Lecture Title and Date

Genomics I 01/27/25

Objectives of the Lecture

The objective of the lecture is to do a global study of how biological information is encoded in the genome sequence. By the end of the lecture, students should be able to:

- Understand the experimental steps and troubleshooting associated with obtaining the samples needed for sequencing
- Knowledgeable of which analysis to apply to the genomic data obtained from experiments and how to properly interpret said data after analysis

Key Concepts and Definitions

- Genome: the complete set of DNA, including all genes and non-coding sequences, in an organism which encodes the genetics instructions for development and function.
- Genomics: how biological information is encoded, read, and interpreted to produce distinct biological outcomes.
- Sequencing: Determining the precise order of nucleotides in a DNA or RNA molecule.
    - Deep sequencing: High-throughput sequencing with increased read depth and uniform coverage.
- General Process of Sequencing:
    - Sample Preparation
        - Isolation: Extraction of DNA or RNA from cells or tissue.
        - Library Construction: Preparation of appropriately sized nucleic acid fragments for sequencing.
            - Ligation: Addition of adapter sequences to DNA ends for platform compatibility.
            - Tagmentation: Simultaneous fragmenting and tagging of DNA with adapter sequences to save time.
    - Sequencing
        - Flow Cell Loading: DNA is bound to the flow cell surface for sequencing.
        - Cluster Generation: DNA fragments are amplified to form clusters, with each cluster representing a single DNA molecule.
        - Sequencing by Synthesis: Nucleotides are read using reversible dye terminators.
        - Processing Image Files: Fluorescent signals captured during synthesis are converted into nucleotide sequences.
        - De-multiplexing Samples: Pooled samples are separated based on barcode sequences.
    - Data Analysis
        - Read Filtering: Removal of low-quality reads and adapter sequences.

        - Alignment to a Genome: Mapping reads to a reference genome.
        - Diverse Analyses: Includes epigenomic mapping, variant calling, etc.
- Methods of Sequencing:
    - One at a Time (First-Generation Sequencing)
        - Maxam-Gilbert: Chemical cleavage is performed at specific bases to sequence DNA. This involves labeling one end of the DNA, followed by selective base-specific cleavage and analysis of fragment lengths.
        - Sanger Sequencing: DNA is sequenced by chain termination using dideoxynucleotides. DNA is amplified, terminated at specific nucleotides, and the fragments are analyzed using gel or capillary electrophoresis.
    - Short Read Deep Sequencing (Next-Generation Sequencing)
        - Illumina Sequencing: Sequencing by synthesis generates millions of short reads. DNA is fragmented, adapters are attached, DNA is amplified on a flow cell, and sequencing is performed using reversible dye terminators.
        - Ion Torrent: Sequencing detects pH changes caused by hydrogen ions released during nucleotide incorporation. DNA is fragmented, attached to beads, amplified, and sequenced by monitoring pH changes during synthesis.
    - Long Read Deep Sequencing (Third-Generation Sequencing)
        - Nanopore-Based: DNA or RNA passes through a nanopore, and changes in electrical current are measured to determine the sequence. This enables real-time sequencing of long fragments without amplification or modification.
        - Pacific Biosciences (PacBio) Sequencing: Single-molecule real-time sequencing is performed by immobilizing a DNA polymerase with the DNA template. Nucleotides labeled with fluorescent dyes are incorporated and detected in real time.
- Spatial Omics
    - DNA Probe-Based: Spatial gene expression is studied using in situ hybridization, where DNA probes labeled with fluorescent dyes bind to complementary sequences in fixed cells or tissue.
    - Antibody Probe-Based: Protein localization and expression are detected using labeled antibodies, providing spatial resolution of protein expression in cells or tissues.

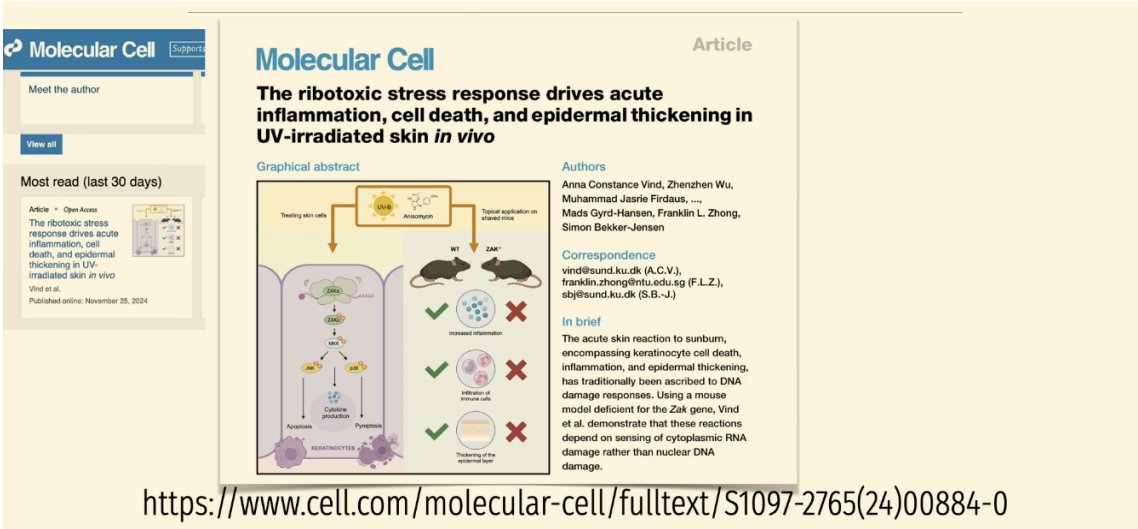
Main Content/Topics

Genomics is the global study of how biological information is encoded in genome sequences, and how this information is read out to produce distinct biological outcomes. This definition is a biological-centered view of genomics, and a lot of people use genomics to talk about anything that is done with deep sequencing.

Today’s lecture is on sequencing technology and genomes, and the next lecture (Genomics II) is focused on applications of sequencing technology.

Overview: will cover sequencing data from wet lab to fastq. Sophisticated use of data from genomics requires an integrated understanding of the biological experiment, sample preparation, and downstream computational analyses of the data.

Importance of genomics data: these data are central to most biomedical and biological applications. Pretty much any topic will have sequencing data in it, such as the following paper:



Raw data can be found in genomics databases.
What is the output of an Illumina sequencing experiment? Fastq uses the following syntax:

1. Read identifier
2. Sequence
3. Quality score identifier (“+”)
4. Quality score

Workflow:

1. Isolation of sample (e.g. isolate DNA and shear)
2. Library preparation (e.g. add known sequences to the ends)
3. Sequencing (e.g. Illumina Novaseq)
4. Analysis (e.g. map to genome and interpret)

Metrics for evaluating sequencing technology:

1. Throughput
    a. Number of high-quality bases per unit of time
    b. Number of independent samples run in parallel
2. Cost
    a. Per run cost
    b. Per base cost
    c. Equipment
    d. Reagents

    e. Labor
    f. Analysis
3. Yield
    a. Number of useful reads per sample
    b. Read length
4. Quality
    a. Accuracy per base

There are many types of sequencing:

1. One-at-a-tie methods
    a. Maxam-Gilbert sequencing
    b. Sanger sequencing
2. Short read deep sequencing
    a. Illumina sequencing
    b. Ion torrent
3. Long-read deep sequencing
    a. Nanopore based
    b. Pacific Bioscience sequencing
4. Spatial-omics
    a. DNA probe-based
    b. Antibody probe based

Sequencing technology	Platform	Data type	Read length (kb)		Read accuracy (%)	Throughput per flow cell (Gb)		Estimated cost per Gb (US$)	Maximum throughput per year (Gb)[a]
			N50	Maximum		Mean	Maximum		
Pacific Biosciences (PacBio)	RS II[b]	CLR	5–15	>60	87–92	0.75–1.5	2	333–933[c]	4,380
	Sequel	CLR	25–50	>100		5–10	20	98–195[d]	17,520
	Sequel II	CLR	30–60	>200		50–100	160	13–26[e]	93,440
		HiFi	10–20	>20	>99	15–30	35	43–86[e]	10,220
Oxford Nanopore Technologies (ONT)	MinION/GridION	Long	10–60	>1,000	87–98	2-20	30	50–500[f]	21,900 (MinION) 109,500 (GridION)
		Ultra-long	100–200	>1,500		0.5–2	2.5	500–2,000[f]	913 (MinION) 4,563 (GridION)
	PromethION	Long	10–60	>1,000		50–100	180	21–42[f]	3,153,600
Illumina	NextSeq 550	Single-end	0.075–0.15	0.15	>99.9	16–30	>30	50–63[g]	>47,782
		Paired-end	0.075–0.15 (×2)	0.15 (×2)		32–120	>120	40–60[g]	>70,080
	NovaSeq 6000	Single-end	0.05–0.25	0.25		65–3,000	>3,000	10–35[h]	>1,194,545
		Paired-end	0.05–0.25 (×2)	0.25 (×2)					

The steps of sequencing experiments:

Sequencing

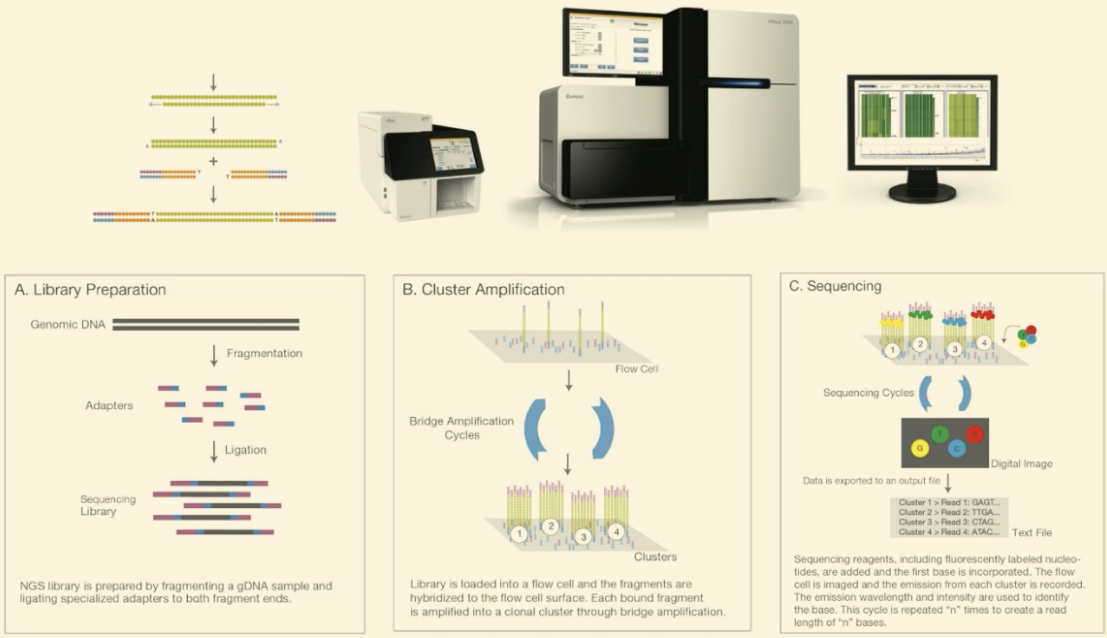
Service	Yale Rate	Non-Yale Rate
MiSeq 500 Cycle	$1,905	$2,485
NextSeq Usage	$1,039	$1,358
NovaSeq X Plus 25B 2x150	$2,936	$3,825
NovaSeq SP 2x150	$2,617	$3,410

…per 1.25 billion reads!

1. Sample preparation
    a. Isolation
    b. Library construction
2. Sequencing
    a. Flow cell loading
    b. Cluster generation
    c. Sequencing
    d. Processing image files
    e. De-multiplexing samples
3. Data analysis
    a. Read filtering
    b. Alignment to a genome
    c. Diverse analysis

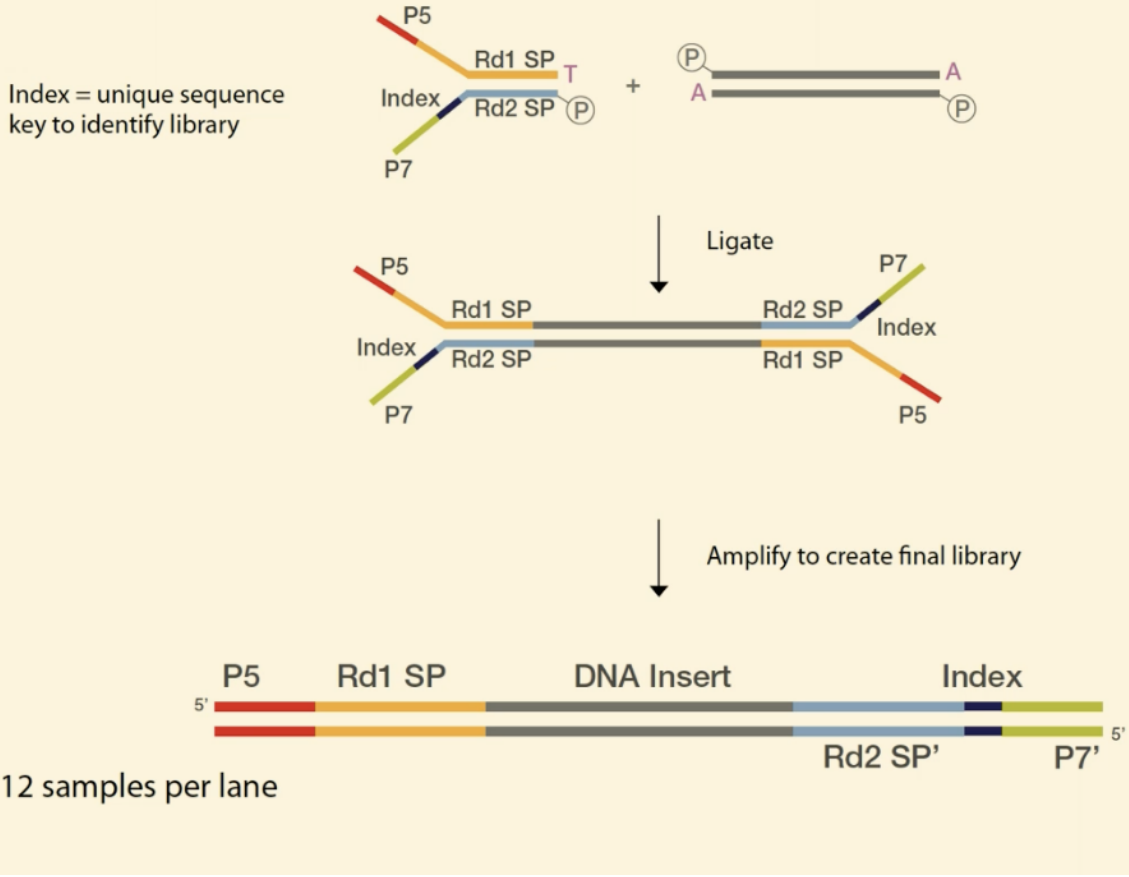
What is the most raw form of data recorded in an Illumina sequencing experiment? Chromatogram, string of letters, series of images, or readout of genomic locations? Answer: series of images.

Where do these reads come from? Must turn DNA into a format that can be read (library preparation):



https://www.illumina.com/content/dam/illumina-marketing/documents/products/illumina_sequencing_introduction.pdf

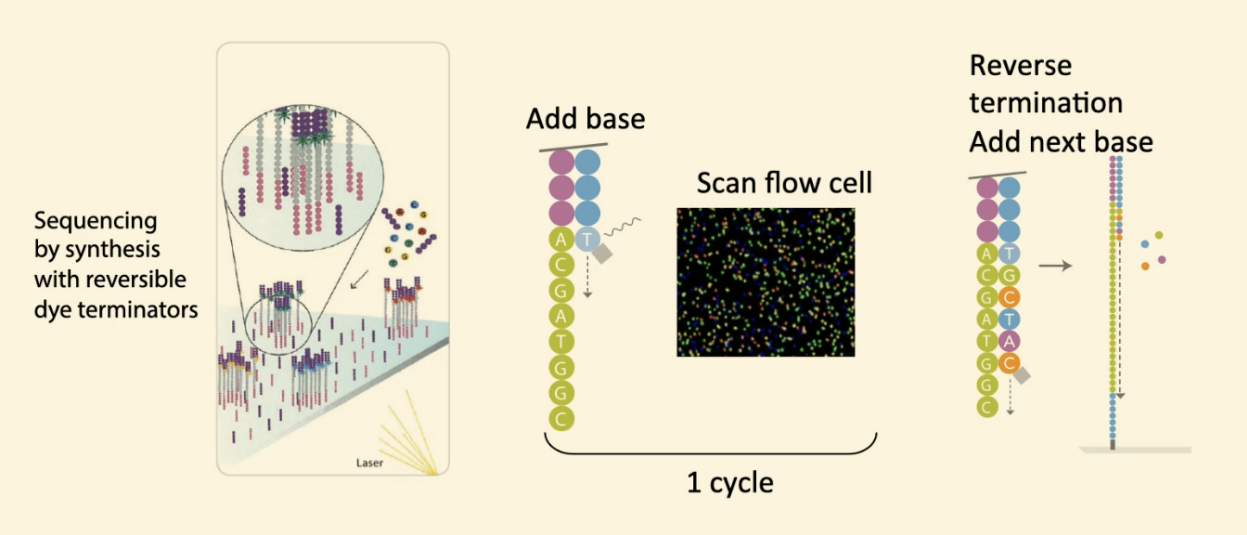
Optional: library preparation using ligation:



Cluster amplification:

1. Separate each individual molecule (randomly)
2. Give each molecule an address (spatial location)
3. Pack as many on as possible but avoid overlaps

Sequencing by synthesis:



The reads are 75 nucleotides long (then 100, now 150). While other technologies can give longer read lengths, Illumina reads are generally 50 nt - 250 nt.

What limits the insert size and read length? Instead of fragmenting to 100, why not 10,000? It only works if they are all in sync. Some of the molecules will be a base off. Therefore, the quality of the sequence decreases.

- For every single end read: incomplete incorporation of bases
- For the size of the insert (especially for paired-end analysis): Ability to get consistent clusters

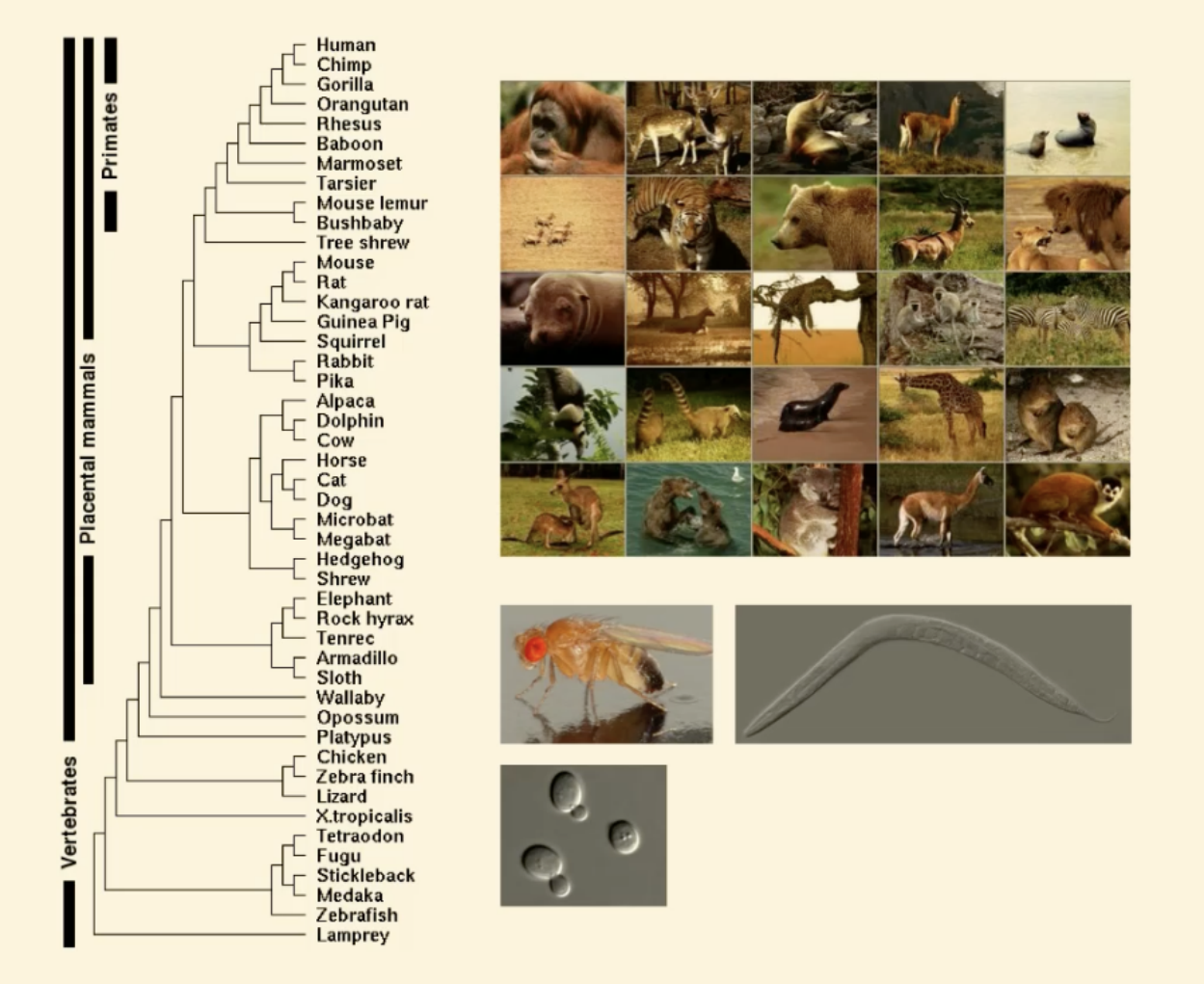
What is the output from an Illumina sequencing experiment? Paired read (fastq format):

1. Read identifier
    a. Flow cell
    b. Read ID
    c. Coordinates
    d. Which read from a paired-end sample
    e. Which index for a multiplexed read
2. Sequence
3. Quality score identifier “+”
4. Quality score

Generally ,~2 billion reads/sequencing lane (for an Illumina NovaSeq with current chemistry)

Next section: what do i do with my sequencing reads?

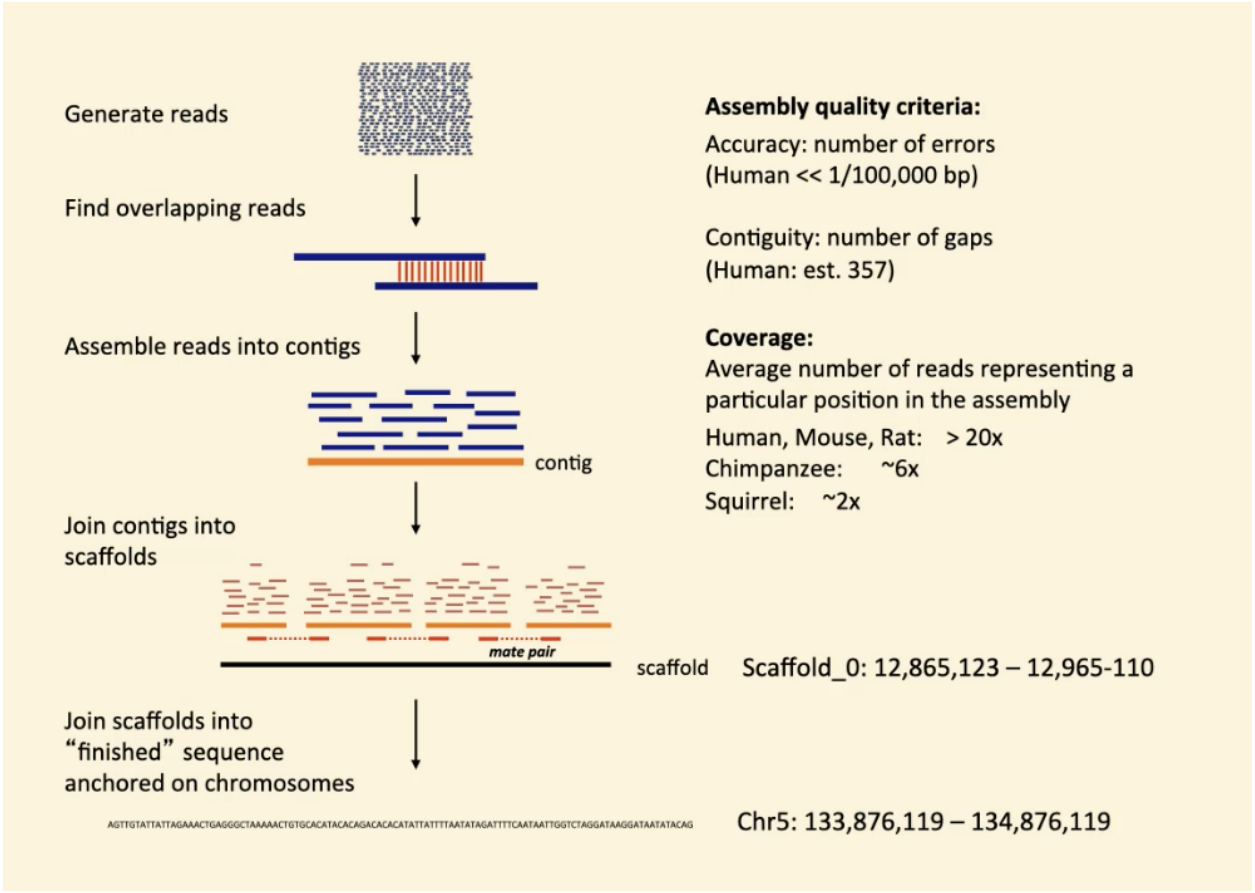
Many reference genomes are available (you can figure out how your sequence relates to an organism that already has a reference sequence:



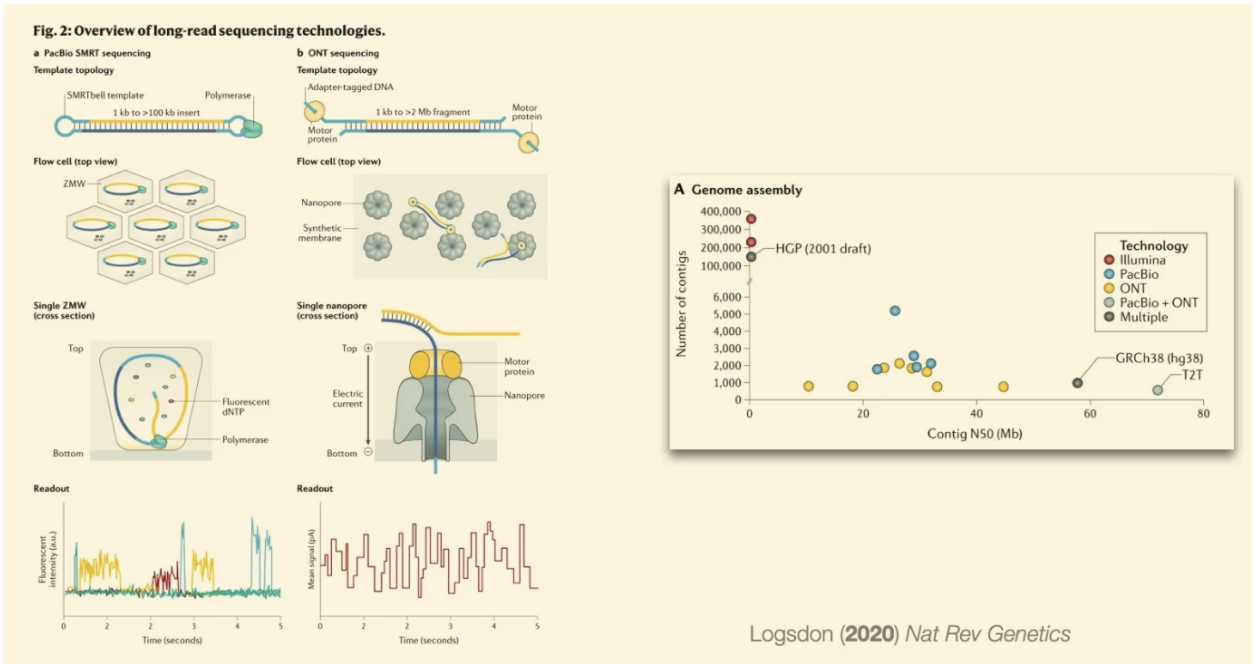
There is a wide range of genome sizes: the human haploid genome has ~3 Gb.

Sequencing of the human genome was done initially in 2003, but in reality, parts were missing. It also cost $3 billion at a rate of 1 Gb/month, now it’s $100.

How to assemble a genome:



The importance of long-read sequencing:



What types of annotation do we want?

1. Genes
    a. Coding, noncoding, miRNA, etc.
    b. Isoforms
    c. Expression
2. Genetic variation
    a. SNPs and CNVs
3. Sequence conservation
4. Regulatory sequences
    a. Promoters
    b. Enhancers
    c. Insulators
5. Epigenetics
    a. DNA methylation
    b. Chromatin

Degrees of genomic annotation vary widely

1. Humans, mouse (fly, worm, yeast)
    a. Chromosome assemblies
    b. Dense gene and regulatory maps, variation, etc.
2. Other models (dog, chicken, zebrafish):
    a. Chromosome assemblies
    b. Partial gene maps; variation; little regulatory data
3. Low-coverage vertebrate genomes:
    a. Scaffold assemblies
    b. Few annotated genes
    c. Used for comparative purposes

Where to look for existing annotations:

1. UCSC genome browser
2. Ensembl
3. Integrative genomics viewer
4. Galaxy

Example of a genome browser track (UCSC):

Using sequencing to annotate the genome

1. Where are the cis-acting regulatory elements in DNA?
    a. DNase I hypersensitivity mapping (DNase-Seq)
    b. FAIRE to map regulatory elements.
2. Where do transcription factors bind?
    a. ChIP-seq of transcription factors (or in high res, ChIP-exo)
    b. Nucleosome mapping (MNase-Seq).
3. Where are different histone modifications found?
    a. ChIP-Seq of histone modifications.
    b. ChIP-Seq of chromatin writers, readers and erasers.
4. Where is RNA polymerase transcribing?
    a. ChIP-Seq of polymerase.
    b. GRO-Seq, NET-Seq and TT-Seq to measure RNA in the polymerase active site.
5. How is the genome organized in 3D?
    a. 4C/5C/Hi-C to measure chromatin conformation.

Applications of sequencing technology covered next class.

Conclusions

1. Sequencing technology is central to our understanding of biology
2. The decrease in cost and increase in throughput make sequencing data increasingly ubiquitous

References ISL/ESL

Obviously, this lecture includes almost no discussion on statistics, however, there are certainly some statistical methods and understanding that are needed when working with genomics that I am sure we will be covering later in the course. However, with that being said, here are some references:

ESL: Chapter 18 of ESL titled *High-Dimensional Problems* is likely a critical reference when working with genomic data. The main focus of this chapter is on exploring what to do when p>N, which is stated to be a common theme when working with high-dimensional genomic data. High variance and overfitting are two big challenges when dealing with genomic data. This chapter emphasizes regularization for both classification and regression and also focuses on feature selection and assessment. The other chapter I would suggest would be Chapter 14 because that chapter discusses dimension reduction, PCA, and nonlinear dimension reduction, all of which are critical to know when dealing with extremely large datasets.

ISL: Similarly, Chapter 6 of ISL discusses a lot about regularization, shrinkage methods, dimension reduction, and considerations for working with high dimensional data. This chapter should be a good additional resource to chapter 18 of ESL. Chapter 12 discusses important unsupervised machine-learning methods that are critical for genomics. Specifically, it briefly touches on centroid linkage which is said to be an important tool for genomics analysis.

In general, there are some great chapters in both books that should better your understanding of the statistical frameworks used for most genome research. Because this lecture is a broad introduction to genomics and there are so many possible ways of analyzing the genome, it might be important to consider reading some of the many other chapters depending on your research interests.

Suggest references for many of the key concepts

Genomics Overview: *A Vision for the Future of Genomics Research* - *https://www.nature.com/articles/nature01626*

DNA Sequencing Technologies: *What is next generation sequencing?* - https://pmc.ncbi.nlm.nih.gov/articles/PMC3841808/

More information on Illumina: https://www.illumina.com/science/technology/next-generation-sequencing.html

The Challenge of Genome Sequence Assembly - https://openbioinformaticsjournal.com/VOLUME/11/PAGE/231/FULLTEXT/

Genome Annotation - https://www.sciencedirect.com/topics/biochemistry-genetics-and-molecular-biology/genome-annotation

NGS Workflow - https://www.idtdna.com/pages/technology/next-generation-sequencing/workflow

Genome Browser (interactive) - genome.ucsc.edu