single nucleotide polymorphisms (SNPs) and receive a functional annotation through association with phenotypic characteristics or dysfunction as found in genetic diseases. If nucleotides happen to be part of a continuous stretch they might get a common function or role assignment as part of a larger region (global assignment; regions may overlap) and at the same time a functional assignment as single nucleotide (a local assignment). For example, the adenine nucleotide involved in lariat formation during eukaryotic transcript splicing is part of the branchpoint (local assignment), which is an essential part of a spliceosomal intron (first level global assignment). This, in turn, is part of a eukaryotic multi-exon gene (second level global assignment), which again might contribute to a phenotype-associated locus (third level global assignment). However, in the human genome, the vast majority of nucleotides do not have a local function but only a role as part of larger genomic regions. This is the case, for example, for most of the nucleotides within the extended inter- and intragenic regions. Their global role is, among others, being a spacer (or "ballast")[68] between genic and regulatory regions or having impact on chromosome structure, while the actual number and specific order of those nucleotides do not matter. Naturally, the density of nucleotides with local function

**Box 3**
**Technological improvements impacting human genome annotation**

The first generation of sequencing technologies utilizing Sanger's method[139,140] provided expressed sequence tags (ESTs), which are short cDNA snippets representing regions of expressed mRNA. EST data expanded the ab initio gene predictions with genome-wide transcriptional information.[141–143] ESTs were not only used to identify expressed regions in the genome, but also to determine exact splice junctions that define exons and provide information on how these are concatenated to transcripts.[144] Similarly, short reads from second generation sequencers yielded deep sampling information across the whole genome and are used to evaluate gene annotation models. The major technology advancements were introduced by the 454, SOLiD, and Illumina sequencing platforms[140,145] leading to a massive increase in RNA sequence information. These data had high impact on the accuracy of genome annotations, although there seems to be a limit upon which accuracy and precision cannot be further improved with current methods.[46] The main challenge in genome annotation thus shifted from identifying expressed regions with high sensitivity to annotating functional genes, alternative transcripts, and exons with high precision. In contrast to the EST/cDNA data, the short read data provides evidence for short exons and exon junctions only. Here, high-throughput long read sequencing allows to sequence transcripts in full-length and to validate combinations of alternative exons. Two technologies are most established for this purpose, Pacific Biosciences[146,147] and Nanopore[148–150] Isoform expression can nowadays also be analyzed at single-cell level[151] using and combining short reads and long-read sequencing methods.[149,152]

highly correlates with the size of a genome, meaning genomes having less and shorter inter- and intragenic regions show higher density.

## 6. Is There Consensus on the Low-Hanging Fruits?

Two major databases, NCBI (RefSeq) and Ensembl (GENCODE), dominated human gene annotations since release of the initial human genome assembly. While both efforts (and any other initiative) agreed on using the assembly provided by the GRC as reference, every annotation effort had and continues to have its own rules and definitions and these rules and definitions even change over time.[69,70] Even for the term "gene" no definition has agreed on (Box 1) so that not only numbers of named categories highly disagree but also categories and subcategories themselves. For this reason, the CCDS project started in 2005 to generate a reliable set of consensus sequences as reference for the scientific community. In its first release, this set contained 12 950 protein-coding genes (**Figure 2**). "Consensus" was defined as "protein-coding regions that agree at the start codon, stop codon, and splice junctions and for which the prediction meets quality assurance benchmarks."[71] The consensus increased rapidly to 18 407 protein-coding genes in 2011 (CCDS version 8) and reached 19 033 genes in the current version (v. 22). Although both the RefSeq and the GENCODE annotations heavily rely on manual gene annotation efforts, there is still a gap of 1200 and 950 genes comparing CCDS to RefSeq and GENCODE, respectively (Figure 2).

In 2011, the Human Proteome Project (HPP) started to map the entire human proteome in a systematic effort.[72] The two main efforts within HPP, neXtProt, and PeptideAtlas, try to match proteomics data with available gene and transcript datasets. They differ mainly in neXtProt including not only evidence from mass spectrometry data but also from Edman sequencing, biochemical studies, posttranslational modifications, protein–protein interactions, antibody-based techniques, 3D structures, and disease mutations.[73] The current releases of neXtProt and PeptideAtlas claim to contain unambiguous evidence for 17 694 and 15 798 proteins, respectively. Although both approaches steadily close the gap to the number of human genes known from transcript data, gene prediction, and evolutionary conservation, they are technically limited in detecting low-abundance proteins, sequences lacking proteolytic cleavage sites and proteins expressed in tissues unavailable for studies.

In summary, major annotation databases currently reach a consensus on about 94% of all protein-coding genes and agree that about 86% of these genes are in fact translated and present in at least one human tissue. These numbers are also in accordance with The Human Protein Atlas project, which combines antibody-based imaging, mass spectrometry-based proteomics, transcriptomics, and systems biology. Currently, these data support 18 899 protein-coding genes in human.[61] Given this wealth of evidence, claims that "one in five human genes still have unresolved coding status"[55] and might rather be noncoding genes or pseudogenes than protein-coding genes seem rather exaggerated.

Apart from consortia-guided gene annotation efforts, there are a few small-scale efforts that have provided impressive results. In one project, ab initio gene prediction was used to generate a set of 8 million candidate transcripts. Subsequent filtering and validation by 26 RNA-seq data sets and shotgun proteomics revealed 36 novel proteins and more than 31 000 new transcripts.[74] For generating the intropolis resource, this approach was taken even further.[48] There, 21 504 RNA-seq samples from the SRA archive were aligned against the human genome. About 57 000 new exon junctions have been identified that are present in at least 1000 samples.[48] In another approach, unstranded and stranded RNA-seq data were not directly mapped to the human genome but first assembled into transcript assemblies and subsequently improved by predicting the orientation of unstranded reads and by integrating information about transcription start, cleavage, and polyadenylation sites.[75] With this pipeline, called CAFE, the BIGTranscriptome dataset was generated adding thousands of potential transcripts

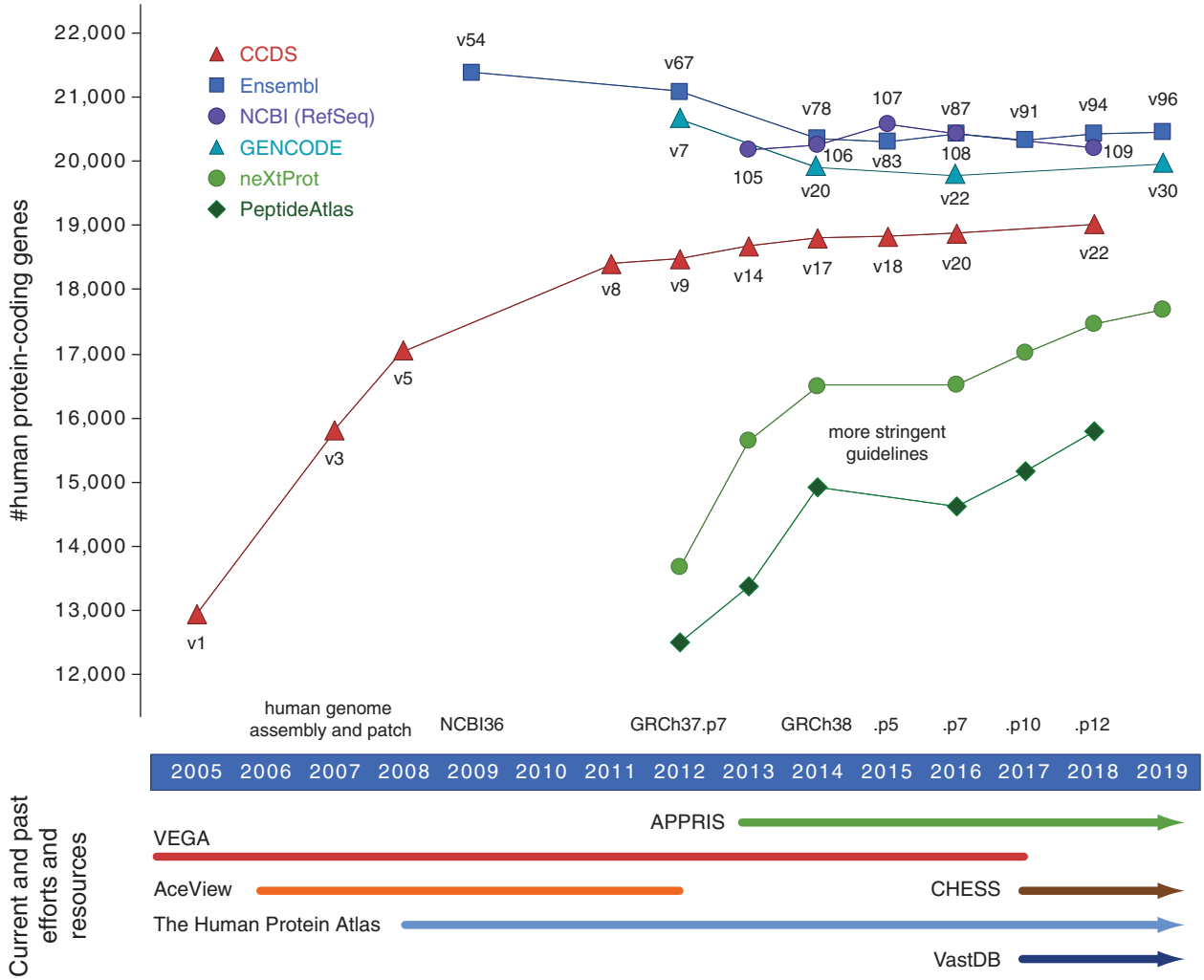


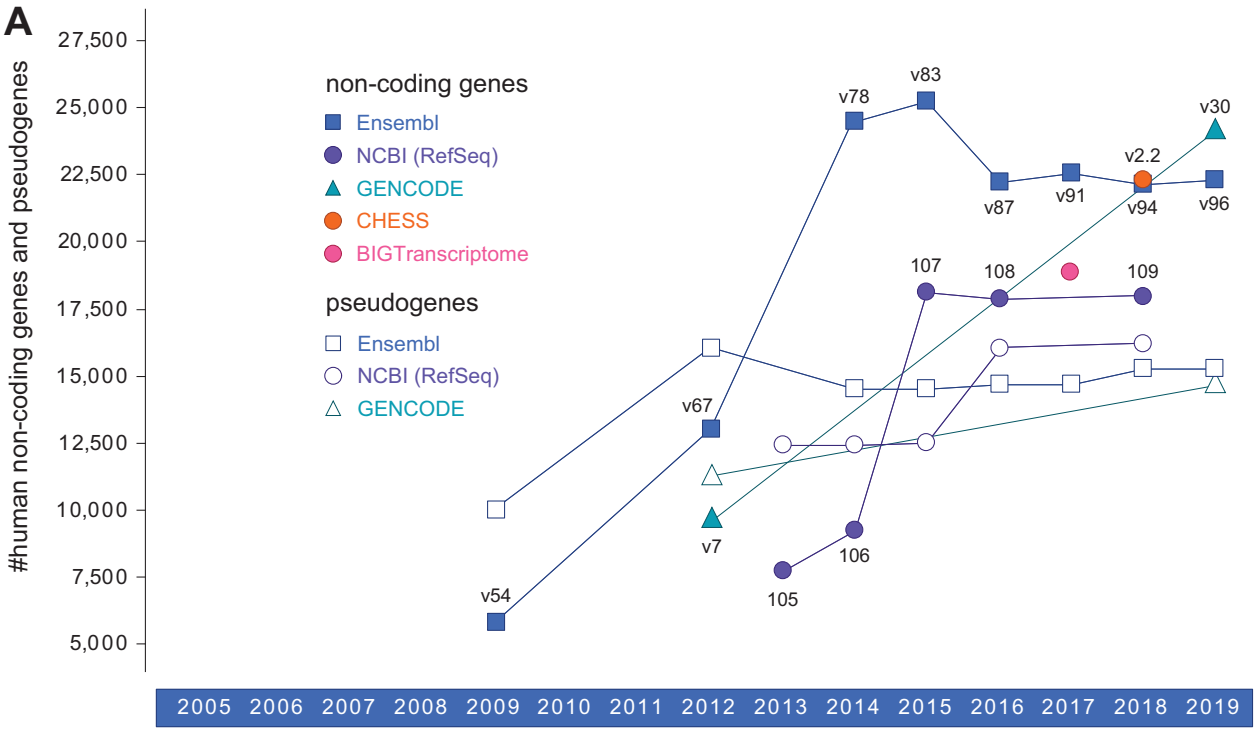

**Figure 2.** Efforts to annotate the human protein-coding genes. The scheme compiles the number of protein-coding genes as given by the respective data releases of the largest human genome annotation efforts. Numbers are different because definitions of terms and categories vary from effort to effort and even from time to time within efforts. New releases not only contain novel protein-coding genes but also drop genes. Thus, more genes from one release to the next are not directly related to the addition of the respective number of genes and likewise fewer genes do not represent the number of genes removed. The Consensus Coding Sequence project (CCDS) is an effort to mark all protein-coding genes with identical genomic coordinates in both the Ensembl and the NCBI annotations and contains 19 033 genes in the latest release. As example for a state-of-the-art analysis integrating transcriptomics and MS data, a recent tissue-specific expression study by Wang et al. detected 18 072 transcripts and 13 640 proteins.[60] Abbreviations of genome annotation efforts and resources: VEGA (The Vertebrate Genome Annotation), CHESS (Comprehensive Human Expressed SequenceS), and APPRIS (annotation of principal and alternative splice isoforms).

to GENCODE and RefSeq annotations (**Figure 3**). Similarly, in the most recent effort termed CHESS (Comprehensive Human Expressed Sequences), 9795 RNA-seq data sets generated by the genotype-tissue expression (GTEx) consortium were assembled resulting in 224 novel protein-coding genes and more than 116 000 novel transcripts (protein-coding and noncoding; Figure 3).[50] Many of the novel protein-coding genes were shown to be conserved in other mammals adding evolutionary evidence. Remarkably, 30 million of the assembled transcripts were annotated as nonfunctional and transcriptional noise.[50]

In summary, the number of protein-coding genes with support at protein, transcript, and homology level seems to stabilize around 20 000.[73,76] In addition, there are about 500 candidate genes whose existence or classification is unclear. While the uncertainty about the number of genes strongly decreased in the past 20 years (compare Figure 1 and Figure 2), the number of generated transcripts remains largely unclear (Figure 3B). It steadily increased in RefSeq, GENCODE, and other databases over the past ten years, from about 60 000 in 2009 to 210 000 in 2018, but it is not clear yet, to which extent these transcripts result from erroneous splicing[50,77–80] or are translated to a significant level, if at all.[62,81–85]

## 7. Finding All Protein-Coding Segments Remains Difficult

Alternative processing of primary RNA transcripts has been found across all eukaryotes and is a characteristic ranging back



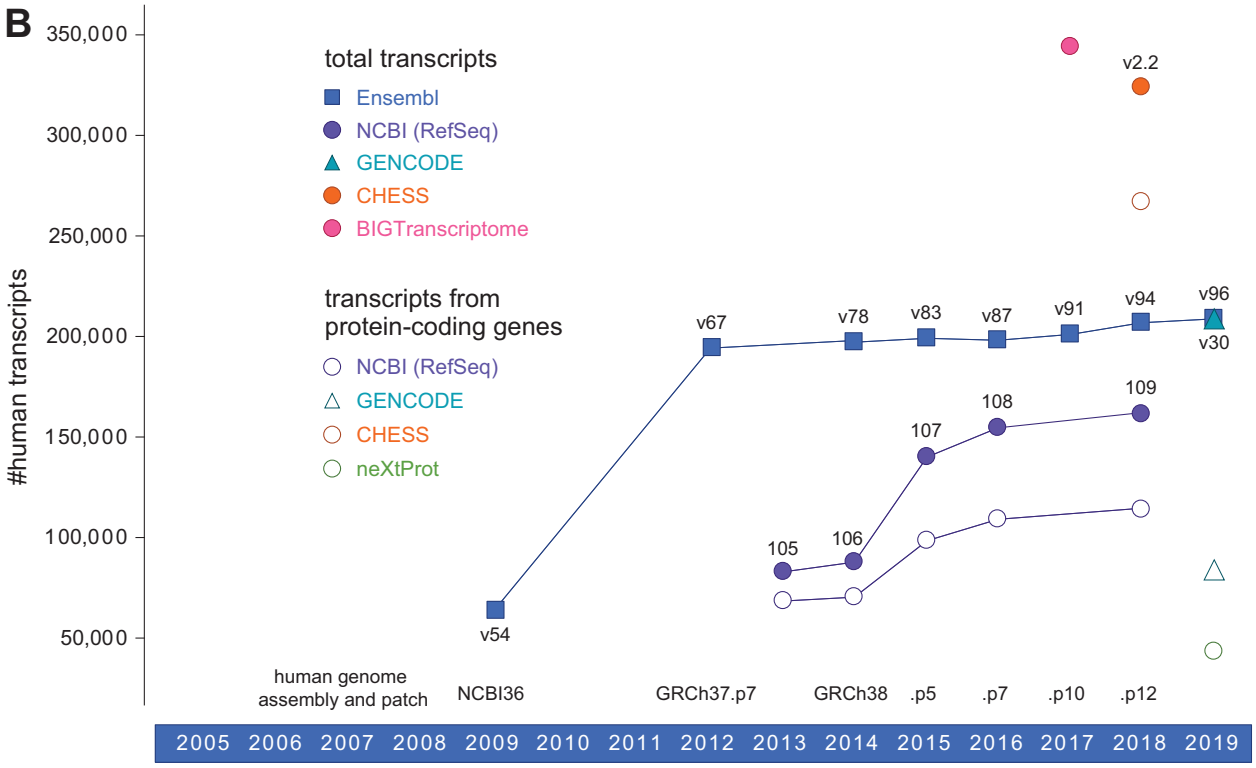




**Figure 3.** Annotation of human noncoding genes, pseudogenes, and transcripts. A) This scheme compiles the number of noncoding genes and pseudogenes as given by respective databases. B) The number of human transcripts, both the total number of transcripts and the fraction generated from protein-coding genes, considerably vary over time and across databases.

to the last eukaryotic common ancestor. It is used to increase proteome diversity and has been shown to be highly regulated in many species. There are many different types of alternative splicing such as differential inclusion of exons, intron retention, or alternative 5'- and 3'-splicing of exons. A particularly interesting case is mutually exclusive splicing, in which neighboring exons are spliced in a mutually exclusive manner into the mature transcript.

### 7.1. Micro-Exons

Micro-exons are very short coding exons and therefore difficult to detect. The maximum length of micro-exons differs between publications, further complicating systematic comparison:[86] Volfovsky et al. defined micro-exons by a maximum length of 25 nucleotides,[87] Wen et al. looked for short alternative splicing events of maximal 51 nucleotides,[88] Irimia et al. characterized micro-exons with lengths of 3–27 nucleotides[89] and Li et al. referred to a length between six and 51 nucleotides.[90] The latter, most exhaustive approach detected unknown micro-exons by mapping RNA-seq data from diverse datasets to the human reference transcriptome from Ensembl.[90] Mapped reads were filtered for insertions, which were defined as short, additional RNA stretches with a maximum length of 51 nucleotides. When occurring at the exon boundaries of a transcript, such insertions were considered candidate micro-exons. Subsequently, corresponding introns were scanned for the canonical splice site pattern GT–AG. Candidate micro-exons were approved if their sequences exactly matched the intronic sequences between AG and GT.[90] With this approach, 310 predicted novel micro-exons were added to the 12 835 Ensembl-annotated micro-exons.[90]

Micro-exons are considered to preserve the reading frame and, if alternatively spliced, to modulate protein structure. Interestingly, micro-exons are highly enriched in transcripts in brain compared to other tissues.[91] About 2500 neural-regulated micro-exons have been identified in each human and mouse[89] and their splicing was shown to be mis-regulated in autism.[89,92,93] Many of the very short micro-exons are well conserved from fish to human. Of the about 150 mammalian, neural-regulated micro-exons with lengths of 3–15 nucleotides at least 55 are deeply conserved in vertebrate species spanning 400–450 million years of evolution.[89] Micro-exon splicing is suggested to be promoted by a specialized domain in an ancestral splicing factor that originated in a common bilaterian ancestor.[94]

### 7.2. Wobble Splicing

Another mechanism introducing small variations to protein isoforms is wobble splicing. Here, a GYN repeat at the donor splice site (5' splice site; Y stands for C or T and N stands for A, C, G, or T) or an NAG repeat at the acceptor splice site (3' splice site) leads to subtle length variations in the spliced transcripts and finally to alternative isoforms differing in few amino acids. The most frequent wobble splicing element is the NAGNAG tandem repeat, which was systematically identified in the human genome using EST data.[95–97] In a first attempt to characterize NAGNAG splicing, the RefSeq annotation was searched, and 7326 candidate acceptor splice sites were discovered, corresponding to NAGNAG splicing in 30% of all RefSeq transcripts.[95] Subsequent EST data mapping confirmed 878 sites. In addition, NAGNAG acceptors show high conservation[98] with 73% being conserved between human and mouse.[95] Interestingly, NAGNAG repeats mainly alternate the protein sequence by one amino acid, but rarely introduce premature stop codons. Later, the same approach was used to identify and verify GYNGYN tandem repeats at donor splice sites in the human genome.[99]

In another approach to identify wobble splice sites the SwissProt database was searched for pairs of protein isoforms with single amino acid differences. The candidate list was then extended by reports in the literature about subtle protein differences.[96] Identified alternative donor splice sites were verified through presence in the human genome and through comparison with cDNA available from NCBI and EBI's AltSplice database. RNA-seq data were also used to investigate the regulation of NAGNAG splicing.[100] Here, sequences flanking the NAGNAG acceptor splice sites identified in the human genome were extracted and mapped with RNA-seq reads requiring at least six nucleotides overhang on each side. Evidence for tissue regulated splicing was found for 73% of NAGNAG acceptor splicing events.[100,101] Still, some events might better be explained by stochastic splicing alone.[102–105] This stochasticity has been described as a physiologically triggered, concerted shift in alternative splicing.[106] A specialized tool to detect and quantify NAGNAG splicing events identifies NAGNAG motifs in splice sites and counts RNA-seq reads mapping to those sites.[107]

### 7.3. Mutually Exclusive Exons

Mutually exclusive splicing means that exactly one exon of a cluster of neighboring exons is spliced into the mature transcript. Although mutually exclusive exons (MXEs) of a cluster are relatively similar, they cannot substitute each other if one is damaged. MXEs have been described in many crucial and essential human genes such as in the $\alpha$-subunits of six of the ten voltage-gated sodium channels (*SCN* genes), in each of the glutamate receptor subunits 1–4 (*GluR1-4*) in which the MXEs are called flip and flop and in *SNAP-25* as part of the neuroexocytosis machinery. Mutations in MXEs have been shown to cause diseases such as Timothy syndrome (missense mutation in the *CACNA1C* gene), cardiomyopathy (defect of the mitochondrial phosphate carrier *SLC25A3*), or cancer (mutations in, e.g., the pyruvate kinase *PKM* and the zinc transporter *SLC39A14*).

To explore the extent of mutually exclusive splicing in humans, we recently predicted 1722 completely novel exons in previously intronic regions in the human genome.[108] Our prediction algorithm is based on criteria derived from biological knowledge and has successfully been applied to plants, worm, and fruit fly before.[109,110] MXEs must be translated in same reading frames and splice sites must be compatible. We expect MXEs to have about the same length, because they code for the same structural region in the resulting protein, and length differences should only be possible in loop regions. Finally, the protein sequences coded by MXEs are supposed to be similar, because they code for the same region in the protein and evolved most

probably through exon duplication. Together with already annotated exons matching the aforementioned criteria, the mentioned 1722 newly predicted MXEs became part of a list of 6541 MXE candidates. By mapping 15 billion RNA-seq reads, representing 515 samples comprising 31 tissues and organs, 12 cell lines, and seven developmental stages, we could show that each novel exon is covered by at least one read. Applying strict criteria requiring reads bridging the respective other MXE and absence of reads joining MXEs, 1399 of the 6541 MXE candidates comprise high-confidence MXEs. This number is about tenfold higher than previous MXE estimates in human (ENCODE, e.g., reported only 14 MXEs in human and other analyses showed a maximum of 147). Mapping high-confidence MXEs onto known protein structures revealed further support for their annotation. The ends of MXE-encoded sequences preferentially match within secondary structural elements excluding that these exons can be spliced as constitutive or differentially included exons.

In order to assess the conservation and evolution of human MXEs across mammals, we identified orthologous proteins in 18 representative species from all major sub-branches spanning 180 million years and predicted MXEs therein.[108] Of the 554 human MXE clusters, 100 (18%) and 86 (15%) are shared between at least 15 and 16 of the 18 species, respectively. Conserved clusters include the annotated MXEs of sodium-channel *SCN* genes, *MAPK8* and *MAPK14*, glutamate-receptors *GRIA* genes, and *KCNMA1*. Dozens of genes contained novel exons such as collagen genes, members of the *SLIT* family of secreted glycoproteins, calcium-channel *CACNA1E*, and many genes of the solute-carrier family. MXEs are of high relevance in human diseases as an overlay of the set of high-confidence MXEs with the ClinVar database showed (35 of the MXEs contained 82 pathogenic SNPs). Disease-associated MXEs show tight developmental and tissue-specific expression with prominent selective expression in heart and brain and in cancer cell lines.

## 8. To Each Individual Its Own Annotation

Every individual has its own genome including parts conserved in families, differences at population level, and a collection of sequences distinguishing us from related hominins. The first two human genome assemblies represented examples of two extremes: a private genome sequencing effort reported the genome of an individual, John Craig Venter,[39] while the international human genome project generated a reference genome representing a mix of cell-lines and various donors.[40] The latest reference genome, GRCh38, provides representations for alternative loci that are alternative sequences found in largely haploid assemblies (earlier also termed "alternative alleles" and "alternative haplotypes").[111,112] However, this reference genome and corresponding reference genome annotation represent a rather artificial consensus of the "human genome." Genomic drift causes random duplication and deletion of genes, which means that every individual genome contains a significant amount of copy number variations (CNV). The first analysis of genomic drift in humans identified CNVs of sensory receptor genes among 270 individuals from the HapMap data demonstrating a difference of at least eleven olfactory receptor genes between randomly chosen individuals.[113] At the level of nations, sequencing and de novo assembly of 150 genomes from Denmark showed large differences at the chromosome scale.[114] Recently, the first pan-genome of a human population, the population of 910 humans of African descent, revealed a collection of sequences totaling about 300 million nucleotides that are not present in the human reference genome assembly but shared among multiple individuals of the African population.[115] Similarly, building and analyzing the pan-genome of Han Chinese detected 29.5 million novel nucleotides and at least 188 novel protein-coding genes.[116]

Human reference genome assembly and annotation are undoubtedly invaluable tools with respect to comparison and evaluation of annotations, tools, approaches, and conclusions. However, the available sequencing tool set should now allow combining analyses of genome, transcriptome, and proteome data at the individual level. Such an approach will also shed light on the reported discrepancies of alternatively spliced isoforms found in transcript and proteomics studies. We also anticipate that some of the loci annotated as pseudogenes in the reference annotation will become protein-coding genes in other annotations. Annotation of "alternate sequences" placed on alternative assembly units in RefSeq and Ensembl is a huge step forward (e.g., Ensembl release 96 contains 2960 coding genes on alternative sequences), but still the annotations are mixed and do not represent different populations, or even individuals.

## 9. Conclusions and Outlook: Are We Approaching Completeness of Human Genome Annotation?

Although the human genome assembly was declared to be completed in 2003, it has seen substantial changes since then. The current reference assembly has representations for alternative regions, from allelic differences and copy number variations and still there are gaps that need to be closed. These gaps (some of which have recently been closed) have a considerable influence on the annotation. For example, a gap of about 56 000 nucleotides within the dynein heavy chain gene *DHC7C* covering about 860 of the 3800 amino acids was closed only in the latest genome version, GRCh38, although contigs spanning the entire gene region were already present and in correct order in the J. C. Venter assembly.[117] Accordingly, the old gene annotations contained separate genes for *DHC7C* N- and C-terminus. These are still present as alternative transcripts in latest genome annotations, although they are artificial and will never result in folded proteins. A big step forward would be to not only search for novel transcripts but also to clean up old annotations. Therefore, it is highly likely that similar to finishing the genome assembly finishing reference genome annotation will also last for many years to come.

In addition to data provided by large consortia the scientific community highly profits from complementary efforts by, for example, CHESS in determining a set of consensus transcript sequences or by the Human Proteome Project and the Human Protein Atlas, which provide strong evidence for tissue-specific presence of proteins. Integrating many layers of evidence from ab initio predictions, RNA-seq data, splicing mechanism, protein structure, and evolutionary conservation is promising in case experimental data generation is limited by, e.g., protein abundance

or due to highly tissue- and developmental stage-specific expression. Such a multi-facet analysis revealed a detailed landscape of the mutually exclusive exome including not only the identification of hundreds of novel exons but also a re-evaluation of the splice type of hundreds of already known exons[108] and could be used as blueprint for the comprehensive analysis of other splice variants.

It has been pointed out already that the "human reference genome" generated from mostly a single individual is as bad a reference for human genomes as a genome of every other individual would be.[118] We suppose that the same will become true for the "human reference annotation", which will be substituted by annotations for individuals, groups, and populations. Genome annotations have many layers and differ from individual to individual and in many cases also from cell to cell. Mapping transcriptional, gene regulatory, and genetic variant data on a reference genome alone considerably limits the use and application of these data.[118] Deep learning approaches, which currently still incorporate steady genome annotations,[119,120] might resolve the issues of such individual annotations in the future.

## Conflict of Interest

The authors declare no conflict of interest. KH is employed by F. Hoffmann-La Roche Ltd.

## Keywords

alternative splicing, human genome annotation, human pan-genome, micro-exon, mutually exclusive exons, protein-coding genes, wobble splicing

Received: April 15, 2019
Revised: August 7, 2019
Published online: September 23, 2019

[1] S. Zhao, *PLoS One* **2014**, *9*, e101374.
[2] G. Chen, C. Wang, L. Shi, X. Qu, J. Chen, J. Yang, C. Shi, L. Chen, P. Zhou, B. Ning, W. Tong, T. Shi, *RNA* **2013**, *19*, 479.
[3] P.-Y. Wu, J. H. Phan, M. D. Wang, *BMC Bioinf.* **2013**, *14*, S8.
[4] S. Zhao, B. Zhang, *BMC Genomics* **2015**, *16*, 97.
[5] N. A. Naryshkin, M. Weetall, A. Dakka, J. Narasimhan, X. Zhao, Z. Feng, K. K. Y. Ling, G. M. Karp, H. Qi, M. G. Woll, G. Chen, N. Zhang, V. Gabbeta, P. Vazirani, A. Bhattacharyya, B. Furia, N. Risher, J. Sheedy, R. Kong, J. Ma, A. Turpoff, C.-S. Lee, X. Zhang, Y.-C. Moon, P. Trifillis, E. M. Welch, J. M. Colacino, J. Babiak, N. G. Almstead, S. W. Peltz, et al., *Science* **2014**, *345*, 688.
[6] M. Sivaramakrishnan, K. D. McCarthy, S. Campagne, S. Huber, S. Meier, A. Augustin, T. Heckel, H. Meistermann, M. N. Hug, P. Birrer, A. Moursy, S. Khawaja, R. Schmucki, N. Berntenis, N. Giroud, S. Golling, M. Tzouros, B. Banfai, G. Duran-Pacheco, J. Lamerz, Y. Hsiu Liu, T. Luebbers, H. Ratni, M. Ebeling, A. Cléry, S. Paushkin, A. R. Krainer, F. H.-T. Allain, F. Metzger, *Nat. Commun.* **2017**, *8*, 1476.
[7] ENCODE Project Consortium, *Nature* **2012**, *489*, 57.
[8] S. R. Eddy, *Curr. Biol.* **2012**, *22*, R898.
[9] W. F. Doolittle, *Proc. Natl. Acad. Sci. USA* **2013**, *110*, 5294.
[10] S. R. Eddy, *Curr. Biol.* **2013**, *23*, R259.
[11] D. Graur, Y. Zheng, N. Price, R. B. R. Azevedo, R. A. Zufall, E. Elhaik, *Genome Biol. Evol.* **2013**, *5*, 578.
[12] D.-K. Niu, L. Jiang, *Biochem. Biophys. Res. Commun.* **2013**, *430*, 1340.
[13] W. F. Doolittle, T. D. P. Brunet, S. Linquist, T. R. Gregory, *Genome Biol. Evol.* **2014**, *6*, 1234.
[14] M. Kellis, B. Wold, M. P. Snyder, B. E. Bernstein, A. Kundaje, G. K. Marinov, L. D. Ward, E. Birney, G. E. Crawford, J. Dekker, I. Dunham, L. L. Elnitski, P. J. Farnham, E. A. Feingold, M. Gerstein, M. C. Giddings, D. M. Gilbert, T. R. Gingeras, E. D. Green, R. Guigo, T. Hubbard, J. Kent, J. D. Lieb, R. M. Myers, M. J. Pazin, B. Ren, J. A. Stamatoyannopoulos, Z. Weng, K. P. White, R. C. Hardison, *Proc. Natl. Acad. Sci. USA* **2014**, *111*, 6131.
[15] D. Graur, Y. Zheng, R. B. R. Azevedo, *Genome Biol. Evol.* **2015**, *7*, 642.
[16] W. F. Doolittle, T. D. P. Brunet, *BMC Biol.* **2017**, *15*, 116.
[17] C. A. Thomas, *Annu. Rev. Genet.* **1971**, *5*, 237.
[18] T. D. P. Brunet, W. F. Doolittle, *Proc. Natl. Acad. Sci. USA* **2014**, *111*, E3365.
[19] M. B. Gerstein, C. Bruce, J. S. Rozowsky, D. Zheng, J. Du, J. O. Korbel, O. Emanuelsson, Z. D. Zhang, S. Weissman, M. Snyder, *Genome Res.* **2007**, *17*, 669.
[20] P.-L. Germain, E. Ratti, F. Boem, *Biol. Philos.* **2014**, *29*, 807.
[21] G. Giaever, A. M. Chu, L. Ni, C. Connelly, L. Riles, S. Véronneau, S. Dow, A. Lucau-Danila, K. Anderson, B. André, A. P. Arkin, A. Astromoff, M. El-Bakkoury, R. Bangham, R. Benito, S. Brachat, S. Campanaro, M. Curtiss, K. Davis, A. Deutschbauer, K.-D. Entian, P. Flaherty, F. Foury, D. J. Garfinkel, M. Gerstein, D. Gotte, U. Güldener, J. H. Hegemann, S. Hempel, Z. Herman, et al., *Nature* **2002**, *418*, 387.
[22] D.-U. Kim, J. Hayles, D. Kim, V. Wood, H.-O. Park, M. Won, H.-S. Yoo, T. Duhig, M. Nam, G. Palmer, S. Han, L. Jeffery, S.-T. Baek, H. Lee, Y. S. Shim, M. Lee, L. Kim, K.-S. Heo, E. J. Noh, A.-R. Lee, Y.-J. Jang, K.-S. Chung, S.-J. Choi, J.-Y. Park, Y. Park, H. M. Kim, S.-K. Park, H.-J. Park, E.-J. Kang, H. B. Kim, et al., *Nat. Biotechnol.* **2010**, *28*, 617.
[23] P. Sulem, H. Helgason, A. Oddson, H. Stefansson, S. A. Gudjonsson, F. Zink, E. Hjartarson, G. T. Sigurdsson, A. Jonasdottir, A. Jonasdottir, A. Sigurdsson, O. T. Magnusson, A. Kong, A. Helgason, H. Holm, U. Thorsteinsdottir, G. Masson, D. F. Gudbjartsson, K. Stefansson, *Nat. Genet.* **2015**, *47*, 448.
[24] A. G. Wouters, *Stud. Hist. Philos. Sci. Part C: Stud. Hist. Philos. Biol. Biomed. Sci.* **2003**, *34*, 633.
[25] J. N. Spuhler, *Science* **1948**, *108*, 279.
[26] H. J. Muller, *Am. J. Hum. Genet.* **1950**, *2*, 111.
[27] F. Vogel, *Nature* **1964**, *201*, 847.
[28] M. Pertea, S. L. Salzberg, *Genome Biol.* **2010**, *11*, 206.
[29] C. Willyard, *Nature* **2018**, *558*, 354.
[30] H. J. Muller, *Am. Nat.* **1966**, *100*, 493.
[31] *Understanding Our Genetic Inheritance: The U.S. Human Genome Project : The First Five Years: Fiscal Years 1991–1995*, U.S. Department of Health and Human Services, Public Health Service, National Institutes of Health, National Center for Human Genome Research; U.S. Department of Energy, Office of Energy Research, Office of Health and Environmental Research, Human Genome Program.
[32] F. Antequera, A. Bird, *Proc. Natl. Acad. Sci. USA* **1993**, *90*, 11995.
[33] C. Fields, M. D. Adams, O. White, J. C. Venter, *Nat. Genet.* **1994**, *7*, 345.
[34] M. D. Adams, A. R. Kerlavage, R. D. Fleischmann, R. A. Fuldner, C. J. Bult, N. H. Lee, E. F. Kirkness, K. G. Weinstock, J. D. Gocayne, O. White, *Nature* **1995**, *377*, 3.
[35] R. P. Wagner, M. P. Maguire, R. L. Stallings, *Chromosomes: A Synthesis*, Wiley, Hoboken, NJ **1993**.
[36] F. Liang, I. Holt, G. Pertea, S. Karamycheva, S. L. Salzberg, J. Quackenbush, *Nat. Genet.* **2000**, *25*, 239.

[37] B. Ewing, P. Green, *Nat. Genet.* **2000**, *25*, 232.
[38] H. Roest Crollius, O. Jaillon, A. Bernot, C. Dasilva, L. Bouneau, C. Fischer, C. Fizames, P. Wincker, P. Brottier, F. Quétier, W. Saurin, J. Weissenbach, *Nat. Genet.* **2000**, *25*, 235.
[39] J. C. Venter, M. D. Adams, E. W. Myers, P. W. Li, R. J. Mural, G. G. Sutton, H. O. Smith, M. Yandell, C. A. Evans, R. A. Holt, J. D. Gocayne, P. Amanatides, R. M. Ballew, D. H. Huson, J. R. Wortman, Q. Zhang, C. D. Kodira, X. H. Zheng, L. Chen, M. Skupski, G. Subramanian, P. D. Thomas, J. Zhang, G. L. Gabor Miklos, C. Nelson, S. Broder, A. G. Clark, J. Nadeau, V. A. McKusick, N. Zinder, et al., *Science* **2001**, *291*, 1304.
[40] E. S. Lander, L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh, R. Funke, D. Gage, K. Harris, A. Heaford, J. Howland, L. Kann, J. Lehoczky, R. LeVine, P. McEwan, K. McKernan, J. Meldrim, J. P. Mesirov, C. Miranda, W. Morris, J. Naylor, C. Raymond, M. Rosetti, R. Santos, A. Sheridan, C. Sougnez, *Nature* **2001**, *409*, 860.
[41] Gambling on the genome. *J. Natl. Cancer Inst.* **2000**, *92*, 1373.
[42] P. Smaglik, *Nature* **2000**, *405*, 264.
[43] J. M. Claverie, *Science* **2001**, *291*, 1255.
[44] R. D. Sleator, *Gene* **2010**, *461*, 1.
[45] Y. Huang, S.-Y. Chen, F. Deng, *Comput. Struct. Biotechnol. J* **2016**, *14*, 298.
[46] T. Steijger, J. F. Abril, P. G. Engström, F. Kokocinski, The RGASP Consortium, T. J. Hubbard, R. Guigó, J. Harrow, P. Bertone, *Nat. Methods* **2013**, *10*, 1177.
[47] R. Guigó, P. Flicek, J. F. Abril, A. Reymond, J. Lagarde, F. Denoeud, S. Antonarakis, M. Ashburner, V. B. Bajic, E. Birney, R. Castelo, E. Eyras, C. Ucla, T. R. Gingeras, J. Harrow, T. Hubbard, S. E. Lewis, M. G. Reese, *Genome Biol.* **2006**, *7*, S2.
[48] A. Nellore, A. E. Jaffe, J.-P. Fortin, J. Alquicira-Hernández, L. Collado-Torres, S. Wang, R. A. Phillips Iii, N. Karbhari, K. D. Hansen, B. Langmead, J. T. Leek, *Genome Biol.* **2016**, *17*, 266.
[49] C. Wilks, P. Gaddipati, A. Nellore, B. Langmead, *Bioinformatics* **2018**, *34*, 114.
[50] M. Pertea, A. Shumate, G. Pertea, A. Varabyou, F. P. Breitwieser, Y.-C. Chang, A. K. Madugundu, A. Pandey, S. L. Salzberg, *Genome Biol.* **2018**, *19*, 208.
[51] S. Kalyana-Sundaram, C. Kumar-Sinha, S. Shankar, D. R. Robinson, Y.-M. Wu, X. Cao, I. A. Asangani, V. Kothari, J. R. Prensner, R. J. Lonigro, M. K. Iyer, T. Barrette, A. Shanmugam, S. M. Dhanasekaran, N. Palanisamy, A. M. Chinnaiyan, *Cell* **2012**, *149*, 1622.
[52] X. Guo, M. Lin, S. Rockowitz, H. M. Lachman, D. Zheng, *PLoS One* **2014**, *9*, e93972.
[53] M. Clamp, B. Fry, M. Kamal, X. Xie, J. Cuff, M. F. Lin, M. Kellis, K. Lindblad-Toh, E. S. Lander, *Proc. Natl. Acad. Sci. USA* **2007**, *104*, 19428.
[54] I. Ezkurdia, D. Juan, J. M. Rodriguez, A. Frankish, M. Diekhans, J. Harrow, J. Vazquez, A. Valencia, M. L. Tress, *Hum. Mol. Genet.* **2014**, *23*, 5866.
[55] F. Abascal, D. Juan, I. Jungreis, L. Martinez, M. Rigau, J. M. Rodriguez, J. Vazquez, M. L. Tress, *Nucleic Acids Res.* **2018**, *46*, 7070.
[56] A. I. Nesvizhskii, *Nat. Methods* **2014**, *11*, 1114.
[57] J. C. Wright, J. Mudge, H. Weisser, M. P. Barzine, J. M. Gonzalez, A. Brazma, J. S. Choudhary, J. Harrow, *Nat. Commun.* **2016**, *7*, 11778.
[58] K. V. Ruggles, K. Krug, X. Wang, K. R. Clauser, J. Wang, S. H. Payne, D. Fenyö, B. Zhang, D. R. Mani, *Mol. Cell. Proteomics* **2017**, *16*, 959.
[59] B. Eraslan, D. Wang, M. Gusic, H. Prokisch, B. M. Hallström, M. Uhlén, A. Asplund, F. Pontén, T. Wieland, T. Hopf, H. Hahne, B. Kuster, J. Gagneur, *Mol. Syst. Biol.* **2019**, *15*, e8513.
[60] D. Wang, B. Eraslan, T. Wieland, B. Hallström, T. Hopf, D. P. Zolg, J. Zecha, A. Asplund, L.-H. Li, C. Meng, M. Frejno, T. Schmidt, K. Schnatbaum, M. Wilhelm, F. Ponten, M. Uhlen, J. Gagneur, H. Hahne, B. Kuster, *Mol. Syst. Biol.* **2019**, *15*, e8503.
[61] M. Uhlén, L. Fagerberg, B. M. Hallström, C. Lindskog, P. Oksvold, A. Mardinoglu, Å. Sivertsson, C. Kampf, E. Sjöstedt, A. Asplund, I. Olsson, K. Edlund, E. Lundberg, S. Navani, C. A.-K. Szigyarto, J. Odeberg, D. Djureinovic, J. O. Takanen, S. Hober, T. Alm, P.-H. Edqvist, H. Berling, H. Tegel, J. Mulder, J. Rockberg, P. Nilsson, J. M. Schwenk, M. Hamsten, K. von Feilitzen, M. Forsberg, et al., *Science* **2015**, *347*, 1260419.
[62] M. L. Tress, F. Abascal, A. Valencia, *Trends Biochem. Sci.* **2017**, *42*, 98.
[63] Y. Zhu, L. M. Orre, H. J. Johansson, M. Huss, J. Boekel, M. Vesterlund, A. Fernandez-Woodbridge, R. M. M. Branca, J. Lehtiö, *Nat. Commun.* **2018**, *9*, 903.
[64] K. Lindblad-Toh, M. Garber, O. Zuk, M. F. Lin, B. J. Parker, S. Washietl, P. Kheradpour, J. Ernst, G. Jordan, E. Mauceli, L. D. Ward, C. B. Lowe, A. K. Holloway, M. Clamp, S. Gnerre, J. Alföldi, K. Beal, J. Chang, H. Clawson, J. Cuff, F. Di Palma, S. Fitzgerald, P. Flicek, M. Guttman, M. J. Hubisz, D. B. Jaffe, I. Jungreis, W. J. Kent, D. Kostka, M. Lara, et al., *Nature* **2011**, *478*, 476.
[65] M. E. Dolan, R. M. Baldarelli, S. M. Bello, L. Ni, M. S. McAndrews, C. J. Bult, J. A. Kadin, J. E. Richardson, M. Ringwald, J. T. Eppig, J. A. Blake, *Mamm. Genome* **2015**, *26*, 305.
[66] S. König, L. W. Romoth, L. Gerischer, M. Stanke, *Bioinformatics* **2016**, *32*, 3388.
[67] J. Armstrong, I. T. Fiddes, M. Diekhans, B. Paten, *Annu. Rev. Anim. Biosci.* **2019**, *7*, 41.
[68] M. Freeling, J. Xu, M. Woodhouse, D. Lisch, *Mol. Plant* **2015**, *8*, 899.
[69] A. Frankish, B. Uszczynska, G. R. S. Ritchie, J. M. Gonzalez, D. Pervouchine, R. Petryszak, J. M. Mudge, N. Fonseca, A. Brazma, R. Guigo, J. Harrow, *BMC Genomics* **2015**, *16*, S2.
[70] T. Weirick, D. John, S. Uchida, *Brief. Bioinform.* **2017**, *18*, 226.
[71] K. D. Pruitt, J. Harrow, R. A. Harte, C. Wallin, M. Diekhans, D. R. Maglott, S. Searle, C. M. Farrell, J. E. Loveland, B. J. Ruef, E. Hart, M.-M. Suner, M. J. Landrum, B. Aken, S. Ayling, R. Baertsch, J. Fernandez-Banet, J. L. Cherry, V. Curwen, M. Dicuccio, M. Kellis, J. Lee, M. F. Lin, M. Schuster, A. Shkeda, C. Amid, G. Brown, O. Dukhanina, A. Frankish, J. Hart, et al., *Genome Res.* **2009**, *19*, 1316.
[72] P. Legrain, R. Aebersold, A. Archakov, A. Bairoch, K. Bala, L. Beretta, J. Bergeron, C. H. Borchers, G. L. Corthals, C. E. Costello, E. W. Deutsch, B. Domon, W. Hancock, F. He, D. Hochstrasser, G. Marko-Varga, G. H. Salekdeh, S. Sechi, M. Snyder, S. Srivastava, M. Uhlén, C. H. Wu, T. Yamamoto, Y.-K. Paik, G. S. Omenn, *Mol. Cell. Proteomics* **2011**, *10*, M111.009993.
[73] G. S. Omenn, L. Lane, C. M. Overall, F. J. Corrales, J. M. Schwenk, Y.-K. Paik, J. E. Van Eyk, S. Liu, M. Snyder, M. S. Baker, E. W. Deutsch, *J. Proteome Res.* **2018**, *17*, 4031.
[74] Z. Hu, H. S. Scott, G. Qin, G. Zheng, X. Chu, L. Xie, D. L. Adelson, B. E. Oftedal, P. Venugopal, M. Babic, C. N. Hahn, B. Zhang, X. Wang, N. Li, C. Wei, *Sci. Rep.* **2015**, *5*, 10940.
[75] B.-H. You, S.-H. Yoon, J.-W. Nam, *Genome Res.* **2017**, *27*, 1050.
[76] A. Frankish, M. Diekhans, A.-M. Ferreira, R. Johnson, I. Jungreis, J. Loveland, J. M. Mudge, C. Sisu, J. Wright, J. Armstrong, I. Barnes, A. Berry, A. Bignell, S. Carbonell Sala, J. Chrast, F. Cunningham, T. Di Domenico, S. Donaldson, I. T. Fiddes, C. García Girón, J. M. Gonzalez, T. Grego, M. Hardy, T. Hourlier, T. Hunt, O. G. Izuogu, J. Lagarde, F. J. Martin, L. Martínez, S. Mohanan, et al., *Nucleic Acids Res.* **2019**, *47*, D766.
[77] S. Light, A. Elofsson, *Curr. Opin. Struct. Biol.* **2013**, *23*, 451.
[78] B. Saudemont, A. Popa, J. L. Parmley, V. Rocher, C. Blugeon, A. Necsulea, E. Meyer, L. Duret, *Genome Biol.* **2017**, *18*, 208.
[79] Y. Wan, D. R. Larson, *Genome Biol.* **2018**, *19*, 86.
[80] W. F. Doolittle, *Genome Biol.* **2018**, *19*, 223.

[81] F. Abascal, I. Ezkurdia, J. Rodriguez-Rivas, J. M. Rodriguez, A. del Pozo, J. Vázquez, A. Valencia, M. L. Tress, *PLoS Comput. Biol.* **2015**, *11*, e1004325.
[82] M. Gonzàlez-Porta, A. Frankish, J. Rung, J. Harrow, A. Brazma, *Genome Biol.* **2013**, *14*, R70.
[83] M. L. Tress, F. Abascal, A. Valencia, *Trends Biochem. Sci.* **2017**, *42*, 408.
[84] B. J. Blencowe, *Trends Biochem. Sci.* **2017**, *42*, 407.
[85] S. A. Bhuiyan, S. Ly, M. Phan, B. Huntington, E. Hogan, C. C. Liu, J. Liu, P. Pavlidis, *BMC Genomics* **2018**, *19*, 637.
[86] D. Ustianenko, S. M. Weyn-Vanhentenryck, C. Zhang, *Wiley Interdiscip. Rev.: RNA* **2017**, *8*, e1418.
[87] N. Volfovsky, B. J. Haas, S. L. Salzberg, *Genome Res.* **2003**, *13*, 1216.
[88] F. Wen, F. Li, H. Xia, X. Lu, X. Zhang, Y. Li, *Trends Genet.* **2004**, *20*, 232.
[89] M. Irimia, R. J. Weatheritt, J. D. Ellis, N. N. Parikshak, T. Gonatopoulos-Pournatzis, M. Babor, M. Quesnel-Vallières, J. Tapial, B. Raj, D. O'Hanlon, M. Barrios-Rodiles, M. J. E. Sternberg, S. P. Cordes, F. P. Roth, J. L. Wrana, D. H. Geschwind, B. J. Blencowe, *Cell* **2014**, *159*, 1511.
[90] Y. I. Li, L. Sanchez-Pulido, W. Haerty, C. P. Ponting, *Genome Res.* **2015**, *25*, 1.
[91] M. Melé, P. G. Ferreira, F. Reverter, D. S. DeLuca, J. Monlong, M. Sammeth, T. R. Young, J. M. Goldmann, D. D. Pervouchine, T. J. Sullivan, R. Johnson, A. V. Segrè, S. Djebali, A. Niarchou, GTEx Consortium, F. A. Wright, T. Lappalainen, M. Calvo, G. Getz, E. T. Dermitzakis, K. G. Ardlie, R. Guigó, *Science* **2015**, *348*, 660.
[92] T. Gonatopoulos-Pournatzis, M. Wu, U. Braunschweig, J. Roth, H. Han, A. J. Best, B. Raj, M. Aregger, D. O'Hanlon, J. D. Ellis, J. A. Calarco, J. Moffat, A.-C. Gingras, B. J. Blencowe, *Mol. Cell* **2018**, *72*, 510.
[93] A. Parras, H. Anta, M. Santos-Galindo, V. Swarup, A. Elorza, J. L. Nieto-González, S. Picó, I. H. Hernández, J. I. Díaz-Hernández, E. Belloc, A. Rodolosse, N. N. Parikshak, O. Peñagarikano, R. Fernández-Chacón, M. Irimia, P. Navarro, D. H. Geschwind, R. Méndez, J. J. Lucas, *Nature* **2018**, *560*, 441.
[94] A. Torres-Méndez, S. Bonnal, Y. Marquez, J. Roth, M. Iglesias, J. Permanyer, I. Almudí, D. O'Hanlon, T. Guitart, M. Soller, A.-C. Gingras, F. Gebauer, F. Rentzsch, B. J. Blencowe, J. Valcárcel, M. Irimia, *Nat. Ecol. Evol.* **2019**, *3*, 691.
[95] M. Hiller, K. Huse, K. Szafranski, N. Jahn, J. Hampe, S. Schreiber, R. Backofen, M. Platzer, *Nat. Genet.* **2004**, *36*, 1255.
[96] K. Tadokoro, M. Yamazaki-Inoue, M. Tachibana, M. Fujishiro, K. Nagao, M. Toyoda, M. Ozaki, M. Ono, N. Miki, T. Miyashita, M. Yamada, *J. Hum. Genet.* **2005**, *50*, 382.
[97] C.-H. Lai, L.-Y. Hu, W. Lin, *Biochem. Biophys. Res. Commun.* **2006**, *342*, 197.
[98] M. Akerman, Y. Mandel-Gutfreund, *Nucleic Acids Res.* **2006**, *34*, 23.
[99] M. Hiller, K. Huse, K. Szafranski, P. Rosenstiel, S. Schreiber, R. Backofen, M. Platzer, *Genome Biol.* **2006**, *7*, R65.
[100] R. K. Bradley, J. Merkin, N. J. Lambert, C. B. Burge, *PLoS Biol.* **2012**, *10*, e1001229.
[101] A. Busch, K. J. Hertel, *Genome Biol.* **2012**, *13*, 143.
[102] T.-M. Chern, E. van Nimwegen, C. Kai, J. Kawai, P. Carninci, Y. Hayashizaki, M. Zavolan, *PLoS Genet.* **2006**, *2*, e45.
[103] M. Hiller, K. Szafranski, R. Backofen, M. Platzer, *PLoS Genet.* **2006**, *2*, e207; author reply e208.
[104] K.-W. Tsai, W.-C. Lin, *Genomics* **2006**, *88*, 855.
[105] R. Sinha, S. Nikolajewa, K. Szafranski, M. Hiller, N. Jahn, K. Huse, M. Platzer, R. Backofen, *Nucleic Acids Res.* **2009**, *37*, 3569.
[106] K. Szafranski, M. Kramer, *RNA Biol.* **2015**, *12*, 115.
[107] X. Yan, G. Sablok, G. Feng, J. Ma, H. Zhao, X. Sun, *FEBS Lett.* **2015**, *589*, 1766.
[108] K. Hatje, R.-U. Rahman, R. O. Vidal, D. Simm, B. Hammesfahr, V. Bansal, A. Rajput, M. E. Mickael, T. Sun, S. Bonn, M. Kollmar, *Mol. Syst. Biol.* **2017**, *13*, 959.
[109] H. Pillmann, K. Hatje, F. Odronitz, B. Hammesfahr, M. Kollmar, *BMC Bioinf.* **2011**, *12*, 270.
[110] K. Hatje, M. Kollmar, *Nat. Commun.* **2013**, *4*, 2460.
[111] V. A. Schneider, T. Graves-Lindsay, K. Howe, N. Bouk, H.-C. Chen, P. A. Kitts, T. D. Murphy, K. D. Pruitt, F. Thibaud-Nissen, D. Albracht, R. S. Fulton, M. Kremitzki, V. Magrini, C. Markovic, S. McGrath, K. M. Steinberg, K. Auger, W. Chow, J. Collins, G. Harden, T. Hubbard, S. Pelan, J. T. Simpson, G. Threadgold, J. Torrance, J. M. Wood, L. Clarke, S. Koren, M. Boitano, P. Peluso, et al., *Genome Res.* **2017**, *27*, 849.
[112] M. Jain, S. Koren, K. H. Miga, J. Quick, A. C. Rand, T. A. Sasani, J. R. Tyson, A. D. Beggs, A. T. Dilthey, I. T. Fiddes, S. Malla, H. Marriott, T. Nieto, J. O'Grady, H. E. Olsen, B. S. Pedersen, A. Rhie, H. Richardson, A. R. Quinlan, T. P. Snutch, L. Tee, B. Paten, A. M. Phillippy, J. T. Simpson, N. J. Loman, M. Loose, *Nat. Biotechnol.* **2018**, *36*, 338.
[113] M. Nozawa, Y. Kawahara, M. Nei, *Proc. Natl. Acad. Sci. USA* **2007**, *104*, 20421.
[114] L. Maretty, J. M. Jensen, B. Petersen, J. A. Sibbesen, S. Liu, P. Villesen, L. Skov, K. Belling, C. Theil Have, J. M. G. Izarzugaza, M. Grosjean, J. Bork-Jensen, J. Grove, T. D. Als, S. Huang, Y. Chang, R. Xu, W. Ye, J. Rao, X. Guo, J. Sun, H. Cao, C. Ye, J. van Beusekom, T. Espeseth, E. Flindt, R. M. Friborg, A. E. Halager, S. Le Hellard, C. M. Hultman, et al., *Nature* **2017**, *548*, 87.
[115] R. M. Sherman, J. Forman, V. Antonescu, D. Puiu, M. Daya, N. Rafaels, M. P. Boorgula, S. Chavan, C. Vergara, V. E. Ortega, A. M. Levin, C. Eng, M. Yazdanbakhsh, J. G. Wilson, J. Marrugo, L. A. Lange, L. K. Williams, H. Watson, L. B. Ware, C. O. Olopade, O. Olopade, R. R. Oliveira, C. Ober, D. L. Nicolae, D. A. Meyers, A. Mayorga, J. Knight-Madden, T. Hartert, N. N. Hansel, M. G. Foreman, et al., *Nat. Genet.* **2019**, *51*, 30.
[116] Z. Duan, Y. Qiao, J. Lu, H. Lu, W. Zhang, F. Yan, C. Sun, Z. Hu, Z. Zhang, G. Li, H. Chen, Z. Xiang, Z. Zhu, H. Zhao, Y. Yu, C. Wei, *Genome Biol.* **2019**, *20*, 149.
[117] M. Kollmar, *Mol. Biol. Evol.* **2016**, *33*, 3249.
[118] X. Yang, W.-P. Lee, K. Ye, C. Lee, *Genome Biol.* **2019**, *20*, 104.
[119] Z. Zhang, Z. Pan, Y. Ying, Z. Xie, S. Adhikari, J. Phillips, R. P. Carstens, D. L. Black, Y. Wu, Y. Xing, *Nat. Methods* **2019**, *16*, 307.
[120] K. Jaganathan, S. Kyriazopoulou Panagiotopoulou, J. F. McRae, S. F. Darbandi, D. Knowles, Y. I. Li, J. A. Kosmicki, J. Arbelaez, W. Cui, G. B. Schwartz, E. D. Chow, E. Kanterakis, H. Gao, A. Kia, S. Batzoglou, S. J. Sanders, K. K.-H. Farh, *Cell* **2019**, *176*, 535.
[121] P. E. Griffiths, K. Stotz, *Theor. Med. Bioethics* **2006**, *27*, 499.
[122] H. Pearson, *Nature* **2006**, *441*, 398.
[123] ENCODE Project Consortium, E. Birney, J. A. Stamatoyannopoulos, A. Dutta, R. Guigó, T. R. Gingeras, E. H. Margulies, Z. Weng, M. Snyder, E. T. Dermitzakis, R. E. Thurman, M. S. Kuehn, C. M. Taylor, S. Neph, C. M. Koch, S. Asthana, A. Malhotra, I. Adzhubei, J. A. Greenbaum, R. M. Andrews, P. Flicek, P. J. Boyle, H. Cao, N. P. Carter, G. K. Clelland, S. Davis, N. Day, P. Dhami, S. C. Dillon, M. O. Dorschner, H. Fiegler, et al., *Nature* **2007**, *447*, 799.
[124] Y. He, C. Yuan, L. Chen, M. Lei, L. Zellmer, H. Huang, D. J. Liao, *Genes* **2018**, *9*, 40.
[125] X. Chen, J. R. Bracht, A. D. Goldman, E. Dolzhenko, D. M. Clay, E. C. Swart, D. H. Perlman, T. G. Doak, A. Stuart, C. T. Amemiya, R. P. Sebra, L. F. Landweber, *Cell* **2014**, *158*, 1187.
[126] M. Rassoulzadegan, V. Grandjean, P. Gounon, S. Vincent, I. Gillot, F. Cuzin, *Nature* **2006**, *441*, 469.
[127] P. Portin, *Hereditas* **2009**, *146*, 112.
[128] L. Perbal, *EMBO Rep.* **2015**, *16*, 777.
[129] P. Portin, A. Wilkins, *Genetics* **2017**, *205*, 1353.
[130] A. E. Mirsky, H. Ris, *J. Gen. Physiol.* **1951**, *34*, 451.

[131] F. Jacob, J. Monod, *J. Mol. Biol.* **1961**, *3*, 318.
[132] B. Lewin, *Genes IV*, Oxford University Press, Oxford, UK **1990**.
[133] F. Antequera, A. Bird, *Nat. Genet.* **1994**, *8*, 114.
[134] G. D. Schuler, M. S. Boguski, E. A. Stewart, L. D. Stein, G. Gyapay, K. Rice, R. E. White, P. Rodriguez-Tomé, A. Aggarwal, E. Bajorek, S. Bentolila, B. B. Birren, A. Butler, A. B. Castle, N. Chiannilkulchai, A. Chu, C. Clee, S. Cowles, P. J. Day, T. Dibling, N. Drouot, I. Dunham, S. Duprat, C. East, C. Edwards, J. B. Fan, N. Fang, C. Fizames, C. Garrett, L. Green, et al., *Science* **1996**, *274*, 540.
[135] M. Das, C. B. Burge, E. Park, J. Colinas, J. Pelletier, *Genomics* **2001**, *77*, 71.
[136] F. A. Wright, W. J. Lemon, W. D. Zhao, R. Sears, D. Zhuo, J. P. Wang, H. Y. Yang, T. Baer, D. Stredney, J. Spitzner, A. Stutz, R. Krahe, B. Yuan, *Genome Biol.* **2001**, *2*, research0025.1.
[137] J. E. Collins, M. E. Goward, C. G. Cole, L. J. Smink, E. J. Huckle, S. Knowles, J. M. Bye, D. M. Beare, I. Dunham, *Genome Res.* **2003**, *13*, 27.
[138] Z. Xuan, J. Wang, M. Q. Zhang, *Genome Biol.* **2002**, *4*, R1.
[139] F. Sanger, S. Nicklen, A. R. Coulson, *Proc. Natl. Acad. Sci. USA* **1977**, *74*, 5463.
[140] J. M. Heather, B. Chain, *Genomics* **2016**, *107*, 1.
[141] M. R. Brent, *Genome Res.* **2005**, *15*, 1777.
[142] M. Stanke, O. Schöffmann, B. Morgenstern, S. Waack, *BMC Bioinf.* **2006**, *7*, 62.
[143] M. R. Brent, *Nat. Biotechnol.* **2007**, *25*, 883.
[144] K. Hatje, B. Hammesfahr, M. Kollmar, *Nucleic Acids Res.* **2013**, *41*, W504.
[145] M. L. Metzker, *Nat. Rev. Genet.* **2010**, *11*, 31.
[146] J. Eid, A. Fehr, J. Gray, K. Luong, J. Lyle, G. Otto, P. Peluso, D. Rank, P. Baybayan, B. Bettman, A. Bibillo, K. Bjornson, B. Chaudhuri, F. Christians, R. Cicero, S. Clark, R. Dalal, A. Dewinter, J. Dixon, M. Foquet, A. Gaertner, P. Hardenbol, C. Heiner, K. Hester, D. Holden, G. Kearns, X. Kong, R. Kuse, Y. Lacroix, S. Lin, et al., *Science* **2009**, *323*, 133.
[147] D. Sharon, H. Tilgner, F. Grubert, M. Snyder, *Nat. Biotechnol.* **2013**, *31*, 1009.
[148] Y. Feng, Y. Zhang, C. Ying, D. Wang, C. Du, *Genomics, Proteomics Bioinf.* **2015**, *13*, 4.
[149] A. Byrne, A. E. Beaudin, H. E. Olsen, M. Jain, C. Cole, T. Palmer, R. M. DuBois, E. C. Forsberg, M. Akeson, C. Vollmers, *Nat. Commun.* **2017**, *8*, 16027.
[150] D. R. Garalde, E. A. Snell, D. Jachimowicz, B. Sipos, J. H. Lloyd, M. Bruce, N. Pantic, T. Admassu, P. James, A. Warland, M. Jordan, J. Ciccone, S. Serra, J. Keenan, S. Martin, L. McNeill, E. J. Wallace, L. Jayasinghe, C. Wright, J. Blasco, S. Young, D. Brocklebank, S. Juul, J. Clarke, A. J. Heron, D. J. Turner, *Nat. Methods* **2018**, *15*, 201.
[151] Á. Arzalluz-Luque, A. Conesa, *Genome Biol.* **2018**, *19*, 110.
[152] K. Karlsson, S. Linnarsson, *BMC Genomics* **2017**, *18*, 126.