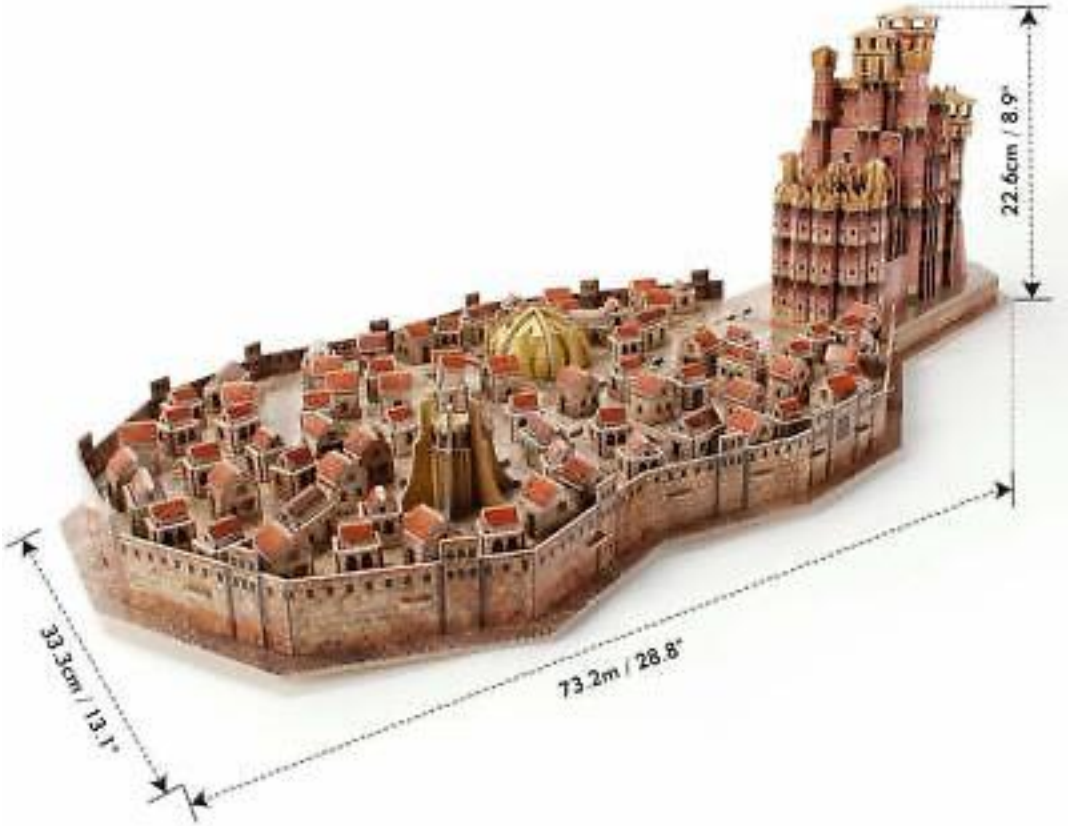
Warm up

… quality filtering/trimming is an important step in the analysis of FASTQ data, as it helps to ensure that the downstream analyses are as accurate and reliable as possible.

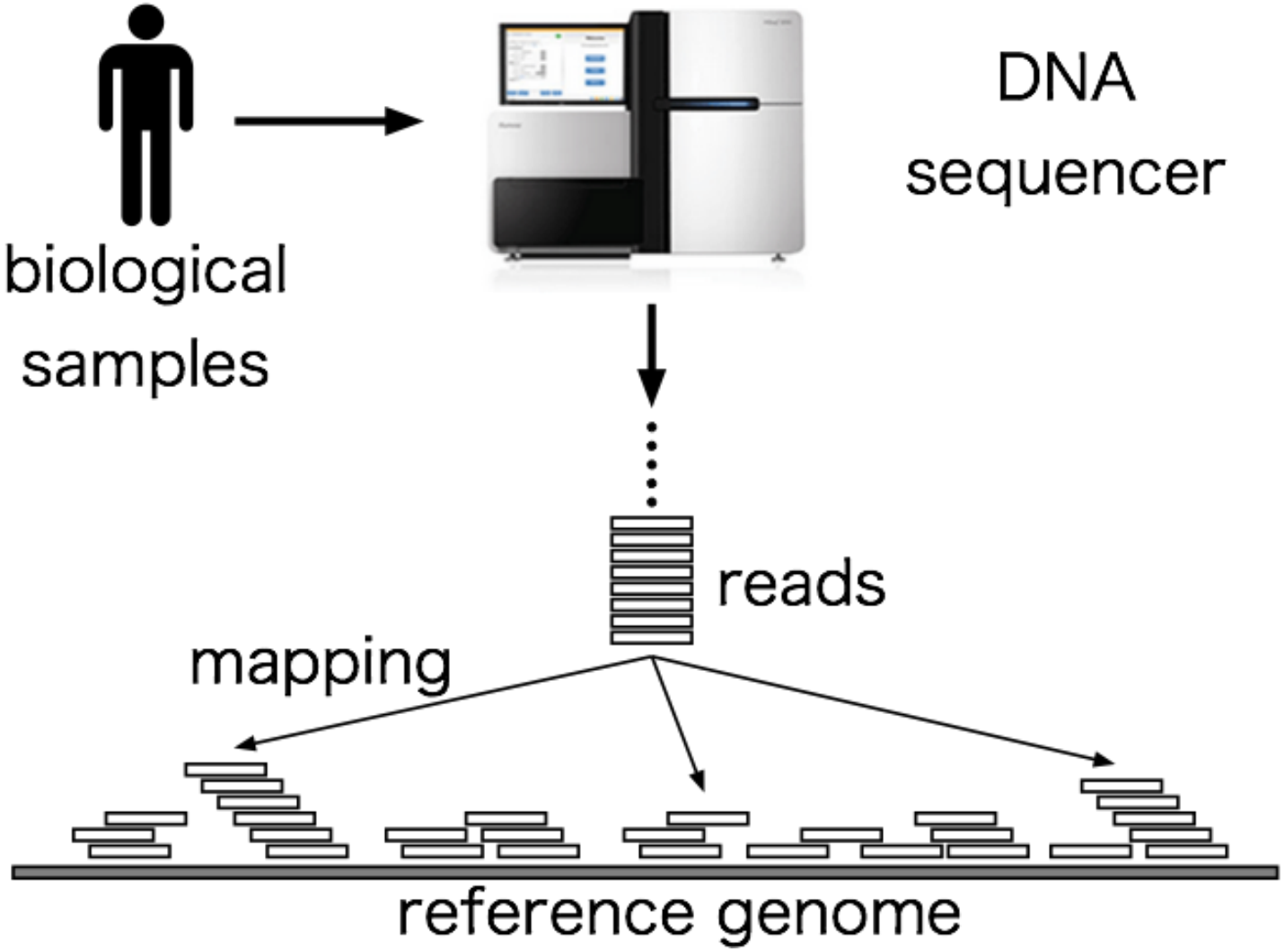
Genomes



Niklaus Zemp
23 June 2024

Genetic Diversity Centre (GDC)
Bioinformatics
ETH Zurich

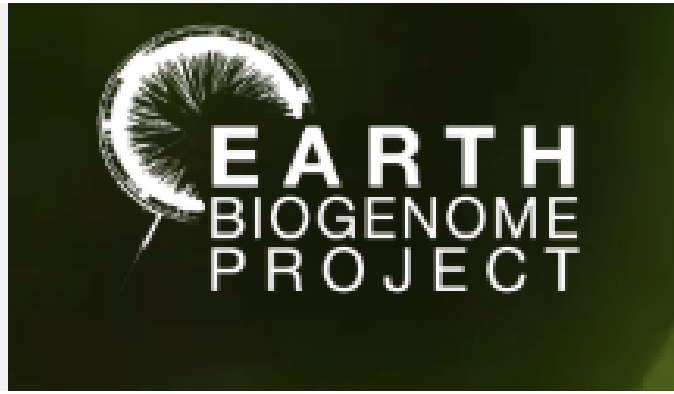
Why we do need a reference?



Reference assembly initiatives

What is the Earth BioGenome Project?

Powerful advances in genome sequencing technology, informatics, automation, and artificial intelligence, have propelled humankind to the threshold of a new beginning in understanding, utilizing, and conserving biodiversity. For the first time in history, it is possible to efficiently sequence the genomes of all known species, and to use genomics to help discover the remaining 80 to 90 percent of species that are currently hidden from science.



Assembly process

3 µg of DNA input per 1 Gb
from a single individual
Pure DNA
20 kbp fragments

Genetic Material

Sequencing

Reads

PacBio (hifi)

Assembler

Hifiasm

Primary assembly

Haplotype 1
Haplotype 2

maternal

parental

related species
HiC
Ultra longONT

Linkage Information such as Mate Pairs, Optical Maps, HiC, Linked Reads

Scaffolder

Scaffold

Reference Genome

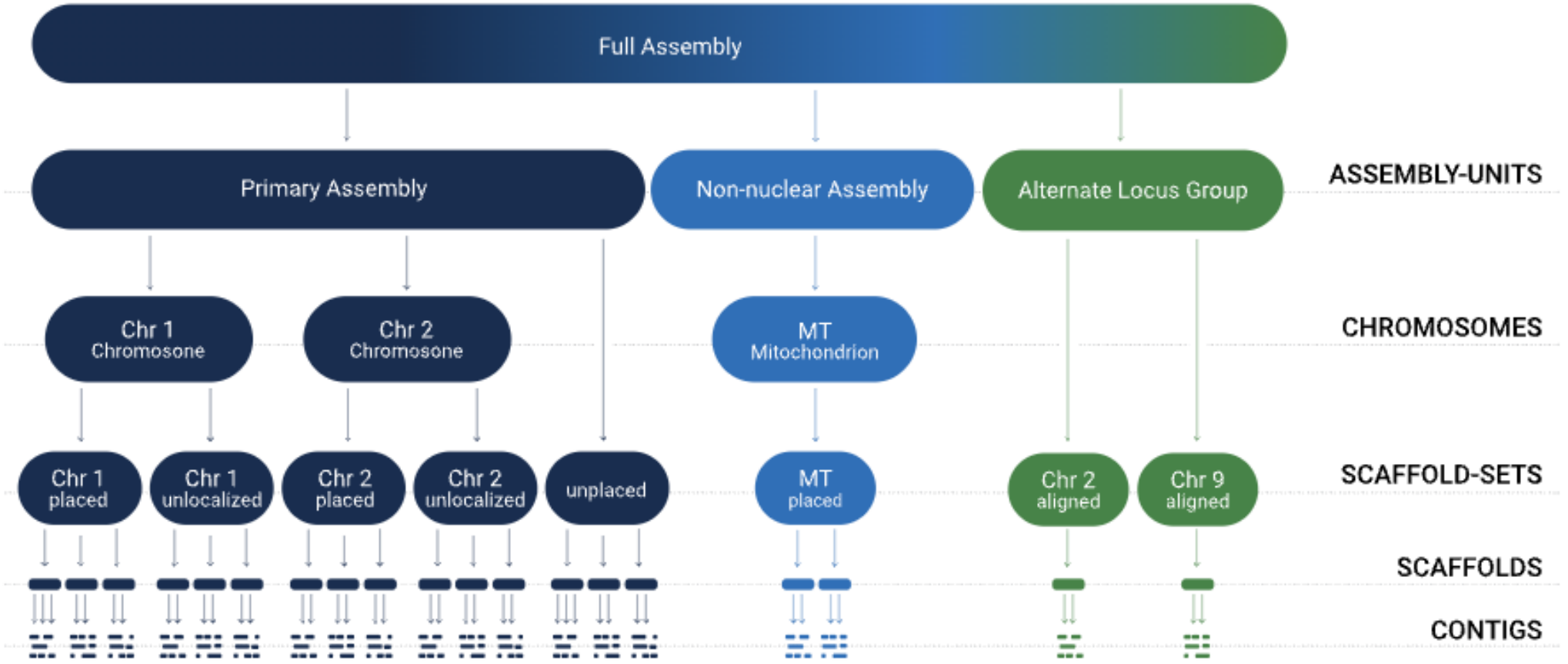
> Science. 2022 Apr;376(6588):44-53. doi: 10.1126/science.abj6987. Epub 2022 Mar 31.

The complete sequence of a human genome

Abstract

Since its initial release in 2000, the human reference genome has covered only the euchromatic fraction of the genome, leaving important heterochromatic regions unfinished. Addressing the remaining 8% of the genome, the Telomere-to-Telomere (T2T) Consortium presents a complete 3.055 billion-base pair sequence of a human genome, T2T-CHM13, that includes gapless assemblies for all chromosomes except Y, corrects errors in the prior references, and introduces nearly 200 million base pairs of sequence containing 1956 gene predictions, 99 of which are predicted to be protein coding. The completed regions include all centromeric satellite arrays, recent segmental duplications, and the short arms of all five acrocentric chromosomes, unlocking these complex regions of the genome to variational and functional studies.

NCBI Genome Assembly



What is a good assembly?

- Used technology
- Genome versus assembly size
- Assembly statistics
- Annotation availbility
- High BUSCO genes scores (conserved single copy genes)
- Similar Kmers composition
- Re-mapping statistics (e.g. short reads)

What is an annotation?

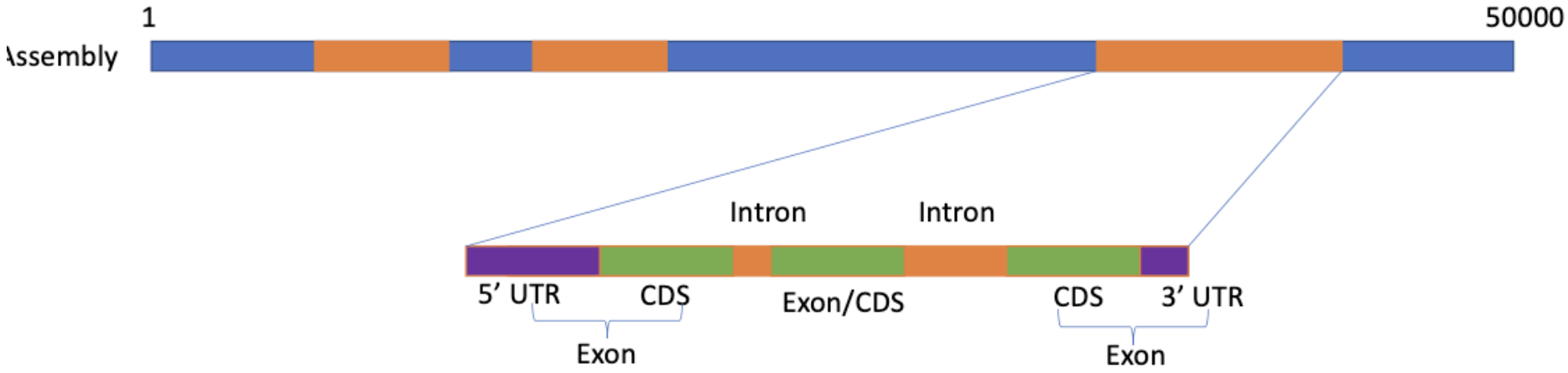
Format: gff/gft

Structural annotation -> ORFs

What is an annotation?

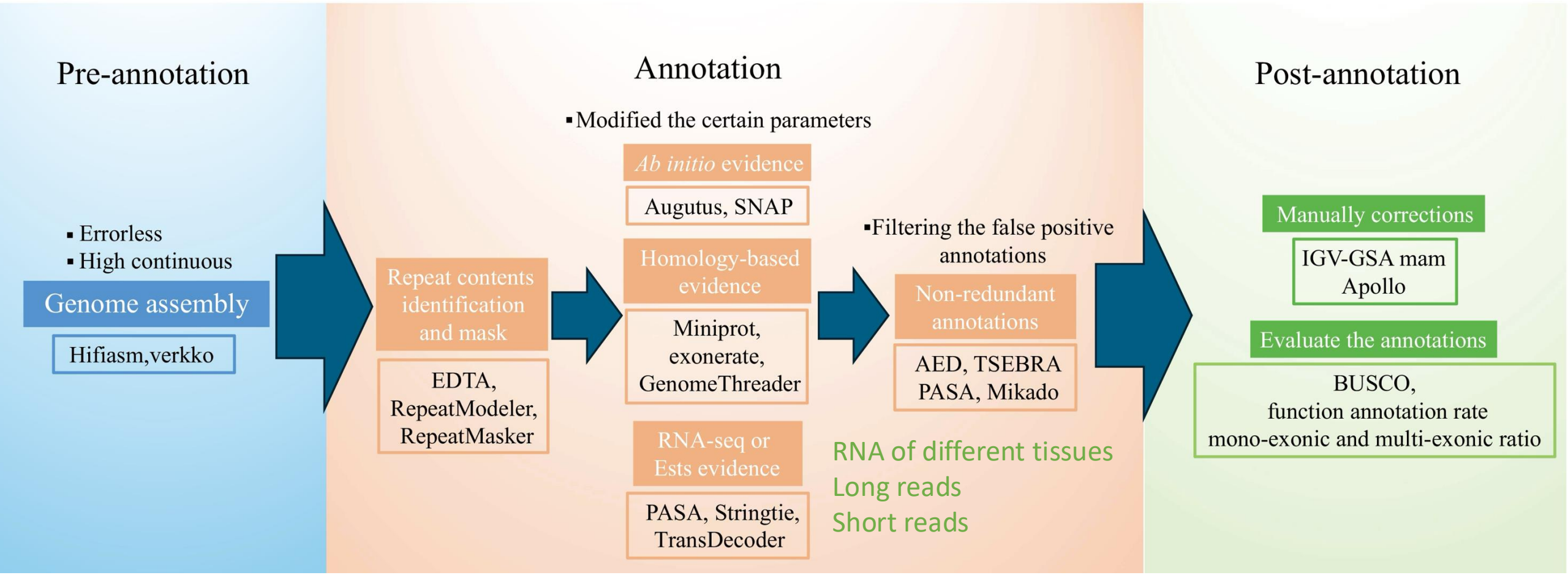
Format: gff/gft

Structural annotation



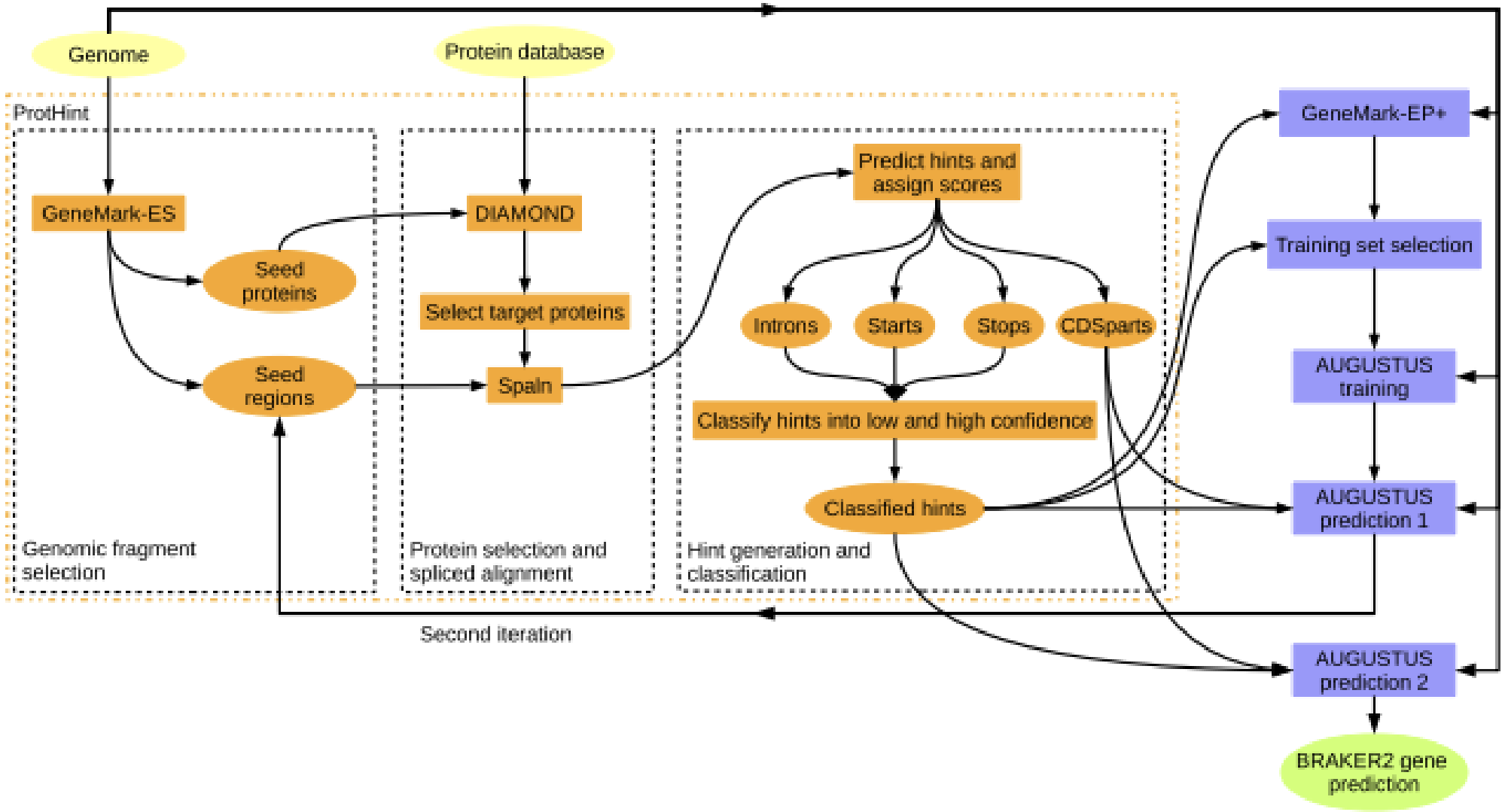
Structural annotation describes the precise location of the different elements in a genome, such as open reading frames (ORFs), coding sequences (CDS), exons, introns, repeats, splice sites, regulatory motifs, start and stop codons, and promoters.

Annotation process



https://www.maxapress.com/article/id/67e36fa6fa6c5866a7a3e810

Annotation process



What is an annotation?

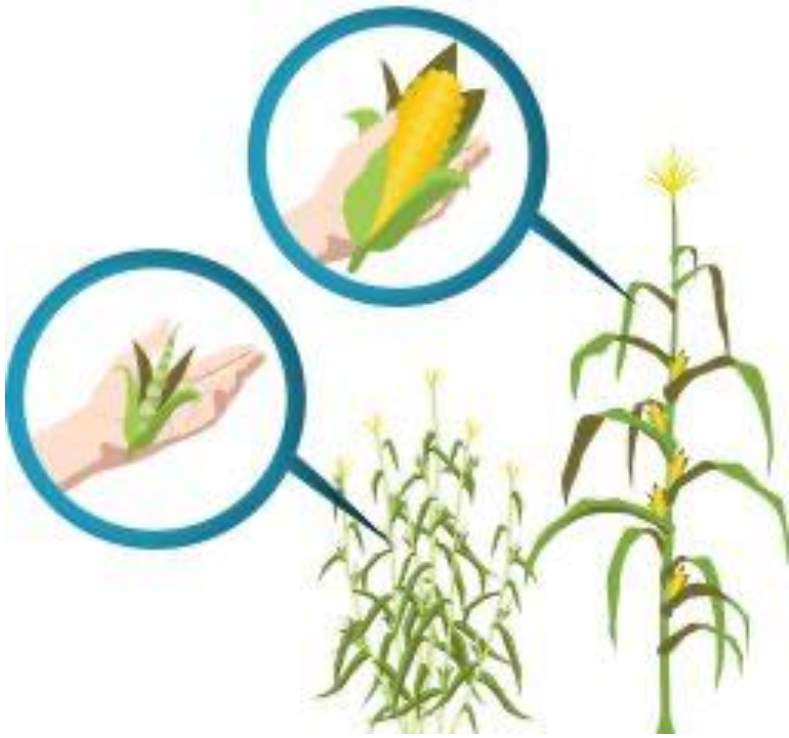
Functional annotation

Functional annotation is defined as the process of collecting information about and describing a gene's biological identity—its various aliases, molecular function, biological role(s), subcellular location, and its expression domains within the plant.

Annotation process

swissprot
NR database

gene name
function
GO term
KEGG term



Genomic location	Chromosome	Orientation	Name	Symbol	Locus Tag	Gene ID
NC_065534.1:113321-114519	1	minus	sphingoid long...	LOC126714477		126714477
NC_065534.1:125021-126085	1	plus	uncharacterize...	LOC126692139		126692139
NC_065534.1:151819-159055	1	plus	uncharacterize...	LOC126692155		126692155
NC_065534.1:188528-192014	1	plus	receptor like pr...	LOC126692164		126692164
NC_065534.1:198872-199389	1	minus	uncharacterize...	LOC126713376		126713376
NC_065534.1:204082-207002	1	minus	uncharacterize...	LOC126692173		126692173
NC_065534.1:223858-229473	1	minus	uncharacterize...	LOC126714485		126714485
NC_065534.1:232377-252965	1	plus	sphingoid long...	LOC126714492		126714492
NC_065534.1:262152-262997	1	minus	uncharacterize...	LOC126692181		126692181
NC_065534.1:267402-268289	1	plus	ferredoxin-thio...	LOC126714498		126714498
NC_065534.1:287470-301598	1	plus	katanin p80 W...	LOC126714506		126714506

White Campion

Genome assembly ASM141213v1 reference

Genome assembly ASM4854445v1 reference

Estimated genome size 2.5 Gbp (1C)

https://www.ncbi.nlm.nih.gov/datasets/genome/GCA_001412135.1/

https://www.ncbi.nlm.nih.gov/datasets/genome/GCA_048544455.1/

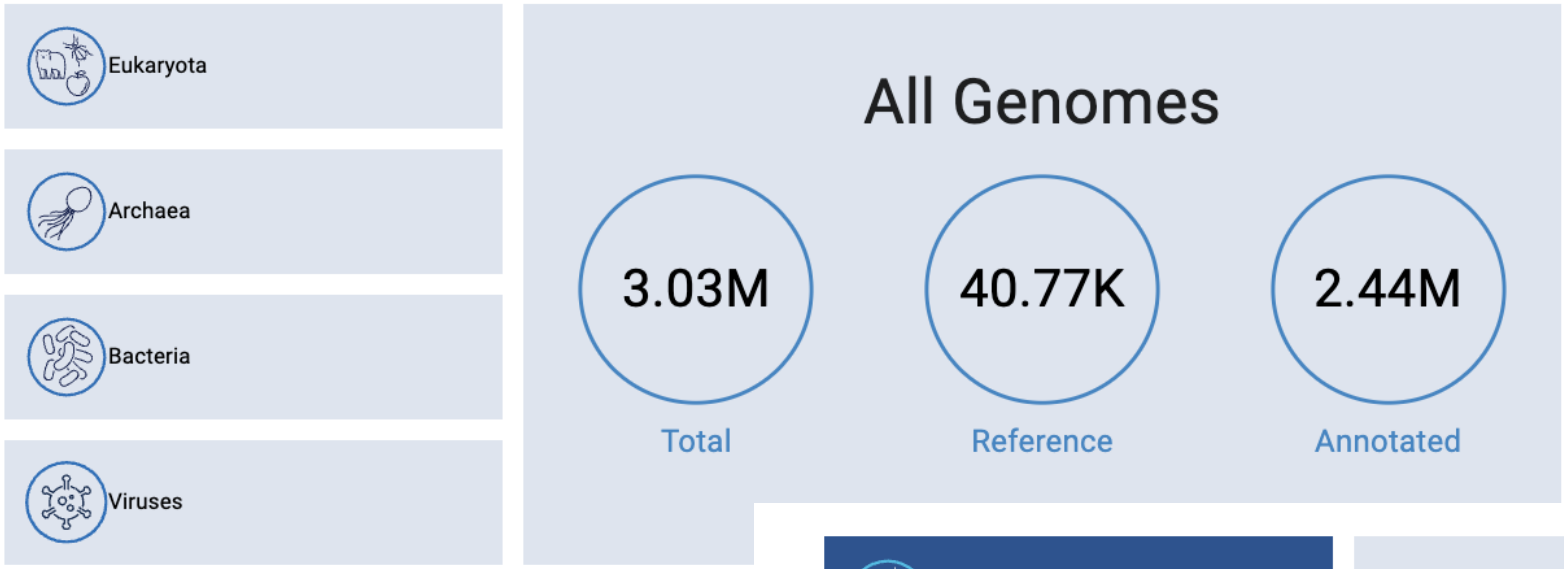
Common toad

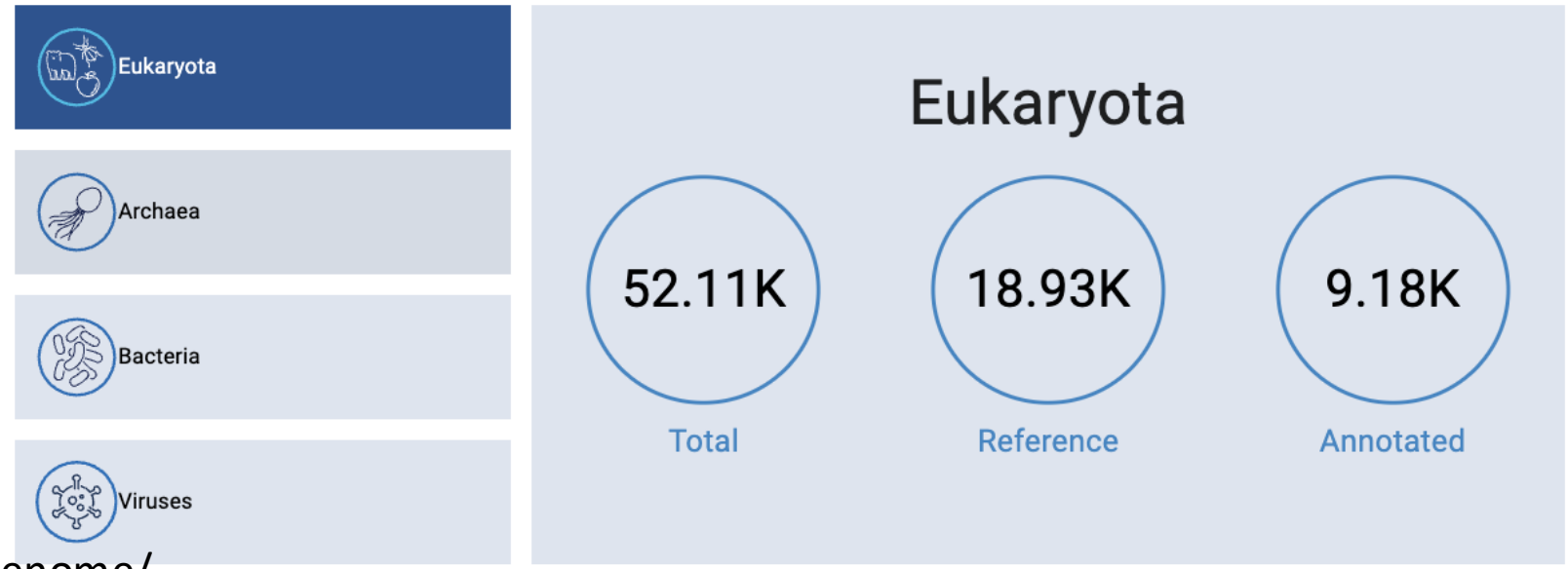
Genome assembly aBufBuf1.1 reference

https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_905171765.1/

Estimated genome size: 5.04 Gb (1C)

How complete are genomes?





https://www.ncbi.nlm.nih.gov/datasets/genome/

Take home message

Genome assemblies are variable in quality.

Know the limits of reference assemblies.

Annotations are “always” wrong.

A genome assembly is the process of piecing together the DNA sequence of an organism's genome. Genomes are typically composed of long strings of nucleotide bases (adenine, thymine, cytosine, and guanine, often abbreviated as A, T, C, and G). However, due to the limitations of sequencing technologies, genomes are rarely sequenced in one continuous stretch. Instead, they are broken down into smaller, overlapping fragments which are sequenced individually. Genome assembly involves taking these fragmented sequences and aligning them to reconstruct the original, complete genome sequence.

Long reads refer to sequencing reads that are substantially longer than traditional short reads. The advantage of long reads in genome sequencing and assembly lies in their ability to span repetitive regions and resolve complex genomic structures more accurately. Here are some of the advantages of long reads:

1. **Resolving repetitive regions**: Long reads can span repetitive regions in the genome that are difficult to resolve with short reads. Repetitive regions pose a challenge for genome assembly because short reads may not be able to uniquely map to these regions, leading to gaps or misassemblies. Long reads provide longer contiguous sequences that can span these repetitive regions, facilitating more accurate assembly.
2. **Capturing structural variations**: Long reads are better able to capture structural variations in the genome, such as insertions, deletions, duplications, and inversions. Short reads may not span these variations entirely, making them challenging to detect accurately. Long reads provide more complete information about the genomic structure, enabling more precise characterization of structural variations.
3. **Simplifying assembly**: Long reads simplify the genome assembly process by providing longer contiguous sequences. This reduces the complexity of the assembly process and mitigates issues such as chimeric contigs and scaffolding errors that can arise from shorter reads. As a result, long-read sequencing often produces more contiguous and accurate genome assemblies.
4. **Facilitating de novo assembly**: Long reads are particularly useful for de novo genome assembly, where the genome of an organism is sequenced without a reference genome. Long reads provide longer sequence information, which is valuable for reconstructing the genome from scratch without relying on a reference sequence.
5. **Improving genome annotation**: Long reads can improve the annotation of genomic features such as genes, promoters, and regulatory elements by providing longer sequence context for accurate identification and characterization of these elements.

Overall, long reads offer several advantages over short reads in genome sequencing and assembly, including improved resolution of repetitive regions, better detection of structural variations, simplified assembly, and enhanced genome annotation. These advantages make long-read sequencing technologies valuable tools for comprehensive genomic analysis in various research fields.

A good genome assembly is one that accurately represents the structure and content of the genome being sequenced. Several criteria contribute to determining the quality of a genome assembly:

1. **Completeness**: A good assembly should cover as much of the genome as possible, with minimal gaps and missing regions. Completeness is often assessed by comparing the assembly to a reference genome or by using metrics such as the percentage of the genome covered and the number of contigs or scaffolds.
2. **Contiguity**: Contiguity refers to the length and continuity of the sequences in the assembly. A highly contiguous assembly consists of long contiguous sequences (contigs or scaffolds) that accurately represent large genomic regions without interruption. Contiguity metrics include N50 (the length of the contig or scaffold at which half of the assembly length is represented) and L50 (the number of contigs or scaffolds needed to cover half of the assembly length).
3. **Accuracy**: Accuracy is crucial for ensuring that the sequences in the assembly are correct. This includes minimizing errors such as base substitutions, insertions, deletions, and misassemblies. Validation methods, such as mapping sequencing reads back to the assembly and using independent experimental data, are used to assess the accuracy of the assembly.
4. **Gene content**: A good assembly should accurately capture the genes and functional elements within the genome. Gene annotation and comparison to known gene sets can help assess the completeness and accuracy of gene representation in the assembly.
5. **Reproducibility**: The assembly should be reproducible, meaning that it can be independently generated and validated by different research groups using the same data and methodologies. Reproducibility ensures the reliability and robustness of the assembly.
6. **Annotation support**: Genome assemblies with accompanying functional annotations, such as gene predictions, regulatory elements, and repetitive sequences, provide valuable insights into the biology of the organism and facilitate downstream analyses.
7. **Biological relevance**: Ultimately, a good genome assembly should be biologically meaningful and relevant to the research questions or applications for which it was generated. This may involve addressing specific genomic features, evolutionary relationships, or functional pathways relevant to the organism under study.

Overall, a good genome assembly balances completeness, contiguity, accuracy, reproducibility, annotation support, and biological relevance to provide a comprehensive and reliable representation of the genome.

Genome annotation is the process of identifying and labeling the functional elements within a genome. These functional elements include genes, coding sequences, regulatory elements, non-coding RNAs, repetitive sequences, and other features. Genome annotation aims to provide a comprehensive understanding of the genetic content and organization of an organism's genome.

The process of genome annotation typically involves several steps:

1. **Gene prediction**: Computational algorithms are used to identify potential protein-coding genes within the genome based on characteristics such as open reading frames (ORFs), sequence homology to known genes, and signals such as start and stop codons. Gene prediction algorithms may also consider evidence from transcriptomic data, such as RNA sequencing (RNA-seq), to refine gene models.
2. **Functional annotation**: Once genes are predicted, their functions are inferred based on similarity to known genes or protein domains. This involves comparing the sequences of predicted genes to databases of annotated genes and proteins using tools such as BLAST (Basic Local Alignment Search Tool) or InterProScan.
3. **Identification of regulatory elements**: Regulatory elements such as promoters, enhancers, and transcription factor binding sites are identified based on sequence motifs, chromatin accessibility data, and other genomic features. These elements play crucial roles in controlling gene expression and are important for understanding the regulation of gene activity within the genome.
4. **Annotation of non-coding RNAs**: Non-coding RNAs (ncRNAs), which do not encode proteins but have regulatory or structural roles in the cell, are also annotated. This includes microRNAs, long non-coding RNAs, and ribosomal RNAs, among others. Computational methods and experimental approaches such as RNA-seq are used to identify and characterize these ncRNAs.
5. **Annotation of repetitive elements**: Repetitive sequences, such as transposable elements and tandem repeats, are annotated to identify their locations and types within the genome. These elements can have important functional implications, such as influencing genome stability and gene regulation.
6. **Integration and visualization**: The annotated genome data are integrated into databases and genome browsers, where researchers can access and visualize the genomic information. Genome browsers provide interactive tools for exploring the genomic features, gene structures, and functional annotations within the genome.

Genome annotation is essential for interpreting the biological significance of genomic sequences, understanding gene function, and conducting comparative genomics studies across different species. It provides a foundation for various fields of research, including genetics, molecular biology, evolutionary biology, and biomedical science.