



BIMM 143

Genome Informatics I

Lecture 13

Barry Grant
UC San Diego

<http://thegrantlab.org/bimm143>

TODAYS 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**

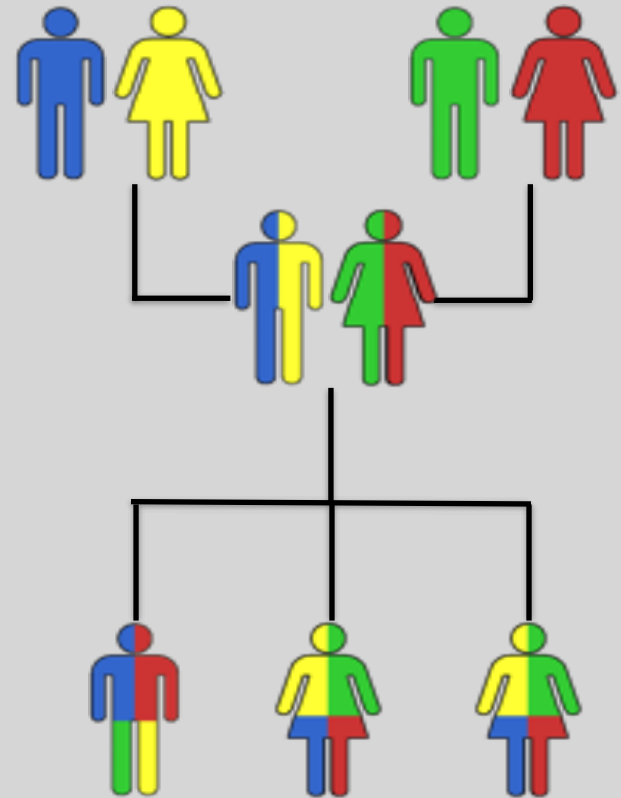
- RNA-Sequencing and Discovering variation

Genetics and Genomics

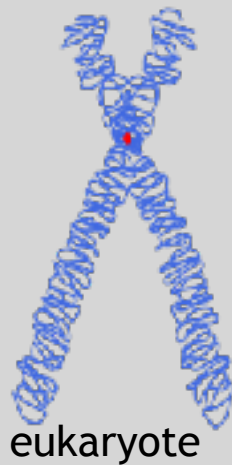
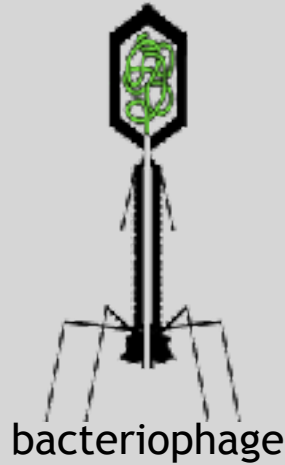
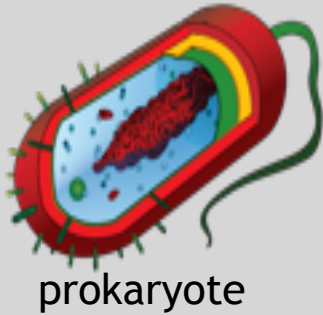
- **Genetics** is primarily the study of individual genes, mutations within those genes, and their inheritance patterns in order to understand specific traits.
- **Genomics** expands upon classical genetics and considers aspects of the entire genome, typically using computer aided approaches.

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

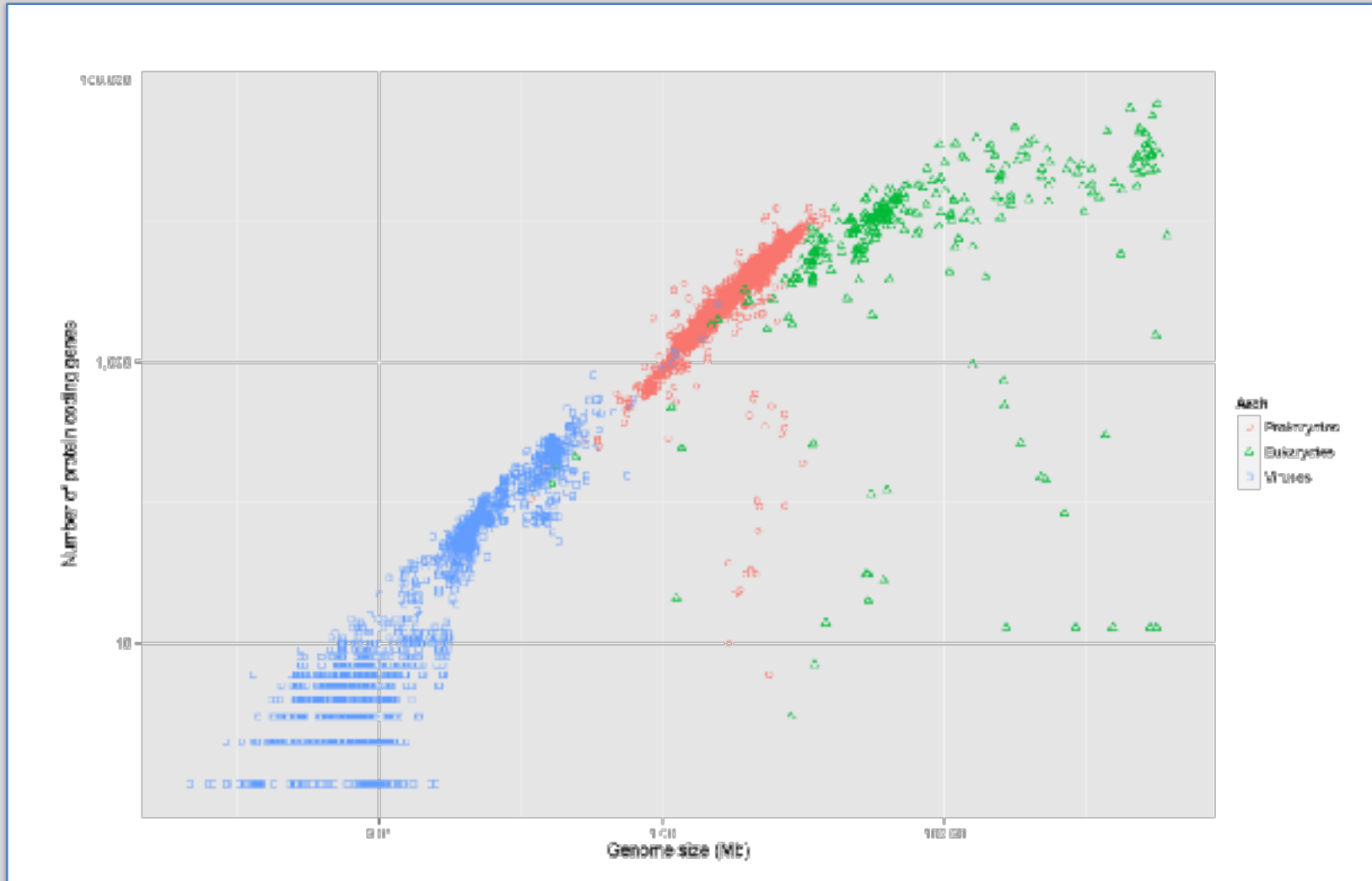


Genomes come in many shapes



- 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)

Genomes come in many sizes



Genome Databases

NCBI Genome:

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

NCBI Resources Help Us

Genome [Help](#)

[Links](#) [Advanced](#)

Genome

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

Using Genomes

- [Help](#)
- [Browse by Organism](#)
- [Download FASTA](#)
- [Download SRA](#)
- [Genome & Proteins](#)

Genome Tools

- [BLAST Entrez Local Genome](#)
- [Maskfile, Maskfile BLAST](#)
- [ToolSet \(GenomeGenomes, CompGen\)](#)

Genetic resources

- [Human genome](#)
- [Mouses](#)
- [Drosophila](#)
- [Yeast](#)
- [PROKARYOTIC GENOMES DATABASE](#)

Other Resources

- [Genetics](#)
- [BioPrint](#)
- [BioSamples](#)
- [BioProject](#)
- [Genome Commons](#)

Genome Annotation and Analysis

- [Eukaryotic Genome Annotation](#)
- [Prokaryotic Genome Annotation](#)
- [RASC \(Rapid Sequence Comparison\)](#)

External Resources

- [GOLD - Genomes On Line Database](#)
- [Ensembl Genome Browser](#)
- [Bacteria Genomes at Sanger](#)
- [Large Scale Genome Sequencing \(LSGS\)](#)

You are here: [NCBI](#) > [Genomes & Maps](#) > [Genomes](#) [Help Us](#)

GETTING STARTED

- [NCBI Education](#)
- [NCBI Help Manual](#)
- [NCBI Handbook](#)
- [Training & Tutorials](#)

RESOURCES

- [Chromosomes & Stems](#)
- [Data & Software](#)
- [DNA & RNA](#)
- [Genome 4-Structures](#)
- [Genes & Expression](#)
- [Genetics & Medicine](#)
- [Genomes & Maps](#)
- [Microbiology](#)
- [Literature](#)
- [Proteins](#)
- [Sequence Analysis](#)
- [Taxonomy](#)
- [Training & Tutorials](#)
- [Visitor](#)

POPULAR

- [PubMed](#)
- [Epubs on](#)
- [PubMed Central](#)
- [PubMed Health](#)
- [BLAST](#)
- [Nucleotide](#)
- [Genome](#)
- [SRA](#)
- [Gene](#)
- [Ensembl](#)
- [PubChem](#)

FEATURED

- [Genetic Testing Registry](#)
- [PubMed Health](#)
- [GenBank](#)
- [RefSeq](#)
- [Gene Expression Omnibus](#)
- [Viral Accession](#)
- [Human Genome](#)
- [Mouse Genome](#)
- [Influenza Virus](#)
- [Omicron BLAST](#)
- [Sequence Read Archive](#)

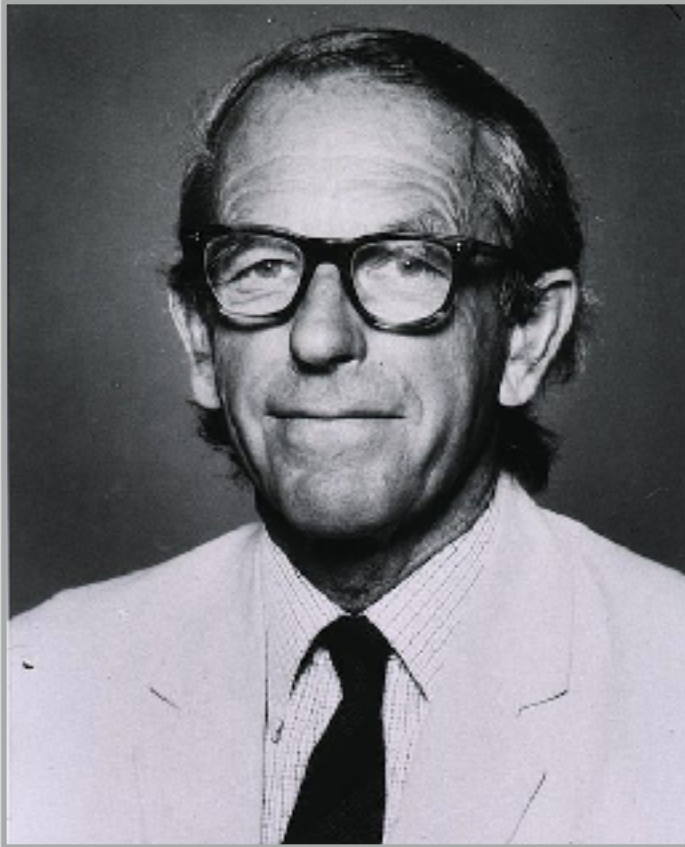
NCBI INFORMATION

- [About NCBI](#)
- [Research at NCBI](#)
- [NCBI Press](#)
- [NCBI FTP Sites](#)
- [NCBI on Facebook](#)
- [NCBI on Twitter](#)
- [NCBI on YouTube](#)

[Copyright](#) | [Disclaimer](#) | [Privacy](#) | [Workshops](#) | [Accessibility](#) | [Contact](#)

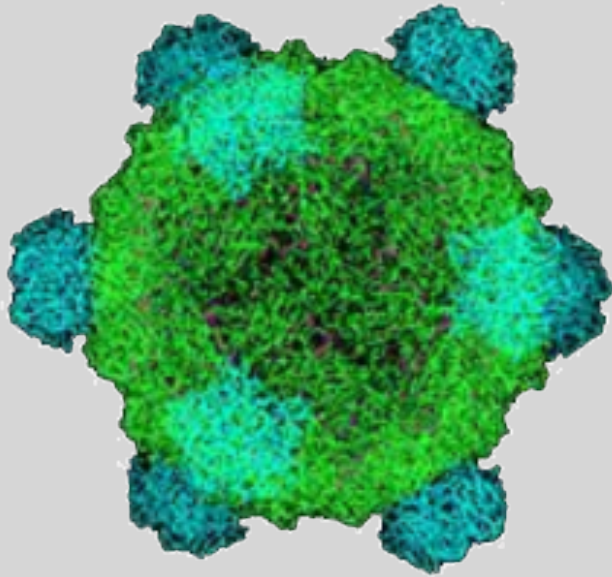
National Center for Biotechnology Information, 101 National Mall, Bethesda, MD 20894 USA

Early Genome Sequencing



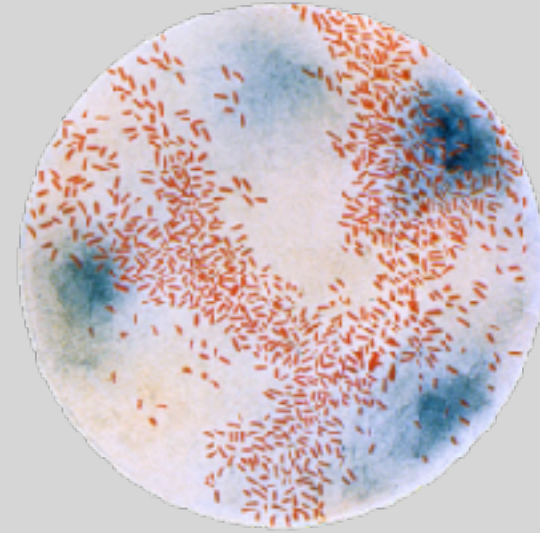
- Chain-termination “Sanger” sequencing was developed in 1977 by Frederick Sanger, colloquially referred to as the “Father of Genomics”
- Sequence reads were typically 750-1000 base pairs in length with an error rate of $\sim 1 / 10000$ bases

The First Sequenced Genomes



Bacteriophage ϕ -X174

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

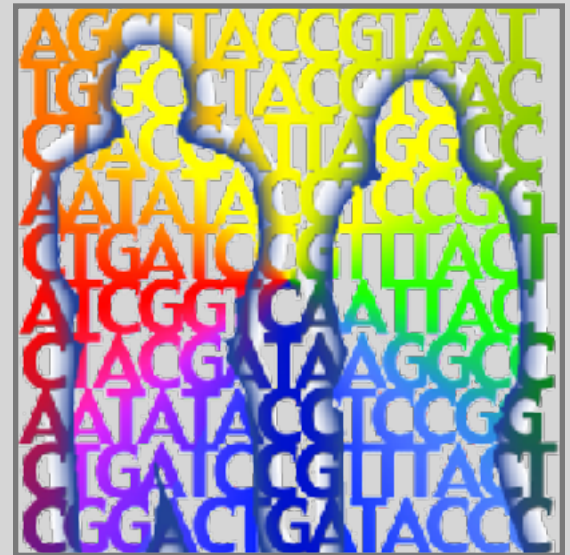


Haemophilus influenzae

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

The Human Genome Project

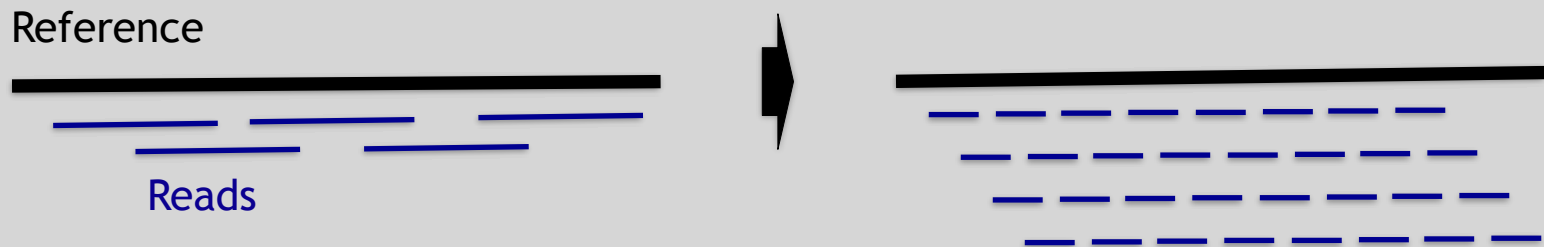
- The Human Genome Project (HGP) was an international, public consortium that began in 1990
 - Initiated by James Watson
 - Primarily led by Francis Collins
 - Eventual Cost: \$2.7 Billion
- Celera Genomics was a private corporation that started in 1998
 - Headed by Craig Venter
 - Eventual Cost: \$300 Million
- Both initiatives released initial drafts of the human genome in 2001
 - ~3.2 Billion base pairs, dsDNA
 - 22 autosomes, 2 sex chromosomes
 - ~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

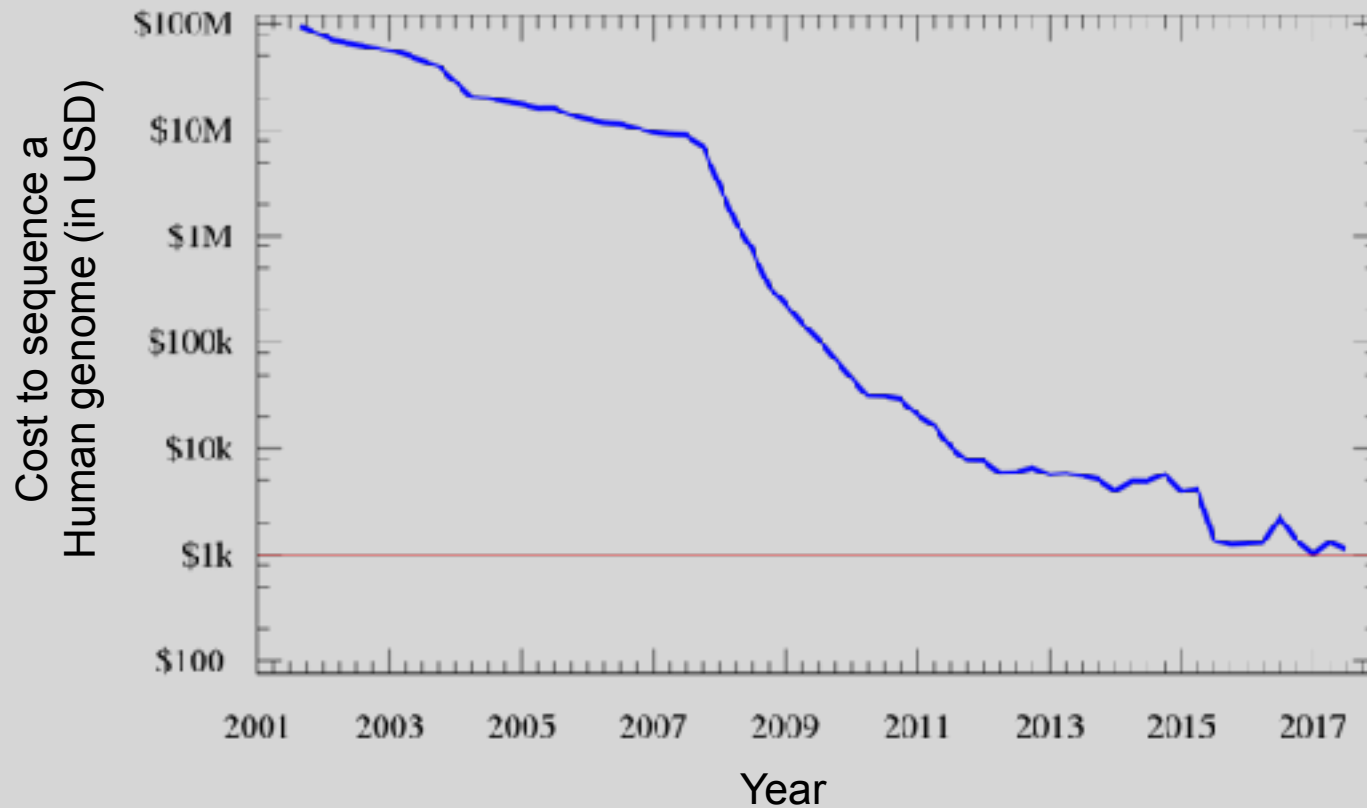


Image source: https://en.wikipedia.org/wiki/Carlson_curve

Rapid progress of genome sequencing

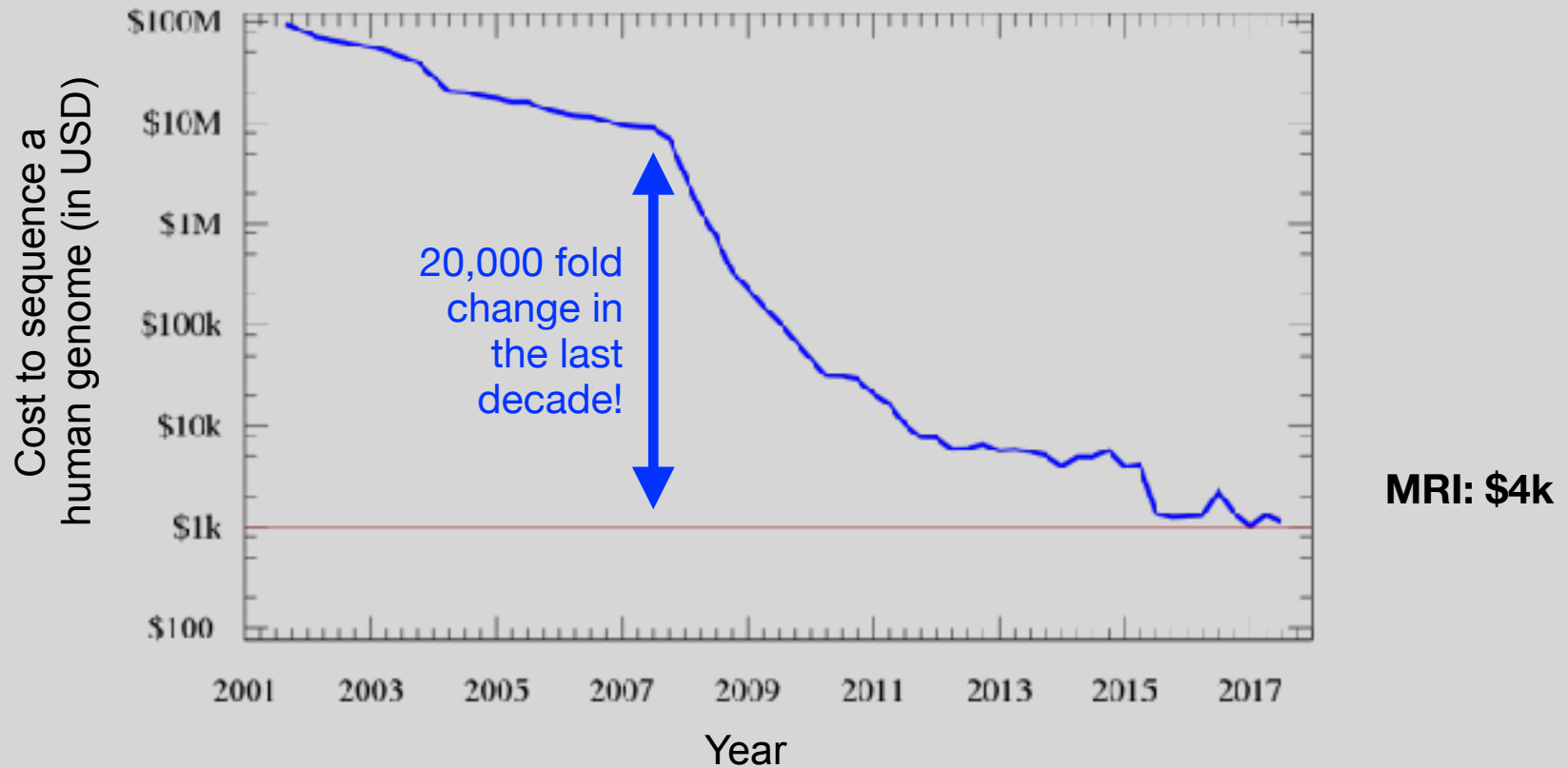
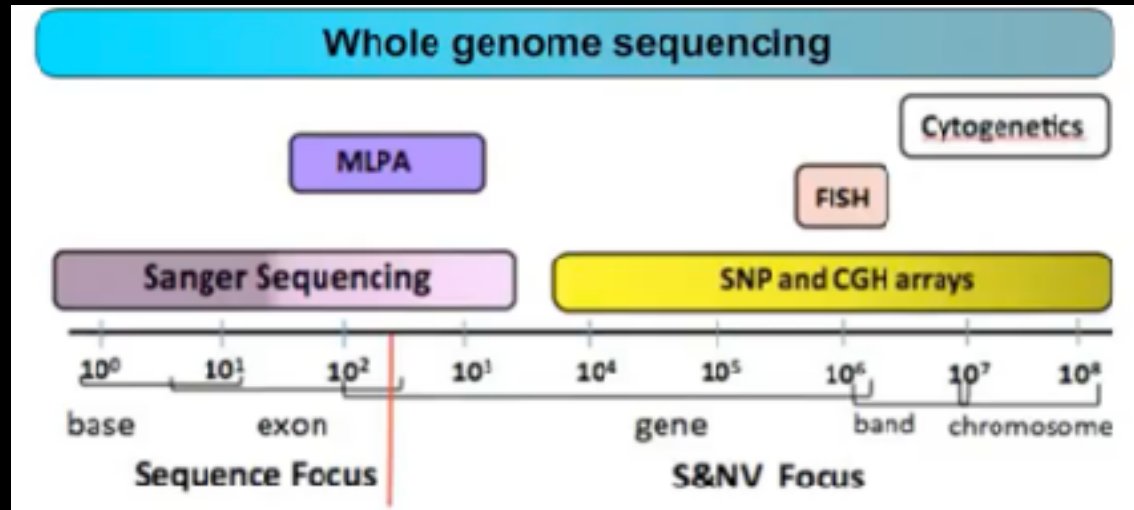


Image source: https://en.wikipedia.org/wiki/Carlson_curve

Whole genome sequencing transforms genetic testing



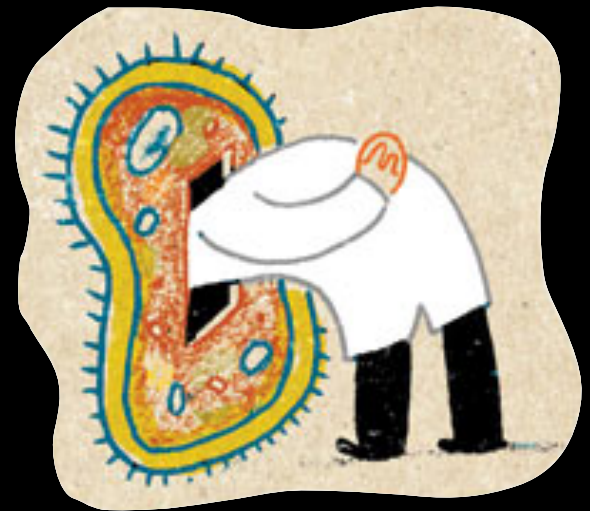
- 1000s of single gene tests
- Structural and copy number variation tests
- Permits hypothesis free diagnosis

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

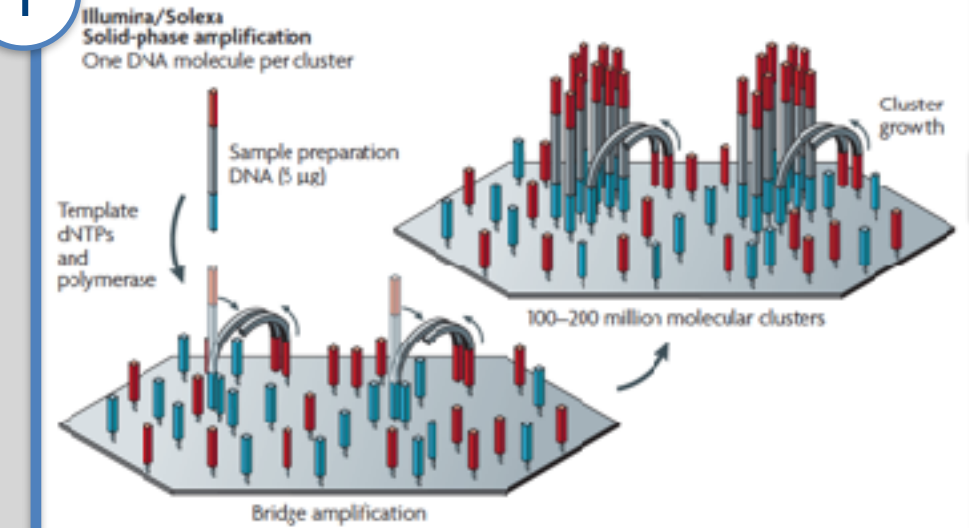
- **Sequencing by Synthesis:** Uses a polymerase to incorporate and assess nucleotides to a primer sequence
 - 1 nucleotide at a time
- **Sequencing by Ligation:** Uses a ligase to attach hybridized sequences to a primer sequence
 - 1 or more nucleotides at a time (e.g. dibase)

Modern NGS Sequencing Platforms

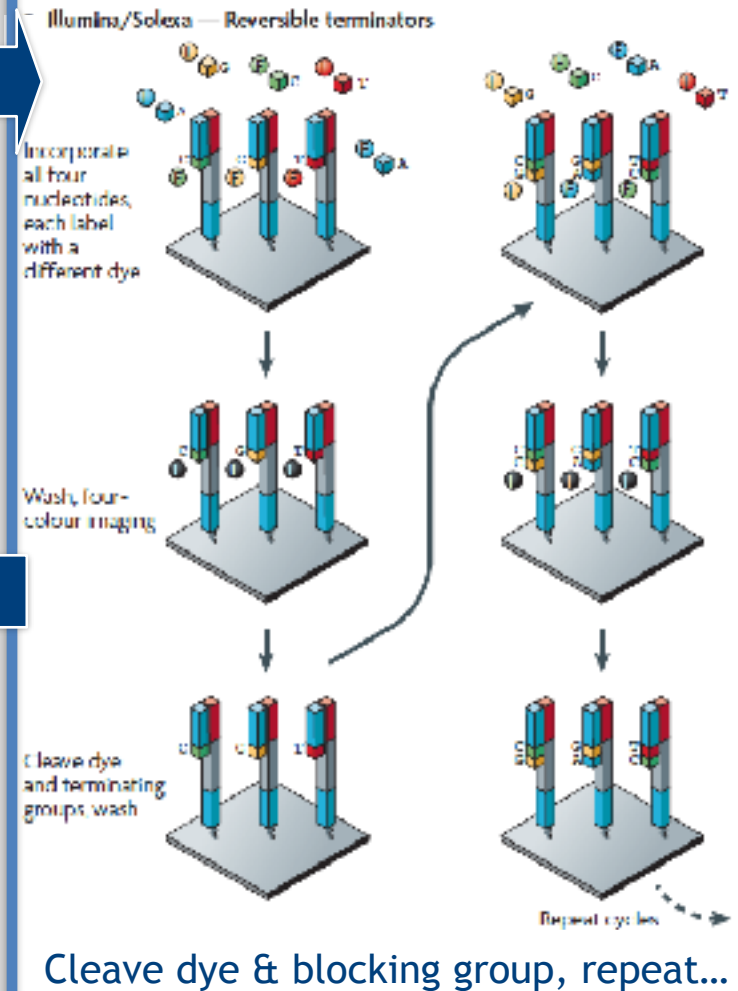
	Roche/454	Life Technologies SOLiD	Illumina Hi-Seq 2000
Library amplification method	emPCR* on bead surface	emPCR* on bead surface	Enzymatic amplification on glass surface
Sequencing method	Polymerase-mediated incorporation of unlabelled nucleotides	Ligase-mediated addition of 2-base encoded fluorescent 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

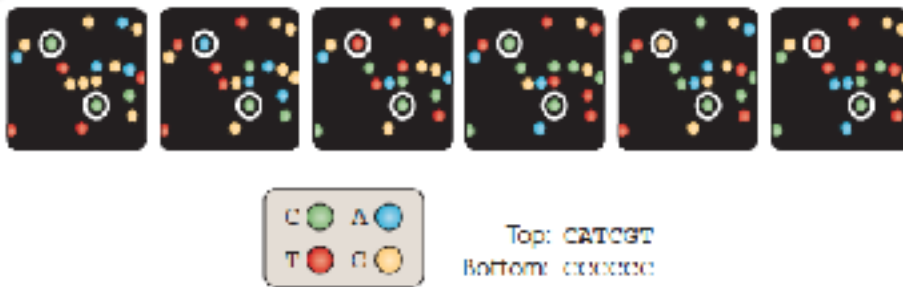
1 Enzymatic amplification on glass surface



2 Polymerase-mediated incorporation of end blocked fluorescent nucleotides



3 Fluorescent emission from incorporated dye-labeled nucleotides



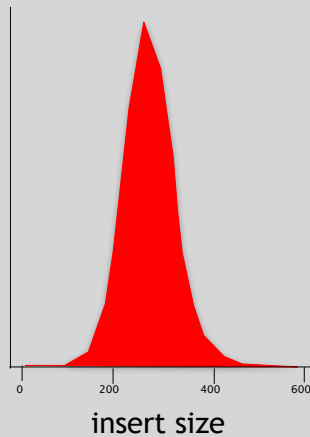
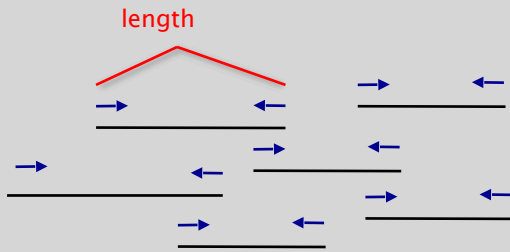
Illumina Sequencing - Video



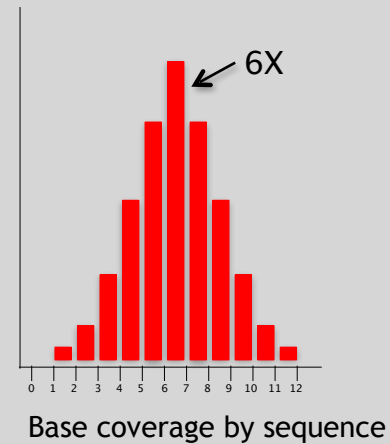
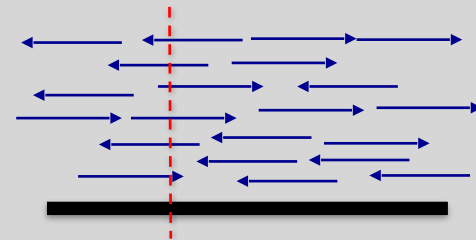
https://www.youtube.com/watch?src_vid=womKfikWlxM&v=fCd6B5HRaZ8

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-seq 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 (300–1000 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](#) [Feed](#) [View source](#) [View history](#) [Go](#) [Items](#)

Software list

Results (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:

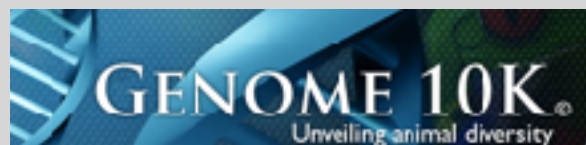
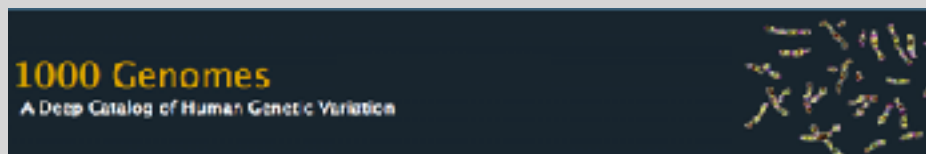
OS:

Name	Summary	Tags	OS	Features	License	OS
Repeat	Aligns overlapping sequencing trace files, multi-reading alignment, RepeatMasker and overlapping sequences.	Sequencing	Sequence analysis		Freeware	Mac OS X
A-Large mode tool	Identifies deviations in clone-read size that indicate three-dimensional structural variations in a read by a reference genome.	INDEL discovery Sequencing	Mapping		GPL	Linux 32
A-Small read Tool	The SOLO™ Small Read Tool processes the raw evidence found in the passing step of the SOLO™ System Analysis Pipeline Tool (Genome 1.0).	INDEL discovery Sequencing	Mapping Alignment		GPL	Linux 32
A-RA	Assembly-based hybrid RNA-seq analysis is a comparative gene assembler, which uses paired-end and unpaired reads from parallel libraries to help build a better assembly.	Genome Assembly	Assembly Scaffolding		Attribution License	Linux
A-Mapper	Maps RNA-seq reads to target genome considering possible multiple mapping locations and splice junctions.	Genome Transcriptomics	Mapping Alignment		GPLv3	Linux
A-BUS	MBUS is a de novo assembler designed for short reads and large genomes.	De novo assembly	Assembly De-bruyn graph	GPL OpenMP	Free for academic use	POSIX Linux Mac OS X
A-Genie	Performs genetic annotations from raw read data.	Genome	Annotation Transcript		Attribution License	Linux 32

**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



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

“Variety’s the very spice of life”

-William Cowper, 1785

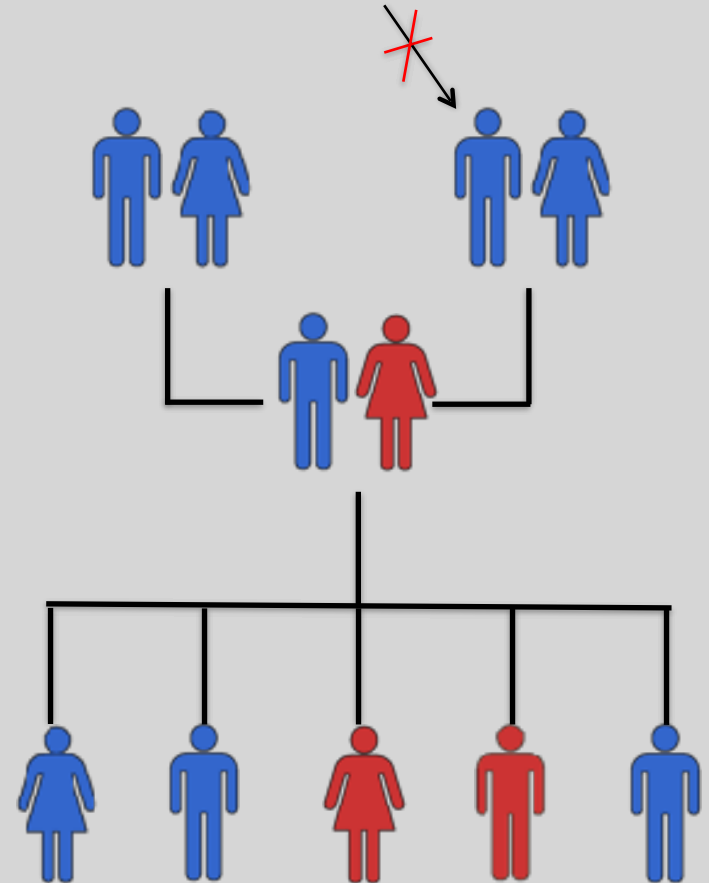
“Variation is the spice of life”

-Kruglyak & Nickerson, 2001

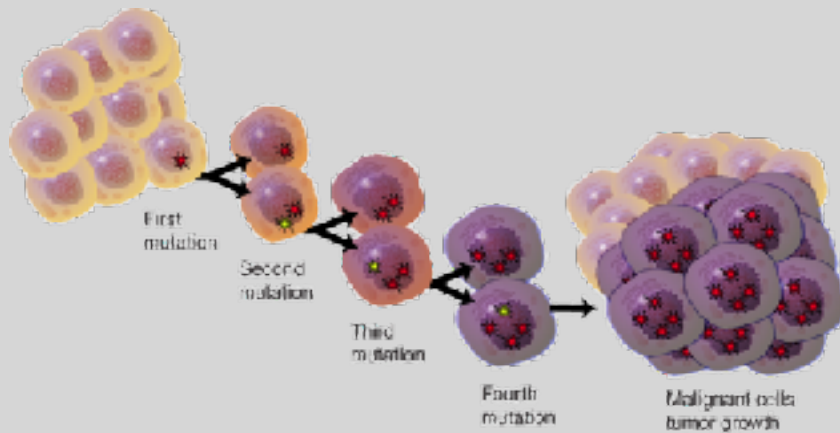
- While the sequencing of the human genome was a great milestone, the DNA from a single person is not representative of the millions of potential differences that can occur between individuals
- These unknown genetic variants could be the cause of many phenotypes such as differing morphology, susceptibility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



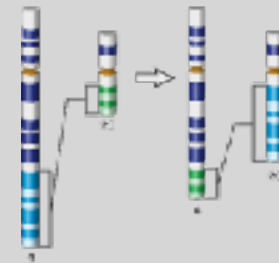
Somatic Variation



- 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

ATCCTGATTCGGTGAACGTTATCGACGATCCGATCGA
ATCCTGATTCGGTGAACGTTATCGACGATCCGATCGA
CGGTGAACGTTATCGACGATCCGATCGAACTGTCAGC
GGTGAACGTTATCGACGTTCCGATCGAACTGTCAGCG
TGAACGTTATCGACGTTCCGATCGAACTGTCATCGGC
TGAACGTTATCGACGTTCCGATCGAACTGTCAGCGGC
TGAACGTTATCGACGTTCCGATCGAACTGTCAGCGGC
GTTATCGACGATCCGATCGAACTGTCAGCGGCAAGCT
TTATCGACGATCCGATCGAACTGTCAGCGGCAAGCT

sequencing error or genetic variant?

reference genome

ATCCTGATTCGGTGAACGTTATCGACGATCCGATCGAACTGTCAGCGGCAAGCTGATCCGATCGATCGATGCTAGTG

TTATCGACGATCCGATCGAACTGTCAGCGGCAAGCT
TCGACGATCCGATCGAACTGTCAGCGGCAAGCTGAT
ATCCGATCGAACTGTCAGCGGCAAGCTGATCGATCGAT
TCCGATCGAACTGTCAGCGGCAAGCTGATCGATCGATCGA
TCCGATCGAACTGTCAGCGGCAAGCTGATCGATCGA
GATCGAACTGTCAGCGGCAAGCTGATCGATCGATCGA
AACTGTCAGCGGCAAGCTGATCGATCGATCGATGCTA
TGTCAGCGGCAAGCTGATCGATCGATCGATGCTAG
TCAGCGGCAAGCTGATCGATCGATCGATGCTAGTG

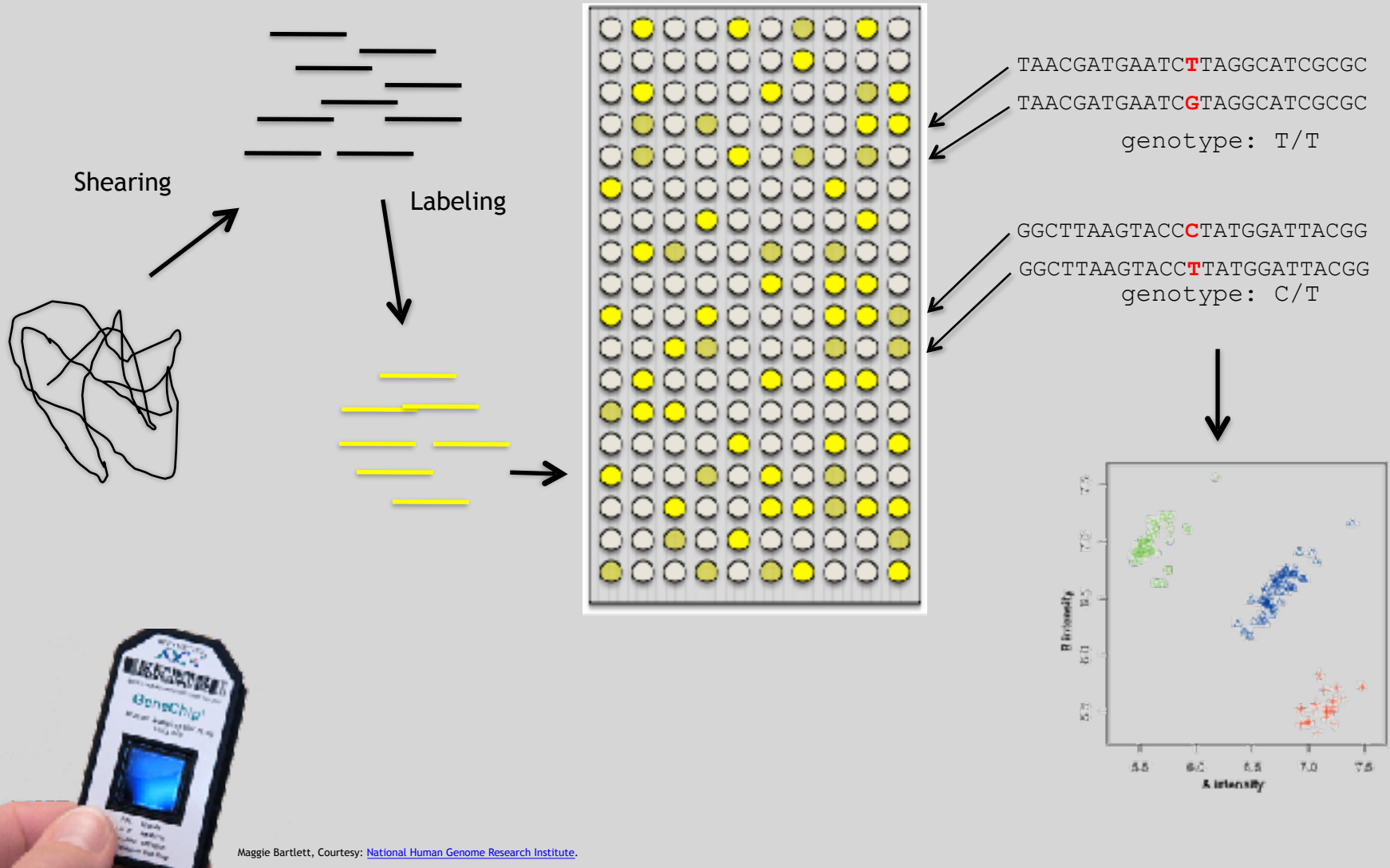
sequencing error or genetic variant?

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

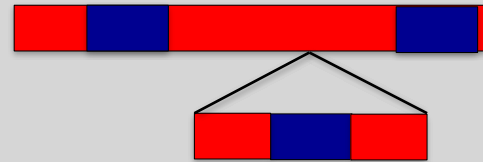
Premature stop codons



TAC → TAA



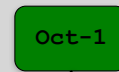
Gene or exon deletion



Frameshift mutation



TAC → T-C



Transcription factor binding disruption



ATGCAAAT → ATGCAGAT

Do it Yourself!

Hand-on time!

https://bioboot.github.io/bimm143_S18/lectures/#13

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 screenshot displays the Ensembl genome browser interface with three main sections:

- Chromosome image:** Shows a chromosome with a highlighted "Region of interest" and "Haplotypes and patches".
- Overview image:** Shows a detailed view of the region with "Genes" and a "Gene or region of interest" highlighted. It includes a "Gene Legend" and navigation controls. Callouts point to "Change or add data tracks" and "Change or add data tracks".
- Zoomable Region image:** Shows a detailed view of the region with "Transcripts (splice variants)" and a "Genome" highlighted. It includes a "Gene Legend" and navigation controls. Callouts point to "Change or add data tracks" and "Zoomable Region image".

Navigation and configuration options are visible on the left side of the interface, including "Configure this page" and "Bookmark this page".

Access a jetstream galaxy instance!

Use assigned IP address

Do it Yourself!

Galaxy

149.165.169.186

Galaxy

Analyse 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

Measblast compare short reads against hg18, m1, and vgs 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: HCO0109_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): papfam1

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

Set read groups information?

Do not set

Specifying readgroup 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.

From the bowtie2 mapping statistics to the history:

History

search datasets

Unnamed history

22 shown, 2 pending, 1 hidden

12.12 MB

25: vireo-count on data 18 and data 17 (no featu...)

24: vireo-count on data 18 and data 17

23: Cufflinks on data 18 and data 16: Skipped Tra...
nscripts

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

20: Cufflinks on data 18 and data 16: transcript e...
xpression

19: Cufflinks on data 18 and data 16: gene expe...
ssion

575 lines

format: tabular, database: hg19

```

cufflinks v2.2.1
cufflinks -q --no-update-check -l
300000 -F 0.100000 -J 0.100000 -p
6 -C /opt/galaxy/galaxy-
apc/database/datasets/000/dataset_4
/opt/galaxy/galaxy-
apc/database/datasets/000/dataset_4

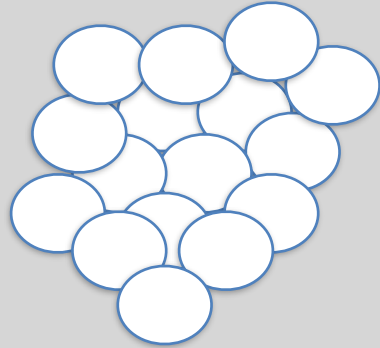
```

1	2	3
tracking_id	class_code	nearest_ref_id
ZZ01	-	-
CY802	-	-
ANKY1	-	-

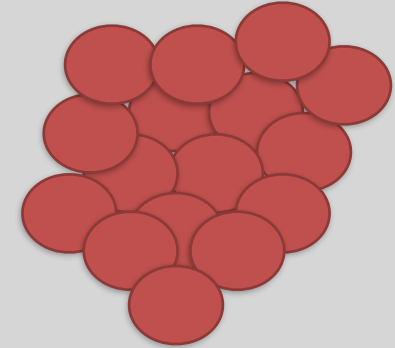
RNA Sequencing

The absolute basics

Normal Cells

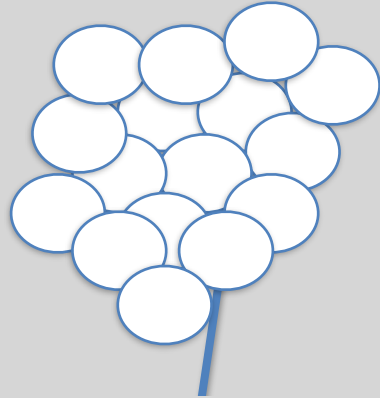


Mutated Cells

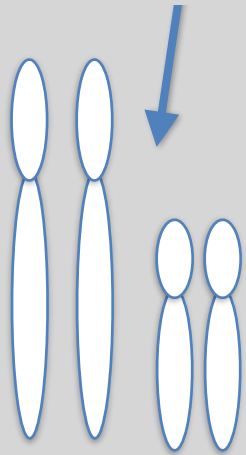


- The **mutated cells** behave differently than the **normal cells**
- We want to know what genetic mechanism is causing the difference
- One way to address this is to examine differences in gene expression via RNA sequencing...

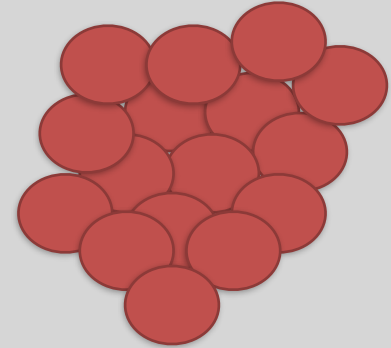
Normal Cells



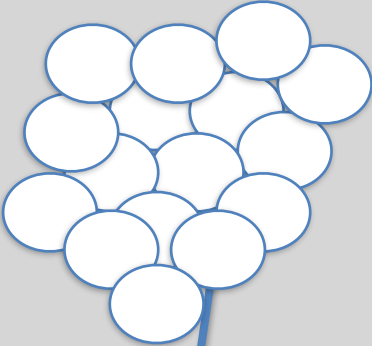
Each cell has a bunch of chromosomes



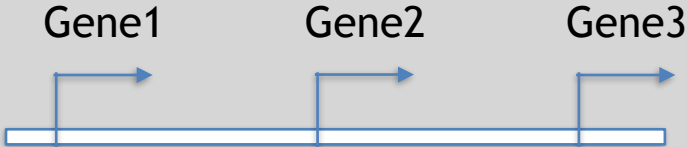
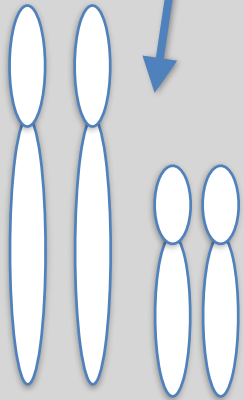
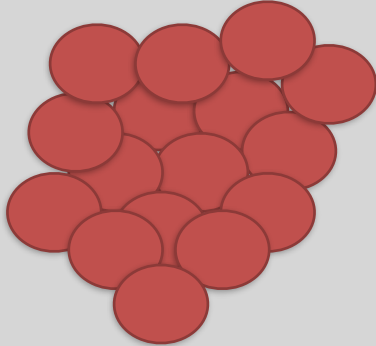
Mutated Cells



Normal Cells

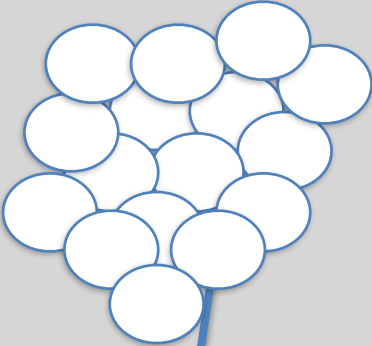


Mutated Cells

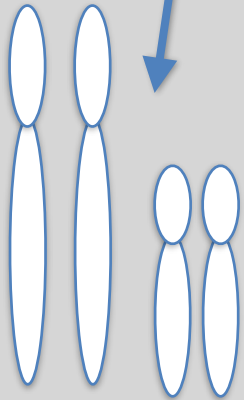
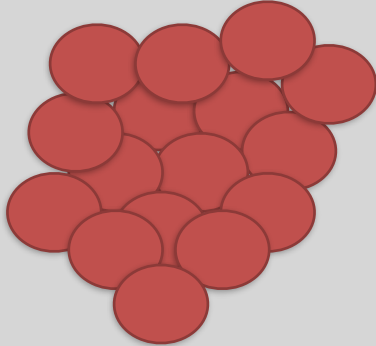


Each chromosome has a bunch of genes

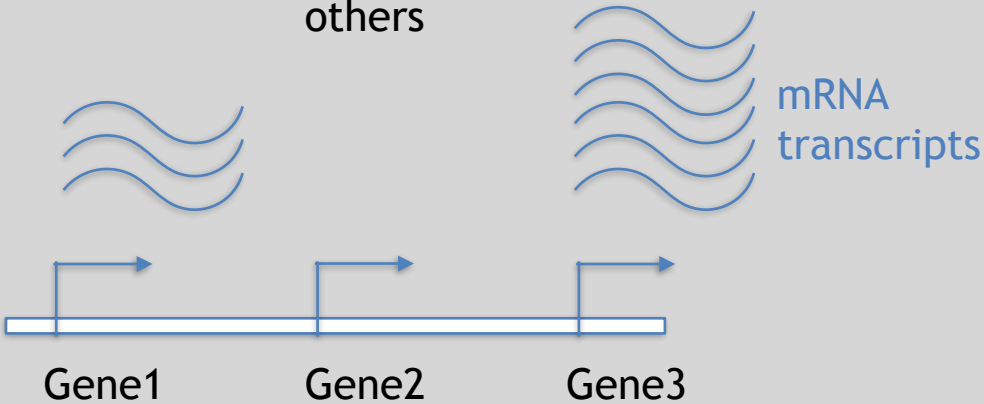
Normal Cells



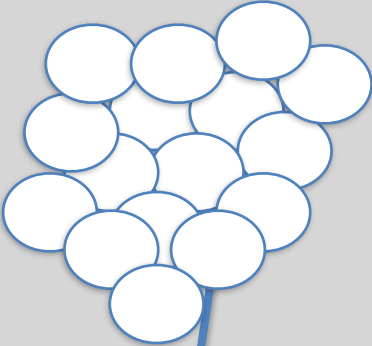
Mutated Cells



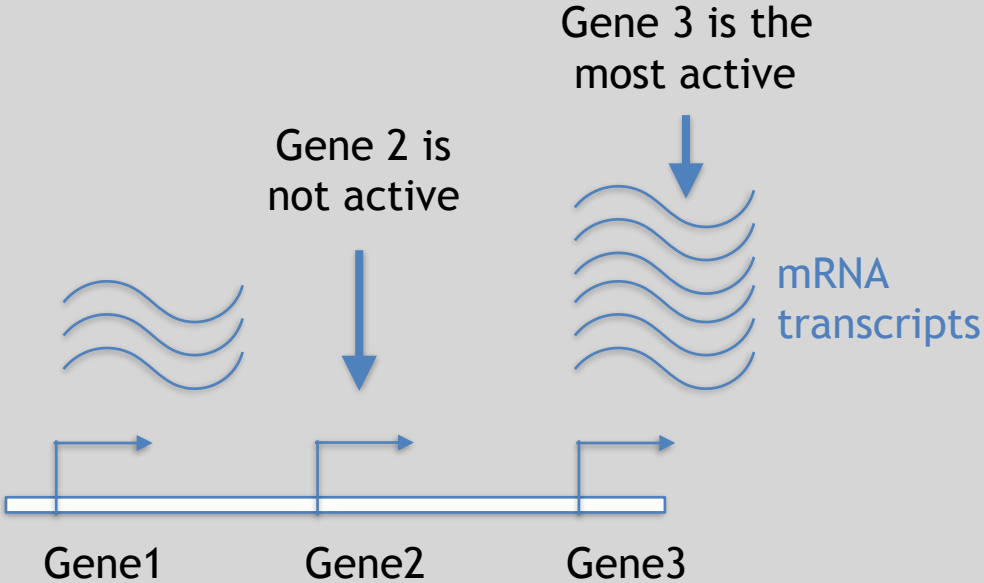
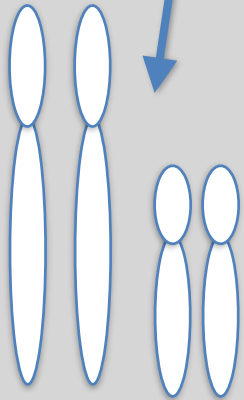
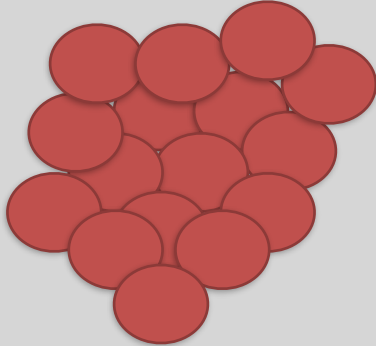
Some genes are active more than others



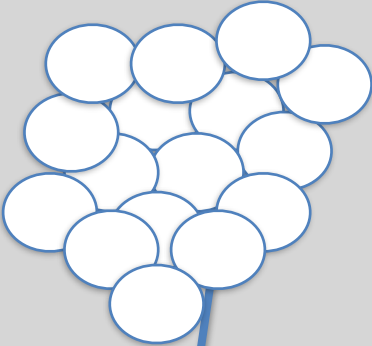
Normal Cells



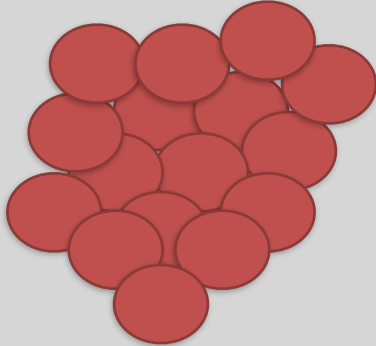
Mutated Cells



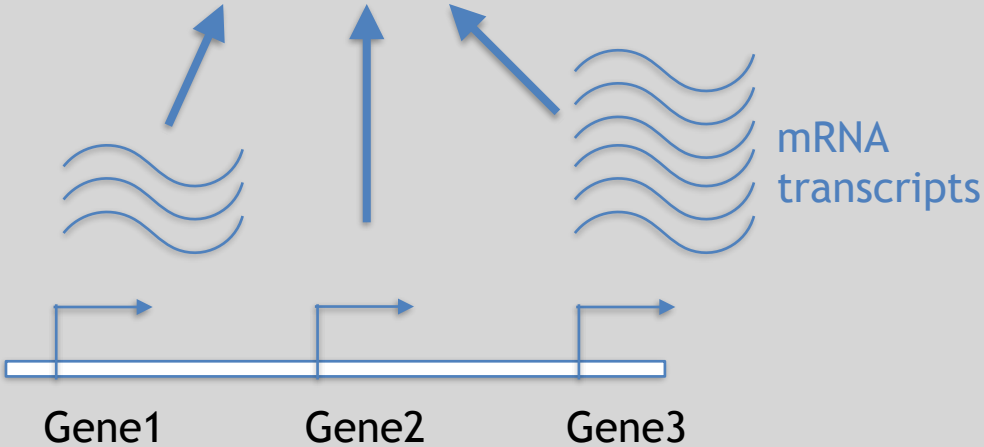
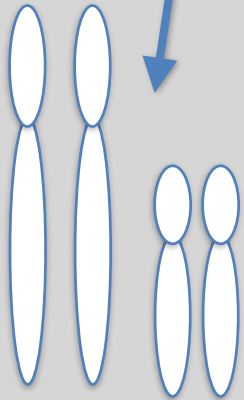
Normal Cells



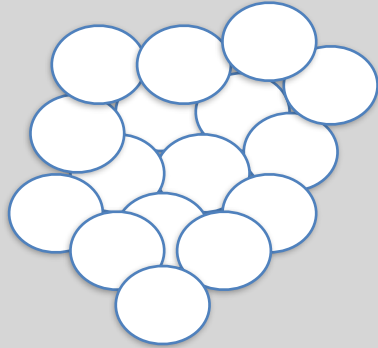
Mutated Cells



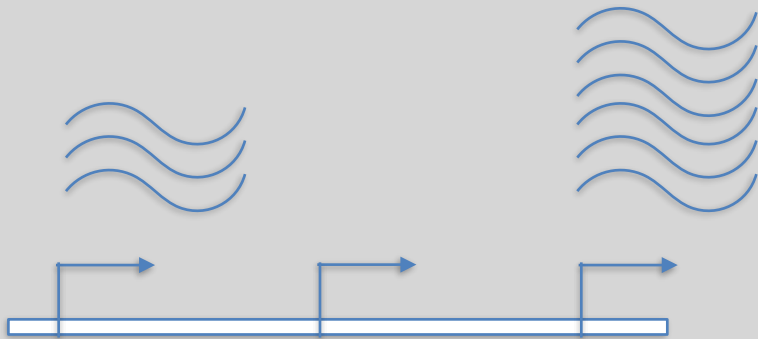
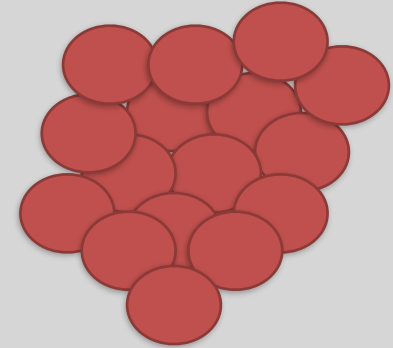
HTS tells us which genes are active, and how much they are transcribed!



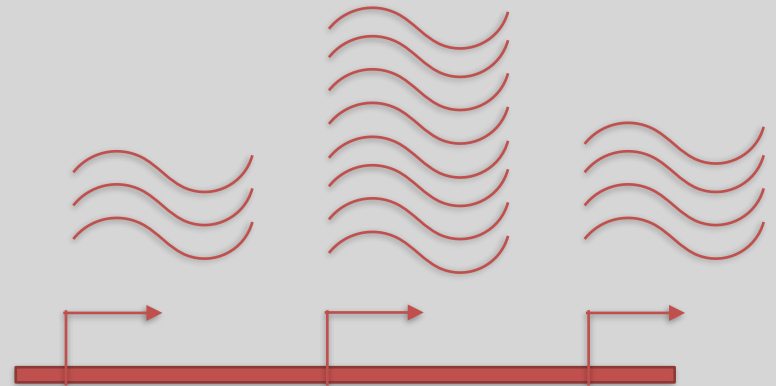
Normal Cells



Mutated Cells

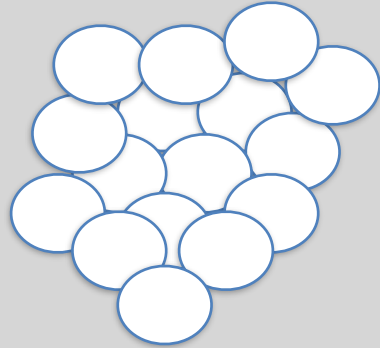


We use RNA-Seq to measure gene expression in normal cells ...

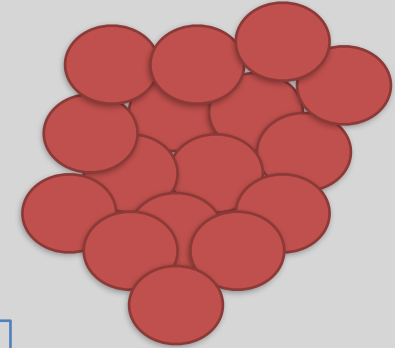


... then use it to measure gene expression in mutated cells

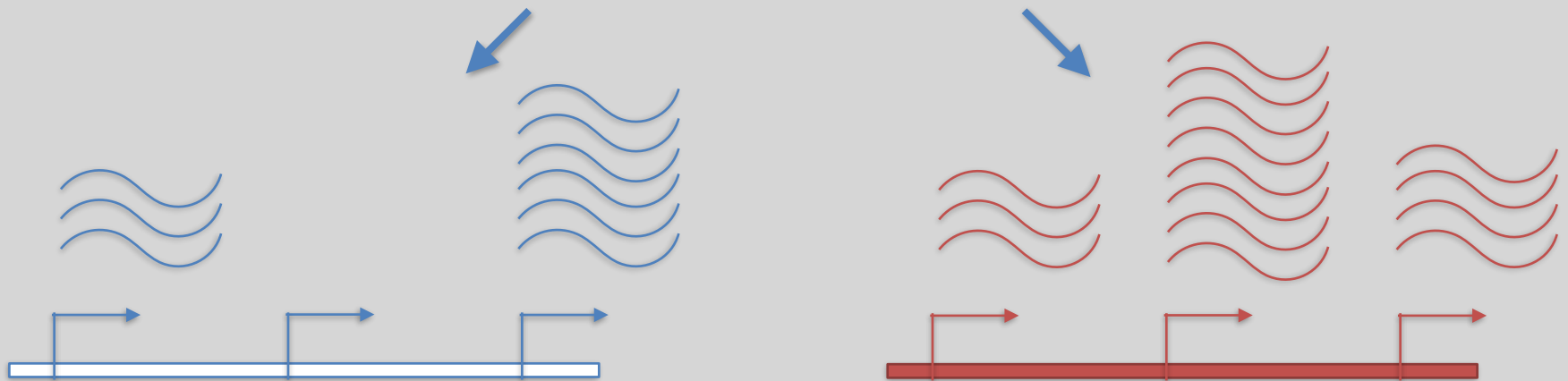
Normal Cells



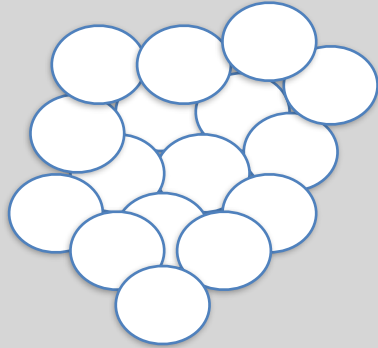
Mutated Cells



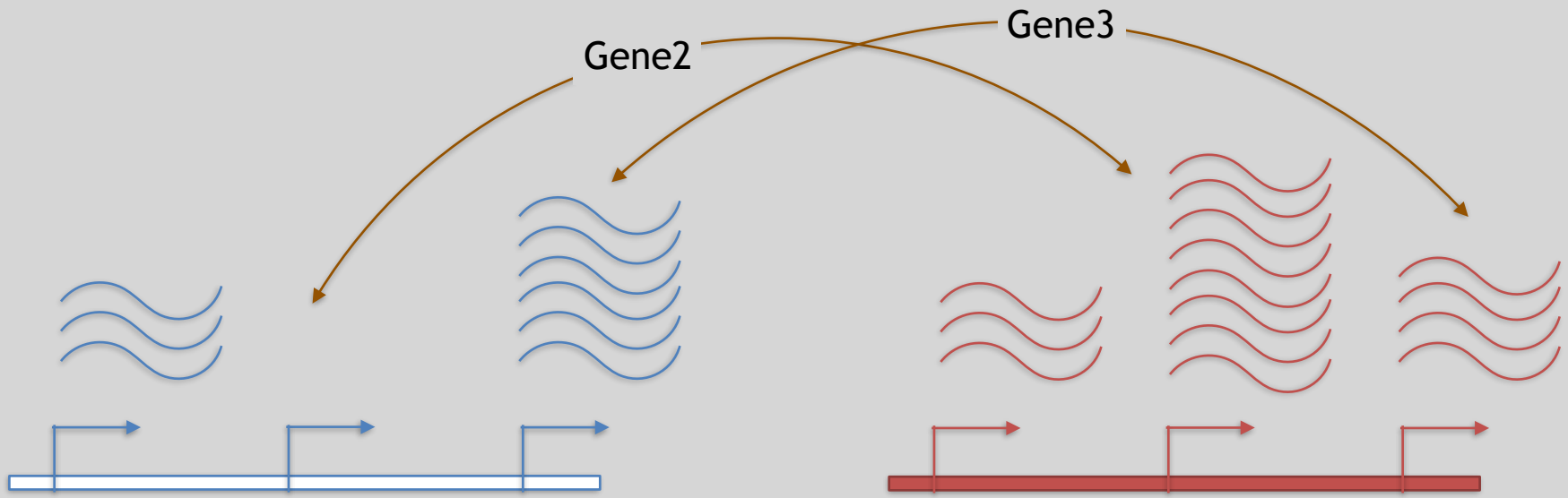
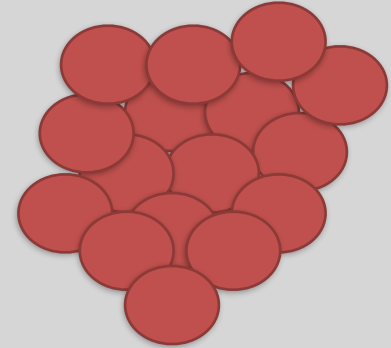
Then we can compare the two cell types to figure out what is different in the mutated cells!



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

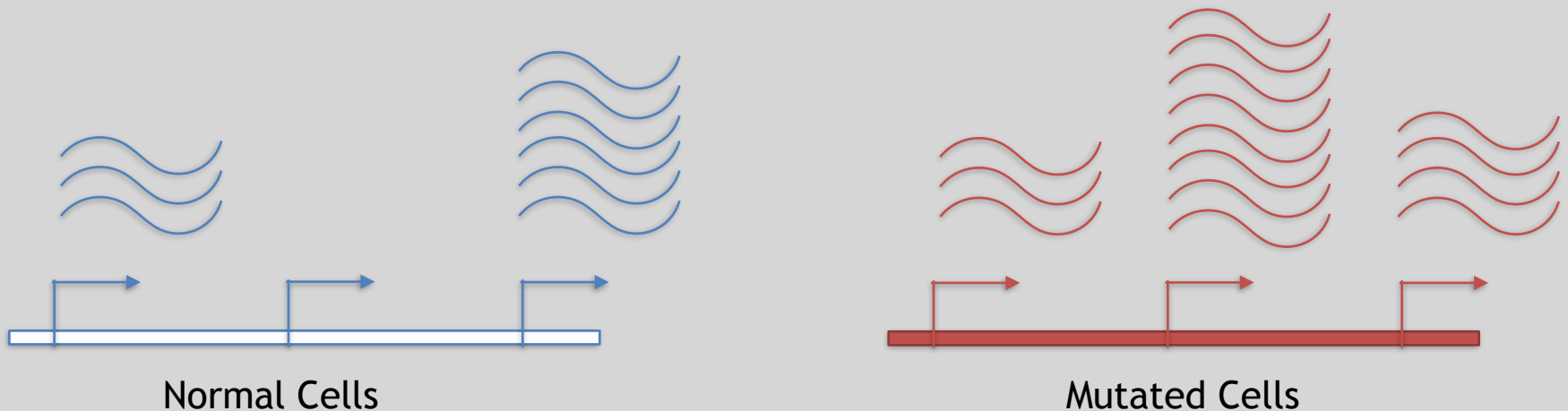
(Often the major bottleneck to overall success!)

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



TODAYS MENU:

▶ **What is a Genome?**

- Genome sequencing and the Human genome project

▶ **What can we do with a Genome?**

- Comparative genomics

▶ **Modern Genome Sequencing**

- 1st, 2nd and 3rd generation sequencing

▶ **Workflow for NGS**

- RNA-Sequencing and discovering variation

Do it Yourself!

Additional Reference Slides

(On FASTQ format, ASCII Encoded Base Qualities, FastQC, Alignment and SAM/BAM formats)

Hands-on worksheet:

https://bioboot.github.io/bimm143_W18/lectures/#13

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

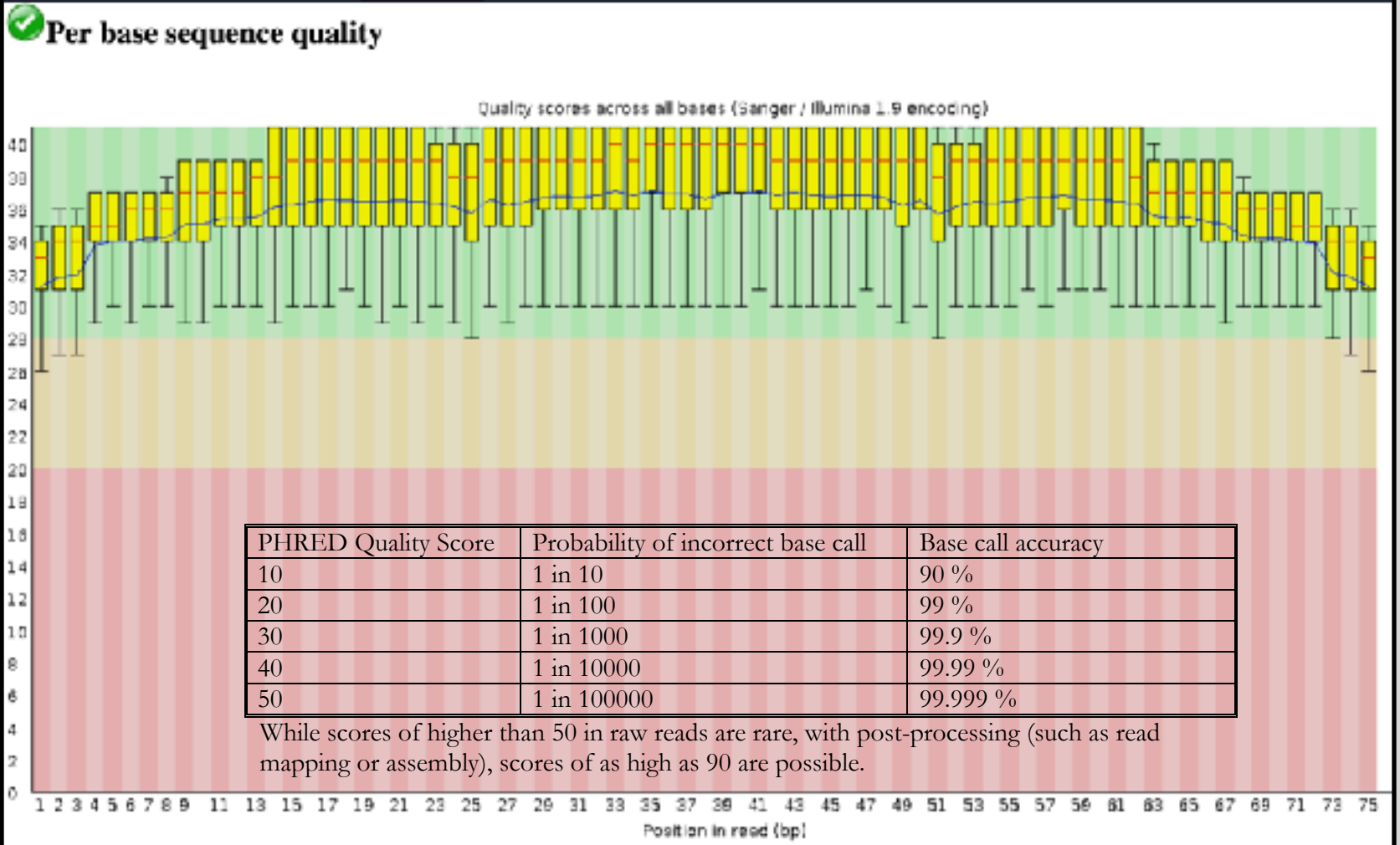
- Each sequence base has a corresponding numeric quality score encoded by a single ASCII character typically on the 4th line (see 4 above)
- ASCII characters represent integers between 0 and 127
- Printable ASCII characters range from 33 to 126
- Unfortunately there are 3 quality score formats that you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

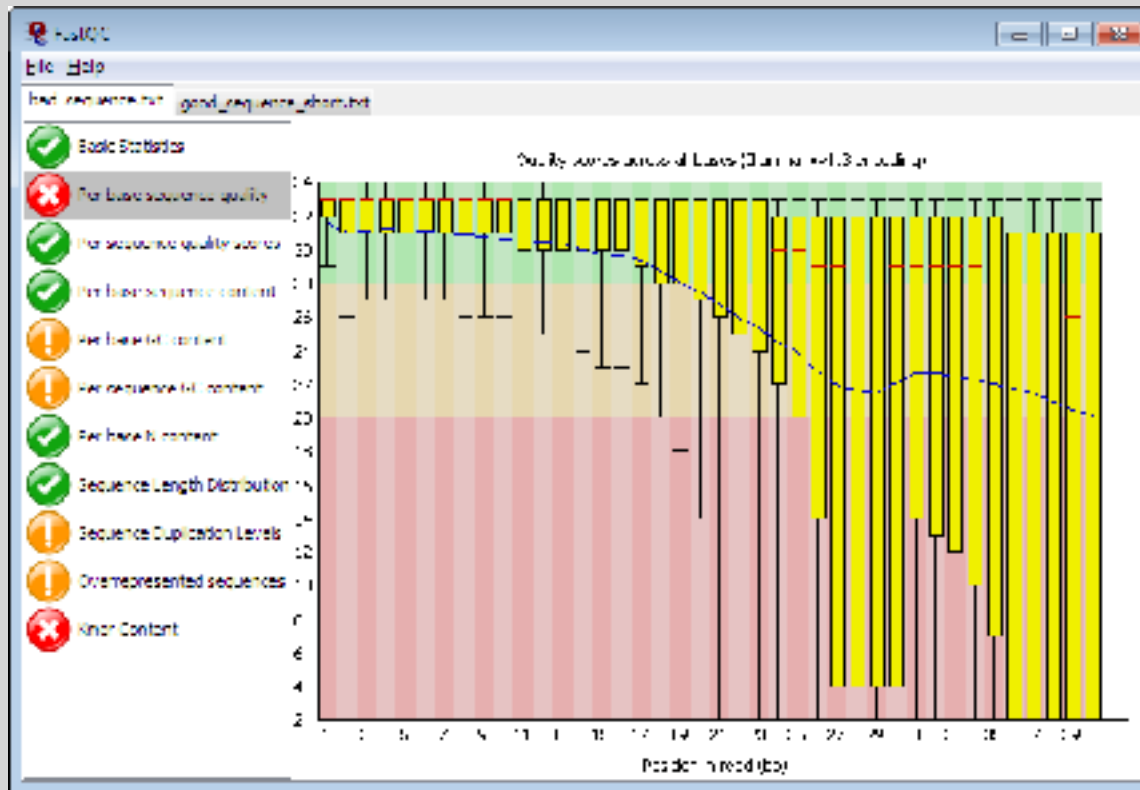
FastQC Report



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

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).
- Because SAM files are plain text (unlike their binary counterpart, BAM), we can take a peek at a few lines of the header with head, See:

https://bioboot.github.io/bggn213_f17/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)

Genome Analysis Toolkit (GATK)

- Developed in part to aid in the analysis of 1000 Genomes Project data
- Includes many tools for manipulating, filtering, and utilizing next generation sequence data
- <http://www.broadinstitute.org/gatk/>

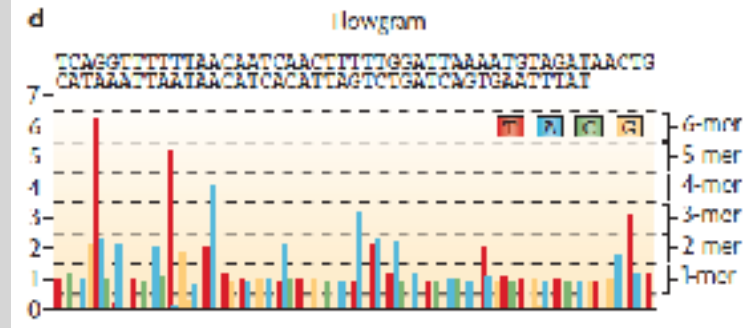
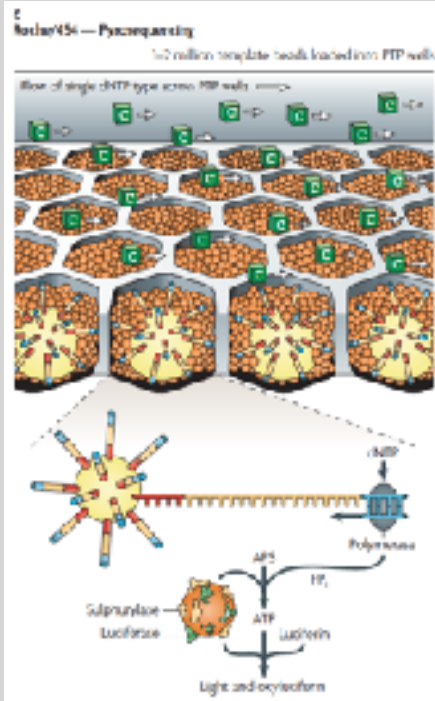
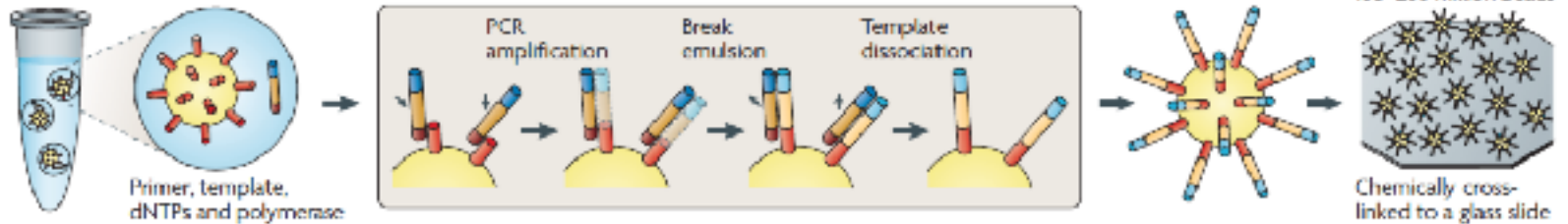
Do it Yourself!

Additional Reference Slides on Sequencing Methods

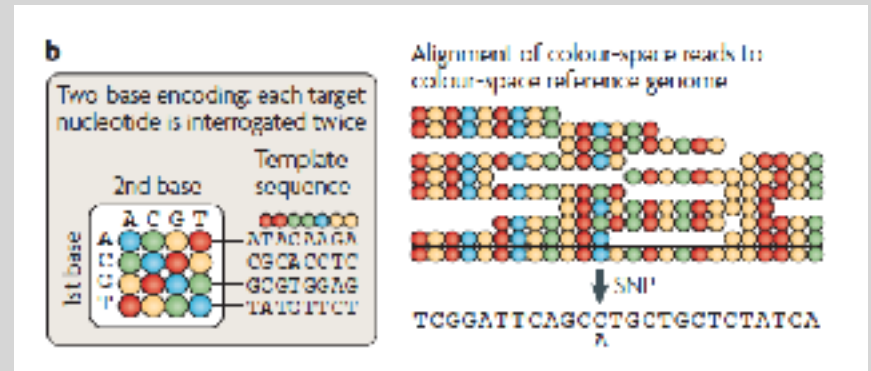
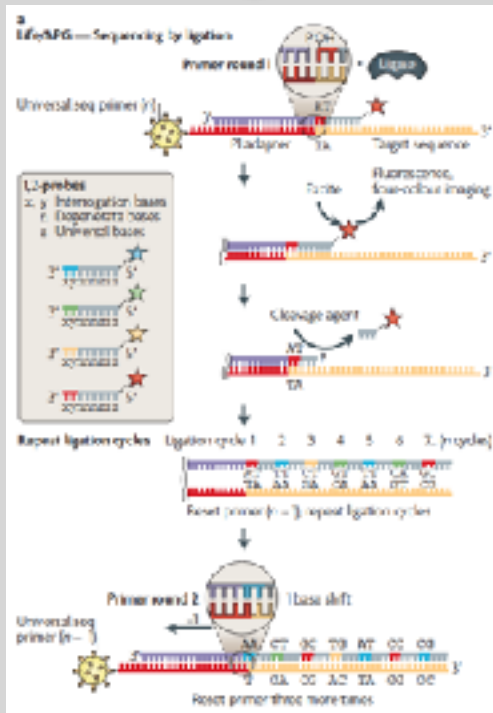
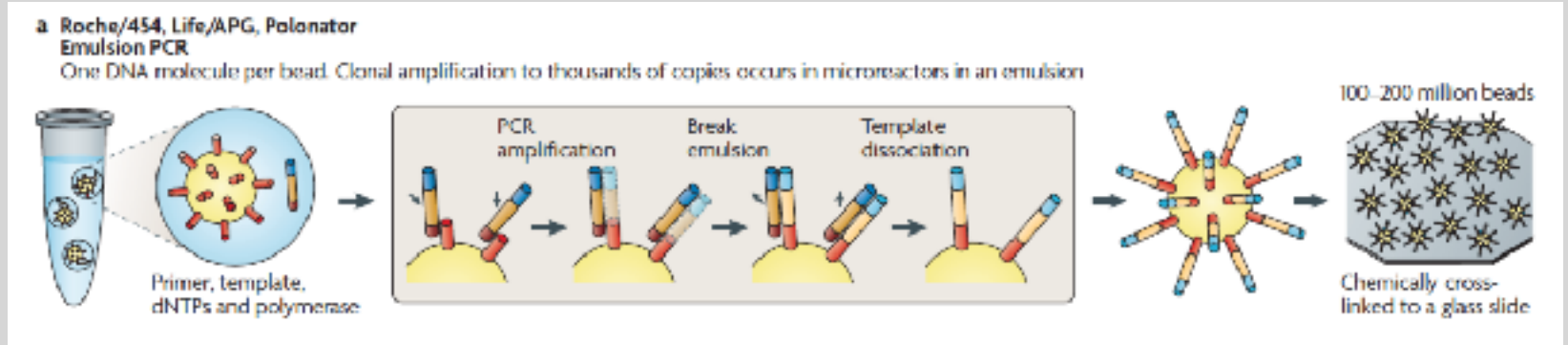
Roche 454 - Pyrosequencing

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

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

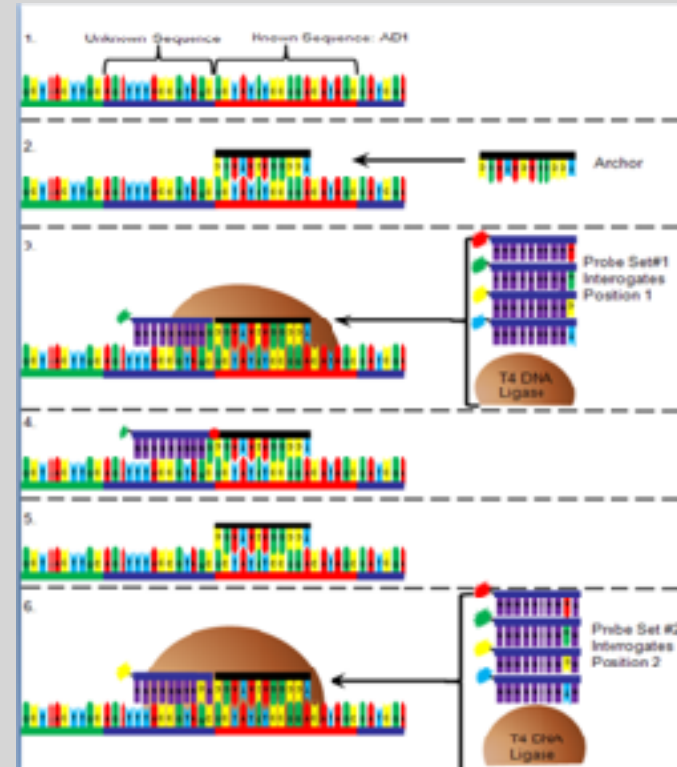
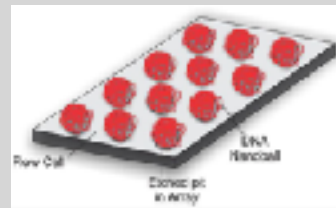
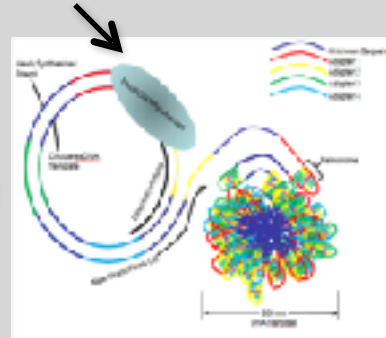
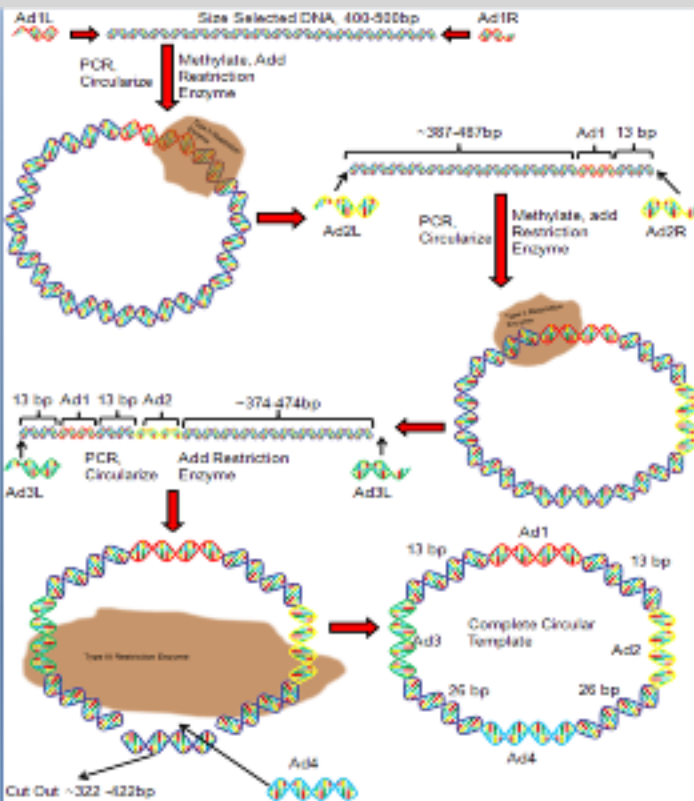


Life Technologies SOLiD - Sequence by Ligation



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

