



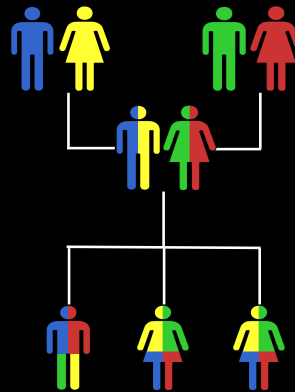
**BIMM 143**  
**Genome Informatics I**  
Lecture 13  
Barry Grant  
UC San Diego  
<http://thegrantlab.org/bimm143>

## 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



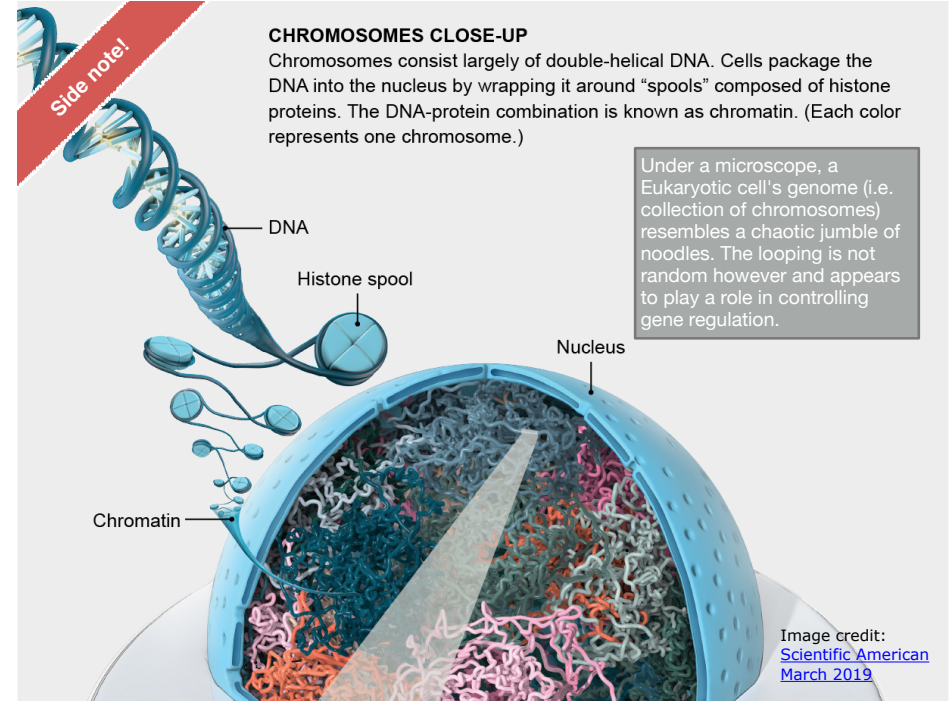
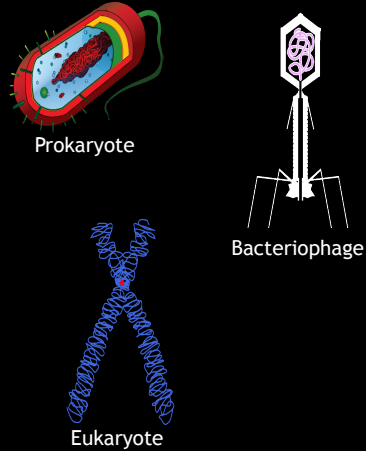
## Genetics and Genomics

- 
- **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!

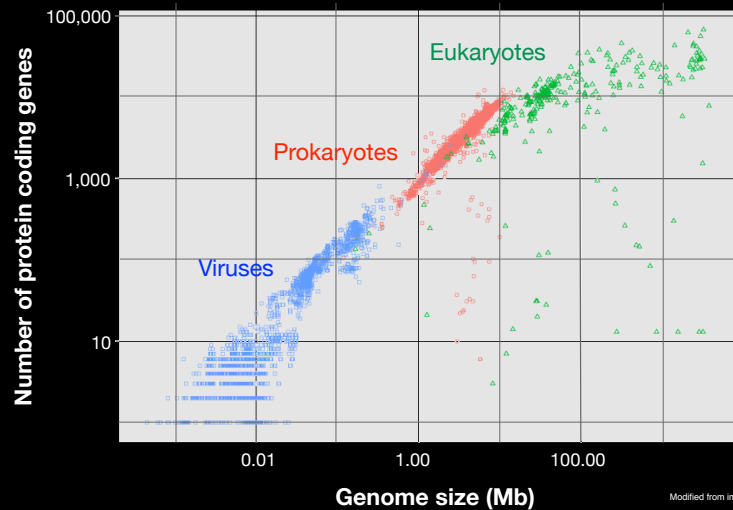
- Primarily DNA, but can be RNA in the case of some viruses
- Some genomes are circular, others linear
- Can be organized into discrete units (chromosomes) or freestanding molecules (plasmids)



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

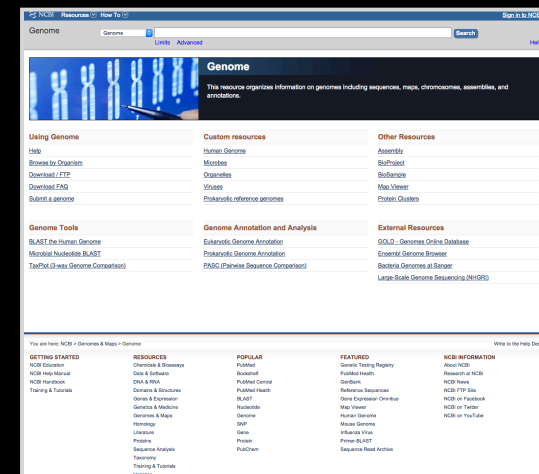


Modified from image by Estévez / CC BY-SA

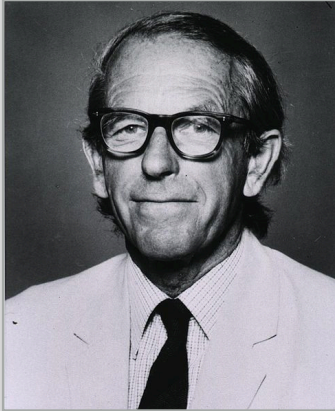
# Genome Databases

NCBI Genome:

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



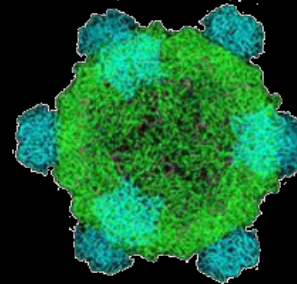
## Early Genome Sequencing



[http://en.wikipedia.org/wiki/Frederick\\_Sanger](http://en.wikipedia.org/wiki/Frederick_Sanger)

- 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

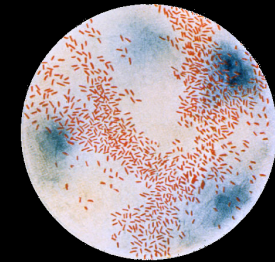
## The First Sequenced Genomes



**Bacteriophage  $\phi$ -X174**

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

[http://en.wikipedia.org/wiki/Phi\\_X\\_174](http://en.wikipedia.org/wiki/Phi_X_174)



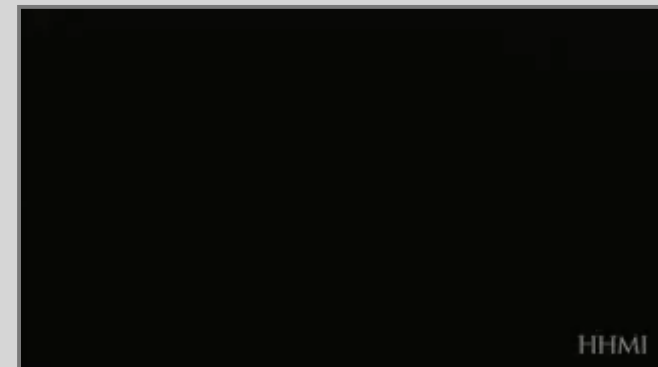
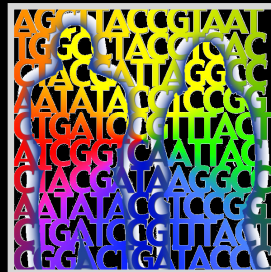
**Haemophilus influenzae**

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

<http://phil.cdc.gov/>

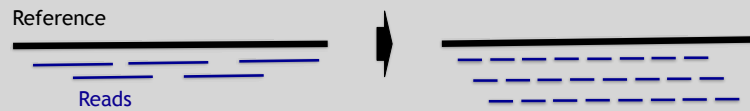
## The Human Genome Project

- 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
  - $\sim 3.2$  Billion base pairs, dsDNA
  - $\sim 20,000$  genes



## 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](https://en.wikipedia.org/wiki/Carlson_curve)

## Rapid progress of genome sequencing

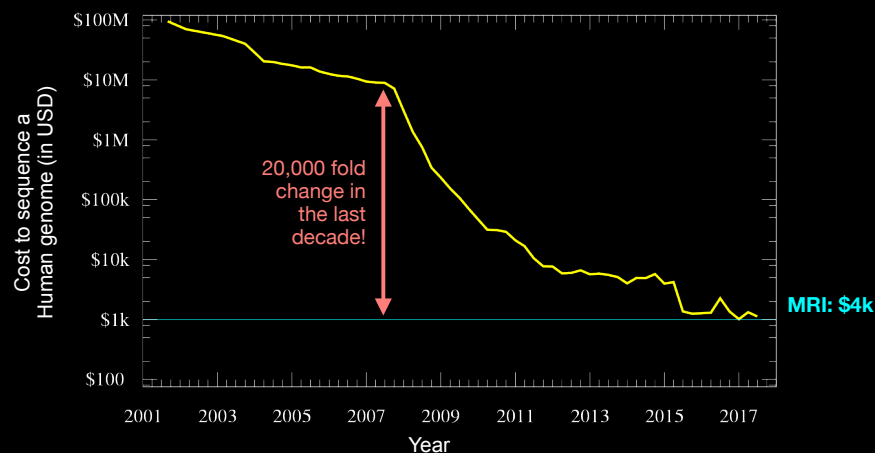


Image source: [https://en.wikipedia.org/wiki/Carlson\\_curve](https://en.wikipedia.org/wiki/Carlson_curve)

## 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
Protein expression	Protein arrays, mass spectrometry

WGS = whole genome sequencing, WXS = whole exome sequencing  
 RRBS = reduced representation bisulfite sequencing, WGBS = whole genome bisulfite sequencing

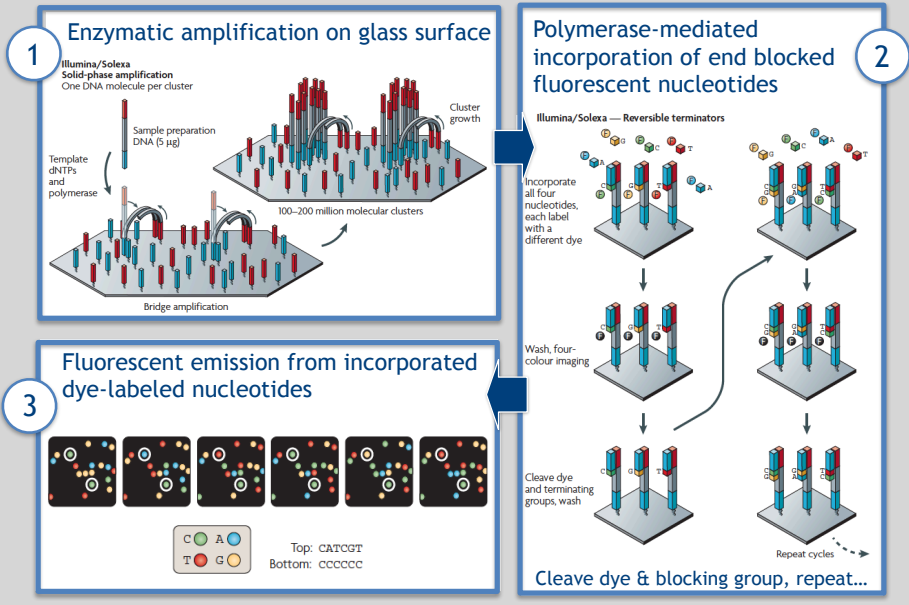
## DNA Sequencing Concepts

- **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

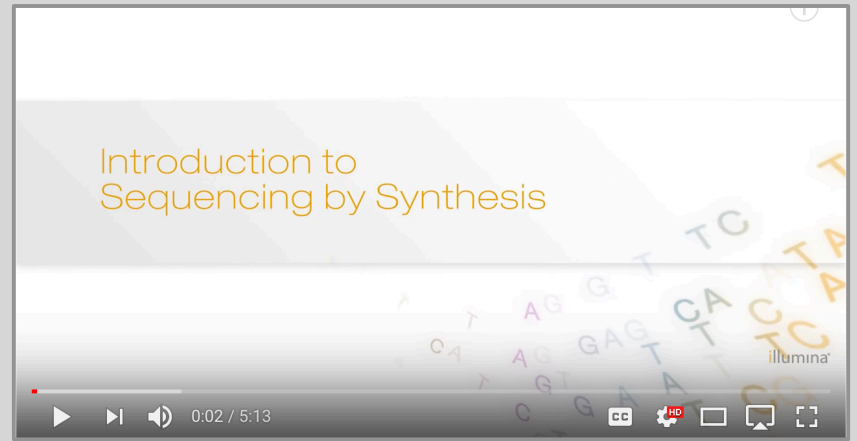
	Roche/454	Life Technologies SOLiD	Illumina Hi-Seq 2000
Library amplification method	emPCR* on bead surface	emPCR* on bead surface	Enzymatic amplification on glass surface
Sequencing method	Polymerase-mediated incorporation of unlabelled nucleotides	Ligase-mediated addition of 2-base encoded fluorescent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



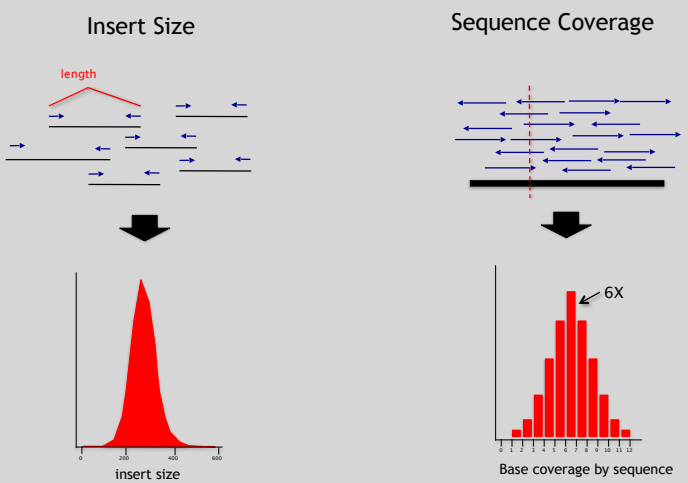
Images adapted from: Metzker, ML (2010), *Nat. Rev. Genet.*, 11, pp. 31–46

# Illumina Sequencing - Video



[https://www.youtube.com/watch?src\\_vid=womKfikWlxM&v=fCd6B5HRaZ8](https://www.youtube.com/watch?src_vid=womKfikWlxM&v=fCd6B5HRaZ8)

# NGS Sequencing Terminology



# Summary: “Generations” of DNA Sequencing

	First generation	Second generation*	Third generation*
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

Schadt, EE et al (2010), *Hum. Mol. Biol.*, 19(R12), pp. R227-R240

## Third Generation Sequencing

- Currently in active development
- Hard to define what “3<sup>rd</sup>” 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:

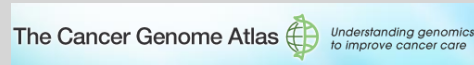
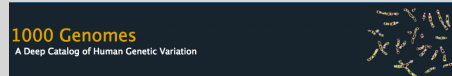
A good repository of analysis software can be found at <http://seqanswers.com/wiki/Software/list>

Name	Summary	Bio Tags	Meth Tags	Features	Language	Licence	OS
ApExts	Allows viewing sequencing trace files, multi-reading trimming, BLAST and reporting 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.	Hi-Res discovery Sequencing	Mapping		Perl	GPL	Linux 64
AB Small InDel Tool	The SQUID™ (Small InDel Tool) processes the raw evidence found in the pairing step of the SQUID™ System Analysis Pipeline Tool (Genome LITE).	Hi-Res discovery Sequencing	Alignment		Perl C++	GPL	Linux 64
ABISA	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.	Genomics Assembly	Assembly Scaffolding			Artistic License	Linux
ABMapper	Maps RNA-Seq reads to target genomes, 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
Arcturic Removal	Removes arcturic homopolymers from raw short read	De novo	Arcturic Removal	Traversal	Java	Custom license	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



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

“Variety’s the very spice of life”

-William Cowper, 1785

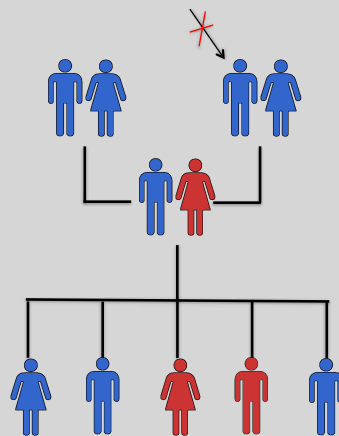
“Variation is the spice of life”

-Kruglyak & Nickerson, 2001

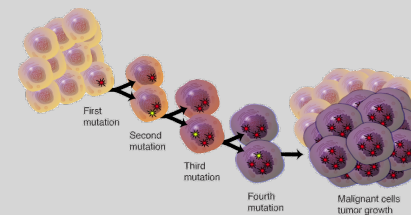
- While the sequencing of the human genome was a great milestone, the DNA from a single person is not representative of the millions of potential differences that can occur between individuals
- These unknown genetic variants could be the cause of many phenotypes such as differing morphology, susceptibility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

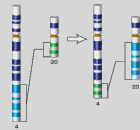
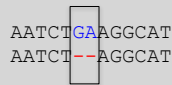


- Mutations in non-germline cells that are not passed along to offspring
- Can occur during mitosis or from the environment itself
- Are an integral part in tumor progression and evolution



# 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



Darryl Leja, Courtesy: National Human Genome Research Institute

# 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!

[ Numbers from: 1000 Genomes Project, Nature, 2012 ]

# Discovering Variation: SNPs and INDELs

SNP

```

ATCCTGATTCGGTGAACGTTATCGACGATCCGATCGA
ATCCTGATTCGGTGAACGTTATCGACGATCCGATCGA
CGGTGAACGTTATCGACGATCCGATCGAACTGTCAGC
GGTGAACGTTATCGACGTTCCGATCGAACTGTCAGCG
TGAACGTTATCGACGTTCCGATCGAACTGTCAGGCC
TGAACGTTATCGACGTTCCGATCGAACTGTCAGCGGC
TGAACGTTATCGACGTTCCGATCGAACTGTCAGCGGC
GTTATCGACGATCCGATCGAACTGTCAGCGGCAAGCT
TTATCGACGATCCGATCGAACTGTCAGCGGCAAGCT
ATCCTGATTCGGTGAACGTTATCGACGATCCGATCGAACTGTCAGCGGCAAGCTGATCGATCGATCGTAGTG
reference genome TTATCGACGATCCGATCGAACTGTCAGCGGCAAGCT
TCGACGATCCGATCGAACTGTCAGCGGCAAGCTGAT
ATCCGATCGAACTGTCAGCGGCAAGCTGATCGATCGAT
TCCGATCGAACTGTCAGCGGCAAGCTGATCGATCGATCGA
TCCGATCGAACTGTCAGCGGCAAGCTGATCGATCGA
GATCGAACTGTCAGCGGCAAGCTGATCGATCGATCGA
AACTGTCAGCGGCAAGCTGATCGATCGATCGATCGA
TGTGAGCGGCAAGCTGATCGATCGATCGATCGTAG
TCAGCGGCAAGCTGATCGATCGATCGATCGTAGTG
    
```

sequencing error or genetic variant?

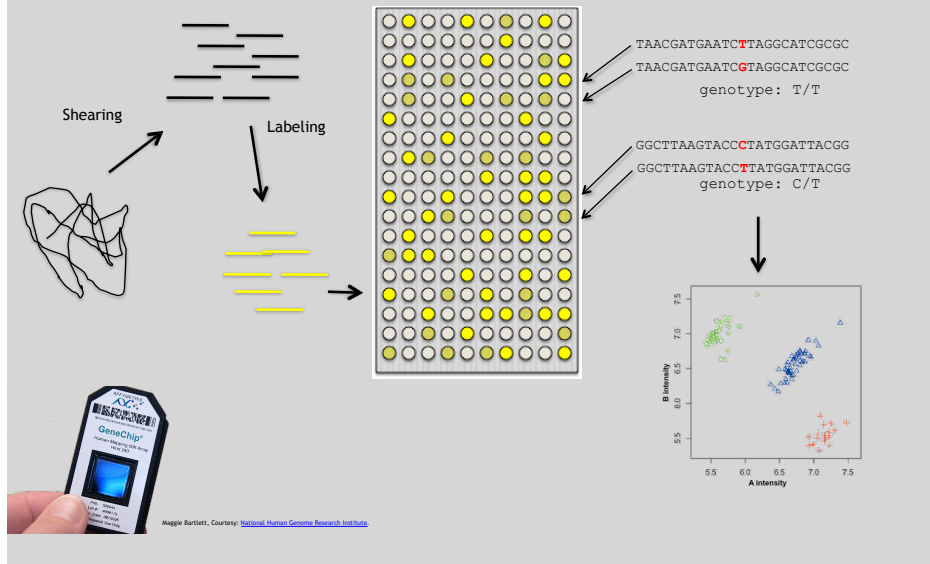
sequencing error or genetic variant?

INDEL

# Genotyping Small Variants

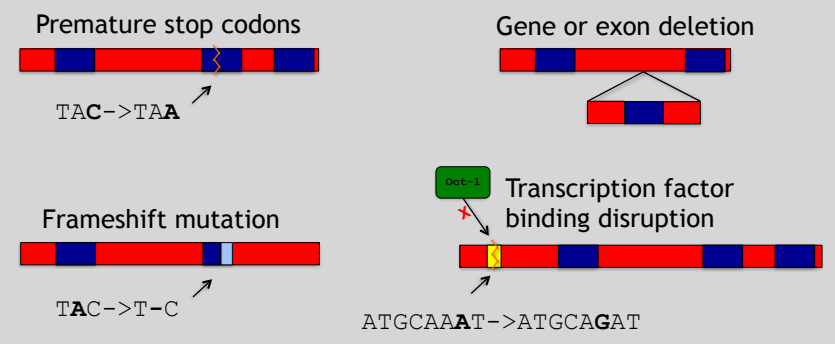
- Once discovered, oligonucleotide probes can be generated with each individual allele of a variant of interest
- A large number can then be assessed simultaneously on microarrays to detect which combination of alleles is present in a sample

## SNP Microarrays



## Impact of Genetic Variation

There are numerous ways genetic variation can exhibit functional effects



Do it Yourself!

## Hand-on time!

[https://bioboot.github.io/bimm143\\_S19/lectures/#14](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>

Region of interest

Haplotypes and patches

Gene or region of interest

Transcripts (splice variants)

Chromosome image

Overview image

Zoomable Region Image

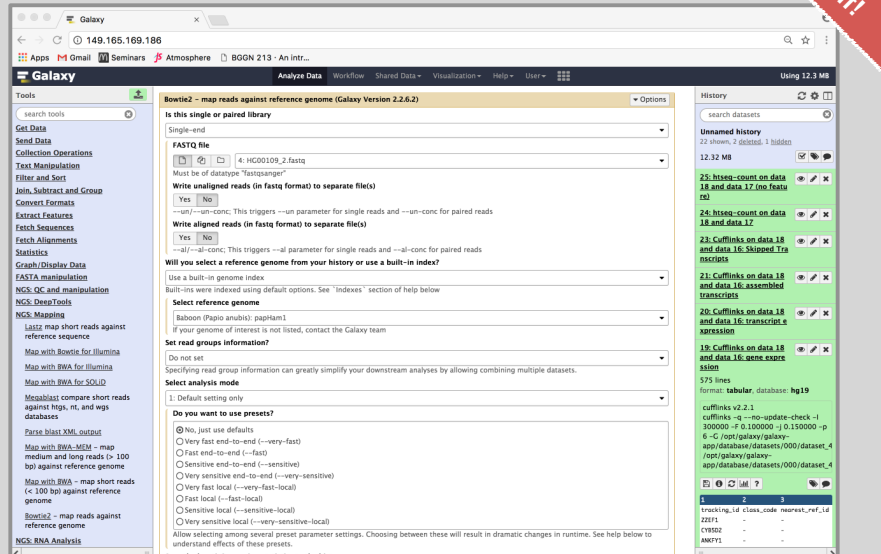
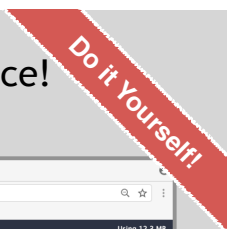
Change or add data tracks

Change or add data tracks

Change or add data tracks

## Access a jetstream galaxy instance!

Use assigned IP address



## Raw data usually in FASTQ format

```
@NS500177:196:HFTTTFAXX:1:11101:10916:1458 2:N:0:CGCGGCTG
ACACGACGATGAGGTGACAGTCACGGAGGATAAGATCAATGCCCTCATTAAGCAGCCGGTGTAA
+
AAAAAAAAEEEEEEEE//AEEEEEEEEEEEEEE/EE/<<EE/AEEEE//EEEEEEEEEA<
```

- 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:HFTTTFAXX:1:11101:10916:1458 2:N:0:CGCGGCTG
ACACGACGATGAGGTGACAGTCACGGAGGATAAGATCAATGCCCTCATTAAGCAGCCGGTGTAA
+
AAAAAAAAEEEEEEEE//AEEEEEEEEEEEEEE/EE/<<EE/AEEEE//EEEEEEEEEA<
```

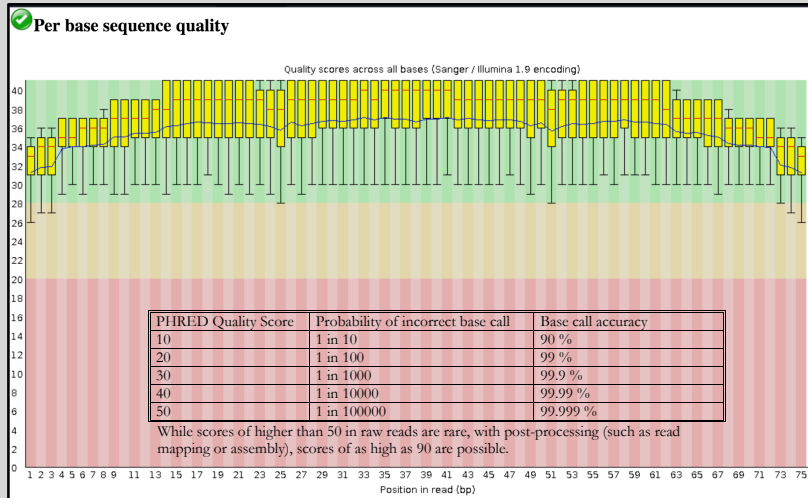
- Each sequence base has a corresponding numeric quality score encoded by a single ASCII character typically on the 4th line (see 4 above)
- ASCII characters represent integers between 0 and 127
- Printable ASCII characters range from 33 to 126
- Unfortunately there are 3 quality score formats that you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

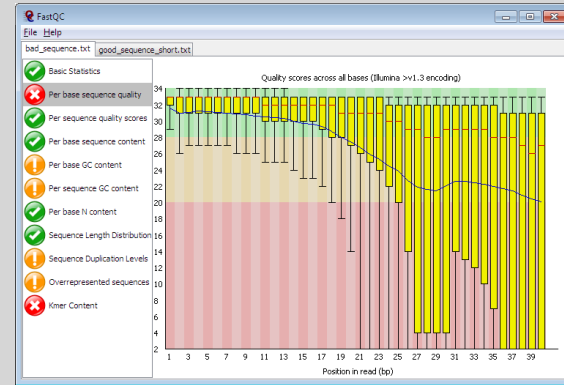
# FastQC Report



# 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

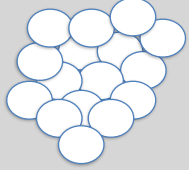
- Once sequence quality has been assessed, the next step is to align the sequence to a reference genome
- There are *many* distinct tools for doing this; which one you choose is often a reflection of your specific experiment and personal preference

BWA	BarraCUDA	RMAP
Bowtie	CASHx	SSAHA
SOAP2	GSNAP	etc
Novoalign	Mosiak	
mr/mrsFast	Stampy	
Eland	SHRiMP	
Blat	SeqMap	
Bfast	SLIDER	

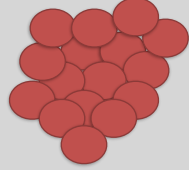
# RNA Sequencing

The absolute basics

Normal Cells

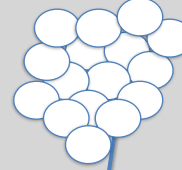


Mutated Cells

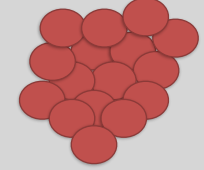


- The **mutated cells** behave differently than the **normal cells**
- We want to know what genetic mechanism is causing the difference
- One way to address this is to examine differences in gene expression via RNA sequencing...

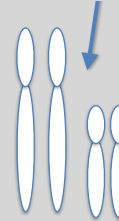
Normal Cells



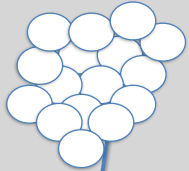
Mutated Cells



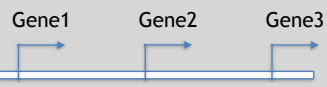
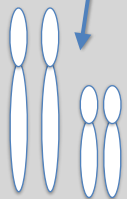
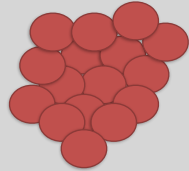
Each cell has a bunch of chromosomes



Normal Cells

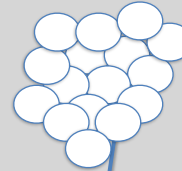


Mutated Cells

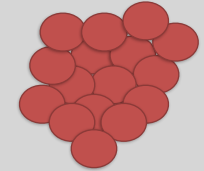


Each chromosome has a bunch of genes

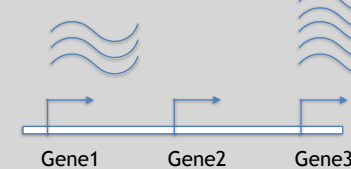
Normal Cells



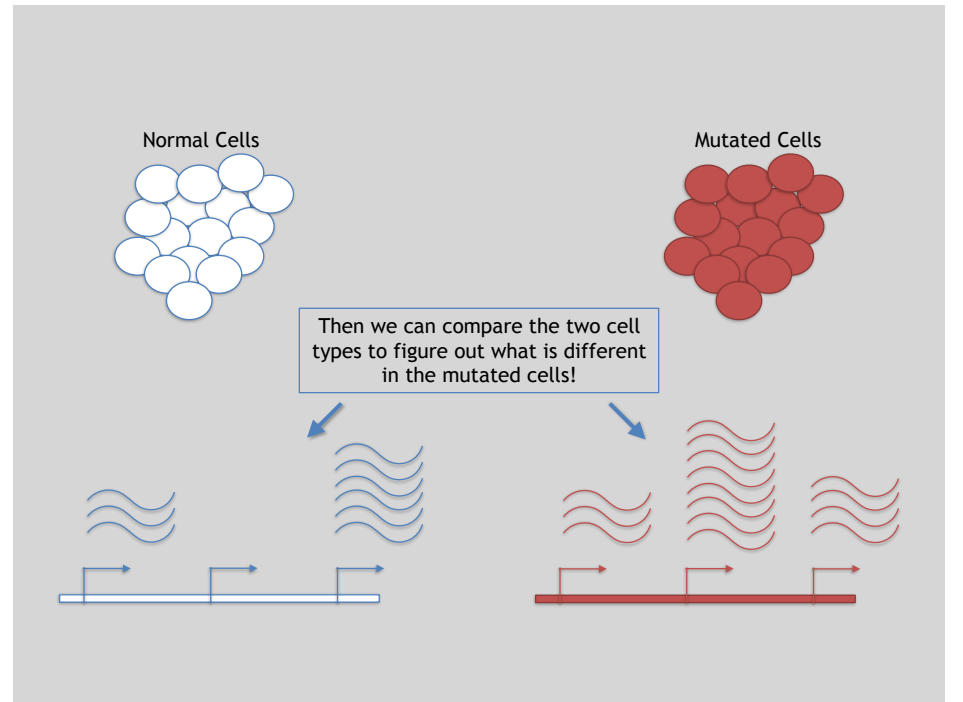
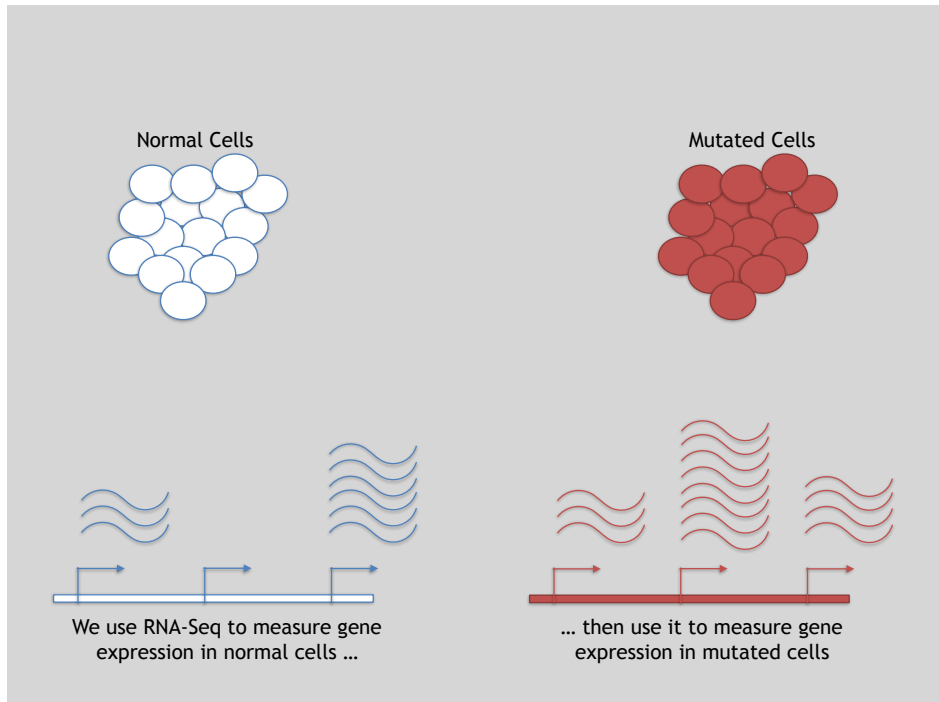
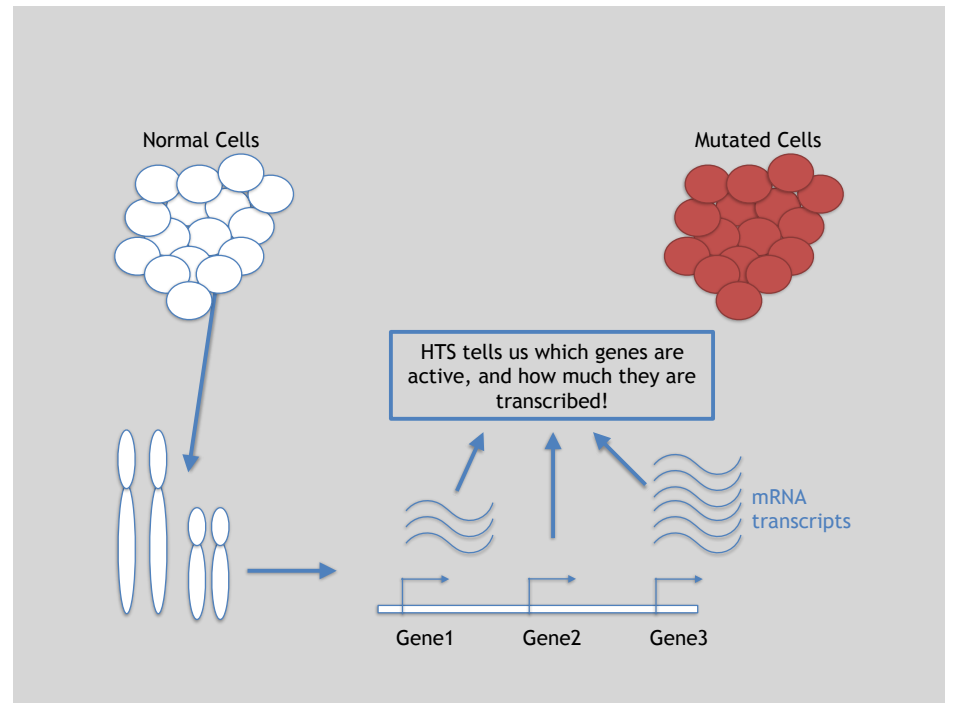
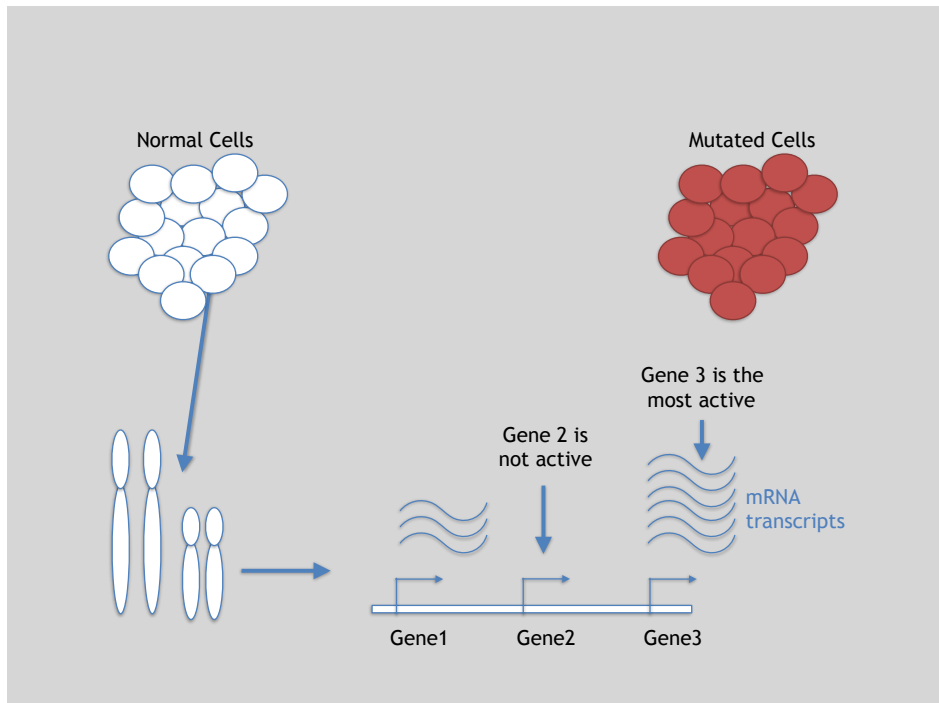
Mutated Cells

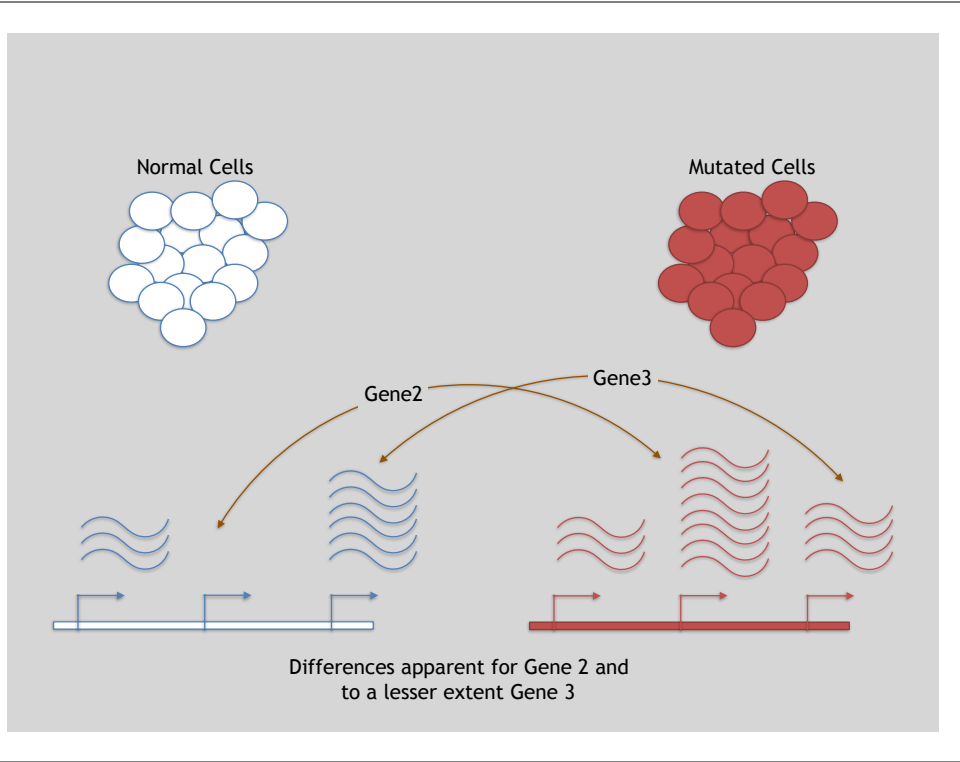


Some genes are active more than others



mRNA transcripts





- ### 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

Normal Cells

Mutated Cells

*Do it Yourself!*

## Hand-on time!

[https://bioboot.github.io/bimm143\\_S19/lectures/#14](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

- 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

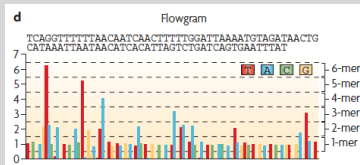
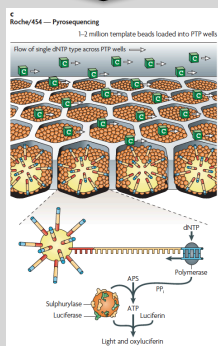
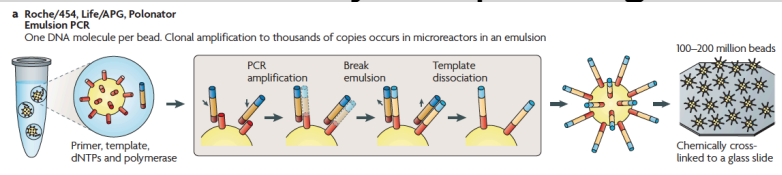
## SAM Format

- 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>



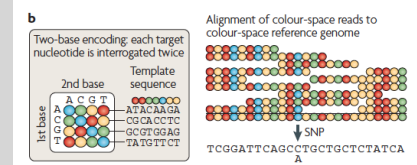
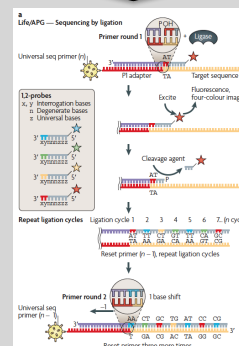
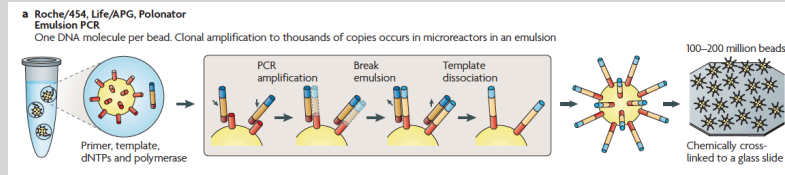


# Roche 454 - Pyrosequencing



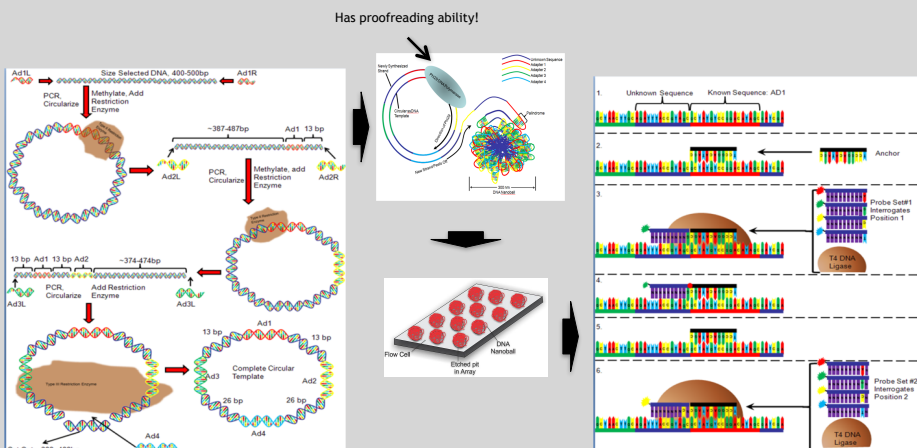
Metzker, ML (2010), *Nat. Rev. Genet.*, 11, pp. 31-46

# Life Technologies SOLiD - Sequence by Ligation



Metzker, ML (2010), *Nat. Rev. Genet.*, 11, pp. 31-46

# Complete Genomics - Nanoball Sequencing



Nedringhaus, TP et al (2011), *Analytical Chem.*, 83, pp. 4327-4341

Wikipedia, "DNA Nanoball Sequencing", September 26, 2012

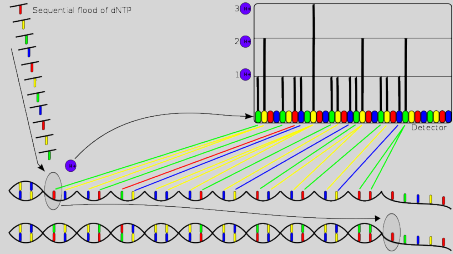
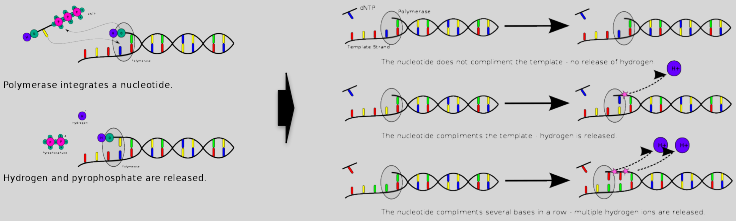
# "Benchtop" Sequencers

- 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 <sup>a,b</sup>	\$225 <sup>c</sup>	10 Mb (100 bases)	3 h	\$22.5	3.3
(316 chip)		\$425	100 Mb <sup>d</sup> (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 x 150 bases)	27 h	\$0.5	55.5

Loman, NJ (2012), *Nat. Biotech.*, 5, pp. 434-439

# PGM - Ion Semiconductor Sequencing



Wikipedia, "Ion Semiconductor Sequencing", September 26, 2012