



BIMM 143

Genome Informatics I

Lecture 13

Barry Grant
UC San Diego

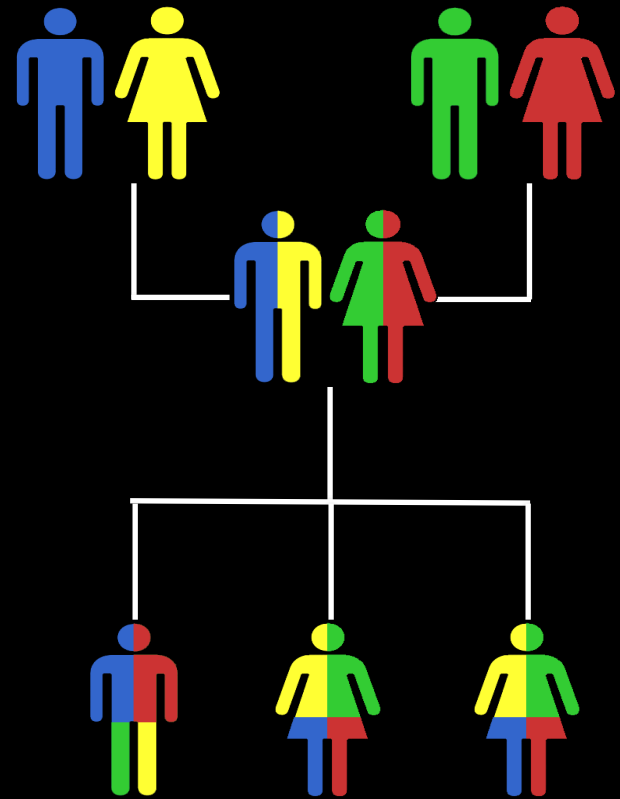
<http://thegrantlab.org/bimm143>

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



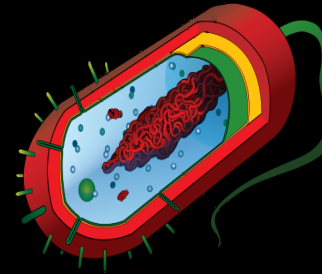
Genetics and Genomics

- **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.

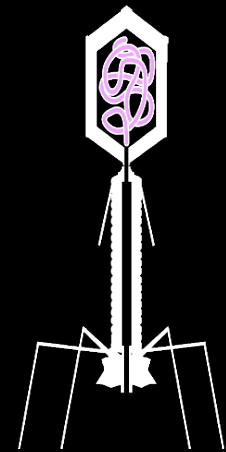
Genomes come in many shapes

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.

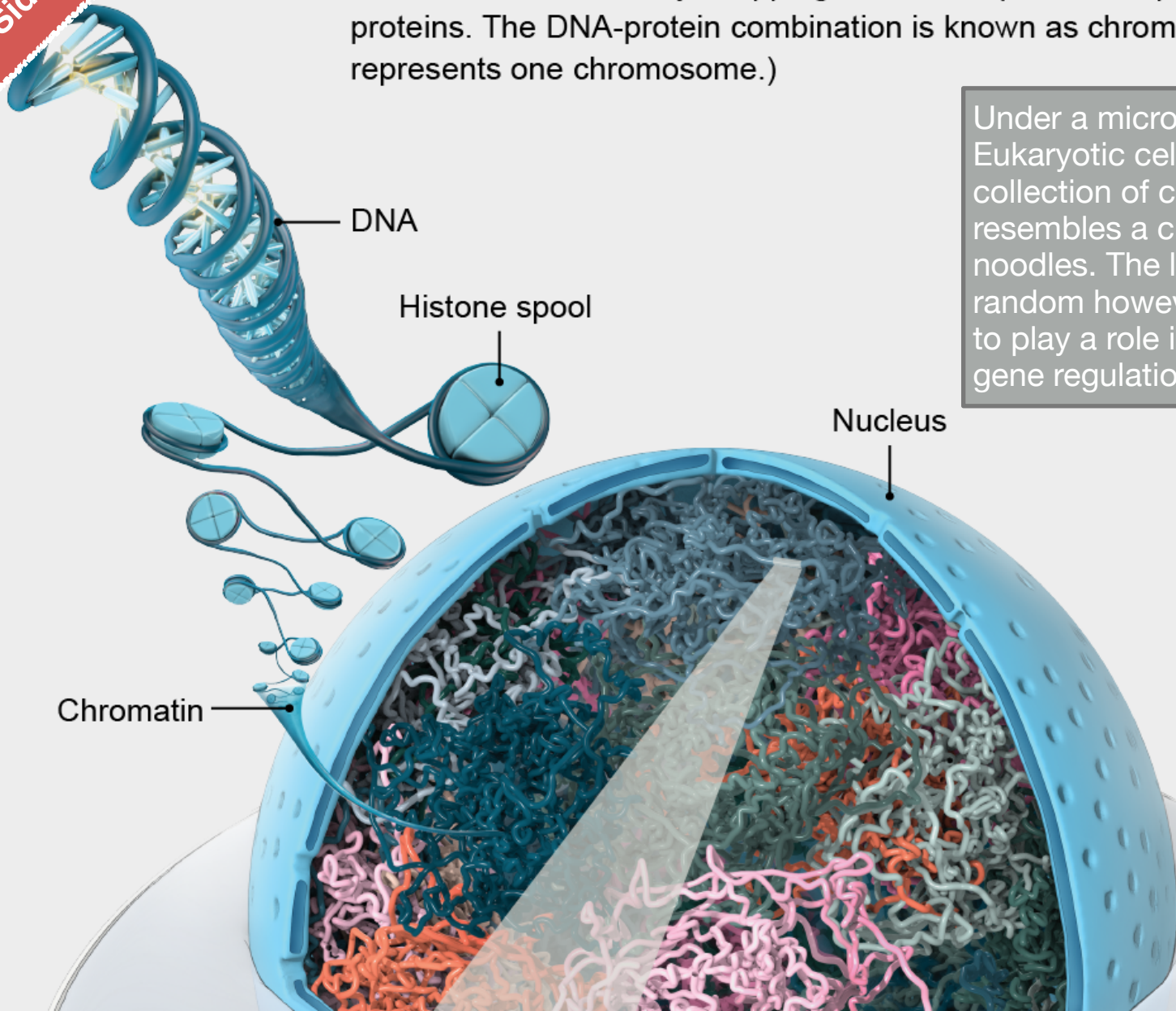
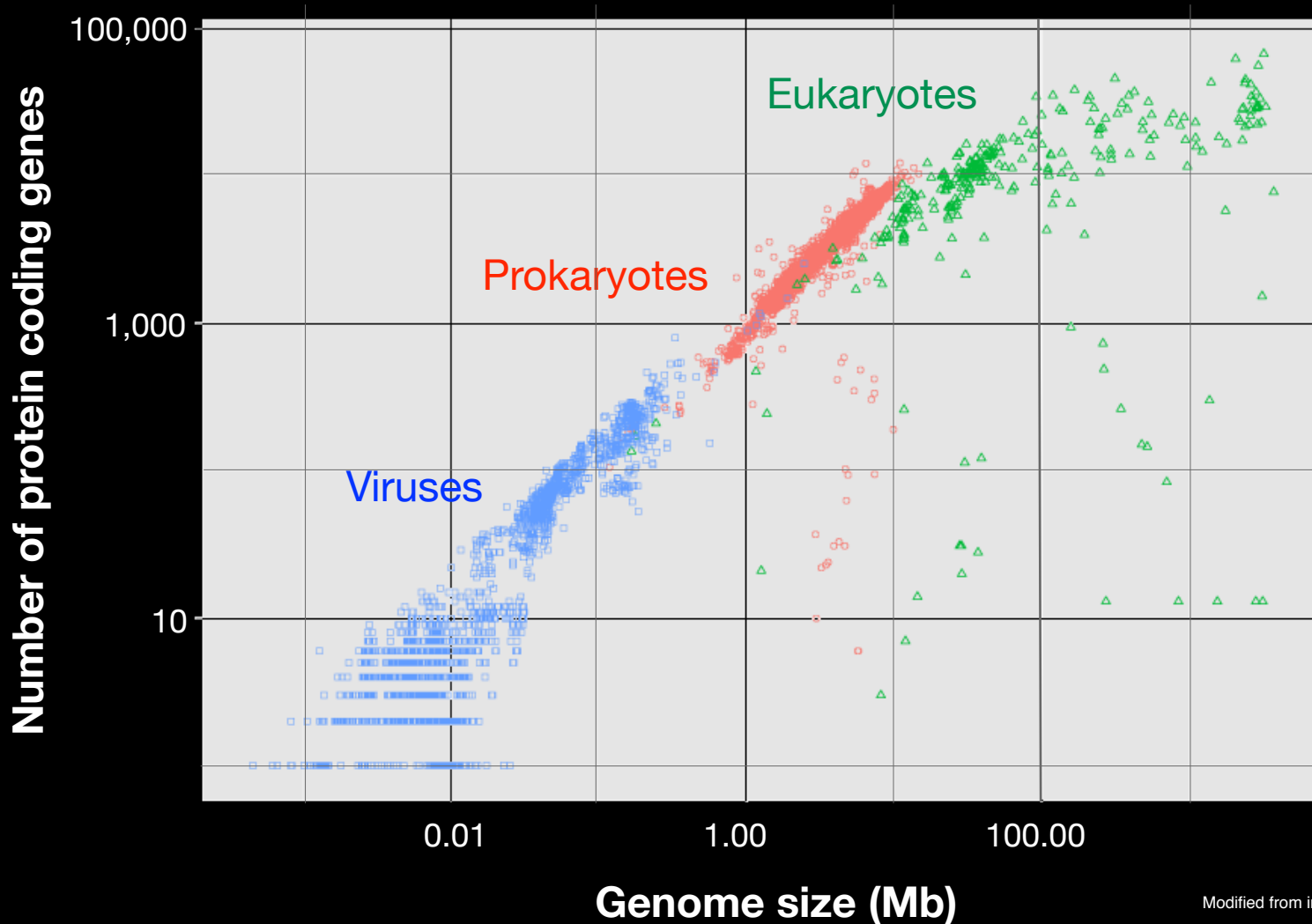


Image credit:
[Scientific American](#)
[March 2019](#)

Genomes come in many sizes



Genome Databases

NCBI Genome:

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

NCBI Resources How To Sign in to NCBI

Genome Genome Search Limits Advanced Help

Genome

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

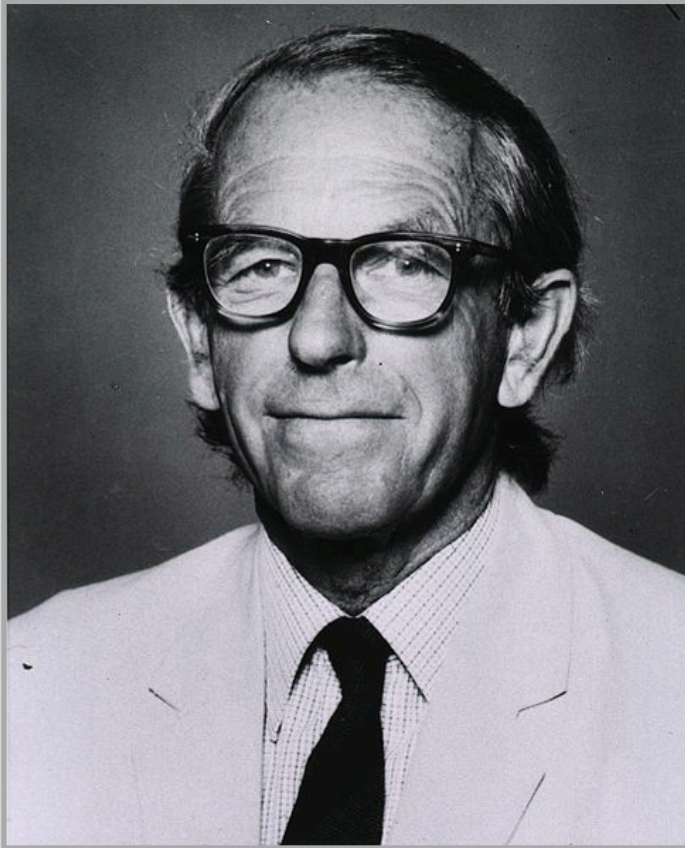
Using Genome	Custom resources	Other Resources
Help	Human Genome	Assembly
Browse by Organism	Microbes	BioProject
Download / FTP	Organelles	BioSample
Download FAQ	Viruses	Map Viewer
Submit a genome	Prokaryotic reference genomes	Protein Clusters

Genome Tools	Genome Annotation and Analysis	External Resources
BLAST the Human Genome	Eukaryotic Genome Annotation	GOLD - Genomes Online Database
Microbial Nucleotide BLAST	Prokaryotic Genome Annotation	Ensembl Genome Browser
TaxPlot (3-way Genome Comparison)	PASC (Pairwise Sequence Comparison)	Bacteria Genomes at Sanger
		Large-Scale Genome Sequencing (NHGRI)

You are here: NCBI > Genomes & Maps > Genome Write to the Help Desk

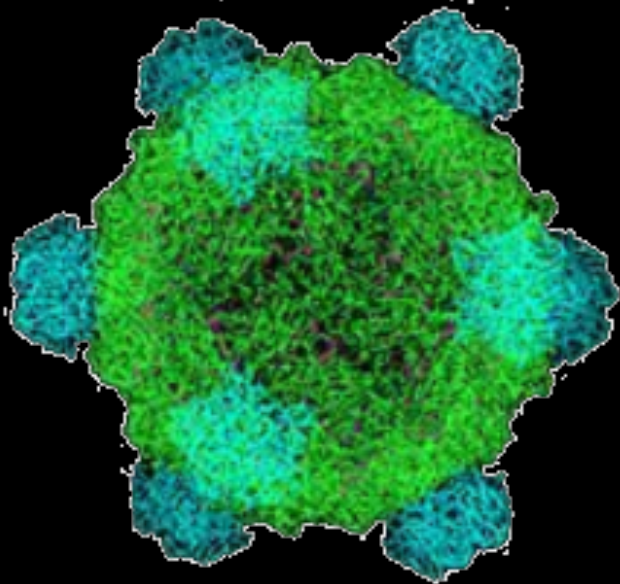
GETTING STARTED	RESOURCES	POPULAR	FEATURED	NCBI INFORMATION
NCBI Education	Chemicals & Bioassays	PubMed	Genetic Testing Registry	About NCBI
NCBI Help Manual	Data & Software	Bookshelf	PubMed Health	Research at NCBI
NCBI Handbook	DNA & RNA	PubMed Central	GenBank	NCBI News
Training & Tutorials	Domains & Structures	PubMed Health	Reference Sequences	NCBI FTP Site
	Genes & Expression	BLAST	Gene Expression Omnibus	NCBI on Facebook
	Genetics & Medicine	Nucleotide	Map Viewer	NCBI on Twitter
	Genomes & Maps	Genome	Human Genome	NCBI on YouTube
	Homology	SNP	Mouse Genome	
	Literature	Gene	Influenza Virus	
	Proteins	Protein	Primer-BLAST	
	Sequence Analysis	PubChem	Sequence Read Archive	
	Taxonomy			
	Training & Tutorials			
	Variation			

Early Genome Sequencing



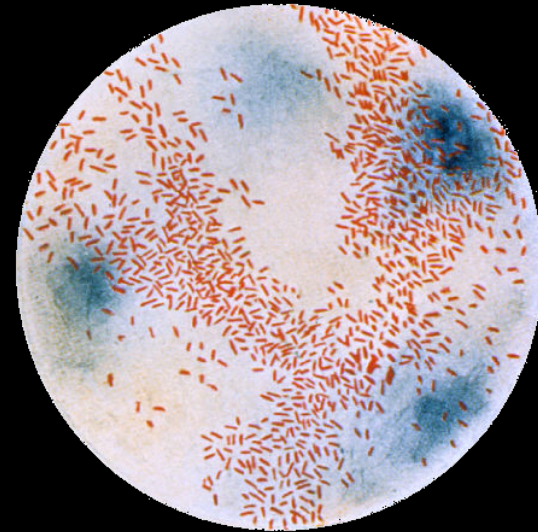
- Chain-termination “**Sanger**” sequencing was developed in 1977 by *Frederick Sanger*, colloquially referred to as the “Father of Genomics”
- Sequence reads were typically 750-1000 base pairs in length with an error rate of $\sim 1 / 10000$ bases

The First Sequenced Genomes



Bacteriophage ϕ -X174

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

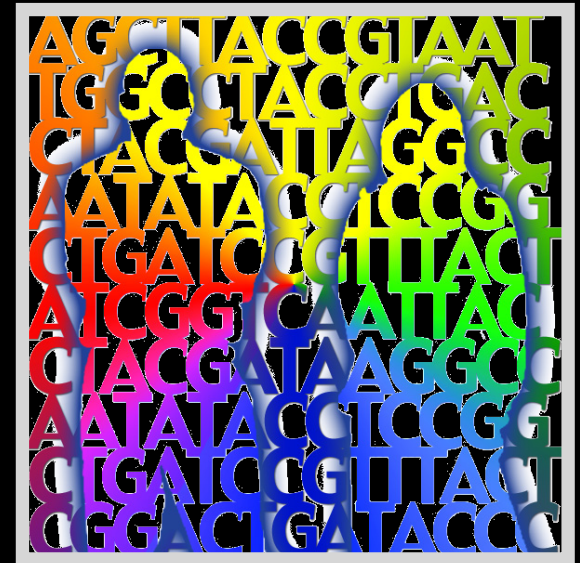


Haemophilus influenzae

- Completed in 1995
- 1,830,140 base pairs, dsDNA
- 1740 genes

The Human Genome Project

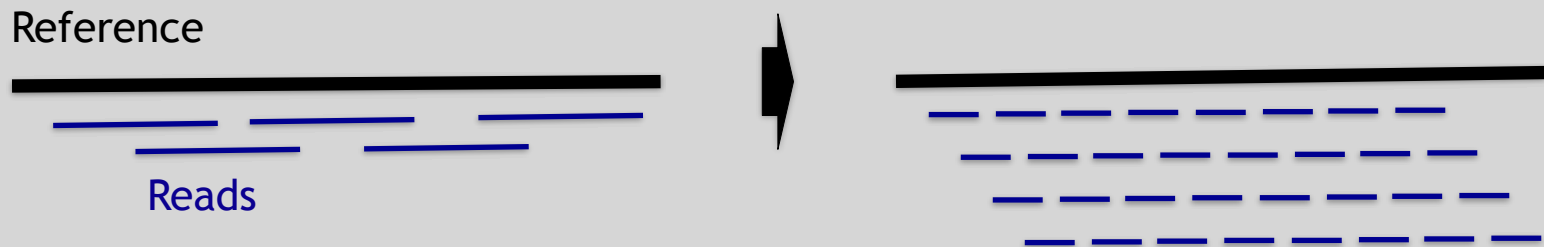
- The Human Genome Project (HGP) was an international, public consortium that began in 1990
 - Initiated by James Watson
 - Primarily led by Francis Collins
 - Eventual Cost: \$2.7 Billion
- Celera Genomics was a private corporation that started in 1998
 - Headed by Craig Venter
 - Eventual Cost: \$300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



HHMI

Modern Genome Sequencing

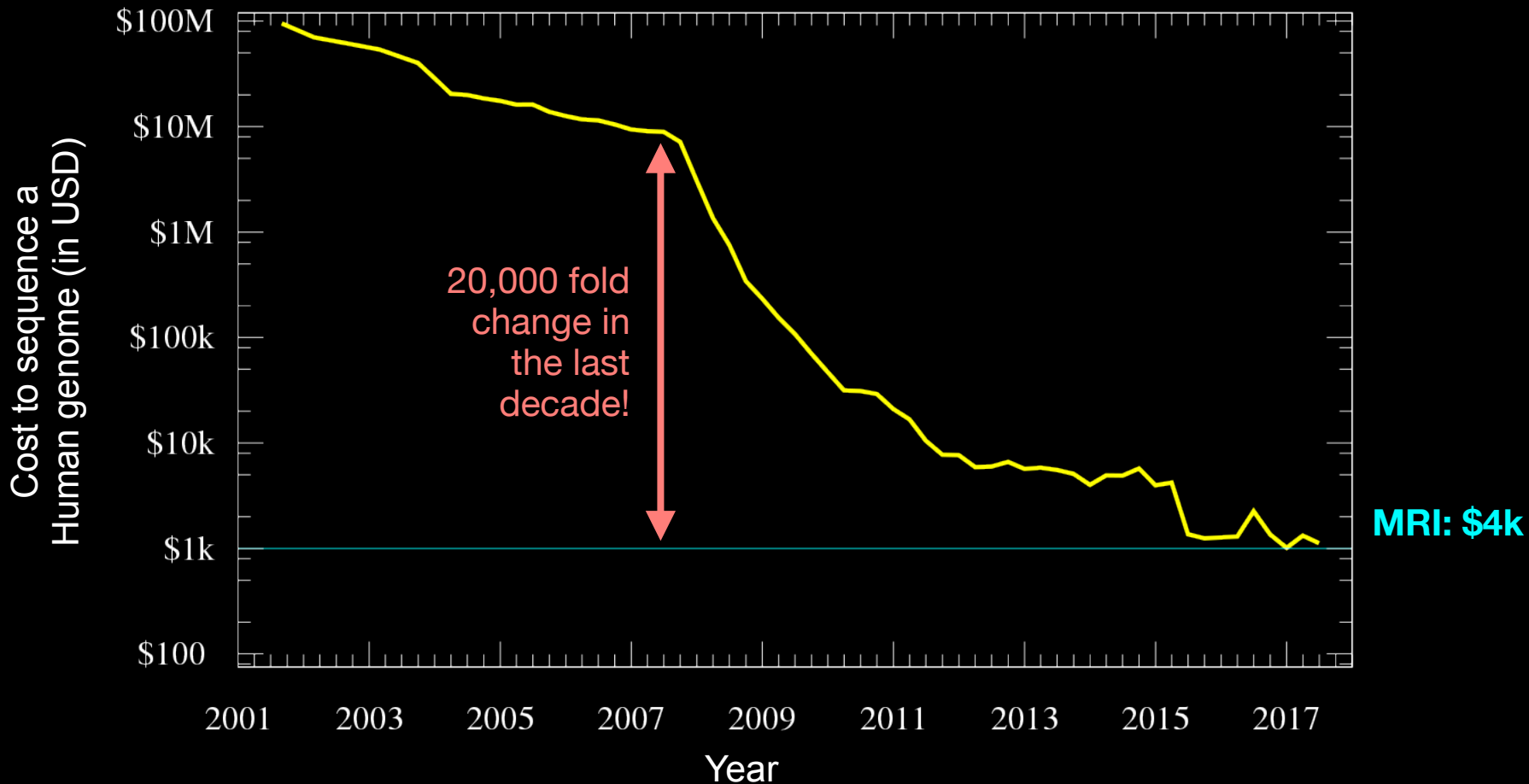
- Next Generation Sequencing (NGS) technologies have resulted in a paradigm shift from long reads at low coverage to short reads at high coverage
- This provides numerous opportunities for new and expanded genomic applications



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.

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)

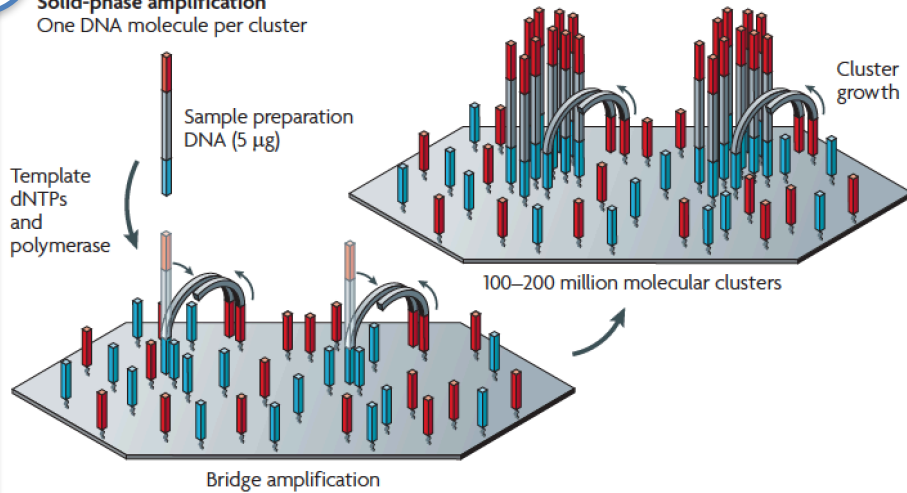
Modern NGS Sequencing Platforms

	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 - 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

2

Illumina/Solexa — Reversible terminators

Incorporate all four nucleotides, each label with a different dye

Wash, four-colour imaging

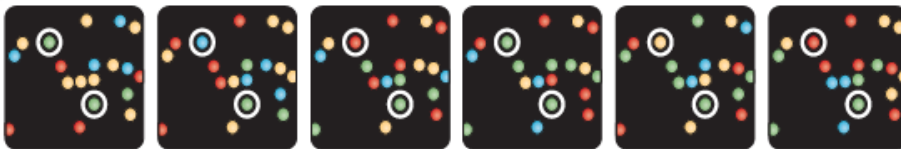
Cleave dye and terminating groups, wash

Repeat cycles

Cleave dye & blocking group, repeat...

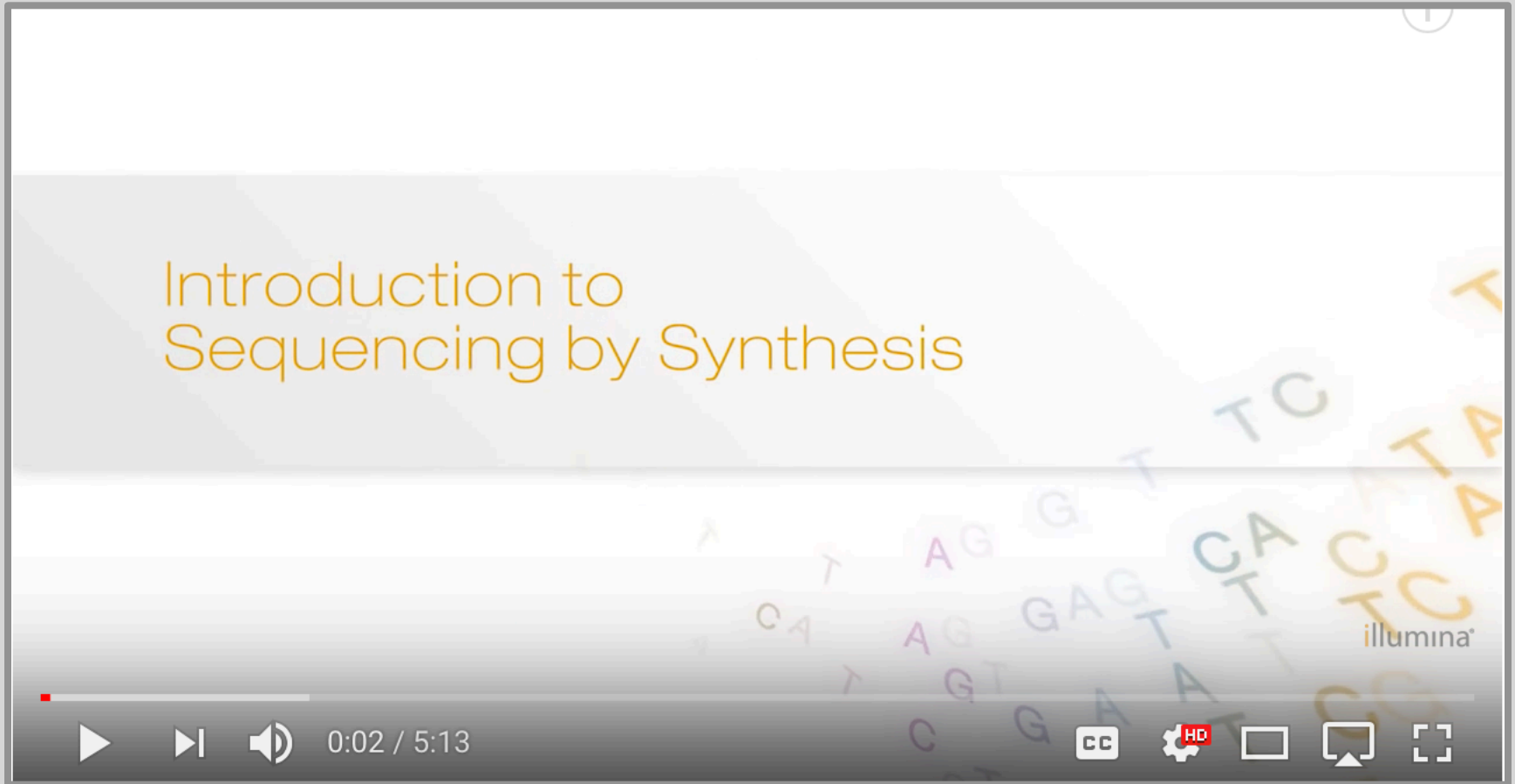
3 Fluorescent emission from incorporated dye-labeled nucleotides

3



Top: CATCGT
Bottom: CCCCCC

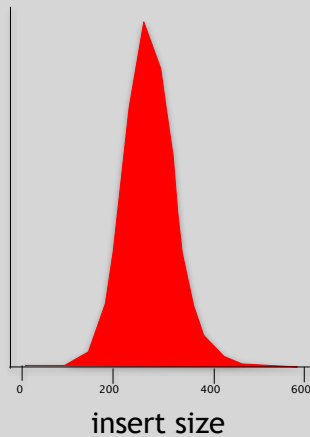
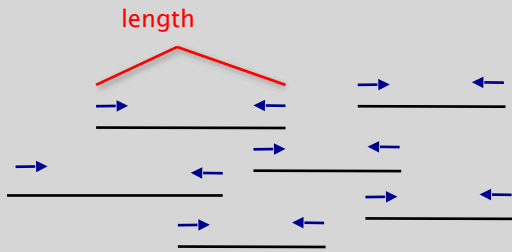
Illumina Sequencing - Video



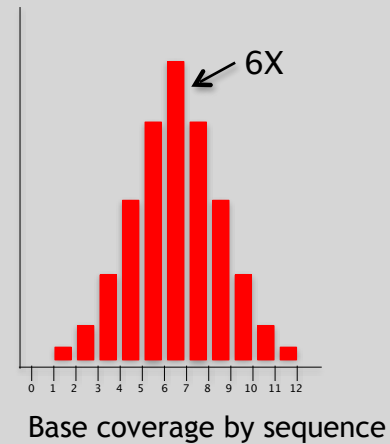
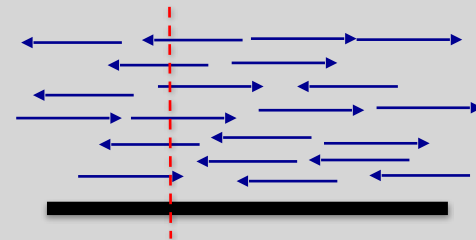
https://www.youtube.com/watch?src_vid=womKfikWlxM&v=fCd6B5HRaZ8

NGS Sequencing Terminology

Insert Size



Sequence Coverage



Summary: “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 (1,000bp+) sequence reads
 - Single molecule (no amplification step)
 - Often associated with nanopore technology
 - But not necessarily!

The first direct RNA sequencing by nanopore

Side-Note:

- For example this new nanopore sequencing method was just published!

<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."

SeqAnswers Wiki

Side-Note:

A good repository of analysis software can be found at <http://seqanswers.com/wiki/Software/list>

Page [Discussion](#) Read [View source](#) [View history](#)

Software/list

[< Software](#)

Below is (one of many possible) dynamic tables of software data, created from pages in the wiki. To add a package to the list, use the following form:

CSV
JSON

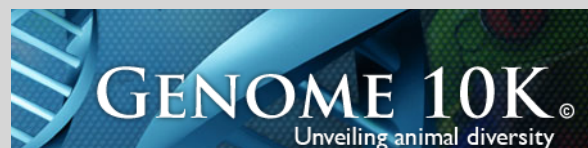
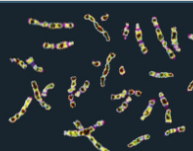
Name	Summary	Bio Tags	Meth Tags	Features	Language	Licence	OS
4peaks	Allows viewing sequencing trace files, motif searching trimming, BLAST and exporting sequences.	Sequencing	Sequence analysis			Freeware	Mac OS X
AB Large Indel Tool	Identifies deviations in clone insert size that indicate intra-chromosomal structural variations compared to a reference genome.	InDel discovery Sequencing	Mapping		Perl	GPL	Linux 64
AB Small Indel Tool	The SOLID™ Small Indel Tool processes the indel evidences found in the pairing step of the SOLID™ System Analysis Pipeline Tool (Corona Lite).	InDel discovery Sequencing	Mapping Alignment		Perl C++	GPL	Linux 64
ABBA	Assembly Boosted By Amino acid sequence is a comparative gene assembler, which uses amino acid sequences from predicted proteins to help build a better assembly	Genomic Assembly	Assembly Scaffolding			Artistic License	Linux
ABMapper	Maps RNA-Seq reads to target genome considering possible multiple mapping locations and splice junctions	Genomics Transcriptomics	Mapping Alignment		C++ Perl	GPLv3	Linux
ABYSS	ABYSS is a de novo sequence assembler designed for short reads and large genomes.	De-novo assembly	Assembly De Bruijn graph	MPI OpenMP	C++	Free for academic use	POSIX Linux Mac OS X
Adaptor Removal	Removes adaptor fragments from raw short read	General	Adaptor Removal	Trimming	Java	Custom Licence	Linux 64

**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

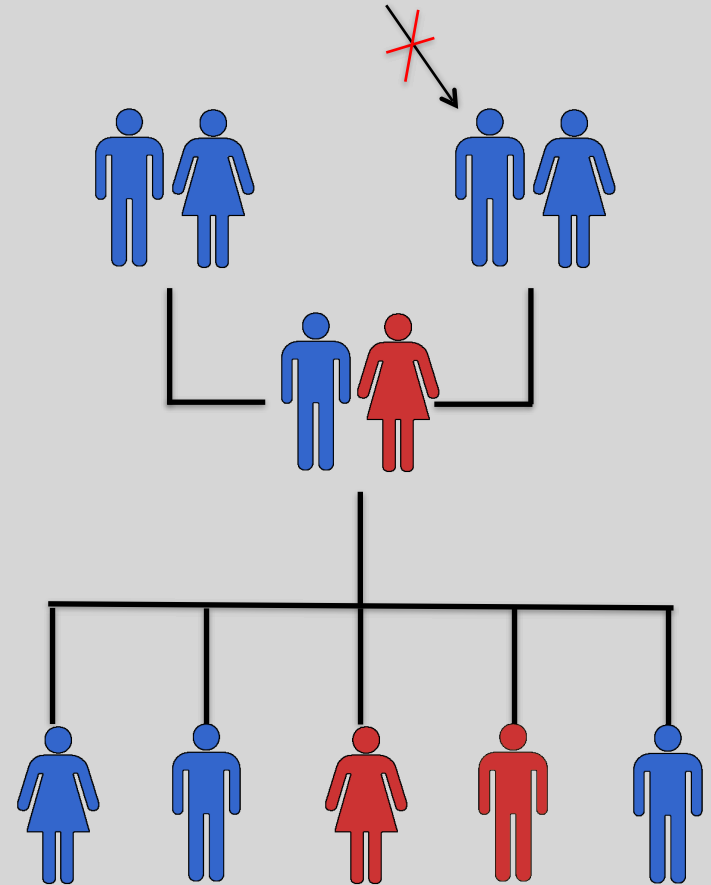
“Variation is the spice of life”

-Kruglyak & Nickerson, 2001

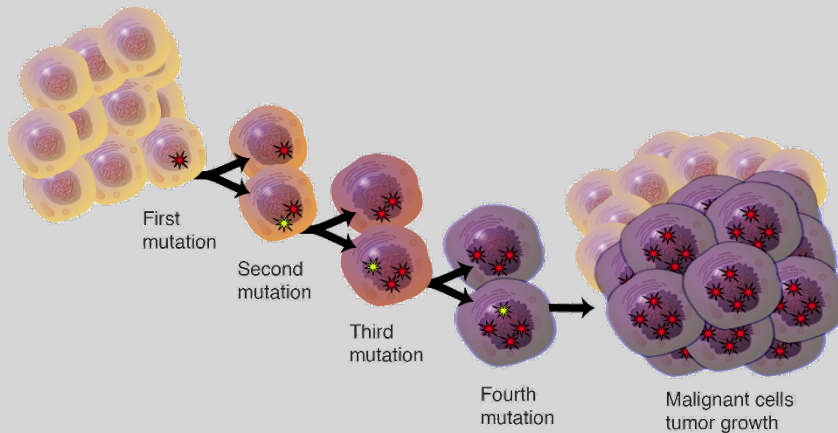
- While the sequencing of the human genome was a great milestone, the DNA from a single person is not representative of the millions of potential differences that can occur between individuals
- These unknown genetic variants could be the cause of many phenotypes such as differing morphology, susceptibility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

Germline Variation

- Mutations in the germline are passed along to offspring and are present in the DNA over every cell
- In animals, these typically occur in meiosis during gamete differentiation



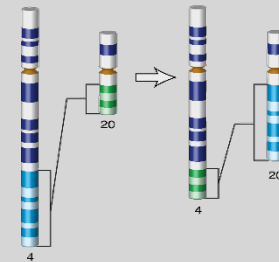
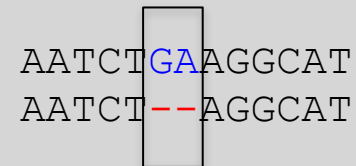
Somatic Variation



- Mutations in non-germline cells that are not passed along to offspring
- Can occur during mitosis or from the environment itself
- Are an integral part in tumor progression and evolution

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



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!

Discovering Variation: SNPs and INDELS

SNP

sequencing error or genetic variant?

```
ATCCTGATTTCGGTGAACGTTATCGACGATCCGATCGA
ATCCTGATTTCGGTGAACGTTATCGACGATCCGATCGA
  CCGTGAACGTTATCGACGATCCGATCGAACTGTCAGC
  GGTGAACGTTATCGACGTTCCGATCGAACTGTCAGCG
  TGAACGTTATCGACGTTCCGATCGAACTGTCATCGGC
  TGAACGTTATCGACGTTCCGATCGAACTGTCAGCGGC
  TGAACGTTATCGACGTTCCGATCGAACTGTCAGCGGC
    GTTATCGACGATCCGATCGAACTGTCAGCGGCAAGCT
    TTATCGACGATCCGATCGAACTGTCAGCGGCAAGCT
```

ATCCTGATTTCGGTGAACGTTATCGACGATCCGATCGAACTGTCAGCGGCAAGCTGATCCGATCGATCGATGCTAGTG

reference genome

```
TTATCGACGATCCGATCGAACTGTCAGCGGCAAGCT
  TCGACGATCCGATCGAACTGTCAGCGGCAAGCTGAT
    ATCCGATCGAACTGTCAGCGGCAAGCTGATCGATCGAT
      TCCGATCGAACTGTCAGCGGCAAGCTGATCGATCGATCG
        TCCGATCGAACTGTCAGCGGCAAGCTGATCGATCGA
          GATCGAACTGTCAGCGGCAAGCTGATCGATCGATCGA
            AACTGTCAGCGGCAAGCTGATCGATCGATCGATGCTA
              TGTCAGCGGCAAGCTGATCGATCGATCGATGCTAG
                TCAGCGGCAAGCTGATCGATCGATCGATGCTAGTG
```

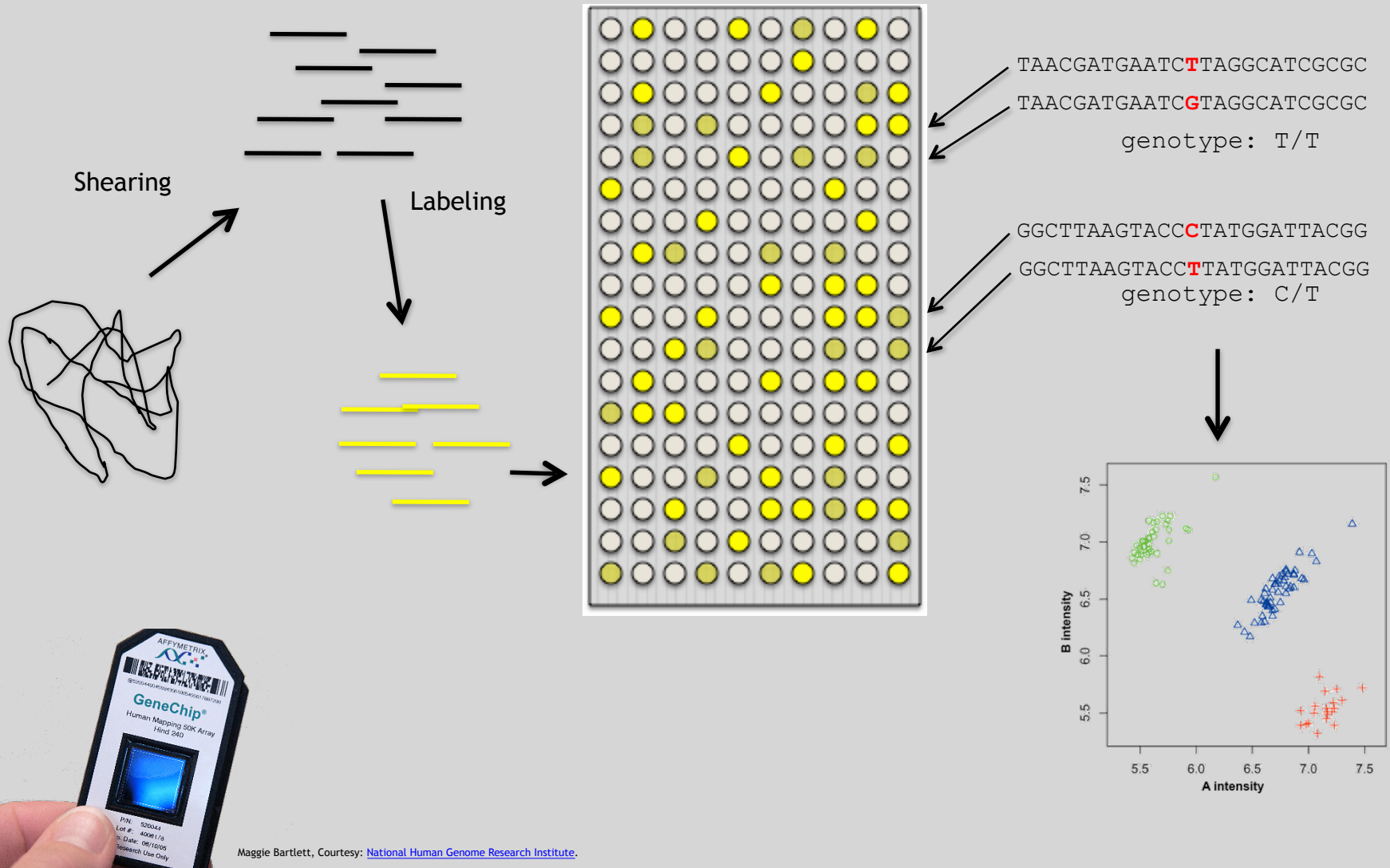
sequencing error or genetic variant?

INDEL

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



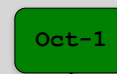
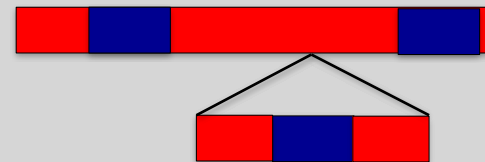
TAC → TAA

Frameshift mutation



TAC → T-C

Gene or exon deletion



Transcription factor binding disruption



ATGCAAAT → ATGCAGAT

Do it Yourself!

Hand-on time!

https://bioboot.github.io/bimm143_S19/lectures/#14

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 (25,200,000 to 25,250,936) and haplotypes and patches.

Overview image: Shows a detailed view of the region, including genes (e.g., RPL11-625L16.3, CASQ1, RPL11-713N11.4, RPL11-295S.3, RPL11-318G8.4) and a gene legend (merged Ensembl/Havana, processed transcript, RNA gene, pseudogene).

Zoomable Region image: Shows a highly detailed view of the region, including transcripts (splice variants) for genes like KRAS-001, KRAS-003, KRAS-002, and KRAS-004, and CDS set (e.g., CCDS1703.1).

Callouts:

- Region of interest
- Haplotypes and patches
- Change or add data tracks
- Change or add data tracks
- Change or add data tracks
- Genes
- Gene or region of interest
- Genome
- Transcripts (splice variants)

Navigation and Configuration:

- Location: 12:25,204,789-25,250,936
- Gene: [input field]
- Go buttons
- Drag/Select: [input field]
- Configure this page
- Bookmark this page

Access a jetstream galaxy instance!

Use assigned IP address

Do it Yourself!

The screenshot displays the Galaxy web interface for the Bowtie2 tool. The browser address bar shows the IP address 149.165.169.186. The tool configuration page is titled "Bowtie2 - map reads against reference genome (Galaxy Version 2.2.6.2)".

Is this single or paired library
Single-end

FASTQ file
4: HG00109_2.fastq

Write unaligned reads (in fastq format) to separate file(s)
Yes No

Write aligned reads (in fastq format) to separate file(s)
Yes No

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

Select reference genome
Baboon (Papio anubis): papHam1

Set read groups information?
Do not set

Select analysis mode
1: Default setting only

Do you want to use presets?
 No, just use defaults

History
search datasets
Unnamed history
22 shown, 2 deleted, 1 hidden
12.32 MB

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

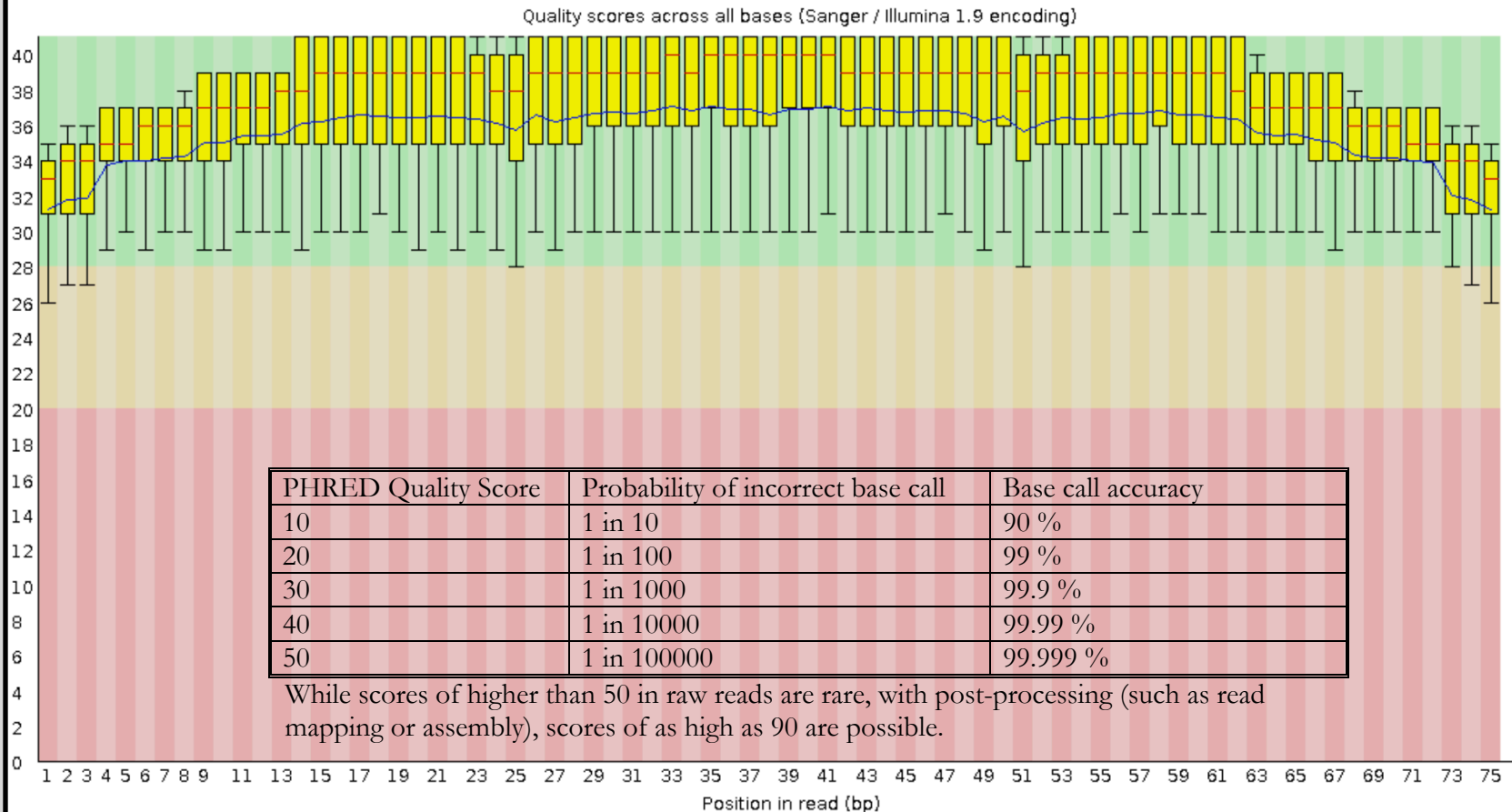
		ASCII Range	Offset	Score Range
Sanger, Illumina (Ver > 1.8)	fastqsanger	33-126	33	0-93
Solexa, Illumina (Ver < 1.3)	fastqsolexa	59-126	64	5-62
Illumina (Ver 1.3 -1.7)	fastqillumina	64-126	64	0-62

```
> library(seqinr)
> library(gtools)
> phred <- asc( s2c("DDDDCDEDCDDDDBBDDCC@") ) - 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)
```

FastQC Report



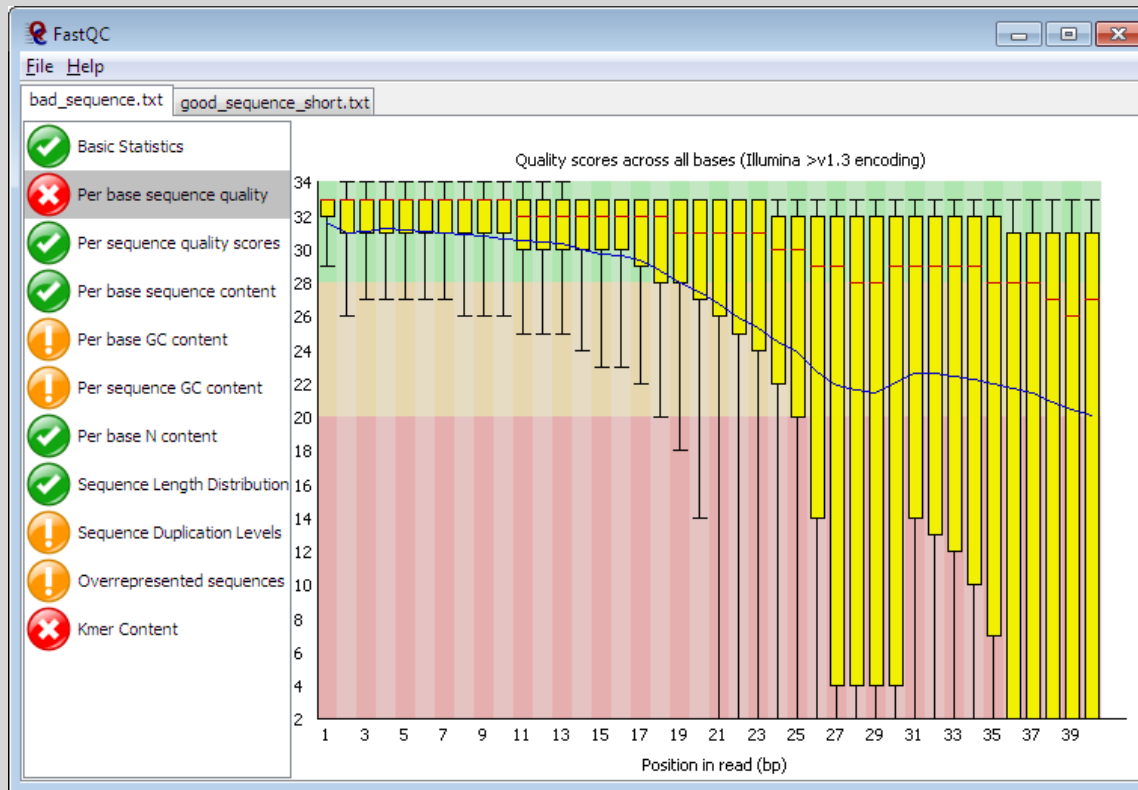
Per base sequence quality



FASTQC

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

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



Sequence Alignment

- Once sequence quality has been assessed, the next step is to align 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

Bowtie

SOAP2

Novoalign

mr/mrsFast

Eland

Blat

Bfast

BarraCUDA

CASHx

GSNAP

Mosiak

Stampy

SHRiMP

SeqMap

SLIDER

RMAP

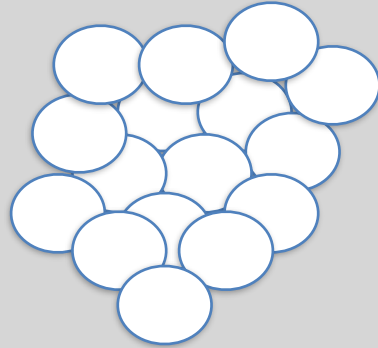
SSAHA

etc

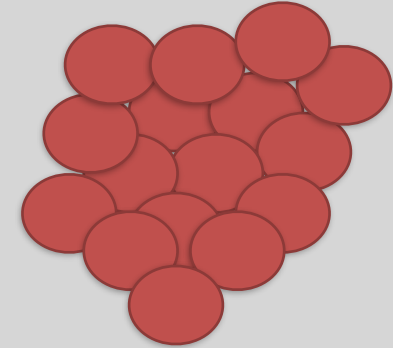
RNA Sequencing

The absolute basics

Normal Cells

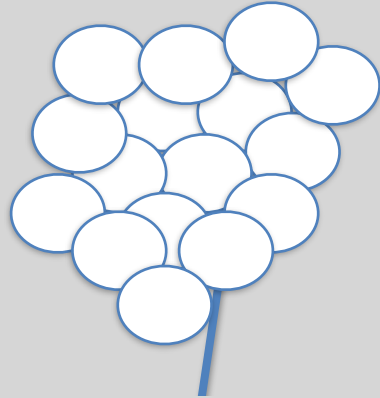


Mutated Cells

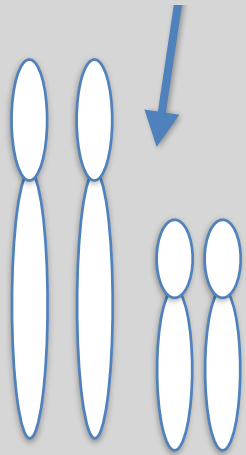


- The **mutated cells** behave differently than the **normal cells**
- We want to know what genetic mechanism is causing the difference
- One way to address this is to examine differences in gene expression via RNA sequencing...

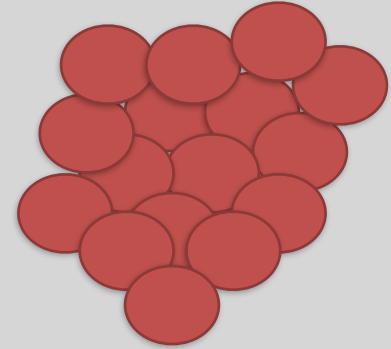
Normal Cells



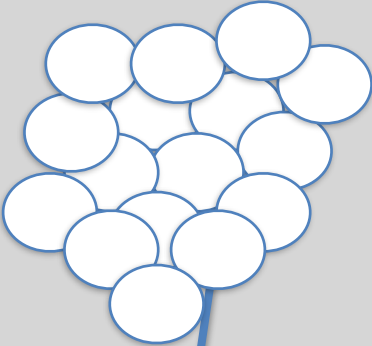
Each cell has a bunch of chromosomes



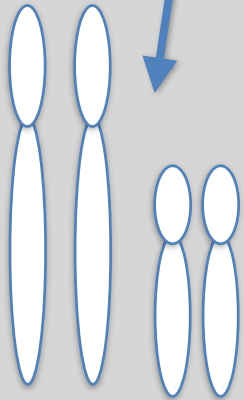
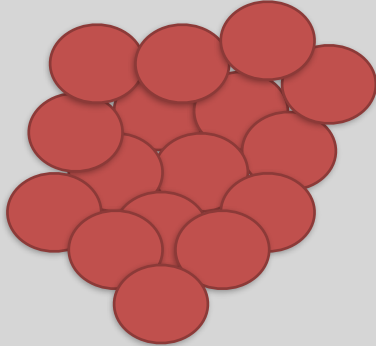
Mutated Cells



Normal Cells

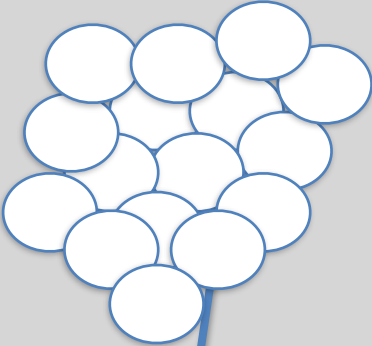


Mutated Cells

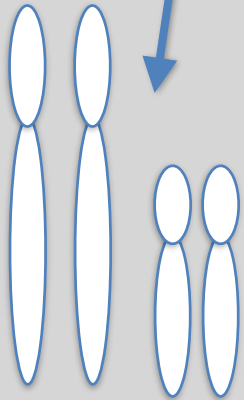
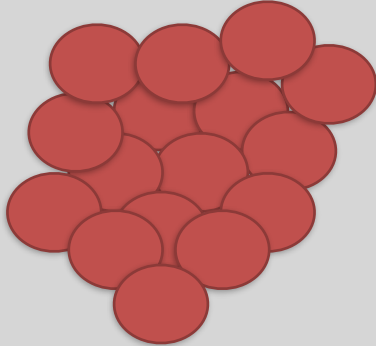


Each chromosome has a bunch of genes

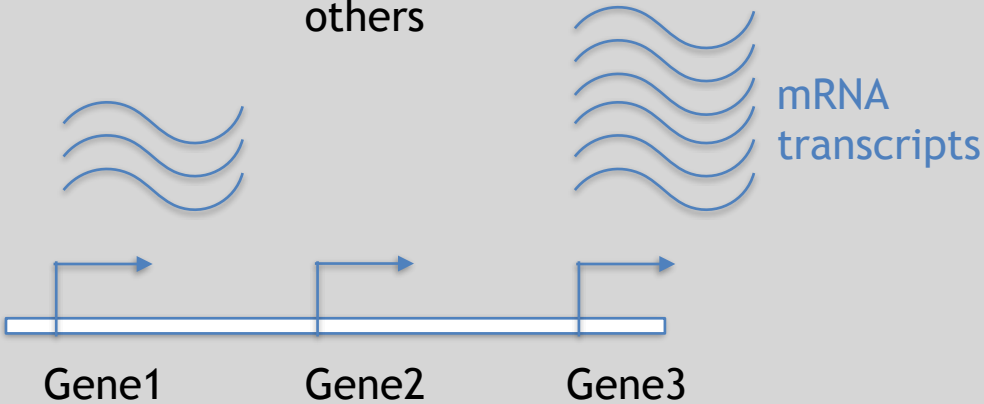
Normal Cells



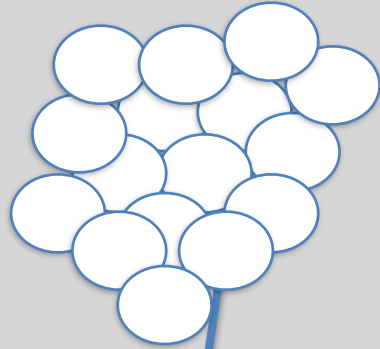
Mutated Cells



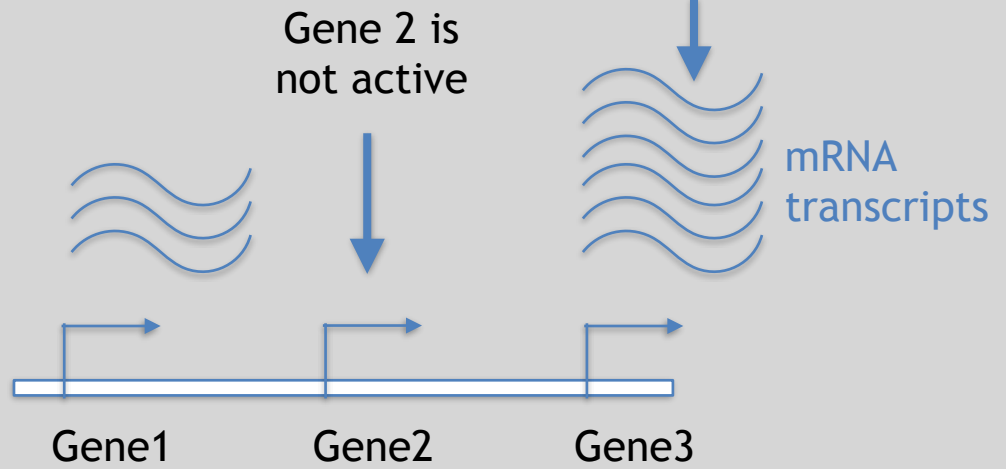
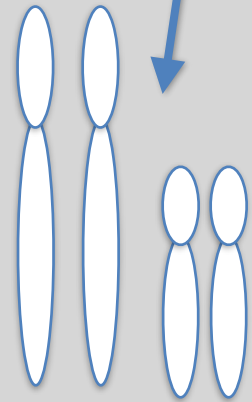
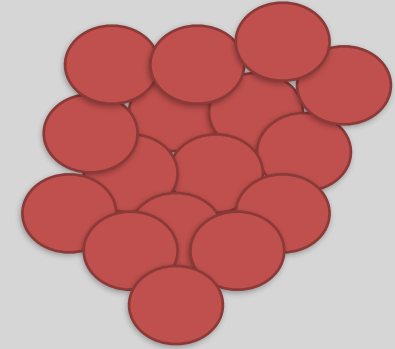
Some genes are active more than others



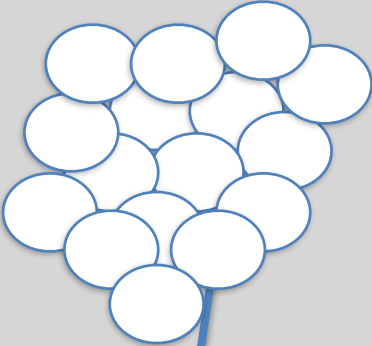
Normal Cells



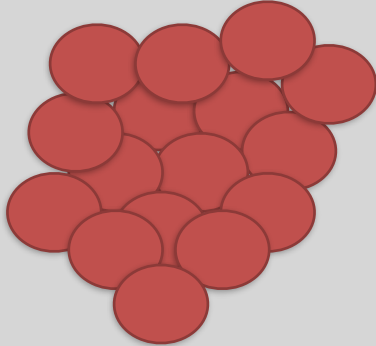
Mutated Cells



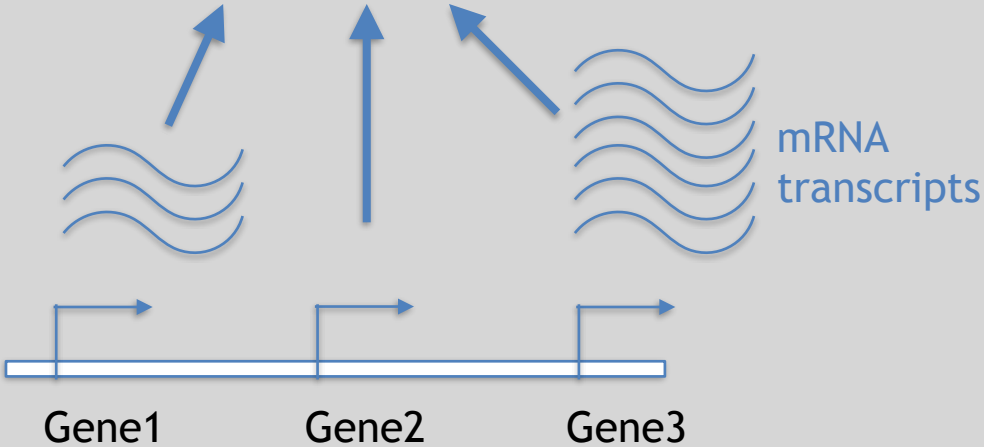
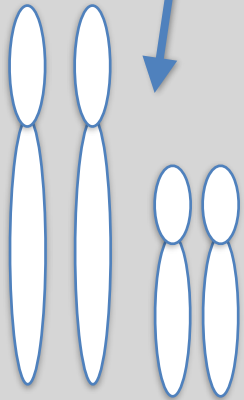
Normal Cells



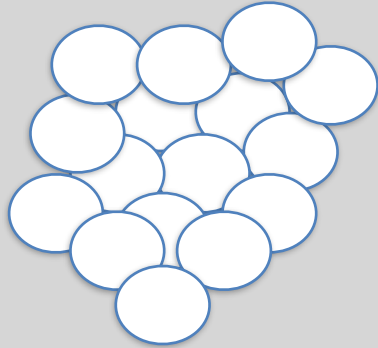
Mutated Cells



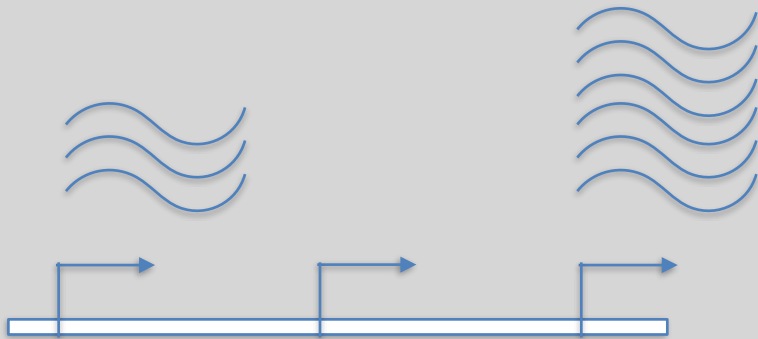
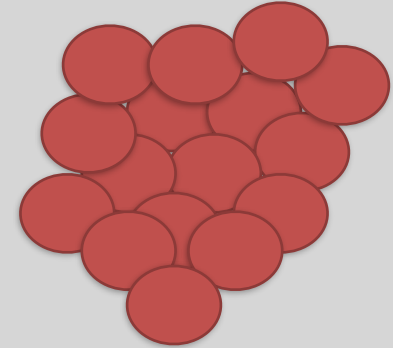
HTS tells us which genes are active, and how much they are transcribed!



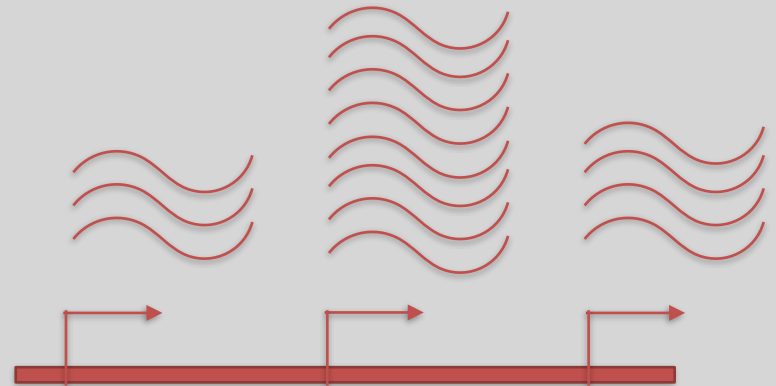
Normal Cells



Mutated Cells

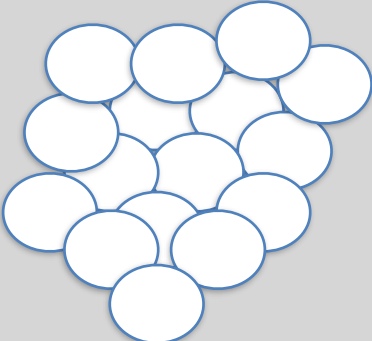


We use RNA-Seq to measure gene expression in normal cells ...

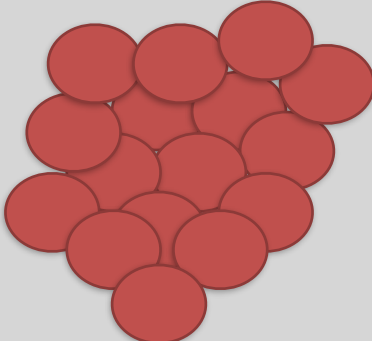


... then use it to measure gene expression in mutated cells

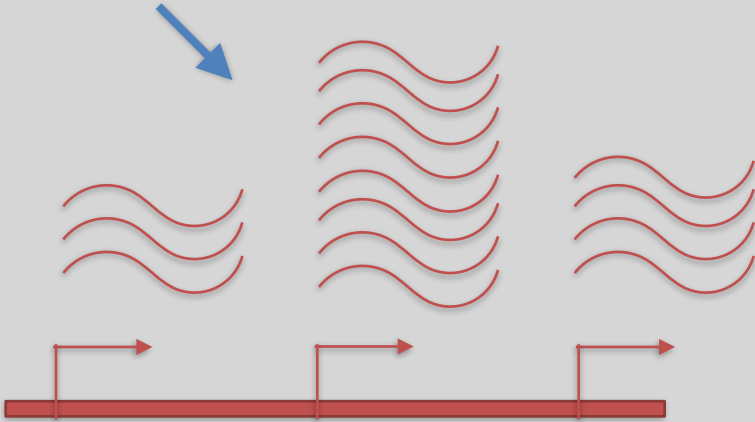
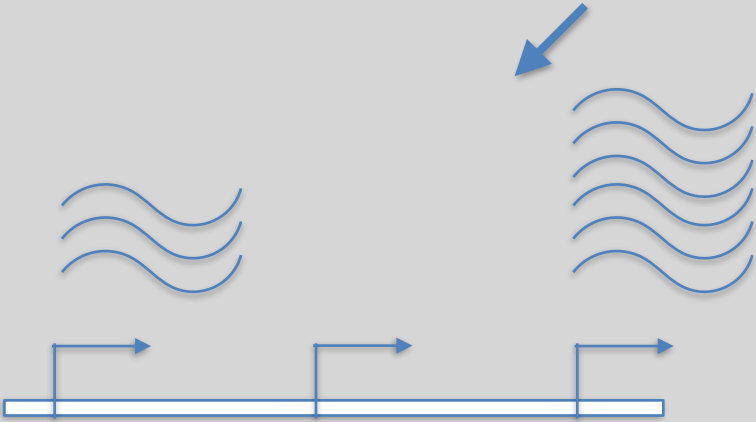
Normal Cells



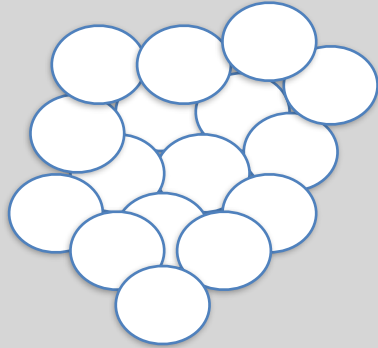
Mutated Cells



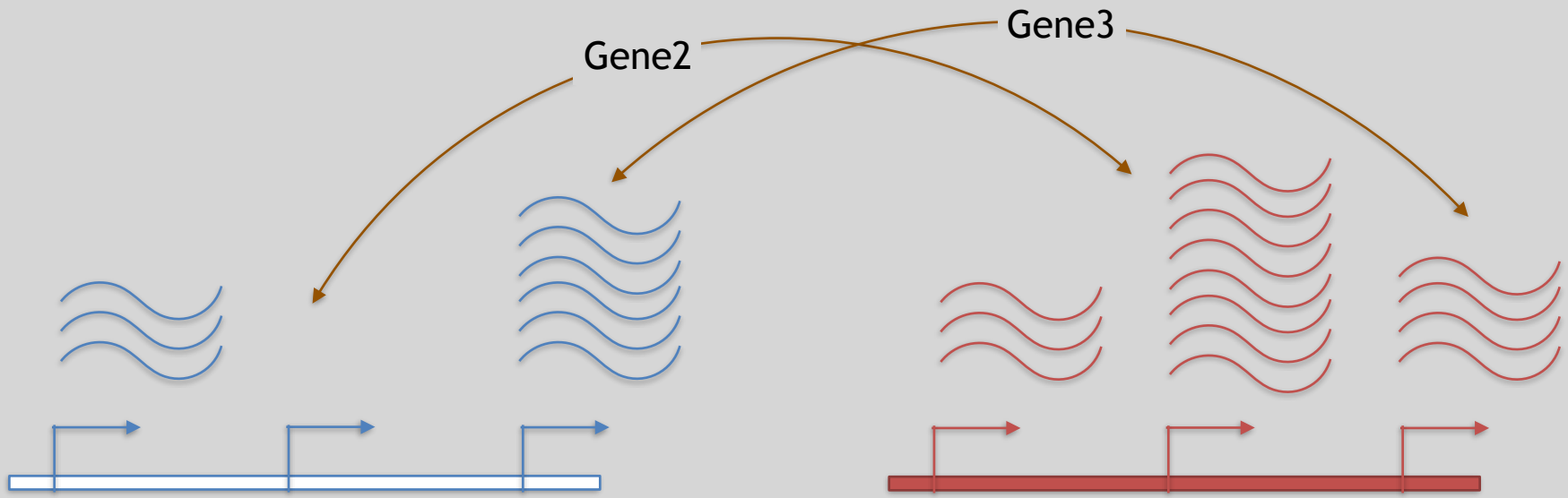
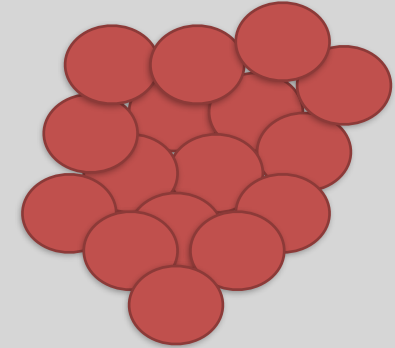
Then we can compare the two cell types to figure out what is different in the mutated cells!



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

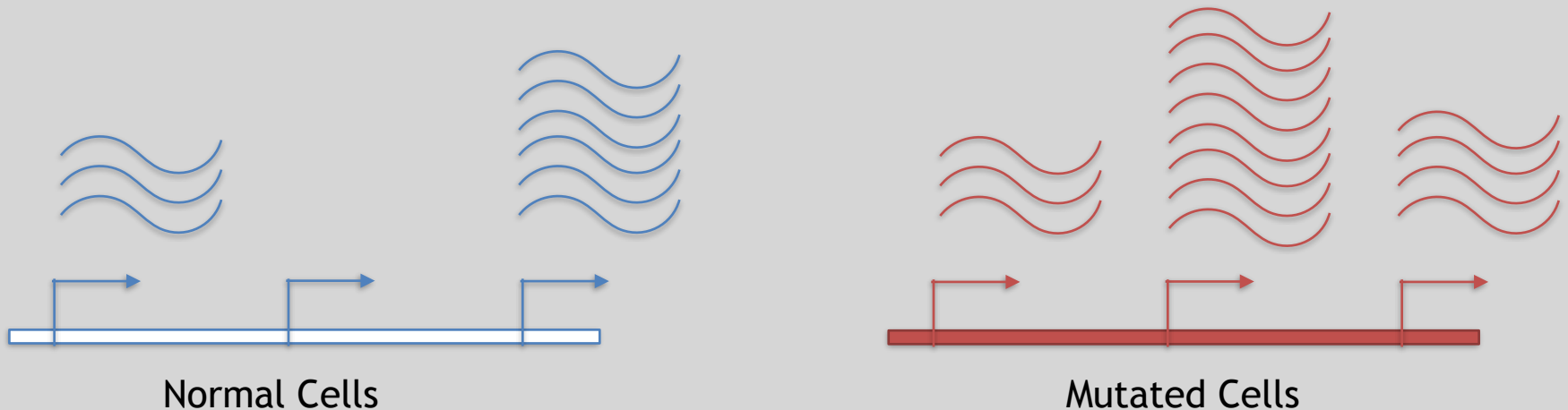
(Often the major bottleneck to overall success!)

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

Today we will get to the 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!

https://bioboot.github.io/bimm143_S19/lectures/#14

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

Feedback:

[Muddy Point Assessment]

Additional Reference Slides on SAM/BAM Format and Sequencing Methods

Sequence Alignment

Reference

- Once sequence quality has been assessed, the next step is to align 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

Bowtie

SOAP2

Novoalign

mr/mrsFast

Eland

Blat

Bfast

BarraCUDA

CASHx

GSNAP

Mosiak

Stampy

SHRiMP

SeqMap

SLIDER

RMAP

SSAHA

etc

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

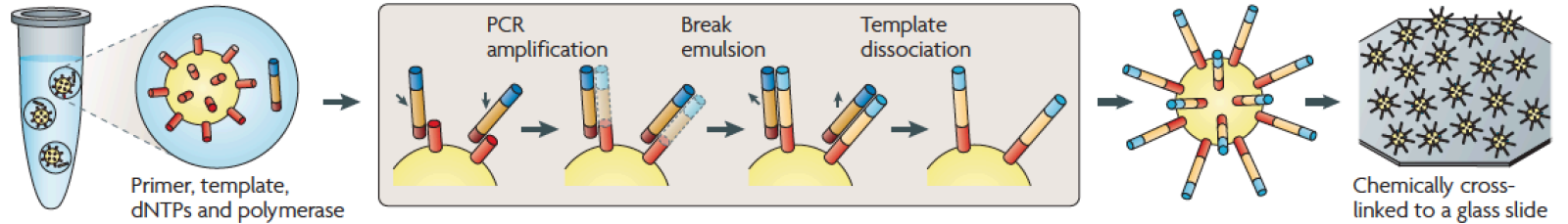
- **Samtools** is a common toolkit for analyzing and manipulating files in SAM/BAM format
 - <http://samtools.sourceforge.net/>
- **Picard** is a another set of utilities that can used to manipulate and modify SAM files
 - <http://picard.sourceforge.net/>
- These can be used for viewing, parsing, sorting,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

Additional Reference Slides on Sequencing Methods

Roche 454 - Pyrosequencing

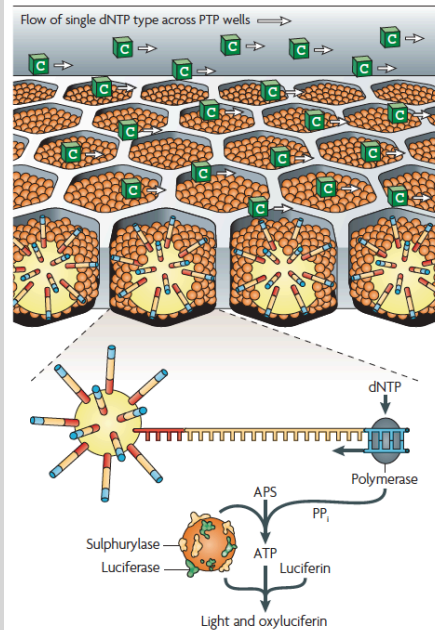
a Roche/454, Life/APG, Polonator Emulsion PCR

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



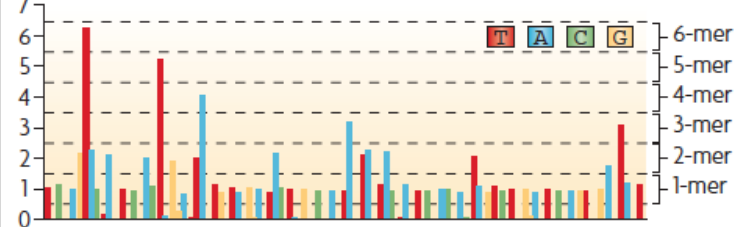
c Roche/454 — Pyrosequencing

1-2 million template beads loaded into PTP wells



d Flowgram

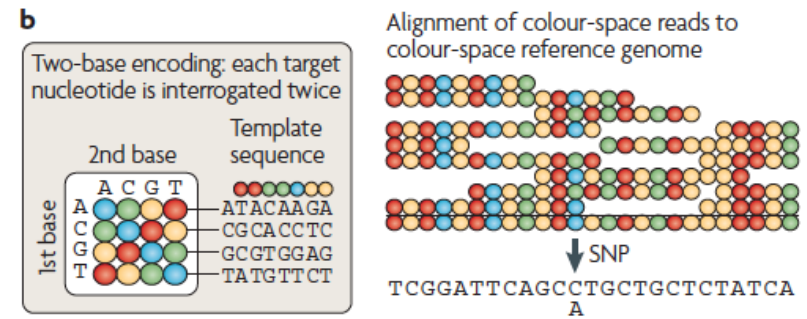
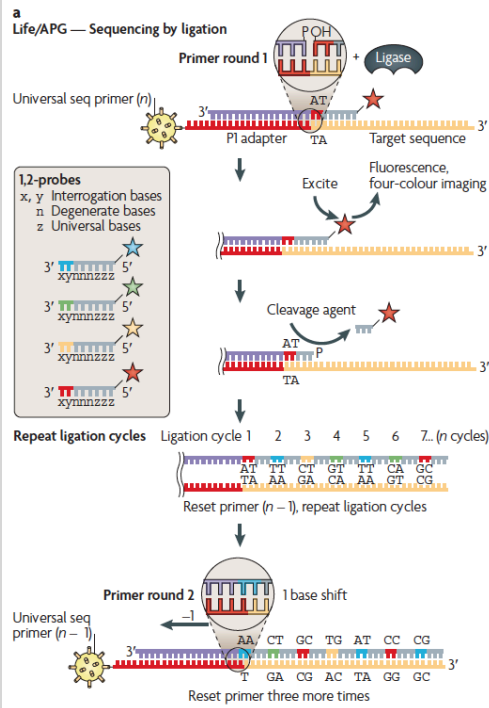
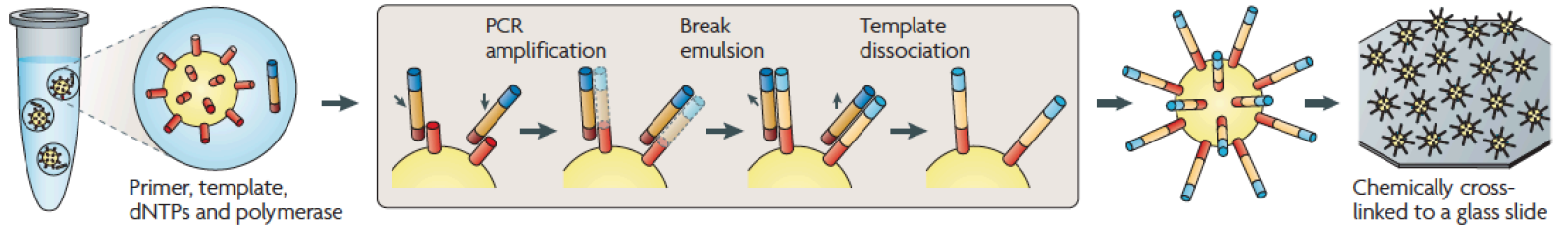
TCAGGTTTTTTTAAACAATCAACTTTTTGGATTAAAAATGTAGATAACTG
CATAAATTAATAACATCACATTAGTCTGATCAGTGAATTTAT



Life Technologies SOLiD - Sequence by Ligation

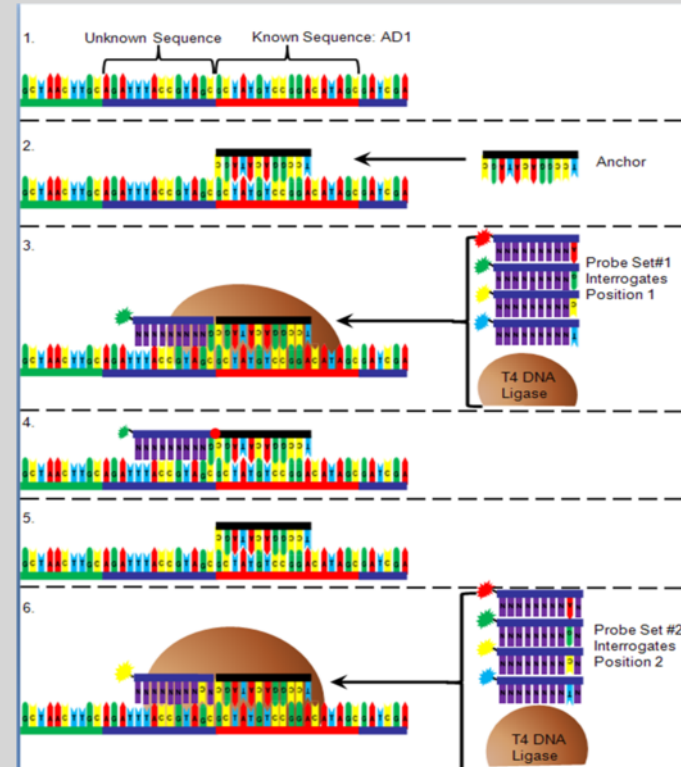
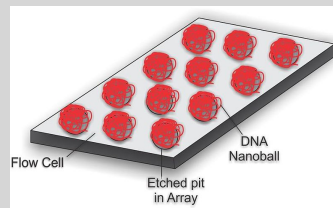
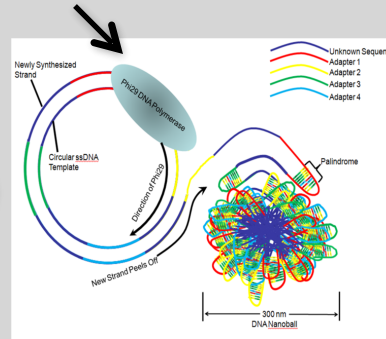
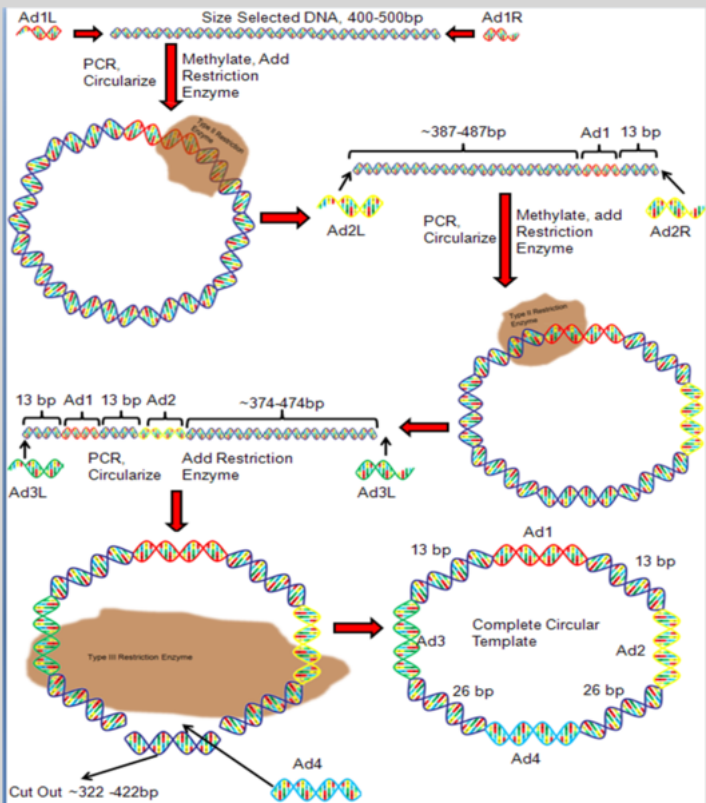
a Roche/454, Life/APG, Polonator Emulsion PCR

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



Complete Genomics - Nanoball Sequencing

Has proofreading ability!

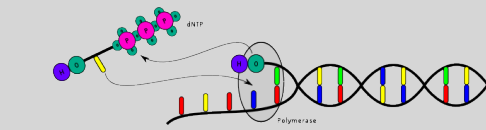


“Benchtop” Sequencers

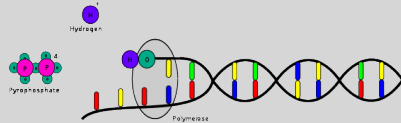
- Lower cost, lower throughput alternative for smaller scale projects
- Currently three significant platforms
 - Roche 454 GS Junior
 - Life Technology Ion Torrent
 - Personal Genome Machine (PGM)
 - Proton
 - Illumina MiSeq

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 compliment the template - no release of hydrogen.



The nucleotide compliments the template - hydrogen is released.



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

