BGGN 213

Genome Informatics I

Lecture 13

Barry Grant

UC San Diego

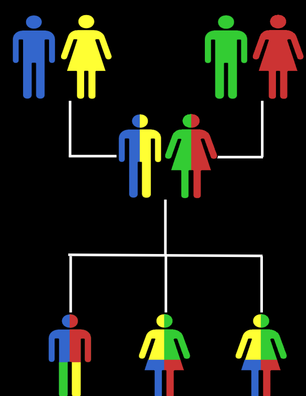
http://thegrantlab.org/bggn213

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

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



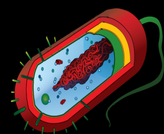
Genetics and Genomics

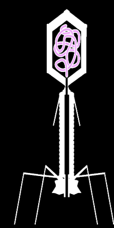
Side note!

- **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

- 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

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.

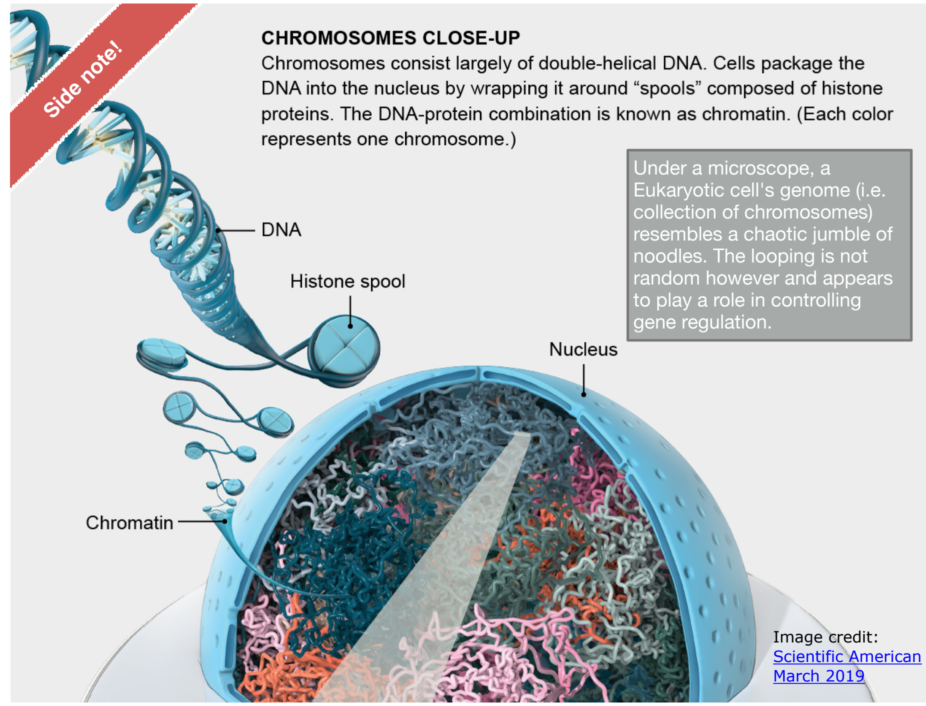
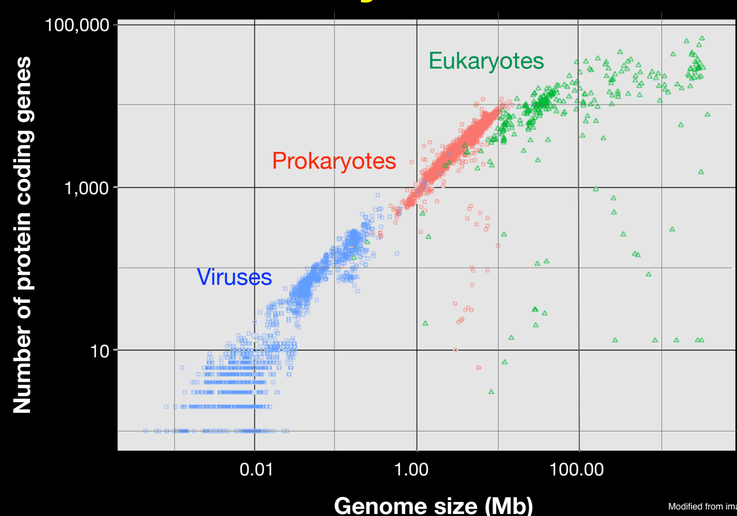


Image credit: [Scientific American March 2019]

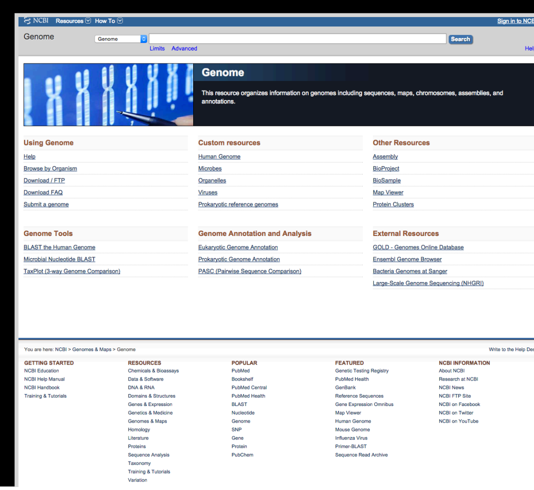
Genomes come in many sizes



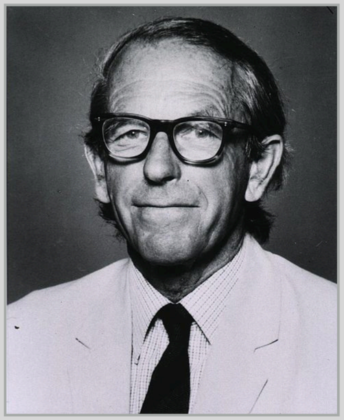
Modified from image by Estevezj / CC BY-SA

Genome Databases

NCBI Genome:
http://www.ncbi.nlm.nih.gov/genome



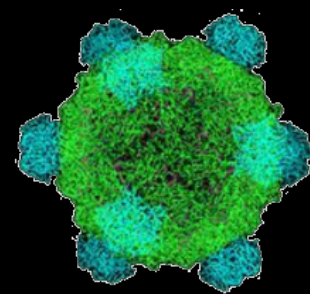
Early Genome Sequencing



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 ~1 / 10000 bases

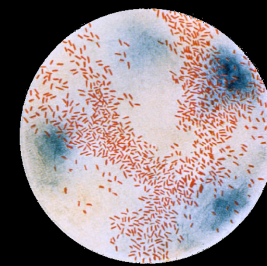
The First Sequenced Genomes



Bacteriophage φ-X174

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

http://en.wikipedia.org/wiki/Phi_X_174



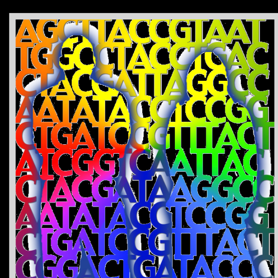
Haemophilus influenzae

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

http://phil.cdc.gov/

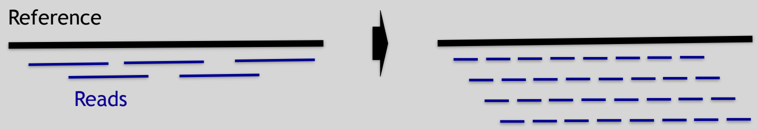
The Human Genome Project

- 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
  - ~3.2 Billion base pairs, dsDNA
  - ~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

Rapid progress of genome sequencing

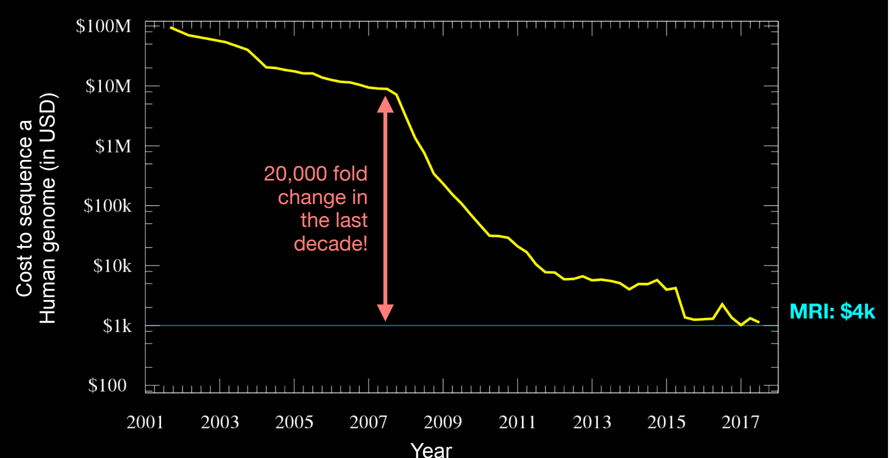


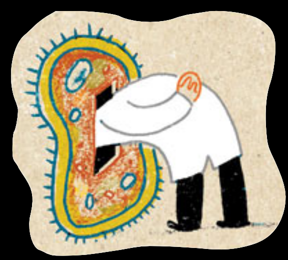
Image source: https://en.wikipedia.org/wiki/Carlson_curve

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

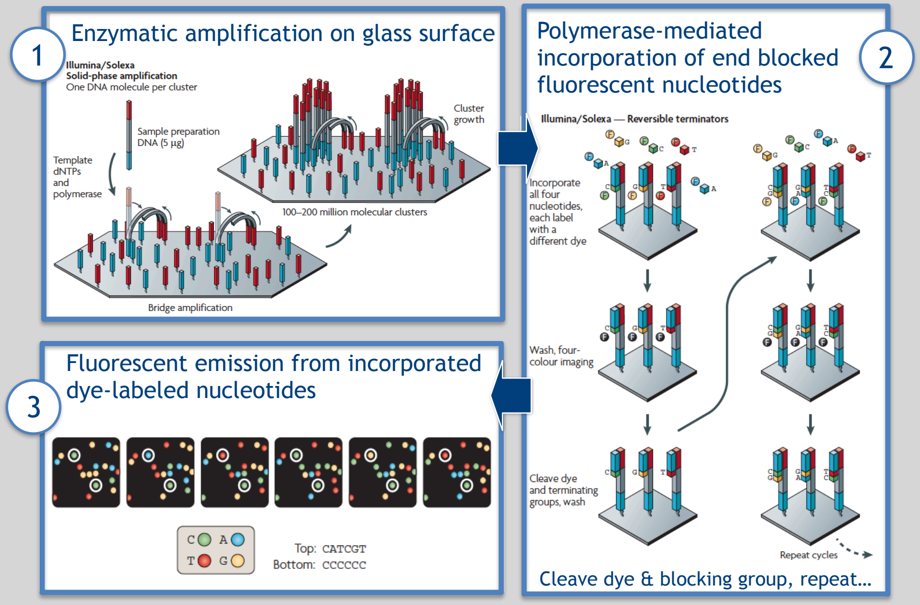
- **Sequencing by Synthesis:** Uses a polymerase to incorporate and assess nucleotides to a primer sequence
  - 1 nucleotide at a time
- **Sequencing by Ligation:** Uses a ligase to attach hybridized sequences to a primer sequence
  - 1 or more nucleotides at a time (e.g. dibase)

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

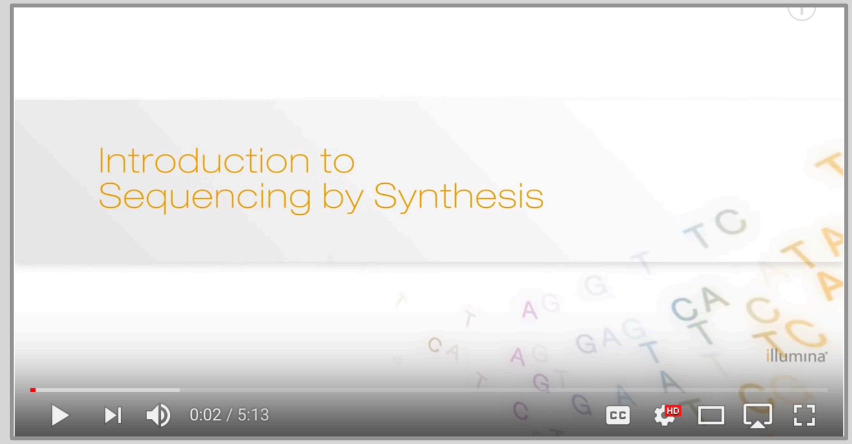
Modified from Mardis, ER (2011), Nature, 470, pp. 198-203

Illumina – Reversible terminators



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

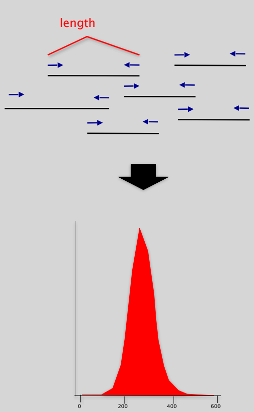
Illumina Sequencing - Video



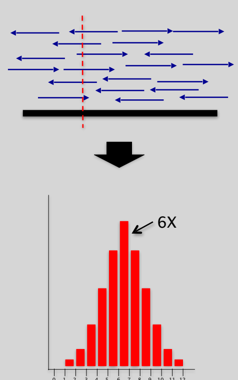
https://www.youtube.com/watch?src_vid=womKfikWlxM&v=fCd6B5HRaZ8

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

Schadt, EE et al (2010), *Hum. Mol. Biol.*, 19(RI2), pp. R227-R240

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

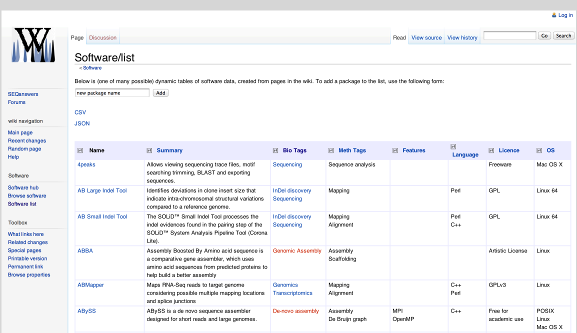
Side-Note:

- 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 & BioStars

Side-Note:

A good repository of analysis software can be found at http://seqanswers.com and https://www.biostars.org/



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

GENOME 10K
Unveiling animal diversity

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

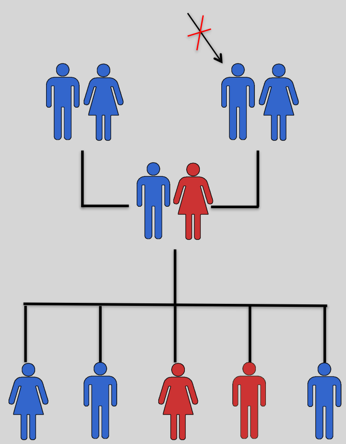
"Variation is the spice of life"

–Kruglyak & Nickerson, 2001

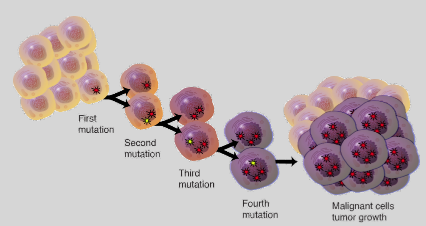
- While the sequencing of the human genome was a great milestone, the DNA from a single person is not representative of the millions of potential differences that can occur between individuals
- These unknown genetic variants could be the cause of many phenotypes such as differing morphology, susceptibility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

Darryl Leja, Courtesy: National Human Genome Research Institute.

Types of Genomic Variation

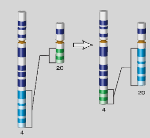
- **Single Nucleotide Polymorphisms** (SNPs) - mutations of one nucleotide to another

```
AATCTGAGGCAT
AATCTCAGGCAT
```

- **Insertion/Deletion Polymorphisms** (INDELs) - small mutations removing or adding one or more nucleotides at a particular locus

```
AATCTGAAGGCAT
AATCT--AGGCAT
```

- **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

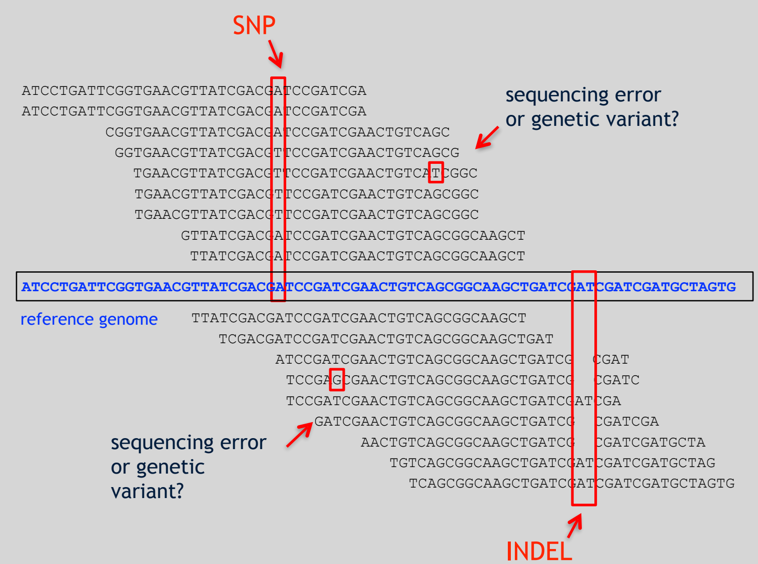
Numbers change depending on ancestry!

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

Discovering Variation: SNPs and INDELs

- Small variants require the use of sequence data to initially be discovered
- Most approaches align sequences to a reference genome to identify differing positions
- The amount of DNA sequenced is proportional to the number of times a region is covered by a sequence read
  - More sequence coverage equates to more support for a candidate variant site

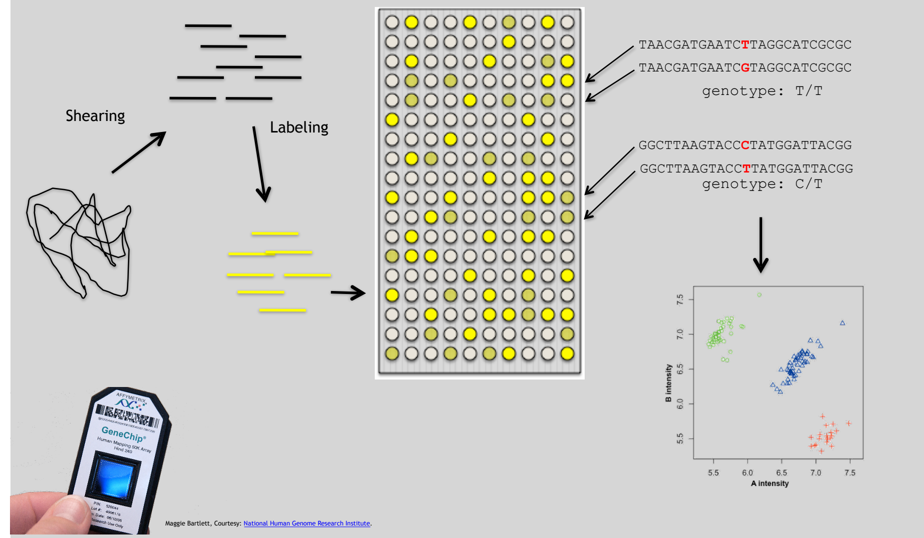
Discovering Variation: SNPs and INDELs



Genotyping Small Variants

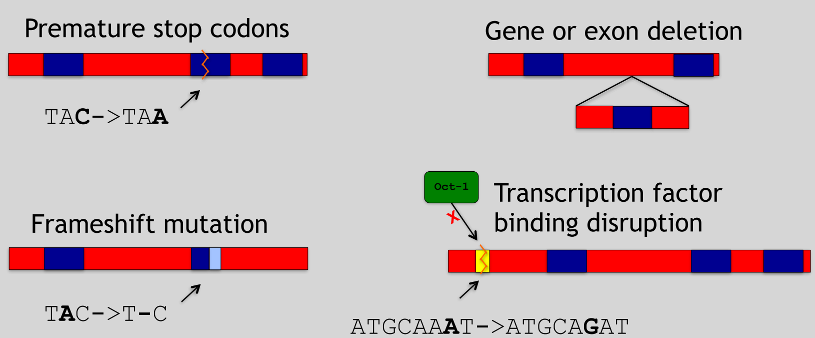
- Once discovered, oligonucleotide probes can be generated with each individual allele of a variant of interest
- A large number can then be assessed simultaneously on microarrays to detect which combination of alleles is present in a sample

SNP Microarrays



Impact of Genetic Variation

There are numerous ways genetic variation can exhibit functional effects



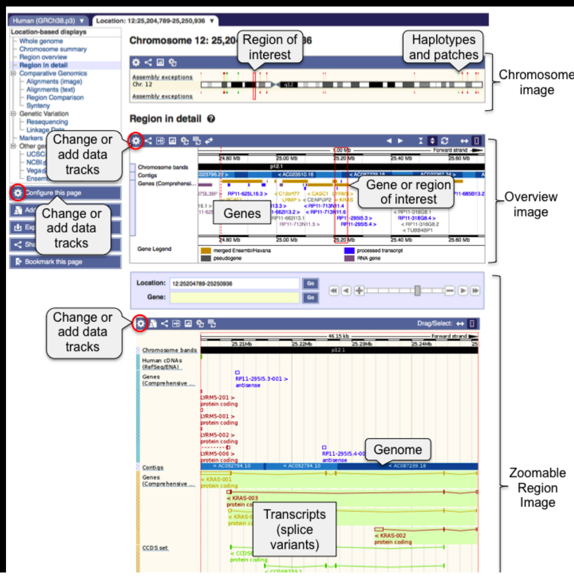
Do it Yourself!

Hand-on time!

https://bioboot.github.io/bggn213_S19/lectures/#13

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

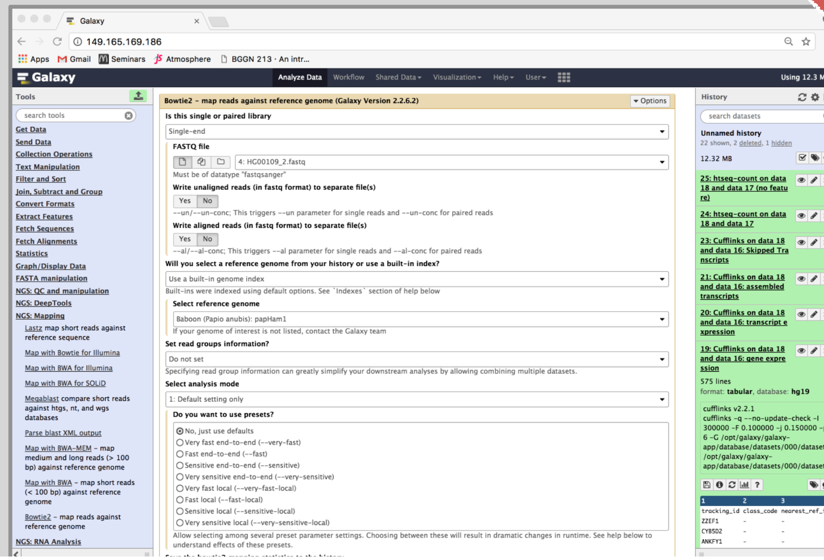
http://uswest.ensembl.org/Help/View?id=140



Access a jetstream galaxy instance!

Do it Yourself!

Use assigned IP address



Raw data usually in FASTQ format

```
@NS500177:196:HFTTTAFXX:1:11101:10916:1458 2:N:0:CGCGGCTG   (1)
ACACGACGATGAGGTGACAGTCACGGAGGATAAGATCAATGCCCTCATTAAAGCAGCCGGTGTAA   (2)
+   (3)
AAAAAEEEEEEEEEEE//AEEEAEEEEEEEEEEE/EE/<<EE/AAEEAEE///EEEEAEEEAEA<   (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 forth 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
+
AAAAAEEEEEEEEEEE//AEEEAEEEEEEEEEEE/EE/<<EE/AAEEAEE///EEEEAEEEAEA<   (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...

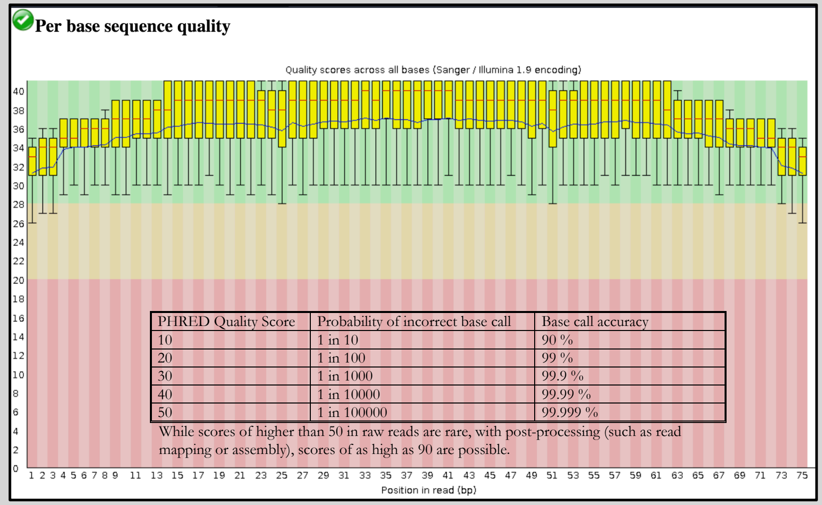
Interpreting Base Qualities in R

		ASCII Range	Offset	Score Range
Sanger, Illumina (Ver > 1.8)	**fastqsanger**	**33-126**	**33**	**0-93**
Solexa, Ilumina (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("DDDDCDEDCDDDDBBDDDCC@") ) - 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)
```

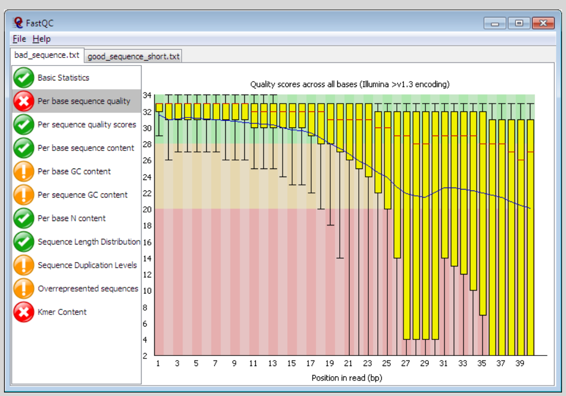
FastQC Report



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

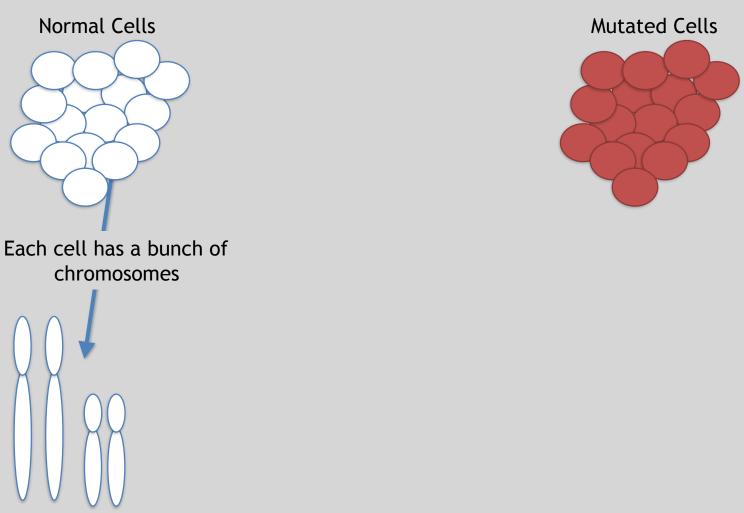
- Once sequence quality has been assessed, the next step is to align the sequence to a reference genome
- There are *many* distinct tools for doing this; which one you choose is often a reflection of your specific experiment and personal preference

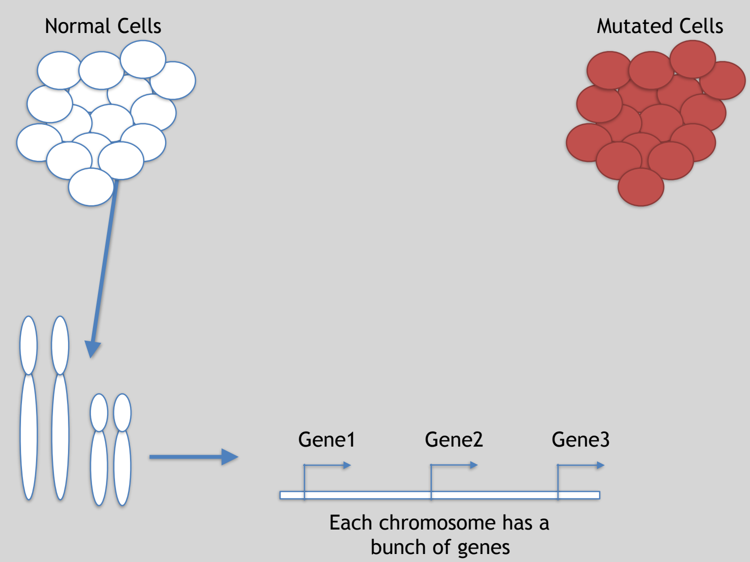
BWA	BarraCUDA	RMAP
Bowtie	CASHx	SSAHA
SOAP2	GSNAP	etc
Novoalign	Mosiak	
mr/mrsFast	Stampy	
Eland	SHRiMP	
Blat	SeqMap	
Bfast	SLIDER	

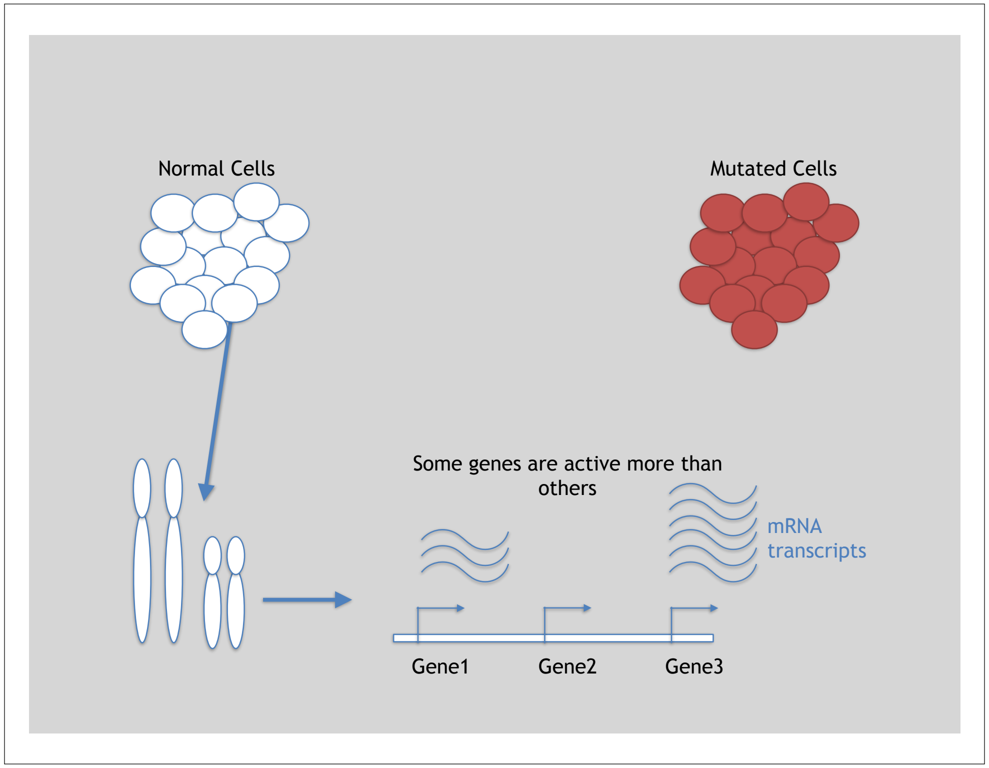
RNA Sequencing

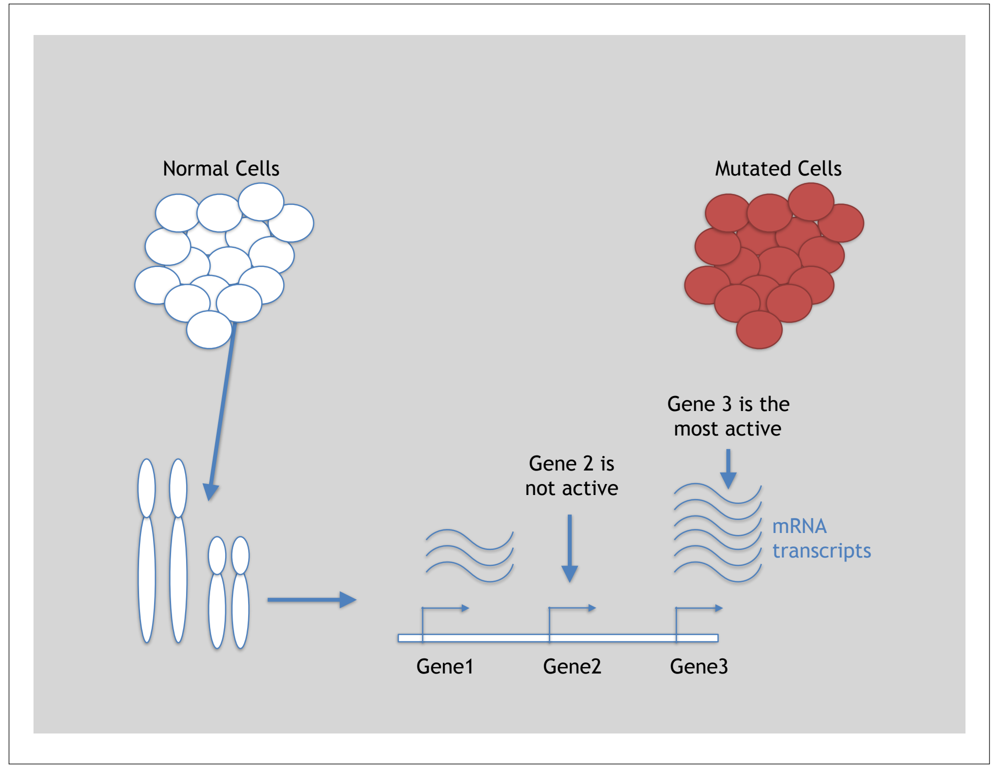
The absolute basics

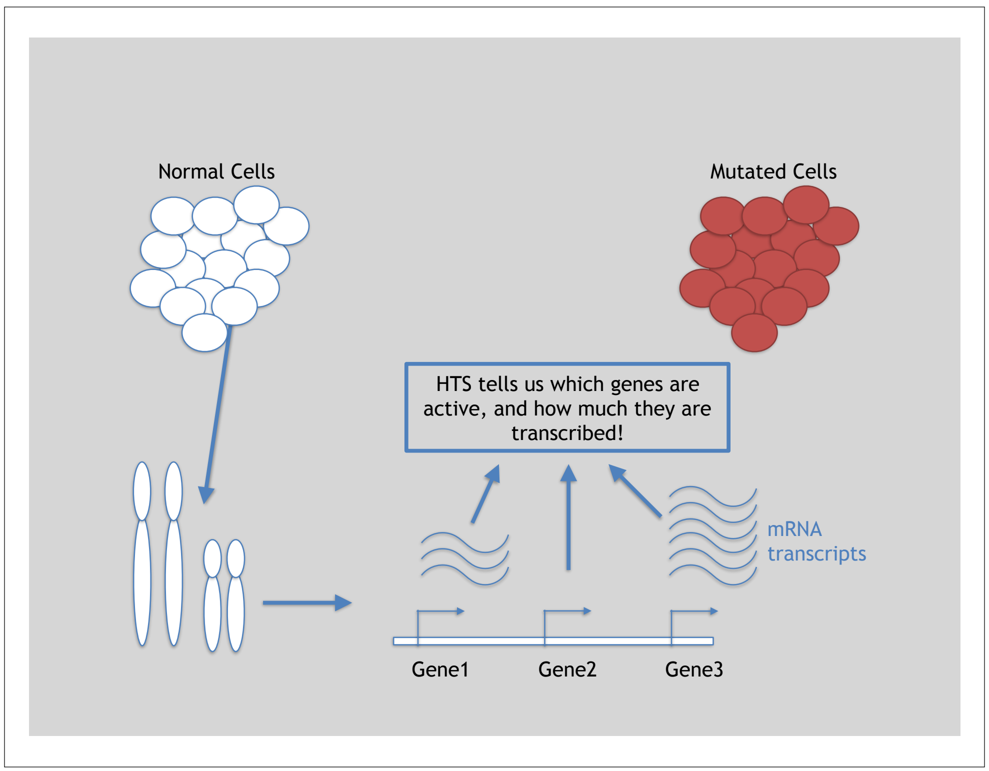
- The mutated cells behave differently than the normal cells
- We want to know what genetic mechanism is causing the difference
- One way to address this is to examine differences in gene expression via RNA sequencing…

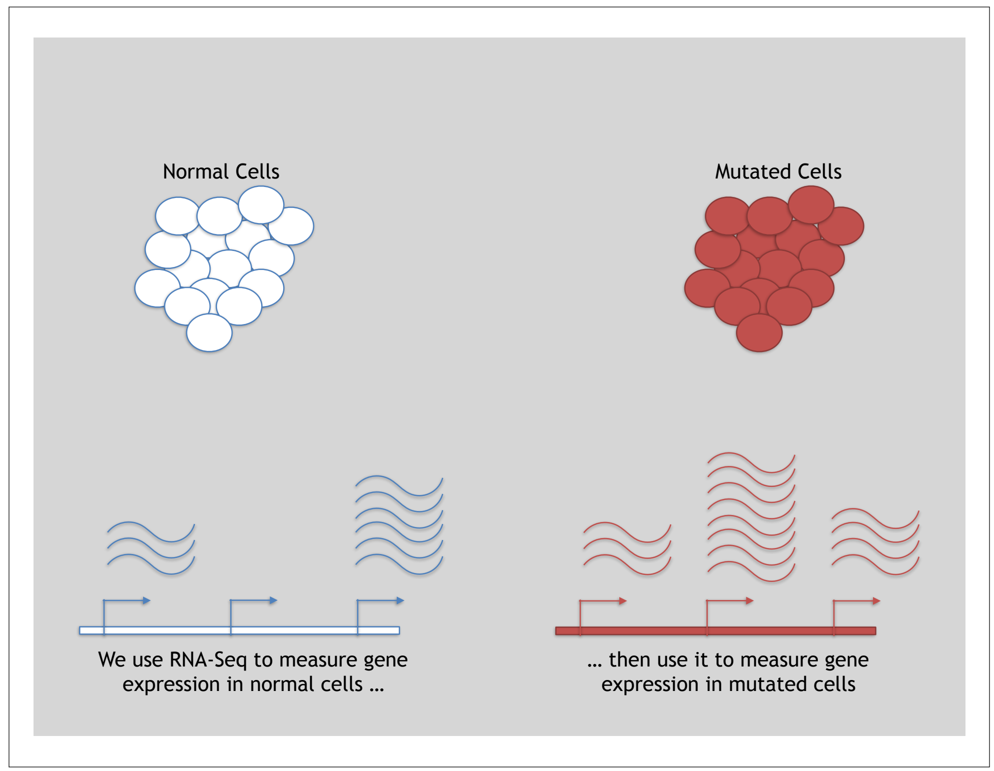


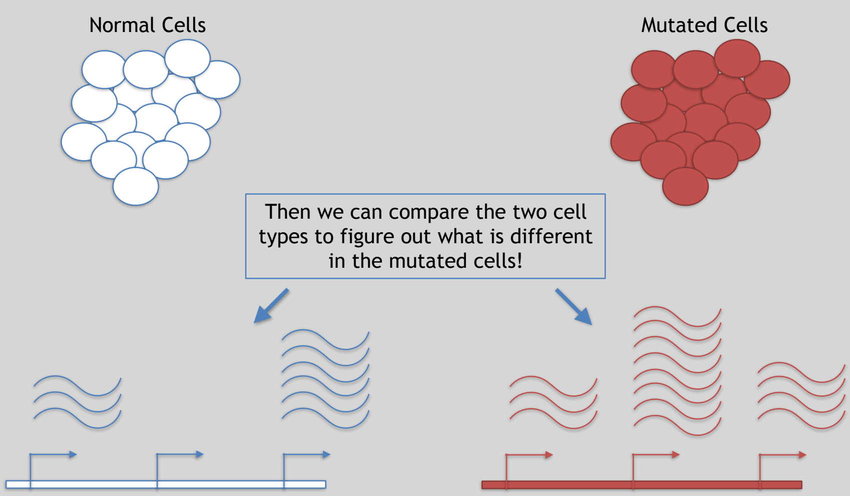


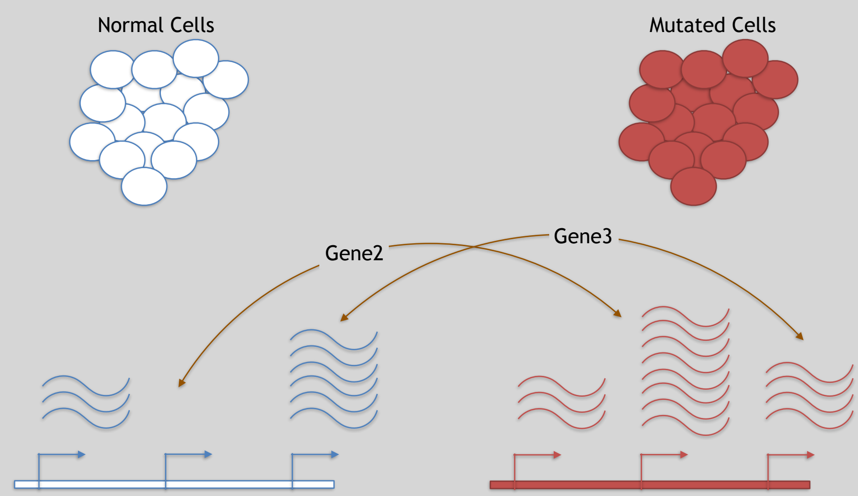

Normal Cells
Mutated Cells
Some genes are active more than others
mRNA transcripts
Gene1
Gene2
Gene3


Normal Cells
Mutated Cells
Gene 3 is the most active
Gene 2 is not active
mRNA transcripts
Gene1
Gene2
Gene3


Normal Cells
Mutated Cells
HTS tells us which genes are active, and how much they are transcribed!
mRNA transcripts
Gene1
Gene2
Gene3


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





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)

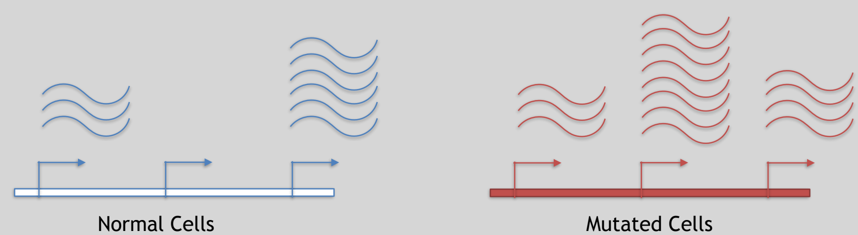
3) **Data analysis**
(Often the major bottleneck to overall success!)

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/bggn213_S19/lectures/#13

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

Feedback:
[Muddy Point Assessment]

Reference

Additional Reference Slides
on SAM/BAM Format and Sequencing Methods

Reference

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	

SAM Format

Reference

- **S**equence **A**lignment/**M**ap (**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

Example SAM File

Reference

- 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/bimm143_F18/class-material/sam_format/

Header section

```
@HD     VN:1.0          SO:coordinate
@SQ     SN:1            LN:249250621    AS:NCBI37       UR:file:/data/local/ref/GATK/human_g1k_v37.fasta        M5:1b22b98cdeb4a9304cb5d48026a85128
@SQ     SN:2            LN:243199373    AS:NCBI37       UR:file:/data/local/ref/GATK/human_g1k_v37.fasta        M5:a0d9851da00400dec1098a9255ac712e
@SQ     SN:3            LN:198022430    AS:NCBI37       UR:file:/data/local/ref/GATK/human_g1k_v37.fasta        M5:fdfd811849cc2fadebc929bb925902e5
@RG     ID:UM0098:1     PL:ILLUMINA     PU:HWUSI-EAS1707-615LHAAXX-L001     LB:80       DT:2010-05-05T20:00:00-0400     SM:SD37743      CN:UMCORE
@RG     ID:UM0098:2     PL:ILLUMINA     PU:HWUSI-EAS1707-615LHAAXX-L002     LB:80       DT:2010-05-05T20:00:00-0400     SM:SD37743      CN:UMCORE
@PG     ID:bwa          VN:0.5.4
```

Alignment section

```
1:497:R:-272+13M17D24M  113     1       497     37      37M     15      100338662       0
CGGGTCTGACCTGAGGAGAACTGTGCTCCGCCTTCAG   0;==-==9;>>>>>=>>>>>>>>>>>=>>>>>>>>>>   XT:A:U  NM:i:0  SM:i:37 AM:i:0  X0:i:1  X1:i:0
XM:i:0  XO:i:0  XG:i:0  MD:Z:37
19:20389:F:275+18M2D19M 99      1       17644   0       37M     =       17919   314
TATGACTGCTAATAATACCTACACATGTTAGAACCAT   >>>>>>>>>>>>>>>>>>>><<>>><<>>4::>>:<9   RG:Z:UM0098:1   XT:A:R  NM:i:0  SM:i:0  AM:i:0  X0:i:4
X1:i:0  XM:i:0  XO:i:0  XG:i:0  MD:Z:37
19:20389:F:275+18M2D19M 147     1       17919   0       18M2D19M        =       17644   -314
GTAGTACCAACTGTAAGTCCTTATCTTCATACTTTGT   ;44999;499<8<8<<<8<<><<<<><7<;<<<>><<   XT:A:R  NM:i:2  SM:i:0  AM:i:0  X0:i:4  X1:i:0
XM:i:0  XO:i:1  XG:i:2  MD:Z:18^CA19
9:21597+10M2I25M:R:-209 83      1       21678   0       8M2I27M =       21469   -244
CACCACATCACATATACCAAGCCTGGCTGTGTCTTCT   <;9<<5><<<<><<<>><<><>><9>><>>>9>>><>   XT:A:R  NM:i:2  SM:i:0  AM:i:0  X0:i:5  X1:i:0
XM:i:0  XO:i:1  XG:i:2  MD:Z:35
```

SAM header section

Reference

- 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

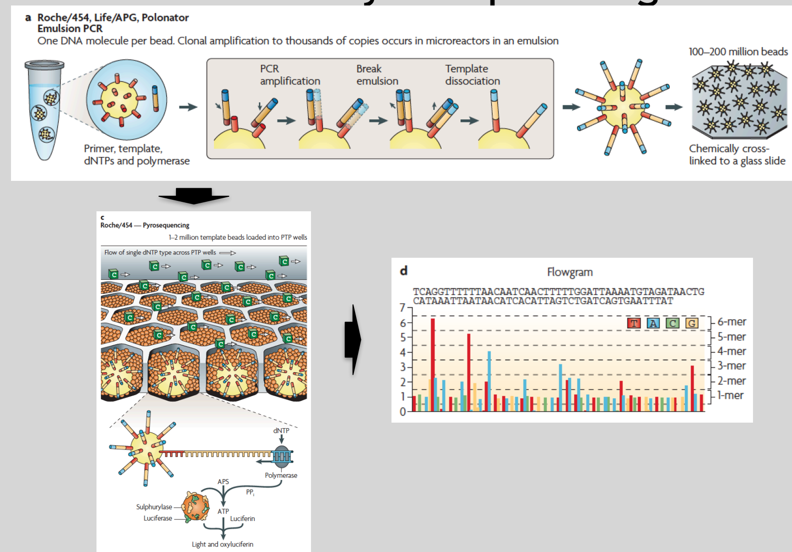
Reference

- **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)

Reference

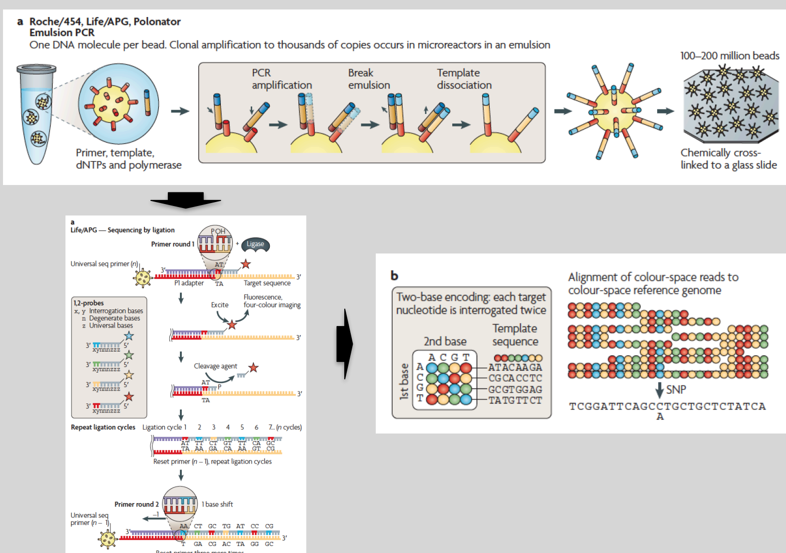
Additional Reference Slides on Sequencing Methods

Roche 454 - Pyrosequencing



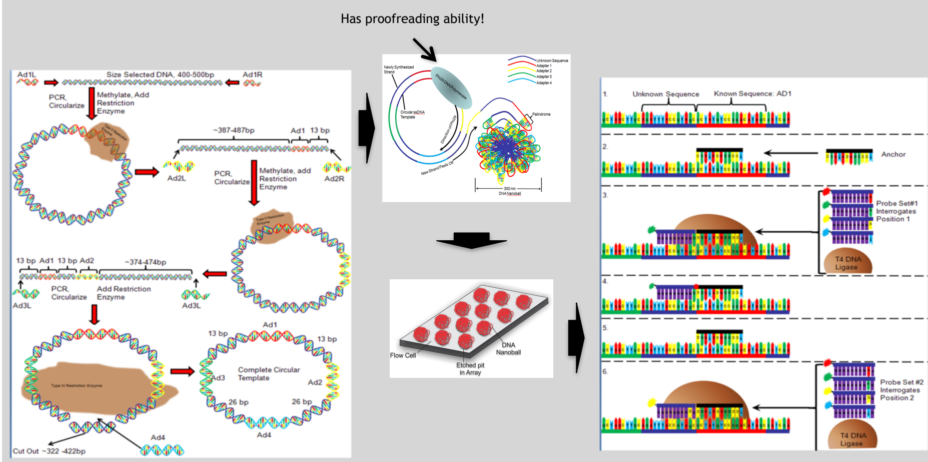
Metzker, ML (2010), *Nat. Rev. Genet*, 11, pp. 31-46

Life Technologies SOLiD – Sequence by Ligation



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

Complete Genomics – Nanoball Sequencing



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

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

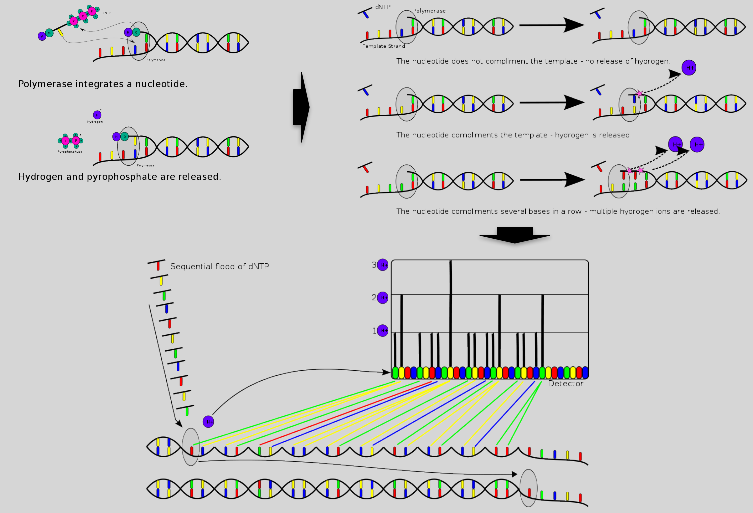
“Benchtop” Sequencers

- 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[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

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

PGM - Ion Semiconductor Sequencing



Wikipedia, “Ion Semiconductor Sequencing”, September 26, 2012