biology

MDPI

Review

Review on the Computational Genome Annotation of Sequences Obtained by Next-Generation Sequencing

Girum Fitihamlak Ejigu and Jaehee Jung *

Department of Information and Communication Engineering, Myongji University, Yongin-si 17058, Gyeonggi-do, Korea; girumfitex@gmail.com

* Correspondence: jhjung@mju.ac.kr

Received: 21 August 2020; Accepted: 16 September 2020; Published: 18 September 2020

Simple Summary: Due to the development of high-throughput sequencing technologies, computational genome annotation of sequences has become one of the principal research area in computational biology. First, we reviewed comparative annotation tools and pipelines for both annotations of structures and functions, which enable us to comprehend gene functions and their genome evolution. Second, we compared genome annotation tools that utilize homology-based and ab initio methods depending on the similarity of sequences or the lack of evidences. Third, we explored visualization tools that aid the annotation process and stressed the need for the quality control of annotations and re-annotations, because misannotations may happen due to experimental errors or missed genes by preceding technologies. Finally, we highlighted how emerging technologies can be used in future annotations.

Abstract: Next-Generation Sequencing (NGS) has made it easier to obtain genome-wide sequence data and it has shifted the research focus into genome annotation. The challenging tasks involved in annotation rely on the currently available tools and techniques to decode the information contained in nucleotide sequences. This information will improve our understanding of general aspects of life and evolution and improve our ability to diagnose genetic disorders. Here, we present a summary of both structural and functional annotations, as well as the associated comparative annotation tools and pipelines. We highlight visualization tools that immensely aid the annotation process and the contributions of the scientific community to the annotation. Further, we discuss quality-control practices and the need for re-annotation, and highlight the future of annotation.

Keywords: structural annotation; functional annotation; ab initio annotation; homology-based annotation

1. Introduction

Next-Generation Sequencing (NGS) has facilitated the generation of vast amount of DNA sequence information from a broad array of lifeforms in amazingly short time [1]. However, information stored in each sequence needs to be extracted, to help us understand the organism itself and evolution in general. NGS has also made possible the investigation of the genetic bases of diseases and gene mapping through large scale screening of genome variation. Therefore, this information benefits processes such as genetic disorder diagnosis and drug design [2]. Annotation is a means of retrieving information encoded within the multitude of different sequence patterns of the four nucleotides (i.e., A, T, C and G). The term genome annotation has evolved from the annotation of protein-coding genes to include the annotation of single nucleotides on thousands of individual genomes. A successful annotation depends on the quality of the genome assembly. Several statistical methods are employed to describe the completeness and contiguity of an assembly [3]. Improvements in sequencing, including long-read [4] and linked-read [5] technology have made high-quality genome assemblies available

at lower prices. The availability of high-quality genome assemblies has provided a robust source for phylogenetic information, and this, in turn, has been leveraged to improve whole-genome alignments and annotations, which have heavily relied on models, from mice and humans [6].

Finding and identifying genes constitutes a big part of genome annotation. Therefore, a comprehensive and accurate gene discovery approach is crucial. It entails the application of multiple independent and complementary analysis tools and methods. The employed approaches should hence utilize information that is intrinsic, e.g., ab initio predictions, and extrinsic, including information on proteins and transcripts. Numerous software tools and methods have been developed to tackle the various problems associated with annotation, but the challenge continues, as the technology develops and the knowledge grows [7–9].

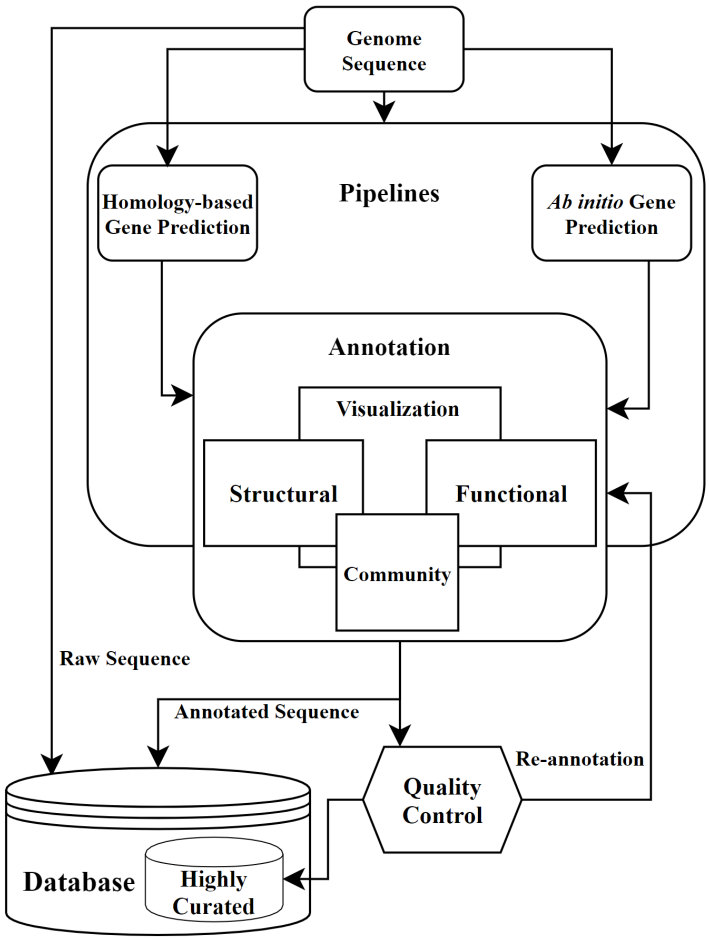

Figure 1. Genome annotation workflow.

In this paper, we summarize the current definitions and tools used for genome annotation. Most genome annotation tools require different types of input formats, and provide various types of outputs. Depending on the research environment, researcher can choose one of applications. We start by highlighting the structural and functional annotation processes, and commonly used programs. The goals of structural and functional annotation can be achieved through the analysis of sequence data, by exploiting either a statistical model approach (the ab initio method) or a sequence similarity technique (homology-based annotation), although these approaches are not mutually exclusive. In parallel, databases that play an integral part in annotation will be discussed. Annotation pipelines that aggregate ab initio and homology-based methods, together with other software components to generate well-annotated genomes, are presented. Even though our focus in this review concerns sequence annotation, we discuss how annotation plays a major role in identifying

a potential disease-causing gene or causal mutation. Therefore, databases and tools used for gene variant annotation are described as well. Annotation is not an easy task and visualization tools are useful in facilitating it. We hence explore different genome browsers that are used for gene structure and function predictions, as well as other visualization tools that aid the analysis of gene function. The annotation data should undergo quality checks, as errors can be easily propagated, and affect downstream annotation and analysis. A quality-check result may necessitate genomic re-annotation, which may go as far as re-sequencing, sometimes discarding the original version. We conclude by examining the future possibilities of annotation. The entire annotation workflow following sequence assembly is summarized in Figure 1. This workflow can also be considered as a graphic summary of this review.

2. Types of Annotation

2.1. Structural Annotation

Finding features of DNA—exons, introns, promoters, transposons, etc.—is known as structural annotation. While structural annotation attempts to find genes in a genomic sequence, gene definition has evolved with the advances in modern genomics. A gene can be defined as "a sequence region necessary for generating functional products" [10]. Functional products of genes are proteins and RNAs. Genes that lead to the production of proteins are called protein-coding genes. Other genes that do not code proteins, but instead functional RNA molecules, are called noncoding genes. Noncoding RNA genes include genes for ribosomal RNA (rRNA), transfer RNA (tRNA), microRNA (miRNA), small nuclear RNA and nucleolar RNA (snRNA and snoRNA, respectively) [11] and long noncoding RNA (lncRNA). Structural annotations also identify pseudogenes. They were initially considered to be functionless and evolutionary dead-ends. We now know that they sometimes participate in gene regulation [12]. Hence, their prediction improves our understanding of genomes.

2.1.1. Repeats

The first step in structural annotation involves repeat masking. DNA repeats occur in both prokaryotic and eukaryotic organisms. The repeats account for 0% to over 42% of the prokaryotic genome [13]. Similarly, eukaryotic genomes can harbor millions of repeats. For instance, repeats account for two-thirds of the human genome [14]. Repeat sequences can be localized in tandem, i.e., adjacent to one other, and are typically found in the centromere [15]. Alternatively, they can be interspersed in different forms of transposable elements, e.g., in long and short interspersed nuclear elements (LINEs and SINEs), DNA transposons, etc. [16]. Identification of the essential features of repetitive elements is still challenging, despite advances in repeat identification. Repeat masking tools rely on databases with lists of already identified repeats. RepeatMasker [17] is a good example of such tool.

Aligning transcript and protein evidence after masking is the second step of structural annotation before gene identification, although it is not mandatory. BLAST [18] or BLAT [19] can be used to align the transcript and protein evidence. Further, RNA-seq evidence can be aligned using TopHat [20] or HISAT [21].

2.1.2. Predictions of Gene and Different Features

Identifying protein-coding genes and other regulatory elements takes center stage in gene annotation. Gene prediction is a complex process, especially for eukaryotic DNA [3]. The varying sizes of introns (noncoding sequences) in-between exons and alternative splice variants make gene structure prediction difficult. Many gene prediction programs exist. They can be categorized into three groups: ab initio methods, homology-based methods, and combined methods. Approaches for gene prediction based on nucleotide sequence are called ab initio methods. Ab initio approaches rely on statistical models, such as the hidden Markov model (HMM), to identify promoters, coding or noncoding regions, and intron–exon

junctions in the genome sequence. The second approach aligns the sequence with expressed sequence tags (EST), complementary DNA (cDNA), or protein evidence, and uses detected similarities for gene prediction. The other group comprises programs that combine ab initio and evidence- or homology-based approaches for gene prediction [22]. In addition, gene prediction programs should be able to predict alternative splicing sites because alternative splicing is a major actor in the regulation of gene expression, and transcriptome and proteome diversity [23]. Accordingly, gene prediction programs use various models to predict splice sites. Since approximately 99% of the introns in sequenced genomes begin with GT and end with AG, these features are denoted as mandatory by most gene prediction systems for splice site detection. In addition, incorporation of a strong splice donor consensus, such as the GC–AG splice site, improves the accuracy of gene prediction programs [24]. Commonly used gene prediction programs and their classification, based on the above discussion, are listed in Table 1.

Table 1. Commonly used gene prediction programs.

Method	Program	Description	URL	Ref
Ab initio	EasyGene	HMM-based automatic gene predictor for prokaryotes that ranks open reading frames (ORFs) by statistical significance	https://services.healthtech.dtu.dk/service.php?EasyGene-1.2	[25]
	FGENESH	HMM-based gene structure prediction	http://www.softberry.com/berry.phtml?topic=fgenesh&group=programs&subgroup=gfind	[26]
	GeneMark	A family of self-training gene prediction programs for bacteria, archaea, metagenomes, metatranscriptomes and eukaryotes	http://opal.biology.gatech.edu/GeneMark/	[27]
	GeneZilla	Generalized hidden Markov model (GHMM) eukaryotic gene finder (formerly known as TIGRscan)	http://www.genezilla.org/	[28]
	GenScan	Algorithm for ab initio prediction of complete gene structures in vertebrate, *Drosophila*, and plant genomic sequences	http://hollywood.mit.edu/GENSCAN.html	[29]
	GlimmerHMM	GHMM-based eukaryotic gene finder that incorporates splice sites from GeneSplicer and decision tree from GlimmerM in Unix environment	http://ccb.jhu.edu/software/glimmerhmm/	[30]
	HMMgene	HMM-based gene predictor for vertebrates and *C. elegans*, full as well as partial genes	https://services.healthtech.dtu.dk/service.php?HMMgene-1.1	[31]
	mGene	Web service for predicting eukaryotic gene structures, including protein-coding genes and untranslated region (UTR) with pre-trained models	https://galaxy.inf.ethz.ch/tool_runner?tool_id=mgenepredict	[32]
	NetGene	Predicts splice sites in human, *C. elegans* and *A. thaliana* DNA	https://services.healthtech.dtu.dk/service.php?NetGene2-2.42	[33]
	RNAmmer	A two level HMM-based predictor of rRNA genes in full genome sequences	http://www.cbs.dtu.dk/services/RNAmmer/	[34]
	SNAP	Semi-HMM general-purpose gene finding program suitable for both eukaryotic and prokaryotic genomes	https://github.com/KorfLab/SNAP	[35]
	tRNAscan-SE	A covariance model-based program that provides genomic coordinates, predicted function, and secondary structure of tRNA genes	http://lowelab.ucsc.edu/tRNAscan-SE/	[36]

Table 1. *Cont.*

Method	Program	Description	URL	Ref
Homology	GeMoMa	A program that uses annotated genes to infer protein-coding genes in a target genome	http://galaxy.informatik.uni-halle.de/	[37]
	GenomeThreader	Uses cDNA, EST and protein sequences to predict gene structures via spliced alignments	http://genomethreader.org/	[38]
	PPFINDER	Identifier of processed pseudogenes incorporated in mammalian genome annotation	https://mblab.wustl.edu/software.html	[39]
	PseudoPipe	A computational pipeline that searches a mammalian genome and identifies pseudogene sequences	http://www.pseudogene.org/pseudopipe/	[40]
	TWINSCAN	GenScan extension, gene structure prediction system that exploits homology of related genomes	https://mblab.wustl.edu/software.html	[41]
Combined	AUGUSTUS	An ab initio gene prediction program that can also incorporate extrinsic sources, e.g., EST alignment, protein alignments and syntetic genome alignments	http://bioinf.uni-greifswald.de/augustus/	[42]
	JIGSAW	Gene model predictor that combines outputs from other gene finders, splice site predictors, and sequence alignments	http://www.cbcb.umd.edu/software/jigsaw/	[43]

2.1.3. Databases for Structural Annotation

Annotations require supporting data that can be used or presented as evidence of predicted assignments. Currently, homology-based methods play a central role in genome annotation because of the huge amount of EST and cDNA sequences available [44]. Homology-based methods depend on DNA, RNA, or protein sequence alignment data, which can easily be retrieved from biological databases. Ab initio annotations, on the other hand, identify genes and their structures using mathematical models. Nonetheless, the ab initio gene predictors have to be trained using high-quality gene models or organism-specific genome traits, such as codon frequency and intron–exon length distribution [45]. Further, ab initio models require ESTs, RNA-seq data, and proteins to improve prediction accuracy. Databases readily provide such data.

Nucleotide and protein sequence or structure can easily be found in comprehensive public-domain databases, e.g., the GenBank [46], European Nucleotide Archive (ENA) [47], and DNA Databank of Japan (DDBJ) [48]. UniProt [49], which is a protein sequence database that combines UniProtKB/Swiss-Prot (over 560,000 manually curated sequences) and UniProtKB/TrEMBL (180 million automatically annotated sequences), provides the scientific community with high-quality and freely accessible protein sequences with the associated functional information. Another great database for protein annotation is InterPro [50], which provides information on protein families, domains, and important sites such as binding sites, active sites, conserved sites, and repeats. The InterPro Consortium has 14 member databases, including Pfam [51], PROSITE [52], TIGRFAM [53], CATH-Gene3D [54], and PANTHER [55].

In addition, some specialized databases are built as comprehensive one-stop points of information on specific topics of interest. For example, databases, such as NONCODE [56], Pseudogene.org [57], Dfam [58], and miRbase [59] have contributed to the structural annotation of noncoding RNAs, pseudogenes, transposable elements, and microRNAs, respectively.

2.2. Functional Annotation

Association of biological information with gene or protein sequences identified by structural annotation is called functional annotation. Protein-coding genes were the focus of traditional functional annotation. However, the many different functions of noncoding genes and untranslated transcripts

are currently recognized. The term "functional" has a very different meaning to evolutionary biologists, who are interested in conservation, and experimental biologists, who are interested in biochemical roles [60]. However, the basic portion of functional annotation involves the association of a functional description with a gene, after identifying a similar sequence using tools, such as BLAST.

Functional annotation is also employed to assess the variation in genes. Annotation of genomic variants is an increasingly important and complex part of the analysis of sequence-based genomic analysis. The goal of variant annotation is to identify and prioritize variants based on their functional impacts [61]. Molecular impacts of genetic variants on phenotypes can be understood by assigning structural and functional knowledge of genomic sequences to variants [62]. Gene variation can happen because of a single nucleotide change in between members of same biological species' genomic DNA, called a single nucleotide polymorphism (SNP), or by structural rearrangements such as insertion, deletion, translocation, and inversion. Insertion and deletion cause variation known as copy number variations (CNVs) [63]. The functional relevance of variants can be explored in databases that have functional annotation information of known and novel variants.

2.2.1. Automatic Functional Annotation

Although manual annotation is still considered as the gold standard, this approach is difficult to scale. This necessitates the use of automated annotation methods, to scale up to match the plethora of genomic data currently being generated with NGS technology. Automatic function prediction can be achieved directly, by using local alignment tools, such as BLAST, where a protein database is searched for high-scoring alignments. The function is then assigned to the unknown query sequence based on a known result sequence, provided that it is the highest scoring alignment from all sequences, above some specified threshold value. The assumption of function transfer on which BLAST-like tools rely is that the function is retained in proteins that have similar sequences and have evolved from a single ancestor. In other words, the tools identify evolutionary relationships by discovering orthologous and paralogous relations between sequences. Orthologs are genes that have originated from a single ancestral gene in the last common ancestor of the compared genomes, whereas paralogs are genes within the same genome that have arisen from duplications [64]. Local alignment-based functional annotations are simple to use and perform well in many cases. However, they have some drawbacks. Examples include database source error, relativity of the alignment threshold, low sensitivity/specificity, and excessive transfer of annotation from an unrelated local region of similarity [65].

2.2.2. Databases for Functional Annotation

Gene Ontology (GO)

The GO resource is the most comprehensive and widely used knowledgebase for gene function [66]. GO covers three aspects of gene function: the molecular function (activity of a gene product at the molecular level), the cellular component (location of the gene product), and the biological process (a biological program, in which a gene's function is used). Gene products (proteins and RNA) should be consistently described to allow a comprehensive coverage of biological concepts. The GO Consortium tries to address this need by developing and maintaining ontology standards, annotating gene products, and developing and maintaining tools to do so [67]. The standard GO annotation comprises gene, GO term, and scientific evidence elements. These only reflect a partial functional description because single GO term annotations represent minimal knowledge determined from few experiments. Hence, a model concept called GO Causal Activity Modeling (GO-CAM) was introduced in 2018 to extend the existing annotation to represent a complex statement that can be scalable and structured [68]. Multiple standard GO annotations are linked to larger models of biological functions by GO-CAM in a semantically structured manner. The explicit relationship between the molecular function, biological process, and cellular component of each gene function is also defined by GO-CAM.

This results in improved quality and consistency of GO annotations [69]. Additional functional databases, other than the GO database, exist. Nonetheless, the GO database is quite popular and different tools have been developed to use its rich ontologies for annotation. Table 2 lists useful functional annotation tools.

Table 2. Ontology based annotation tools.

Program	Description	URL	Ref
BLAST2GO	A comprehensive bioinformatics tool for functional annotation of sequences and data mining on annotation results	https://www.blast2go.com/	[70]
FastAnnotator	An integration of well-established annotation tools for annotation of transcripts, which assigns GO terms, enzyme commission numbers, and functional domains	http://fastannotator.cgu.edu.tw/	[71]
GO FEAT	Homology-based functional annotation tool for genomic and transcriptomic data	http://computationalbiology.ufpa.br/gofeat/	[72]
GOtcha	A method that predicts gene product function by annotation with GO terms	http://www.compbio.dundee.ac.uk/gotcha/gotcha.php	[73]
PANNZER2	A fully automated service for functional annotation of prokaryotic and eukaryotic proteins of unknown function that provides both GO annotations and free text description predictions	http://ekhidna2.biocenter.helsinki.fi/sanspanz/	[74]
PoGO	A statistical pattern recognition method that assigns GO terms for fungal proteins	http://bioinformatica.vil.usal.es/lab_resources/pogo/	[75]

Kyoto Encyclopedia of Gene and Genomes (KEGG) [76] acts as a link between genomic data and higher-order functional information, which are stored in the GENES and PATHWAY databases, respectively. It affords understanding of high-level functions and utilities of the biological system, such as the cell, organism, and ecosystem, from genomic- and molecular-level information. The KEGG Orthology (KO) database links genes to high-level functions [77]. The KO system is the basis for genome annotation and KEGG mapping, which replaced the EC number that linked genomes to metabolic pathways.

The Reactome Pathway Knowledgebase [78] focuses on *Homo sapiens*, linking human proteins to their molecular processes. The Reactome data model presents molecular details and processes of signal transduction, transport, DNA replication, and metabolism as an ordered network of molecular transformations. Reaction is the core unit in the Reactome data model. Nucleic acids, proteins, and other molecules participate in reactions, forming a network of interactions, and are grouped into pathways, such as metabolism, regulation, and disease. The recent addition of a new drug class to the database extended the annotation process to human diseases [79].

Rhea [80] enables the functional annotation of enzymes and description of metabolic pathways based on an expert-curated non-redundant resource of biochemical reactions.

ChEBI [81] contains manually curated data of chemical entities that are classified into two sub-ontologies. The chemical entity ontology classification is based on common structural features and the role ontology considers activities in biological and chemical systems, or applicability.

NCBI's Conserved Domain Database (CDD) [82] comprises protein domains conserved during molecular evolution, and provides conserved domain footprints, along with conserved functional site annotations of protein sequences. Evolutionarily conserved domains help transfer the functional annotation from a known domain model to protein sequence. The Conserved Domain Architecture Retrieval Tool (CDART) groups proteins into superfamilies, while the Subfamily Protein Architecture Labeling Engine (SPARCLE) groups them according to subfamily domain architectures.

The Database of Genomic Variants (DGV) [83] provides a comprehensive summary of genomic variations (structural variations) that are larger than 50 bp (base pairs) from DNA segments of the

human genome. It contains structural variations identified in healthy control samples and provides a useful catalog of control data for studies aiming to correlate genomic variation with phenotypic data.

dbVar [84] is a human genomic structural variation database from the NCBI. It contains more than six million submitted structural variants and has the same data model as DGV that allows the implementation of standardized terminology. For the functional analysis of SNPs, the NCBI also provides another database called the dbSNP [85].

The Human Gene Mutation Database (HGMD) [86] constitutes information about gene mutations associated with human inherited disease and functional SNPs.

HGVbase [87] is a human sequence variation database that has high quality and non-redundant variation data and it mostly comprises SNPs.

The International Genome Sample Resource (IGSR) [88] is an extension of the 1000 Genomes Project [89] data, which served as a reference set of human variation. It maintains and updates to 1000 Genomes Project resources to the GRCh38 (Genome Reference Consortium human assembly). The web-based portal includes samples that were not part of the 1000 Genome Project and presents a unified view of data from multiple studies.

3. Comparative Annotation Methods

Genome annotation achieved by comparison of genes and genomes across species can be a reliable information source for understanding genome evolution. Comparative annotation allows annotations of a well-studied genome to be projected onto an evolutionarily close species. It often focuses on the coding genes. Valuable information for comparative annotation can be found from genome alignment. A well-aligned genome will yield sound data for comparative annotation [90]. Approaches to comparative annotation of genomes can be categorized into ab initio methods and homology-based methods, considering the input information used for annotation, i.e., either a statistical model of genes, or protein sequence, EST, and cDNA, accordingly. Ab initio approaches are preferred for genes that are weakly or not at all represented in RNA-seq library and have insufficient similarity to any known protein and lack other evidence.

Several comparative annotation methods have also been developed for variant calling purposes [91]. Like sequence annotation, variant calling annotation starts with alignment against a reference genome. Various tools exist to perform the variant calling and they produce a variant calling format (VCF) file for further downstream analysis.

3.1. Ab Initio Annotation

Ab initio annotation relies on ab initio gene predictors, which in turn rely on training data to construct an algorithm or model. Prediction is done based on the genomic sequence in question, using statistical analysis and other gene signals such as k-mer statistics and frame length. Some popular ab initio gene predictors are discussed below.

AUGUSTUS [42] defines the probability distributions for eukaryotic genome sequences based on GHMM. AUGUSTUS is re-trainable and it can predict alterative splicing, and the 5′UTR and 3′UTR, including introns. AUGUSTUS is one of the most accurate ab initio gene prediction programs for the species it has been trained for [92].

FGENESH [93] is an HMM-based, very fast, and accurate ab initio gene structure prediction program for humans, *Drosophila*, plants, yeasts, and nematodes. When applied to single-gene sequences, FGENESH predicts approximately 93% of all coding exon bases, as well as 80% of human exons, in 1.5 min. This renders it the fastest tool among HMM-based gene finding programs [26].

GENSCAN [29] is another HMM-based ab initio tool for predicting locations and exon–intron structures of genes in genomic sequences of a variety of organisms. Vertebrate and invertebrate versions of GENSCAN are available. The accuracy of the latter is lower because the original tool was primarily designed for the detection of genes in human and vertebrate genomic sequences.

It is becoming a common practice to use ab initio annotation methods in combination with transcriptome information such as that provided by RNA-seq. [60], particularly for higher eukaryotes. This can be viewed as an evidence-based or extrinsic approach. For example, a newer version of AUGUSTUS can incorporate information from EST and protein alignments. In addition, a variant of FGENESH called FGENESH-C [94] uses HMM and cDNA for predictions, while GenomeScan (an extension of GENSCAN) uses extrinsic information of protein BLAST alignments [95] for gene structure prediction.

3.2. Homology-Based Annotation

According to the molecular evolution principle, the rate of evolution of functionally important portions of the genome is slower than the rest of the cellular molecular regions. Hence, gene sequences that are useful for survival and other crucial functions are conserved [96], especially in closely related species. Homology-based annotations exploit this fact, to predict and annotate genes by identifying significant matches from a well annotated genome sequence by employing alignment tools such as BLAST. Homology-based annotations use the coding sequences (CDS), usually protein sequences and sometimes transcripts in the form of mRNA, cDNA, or EST to predict genes, assuming similar sequence regions encode homologous proteins. Tools like Exonerate [97] and DIALIGN [98] can be used for sequence alignment; GenomeThreader [38] and AGenDA [99] are used for gene predictions. Increased evolutionary distance between the input protein and the target protein reduces the accuracy of homology-based gene finding [41]. This happens because of heavy reliance on the alignment and information derived from the already known genes, which creates a challenge in identifying genes whose properties are different from those of referenced genes. However, newer comparative approaches solve this issue by relying to a greater degree on sequence conservation, which enables them to identify genes with new features and different statistical composition. TWINSCAN [41] and SGP2 [100] are examples of tools in which gene prediction uses the analysis of sequence conservation patterns between genomic sequences of evolutionarily related organisms [101]. Additional gene predictors used for homology-based annotations are listed in Table 1.

Table 3 below summarizes and compares ab initio and homology-based annotations discussed above.

Table 3. Ab initio and Homology-based annotation tools summary.

	Gene Prediction	**Source of Data**	**Evolutionary Distance Effect**	**Strength**
Ab initio	Rely on statistical model and gene signal	Models (HMM, GHMM, WAM) that can be trained supervised or unsupervised	Medium	Fast and easy means to identify and novel genes
Homology	Rely on sequence alignment	Proteins, EST, cDNA	High	Better accuracy, suitable for functional annotations

3.3. Variant Annotation

SnpEff [102] annotates and predicts the effects of variants on genes. It categorizes the effects of SNPs and other variants such as multiple nucleotide polymorphisms (MNPs). SnpEff accepts predicted variants in VCF and annotates the variants plus the effects (e.g. amino acid changes) they produce on known genes.

Ensembl Variant Effect Predictor (VEP) [103] performs annotation and analysis on genomic variants of coding and noncoding regions. It is a flexible tool for the identification of genes and transcripts affected by variants along with their location.

GEMINI [104] is a framework that allows exploring all forms of human genetic variation. GEMINI integrates genetic variation with diverse genome annotation from databases such as dbSNP and KEGG. It accepts a VCF file automatically and annotates by comparing with annotation sources.

SeattleSeq [105] provides annotation of SNVs and small indels, both known and novel. The annotations include dbSNP reference tags, gene names and accession numbers, variation functions, protein positions and amino acid changes, conservation scores, and clinical associations.

SNPnexus [106] is a web-based solution for functional annotation of novel and public domain variations. It allows assessing the potential significance of variants from broad range of annotation categories.

4. Annotation Pipelines

Analysis of large amounts of data generated by the NGS requires multiple computationally-intensive steps [107]. Sets of algorithms that process NGS data and are executed in a predefined order are called a bioinformatic pipelines. Pipelines process massive amounts of sequence data and the associated metadata using multiple software components, databases, and environments [108]. They are comprehensive, holistic packages that try to exploit relevant information provided by both ab initio and similarity-based gene predictors.

4.1. Structural Pipelines

MAKER2 [109] is a multi-threaded, parallelized genome annotation and data management application, which builds up on MAKER [110]. MAKER2 is designed for second-generation genome projects, which lack pre-existing gene models to train gene finders, but it also performs well with first-generation genome projects. Ab initio gene prediction tools SNAP, AUGUSTUS, and GenMark-ES are integrated in MAKER2. Novel genomes with limited training data available can be annotated with MAKER2. The tool can also be used to improve annotation quality by integrating mRNA-seq data. Further, it can be used to update legacy annotations.

NCBI Eukaryotic Annotation Pipeline [111] is an automated pipeline for eukaryotes, in which coding and noncoding genes, transcripts, and proteins in both finished and draft genomes can be annotated. This pipeline uses Splign [112] and ProSplign for alignment. It also has its own gene prediction tool called GNOMON which combines HMM-based ab initio models and homology search information extracted from experimental evidence.

Comparative Annotation Toolkit (CAT) [113] is a fully open-source software toolkit for end-to-end annotation. CAT uses Progressive Cactus [114] for multiple alignments. Its output, together with previously annotated genomes, is used to project annotations using TransMAP [115]. CAT uses AUGUSTUS for gene prediction both from transMap projections and for ab initio gene prediction, integrating extrinsic information from RNA-seq and Iso-Seq transcripts. All sources of transcript evidence are combined in an annotation set using a consensus-finding algorithm within CAT. CAT was developed by the GENCODE [116], and was utilized for the annotation of genomes of laboratory mouse strains [117] and great apes [118].

BRAKER1 [119] is a fully automated and highly accurate unsupervised RNA-seq–based genome annotation pipeline for eukaryotic genomes. It merges the complementary strengths of GeneMark-ET [120], which generates initial ab initio gene structure predictions via unsupervised iterative training of unassembled RNA-seq reads, and AUGUSTUS, which uses the predicted genes as a training set, to predict genes, utilizing mapped unassembled RNA-seq read information. BRAKER1-based gene predictions are more accurate than those of MAKER2, when RNA-seq data only are used. BRAKER1 has been expanded to integrate cross-species proteins, along with RNA-seq and protein alignment data, as heterogeneous extrinsic evidence. Furthermore, it is capable of whole-genome annotation [121].

4.2. Functional Pipelines

Prokka [122] is Unix-based command line software that can be used for rapid annotation of prokaryotic genomes. It identifies the coordinates of genomic features within contigs using external feature prediction tools, such as RNAmmer and Prodigal [123]. Prokka uses a hierarchical method for annotation, starting with a smaller trustworthy database; then, it moves to a medium-sized domain-specific database, and finally to curated models of protein families, including Pfam and TIGRFAMs.

Rapid Annotations using Subsystems Technology (RAST) [124] is a fully automated pipeline for bacterial and archaeal genome annotation. It achieves accuracy, consistency, and completeness by utilizing a library of subsystems, which are functional roles (abstract protein function) that implement a specific biological process or structural complex [125], together with protein families derived from subsystems. Gene function assertions are made based on both subsystem recognition of functional variants called "subsystem-based assertions" and integration of evidence from different tools called "nonsubsystem-based assertions".

4.3. Combined Pipelines

NCBI Prokaryotic Genome Annotation Pipeline (PGAP) [126] is an aggregation of alignment-based methods with a specialized search tool and ab initio gene finding tool called GeneMarkS+. GeneMarkS+ (a new self-training gene finder that is an extension of GeneMarkS [127] developed for use in PGAP) integrates alignment-based protein predictions, RNA predictions, and other extrinsic information with intrinsic information on genome-specific sequence patterns of protein-coding regions. PGAP uses statistical gene predictions when external evidence is insufficient and capitalizes on sequence similarity if enough comparative data are available.

DFAST [128] is a prokaryotic genome annotation pipeline that supports genome submission to the public database DDBJ. DFAST uses GHOSTX algorithm [129] for homology search with referenced databases. LAST [130] (an adaptive local alignment tool that enables fast and sensitive comparison of large sequences) is used for pseudogene detection by re-aligning CDSs with their subject protein sequences. An additional tool called hmmscan [131] searches for profile HMMs against TIGRFAM database. Clusters of Orthologous Groups (COG) from NCBI are searched for the assignment of COG categories using RPS-BLAST [132]. The DFAST workflow supports both structural and functional annotations, which are implemented as a module with common interfaces allowing flexible annotation.

Genome Sequence Annotation Serve (GenSAS) [133] is an online pipeline that provides both structural and functional annotations for eukaryotic and prokaryotic genomes. In addition to annotations, the GenSAS pipeline enables repeat identification and masking, evidence alignment, optional manual editing of gene models, and creation of final annotation files. It uses more than 25 tools for gene prediction, alignment, and annotation, and integrates some genome browsers.

4.4. Variant Pipelines

ANNOVAR [134] annotates SNPs and CNVs and examines their functional consequences on genes. It also performs genomic region-based annotation and compares variants to variation databases. ANNOVAR can evaluate and filter out variants against user dataset or variants that are not reported in public databases. To address personal genome annotation by biologists and clinicians, a web server called wANNOVAR [135] was developed using ANNOVAR as a backend annotation engine.

AnnoGen [136] allows the annotation of chemical binding energy, sequence information entropy, and homology score features for the GRCh38 framework.

annotatr [137] provides genomic annotations and set of functions to read, intersect, summarize, and visualize genomic regions in the context of genomic annotations. It is a Bioconductor package that provides insights into how characteristics of the region differ across annotations.

5. Annotation Visualization

5.1. File Formats

Most bioinformatic tools use the FASTA format as a standard for sequence data sharing. The FASTA format is used for searching sequence databases, evaluating similarity scores, and identification of periodic similarity scores. The format can also be used to compare a protein sequence with information in a DNA sequence database, with the DNA database translated while the search is performed [138]. Nevertheless, FASTA is a simple data file format that cannot handle all the information that might be added in the course of an annotation. Other standard file formats exist that can accommodate additional information and can be used by different programs, and interpreted by human users. The most common of these are the GenBank file format of NCBI, DDBJ format of DDBJ, EMBL format of ENA, and general feature format (GFF) and GTF.

The GFF [139] especially has become the de facto reference format for annotations. It stores genomic features in a standard text file format. Its new extended GFF3 format is a nine-column tab-delimited plain text file that addresses deficiencies of the previous versions GFF2/GTF. GFF3 allows flexibility, which enables storage of a wide variety of information. It is widely used for data exchange and genomic data representation.

5.2. Genome Browsers

Annotation yields gene structure, gene function, gene expression, regulation, variation, and additional information by employing multiple tools and information sources. Researchers and users utilize genome browsers to integrate various types of information, as well as analyze and visualize data related to annotation. Genome browsers are usually used to efficiently and conveniently browse, search, retrieve, and examine genomic sequence and annotation data, via a graphical interface [140]. They have been deployed since the initial sequence set generated by the Human Genome Project [141]. Although some standalone genome browsers exist, most genome browsers are web-based and can be classified as general and species-specific genome browsers.

General genome browsers host multiple richly annotated genomes from different species and enable comparative analysis. The UCSC Genome Browser [142] is the most commonly used genome browser; many visualization tools are modeled based on this tool. Although its user base focuses on human and mouse research, the UCSC genome browser database, which was founded in 2001, currently hosts more than 105 different species. The Ensembl genome browser [143] is another widely used genome browser for vertebrate genomes, which supports comparative genomics, sequence variation analysis, and transcriptional regulation analysis. The NCBI Genome Data Viewer (GDV), previously known as the Map Viewer, is another browser that supports visualized exploration and analysis of eukaryotic NCBI's Reference sequence (RefSeq) genome assemblies. Nonetheless, generalized genome browsers cannot handle diverse analyses and the increasing customized visualization in each species-specific area.

Species-specific genome browsers focus on one model organism and help to visualize genomic, epigenomic, and transcriptomics data for that specific organism. As an example, Wormbase [144], Flybase [145], and MaizeGDB [146] provide species-specific browsers, which are based on the GBrowse framework of Generic Model Organism Database (GMOD).

GMOD is a collection of interconnected open-source software tools and databases for managing, visualizing, storing, and sharing genetic and genomic information. The most popular component of GMOD is GBrowse (a generic genome browser) [147], which is a web application for displaying genomic annotations and other features. Customizable design has made GBrowse suitable to act as a building block for databases for many model organisms, including Wormbase, Flybase, and many more. A very fast and scalable successor of GBrowse is the JBrowse genome browser [148]. JBrowse is built with JavaScript and HTML5, and can run standalone analyses or can be embedded in a website. The functionality of JBrowse is greater than that of GBrowse, with greater speed and responsiveness,

and click-and-drag navigation, including same-screen track selection. JBrowse is the next-generation of this genome browser, constantly expanded by data migrated from other databases. WebApollo, or simply Apollo [149], is a plugin for JBrowse for viewing and manual annotation of genomes. It allows real-time collaborative editing. Apollo enables concurrent editing by multiple users via WebSockets, which are supported by most web browsers. GMOD software tools have been designed, developed, and tested by many developers, scientists, and laboratories over the years, and have high demands by biologists on account of their interconnectedness . In general, the GMOD project is directed by its user base, who are mostly biologists.

JBrowse was the first genome browser that utilized client-side technology for retrieving and processing data through in-browser JavaScript programming. This enabled the user to cease to entirely rely on a web-server to preprocess data. ABrowse [150] enhanced and extended the interactivity of the JBrowse model by allowing access to more data sources and enabling non-real time commenting and annotation. Likewise, Genome Maps [151], ChromoZoom [152], and PBrowse [153] also try to improve the user experience, with a focus on implementing improved web technologies to handle high-volume data.

5.3. Functional Analysis Visualization Tools

Functional annotation of large gene sets (gene lists) is the final step of omics data analysis. It serves for the identification of transcriptional networks, the building of predictive models, and the discovery of candidate biomarkers. This differential analysis is challenging because of the high-dimensional nature of functional gene profiles derived from multiple experiments. Multiple tools are available for graphic representation and analysis of enriched functional annotations. First, explanatory data analysis methods are used to reveal the structure, and then statistical methods are applied to detect biological process patterns. Mapping molecules to biological annotations is a common approach using hierarchical structures of terms from the KEGG pathways, Reactome pathways, and GO terms. The majority of available tools are web services or are implemented in R. We discuss some such tools below.

Database for Annotation, Visualization and Integrated Discovery (DAVID) [154] is a program that facilitates the functional annotation and analysis of large gene lists. DAVID is linked to rich biological annotation sources, which facilitate biological discovery by biochemical pathway mapping, functional classification, and analysis of conserved protein domain architectures. DAVID combines annotation with graphical representation and produces tabular output with query-based access to functional annotation. In addition, DAVID can cluster redundant annotation terms, explore gene names in batch, and execute gene-enrichment analysis, particularly for GO terms.

g:Profiler [155] is a tool for functional enrichment analysis and additional information mining. The web server analyzes gene lists for enriched features, converts different class gene identifiers, maps genes to orthologous genes, and searches similarly expressed genes in public microarray datasets. g:Profiler uses gene annotations and identifiers from Ensembl, and ontologies from the GO website.

GOPlot [156] is an R package for functional analysis that follows deductive reasoning. GOPlot generates a visual representation (plot), from a general identification of most enriched categories to detailed molecule displays, in a specified set of categories.

FunMappOne [157] accepts input of gene lists and modifications and enables a graphical visualization of enriched terms. The output is provided with interactive navigation. Over-represented biological terms from GO, KEGG, or Reactome datasets are evaluated statistically by the functionalities offered by FunMappOne, to graphically summarize and navigate them within super-classes. FunMappOne exploits hierarchical structure of functional annotations of KEGG, GO, and Reactome and homogenizes them to offer three levels of summarization, from terms-root.

Gene Annotation Easy Viewer (GAEV) [158] has been developed to construct the complete set of molecular pathways for non-model species using resources at KEGG, i.e., by integrating KO annotation

and KEGG pathway mapping. GAEV software can be run on Windows and Linux machines, and it provides gene function summaries and the association of molecular pathways with genes.

5.4. Other Visualization Tools

Visualization plays a major role in displaying the finalized records of organellar (mitochondrial and plastid) genomes. Organellar genomes are the focus of taxonomic relationship studies because they are inherited from single parent and abundant in a cell. These genomes are small, informative, and can be easily sequenced. OGDRAW [159] is currently the standard tool for the generation of graphical maps of organellar genomes. GeneBank file formats are used as an input, and graphical maps of both circular and linear genomes can be drawn using this tool. Visualization of coding regions and other feature-bearing regions, together with gene expression data and cut sites of restriction enzymes can be displayed in OGDRAW. Organellar annotation programs, such as AGORA [160], which uses BLAST-based homology searches for organellar annotation from user or NCBI database reference, and GeSeq [161], an annotator for organellar genomes (particularly for the chloroplast), which identifies genes (by BLAT-based homology search), proteins (by HMM search), and rRNA-coding genes (by de novo prediction), use OGDRAW to visually display the annotation output.

An aesthetically appealing tool called Circos [162] displays genomic interval relationships in a circular ideogram layout. It facilitates the identification and analysis of similarities and differences in large volumes of genomic data. Circos plots are widely used in various genomics studies to demonstrate various genomic data, such as gene rearrangements [163]. However, its use is limited by the installation and command line-based usage. Several tools with different objectives have been developed to address these issues, including CircosVCF [164], MISTIC [165], J-Circos [166], and shinyCircos [167]. Although it is mainly used as a multiple genome alignment tool that identifies conserved regions, rearrangements, and inversions, MAUVE [168] has a simple viewing system that can display structural rearrangements of genomes. The Mauve rearrangement viewer has an interactive feature that enables searching and zooming into regions of interest in aligned genomes. A similar interactive online tool that is used for the display, manipulation, and annotation of phylogenetic trees is the Interactive Tree of Life (iTOL) [169]. This tool can be used to represent phylogenetic trees in several tree formats, including the circular (radial) mode.

6. Community Annotation and Quality Control in Annotation

6.1. Community Annotation

The advent of NGS technologies has resulted in a large volume of sequencing data and turned the research focus toward genome annotation. Gene databases, such as Ensembl and GenBank, and model organism databases, such as FlyBase provide annotations. However, these sites are authoritative because of the high degree of oversight by expert curators [170]. Continuous sequence annotation is a challenge, especially when only a limited number of professional annotators are available for the databases mentioned above. As an alternative, community engagement helps to deal with annotation bottlenecks. In 2001, Lincoln [171] proposed four models to describe the "sociology of genome annotation." The first model is the so-called "Factory model," which involves a high degree of automated genome analysis for finding genes and identifying structural landmarks. Genome browsers usually use this model. The second model, "museum model," focuses on the functional roles of genes and requires considerable manual input from expert curators, which makes it a good choice for model organism analyses. The third model is the "cottage industry model," and involves decentralized effort from curators at different laboratories. The last model, the "party" or "jamboree model," assembles expert biocurators for a specific time period, usually a week. This model has been famously used for the annotation of *Drosophila melanogaster* [172] and cDNA annotation of the mouse genome [173]. Additional models, the "blessed annotator" and "gatekeeper approach," were presented in 2012 by the Wellcome Trust Sanger Institute [174]. The Blessed annotator is a variation of the Museum approach

and is used for the Knockout Mouse Project (KOMP), while the Gatekeeper approach is an extension and refinement of the party and cottage industry models, and is used for the analysis of data for multiple species.

Community annotations can take forms different from the ones described above. For instance, for supervised dispersed-community annotations, experts in a field annotate specific items in response to request from a coordinator [175]. Currently, a community annotation jamboree can take place virtually. One variant of the annotation jamboree is student community annotation, where students are taught during a class or workshop about annotation, and then perform the annotation. This is a mutually beneficial scheme for both the students' education and genomic resources. Another type of community annotation, which requires the least engagement, is unsupervised dispersed-community annotation or Gene Wiki (Wikis), where anyone can login and annotate an entry of choice, and which is based on the open data model of Wikipedia.

6.2. Quality Control for Annotation

Quality of annotation is directly affected by the input genome sequence quality. Although NGS technologies have enabled the generation of sequences in cost effective and short period, they produce reads, ranging from dozens to thousands of consecutive bases, that should be assembled to make a complete sequence. Hence, assessing quality of a sequence assembly is vital before subsequent annotation. Tools such as MaGuS [176], QUAST [177], and BUSCO [178] can be employed to check the quality, contiguity and completeness of genome assemblies.

Manual annotation requires considerable infrastructure and specific tools, which make it costly. Nonetheless, it provides an accurate gene set, which serves as a solid reference for a variety of studies. Manual curation has been held as a gold standard for functional annotation, but newer automatic systems might perform as well as teams of sequence-annotating experts [179].

Automated systems are necessary for meeting the challenge of extracting information from the mountain of genomes generated by sequencing [180]. Annotation-scoring schemes for automatic annotation methods are needed to allow the reduction of the cost of genome annotation. One proposed method that uses a genome comparison approach is the annotation confidence score (ACS) [181]. ACS attempts to combine sequence and textual similarity to denote the quality of annotation. The quality score is derived from a summary of sequence homology, taxonomic distance, and textual similarity analysis. Another example is a semi-automated genome annotation comparison and integration scheme [182]. This scheme compares annotations and arrives at a consensus annotation based on the comparison outcome. An automated tool for Bacterial Genome Annotation Comparison (BEACON) [183] similarly enables a fully automated, simple, and quick comparison of genome annotations generated by multiple annotation methods. It yields analytical results that are comprehensive and informative. Functional annotation of prokaryotic genomes obtained by different annotation methods can be compared in BEACON, and it can be used to combine and extend annotations from different annotation methods.

Annotation edit distance (AED) is a different measure for annotation comparisons, which aims to evaluate changes across annotation releases [184]. It was introduced in 2009. It complements a similar measure, called Annotation Turnover, which tracks the addition and deletion of gene annotations between releases. AED determines structural changes to an annotation, such as alternative splicing, which cannot be reported by using conventional measures, such as sensitivity and specificity. AED is used as a quality-control measure in MAKER2, with some adaptation. MAKER2 uses AED to show alignment between a gene and the supporting evidence used. An AED of 0 in MAKER2 indicates a perfect match between the intron–exon coordinates of annotation and the used evidence, such as EST, protein, and mRNA-seq data. On the other hand, an AED of 1 indicates no evidence-based support. As implemented in MAKER2, AED can be used as a quality-measure tool. This was confirmed by investigations of the annotations of human and mouse genomes from RefSeq, which revealed an agreement between AED scores and domain contents in Pfam. In addition, the International Nucleotide

Sequence Database Collaboration (INSDC) [185] has designed quality control procedures to be used in annotation pipelines, such as NCBI's PGAP. The quality control matrices within PGAP are generated automatically, facilitating annotation submission to GenBank.

RefSeq is a highly curated collection of annotated genomes and transcripts that is widely used as a reference for genome projects and different analysis tools and is considered to contain high-quality annotations [186]. It is a collection of comprehensive and non-redundant, explicitly linked genomes, transcripts, and protein sequence records, with publications, informative nomenclature, and standardized and expanded annotations available. Quality assurance checks for different data types are applied to all RefSeq data. This renders the RefSeq data consistent and allows it to serve as a baseline for multiple gene-specific reporting and cross-species comparisons. For annotation, the RefSeq dataset uses both computational methodology and manual curation by NCBI scientific staff. The RefSeq dataset is freely accessible. Its 201st release contains data on more than 103,000 organisms and can be accessed using NCBI's nucleotide and protein databases, BLAST databases, and through FTP.

7. Re-Annotation and Future of Annotation

7.1. Re-Annotation

We have seen that as a result of the increasing volume of data from genome sequencing projects, computational analysis methods have become a considerable element of genome annotations. However, this has led to high levels of misannotation in public databases [187,188]. Since annotations are used as a resource in other annotation projects, researchers have to be presented with high-quality data. To ensure such high-quality data, NCBI and other sequencing centers have developed international annotation standards [189]. While re-annotation is crucial for correcting some missannotations [190], other main motivations for re-annotation are the discovery of new genes or protein functions, comparison of new and existing annotation methods, and assessment of annotation reproducibility [191]. Re-annotation benefits the end-user by providing the latest resources. Updating and re-annotating genome annotations is necessary for the provision of accurate and relevant information, because the knowledge of gene products is expanded each day by downstream research, such as comparative genomics, transcriptomics, proteomics, and metabolomics. Updating a previously annotated genome can be seen as re-annotation [192]. Automated annotations save time and resources, but manual annotations, although time-consuming, are better than automated annotations. Hence, aggregating and comparing multiple automated annotations side-by-side, followed by manual curation, will greatly reduce subsequent error propagation. Annotations of many of the first-generation genomes published were limited because of limited information and small numbers of references. Currently, because of the falling cost of sequencing, new assemblies, other evidence (such as RNA sequencing), and other genome technologies, these genomes are being re-annotated and updated [193–195].

In some ways, comparative studies are becoming more difficult because of the diverse annotation strategies and updates. Re-annotation can be used to create large complete genomes, and indeed, there are tools that can be used for this purpose. Restauro-G [196] is rapid bacterial genome re-annotation software that utilizes a BLAST-like alignment tool for re-annotation. MAKER2 incorporates an external annotation pass-through mechanism that accepts pre-existing genome annotations and aligned experimental evidence in GFF3 format as an input. This mechanism allows annotations from reference genomes to be done over the legacy annotation and creates a non-redundant consensus dataset after merging. Although annotations must be recomputed using the latest software and databases, there are no standard means to do this.

Yet another proposed approach is the Wiki solution, which is an open-editing framework for websites and data, where anyone can edit a shared resource [197]. Wiki-based sites have been proven successful in providing accurate, useful, and updated information, despite the fear of being filled

with unreliable and inaccurate data. Currently, new information emerges from different corners of bioinformatic fields, which impacts gene annotation, rendering re-annotation a never-ending process, to some degree.

7.2. The Future of Annotation

The scope of genome annotation has expanded since the first complete annotation of the *Haemophilus influenzae* genome in 1995 [198]. The scope has widened to include information about noncoding RNAs [199], promoters and enhancers [200], pseudogenes [201], and many other features. Annotation will keep expanding, as the sequencing technology and knowledge related to genomics continues to evolve. Direct sequencing of RNA using Oxford Nanopore technology has been recently introduced [202]. Indeed, long-read RNA sequences improved human poly(A) RNA isoform characterization and allele specificity analysis. The method addresses the loss of information in high-throughput complementary DNA sequencing, which frequently copies biological RNA as short reads. Although the nanopore RNA sequencing technology is in its infant stage, it may soon allow low-cost sequencing of full transcripts. It could thus play a major role in future annotations [203]. As another suggestion for the future of annotation [204], the standard automated annotation practice relies on the majority rule that follows "the sequence tells the structure tells the function" stance, which hinders progress, because it generates and propagates errors. This inductive process discourages the discovery of novelty. It therefore argues for annotation systems that support multiple models of inference, such as deductive and abductive (trial and error method) inferences, in addition to the inductive processes used.

A different perspective on the future of annotation is the anticipation of multiple dimensions in characterizing genome-scale function [205]. From this perspective, identification of genes and assigning their functionality is considered to be a one-dimensional genome annotation, while specifying the cellular component and their interactions is a two-dimensional annotation. Three-dimensional annotation considers the effects of cellular packing and localization, i.e., the intracellular arrangement of chromosomes and other cell components. A fourth dimension would be the investigation of changes driven by adaptive evolution. While only one- and two-dimensional annotations are currently feasible, the higher-dimension annotations listed above and beyond should be possible in the future.

A popular computational approach that can be applied in annotation is machine learning. Machine learning constructs a mathematical model for a specific concept and identifies data patterns. This property is useful for genome annotation. Machine learning has already been implemented in finding functional elements in the human genome via unsupervised learning [206]. Machine-learning methods can be used to integrate multiple and heterogeneous datasets by applying complex functions, but the lack of training examples and the context specificity of models poses some challenges [207]. Although annotation tasks, such as protein function predictions [208] are challenging for machine-learning models, these models are likely to play a big role in future annotations considering the constant increase in the available data.

8. Conclusions

Understanding the structure of a gene is a crucial step in comprehending its function and the significance of variations. Computational annotation approaches, such as ab initio and homology-based annotations, enable such endeavors to be carried out automatically. Using automatic annotation systems and pipelines is imperative, considering the large amounts of sequence data generated by NGS. The importance of genome annotation ranges from answering in-depth questions about evolution to current applications, such as diagnosis of genetic disorders and drug design. This necessitates annotation quality-control, as errors can be easily propagated downstream. Quality-control methods and community annotations will help in avoiding such errors. Further, re-annotation is needed to correct faulty annotations or to update older annotations and even in some cases of well-studied genes,

to identify novel features that were missed by preceding technologies. These factors frame annotation as an incessant journey, as new perspectives and technologies emerge every day.

Author Contributions: Conceptualization, G.F.E. and J.J.; writing—original draft preparation, G.F.E.; writing—review and editing, J.J.; supervision, J.J. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT and Future Planning (NRF-2019R1A2C1084308).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Mardis, E.R. The impact of next-generation sequencing technology on genetics. *Trends Genet.* **2008**, *24*, 133–141. [CrossRef]
2. Steward, C.A.; Parker, A.P.J.; Minassian, B.A.; Sisodiya, S.M.; Frankish, A.; Harrow, J. Genome annotation for clinical genomic diagnostics: Strengths and weaknesses. *Genome Med.* **2017**, *9*, 49. [CrossRef] [PubMed]
3. Yandell, M.; Ence, D. A beginner's guide to eukaryotic genome annotation. *Nat. Rev. Genet.* **2012**, *13*, 329–342. [CrossRef] [PubMed]
4. English, A.C.; Richards, S.; Han, Y.; Wang, M.; Vee, V.; Qu, J.; Qin, X.; Muzny, D.M.; Reid, J.G.; Worley, K.C.; et al. Mind the gap: Upgrading genomes with Pacific Biosciences RS long-read sequencing technology. *PLoS ONE* **2012**, *7*, e47768. [CrossRef]
5. Weisenfeld, N.I.; Kumar, V.; Shah, P.; Church, D.M.; Jaffe, D.B. Direct determination of diploid genome sequences. *Genome Res.* **2017**, *27*, 757–767. [CrossRef] [PubMed]
6. Armstrong, J.; Fiddes, I.T.; Diekhans, M.; Paten, B. Whole-genome alignment and comparative annotation. *Annu. Rev. Anim. Biosci.* **2019**, *7*, 41–64. [CrossRef]
7. Brent, M.R. Genome annotation past, present, and future: How to define an ORF at each locus. *Genome Res.* **2005**, *15*, 1777–1786. [CrossRef]
8. Li, F.; Zhao, X.; Li, M.; He, K.; Huang, C.; Zhou, Y.; Li, Z.; Walters, J.R. Insect genomes: Progress and challenges. *Insect Mol. Biol.* **2019**, *28*, 739–758. [CrossRef]
9. Mishra, S.; Rastogi, Y.P.; Jabin, S.; Kaur, P.; Amir, M.; Khatoon, S. A bacterial phyla dataset for protein function prediction. *Data Brief* **2020**, *28*, 105002. [CrossRef]
10. Spieth, J.; Lawson, D. Overview of gene structure. *Genome Biol. Evol.* **2005**, doi:10.1895/wormbook.1.65.1. [CrossRef]
11. Zhang, B.; Han, D.; Korostelev, Y.; Yan, Z.; Shao, N.; Khrameeva, E.; Velichkovsky, B.M.; Chen, Y.P.P.; Gelfand, M.S.; Khaitovich, P. Changes in snoRNA and snRNA abundance in the human, chimpanzee, macaque, and mouse brain. *Genome Biol. Evol.* **2016**, *8*, 840–850. [CrossRef] [PubMed]
12. Xiao, J.; Sekhwal, M.K.; Li, P.; Ragupathy, R.; Cloutier, S.; Wang, X.; You, F.M. Pseudogenes and their genome-wide prediction in plants. *Int. J. Mol. Sci.* **2016**, *17*, 1991. [CrossRef] [PubMed]
13. Treangen, T.J.; Abraham, A.L.; Touchon, M.; Rocha, E.P.C. Genesis, effects and fates of repeats in prokaryotic genomes. *FEMS Microbiol. Rev.* **2009**, *33*, 539–571. [CrossRef] [PubMed]
14. de Koning, A.P.J.; Gu, W.; Castoe, T.A.; Batzer, M.A.; Pollock, D.D. Repetitive elements may comprise over two-thirds of the human genome. *PLoS Genet* **2011**, *7*, e1002384. [CrossRef]
15. Barra, V.; Fachinetti, D. The dark side of centromeres: Types, causes and consequences of structural abnormalities implicating centromeric DNA. *Nat. Commun.* **2018**, *9*, 1–17. [CrossRef]
16. Bourque, G.; Burns, K.H.; Gehring, M.; Gorbunova, V.; Seluanov, A.; Hammell, M.; Imbeault, M.; Izsvák, Z.; Levin, H.L.; Macfarlan, T.S.; et al. Ten things you should know about transposable elements. *Genome Biol.* **2018**, *19*, 1–12. [CrossRef]
17. Smit, A.F.; Hubley, R.; Green, P. RepeatMasker, 1996. 4.1.1 Released. Available online: http://www.repeatmasker.org/ (accessed on 3 September 2020)
18. Altschul, S.F.; Gish, W.; Miller, W.; Myers, E.W.; Lipman, D.J. Basic local alignment search tool. *J. Mol. Biol.* **1990**, *215*, 403–410. [CrossRef]
19. Kent, W.J. BLAT—the BLAST-like alignment tool. *Genome Res.* **2002**, *12*, 656–664. [CrossRef]

20. Kim, D.; Pertea, G.; Trapnell, C.; Pimentel, H.; Kelley, R.; Salzberg, S.L. TopHat2: Accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* **2013**, *14*, R36. [CrossRef]
21. Kim, D.; Langmead, B.; Salzberg, S.L. HISAT: A fast spliced aligner with low memory requirements. *Nat. Methods* **2015**, *12*, 357–360. [CrossRef]
22. Yu, Y.; Santat, L.A.; Choi, S. Bioinformatics packages for sequence analysis. In *Applied Mycology and Biotechnology*; Elsevier: Amsterdam, The Netherlands, 2006; Volume 6, pp. 143–160.
23. Modrek, B.; Lee, C. A genomic view of alternative splicing. *Nat. Genet.* **2002**, *30*, 13–19. [CrossRef] [PubMed]
24. Brent, M.R.; Guigo, R. Recent advances in gene structure prediction. *Curr. Opin. Struct. Biol.* **2004**, *14*, 264–272. [CrossRef] [PubMed]
25. Larsen, T.S.; Krogh, A. EasyGene—A prokaryotic gene finder that ranks ORFs by statistical significance. *BMC Bioinform.* **2003**, *4*, 21. [CrossRef] [PubMed]
26. Solovyev, V.; Kosarev, P.; Seledsov, I.; Vorobyev, D. Automatic annotation of eukaryotic genes, pseudogenes and promoters. *Genome Biol.* **2006**, *7*, S10. [CrossRef] [PubMed]
27. Besemer, J.; Borodovsky, M. GeneMark: Web software for gene finding in prokaryotes, eukaryotes and viruses. *Nucleic Acids Res.* **2005**, *33*, W451–W454. [CrossRef] [PubMed]
28. Majoros, W.H.; Pertea, M.; Delcher, A.L.; Salzberg, S.L. Efficient decoding algorithms for generalized hidden Markov model gene finders. *BMC Bioinform.* **2005**, *6*, 1–13.
29. Burge, C.; Karlin, S. Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.* **1997**, *268*, 78–94. [CrossRef]
30. Majoros, W.H.; Pertea, M.; Salzberg, S.L. TigrScan and GlimmerHMM: Two open source ab initio eukaryotic gene-finders. *Bioinformatics* **2004**, *20*, 2878–2879. [CrossRef]
31. Krogh, A. Two methods for improving performance of an HMM and their application for gene finding. *Cent. Biol. Seq. Analysis. Phone* **1997**, *45*, 4525.
32. Schweikert, G.; Zien, A.; Zeller, G.; Behr, J.; Dieterich, C.; Ong, C.S.; Philips, P.; De Bona, F.; Hartmann, L.; Bohlen, A.; et al. mGene: Accurate SVM-based gene finding with an application to nematode genomes. *Genome Res.* **2009**, *19*, 2133–2143. [CrossRef]
33. Hebsgaard, S.M.; Korning, P.G.; Tolstrup, N.; Engelbrecht, J.; Rouzé, P.; Brunak, S. Splice site prediction in Arabidopsis thaliana pre-mRNA by combining local and global sequence information. *Nucleic Acids Res.* **1996**, *24*, 3439–3452. [CrossRef] [PubMed]
34. Lagesen, K.; Hallin, P.; Rødland, E.A.; Stærfeldt, H.H.; Rognes, T.; Ussery, D.W. RNAmmer: Consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Res.* **2007**, *35*, 3100–3108. [CrossRef]
35. Korf, I. Gene finding in novel genomes. *BMC Bioinform.* **2004**, *5*, 59. [CrossRef] [PubMed]
36. Chan, P.P.; Lowe, T.M. tRNAscan-SE: Searching for tRNA genes in genomic sequences. In *Gene Prediction*; Springer: New York, NY, USA, 2019; pp. 1–14.
37. Keilwagen, J.; Hartung, F.; Grau, J. GeMoMa: Homology-based gene prediction utilizing intron position conservation and RNA-seq data. In *Gene Prediction*; Springer: New York, NY, USA, 2019; pp. 161–177.
38. Gremme, G.; Brendel, V.; Sparks, M.E.; Kurtz, S. Engineering a software tool for gene structure prediction in higher organisms. *Inf. Softw. Technol.* **2005**, *47*, 965–978. [CrossRef]
39. Van Baren, M.J.; Brent, M.R. Iterative gene prediction and pseudogene removal improves genome annotation. *Genome Res.* **2006**, *16*, 678–685. [CrossRef] [PubMed]
40. Zhang, Z.; Carriero, N.; Zheng, D.; Karro, J.; Harrison, P.M.; Gerstein, M. PseudoPipe: An automated pseudogene identification pipeline. *Bioinformatics* **2006**, *22*, 1437–1439. [CrossRef]
41. Korf, I.; Flicek, P.; Duan, D.; Brent, M.R. Integrating genomic homology into gene structure prediction. *Bioinformatics* **2001**, *17*, S140–S148. [CrossRef]
42. Stanke, M.; Morgenstern, B. AUGUSTUS: A web server for gene prediction in eukaryotes that allows user-defined constraints. *Nucleic Acids Res.* **2005**, *33*, W465–W467. [CrossRef]
43. Allen, J.E.; Majoros, W.H.; Pertea, M.; Salzberg, S.L. JIGSAW, GeneZilla, and GlimmerHMM: Puzzling out the features of human genes in the ENCODE regions. *Genome Biol.* **2006**, *7*, S9. [CrossRef]
44. Sagot, M.F.; Schiex, T.; Rouze, P.; Mathe, C. Current methods of gene prediction, their strengths and weaknesses. *Nucleic Acids Res.* **2002**, *30*, 4103–4117.
45. Wang, Y.; Chen, L.; Song, N.; Lei, X. GASS: Genome structural annotation for eukaryotes based on species similarity. *BMC Genom.* **2015**, *16*, 150. [CrossRef] [PubMed]

46. Sayers, E.W.; Cavanaugh, M.; Clark, K.; Ostell, J.; Pruitt, K.D.; Karsch-Mizrachi, I. GenBank. *Nucleic Acids Res.* **2019**, *47*, D94–D99. [CrossRef] [PubMed]
47. Brooksbank, C.; Bergman, M.T.; Apweiler, R.; Birney, E.; Thornton, J. The european bioinformatics institute's data resources 2014. *Nucleic Acids Res.* **2014**, *42*, D18–D25. [CrossRef] [PubMed]
48. Kodama, Y.; Mashima, J.; Kosuge, T.; Kaminuma, E.; Ogasawara, O.; Okubo, K.; Nakamura, Y.; Takagi, T. DNA data bank of Japan: 30th anniversary. *Nucleic Acids Res.* **2018**, *46*, D30–D35. [CrossRef]
49. Consortium, U. UniProt: A worldwide hub of protein knowledge. *Nucleic Acids Res.* **2019**, *47*, D506–D515. [CrossRef]
50. Mitchell, A.L.; Attwood, T.K.; Babbitt, P.C.; Blum, M.; Bork, P.; Bridge, A.; Brown, S.D.; Chang, H.Y.; El-Gebali, S.; Fraser, M.I.; et al. InterPro in 2019: Improving coverage, classification and access to protein sequence annotations. *Nucleic Acids Res.* **2019**, *47*, D351–D360. [CrossRef]
51. El-Gebali, S.; Mistry, J.; Bateman, A.; Eddy, S.R.; Luciani, A.; Potter, S.C.; Qureshi, M.; Richardson, L.J.; Salazar, G.A.; Smart, A.; et al. The Pfam protein families database in 2019. *Nucleic Acids Res.* **2019**, *47*, D427–D432. [CrossRef]
52. Sigrist, C.J.A.; De Castro, E.; Cerutti, L.; Cuche, B.A.; Hulo, N.; Bridge, A.; Bougueleret, L.; Xenarios, I. New and continuing developments at PROSITE. *Nucleic Acids Res.* **2012**, *41*, D344–D347. [CrossRef]
53. Haft, D.H.; Selengut, J.D.; Richter, R.A.; Harkins, D.; Basu, M.K.; Beck, E. TIGRFAMs and genome properties in 2013. *Nucleic Acids Res.* **2012**, *41*, D387–D395. [CrossRef]
54. Lewis, T.E.; Sillitoe, I.; Dawson, N.; Lam, S.D.; Clarke, T.; Lee, D.; Orengo, C.; Lees, J. Gene3D: Extensive prediction of globular domains in proteins. *Nucleic Acids Res.* **2018**, *46*, D435–D439. [CrossRef]
55. Mi, H.; Huang, X.; Muruganujan, A.; Tang, H.; Mills, C.; Kang, D.; Thomas, P.D. PANTHER version 11: Expanded annotation data from Gene Ontology and Reactome pathways, and data analysis tool enhancements. *Nucleic Acids Res.* **2017**, *45*, D183–D189. [CrossRef]
56. Fang, S.; Zhang, L.; Guo, J.; Niu, Y.; Wu, Y.; Li, H.; Zhao, L.; Li, X.; Teng, X.; Sun, X.; et al. NONCODEV5: A comprehensive annotation database for long non-coding RNAs. *Nucleic Acids Res.* **2018**, *46*, D308–D314. [CrossRef] [PubMed]
57. Karro, J.E.; Yan, Y.; Zheng, D.; Zhang, Z.; Carriero, N.; Cayting, P.; Harrrison, P.; Gerstein, M. Pseudogene. org: A comprehensive database and comparison platform for pseudogene annotation. *Nucleic Acids Res.* **2007**, *35*, D55–D60. [CrossRef] [PubMed]
58. Hubley, R.; Finn, R.D.; Clements, J.; Eddy, S.R.; Jones, T.A.; Bao, W.; Smit, A.F.A.; Wheeler, T.J. The Dfam database of repetitive DNA families. *Nucleic Acids Res.* **2016**, *44*, D81–D89. [CrossRef] [PubMed]
59. Kozomara, A.; Birgaoanu, M.; Griffiths-Jones, S. miRBase: From microRNA sequences to function. *Nucleic Acids Res.* **2019**, *47*, D155–D162. [CrossRef] [PubMed]
60. Mudge, J.M.; Harrow, J. The state of play in higher eukaryote gene annotation. *Nat. Rev. Genet.* **2016**, *17*, 758. [CrossRef] [PubMed]
61. Cutting, G.R. Annotating DNA variants is the next major goal for human genetics. *Am. J. Hum. Genet.* **2014**, *94*, 5–10. [CrossRef]
62. Butkiewicz, M.; Bush, W.S. In silico functional annotation of genomic variation. *Curr. Protoc. Hum. Genet.* **2016**, *88*, 6–15. [CrossRef]
63. Pavlopoulos, G.A.; Oulas, A.; Iacucci, E.; Sifrim, A.; Moreau, Y.; Schneider, R.; Aerts, J.; Iliopoulos, I. Unraveling genomic variation from next generation sequencing data. *BioData Min.* **2013**, *6*, 13. [CrossRef]
64. Koonin, E.V. Orthologs, paralogs, and evolutionary genomics. *Annu. Rev. Genet.* **2005**, *39*, 309–338. [CrossRef]
65. Sasson, O.; Kaplan, N.; Linial, M. Functional annotation prediction: All for one and one for all. *Protein Sci.* **2006**, *15*, 1557–1562. [CrossRef] [PubMed]
66. Botstein, D.; Cherry, J.M.; Ashburner, M.; Ball, C.A.; Blake, J.A.; Butler, H.; Davis, A.P.; Dolinski, K.; Dwight, S.S.; Eppig, J.T.; et al. Gene Ontology: Tool for the unification of biology. *Nat. Genet.* **2000**, *25*, 25–9.
67. Gene Ontology Consortium. Gene ontology consortium: Going forward. *Nucleic Acids Res* **2015**, *43*, D1049–56. [CrossRef] [PubMed]
68. Consortium, G.O. The gene ontology resource: 20 years and still GOing strong. *Nucleic Acids Res.* **2019**, *47*, D330–D338.

69. Thomas, P.D.; Hill, D.P.; Mi, H.; Osumi-Sutherland, D.; Van Auken, K.; Carbon, S.; Balhoff, J.P.; Albou, L.P.; Good, B.; Gaudet, P.; et al. Gene Ontology Causal Activity Modeling (GO-CAM) moves beyond GO annotations to structured descriptions of biological functions and systems. *Nat. Genet.* **2019**, *51*, 1429–1433. [CrossRef]
70. Conesa, A.; Götz, S. Blast2GO: A comprehensive suite for functional analysis in plant genomics. *Int. J. Plant Genom.* **2008**. [CrossRef]
71. Chen, T.W.; Gan, R.C.R.; Wu, T.H.; Huang, P.J.; Lee, C.Y.; Chen, Y.Y.M.; Chen, C.C.; Tang, P. FastAnnotator-an efficient transcript annotation web tool. *BMC Genom.* **2012**, *13*, S9. [CrossRef]
72. Araujo, F.A.; Barh, D.; Silva, A.; Guimaraes, L.; Ramos, R.T.J. GO FEAT: A rapid web-based functional annotation tool for genomic and transcriptomic data. *Sci. Rep.* **2018**, *8*, 1794. [CrossRef]
73. Martin, D.M.A.; Berriman, M.; Barton, G.J. GOtcha: A new method for prediction of protein function assessed by the annotation of seven genomes. *BMC Bioinform.* **2004**, *5*, 178.
74. Törönen, P.; Medlar, A.; Holm, L. PANNZER2: A rapid functional annotation web server. *Nucleic Acids Res.* **2018**, *46*, W84–W88. [CrossRef]
75. Jung, J.; Yi, G.; Sukno, S.A.; Thon, M.R. PoGO: Prediction of Gene Ontology terms for fungal proteins. *BMC Bioinform.* **2010**, *11*, 215. [CrossRef] [PubMed]
76. Kanehisa, M.; Goto, S. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **2000**, *28*, 27–30. [CrossRef] [PubMed]
77. Kanehisa, M.; Sato, Y.; Kawashima, M.; Furumichi, M.; Tanabe, M. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res.* **2016**, *44*, D457–D462. [CrossRef] [PubMed]
78. Fabregat, A.; Jupe, S.; Matthews, L.; Sidiropoulos, K.; Gillespie, M.; Garapati, P.; Haw, R.; Jassal, B.; Korninger, F.; May, B.; et al. The reactome pathway knowledgebase. *Nucleic Acids Res.* **2018**, *46*, D649–D655. [CrossRef]
79. Jassal, B.; Matthews, L.; Viteri, G.; Gong, C.; Lorente, P.; Fabregat, A.; Sidiropoulos, K.; Cook, J.; Gillespie, M.; Haw, R.; et al. The reactome pathway knowledgebase. *Nucleic Acids Res.* **2020**, *48*, D498–D503. [CrossRef]
80. Morgat, A.; Lombardot, T.; Axelsen, K.B.; Aimo, L.; Niknejad, A.; Hyka-Nouspikel, N.; Coudert, E.; Pozzato, M.; Pagni, M.; Moretti, S.; et al. Updates in Rhea-an expert curated resource of biochemical reactions. *Nucleic Acids Res.* **2017**, *45*, D415–D418. [CrossRef]
81. Hastings, J.; Owen, G.; Dekker, A.; Ennis, M.; Kale, N.; Muthukrishnan, V.; Turner, S.; Swainston, N.; Mendes, P.; Steinbeck, C. ChEBI in 2016: Improved services and an expanding collection of metabolites. *Nucleic Acids Res.* **2016**, *44*, D1214–D1219. [CrossRef]
82. Lu, S.; Wang, J.; Chitsaz, F.; Derbyshire, M.K.; Geer, R.C.; Gonzales, N.R.; Gwadz, M.; Hurwitz, D.I.; Marchler, G.H.; Song, J.S.; et al. CDD/SPARCLE: The conserved domain database in 2020. *Nucleic Acids Res.* **2020**, *48*, D265–D268. [CrossRef]
83. MacDonald, J.R.; Ziman, R.; Yuen, R.K.C.; Feuk, L.; Scherer, S.W. The Database of Genomic Variants: A curated collection of structural variation in the human genome. *Nucleic Acids Res.* **2014**, *42*, D986–D992. [CrossRef]
84. Lappalainen, I.; Lopez, J.; Skipper, L.; Hefferon, T.; Spalding, J.D.; Garner, J.; Chen, C.; Maguire, M.; Corbett, M.; Zhou, G.; et al. DbVar and DGVa: Public archives for genomic structural variation. *Nucleic Acids Res.* **2012**, *41*, D936–D941. [CrossRef]
85. Sherry, S.T.; Ward, M.H.; Kholodov, M.; Baker, J.; Phan, L.; Smigielski, E.M.; Sirotkin, K. dbSNP: The NCBI database of genetic variation. *Nucleic Acids Res.* **2001**, *29*, 308–311. [CrossRef] [PubMed]
86. Stenson, P.D.; Ball, E.V.; Mort, M.; Phillips, A.D.; Shaw, K.; Cooper, D.N. The Human Gene Mutation Database (HGMD) and its exploitation in the fields of personalized genomics and molecular evolution. *Curr. Protoc. Bioinform.* **2012**, *39*, 1–13. [CrossRef] [PubMed]
87. Fredman, D.; Siegfried, M.; Yuan, Y.P.; Bork, P.; Lehväslaiho, H.; Brookes, A.J. HGVbase: A human sequence variation database emphasizing data quality and a broad spectrum of data sources. *Nucleic Acids Res.* **2002**, *30*, 387–391. [CrossRef] [PubMed]
88. Fairley, S.; Lowy-Gallego, E.; Perry, E.; Flicek, P. The international genome sample resource (IGSR) collection of open human genomic variation resources. *Nucleic Acids Res.* **2020**, *48*, D941–D947. [CrossRef]
89. Clarke, L.; Zheng-Bradley, X.; Smith, R.; Kulesha, E.; Xiao, C.; Toneva, I.; Vaughan, B.; Preuss, D.; Leinonen, R.; Shumway, M.; et al. The 1000 Genomes Project: Data management and community access. *Nat. Methods* **2012**, *9*, 459–462. [CrossRef]

90. Sharma, V.; Hiller, M. Increased alignment sensitivity improves the usage of genome alignments for comparative gene annotation. *Nucleic Acids Res.* **2017**, *45*, 8369–8377. [CrossRef]
91. Tian, R.; Basu, M.K.; Capriotti, E. Computational methods and resources for the interpretation of genomic variants in cancer. *BMC Genom.* **2015**, *16*, S7. [CrossRef]
92. Coghlan, A.; Fiedler, T.J.; McKay, S.J.; Flicek, P.; Harris, T.W.; Blasiar, D.; Stein, L.D.; nGASP Consortium; et al. nGASP–the nematode genome annotation assessment project. *BMC Bioinform.* **2008**, *9*, 549. [CrossRef]
93. Salamov, A.A.; Solovyev, V.V. Ab initio gene finding in Drosophila genomic DNA. *Genome Res.* **2000**, *10*, 516–522. [CrossRef]
94. Solovyev, V.; of Bioinformatics, D. Statistical approaches in eukaryotic gene prediction. *Handb. Stat. Genet.* **2004**.
95. Yeh, R.F.; Lim, L.P.; Burge, C.B. Computational inference of homologous gene structures in the human genome. *Genome Res.* **2001**, *11*, 803–816. [CrossRef] [PubMed]
96. Clark, D.P.; Pazdernik, N.J.; McGehee, M.R. Chapter 29—Molecular Evolution. In *Molecular Biology*, 3rd ed.; Academic Press: London, United Kingdom, 2019; pp. 925–969. [CrossRef]
97. Slater, G.S.C.; Birney, E. Automated generation of heuristics for biological sequence comparison. *BMC Bioinform.* **2005**, *6*, 31. [CrossRef] [PubMed]
98. Morgenstern, B. DIALIGN: Multiple DNA and protein sequence alignment at BiBiServ. *Nucleic Acids Res.* **2004**, *32*, W33–W36. [CrossRef]
99. Taher, L.; Rinner, O.; Garg, S.; Sczyrba, A.; Brudno, M.; Batzoglou, S.; Morgenstern, B. AGenDA: Homology-based gene prediction. *Bioinformatics* **2003**, *19*, 1575–1577. [CrossRef] [PubMed]
100. Parra, G.; Agarwal, P.; Abril, J.F.; Wiehe, T.; Fickett, J.W.; Guigó, R. Comparative gene prediction in human and mouse. *Genome Res.* **2003**, *13*, 108–117. [CrossRef]
101. Guigó, R.; Flicek, P.; Abril, J.F.; Reymond, A.; Lagarde, J.; Denoeud, F.; Antonarakis, S.; Ashburner, M.; Bajic, V.B.; Birney, E.; et al. EGASP: The human ENCODE genome annotation assessment project. *Genome Biol.* **2006**, *7*, S2. [CrossRef]
102. Cingolani, P.; Platts, A.; Wang, L.L.; Coon, M.; Nguyen, T.; Wang, L.; Land, S.J.; Lu, X.; Ruden, D.M. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly* **2012**, *6*, 80–92. [CrossRef]
103. McLaren, W.; Gil, L.; Hunt, S.E.; Riat, H.S.; Ritchie, G.R.S.; Thormann, A.; Flicek, P.; Cunningham, F. The ensembl variant effect predictor. *Genome Biol.* **2016**, *17*, 122. [CrossRef]
104. Paila, U.; Chapman, B.A.; Kirchner, R.; Quinlan, A.R. GEMINI: Integrative exploration of genetic variation and genome annotations. *PLoS Comput. Biol.* **2013**, *9*, e1003153. [CrossRef]
105. Ng, S.B.; Turner, E.H.; Robertson, P.D.; Flygare, S.D.; Bigham, A.W.; Lee, C.; Shaffer, T.; Wong, M.; Bhattacharjee, A.; Eichler, E.E.; et al. Targeted capture and massively parallel sequencing of 12 human exomes. *Nature* **2009**, *461*, 272–276. [CrossRef]
106. Dayem Ullah, A.Z.; Lemoine, N.R.; Chelala, C. A practical guide for the functional annotation of genetic variations using SNPnexus. *Brief. Bioinform.* **2013**, *14*, 437–447. [CrossRef]
107. Metzker, M.L. Sequencing technologies—the next generation. *Nat. Rev. Genet.* **2010**, *11*, 31–46. [CrossRef] [PubMed]
108. Roy, S.; Coldren, C.; Karunamurthy, A.; Kip, N.S.; Klee, E.W.; Lincoln, S.E.; Leon, A.; Pullambhatla, M.; Temple-Smolkin, R.L.; Voelkerding, K.V.; et al. Standards and guidelines for validating next-generation sequencing bioinformatics pipelines: A joint recommendation of the Association for Molecular Pathology and the College of American Pathologists. *J. Mol. Diagn.* **2018**, *20*, 4–27. [CrossRef] [PubMed]
109. Holt, C.; Yandell, M. MAKER2: An annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinform.* **2011**, *12*, 491. [CrossRef] [PubMed]
110. Cantarel, B.L.; Korf, I.; Robb, S.M.C.; Parra, G.; Ross, E.; Moore, B.; Holt, C.; Alvarado, A.S.; Yandell, M. MAKER: An easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Res.* **2008**, *18*, 188–196. [CrossRef] [PubMed]
111. Thibaud-Nissen, F.; Souvorov, A.; Murphy, T.; DiCuccio, M.; Kitts, P. Eukaryotic genome annotation pipeline. In *The NCBI Handbook*, 2nd ed.; National Center for Biotechnology Information (US): Available online: https://www.ncbi.nlm.nih.gov/sites/books/NBK169439/(accessed on 14 November 2013). .
112. Kapustin, Y.; Souvorov, A.; Tatusova, T.; Lipman, D. Splign: Algorithms for computing spliced alignments with identification of paralogs. *Biol. Direct* **2008**, *3*, 20. [CrossRef]

113. Fiddes, I.T.; Armstrong, J.; Diekhans, M.; Nachtweide, S.; Kronenberg, Z.N.; Underwood, J.G.; Gordon, D.; Earl, D.; Keane, T.; Eichler, E.E.; et al. Comparative Annotation Toolkit (CAT)—simultaneous clade and personal genome annotation. *Genome Res.* **2018**, *28*, 1029–1038. [CrossRef]
114. Paten, B.; Earl, D.; Nguyen, N.; Diekhans, M.; Zerbino, D.; Haussler, D. Cactus: Algorithms for genome multiple sequence alignment. *Genome Res.* **2011**, *21*, 1512–1528. [CrossRef]
115. Stanke, M.; Diekhans, M.; Baertsch, R.; Haussler, D. Using native and syntenically mapped cDNA alignments to improve de novo gene finding. *Bioinformatics* **2008**, *24*, 637–644. [CrossRef]
116. Frankish, A.; Diekhans, M.; Ferreira, A.M.; Johnson, R.; Jungreis, I.; Loveland, J.; Mudge, J.M.; Sisu, C.; Wright, J.; Armstrong, J.; et al. GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res.* **2019**, *47*, D766–D773. [CrossRef]
117. Lilue, J.; Doran, A.G.; Fiddes, I.T.; Abrudan, M.; Armstrong, J.; Bennett, R.; Chow, W.; Collins, J.; Collins, S.; Czechanski, A.; et al. Sixteen diverse laboratory mouse reference genomes define strain-specific haplotypes and novel functional loci. *Nat. Genet.* **2018**, *50*, 1574–1583. [CrossRef]
118. Kronenberg, Z.N.; Fiddes, I.T.; Gordon, D.; Murali, S.; Cantsilieris, S.; Meyerson, O.S.; Underwood, J.G.; Nelson, B.J.; Chaisson, M.J.P.; Dougherty, M.L.; et al. High-resolution comparative analysis of great ape genomes. *Science* **2018**, *360*, 6343. [CrossRef] [PubMed]
119. Hoff, K.J.; Lange, S.; Lomsadze, A.; Borodovsky, M.; Stanke, M. BRAKER1: Unsupervised RNA-Seq-based genome annotation with GeneMark-ET and AUGUSTUS. *Bioinformatics* **2016**, *32*, 767–769. [CrossRef] [PubMed]
120. Lomsadze, A.; Burns, P.D.; Borodovsky, M. Integration of mapped RNA-Seq reads into automatic training of eukaryotic gene finding algorithm. *Nucleic Acids Res.* **2014**, *42*, e119–e119. [CrossRef] [PubMed]
121. Hoff, K.J.; Lomsadze, A.; Borodovsky, M.; Stanke, M. Whole-genome annotation with BRAKER. In *Gene Prediction*; Springer: New York, NY, USA, 2019; pp. 65–95.
122. Seemann, T. Prokka: Rapid prokaryotic genome annotation. *Bioinformatics* **2014**, *30*, 2068–2069. [CrossRef] [PubMed]
123. Hyatt, D.; Chen, G.L.; LoCascio, P.F.; Land, M.L.; Larimer, F.W.; Hauser, L.J. Prodigal: Prokaryotic gene recognition and translation initiation site identification. *BMC Bioinform.* **2010**, *11*, 119. [CrossRef]
124. Aziz, R.K.; Bartels, D.; Best, A.A.; DeJongh, M.; Disz, T.; Edwards, R.A.; Formsma, K.; Gerdes, S.; Glass, E.M.; Kubal, M.; et al. The RAST Server: Rapid annotations using subsystems technology. *BMC Genom.* **2008**, *9*, 1–15. [CrossRef]
125. Overbeek, R.; Begley, T.; Butler, R.M.; Choudhuri, J.V.; Chuang, H.Y.; Cohoon, M.; de Crécy-Lagard, V.; Diaz, N.; Disz, T.; Edwards, R.; et al. The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. *Nucleic Acids Res.* **2005**, *33*, 5691–5702. [CrossRef]
126. Tatusova, T.; DiCuccio, M.; Badretdin, A.; Chetvernin, V.; Nawrocki, E.P.; Zaslavsky, L.; Lomsadze, A.; Pruitt, K.D.; Borodovsky, M.; Ostell, J. NCBI prokaryotic genome annotation pipeline. *Nucleic Acids Res.* **2016**, *44*, 6614–6624. [CrossRef]
127. Besemer, J.; Lomsadze, A.; Borodovsky, M. GeneMarkS: A self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions. *Nucleic Acids Res.* **2001**, *29*, 2607–2618. [CrossRef]
128. Tanizawa, Y.; Fujisawa, T.; Nakamura, Y. DFAST: A flexible prokaryotic genome annotation pipeline for faster genome publication. *Bioinformatics* **2018**, *34*, 1037–1039. [CrossRef]
129. Suzuki, S.; Kakuta, M.; Ishida, T.; Akiyama, Y. GHOSTX: An improved sequence homology search algorithm using a query suffix array and a database suffix array. *PLoS ONE* **2014**, *9*, e103833. [CrossRef] [PubMed]
130. Kiełbasa, S.M.; Wan, R.; Sato, K.; Horton, P.; Frith, M.C. Adaptive seeds tame genomic sequence comparison. *Genome Res.* **2011**, *21*, 487–493. [CrossRef] [PubMed]
131. Finn, R.D.; Clements, J.; Eddy, S.R. HMMER web server: Interactive sequence similarity searching. *Nucleic Acids Res.* **2011**, *39*, W29–W37. [CrossRef]
132. Boratyn, G.M.; Schäffer, A.A.; Agarwala, R.; Altschul, S.F.; Lipman, D.J.; Madden, T.L. Domain enhanced lookup time accelerated BLAST. *Biol. Direct* **2012**, *7*, 12. [CrossRef] [PubMed]
133. Humann, J.L.; Lee, T.; Ficklin, S.; Main, D. Structural and functional annotation of eukaryotic genomes with GenSAS. Methods Mol Biol.; **2019**, *1962*, 29–51.In *Gene Prediction*; Springer: New York, NY, USA, 2019; pp. 29–51. [PubMed]

134. Wang, K.; Li, M.; Hakonarson, H. ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **2010**, *38*, e164–e164. [CrossRef]
135. Chang, X.; Wang, K. wANNOVAR: Annotating genetic variants for personal genomes via the web. *J. Med. Genet.* **2012**, *49*, 433–436. [CrossRef]
136. Sheng, Q.; Yu, H.; Oyebamiji, O.; Wang, J.; Chen, D.; Ness, S.; Zhao, Y.Y.; Guo, Y. AnnoGen: Annotating genome-wide pragmatic features. *Bioinformatics* **2020**, *36*, 2899–2901. [CrossRef]
137. Cavalcante, R.G.; Sartor, M.A. Annotatr: Genomic regions in context. *Bioinformatics* **2017**, *33*, 2381–2383. [CrossRef]
138. Pearson, W.R.; Lipman, D.J. Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci. USA* **1988**, *85*, 2444–2448. [CrossRef]
139. Norling, M.; Jareborg, N.; Dainat, J. EMBLmyGFF3: A converter facilitating genome annotation submission to European Nucleotide Archive. *BMC Res. Notes* **2018**, *11*, 1–5. [CrossRef] [PubMed]
140. Wang, J.; Kong, L.; Gao, G.; Luo, J. A brief introduction to web-based genome browsers. *Brief. Bioinform.* **2013**, *14*, 131–143. [CrossRef] [PubMed]
141. Lander, E.S.; Linton, L.M.; Birren, B.; Nusbaum, C.; Zody, M.C.; Baldwin, J.; Devon, K.; Dewar, K.; Doyle, M.; FitzHugh, W.; et al. Initial sequencing and analysis of the human genome. *Nature* **2001**, *409*, 860–921. [PubMed]
142. Haeussler, M.; Zweig, A.S.; Tyner, C.; Speir, M.L.; Rosenbloom, K.R.; Raney, B.J.; Lee, C.M.; Lee, B.T.; Hinrichs, A.S.; Gonzalez, J.N.; et al. The UCSC genome browser database: 2019 update. *Nucleic Acids Res.* **2019**, *47*, D853–D858. [CrossRef] [PubMed]
143. Cunningham, F.; Achuthan, P.; Akanni, W.; Allen, J.; Amode, M.R.; Armean, I.M.; Bennett, R.; Bhai, J.; Billis, K.; Boddu, S.; et al. Ensembl 2019. *Nucleic Acids Res.* **2019**, *47*, D745–D751. [CrossRef]
144. Harris, T.W.; Arnaboldi, V.; Cain, S.; Chan, J.; Chen, W.J.; Cho, J.; Davis, P.; Gao, S.; Grove, C.A.; Kishore, R.; et al. WormBase: A modern model organism information resource. *Nucleic Acids Res.* **2020**, *48*, D762–D767. [CrossRef]
145. Thurmond, J.; Goodman, J.L.; Strelets, V.B.; Attrill, H.; Gramates, L.S.; Marygold, S.J.; Matthews, B.B.; Millburn, G.; Antonazzo, G.; Trovisco, V.; et al. FlyBase 2.0: The next generation. *Nucleic Acids Res.* **2019**, *47*, D759–D765. [CrossRef]
146. Portwood, J.L.; Woodhouse, M.R.; Cannon, E.K.; Gardiner, J.M.; Harper, L.C.; Schaeffer, M.L.; Walsh, J.R.; Sen, T.Z.; Cho, K.T.; Schott, D.A.; et al. MaizeGDB 2018: The maize multi-genome genetics and genomics database. *Nucleic Acids Res.* **2019**, *47*, D1146–D1154. [CrossRef]
147. Stein, L.D.; Mungall, C.; Shu, S.; Caudy, M.; Mangone, M.; Day, A.; Nickerson, E.; Stajich, J.E.; Harris, T.W.; Arva, A.; et al. The generic genome browser: A building block for a model organism system database. *Genome Res.* **2002**, *12*, 1599–1610. [CrossRef]
148. Buels, R.; Yao, E.; Diesh, C.M.; Hayes, R.D.; Munoz-Torres, M.; Helt, G.; Goodstein, D.M.; Elsik, C.G.; Lewis, S.E.; Stein, L.; et al. JBrowse: A dynamic web platform for genome visualization and analysis. *Genome Biol.* **2016**, *17*, 1–12. [CrossRef]
149. Dunn, N.A.; Unni, D.R.; Diesh, C.; Munoz-Torres, M.; Harris, N.L.; Yao, E.; Rasche, H.; Holmes, I.H.; Elsik, C.G.; Lewis, S.E. Apollo: Democratizing genome annotation. *PLoS Comput. Biol.* **2019**, *15*, e1006790. [CrossRef] [PubMed]
150. Kong, L.; Wang, J.; Zhao, S.; Gu, X.; Luo, J.; Gao, G. ABrowse-a customizable next-generation genome browser framework. *BMC Bioinform.* **2012**, *13*, 1–8. [CrossRef] [PubMed]
151. Medina, I.; Salavert, F.; Sanchez, R.; de Maria, A.; Alonso, R.; Escobar, P.; Bleda, M.; Dopazo, J. Genome Maps, a new generation genome browser. *Nucleic Acids Res.* **2013**, *41*, W41–W46. [CrossRef]
152. Pak, T.R.; Roth, F.P. ChromoZoom: A flexible, fluid, web-based genome browser. *Bioinformatics* **2013**, *29*, 384–386. [CrossRef]
153. Szot, P.S.; Yang, A.; Wang, X.; Parsania, C.; Röhm, U.; Wong, K.H.; Ho, J.W.K. PBrowse: A web-based platform for real-time collaborative exploration of genomic data. *Nucleic Acids Res.* **2017**, *45*, e67. [CrossRef]
154. Dennis, G.; Sherman, B.T.; Hosack, D.A.; Yang, J.; Gao, W.; Lane, H.C.; Lempicki, R.A. DAVID: Database for annotation, visualization, and integrated discovery. *Genome Biol.* **2003**, *4*, 1–11. [CrossRef]
155. Reimand, J.; Arak, T.; Adler, P.; Kolberg, L.; Reisberg, S.; Peterson, H.; Vilo, J. g: Profiler—A web server for functional interpretation of gene lists (2016 update). *Nucleic Acids Res.* **2016**, *44*, W83–W89. [CrossRef] [PubMed]

156. Walter, W.; Sánchez-Cabo, F.; Ricote, M. GOplot: An R package for visually combining expression data with functional analysis. *Bioinformatics* **2015**, *31*, 2912–2914. [CrossRef]
157. Scala, G.; Serra, A.; Marwah, V.S.; Saarimäki, L.A.; Greco, D. FunMappOne: A tool to hierarchically organize and visually navigate functional gene annotations in multiple experiments. *BMC Bioinform.* **2019**, *20*, 1–7. [CrossRef] [PubMed]
158. Huynh, T.; Xu, S. Gene Annotation Easy Viewer (GAEV): Integrating KEGG's Gene Function Annotations and Associated Molecular Pathways. *F1000Research* **2018**, *7*. [CrossRef]
159. Greiner, S.; Lehwark, P.; Bock, R. OrganellarGenomeDRAW (OGDRAW) version 1.3. 1: Expanded toolkit for the graphical visualization of organellar genomes. *Nucleic Acids Res.* **2019**, *47*, W59–W64. [CrossRef] [PubMed]
160. Jung, J.; Kim, J.I.; Jeong, Y.S.; Yi, G. AGORA: Organellar genome annotation from the amino acid and nucleotide references. *Bioinformatics* **2018**, *34*, 2661–2663. [CrossRef] [PubMed]
161. Tillich, M.; Lehwark, P.; Pellizzer, T.; Ulbricht-Jones, E.S.; Fischer, A.; Bock, R.; Greiner, S. GeSeq–versatile and accurate annotation of organelle genomes. *Nucleic Acids Res.* **2017**, *45*, W6–W11. [CrossRef] [PubMed]
162. Krzywinski, M.; Schein, J.; Birol, I.; Connors, J.; Gascoyne, R.; Horsman, D.; Jones, S.J.; Marra, M.A. Circos: An information aesthetic for comparative genomics. *Genome Res.* **2009**, *19*, 1639–1645. [CrossRef]
163. Pabinger, S.; Dander, A.; Fischer, M.; Snajder, R.; Sperk, M.; Efremova, M.; Krabichler, B.; Speicher, M.R.; Zschocke, J.; Trajanoski, Z. A survey of tools for variant analysis of next-generation genome sequencing data. *Brief. Bioinform.* **2014**, *15*, 256–278. [CrossRef]
164. Drori, E.; Levy, D.; Smirin-Yosef, P.; Rahimi, O.; Salmon-Divon, M. CircosVCF: Circos visualization of whole-genome sequence variations stored in VCF files. *Bioinformatics* **2017**, *33*, 1392–1393. [CrossRef]
165. Simonetti, F.L.; Teppa, E.; Chernomoretz, A.; Nielsen, M.; Marino Buslje, C. MISTIC: Mutual information server to infer coevolution. *Nucleic Acids Res.* **2013**, *41*, W8–W14. [CrossRef]
166. An, J.; Lai, J.; Sajjanhar, A.; Batra, J.; Wang, C.; Nelson, C.C. J-Circos: An interactive Circos plotter. *Bioinformatics* **2015**, *31*, 1463–1465. [CrossRef]
167. Yu, Y.; Ouyang, Y.; Yao, W. shinyCircos: An R/Shiny application for interactive creation of Circos plot. *Bioinformatics* **2018**, *34*, 1229–1231. [CrossRef]
168. Darling, A.C.E.; Mau, B.; Blattner, F.R.; Perna, N.T. Mauve: Multiple alignment of conserved genomic sequence with rearrangements. *Genome Res.* **2004**, *14*, 1394–1403. [CrossRef] [PubMed]
169. Letunic, I.; Bork, P. Interactive Tree Of Life (iTOL) v4: Recent updates and new developments. *Nucleic Acids Res.* **2019**, *47*, W256–W259. [CrossRef]
170. Huss III, J.W.; Orozco, C.; Goodale, J.; Wu, C.; Batalov, S.; Vickers, T.J.; Valafar, F.; Su, A.I. A gene wiki for community annotation of gene function. *PLoS Biol.* **2008**, *6*, e175. [CrossRef] [PubMed]
171. Stein, L. Genome annotation: From sequence to biology. *Nat. Rev. Genet.* **2001**, *2*, 493–503. [CrossRef] [PubMed]
172. Pennisi, E. Ideas fly at gene-finding jamboree. *Science* **2000**, *287*, 2182–2184. [CrossRef]
173. Kawai, J.; Shinagawa, A.; Shibata, K.; Yoshino, M.; Itoh, M.; Ishii, Y.; Arakawa, T.; Hara, A.; Fukunishi, Y.; Konno, H.; et al. Functional annotation of a full-length mouse cDNA collection. *Nature* **2001**, *409*, 685–689.
174. Loveland, J.E.; Gilbert, J.G.R.; Griffiths, E.; Harrow, J.L. Community gene annotation in practice. *Database* **2012**, *2012*. [CrossRef]
175. Mazumder, R.; Natale, D.A.; Julio, J.A.E.; Yeh, L.S.; Wu, C.H. Community annotation in biology. *Biol. Direct* **2010**, *5*, 1–7. [CrossRef]
176. Madoui, M.A.; Dossat, C.; d'Agata, L.; van Oeveren, J.; van der Vossen, E.; Aury, J.M. MaGuS: A tool for quality assessment and scaffolding of genome assemblies with Whole Genome Profiling™ Data. *BMC Bioinform.* **2016**, *17*, 115. [CrossRef]
177. Gurevich, A.; Saveliev, V.; Vyahhi, N.; Tesler, G. QUAST: Quality assessment tool for genome assemblies. *Bioinformatics* **2013**, *29*, 1072–1075. [CrossRef]
178. Simao, F.A.; Waterhouse, R.M.; Ioannidis, P.; Kriventseva, E.V.; Zdobnov, E.M. BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **2015**, *31*, 3210–3212. [CrossRef] [PubMed]
179. Iliopoulos, I.; Tsoka, S.; Andrade, M.A.; Enright, A.J.; Carroll, M.; Poullet, P.; Promponas, V.; Liakopoulos, T.; Palaios, G.; Pasquier, C.; et al. Evaluation of annotation strategies using an entire genome sequence. *Bioinformatics* **2003**, *19*, 717–726. [CrossRef] [PubMed]

180. Kasukawa, T.; Furuno, M.; Nikaido, I.; Bono, H.; Hume, D.A.; Bult, C.; Hill, D.P.; Baldarelli, R.; Gough, J.; Kanapin, A.; et al. Development and evaluation of an automated annotation pipeline and cDNA annotation system. *Genome Res.* **2003**, *13*, 1542–1551. [CrossRef] [PubMed]
181. Yang, Y.; Gilbert, D.; Kim, S. Annotation confidence score for genome annotation: A genome comparison approach. *Bioinformatics* **2010**, *26*, 22–29. [CrossRef] [PubMed]
182. Liu, Z.; Ma, H.; Goryanin, I. A semi-automated genome annotation comparison and integration scheme. *BMC Bioinform.* **2013**, *14*, 1–12. [CrossRef]
183. Kalkatawi, M.; Alam, I.; Bajic, V.B. BEACON: Automated tool for bacterial GEnome annotation ComparisON. *BMC Genom.* **2015**, *16*, 616. [CrossRef]
184. Eilbeck, K.; Moore, B.; Holt, C.; Yandell, M. Quantitative measures for the management and comparison of annotated genomes. *BMC Bioinform.* **2009**, *10*, 67. [CrossRef]
185. Cochrane, G.; Karsch-Mizrachi, I.; Takagi, T.; Sequence Database Collaboration, I.N. The international nucleotide sequence database collaboration. *Nucleic Acids Res.* **2016**, *44*, D48–D50. [CrossRef]
186. O'Leary, N.A.; Wright, M.W.; Brister, J.R.; Ciufo, S.; Haddad, D.; McVeigh, R.; Rajput, B.; Robbertse, B.; Smith-White, B.; Ako-Adjei, D.; et al. Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* **2016**, *44*, D733–D745. [CrossRef]
187. Schnoes, A.M.; Brown, S.D.; Dodevski, I.; Babbitt, P.C. Annotation error in public databases: Misannotation of molecular function in enzyme superfamilies. *PLoS Comput. Biol.* **2009**, *5*, e1000605. [CrossRef]
188. Jones, C.E.; Brown, A.L.; Baumann, U. Estimating the annotation error rate of curated GO database sequence annotations. *BMC Bioinform.* **2007**, *8*, 1–9. [CrossRef] [PubMed]
189. Klimke, W.; O'Donovan, C.; White, O.; Brister, J.R.; Clark, K.; Fedorov, B.; Mizrachi, I.; Pruitt, K.D.; Tatusova, T. Solving the problem: Genome annotation standards before the data deluge. *Stand. Genom. Sci.* **2011**, *5*, 168–193. [CrossRef] [PubMed]
190. Nobre, T.; Campos, M.D.; Lucic-Mercy, E.; Arnholdt-Schmitt, B. Misannotation awareness: A tale of two gene-groups. *Front. Plant Sci.* **2016**, *7*, 868. [CrossRef] [PubMed]
191. Ouzounis, C.A.; Karp, P.D. The past, present and future of genome-wide re-annotation. *Genome Biol.* **2002**, *3*,
192. Siezen, R.J.; Van Hijum, S.A.F.T. Genome (re-) annotation and open-source annotation pipelines. *Microb. Biotechnol.* **2010**, *3*, 362. [CrossRef]
193. Yang, H.; Jaime, M.; Polihronakis, M.; Kanegawa, K.; Markow, T.; Kaneshiro, K.; Oliver, B. Re-annotation of eight Drosophila genomes. *Life Sci. Alliance* **2018**, *1*. [CrossRef]
194. Cormier, A.; Avia, K.; Sterck, L.; Derrien, T.; Wucher, V.; Andres, G.; Monsoor, M.; Godfroy, O.; Lipinska, A.; Perrineau, M.M.; et al. Re-annotation, improved large-scale assembly and establishment of a catalogue of noncoding loci for the genome of the model brown alga Ectocarpus. *New Phytol.* **2017**, *214*, 219–232. [CrossRef]
195. Cheng, C.Y.; Krishnakumar, V.; Chan, A.P.; Thibaud-Nissen, F.; Schobel, S.; Town, C.D. Araport11: A complete reannotation of the Arabidopsis thaliana reference genome. *Plant J.* **2017**, *89*, 789–804. [CrossRef]
196. Tamaki, S.; Arakawa, K.; Kono, N.; Tomita, M. Restauro-G: A rapid genome re-annotation system for comparative genomics. *Genom. Proteom. Bioinform.* **2007**, *5*, 53–58. [CrossRef]
197. Salzberg, S.L. Genome re-annotation: A wiki solution? *Genome Biol.* **2007**, *8*, 1–5. [CrossRef]
198. Fleischmann, R.D.; Adams, M.D.; White, O.; Clayton, R.A.; Kirkness, E.F.; Kerlavage, A.R.; Bult, C.J.; Tomb, J.F.; Dougherty, B.A.; Merrick, J.M.; et al. Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science* **1995**, *269*, 496–512. [CrossRef] [PubMed]
199. Lagarde, J.; Uszczynska-Ratajczak, B.; Carbonell, S.; Pérez-Lluch, S.; Abad, A.; Davis, C.; Gingeras, T.R.; Frankish, A.; Harrow, J.; Guigo, R.; et al. High-throughput annotation of full-length long noncoding RNAs with capture long-read sequencing. *Nat. Genet.* **2017**, *49*, 1731–1740. [CrossRef]
200. Robert, C.; Kapetanovic, R.; Beraldi, D.; Watson, M.; Archibald, A.L.; Hume, D.A. Identification and annotation of conserved promoters and macrophage-expressed genes in the pig genome. *BMC Genom.* **2015**, *16*, 970. [CrossRef]
201. Li, W.; Yang, W.; Wang, X.J. Pseudogenes: Pseudo or real functional elements? *J. Genet. Genom.* **2013**, *40*, 171–177. [CrossRef] [PubMed]
202. Workman, R.E.; Tang, A.D.; Tang, P.S.; Jain, M.; Tyson, J.R.; Razaghi, R.; Zuzarte, P.C.; Gilpatrick, T.; Payne, A.; Quick, J.; et al. Nanopore native RNA sequencing of a human poly (A) transcriptome. *Nat. Methods* **2019**, *16*, 1297–1305. [CrossRef] [PubMed]

203. Salzberg, S.L. Next-generation genome annotation: We still struggle to get it right, 2019. *Genome Biol* **2019**, *20* [CrossRef]
204. Danchin, A.; Ouzounis, C.; Tokuyasu, T.; Zucker, J.D. No wisdom in the crowd: Genome annotation in the era of big data–current status and future prospects. *Microb. Biotechnol.* **2018**, *11*, 588–605. [CrossRef]
205. Reed, J.L.; Famili, I.; Thiele, I.; Palsson, B.O. Towards multidimensional genome annotation. *Nat. Rev. Genet.* **2006**, *7*, 130–141. [CrossRef]
206. Hoffman, M.M.; Buske, O.J.; Wang, J.; Weng, Z.; Bilmes, J.A.; Noble, W.S. Unsupervised pattern discovery in human chromatin structure through genomic segmentation. *Nat. Methods* **2012**, *9*, 473. [CrossRef]
207. Yip, K.Y.; Cheng, C.; Gerstein, M. Machine learning and genome annotation: A match meant to be? *Genome Biol.* **2013**, *14*, 1–10. [CrossRef]
208. Nakano, F.K.; Lietaert, M.; Vens, C. Machine learning for discovering missing or wrong protein function annotations. *BMC Bioinform.* **2019**, *20*, 485. [CrossRef] [PubMed]