



BGGN 213

Genome Informatics I

Lecture 14

Barry Grant
UC San Diego

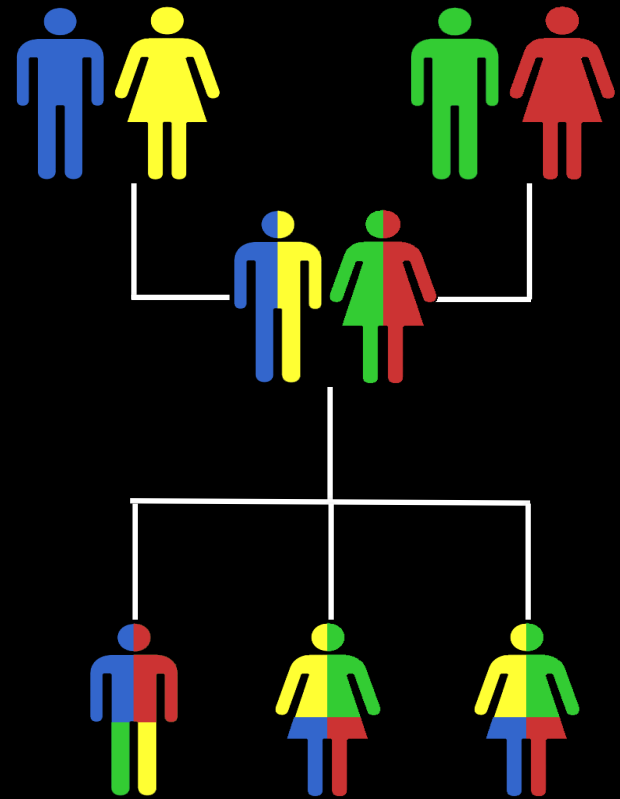
<http://thegrantlab.org/bggn213>

Today's Menu:

- **What is a Genome?**
 - Genome sequencing and the Human genome project
- **What can we do with a Genome?**
 - Compare, model, mine and edit
- **Modern Genome Sequencing**
 - 1st, 2nd and 3rd generation sequencing
- **Workflow for NGS**
 - RNA-Sequencing and Discovering variation

What is a genome?

The total genetic material of an organism by which individual traits are encoded, controlled, and ultimately passed on to future generations



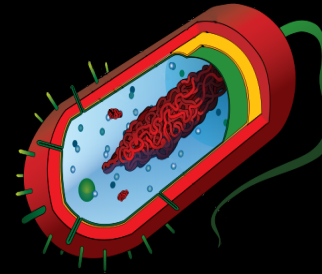
Genetics and Genomics

- **Genetics** is primarily the study of *individual genes*, mutations within those genes, and their inheritance patterns in order to understand specific traits.
- **Genomics** expands upon classical genetics and considers aspects of the *entire genome*, typically using computer aided approaches.

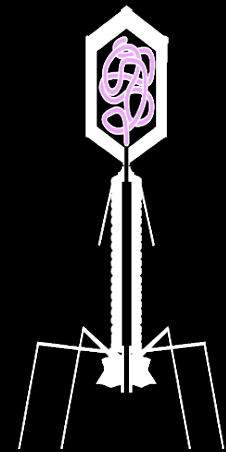
Genomes come in many shapes

Side note!

- Primarily DNA, but can be RNA in the case of some viruses
- Some genomes are circular, others linear
- Can be organized into discrete units (chromosomes) or freestanding molecules (plasmids)



Prokaryote



Bacteriophage



Eukaryote

Side note!

CHROMOSOMES CLOSE-UP

Chromosomes consist largely of double-helical DNA. Cells package the DNA into the nucleus by wrapping it around “spools” composed of histone proteins. The DNA-protein combination is known as chromatin. (Each color represents one chromosome.)

Under a microscope, a Eukaryotic cell's genome (i.e. collection of chromosomes) resembles a chaotic jumble of noodles. The looping is not random however and appears to play a role in controlling gene regulation.

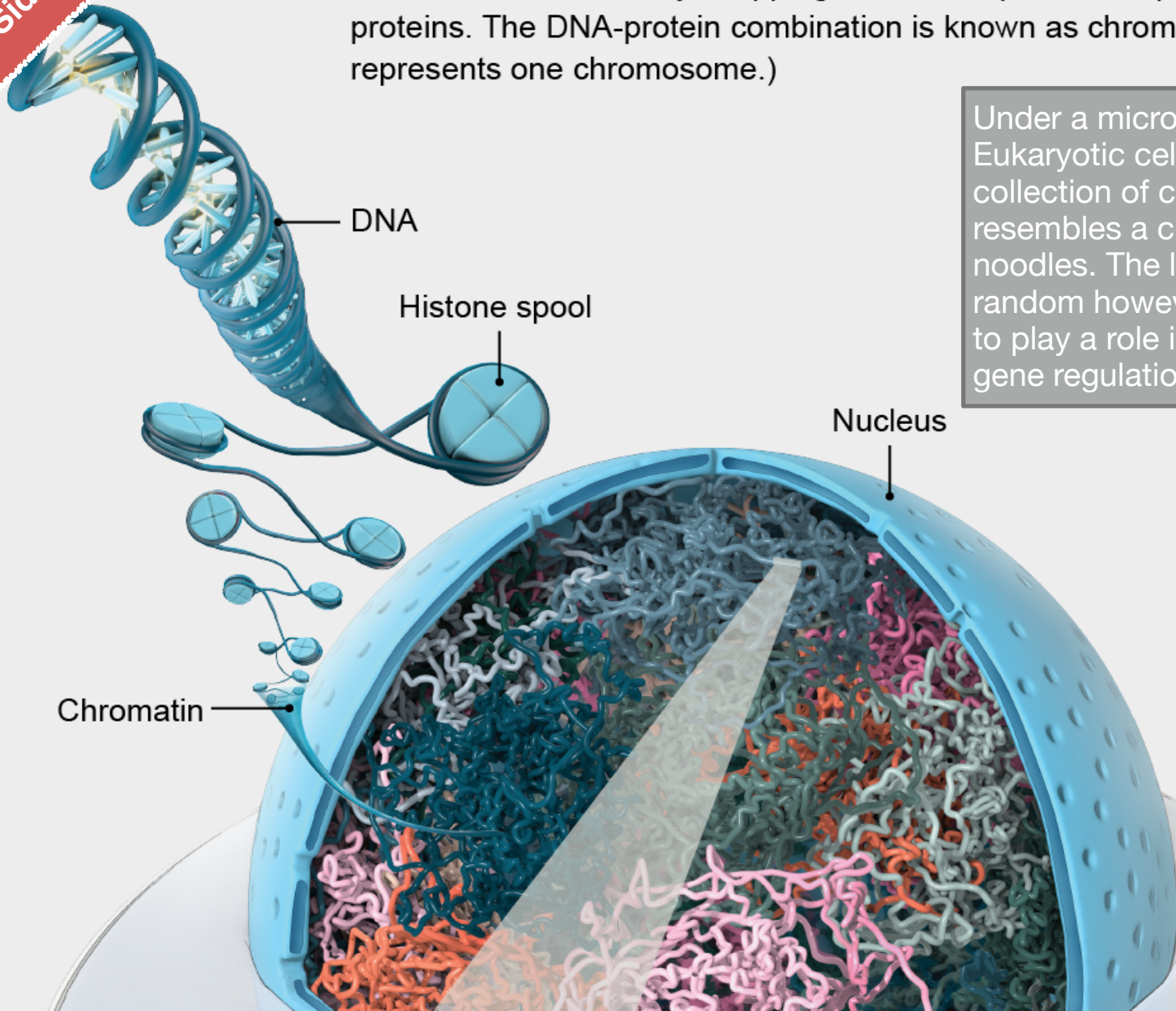
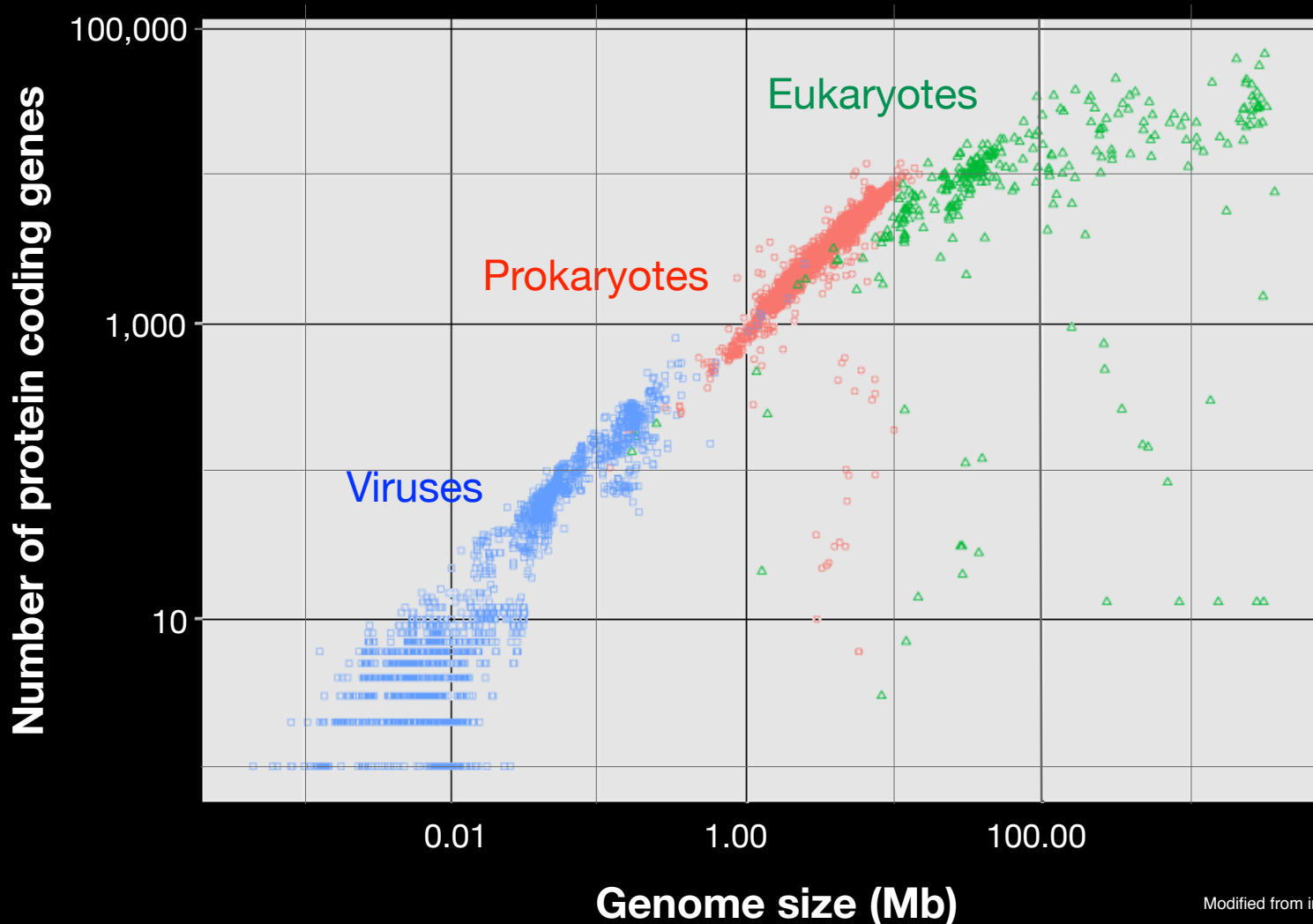


Image credit:
[Scientific American](#)
March 2019

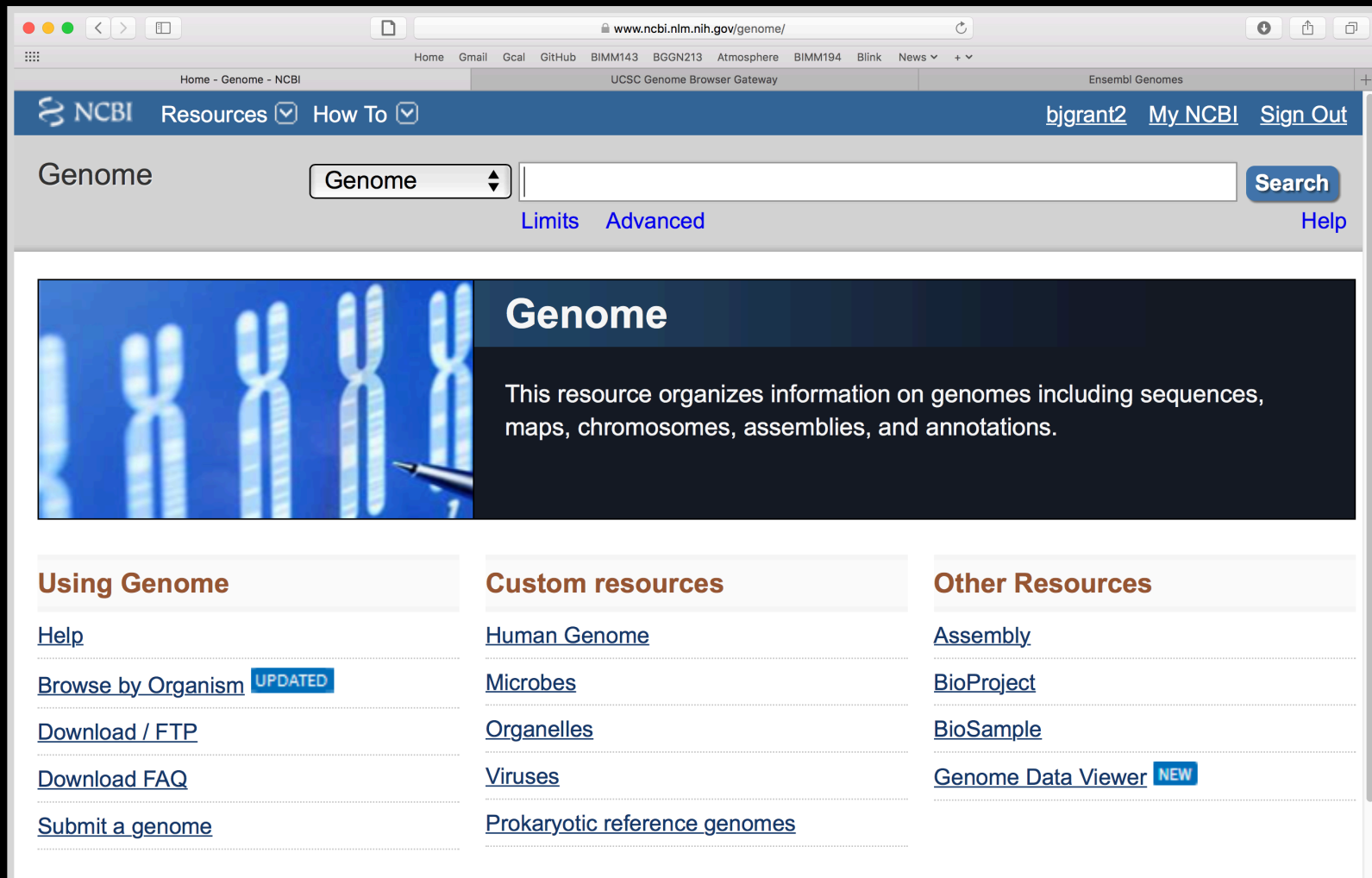
Genomes come in many sizes



Genome Databases

NCBI Genome:

<http://www.ncbi.nlm.nih.gov/genome>



The screenshot shows the NCBI Genome website interface. At the top, there is a navigation bar with the NCBI logo, "Resources" and "How To" dropdown menus, and user links for "bjgrant2", "My NCBI", and "Sign Out". Below this is a search bar with a "Genome" dropdown menu, a search input field, and a "Search" button. There are also links for "Limits" and "Advanced" search options, and a "Help" link. The main content area features a large banner with a blue background and white text. The banner includes an image of chromosomes and a pen, and the heading "Genome". Below the heading, a paragraph states: "This resource organizes information on genomes including sequences, maps, chromosomes, assemblies, and annotations." Below the banner, there are three columns of links. The first column, titled "Using Genome", includes links for "Help", "Browse by Organism" (with an "UPDATED" badge), "Download / FTP", "Download FAQ", and "Submit a genome". The second column, titled "Custom resources", includes links for "Human Genome", "Microbes", "Organelles", "Viruses", and "Prokaryotic reference genomes". The third column, titled "Other Resources", includes links for "Assembly", "BioProject", "BioSample", and "Genome Data Viewer" (with a "NEW" badge).

Home - Genome - NCBI

UCSC Genome Browser Gateway

Ensembl Genomes

NCBI Resources How To

bjgrant2 My NCBI Sign Out

Genome

Genome

Search

Limits Advanced Help

Genome

This resource organizes information on genomes including sequences, maps, chromosomes, assemblies, and annotations.

Using Genome

[Help](#)

[Browse by Organism](#) **UPDATED**

[Download / FTP](#)

[Download FAQ](#)

[Submit a genome](#)

Custom resources

[Human Genome](#)

[Microbes](#)

[Organelles](#)

[Viruses](#)

[Prokaryotic reference genomes](#)

Other Resources

[Assembly](#)

[BioProject](#)

[BioSample](#)

[Genome Data Viewer](#) **NEW**

Genome Databases

(EBI) Ensembl Genomes:

<http://ensemblgenomes.org>

The screenshot shows a web browser window displaying the Ensembl Genomes website. The browser's address bar shows 'ensemblgenomes.org'. The website's header includes the Ensembl Genomes logo and navigation links: 'About us', 'Genomes', 'Data types', 'Data access', and 'FAQs'. A secondary navigation bar lists taxonomic groups: 'Bacteria | Protists | Fungi | Plants | Metazoa | Vertebrates'. The main content area features a blue box with the text 'Ensembl Genomes: Extending Ensembl across the taxonomic space.' Below this is a grid of 15 small images representing various organisms. To the right, there is a section titled 'What's New in Release 46 (January 2020)' with sub-sections for 'Ensembl Bacteria', 'Ensembl Fungi', and 'Ensembl Metazoa'. A yellow box on the right contains the text 'Have a question?' followed by 'Frequently Asked Questions (FAQs) are now available for all domains of Ensembl Genomes. Have a question? Check if it's been asked before! If there is a FAQ missing, contact us.'

ensemblgenomes.org

Home - Genome - NCBI UCSC Genome Browser Gateway Ensembl Genomes

e!EnsemblGenomes

About us | Genomes | Data types | Data access | FAQs

Bacteria | Protists | Fungi | Plants | Metazoa | Vertebrates

Ensembl Genomes: Extending Ensembl across the taxonomic space.

What's New in Release 46 (January 2020)

Ensembl Bacteria

Release 46 of [EnsemblBacteria](#) has updated pan-taxonomic gene trees and homologies (which includes key bacterial species). There are no other significant changes from the last release to the genomes and genes.

Ensembl Fungi

Release 46 of Ensembl Fungi has updated protein features, BioMarts and pan-taxonomic compara data.

Ensembl Metazoa

Release 46 of Ensembl Metazoa adds the genomes from [VectorBase.org](#) and *Drosophila melanogaster* (BDGP6.28 FB2019_03) update.

Ensembl Plants

Have a question?

Frequently Asked Questions ([FAQs](#)) are now available for all domains of Ensembl Genomes. Have a question? Check if it's been asked before! If there is a FAQ missing, [contact us](#).

Genome Databases

UCSC Genome Browser Gateway:

<https://genome.ucsc.edu/>

The screenshot shows the UCSC Genome Browser Gateway website. The browser's address bar displays `genome.ucsc.edu/cgi-bin/hgGateway`. The page header includes the University of California Santa Cruz Genomics Institute logo and the UCSC Genome Browser Gateway title. A navigation menu contains links for Genomes, Genome Browser, Tools, Mirrors, Downloads, My Data, Projects, Help, and About Us. The main content area is divided into two sections: "Browse/Select Species" and "Find Position".

Browse/Select Species

POPULAR SPECIES

- Human
- Mouse
- Rat
- Zebrafish
- Fruitfly
- Worm
- Yeast

Enter species or common name

REPRESENTED SPECIES

- Human
- Chimp
- Bonobo
- Gorilla
- Orangutan
- Gibbon
- Green monkey
- Crab-eating macaque
- Rhesus
- Baboon (anubis)
- Baboon (hamadryas)
- Proboscis monkey
- Golden snub-nosed monkey
- Marmoset
- Squirrel monkey
- Tarsier
- Mouse lemur
- Bushbaby
- Mouse
- Rat

Find Position

Human Assembly

Feb. 2009 (GRCh37/hg19)

Position/Search Term

Enter position, gene symbol or search terms

Current position: chr17:38,007,296-38,170,000

Human Genome Browser - hg19 assembly [view sequences](#)

The February 2009 human reference sequence (GRCh37) was produced by the **Genome Reference Consortium**. For more information about this assembly, see **GRCh37** in the NCBI Assembly database.

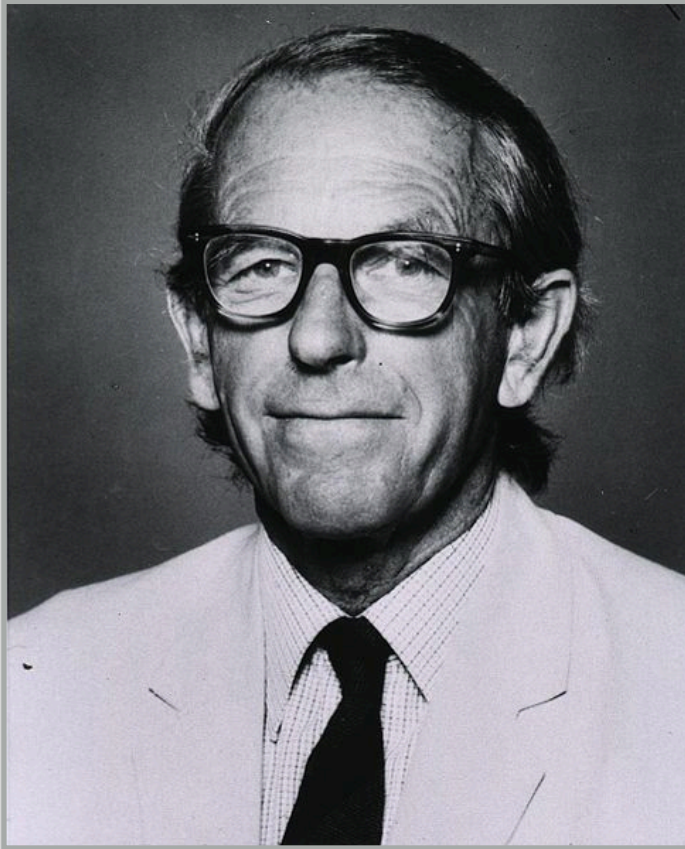
Sample position queries

A genome position can be specified by the accession number of a sequenced genomic clone, an mRNA or EST or STS marker, a chromosomal coordinate range, or keywords from the GenBank description of an mRNA. The following list shows examples of valid position queries for the human genome. See the **User's Guide** for more information.

Request:	Genome Browser Response:
chr7	Displays all of chromosome 7
chrUn_gl000212	Displays all of the unplaced contig gl000212
20p13	Displays region for band p13 on chr 20
chr3:1-1000000	Displays first million bases of chr 3, counting from p-arm telomere
chr3:1000000+2000	Displays a region of chr3 that spans 2000 bases, starting with position 1000000
RH18061;RH80175	Displays region between genome landmarks, such as the STS markers RH18061 and RH80175, or chromosome bands 15q11 to 15q13, or SNPs rs1042522 and rs1800370. This syntax may also be used for other range queries, such as between uniquely determined ESTs, mRNAs, refSeqs, etc.
15q11:15q13	
rs1042522;rs1800370	
D16S3046	Displays region around STS marker D16S3046 from the Genethon/Marshfield maps.

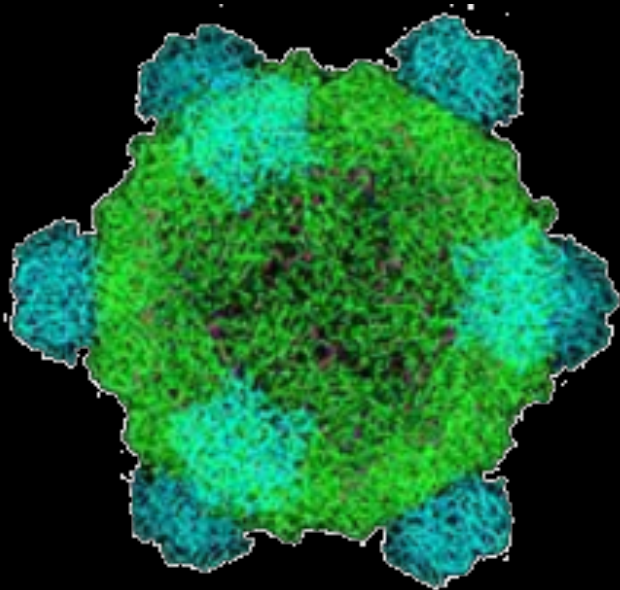
Homo sapiens
(Graphic courtesy of CBSE)

Early Genome Sequencing



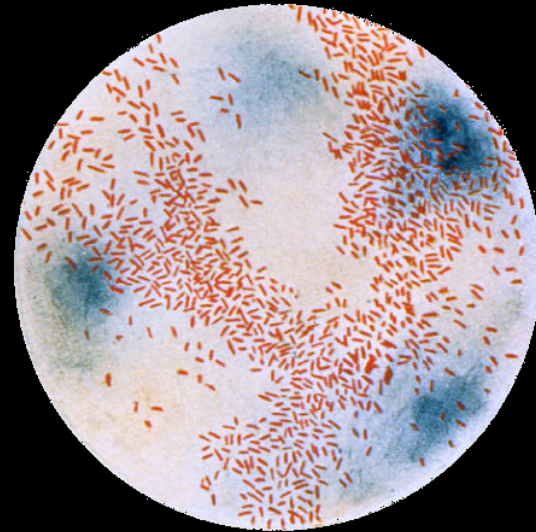
- Chain-termination “**Sanger**” sequencing was developed in 1977 by *Frederick Sanger*, colloquially referred to as the “Father of Genomics”
- Sequence reads were typically 750-1000 base pairs in length with an error rate of $\sim 1 / 10000$ bases

The First Sequenced Genomes



Bacteriophage φ-X174

- Completed in 1977
- 5,386 base pairs, ssDNA
- 11 genes

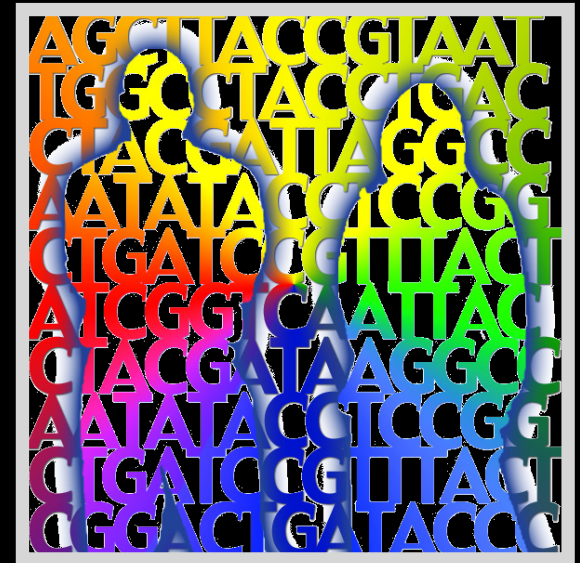


Haemophilus influenzae

- Completed in 1995
- 1,830,140 base pairs, dsDNA
- 1,740 genes

The Human Genome Project

- The Human Genome Project (HGP) was an international, public consortium that began in 1990
 - Initiated by James Watson
 - Primarily led by Francis Collins
 - Eventual Cost: \$2.7 Billion
- Celera Genomics was a private corporation that started in 1998
 - Headed by Craig Venter
 - Eventual Cost: \$300 Million
- Both initiatives released initial drafts of the human genome in 2001
 - ~3.2 Billion base pairs, dsDNA
 - ~20,400 coding (& ~24,000 non-coding) genes*



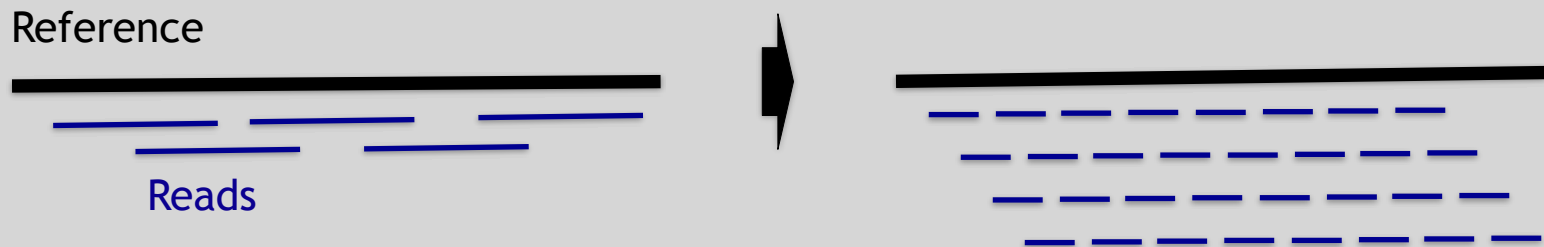
HHMI



DeCode Genetics INC.

Modern Genome Sequencing

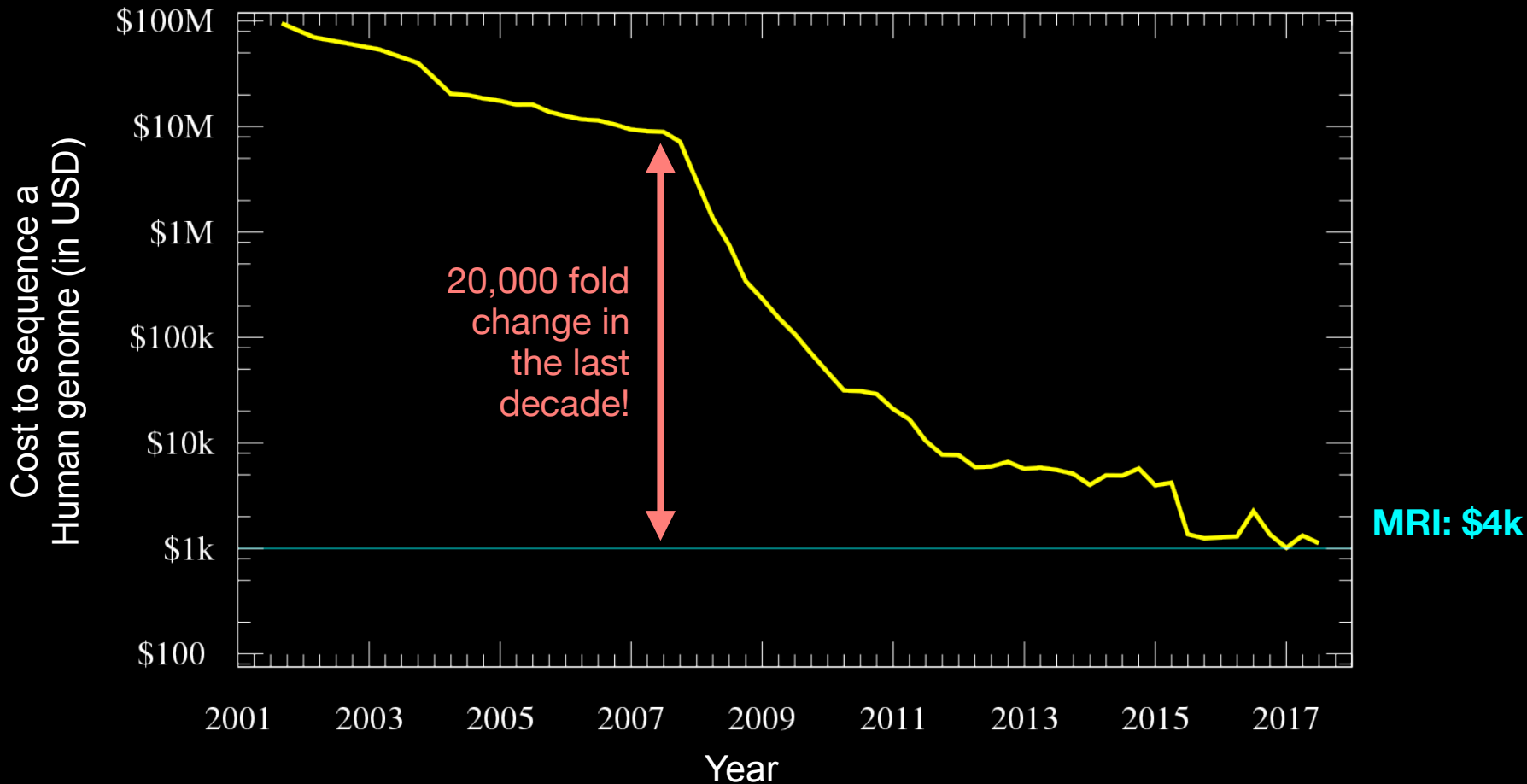
- Next Generation Sequencing (NGS) technologies have resulted in a paradigm shift from long reads at low coverage to short reads at high coverage
- This provides numerous opportunities for new and expanded genomic applications



Rapid progress of genome sequencing



Rapid progress of genome sequencing



Major impact areas for genomic medicine

- **Cancer**: Identification of driver mutations and drugable variants, Molecular stratification to guide and monitor treatment, Identification of tumor specific variants for personalized immunotherapy approaches (precision medicine).
- **Genetic disease diagnose**: Rare, inherited and so-called 'mystery' disease diagnose.
- **Health management**: Predisposition testing for complex diseases (e.g. cardiac disease, diabetes and others), optimization and avoidance of adverse drug reactions.
- **Health data analytics**: Incorporating genomic data with additional health data for improved healthcare delivery.
- Prenatal testing, transplant rejection, pathogen detection, microbiome etc.

Goals of Cancer Genome Research

- Identify changes in the genomes of tumors that drive cancer progression
- Identify new targets for therapy
- Select drugs based on the genomics of the tumor
- Provide early cancer detection and treatment response monitoring
- Utilize cancer specific mutations to derive neoantigen immunotherapy approaches



What can go wrong in cancer genomes?

Type of change	Some common technology to study changes
DNA mutations	WGS, WXS
DNA structural variations	WGS
Copy number variation (CNV)	CGH array, SNP array, WGS
DNA methylation	Methylation array, RRBS, WGBS
mRNA expression changes	mRNA expression array, RNA-seq
miRNA expression changes	miRNA expression array, miRNA-seq
<i>Protein expression</i>	Protein arrays, mass spectrometry

WGS = whole genome sequencing, WXS = whole exome sequencing

RRBS = reduced representation bisulfite sequencing, WGBS = whole genome bisulfite sequencing

DNA Sequencing Concepts

- **Sequencing by Synthesis:** Uses a polymerase to incorporate and assess nucleotides to a primer sequence
 - 1 nucleotide at a time
- **Sequencing by Ligation:** Uses a ligase to attach hybridized sequences to a primer sequence
 - 1 or more nucleotides at a time (e.g. dibase)

Modern NGS Sequencing Platforms

	Roche/454	Life Technologies SOLiD	Illumina Hi-Seq 2000
Library amplification method	emPCR* on bead surface	emPCR* on bead surface	Enzymatic amplification on glass surface
Sequencing method	Polymerase-mediated incorporation of unlabelled nucleotides	Ligase-mediated addition of 2-base encoded fluorescent oligonucleotides	Polymerase-mediated incorporation of end-blocked fluorescent nucleotides
Detection method	Light emitted from secondary reactions initiated by release of PPI	Fluorescent emission from ligated dye-labelled oligonucleotides	Fluorescent emission from incorporated dye-labelled nucleotides
Post incorporation method	NA (unlabelled nucleotides are added in base-specific fashion, followed by detection)	Chemical cleavage removes fluorescent dye and 3' end of oligonucleotide	Chemical cleavage of fluorescent dye and 3' blocking group
Error model	Substitution errors rare, insertion/deletion errors at homopolymers	End of read substitution errors	End of read substitution errors
Read length (fragment/paired end)	400 bp/variable length mate pairs	75 bp/50+25 bp	150 bp/100+100 bp

Illumina now dominates the sequencing market

- Today more than 90% of all sequencing is done on illumina machines
- Generating millions to billions of reads per run (machine dependent)
- High fidelity (>99.9% accuracy for short ~300 bp reads)
- \$1,000 per human genome in 48 hours*

Illumina now dominates the sequencing market

- Today more than 90% of all sequencing is done on illumina machines

MiSeq



(30 million read)

NextSeq



(3 billion reads)

NovaSeq



(13 billion reads)

Illumina Flow Cells



- MiSeq (1-30 million read)
- NextSeq (3 billion reads)
- NovaSeq (13 billion reads)

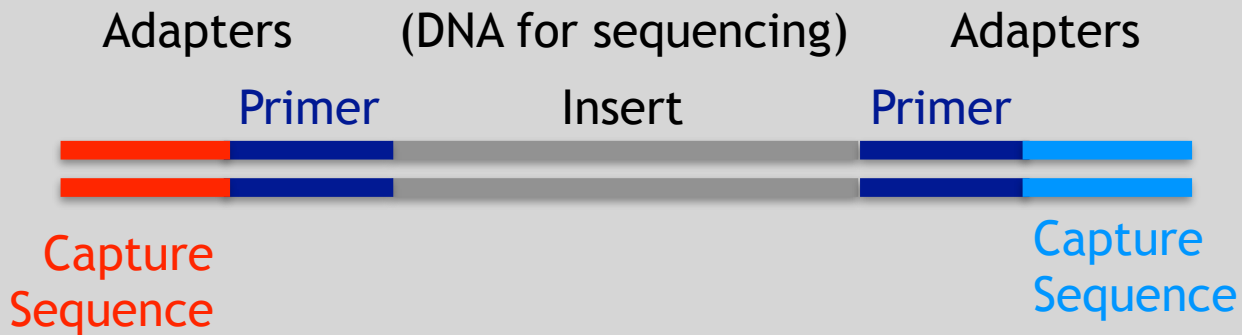
Preparing Samples

(DNA for sequencing)

Insert



Preparing Samples



Adapters are required for sequencing

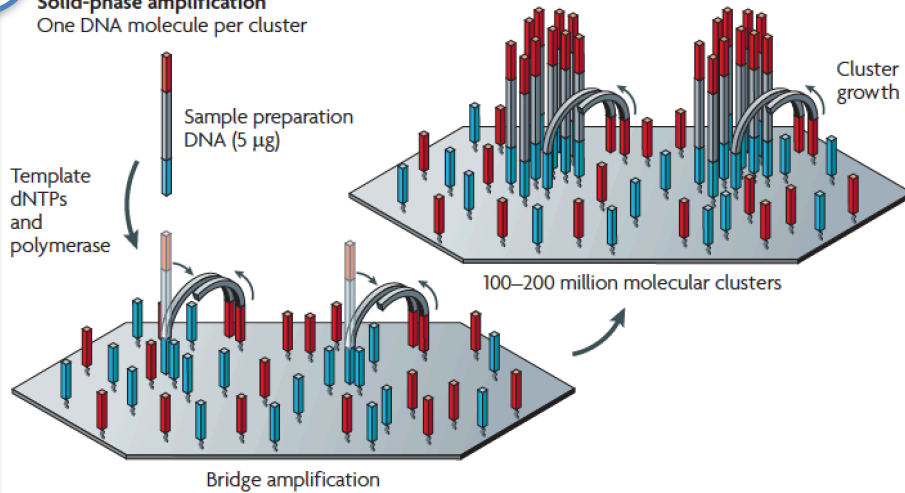
Adapter sequences include primer binding sites and capture sequences

Illumina - Reversible terminators

1

Enzymatic amplification on glass surface

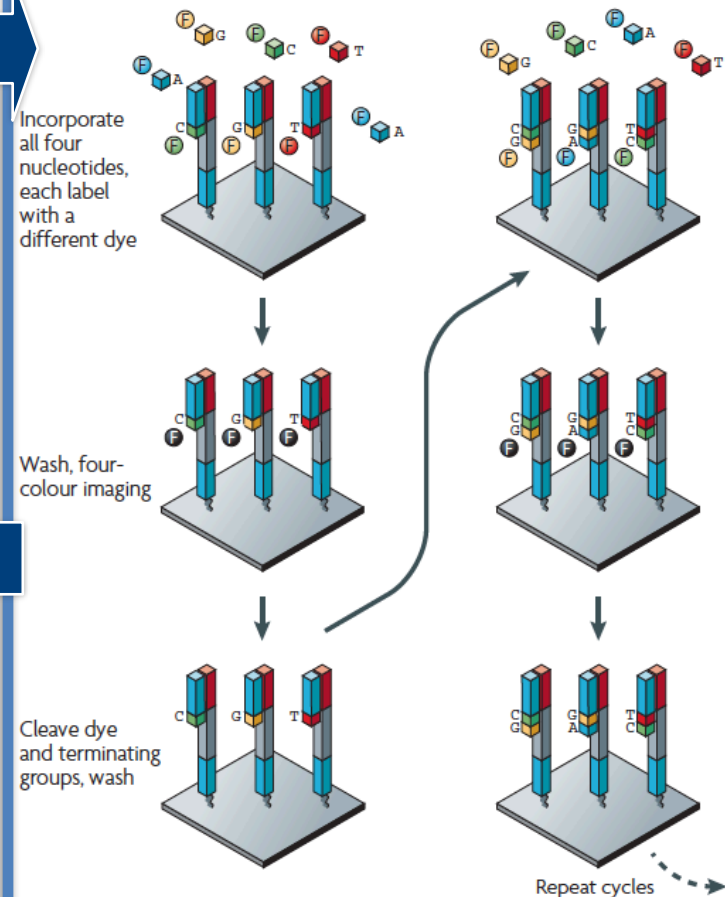
Illumina/Solexa
Solid-phase amplification
One DNA molecule per cluster



2

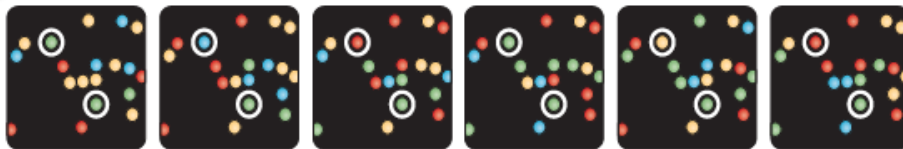
Polymerase-mediated incorporation of end blocked fluorescent nucleotides

Illumina/Solexa — Reversible terminators



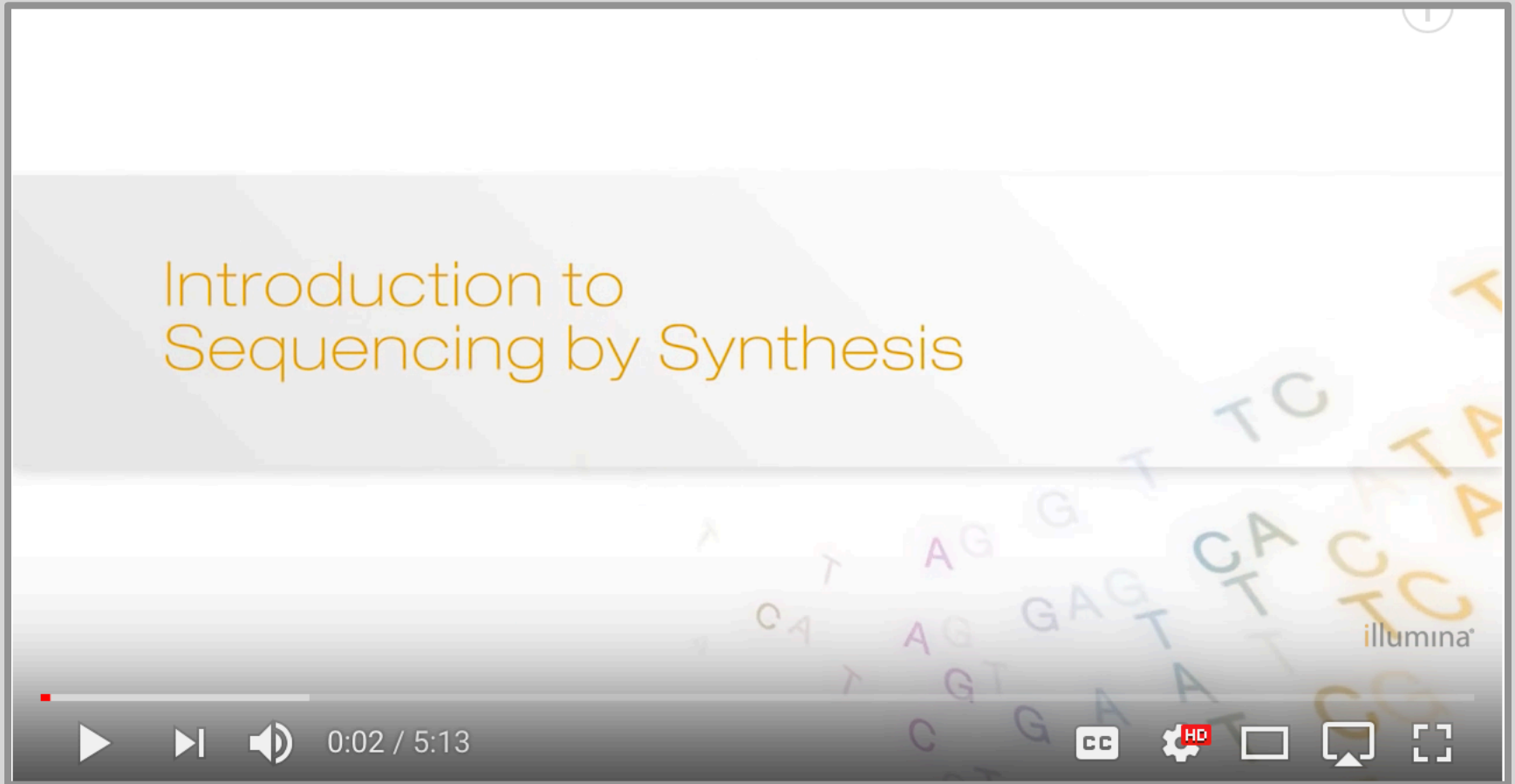
3

Fluorescent emission from incorporated dye-labeled nucleotides



Top: CATCGT
Bottom: CCCCCC

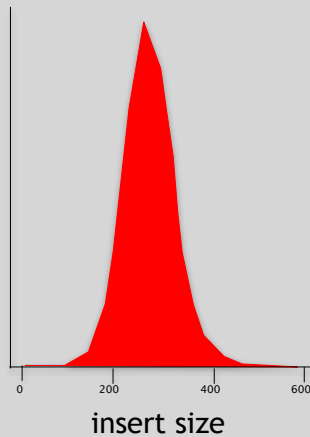
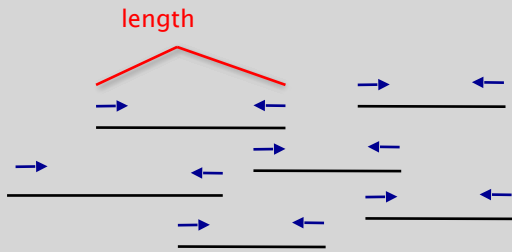
Illumina Sequencing - Video



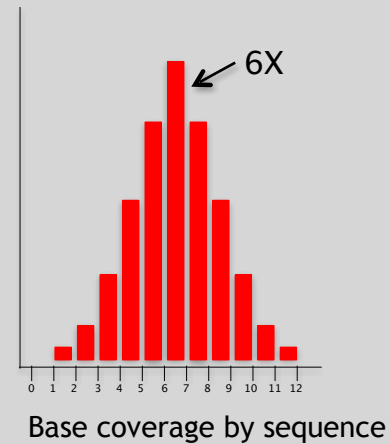
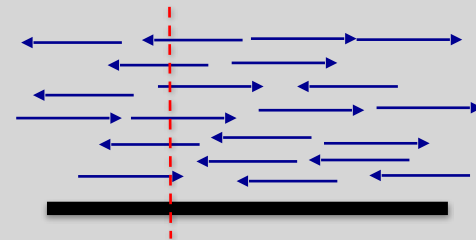
https://www.youtube.com/watch?src_vid=womKfikWlxM&v=fCd6B5HRaZ8

NGS Sequencing Terminology

Insert Size



Sequence Coverage



Terminology: “Generations” of DNA Sequencing

	First generation	Second generation ^a	Third generation ^a
Fundamental technology	Size-separation of specifically end-labeled DNA fragments, produced by SBS or degradation	Wash-and-scan SBS	SBS, by degradation, or direct physical inspection of the DNA molecule
Resolution	Averaged across many copies of the DNA molecule being sequenced	Averaged across many copies of the DNA molecule being sequenced	Single-molecule resolution
Current raw read accuracy	High	High	Moderate
Current read length	Moderate (800–1000 bp)	Short, generally much shorter than Sanger sequencing	Long, 1000 bp and longer in commercial systems
Current throughput	Low	High	Moderate
Current cost	High cost per base Low cost per run	Low cost per base High cost per run	Low-to-moderate cost per base Low cost per run
RNA-sequencing method	cDNA sequencing	cDNA sequencing	Direct RNA sequencing and cDNA sequencing
Time from start of sequencing reaction to result	Hours	Days	Hours
Sample preparation	Moderately complex, PCR amplification not required	Complex, PCR amplification required	Ranges from complex to very simple depending on technology
Data analysis	Routine	Complex because of large data volumes and because short reads complicate assembly and alignment algorithms	Complex because of large data volumes and because technologies yield new types of information and new signal processing challenges
Primary results	Base calls with quality values	Base calls with quality values	Base calls with quality values, potentially other base information such as kinetics

Third Generation Sequencing

- Currently in active development
- Hard to define what “3rd” generation means
- Typical characteristics:
 - Long sequence reads (1,000bp+)
 - Single molecule (no PCR amplification step required)
 - Often associated with "nanopore technology" (e.g. *Oxford Nanopore's MinION USB sequencer*)
 - Note that other approaches are being developed...



The first direct RNA sequencing by nanopore

Side-Note:

- For example this new nanopore direct RNA-sequencing method was published last year:
<https://www.nature.com/articles/nmeth.4577>
- *"Sequencing the RNA in a biological sample can unlock a wealth of information, including the identity of bacteria and viruses, the nuances of alternative splicing or the transcriptional state of organisms. However, **current methods have limitations due to short read lengths and reverse transcription or amplification biases**. Here we demonstrate nanopore direct RNA-seq, a highly parallel, real-time, single-molecule method that circumvents reverse transcription or amplification steps."*

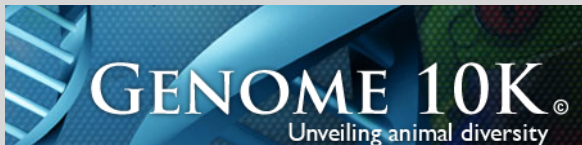
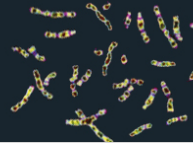
**What can we do with all
this sequence information?**

Population Scale Analysis

We can now begin to assess genetic differences on a very large scale, both as naturally occurring variation in human and non-human populations as well somatically within tumors

1000 Genomes

A Deep Catalog of Human Genetic Variation



The Cancer Genome Atlas



*Understanding genomics
to improve cancer care*

The 100,000 Genomes Project

Genomics England & Partners



<https://www.genomicsengland.co.uk/the-100000-genomes-project/>

“Variety’s the very spice of life”

-William Cowper, 1785

“Variation is the spice of life”

-Kruglyak & Nickerson, 2001

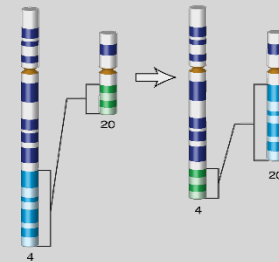
- While the sequencing of the human genome was a great milestone, the DNA from a single person is not representative of the millions of potential differences that can occur between individuals
- These unknown genetic variants could be the cause of many phenotypes such as differing morphology, susceptibility to disease, or be completely benign.

Types of Genomic Variation

- **Single Nucleotide Polymorphisms (SNPs)** - mutations of one nucleotide to another
- **Insertion/Deletion Polymorphisms (INDELs)** - small mutations removing or adding one or more nucleotides at a particular locus
- **Structural Variation (SVs)** - medium to large sized rearrangements of chromosomal DNA

AATCTGAGGCAT
AATCTCAGGCAT

AATCTGAGGCAT
AATCT--AGGCAT



Differences Between Individuals

The average number of genetic differences in the germline between two random humans can be broken down as follows:

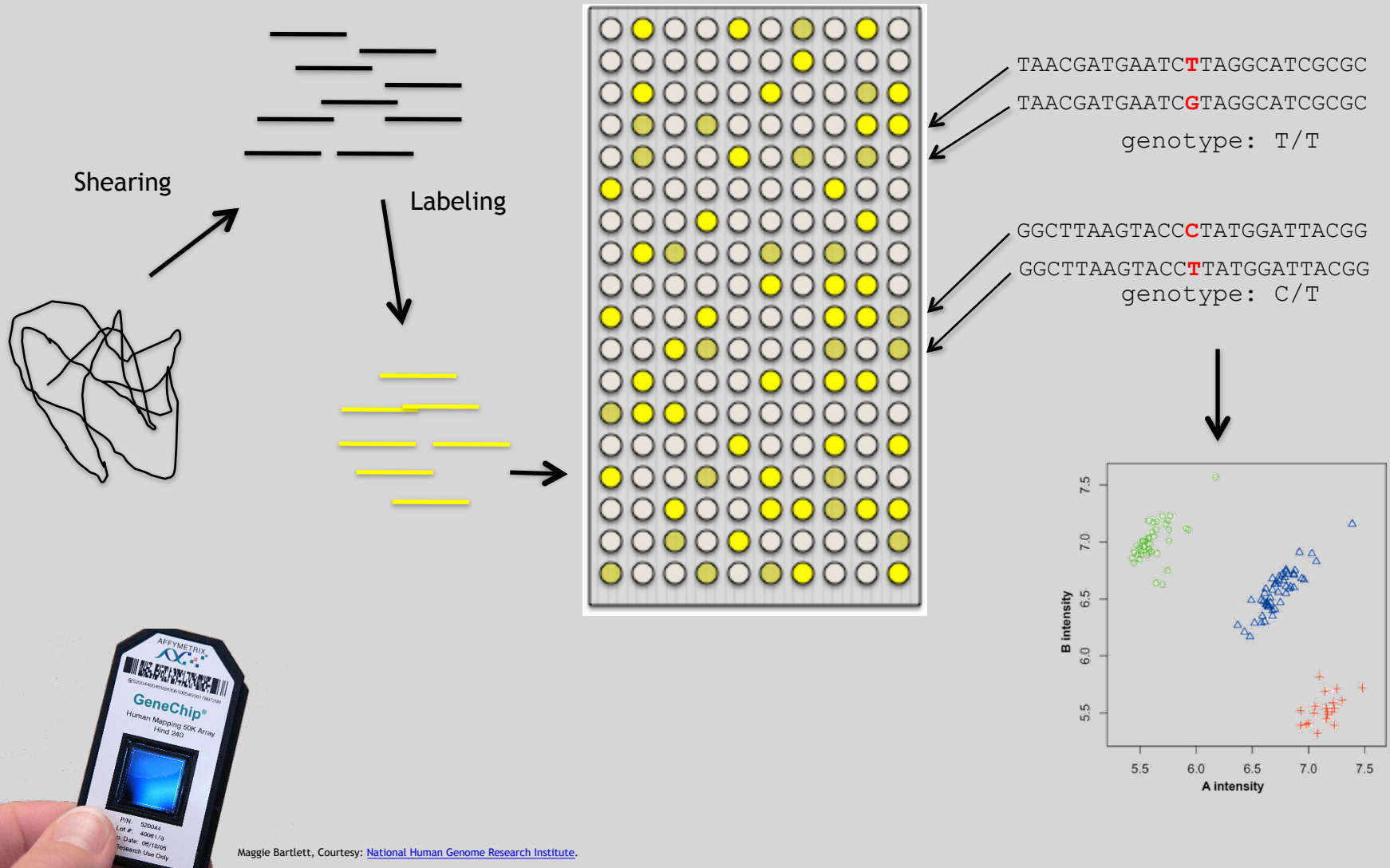
- 3,600,000 single nucleotide differences
- 344,000 small insertion and deletions
- 1,000 larger deletion and duplications

Numbers change depending on ancestry!

Genotyping Small Variants

- Once discovered, oligonucleotide probes can be generated with each individual allele of a variant of interest
- A large number can then be assessed simultaneously on microarrays to detect which combination of alleles is present in a sample

SNP Microarrays



Impact of Genetic Variation

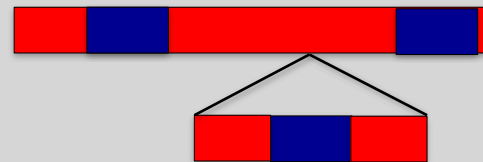
There are numerous ways genetic variation can exhibit functional effects

Premature stop codons



TAC -> TAA

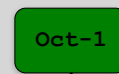
Gene or exon deletion



Frameshift mutation



TAC -> T-C



Transcription factor binding disruption



ATGCAAAT -> ATGCAGAT

Do it Yourself!

Hand-on time!

Sections **1** to **3** please (up to running Read Alignment)
See IP address on website for **your** Galaxy server

<http://uswest.ensembl.org/Help/View?id=140>

The image displays the Ensembl genome browser interface, showing three levels of genomic detail for a region on Chromosome 12 (GRCh38.p3).

Chromosome image: Shows the whole chromosome with a highlighted region of interest (12:25,204,789-25,250,936) and haplotypes/patches. Callouts: "Region of interest", "Haplotypes and patches".

Overview image: Shows a detailed view of the region, including chromosome bands, contigs, and genes. Callouts: "Change or add data tracks" (pointing to the gear icon), "Genes", "Gene or region of interest".

Zoomable Region image: Shows a highly detailed view of the region, including transcripts (splice variants) and protein coding regions. Callouts: "Change or add data tracks" (pointing to the gear icon), "Genome", "Transcripts (splice variants)".

Left sidebar: Contains navigation and configuration options, including "Location-based displays", "Comparative Genomics", "Genetic Variation", "Markers", "Other genes", "UCSC", "NCBI", "Vega", "Ensembl", "Configure this page", "Add", "Export", "Share", and "Bookmark this page".

Access a jetstream galaxy instance!

Use assigned IP address

Do it Yourself!

Galaxy

149.165.169.186

Apps Gmail Seminars Atmosphere BGGN 213 · An intr...

Galaxy Analyze Data Workflow Shared Data Visualization Help User Using 12.3 MB

Tools

search tools

Get Data

Send Data

Collection Operations

Text Manipulation

Filter and Sort

Join, Subtract and Group

Convert Formats

Extract Features

Fetch Sequences

Fetch Alignments

Statistics

Graph/Display Data

FASTA manipulation

NGS: QC and manipulation

NGS: DeepTools

NGS: Mapping

Lastz map short reads against reference sequence

Map with Bowtie for Illumina

Map with BWA for Illumina

Map with BWA for SOLiD

Meqablast compare short reads against htgs, nt, and wgs databases

Parse blast XML output

Map with BWA-MEM - map medium and long reads (> 100 bp) against reference genome

Map with BWA - map short reads (< 100 bp) against reference genome

Bowtie2 - map reads against reference genome

NGS: RNA Analysis

Bowtie2 - map reads against reference genome (Galaxy Version 2.2.6.2)

Options

Is this single or paired library

Single-end

FASTQ file

4: HG00109_2.fastq

Must be of datatype "fastqsanger"

Write unaligned reads (in fastq format) to separate file(s)

Yes No

--un/--un-conc; This triggers --un parameter for single reads and --un-conc for paired reads

Write aligned reads (in fastq format) to separate file(s)

Yes No

--al/--al-conc; This triggers --al parameter for single reads and --al-conc for paired reads

Will you select a reference genome from your history or use a built-in index?

Use a built-in genome index

Built-ins were indexed using default options. See `Indexes` section of help below

Select reference genome

Baboon (Papio anubis): papHam1

If your genome of interest is not listed, contact the Galaxy team

Set read groups information?

Do not set

Specifying read group information can greatly simplify your downstream analyses by allowing combining multiple datasets.

Select analysis mode

1: Default setting only

Do you want to use presets?

No, just use defaults

Very fast end-to-end (--very-fast)

Fast end-to-end (--fast)

Sensitive end-to-end (--sensitive)

Very sensitive end-to-end (--very-sensitive)

Very fast local (--very-fast-local)

Fast local (--fast-local)

Sensitive local (--sensitive-local)

Very sensitive local (--very-sensitive-local)

Allow selecting among several preset parameter settings. Choosing between these will result in dramatic changes in runtime. See help below to understand effects of these presets.

History

search datasets

Unnamed history

22 shown, 2 deleted, 1 hidden

12.32 MB

25: htseq-count on data 18 and data 17 (no feature)

24: htseq-count on data 18 and data 17

23: Cufflinks on data 18 and data 16: Skipped Transcripts

21: Cufflinks on data 18 and data 16: assembled transcripts

20: Cufflinks on data 18 and data 16: transcript expression

19: Cufflinks on data 18 and data 16: gene expression

575 lines

format: tabular, database: hg19

```
cufflinks v2.2.1
cufflinks -q --no-update-check -l
300000 -F 0.100000 -j 0.150000 -p
6 -G /opt/galaxy/galaxy-
app/database/datasets/000/dataset_4
/opt/galaxy/galaxy-
app/database/datasets/000/dataset_4
```

1	2	3
tracking_id	class_code	nearest_ref_id
ZZEF1	-	-
CYB5D2	-	-
ANKFY1	-	-

Raw data usually in FASTQ format

```
@NS500177:196:HFTTTAFXX:1:11101:10916:1458 2:N:0:CGCGGCTG
ACACGACGATGAGGTGACAGTCACGGAGGATAAGATCAATGCCCTCATTAAAGCAGCCGGTGTAA
+
AAAAAEEEEEEEEEEEEEE//AEEEEEEEEEEEEEEEEEE/EE/<<EE/AEEFAEE///EEEEEEEEAEA<
```

1

2

3

4

Each sequencing “read” consists of 4 lines of data :

- 1 The first line (which always starts with ‘@’) is a unique ID for the sequence that follows
- 2 The second line contains the bases called for the sequenced fragment
- 3 The third line is always a “+” character
- 4 The fourth line contains the quality scores for each base in the sequenced fragment (these are ASCII encoded...)

ASCII Encoded Base Qualities

```
@NS500177:196:HFTTTAFXX:1:11101:10916:1458 2:N:0:CGCGGCTG
ACACGACGATGAGGTGACAGTCACGGAGGATAAGATCAATGCCCTCATTAAAGCAGCCGGTGTAA
+
AAAAAEEEEEEEEEEEEEE//AEEEEEEEEEEEEEEEE/EE/<<EE/AEEFAEE///EEEEEEEEAEA<
```

4

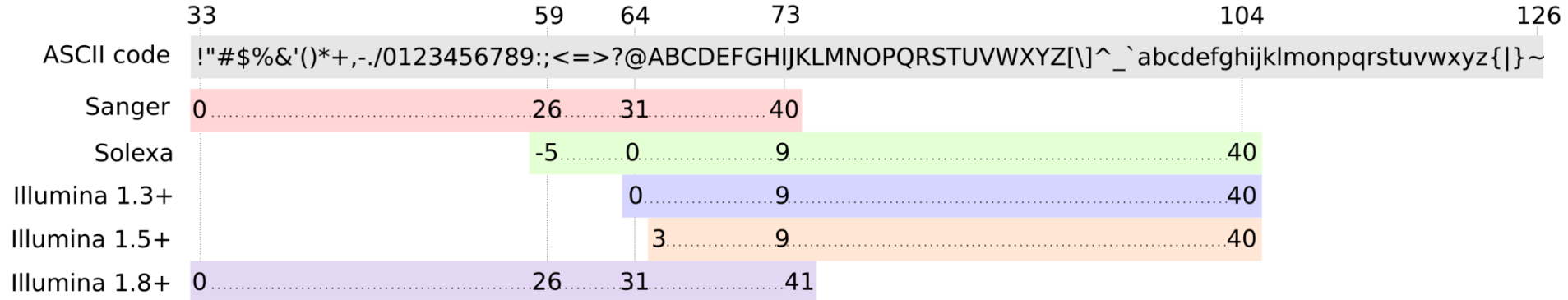
- Each sequence base has a corresponding numeric quality score encoded by a single ASCII character typically on the 4th line (see 4 above)
- ASCII characters represent integers between 0 and 127
- Printable ASCII characters range from 33 to 126
- Unfortunately there are 3 quality score formats that you may come across...

Interpreting Base Qualities in R

		ASCII Range	Offset	Score Range
Sanger, Illumina (Ver > 1.8)	fastqsanger	33-126	33	0-93
Solexa, Illumina (Ver < 1.3)	fastqsolexa	59-126	64	5-62
Illumina (Ver 1.3 -1.7)	fastqillumina	64-126	64	0-62

```
> library(seqinr)
> library(gtools)
> phred <- asc( s2c("DDDDCDEDCDDDBBDDCC@") ) - 33
> phred
## D D D D C D E D C D D D D B B D D D C C @
## 35 35 35 35 34 35 36 35 34 35 35 35 35 33 33 35 35 35 34 34 31
> prob <- 10**(-phred/10)
```

Interpreting Base Qualities in R



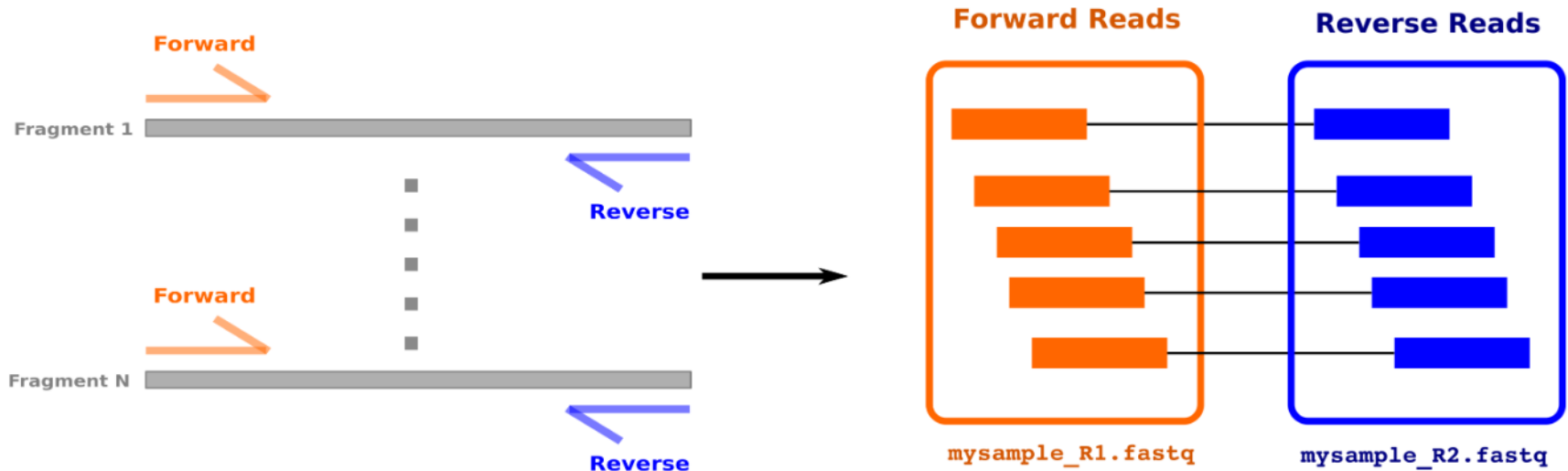
```

> library(seqinr)
> library(gtools)
> phred <- asc( s2c("DDDDCDEDCDDDBBDDCC@") ) - 33
> phred
## D D D D C D E D C D D D D B B D D D C C @
## 35 35 35 35 34 35 36 35 34 35 35 35 35 33 33 35 35 35 34 34 31
> prob <- 10**(-phred/10)

```

Paired-end FASTQ files

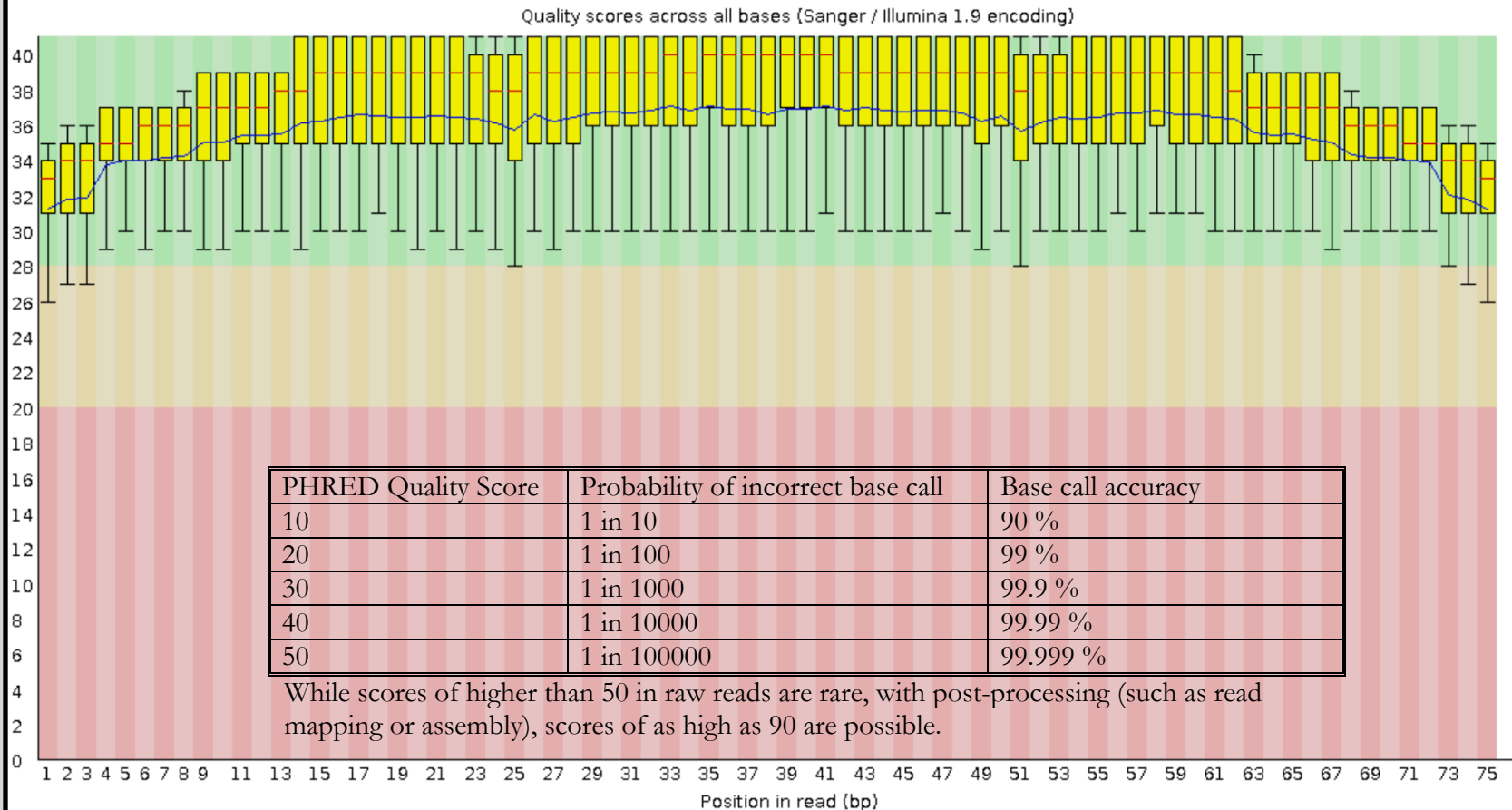
- Sequencer produces two FASTQ files:
 - **Forward** reads (usually **_1** or **_R1** in file name)
 - **Reverse** reads (usually **_2** or **_R2** in file name)



FastQC Report



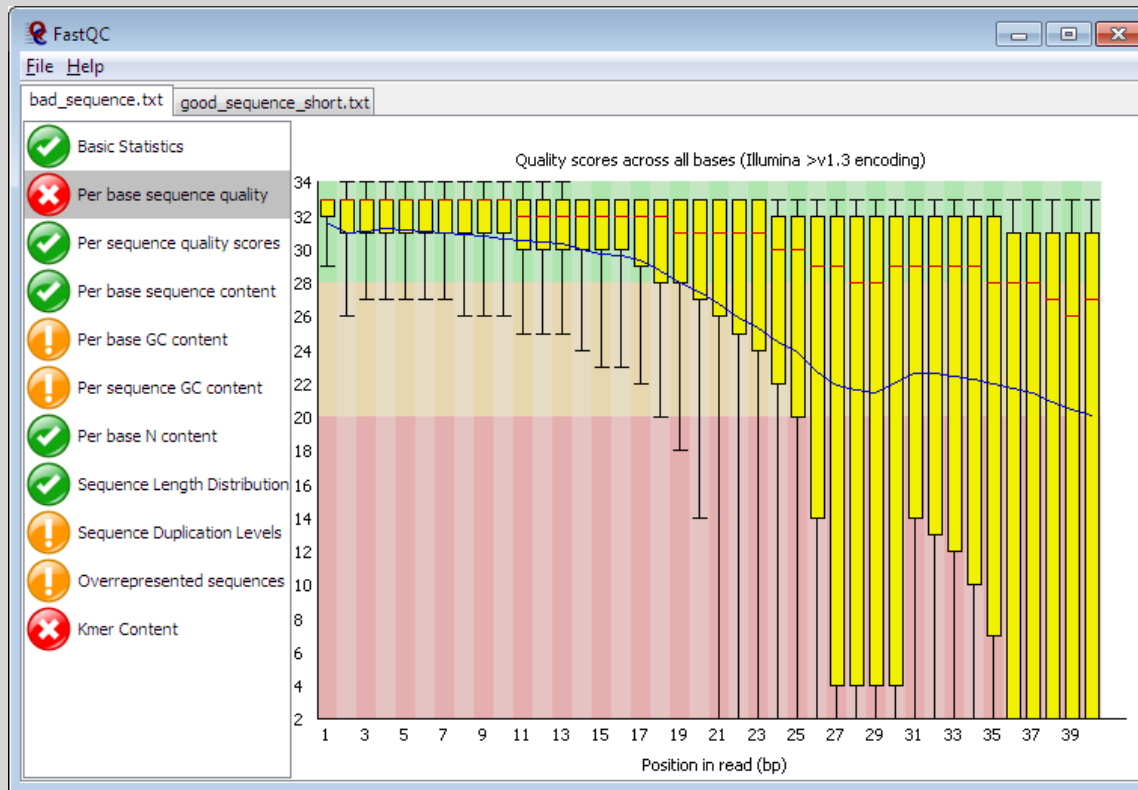
Per base sequence quality



FASTQC

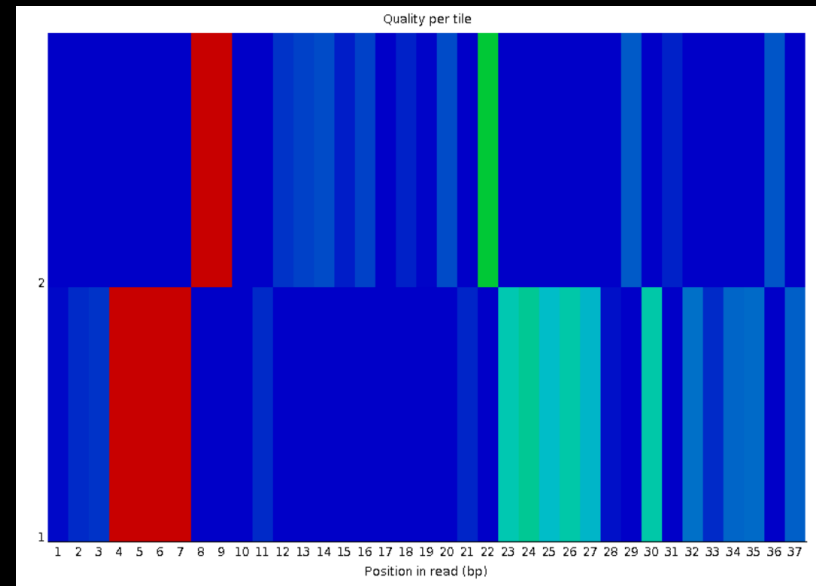
FASTQC is one approach which provides a visual interpretation of the raw sequence reads

- <http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>



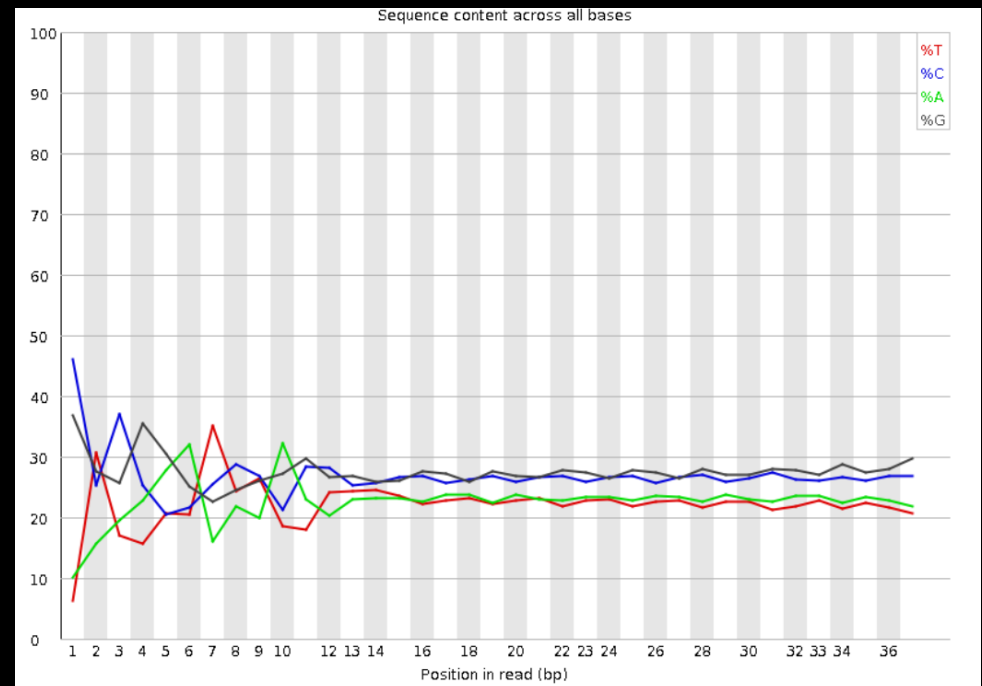
Per Tile Quality shows shows the deviation from the average quality for each tile

- In Illumina libraries the sequence identifier encodes the flowcell tile from which each read came.
- "Hot" colors indicate that a tile had worse quality reads than other tiles for that base
- Suggesting transient problems such as bubbles going through the flowcell, smudges or debris inside the flowcell lane.



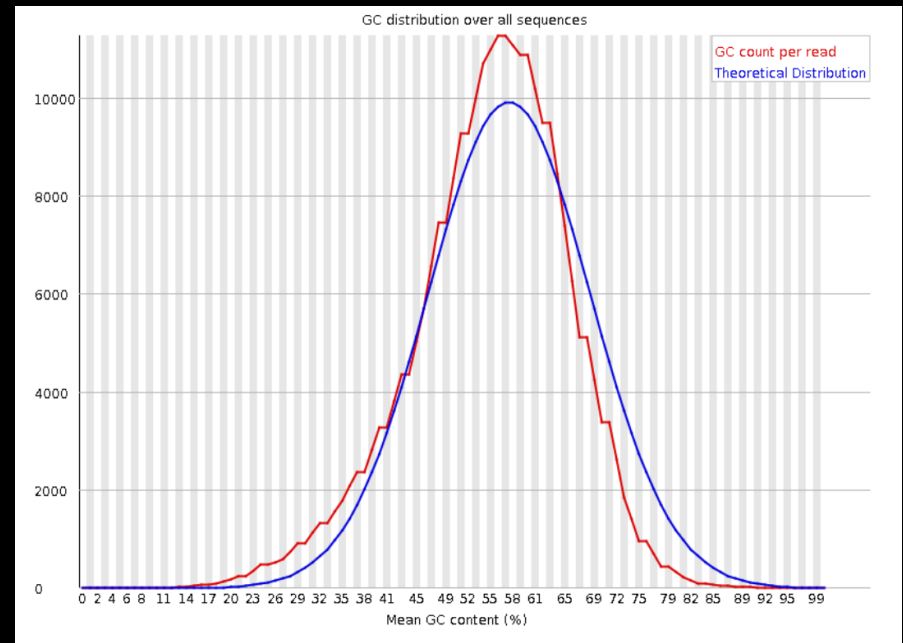
Per-base sequence content highlights the proportion of each base in each position

- In a random library there would be little to no difference between the different bases of a sequence run.
- Note that some types of libraries (e.g. RNA-Seq) will nearly always produce biased sequence composition at the start of the read.



GC content should follow a normal distribution

- An unusually shaped distribution could indicate a contaminated library or some other kinds of biased subset (frequent in metagenomic data sets).
- Sharp peaks on an otherwise smooth distribution are normally the result of a specific contaminant (e.g. adapter dimers)



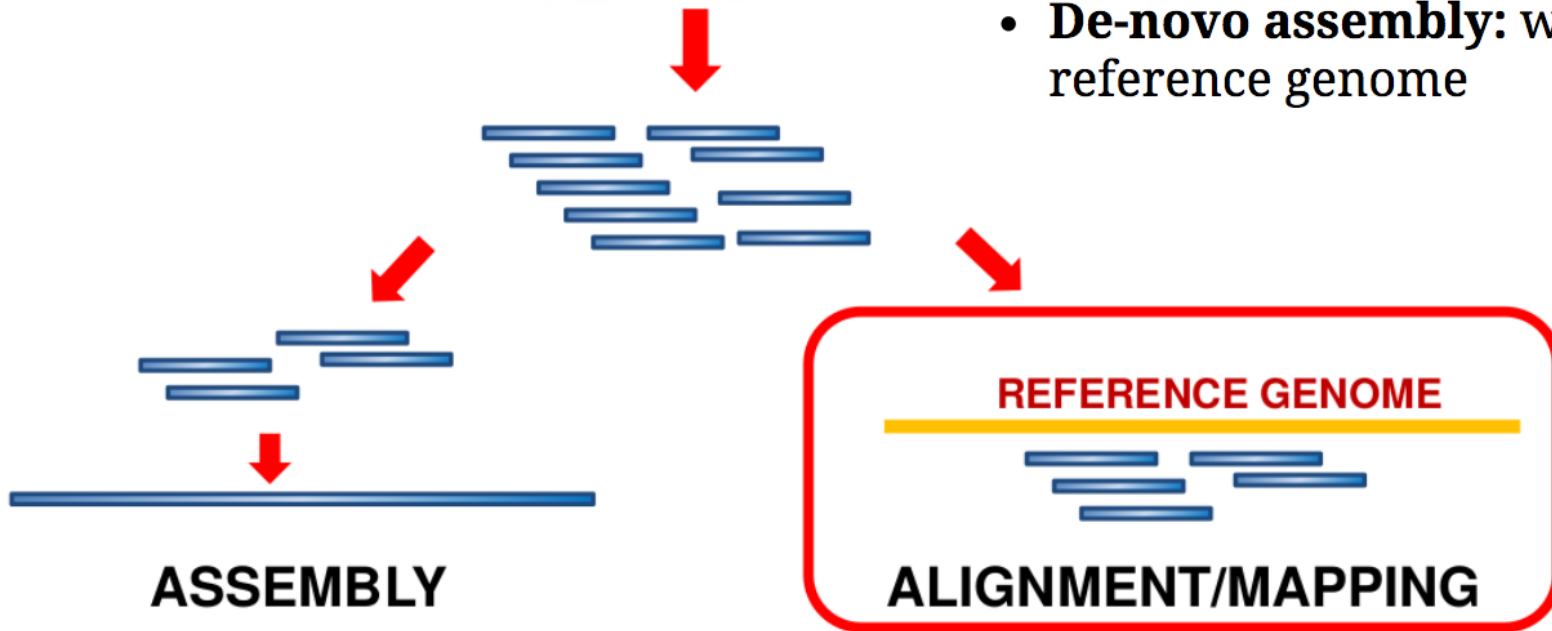
Increasing the quality of sequences

- **Filtering of sequences (i.e. removing sequences):**
 - with small mean quality score
 - with too many N bases
 - based on their GC content
- **Cutting/Trimming sequences from low quality score parts (i.e the tails/ends of reads)**
- Re-run your sequencing job

What is mapping?



- Short reads must be combined into longer fragments
- **Mapping:** use a reference genome as a guide
- **De-novo assembly:** without reference genome



Sequence Alignment

- Once sequence quality has been assessed, the next step is to **align/map** the sequence to a reference genome
- There are *many* distinct tools for doing this; which one you choose is often a reflection of your specific experiment and personal preference

BWA

Bowtie2

SOAP2

Novoalign

mr/mrsFast

Eland

Blat

Bfast

BarraCUDA

CASHx

GSNAP

Mosiak

Stampy

SHRiMP

SeqMap

SLIDER

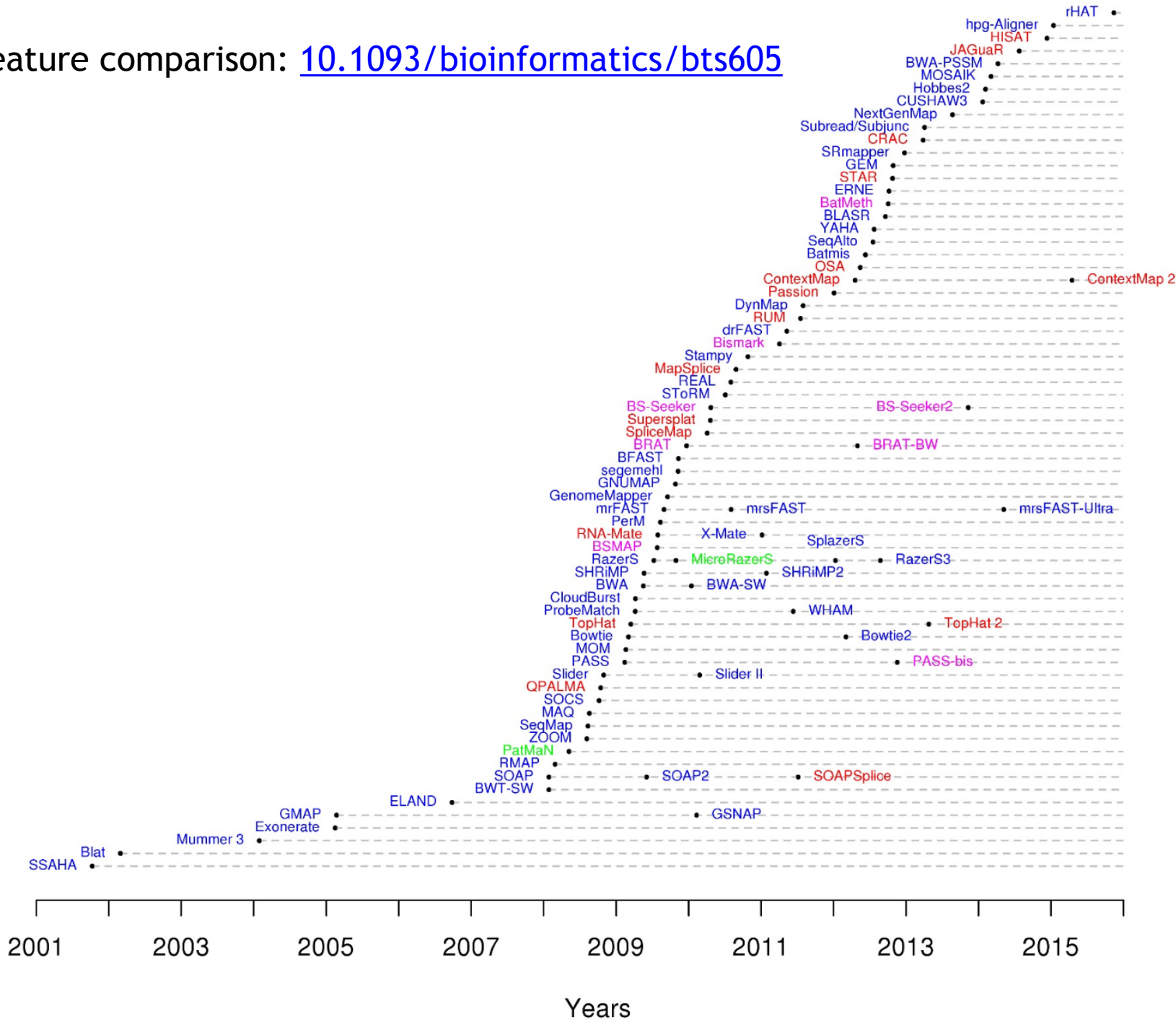
RMAP

SSAHA

etc

Feature comparison: [10.1093/bioinformatics/bts605](https://www.coursera.org/learn/bioinformatics-605)

Feature comparison: 10.1093/bioinformatics/bts605



Inputs

Control

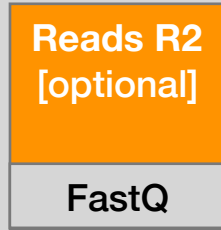
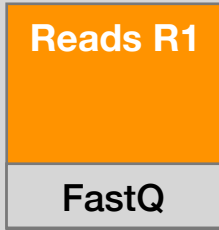
Reads R1
FastQ

Treatment

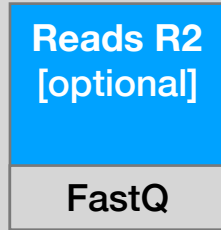
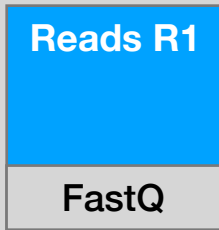
Reads R1
FastQ

Inputs

Control



Treatment

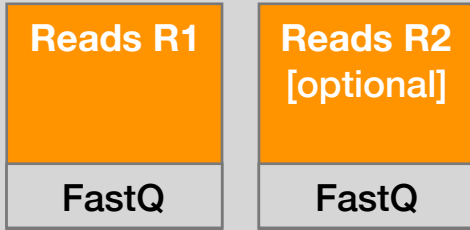


Optional Replicates

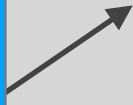
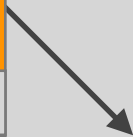
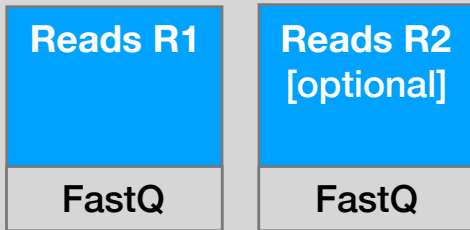
Inputs

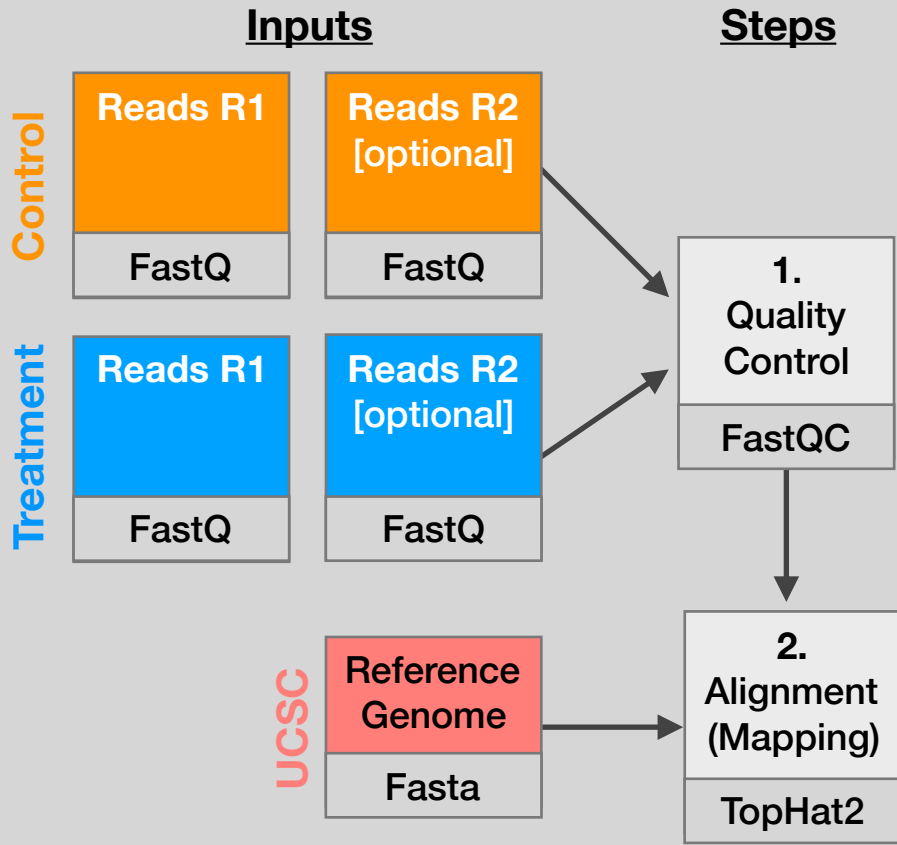
Steps

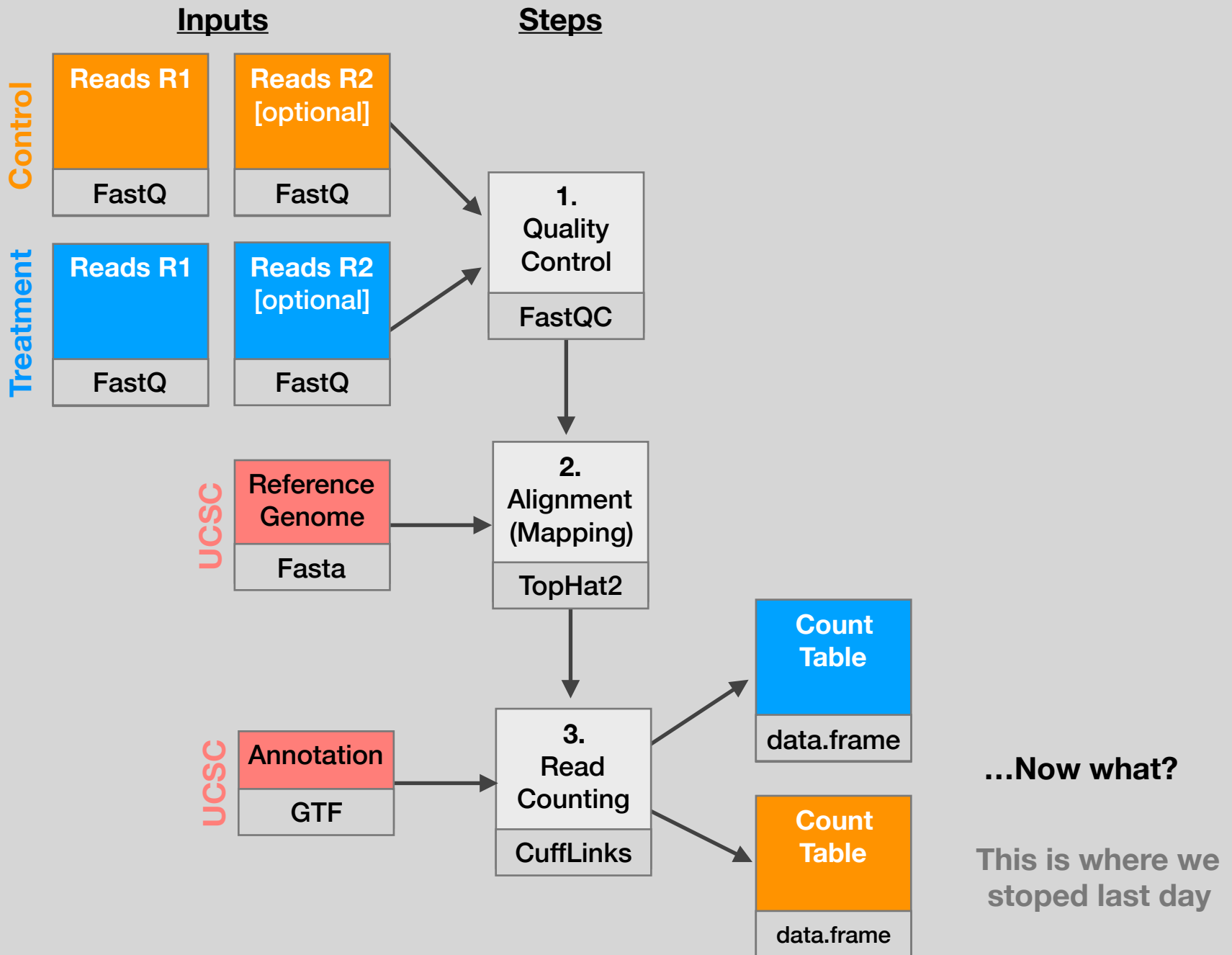
Control

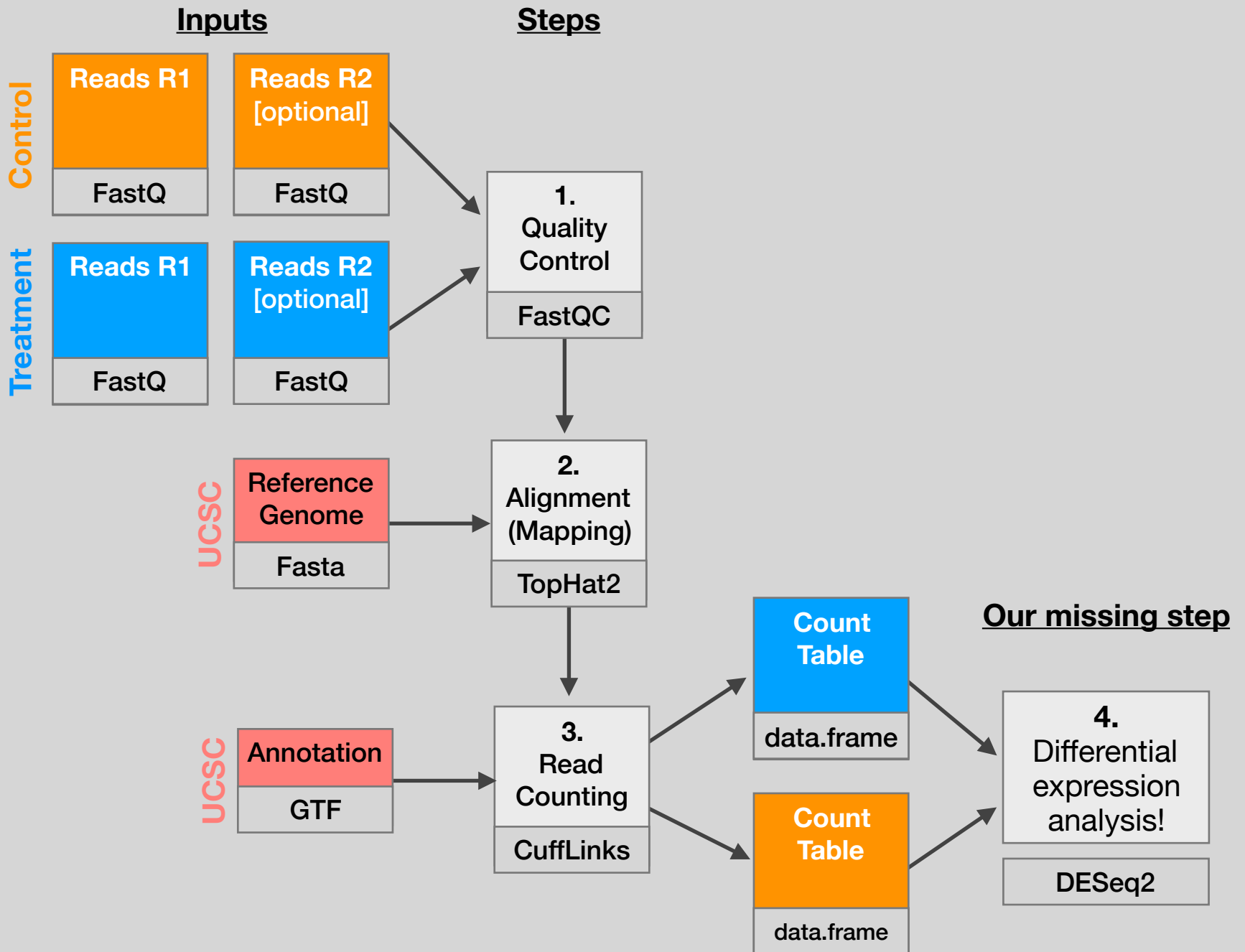


Treatment





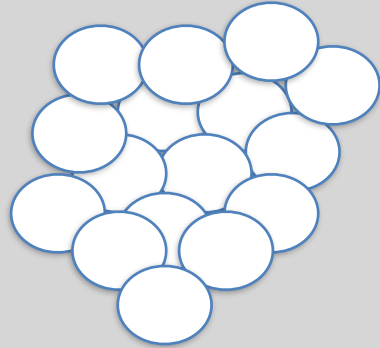




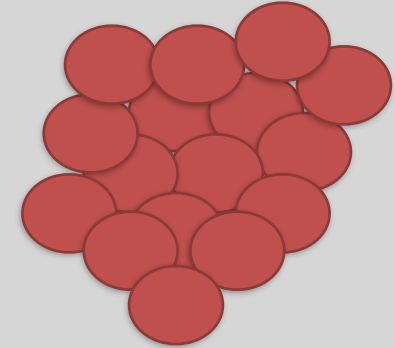
RNA Sequencing

The absolute basics

Normal Cells

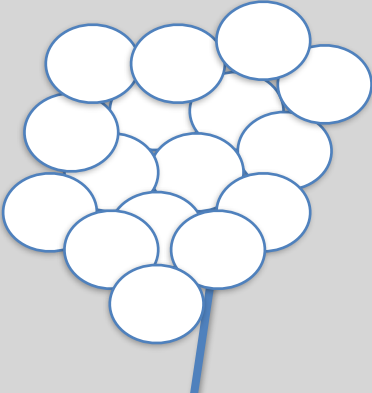


Mutated Cells

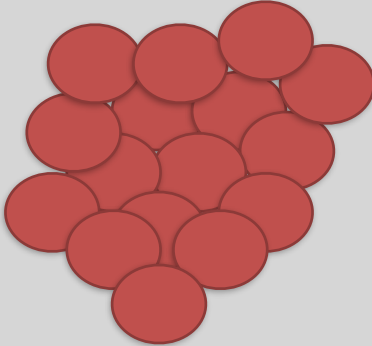


- The **mutated cells** behave differently than the **normal cells**
- We want to know what genetic mechanism is causing the difference
- One way to address this is to examine differences in gene expression via RNA sequencing...

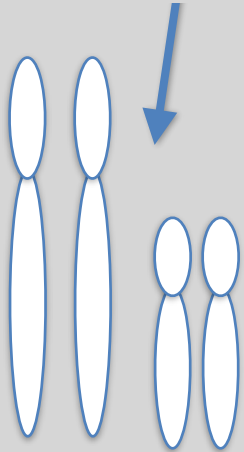
Normal Cells



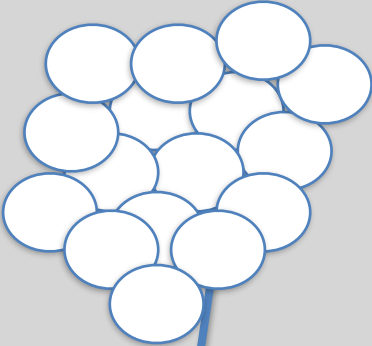
Mutated Cells



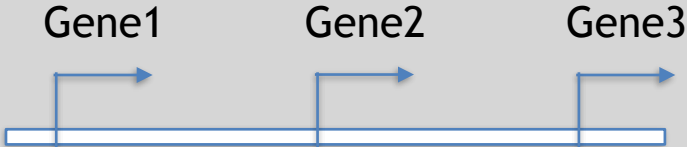
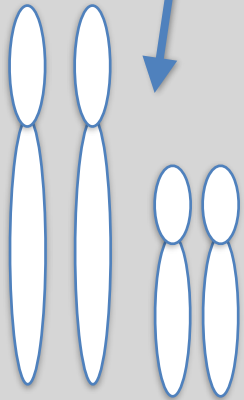
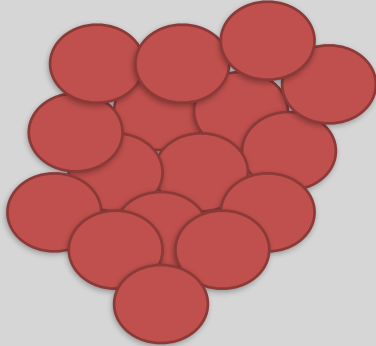
Each cell has a bunch of chromosomes



Normal Cells

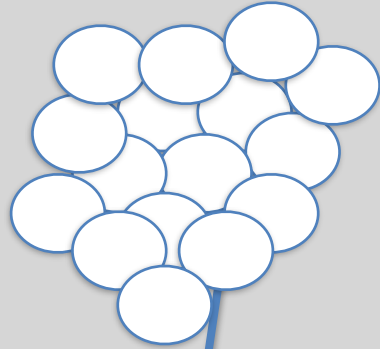


Mutated Cells

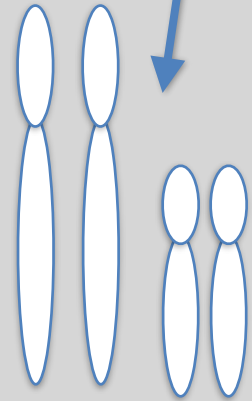
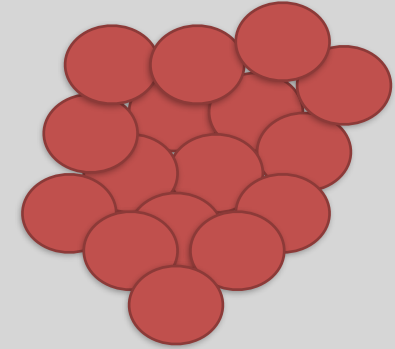


Each chromosome has a bunch of genes

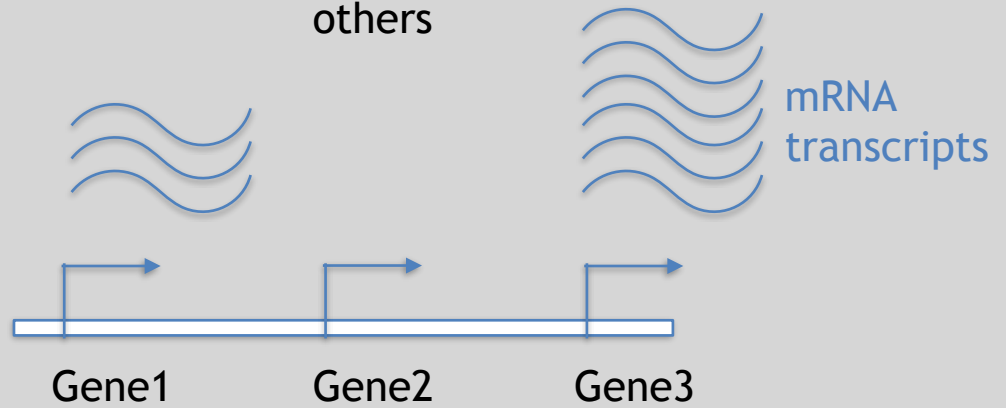
Normal Cells



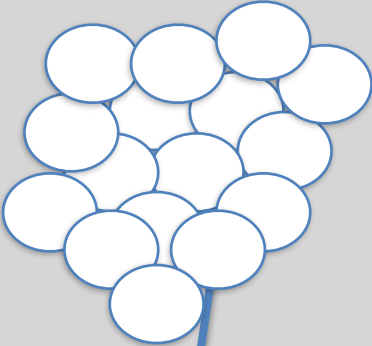
Mutated Cells



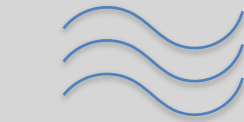
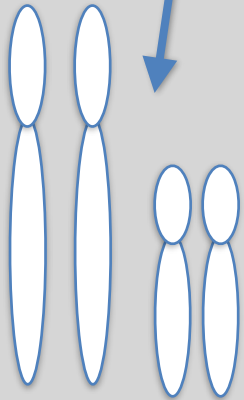
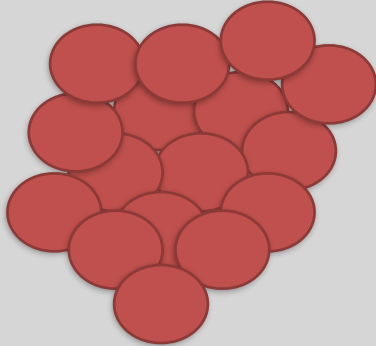
Some genes are active more than others



Normal Cells



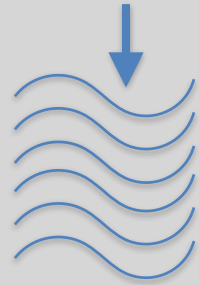
Mutated Cells



Gene 2 is not active



Gene 3 is the most active



mRNA transcripts

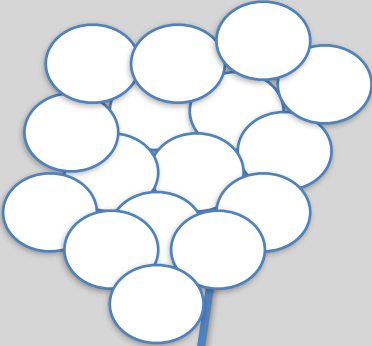


Gene1

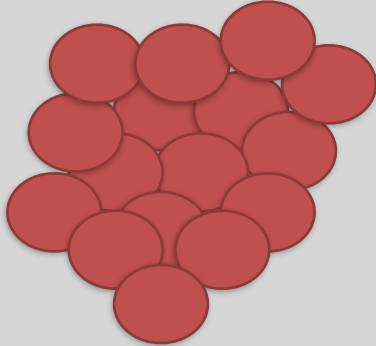
Gene2

Gene3

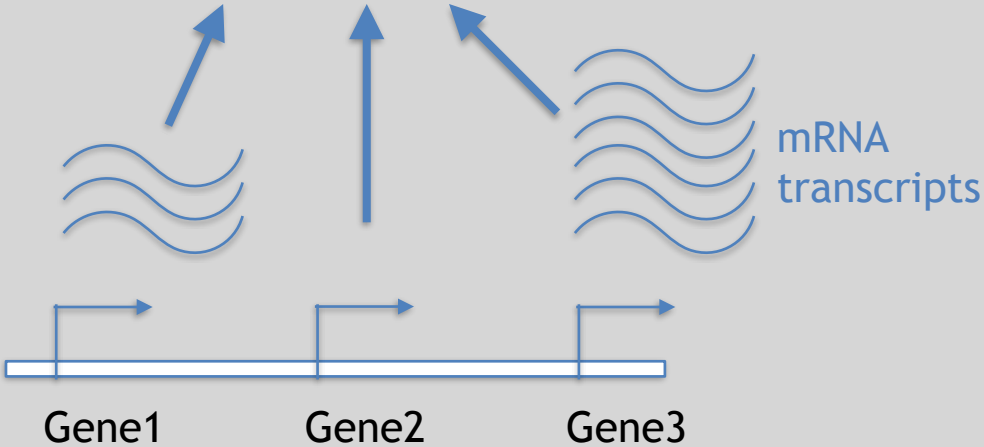
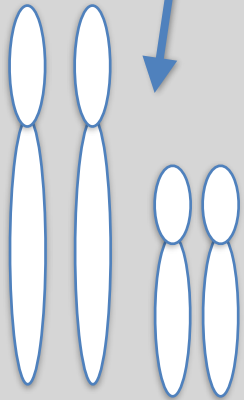
Normal Cells



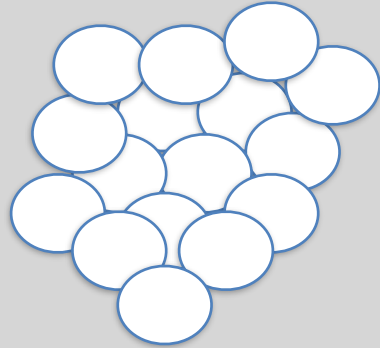
Mutated Cells



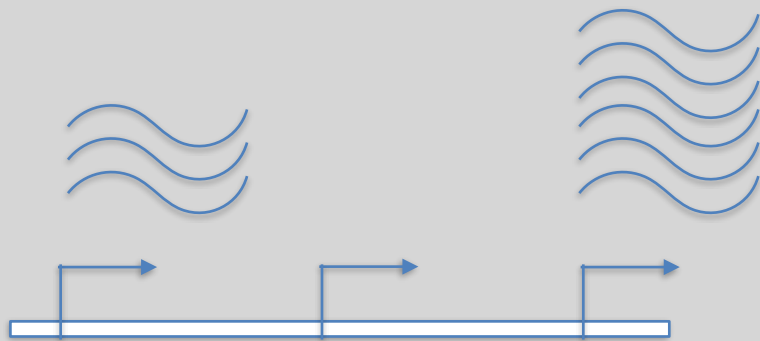
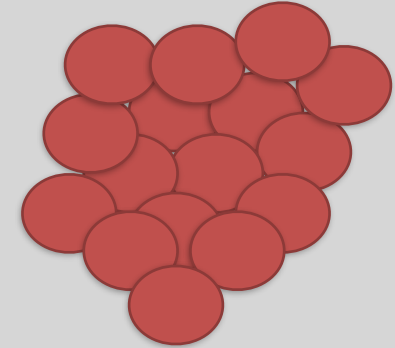
HTS tells us which genes are active, and how much they are transcribed!



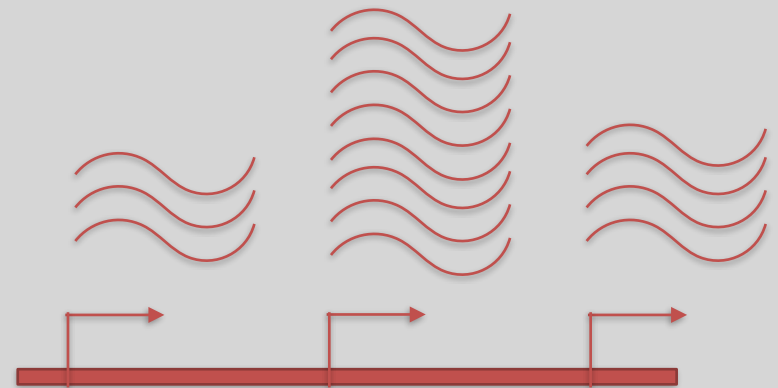
Normal Cells



Mutated Cells

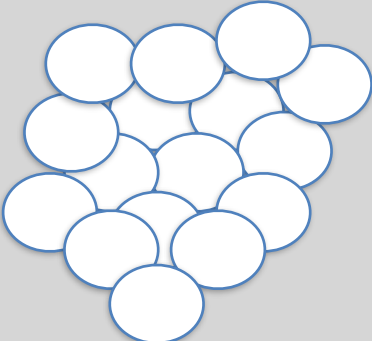


We use RNA-Seq to measure gene expression in normal cells ...

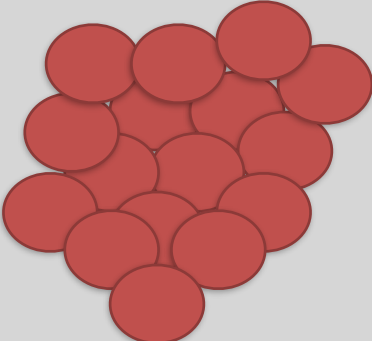


... then use it to measure gene expression in mutated cells

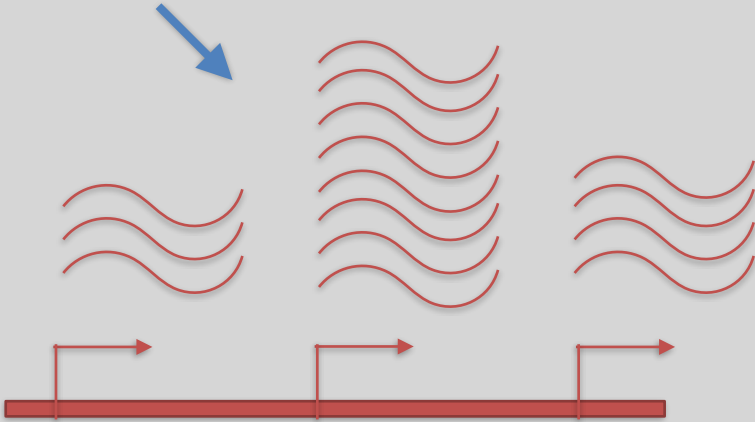
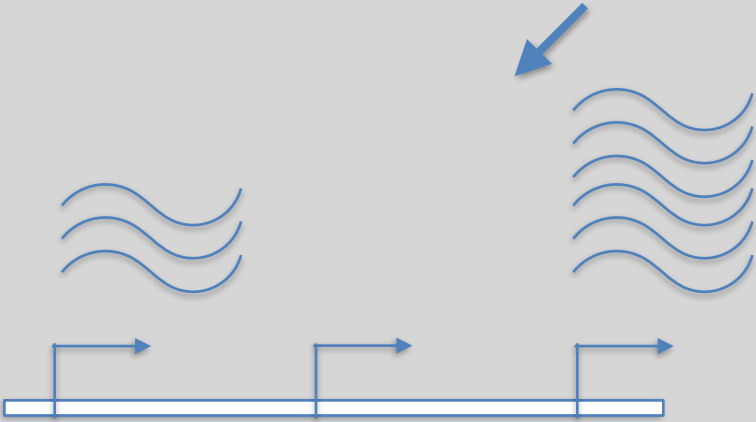
Normal Cells



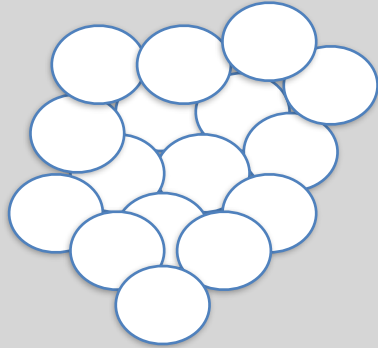
Mutated Cells



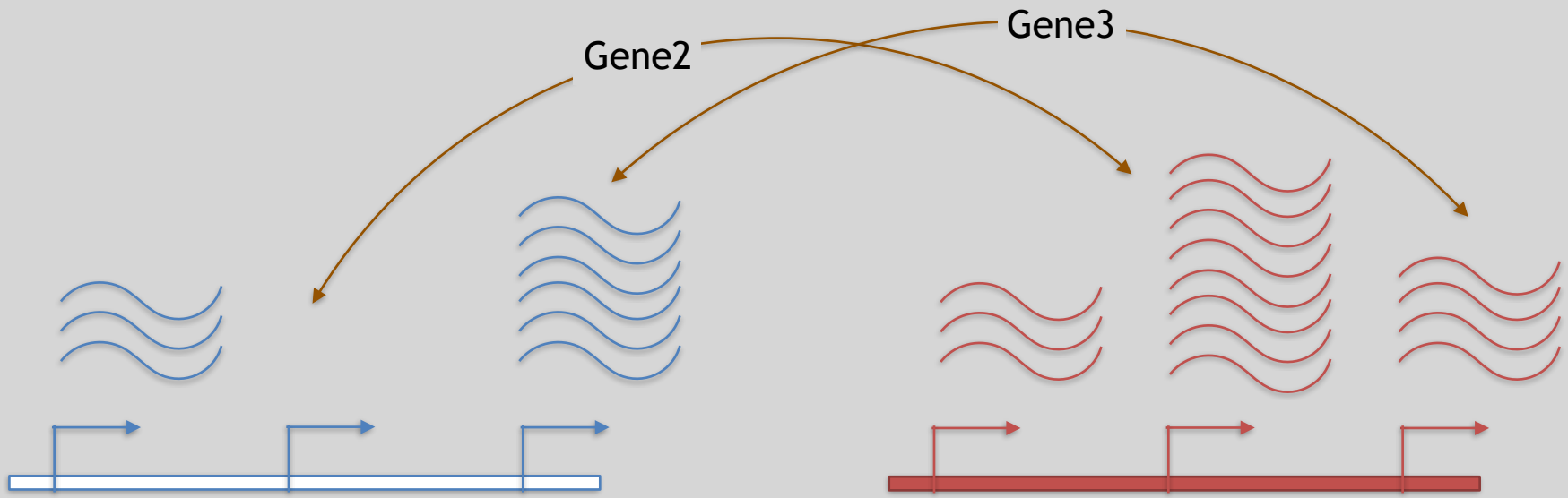
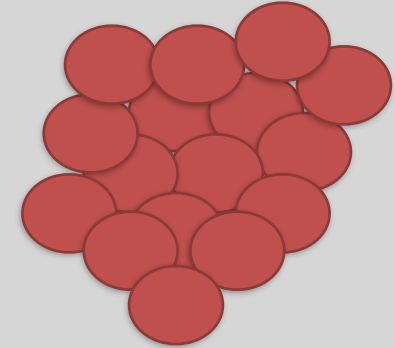
Then we can compare the two cell types to figure out what is different in the mutated cells!



Normal Cells



Mutated Cells



Differences apparent for Gene 2 and
to a lesser extent Gene 3

3 Main Steps for RNA-Seq:

1) Prepare a sequencing library

(RNA to cDNA conversion via reverse transcription)

2) Sequence

(Using the same technologies as DNA sequencing)

3) Data analysis

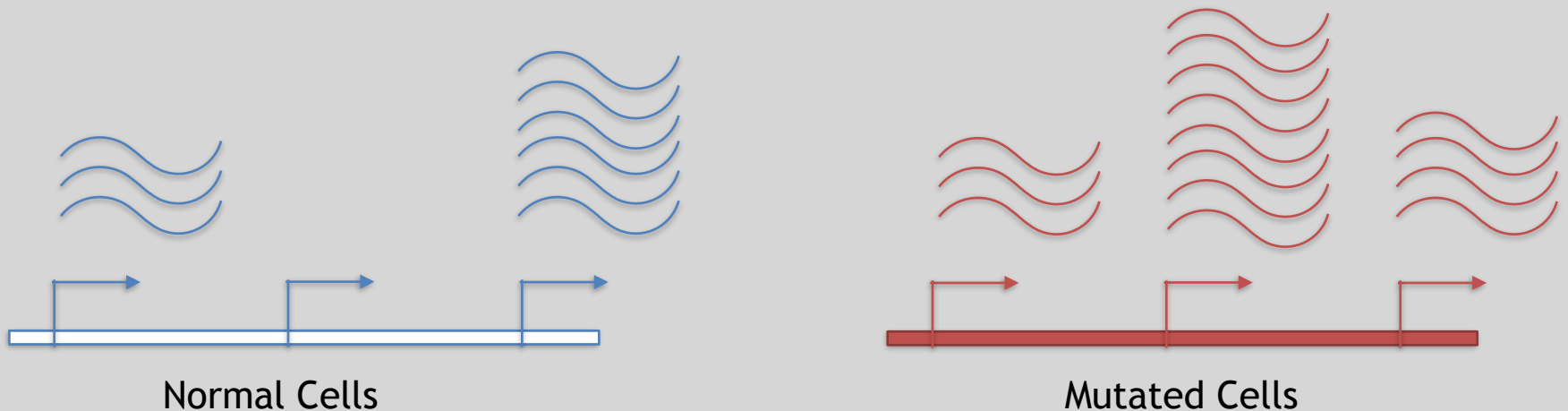
(Often the major bottleneck to overall success!)

We will discuss each of these steps in detail
(particularly the 3rd) next day!

Today we will get start of step 3!

Gene	WT-1	WT-2	WT-3	...
A1BG	30	5	13	...
AS1	24	10	18	...
...

We sequenced, aligned, counted the reads per gene in each sample to arrive at our data matrix



Do it Yourself!

Hand-on time!

Focus on **Sections 4** please
(After your Alignment is finished)

Feedback:

[Muddy Point Assessment]

Additional Reference Slides on SAM/BAM Format and Sequencing Methods

SAM Format

- Sequence Alignment/Map (SAM) format is the almost-universal sequence alignment format for NGS
 - binary version is BAM
- It consists of a header section (lines start with '@') and an alignment section
- The official specification can be found here:
 - <http://samtools.sourceforge.net/SAM1.pdf>

SAM header section

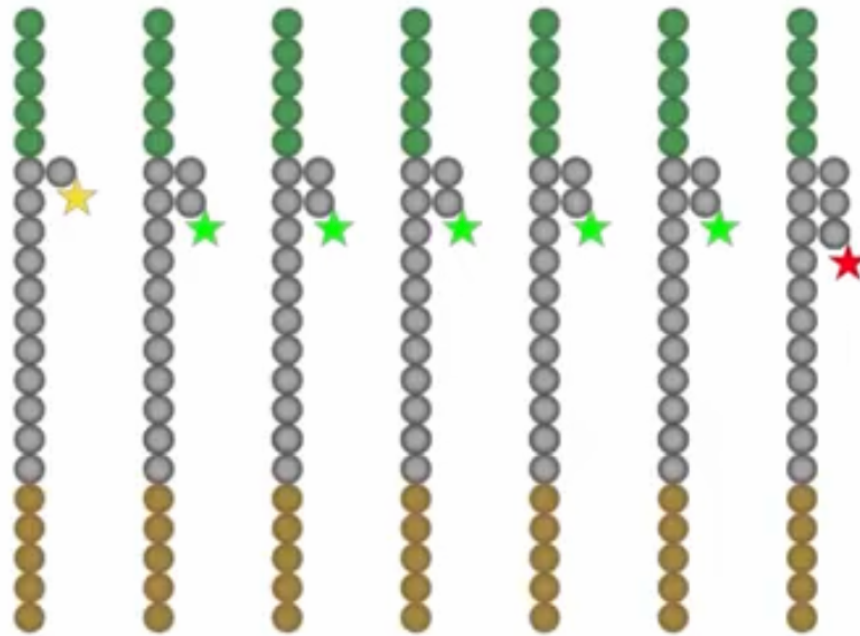
- Header lines contain vital metadata about the reference sequences, read and sample information, and (optionally) processing steps and comments.
- Each header line begins with an @, followed by a two-letter code that distinguishes the different type of metadata records in the header.
- Following this two-letter code are tab-delimited key-value pairs in the format **KEY:VALUE** (the SAM format specification names these tags and values).

https://bioboot.github.io/bimm143_F18/class-material/sam_format/

SAM Utilities

- **Samtools** is a common toolkit for analyzing and manipulating files in SAM/BAM format
 - <http://samtools.sourceforge.net/>
- **Picard** is a another set of utilities that can used to manipulate and modify SAM files
 - <http://picard.sourceforge.net/>
- These can be used for viewing, parsing, sorting, and filtering SAM files as well as adding new information (e.g. Read Groups)

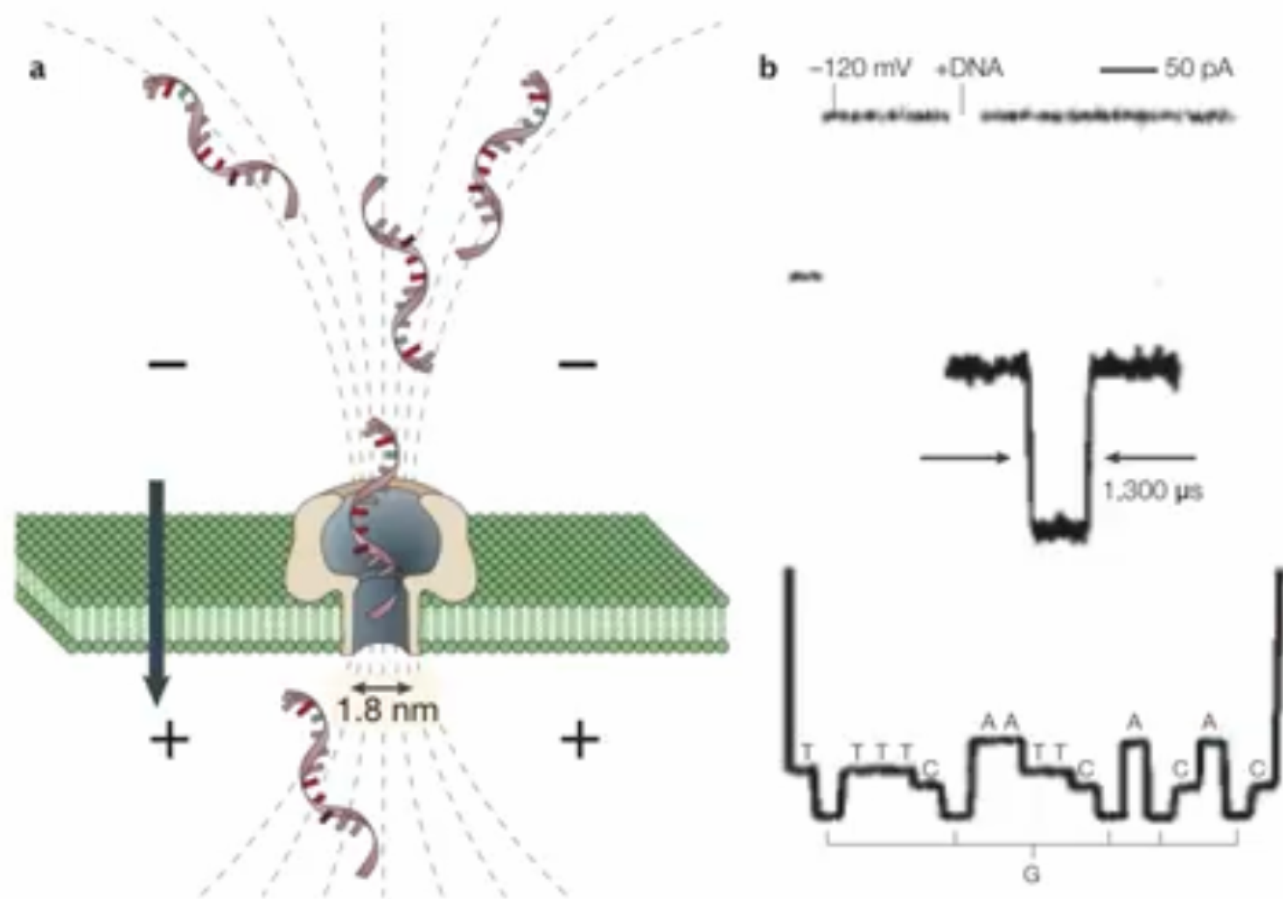
Length limits for Illumina Sequencing



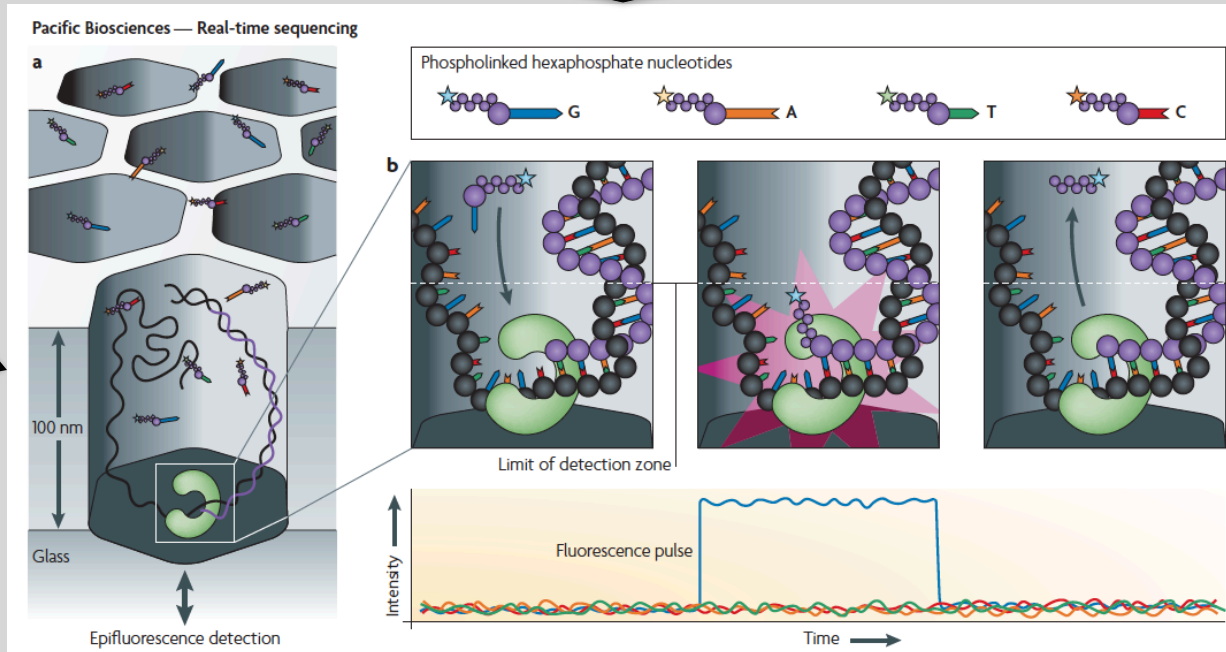
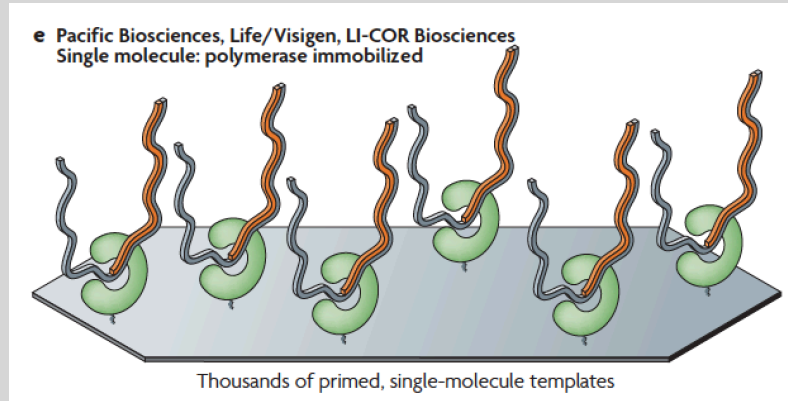
- Errors from chemistry add up.
- Limits reads to 300 bases

Additional Reference Slides on Sequencing Methods

Oxford Nanopore

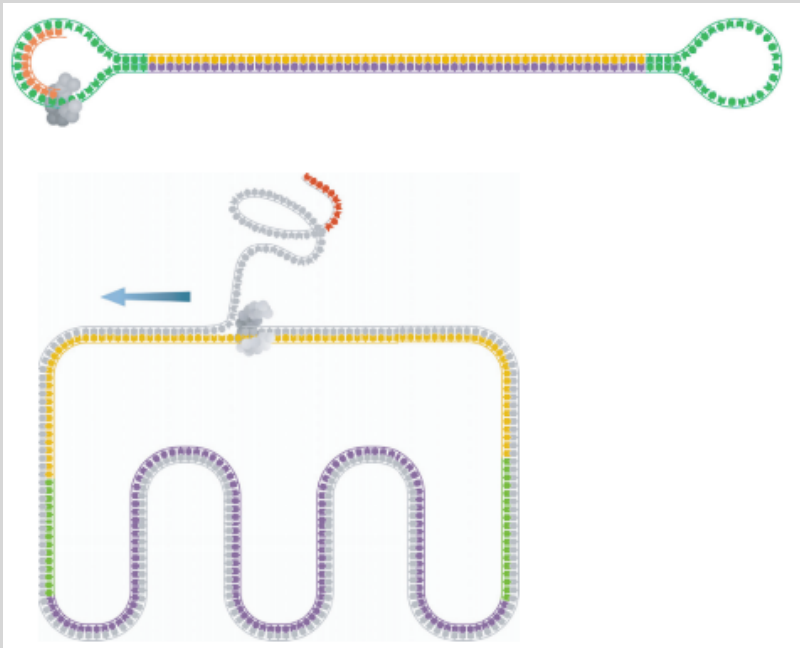


Pacific Biosystems - Real Time Sequencing

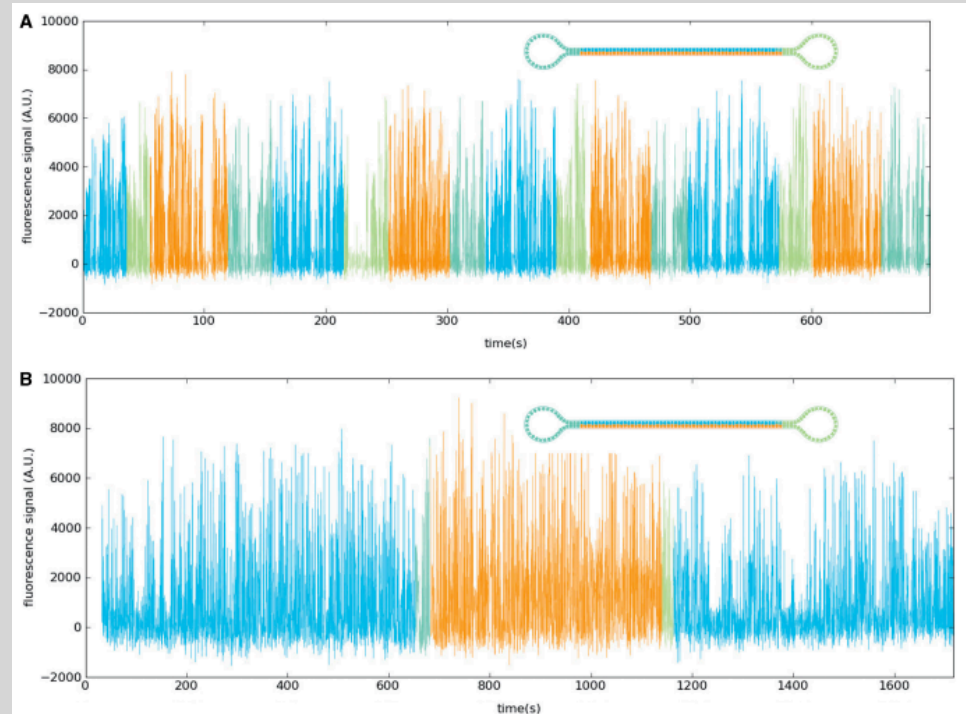


Pacific Biosystems - Circular Consensus

SMRTbell template



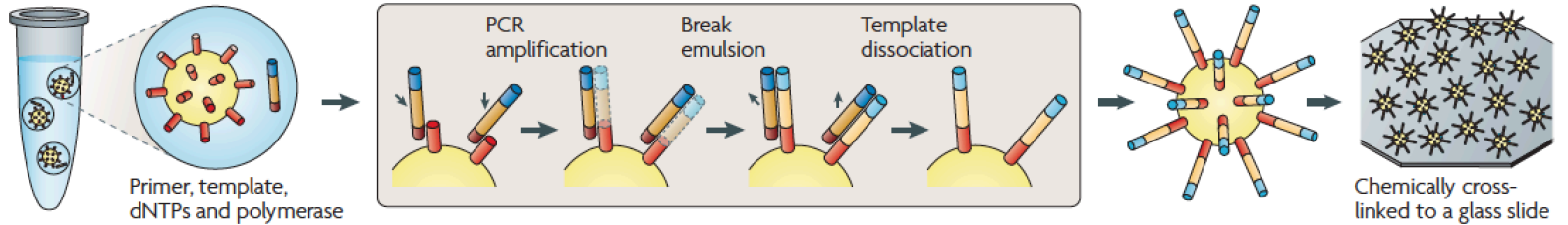
Subread Consensus Sequencing



Roche 454 - Pyrosequencing

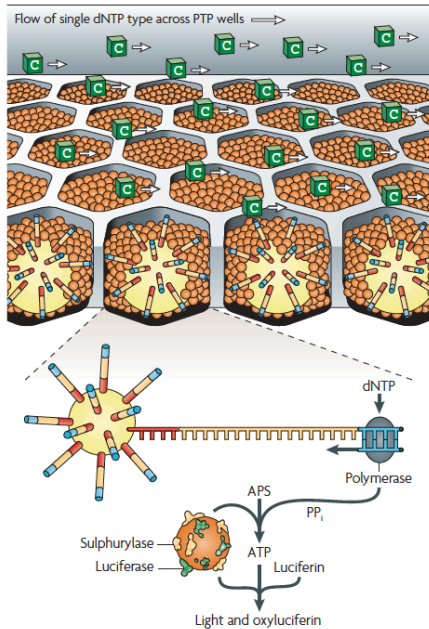
a Roche/454, Life/APG, Polonator Emulsion PCR

One DNA molecule per bead. Clonal amplification to thousands of copies occurs in microreactors in an emulsion



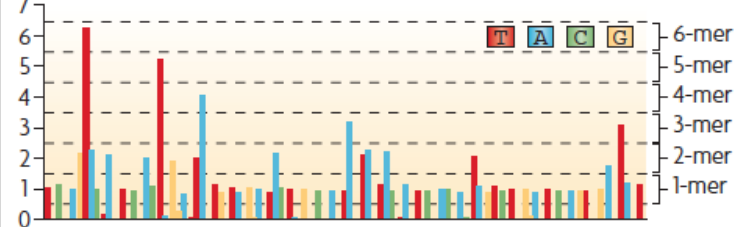
c Roche/454 — Pyrosequencing

1-2 million template beads loaded into PTP wells



d Flowgram

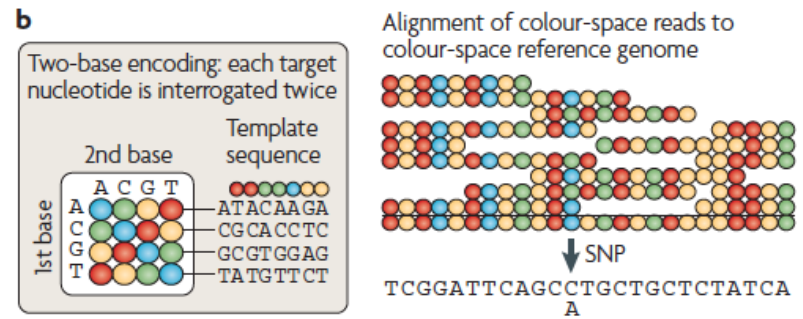
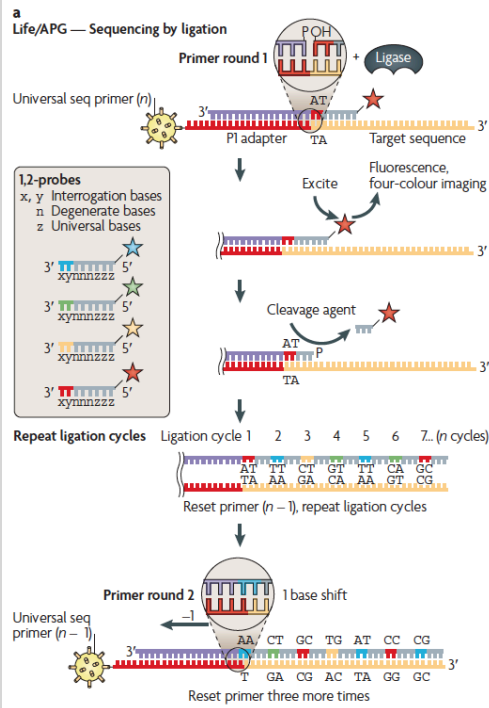
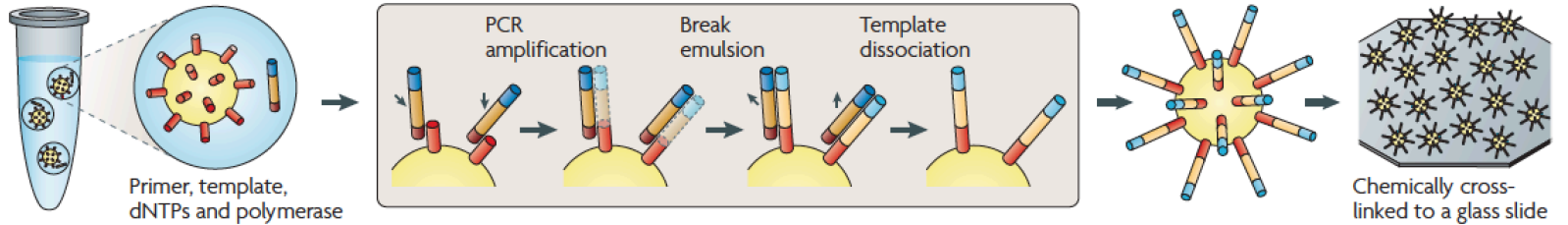
TCAGGTTTTTTTAAACAATCAACTTTTTGGATTAAAAATGTAGATAACTG
CATAAATTAATAACATCACATTAGTCTGATCAGTGAATTTAT



Life Technologies SOLiD - Sequence by Ligation

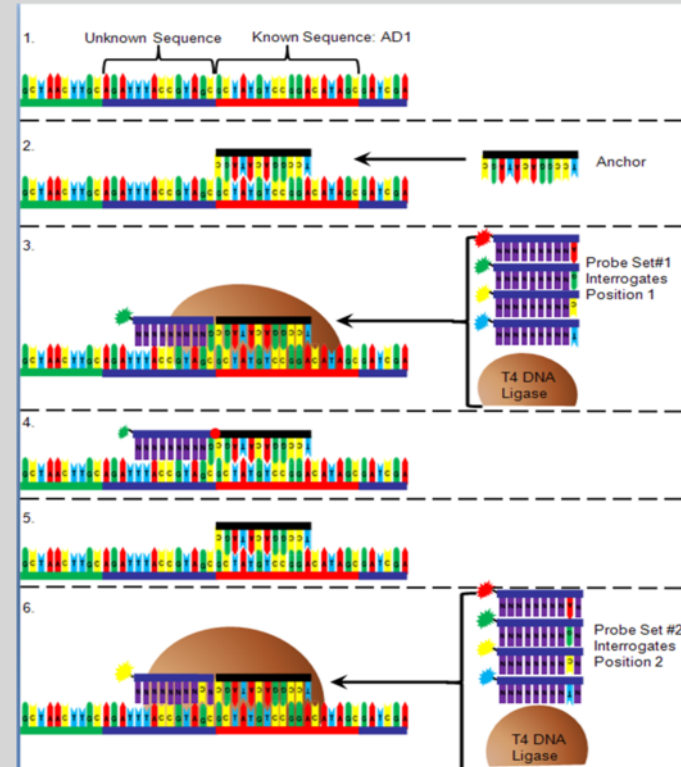
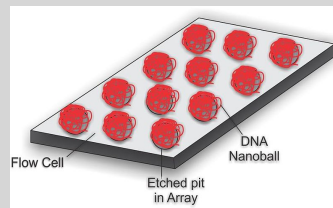
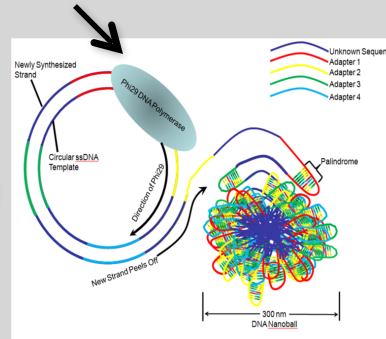
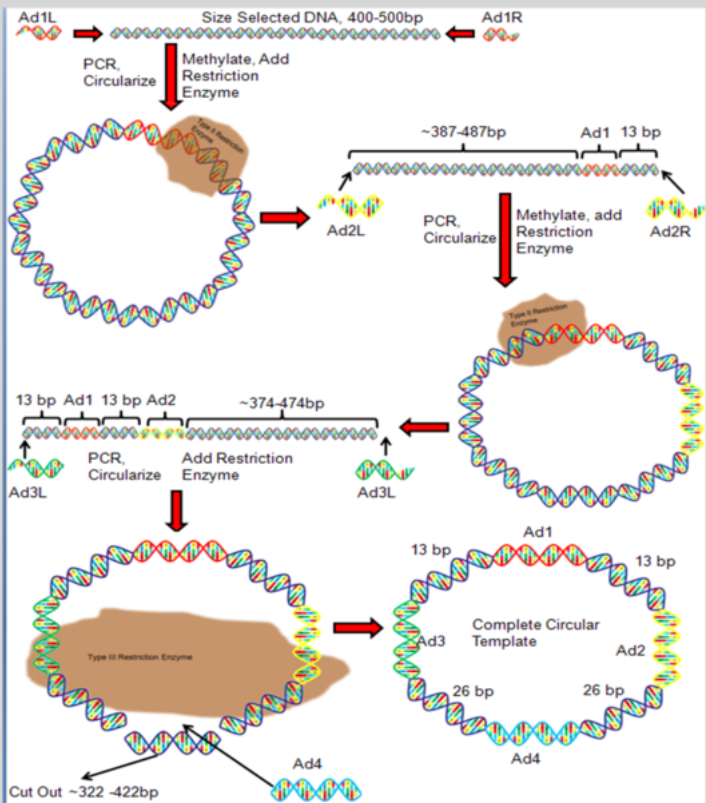
a Roche/454, Life/APG, Polonator Emulsion PCR

One DNA molecule per bead. Clonal amplification to thousands of copies occurs in microreactors in an emulsion



Complete Genomics - Nanoball Sequencing

Has proofreading ability!

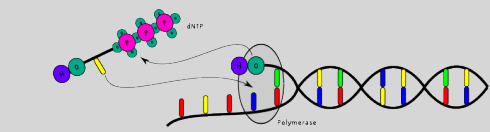


“Benchtop” Sequencers

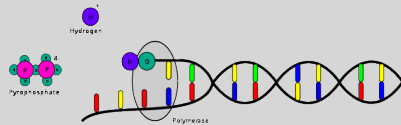
- Lower cost, lower throughput alternative for smaller scale projects
- Currently three significant platforms
 - Roche 454 GS Junior
 - Life Technology Ion Torrent
 - Personal Genome Machine (PGM)
 - Proton
 - Illumina MiSeq

Platform	List price	Approximate cost per run	Minimum throughput (read length)	Run time	Cost/Mb	Mb/h
454 GS Junior	\$108,000	\$1,100	35 Mb (400 bases)	8 h	\$31	4.4
Ion Torrent PGM						
(314 chip)	\$80,490 ^{a,b}	\$225 ^c	10 Mb (100 bases)	3 h	\$22.5	3.3
(316 chip)		\$425	100 Mb ^d (100 bases)	3 h	\$4.25	33.3
(318 chip)		\$625	1,000 Mb (100 bases)	3 h	\$0.63	333.3
MiSeq	\$125,000	\$750	1,500 Mb (2 × 150 bases)	27 h	\$0.5	55.5

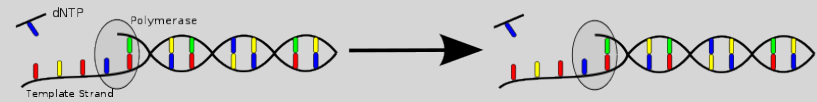
PGM - Ion Semiconductor Sequencing



Polymerase integrates a nucleotide.



Hydrogen and pyrophosphate are released.



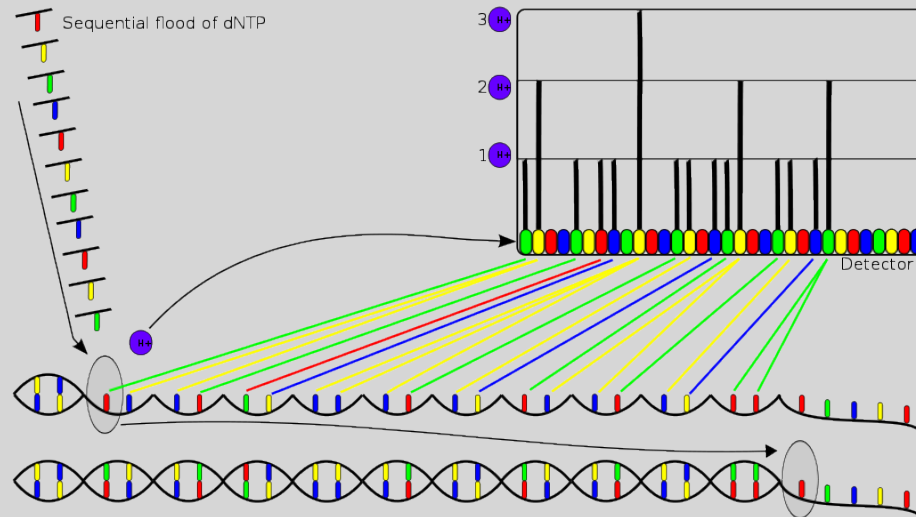
The nucleotide does not complement the template - no release of hydrogen.



The nucleotide complements the template - hydrogen is released.



The nucleotide complements several bases in a row - multiple hydrogen ions are released.



Normalization

- Normalization is required to make comparisons in gene expression - Between 2+ genes in one sample - Between genes in 2+ samples
- Genes will have more reads mapped in sample with high coverage than with low read coverage - $2x$ depth \approx $2x$ expression
- Longer genes will have more reads mapped than shorter genes - $2x$ length \approx $2x$ more reads

Normalization: RPKM, FPKM and TPM

- **N.B.** Some tools for differential expression analysis such as edgeR and DESeq2 want raw read counts - i.e. non normalized input!
- However, often for your manuscripts and reports you will want to report normalized counts - e.g. plots of Log(FoldChange) vs Transcripts Per Million (or TPM)
- RPKM, FPKM and TPM all aim to normalize for sequencing depth and gene length.
- RPKM was made for single-end RNA-seq and stands for Reads per :
 - Count up the total reads in a sample and divide that number by 1,000,000 - this is our “per million” scaling factor.
 - Divide the read counts by the “per million” scaling factor. This normalizes for sequencing depth, giving you reads per million (RPM)
 - Divide the RPM values by the length of the gene, in kilobases. This gives you RPKM.

- FPKM was made for paired-end RNA-seq
- With paired-end RNA-seq, two reads can correspond to a single fragment
- The only difference between RPKM and FPKM is that FPKM takes into account that two reads can map to one fragment (and so it doesn't count this fragment twice).

- TPM is very similar to RPKM and FPKM. The only difference is the order of operations. Here's how you calculate TPM:
 - Divide the read counts by the length of each gene in kilobases. This gives you reads per kilobase (RPK).
 - Count up all the RPK values in a sample and divide this number by 1,000,000. This is your “per million” scaling factor.
 - Divide the RPK values by the “per million” scaling factor. This gives you TPM.
- So you see, when calculating TPM, the only difference is that you normalize for gene length first, and then normalize for sequencing depth second. However, the effects of this difference are quite profound.

- When you use TPM, the sum of all TPMs in each sample are the same. This makes it easier to compare the proportion of reads that mapped to a gene in each sample. In contrast, with RPKM and FPKM, the sum of the normalized reads in each sample may be different, and this makes it harder to compare samples directly.