BIMM 143

Genome Informatics I

Lecture 13

Barry Grant
UC San Diego

http://thegrantlab.org/bimm143

TODAYS MENU:

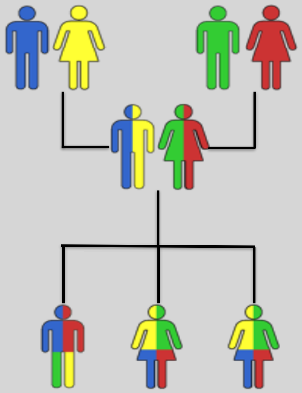
- **What is a Genome?**
  - Genome sequencing and the Human genome project
- **What can we do with a Genome?**
  - Compare, model, mine and edit
- **Modern Genome Sequencing**
  - 1st, 2nd and 3rd generation sequencing
- **Workflow for NGS**
  - RNA-Sequencing and Discovering variation

Genetics and Genomics

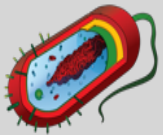
- **Genetics** is primarily the study of individual genes, mutations within those genes, and their inheritance patterns in order to understand specific traits.
- **Genomics** expands upon classical genetics and considers aspects of the entire genome, typically using computer aided approaches.

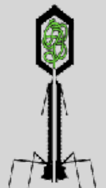
What is a Genome?

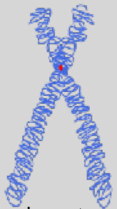
The total genetic material of an organism by which individual traits are encoded, controlled, and ultimately passed on to future generations



Genomes come in many shapes


prokaryote

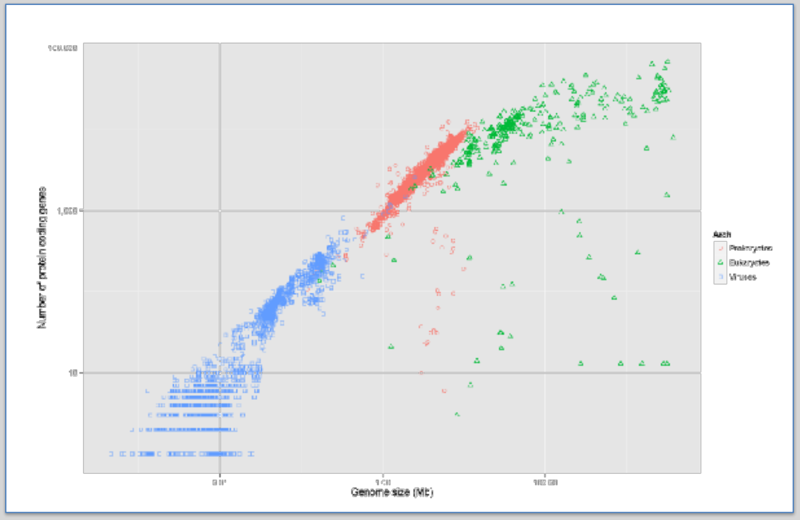

bacteriophage


eukaryote

- 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 by Mariana Ruiz Villarreal| Bacteriophage image by Splette / CC BY-SA | Eukaryote image by Magnus Manske/ CC BY-SA

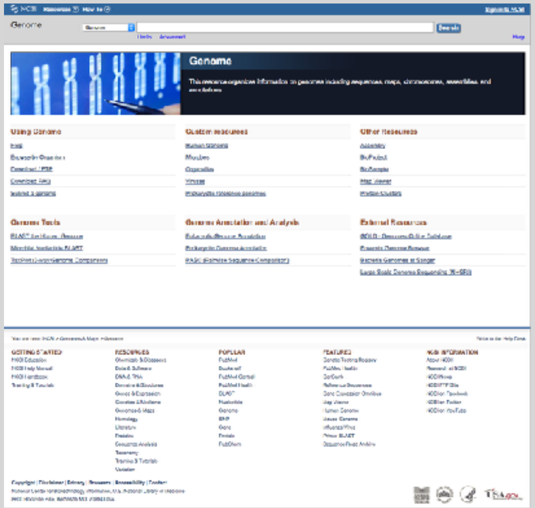
Genomes come in many sizes



Modified from image by Estevezj / CC BY-SA

Genome Databases

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



Genome

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

Using Genome
- Help
- Browse by Organism
- Download / FTP
- Download FAQ
- Submit a genome

Custom resources
- Human Genome
- Microbes
- Organelles
- Viruses
- Prokaryotic reference genomes

Other Resources
- Assembly
- BioProject
- BioSample
- Map Viewer
- Protein Clusters

Genome Tools
- BLAST the Human Genome
- Microbial Nucleotide BLAST
- TaxPlot (3-way Genome Comparisons)

Genome Annotation and Analysis
- Eukaryotic Genome Annotation
- Prokaryotic Genome Annotation
- PASC (Pairwise Sequence Comparison)

External Resources
- GOLD - Genomes OnLine Database
- Ensembl Genome Browser
- Pathogen Genomes at Sanger
- Large Scale Genome Sequencing (W-GS)

You are here: NCBI > Genomes & Maps > Genome

Write to the Help Desk

GETTING STARTED
- NCBI Education
- NCBI Help Manual
- NCBI Handbook
- Training & Tutorials

RESOURCES
- Chemicals & Bioassays
- Data & Software
- DNA & RNA
- Domains & Structures
- Genes & Expression
- Genetics & Medicine
- Genomes & Maps
- Homology
- Literature
- Proteins
- Sequence Analysis
- Taxonomy
- Training & Tutorials
- Variation

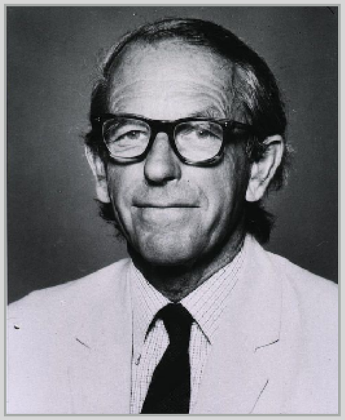
POPULAR
- PubMed
- Bookshelf
- PubMed Central
- PubMed Health
- BLAST
- Nucleotide
- Genome
- SNP
- Gene
- Protein
- PubChem

FEATURED
- Genetic Testing Registry
- PubMed Health
- GenBank
- Reference Sequences
- Gene Expression Omnibus
- Map Viewer
- Human Genome
- Mouse Genome
- Influenza Virus
- Primer-BLAST
- Sequence Read Archive

NCBI INFORMATION
- About NCBI
- Research at NCBI
- NCBI News
- NCBI FTP Site
- NCBI on Facebook
- NCBI on Twitter
- NCBI on YouTube

Copyright | Disclaimer | Privacy | Browsers | Accessibility | Contact
National Center for Biotechnology Information, U.S. National Library of Medicine
8600 Rockville Pike, Bethesda MD, 20894 USA

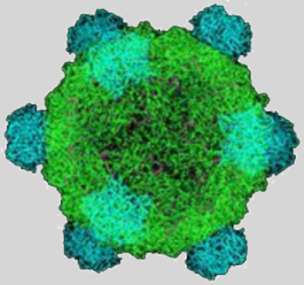
Early Genome Sequencing



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

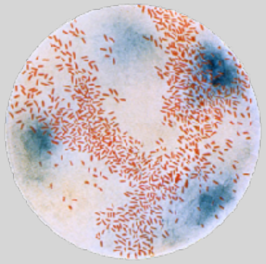
- Chain-termination “Sanger” sequencing was developed in 1977 by Frederick Sanger, colloquially referred to as the “Father of Genomics”
- Sequence reads were typically 750-1000 base pairs in length with an error rate of ~1 / 10000 bases

The First Sequenced Genomes



Bacteriophage φ-X174

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



Haemophilus influenzae

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

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

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
  - 22 autosomes, 2 sex chromosomes
  - ~20,000 genes

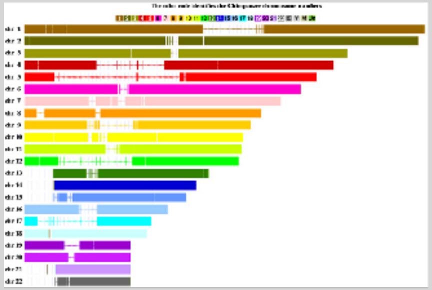
Jane Ades, Courtesy: National Human Genome Research Institute.

What can we do with a Genome?

- We can ***compare*** genomes, both within and between species, to identify regions of variation and of conservation
- We can ***model*** genomes, to find interesting patterns reflecting functional characteristics
- We can ***mine*** genomes, to find mutations and epigenetic correlations with disease, drug sensitivity, treatment efficacy and other phenotypic characteristics
- We can ***edit*** genomes, to add, remove, or modify genes and other regions for adjusting individual traits

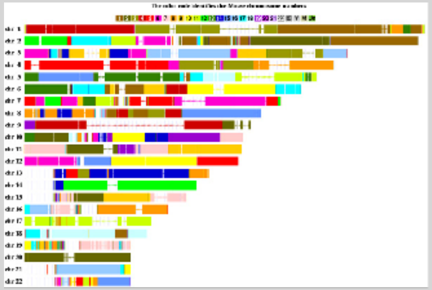
Comparative Genomics

~6-7 million years



Chimpanzee

~60-70 million years



Mouse

http://cbse.soe.ucsc.edu/research/comp_genomics/human_chimp_mouse

Conservation Suggests Function

- Functional regions of the genome tend to mutate slower than nonfunctional regions due to selective pressures
- Comparing genomes can therefore indicate segments of high similarity that have remained conserved across species as candidate genes or regulatory regions

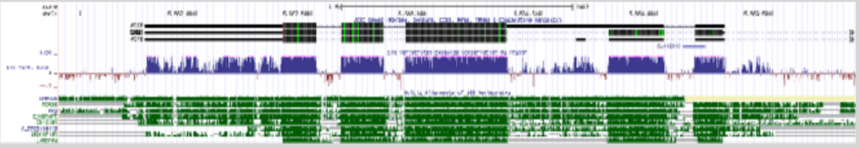


figure generated from: http://genome.ucsc.edu/

Conservation Indicates Loss

- Comparing genomes allows us to also see what we have lost over evolutionary time
- A model example of this is the loss of “penile spines” in the human lineage due to a human-specific deletion of an enhancer for the androgen receptor gene (McLean et al, Nature, 2011)

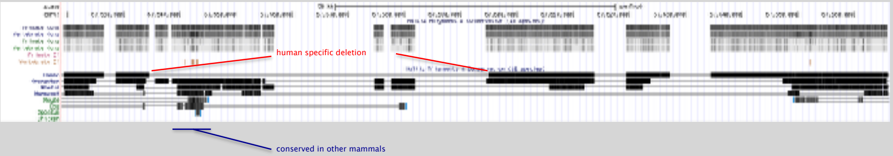
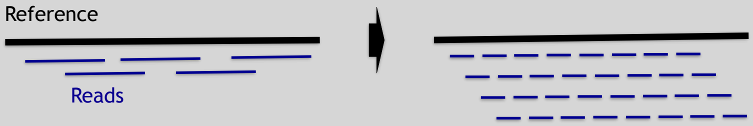


figure generated from: http://genome.ucsc.edu/

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

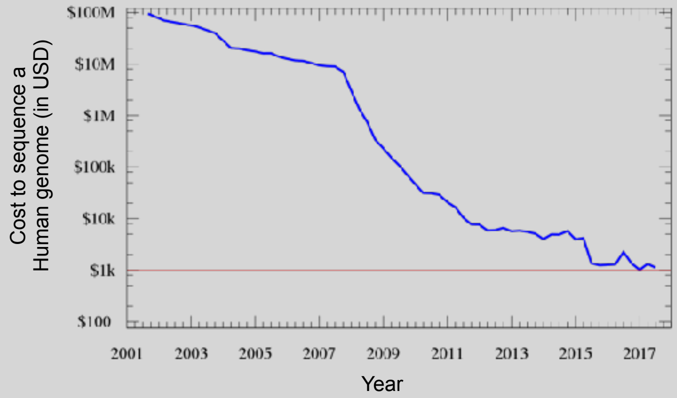
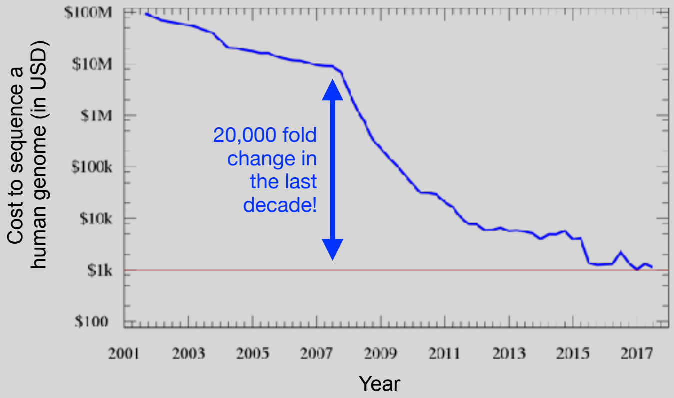


Image source: https://en.wikipedia.org/wiki/Carlson_curve

Rapid progress of genome sequencing

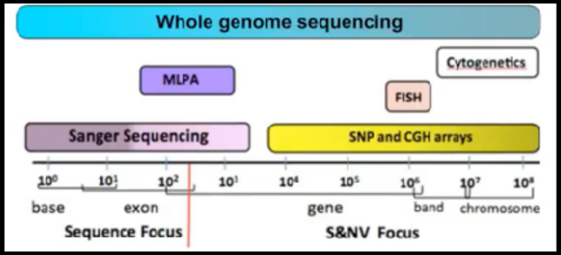


20,000 fold change in the last decade!

MRI: $4k

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 <u>medicine</u>

- 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.

Solving mystery diseases

- Diseases with a genetic origin effect 16 million people in the US and 23% of all pediatric admissions to hospital are for ‘rare' genetic disorders.
- Most are “mystery diseases” in terms of their genetic origin
- Before the recent adoption of exom and genome sequencing these patients faced extensive periods of testing and inappropriate treatment (with cost estimates of $5 million per person) before the basis of their disease was understood.
- Sequencing can thus help realize enormous savings in healthcare costs and spare patients and their families unnecessary, stressful, and time-consuming testing.

How many Mendelian diseases are there?

- As of 01/10/18 ~7,800 Mendelian diseases have been described.
- For 3,963 of these, the likely disease gene is known.
- For many genes, different genetic variants can have distinct effects on the encoded protein, leading to distinct disease characteristics.
- Indeed, the 3,963 unique diseases that have been solved affect only 2,776 genes because different mutations in the same gene can cause different disease characteristics.

How many Mendelian diseases are there?

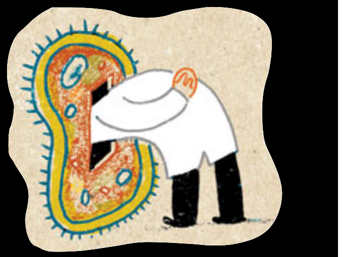
- It is probable that many more Mendelian diseases will be “solved” as genomic analysis becomes more integrated into clinical practice.
- There are ~20,000 protein coding genes and and variants in many of these genes would be expected to cause human disease.
- **Q:** How are genes responsible for genetic diseases currently identified?
  - **Exome** or **whole genome sequencing**

Currently disease causing mutations are found in only ~30% of cases

- For the majority of these cases finding disease causing mutations often does not lead to effective treatments.
- However, the information can still be helpful for guiding patient management, reproductive choices and future certainty. For example:
  - Can bring relief for patients and their families
  - Can be helpful for planning future pregnancies (e.g. IVF and genetic testing for embryo selection)
  - Predicting the possible disease course and long-term prognosis

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?

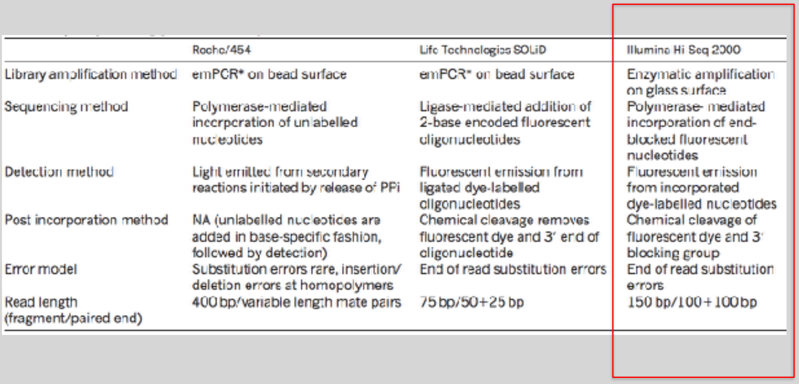
Type of change	Some common technology to study changes
DNA mutations	WGS, WXS
DNA structural variations	WGS
Copy number variation (CNV)	CGH array, SNP array, WGS
DNA methylation	Methylation array, RRBS, WGBS
mRNA expression changes	mRNA expression array, RNA-seq
miRNA expression changes	miRNA expression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

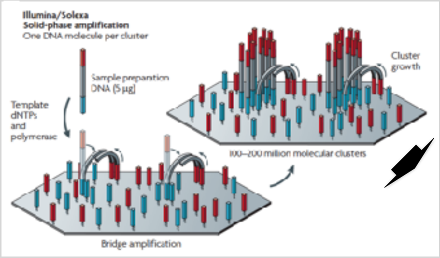
Modern NGS Sequencing Platforms

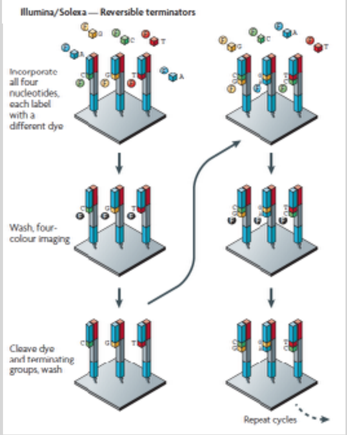


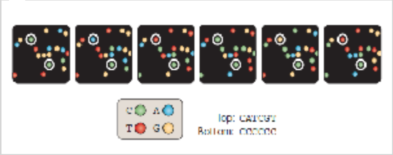
	Roche/454	Life Technologies SOLiD	Illumina Hi Seq 2000
Library amplification method	emPCR* on bead surface	emPCR* on bead surface	Enzymatic amplification on glass surface
Sequencing method	Polymerase-mediated incorporation of unlabelled nucleotides	Ligase-mediated addition of 2-base encoded fluorescent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


Illumina/Solexa
Solid-phase amplification
One DNA molecule per cluster
Sample preparation DNA (5 µg)
Template dNTPs and polymerase
Cluster growth
100–200 million molecular clusters
Bridge amplification

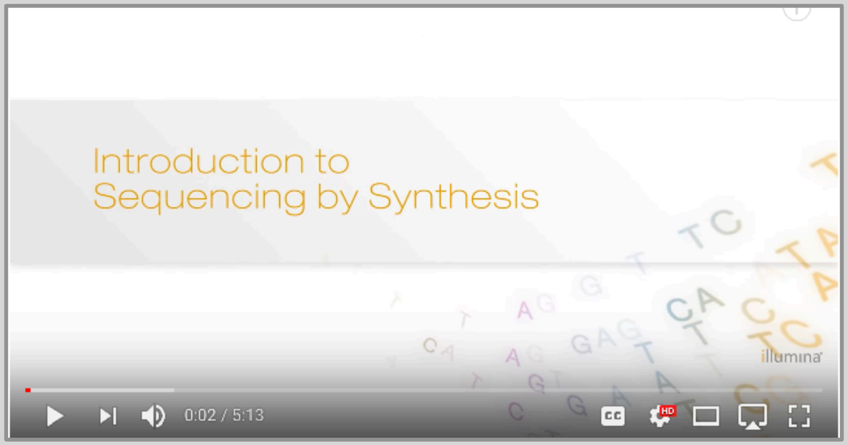

Illumina/Solexa — Reversible terminators
Incorporate all four nucleotides, each label with a different dye
Wash, four-colour imaging
Cleave dye and terminating groups, wash
Repeat cycles
A
C
G
T



(other sequencing platforms summarized at end of slide set)

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

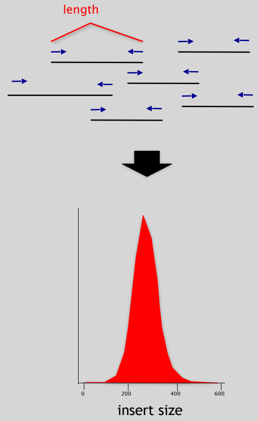
Illumina Sequencing - Video


Introduction to Sequencing by Synthesis

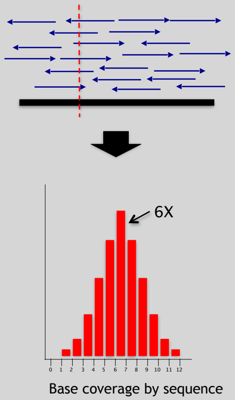
https://www.youtube.com/watch?src_vid=womKfikWlxM&v=fCd6B5HRaZ8

NGS Sequencing Terminology

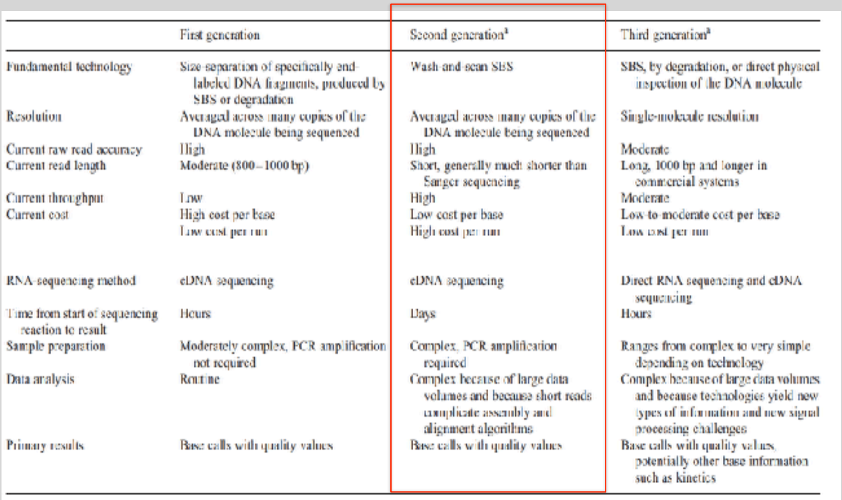
Insert Size



Sequence Coverage



Summary: "Generations" of DNA Sequencing



	First generation	Second generation[a]	Third generation[b]
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!

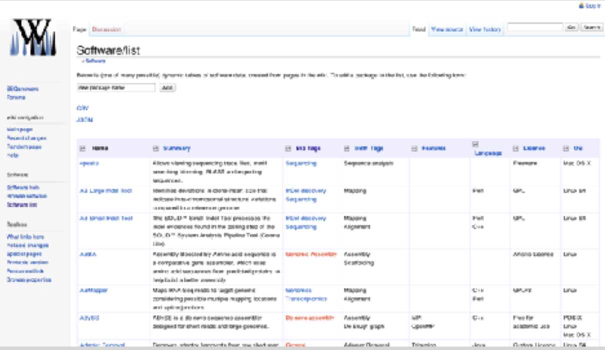
The first direct RNA sequencing by nanopore

Side-Note:

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

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



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/

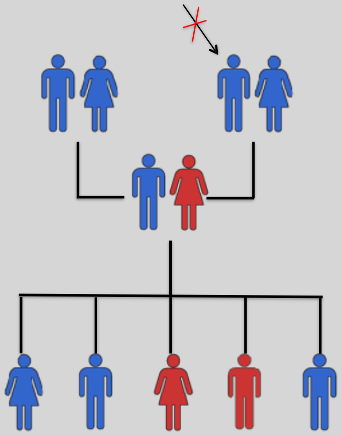
> “Variety’s the very spice of life”
>
> –William Cowper, 1785

> “Variation is the spice of life”
>
> –Kruglyak & Nickerson, 2001

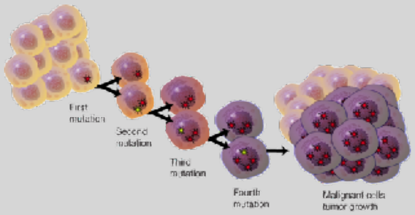
- While the sequencing of the human genome was a great milestone, the DNA from a single person is not representative of the millions of potential differences that can occur between individuals
- These unknown genetic variants could be the cause of many phenotypes such as differing morphology, susceptibility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

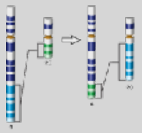
- **Single Nucleotide Polymorphisms** (SNPs) - mutations of one nucleotide to another

```
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

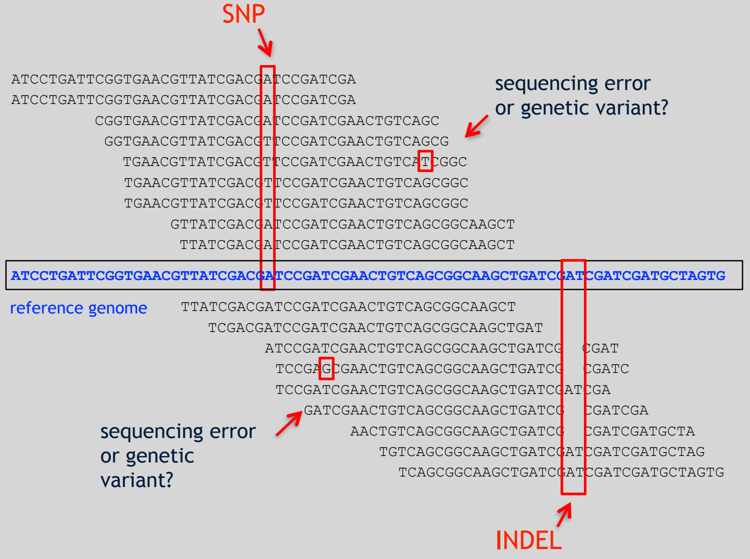
Numbers change depending on ancestry!

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



SNP

sequencing error or genetic variant?

```
ATCCTGATTCGGTGAACGTTATCGACGATCCGATCGA
ATCCTGATTCGGTGAACGTTATCGACGATCCGATCGA
         CGGTGAACGTTATCGACGATCCGATCGAACTGTCAGC
          GGTGAACGTTATCGACGTTCCGATCGAACTGTCAGCG
            TGAACGTTATCGACGTTCCGATCGAACTGTCATCGGC
            TGAACGTTATCGACGTTCCGATCGAACTGTCAGCGGC
            TGAACGTTATCGACGTTCCGATCGAACTGTCAGCGGC
                 GTTATCGACGATCCGATCGAACTGTCAGCGGCAAGCT
                  TTATCGACGATCCGATCGAACTGTCAGCGGCAAGCT
ATCCTGATTCGGTGAACGTTATCGACGATCCGATCGAACTGTCAGCGGCAAGCTGATCGATCGATCGATGCTAGTG
                  TTATCGACGATCCGATCGAACTGTCAGCGGCAAGCT
                     TCGACGATCCGATCGAACTGTCAGCGGCAAGCTGAT
                           ATCCGATCGAACTGTCAGCGGCAAGCTGATCG  CGAT
                            TCCGAGCGAACTGTCAGCGGCAAGCTGATCG  CGATC
                            TCCGATCGAACTGTCAGCGGCAAGCTGATCGATCGA
                               GATCGAACTGTCAGCGGCAAGCTGATCG  CGATCGA
                                    AACTGTCAGCGGCAAGCTGATCG  CGATCGATGCTA
                                       TGTCAGCGGCAAGCTGATCGATCGATCGATGCTAG
                                         TCAGCGGCAAGCTGATCGATCGATCGATGCTAGTG
```

reference genome

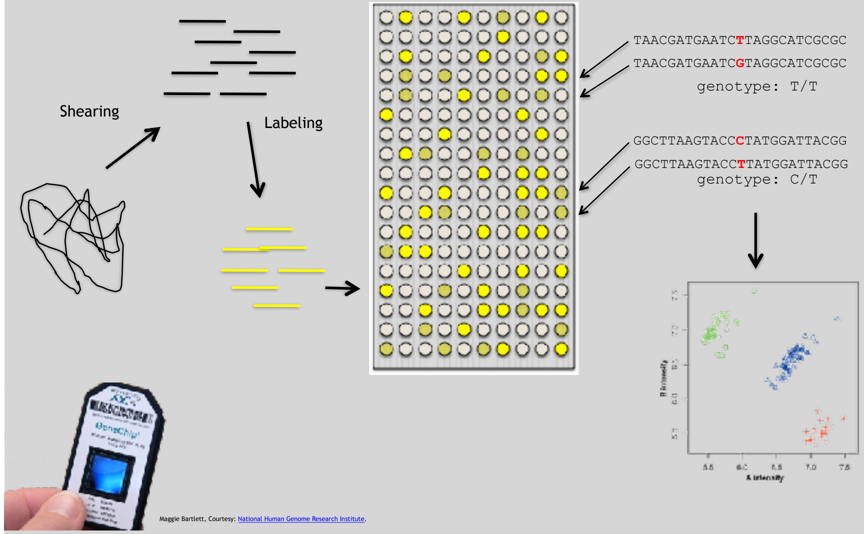
sequencing error or genetic variant?

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



Shearing

Labeling

TAACGATGAATC**T**TAGGCATCGCGC

TAACGATGAATC**G**TAGGCATCGCGC

genotype: T/T

GGCTTAAGTACC**C**TATGGATTACGG

GGCTTAAGTACC**T**TATGGATTACGG

genotype: C/T

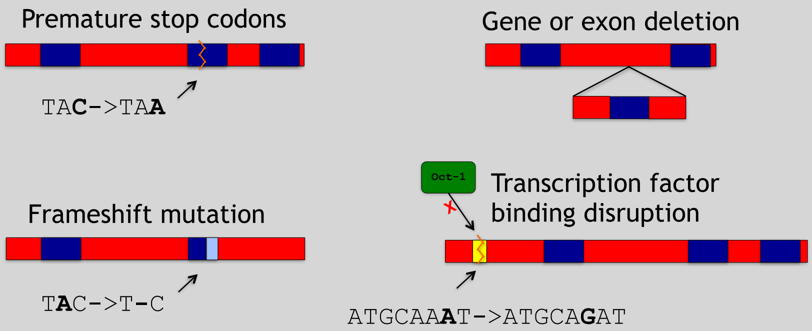
Maggie Bartlett, Courtesy: National Human Genome Research Institute.

Discovering Variation: SVs

- Structural variants can be discovered by both sequence and microarray approaches
- Microarrays can only detect genomic imbalances, specifically copy number variants (CNVs)
- Sequence based approaches can, in principle, identify all types of structural rearrangements

Impact of Genetic Variation

There are numerous ways genetic variation can exhibit functional effects



Premature stop codons

TA**C**->TA**A**

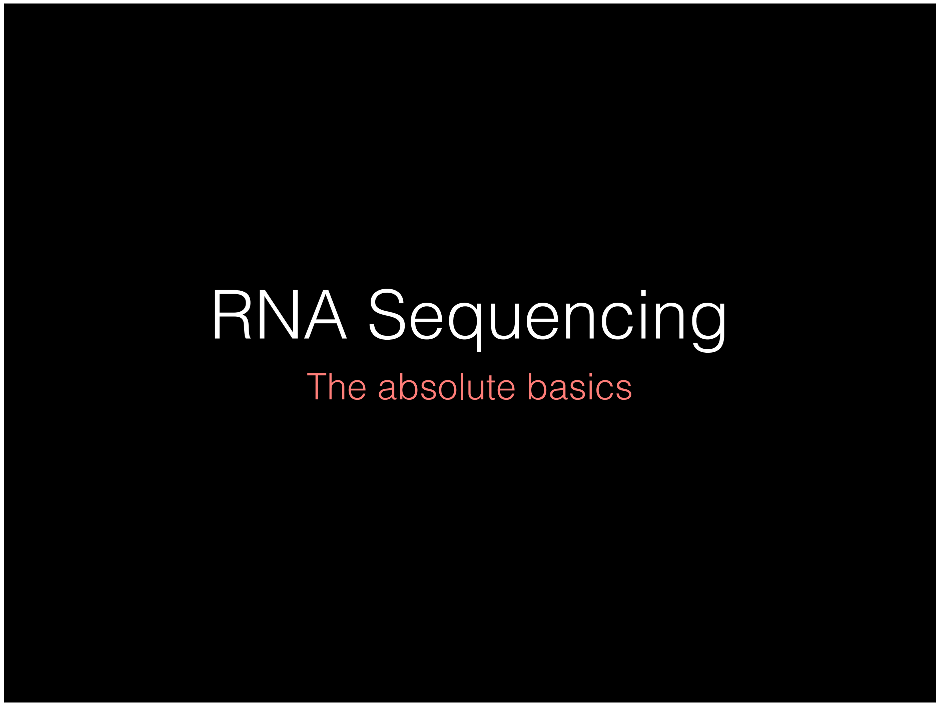
Gene or exon deletion

Frameshift mutation

T**A**C->T-C

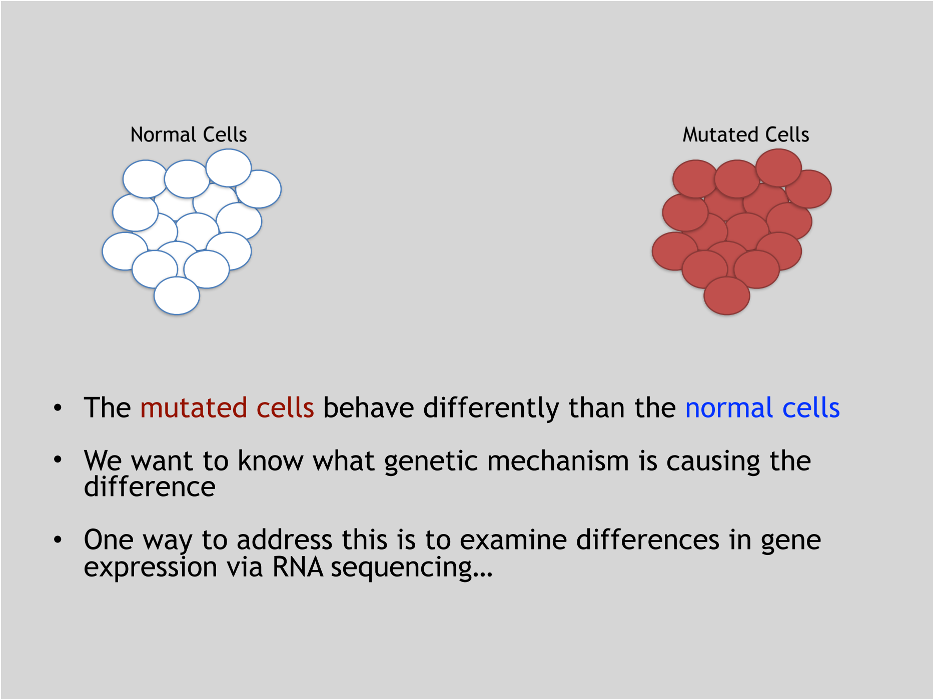
Transcription factor binding disruption

ATGCAA**A**T->ATGCA**G**AT



RNA Sequencing

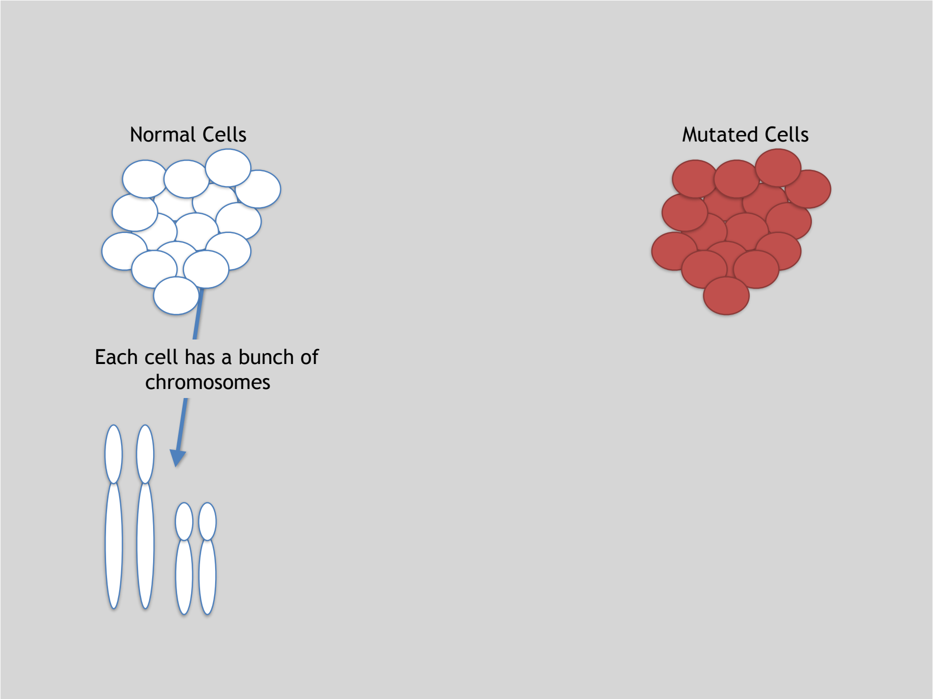
The absolute basics



Normal Cells

Mutated Cells

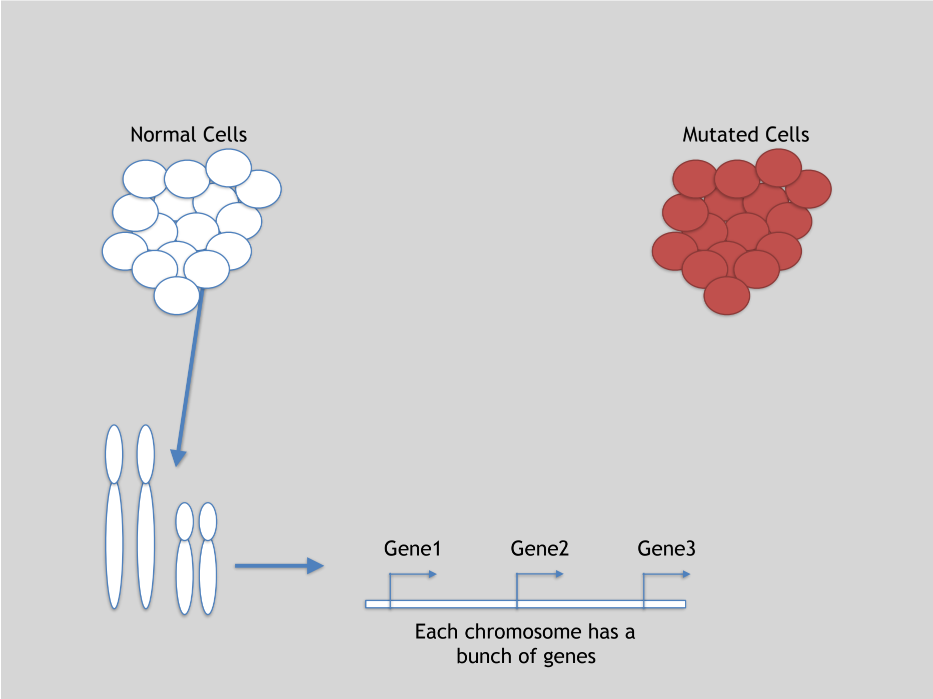
- The mutated cells behave differently than the normal cells
- We want to know what genetic mechanism is causing the difference
- One way to address this is to examine differences in gene expression via RNA sequencing…



Normal Cells

Mutated Cells

Each cell has a bunch of chromosomes



Normal Cells

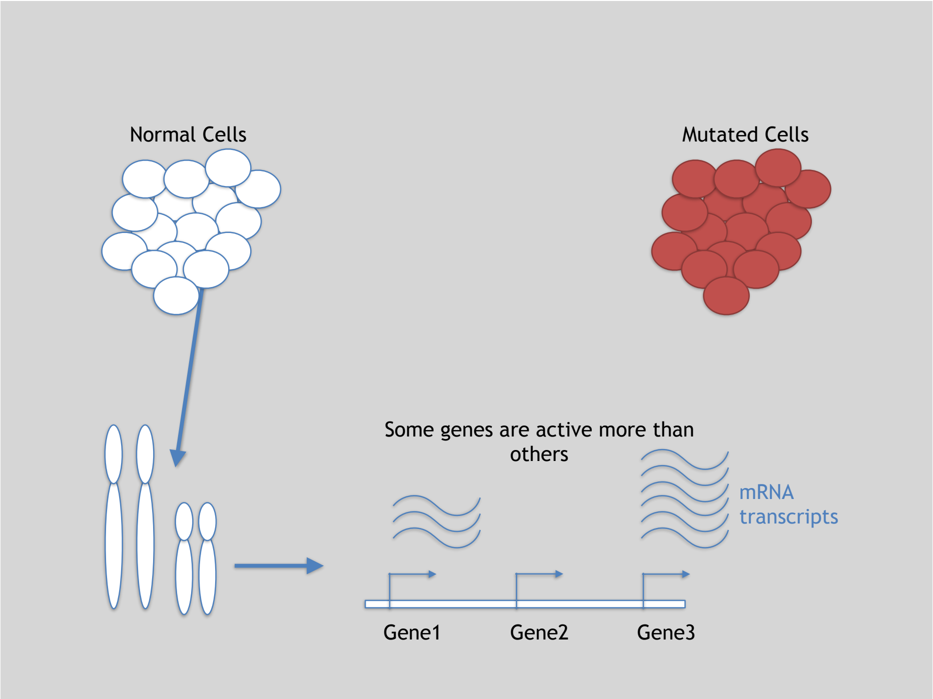
Mutated Cells

Gene1

Gene2

Gene3

Each chromosome has a bunch of genes



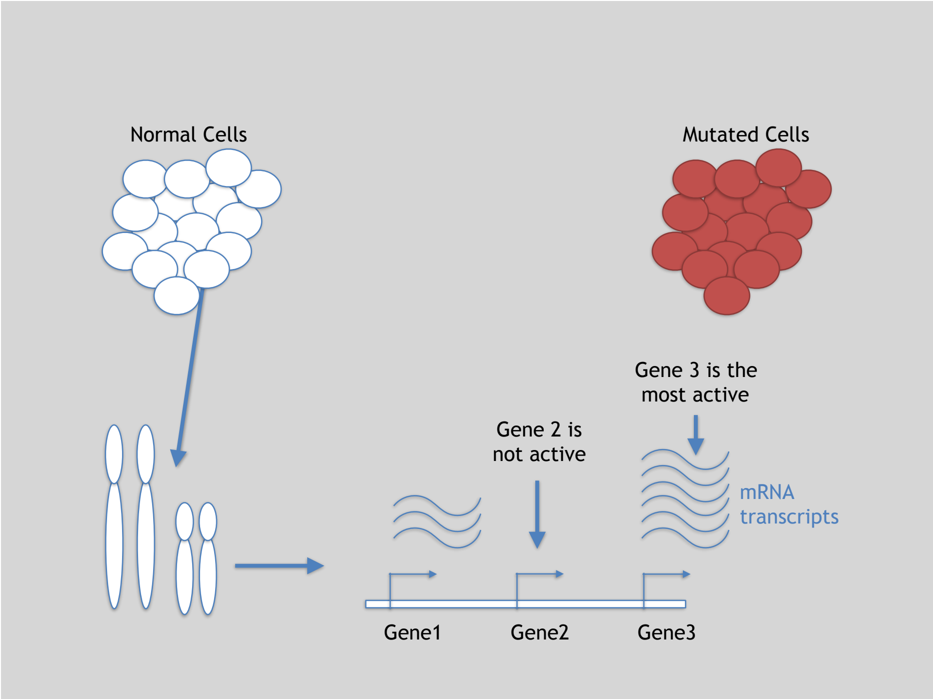
Normal Cells

Mutated Cells

Some genes are active more than others

mRNA transcripts

Gene1 Gene2 Gene3



Normal Cells

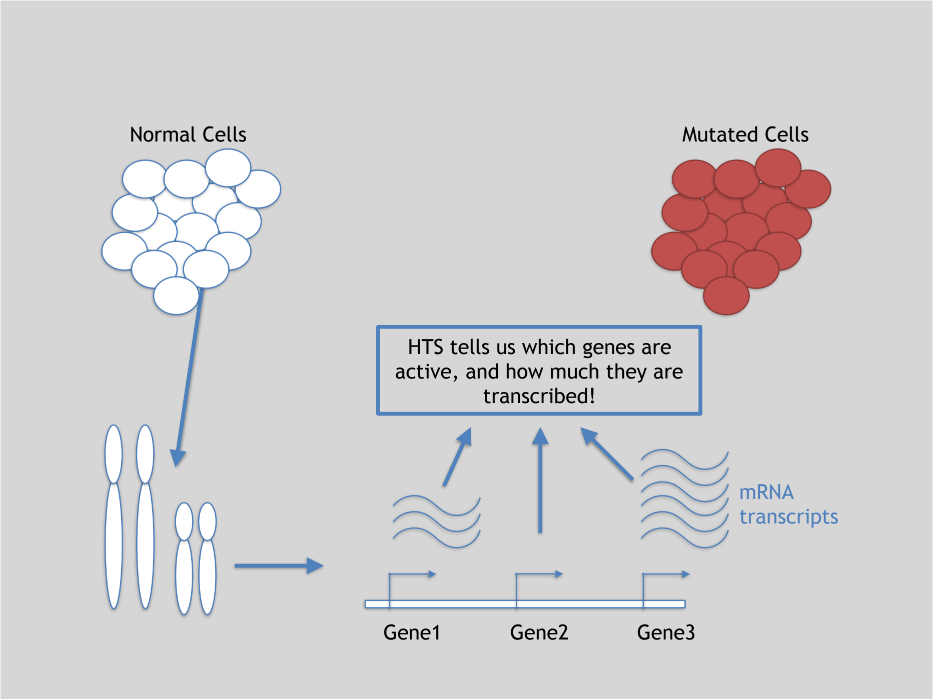
Mutated Cells

Gene 3 is the most active

Gene 2 is not active

mRNA transcripts

Gene1 Gene2 Gene3



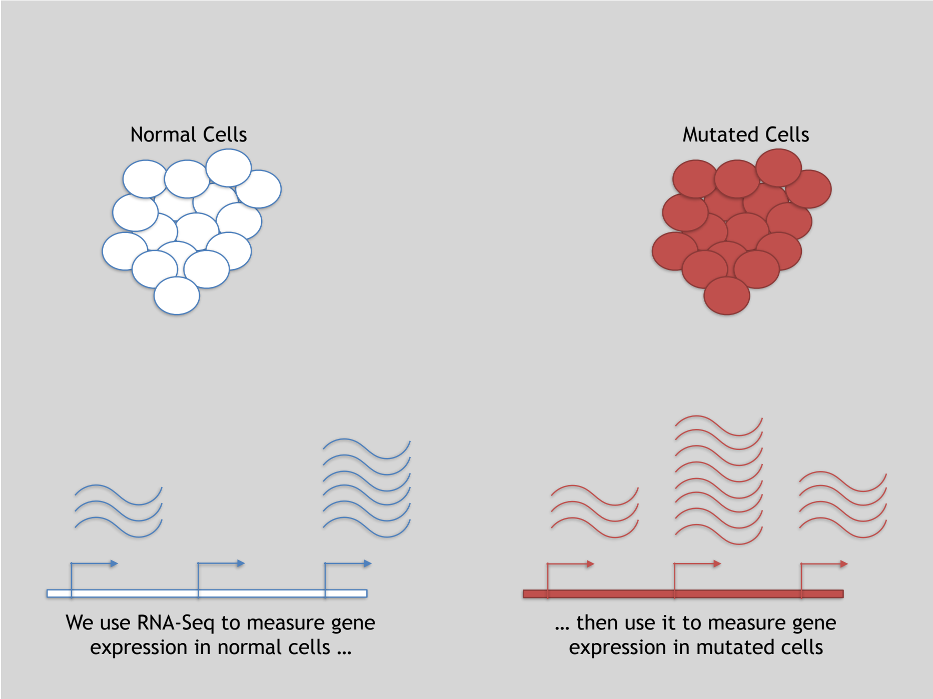
Normal Cells

Mutated Cells

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

mRNA transcripts

Gene1 Gene2 Gene3

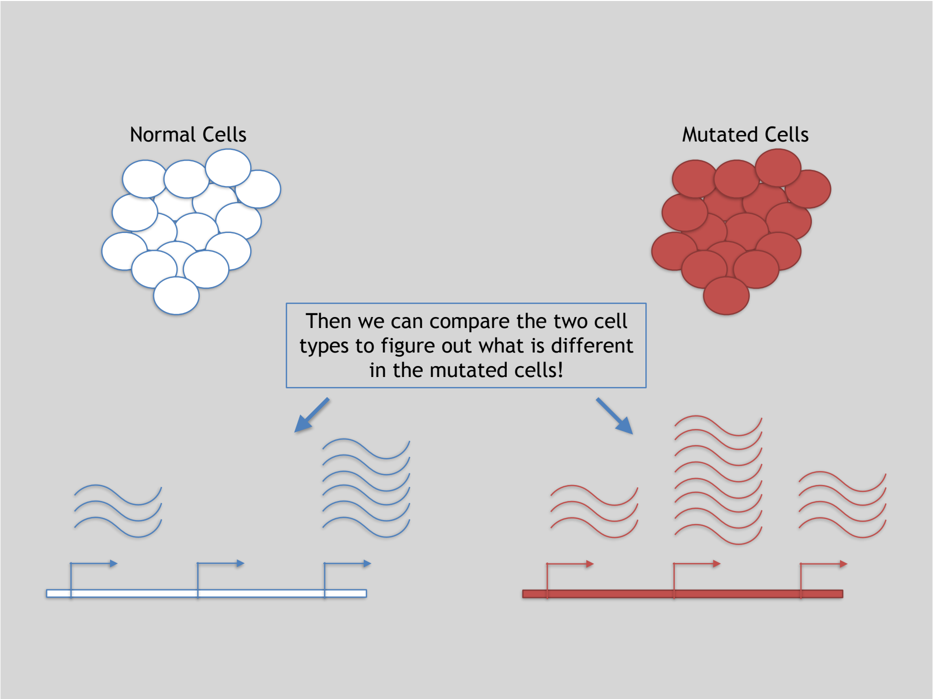


Normal Cells

Mutated Cells

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

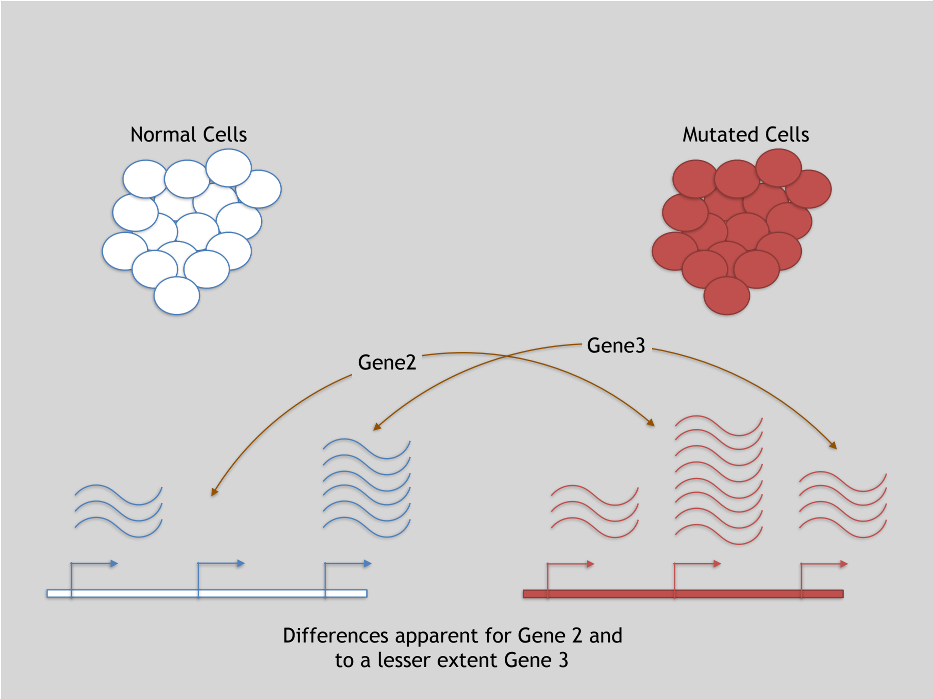
… then use it to measure gene expression in mutated cells



Normal Cells

Mutated Cells

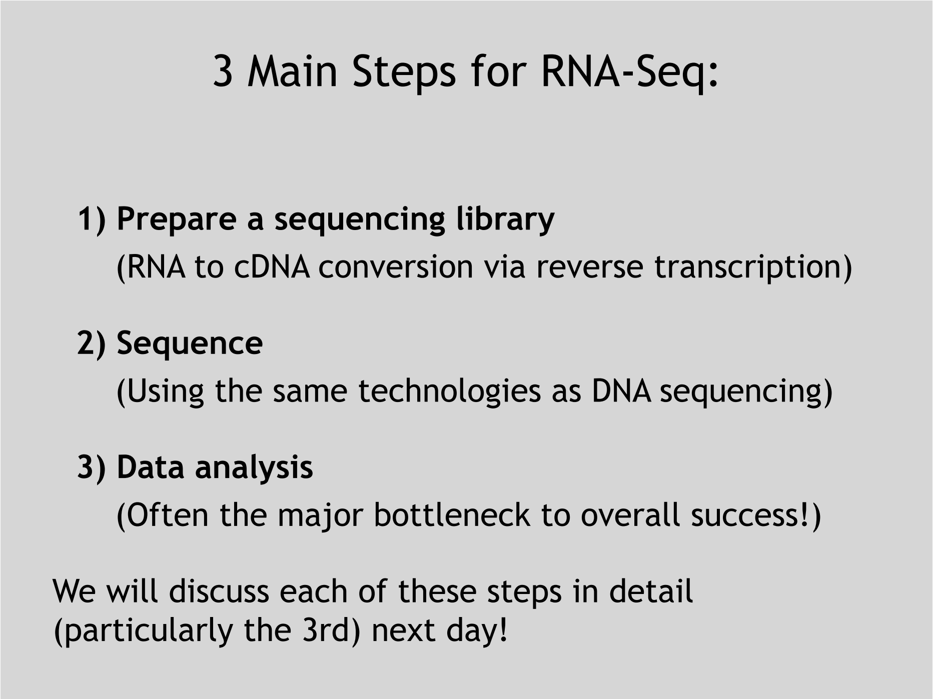
Then we can compare the two cell types to figure out what is different in the mutated cells!



Normal Cells

Mutated Cells

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



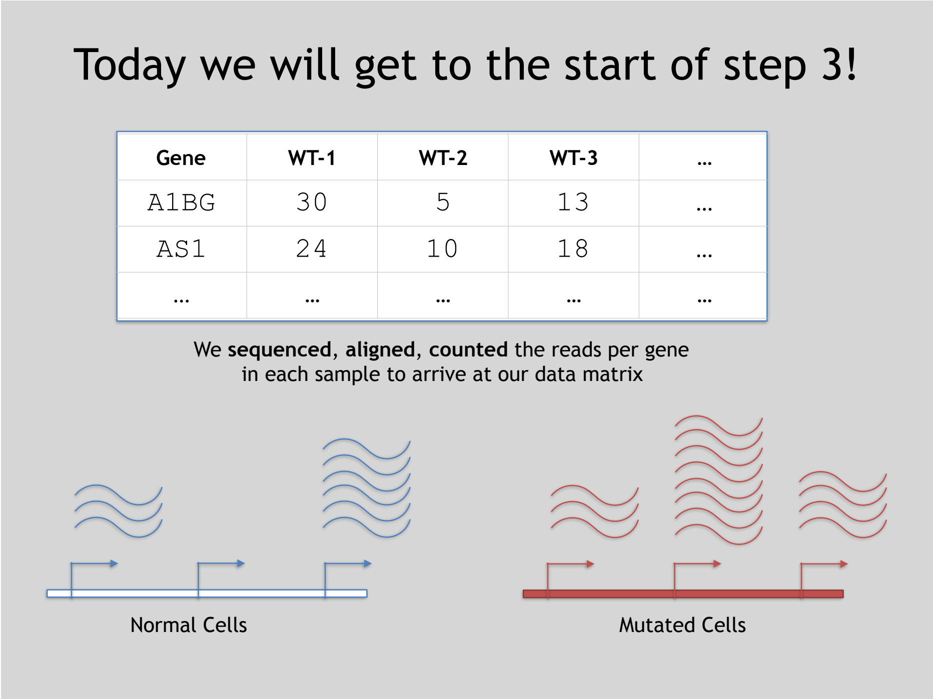
3 Main Steps for RNA-Seq:

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

Normal Cells

Mutated Cells

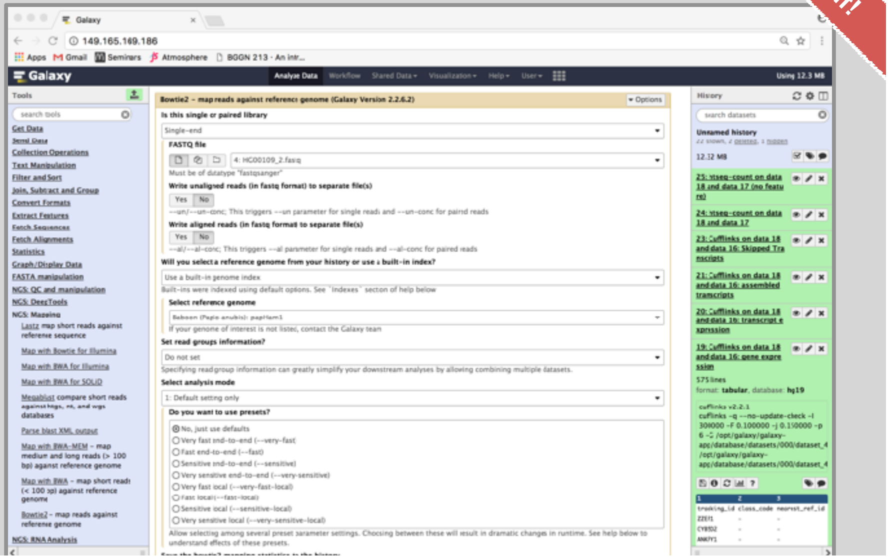
TODAYS MENU:

- **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

Access a jetstream galaxy instance!

Do it Yourself!

Use assigned IP address



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

Is this single or paired library
Single-end

FASTQ file
4: HG00109_2.fastq
Must be of datatype "fastqsanger"

Write unaligned reads (in fastq format) to separate file(s)
Yes [✓ No]
--un/--un-conc; This triggers --un parameter for single reads and --un-conc for paired reads

Write aligned reads (in fastq format) to separate file(s)
Yes [✓ No]
--al/--al-conc; This triggers --al parameter for single reads and --al-conc for paired reads

Will you select a reference genome from your history or use a built-in index?
Use a built-in genome index
Built-ins were indexed using default options. See `Indexes` section of help below

Select reference genome
Baboon (Papio anubis): papHam1
If your genome of interest is not listed, contact the Galaxy team

Set read groups information?
Do not set
Specifying 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.

Unnamed history
12.12 MB

25: htseq-count on data 18 and data 17 (no feature)
24: htseq-count on data 18 and data 17
23: Cufflinks on data 18 and data 16: Skipped Transcripts
21: Cufflinks on data 18 and data 16: assembled transcripts
20: Cufflinks on data 18 and data 16: transcript expression
19: Cufflinks on data 18 and data 16: gene expression
575 lines
format: **tabular**, database: **hg19**

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

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

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
+
AAAAAEEEEEEEEEEEE//AEEEAEEEEEEEEEEEE/EE/<<EE/AAEEAEE///EEEEAEEEEAEA<
```

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///EEEEAEEEEAEA< 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…

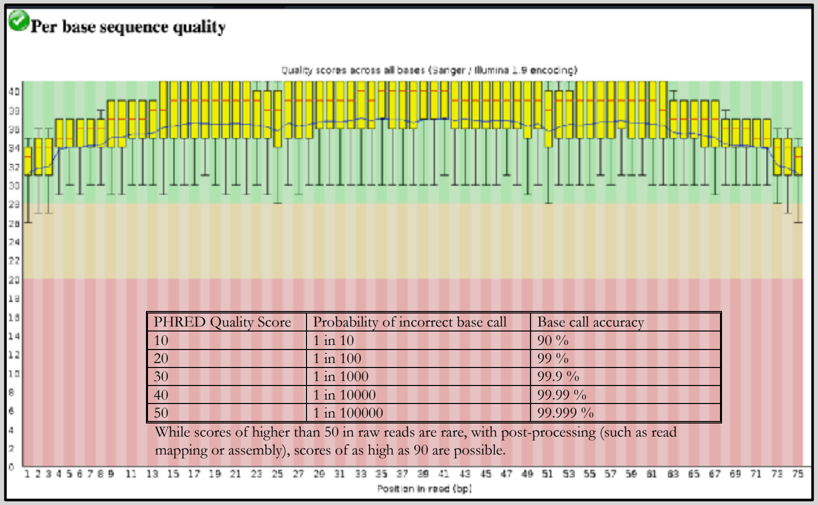
Interpreting Base Qualities in R

		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



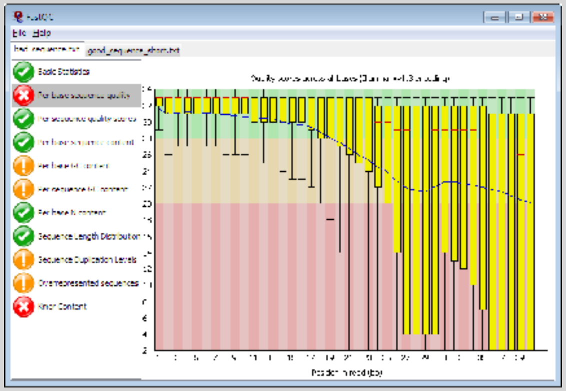
PHRED Quality Score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90 %
20	1 in 100	99 %
30	1 in 1000	99.9 %
40	1 in 10000	99.99 %
50	1 in 100000	99.999 %

While scores of higher than 50 in raw reads are rare, with post-processing (such as read mapping or assembly), scores of as high as 90 are possible.

FASTQC

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

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


File Help
good_sequence_short.txt
Basic Statistics
Sequence Length Distribution
Sequence Duplication Levels
Overrepresented sequences
Kmer Content

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

- **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

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
```

http://genome.sph.umich.edu/wiki/SAM

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)

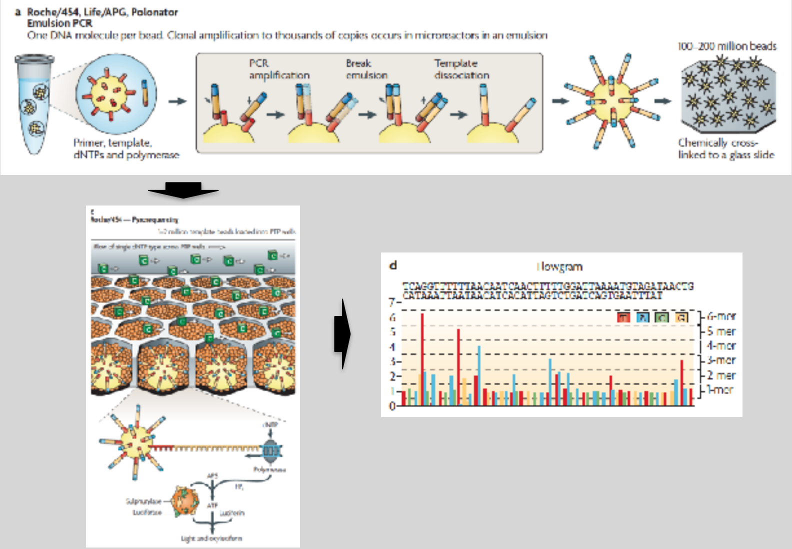
Genome Analysis Toolkit (**GATK**)

- 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!

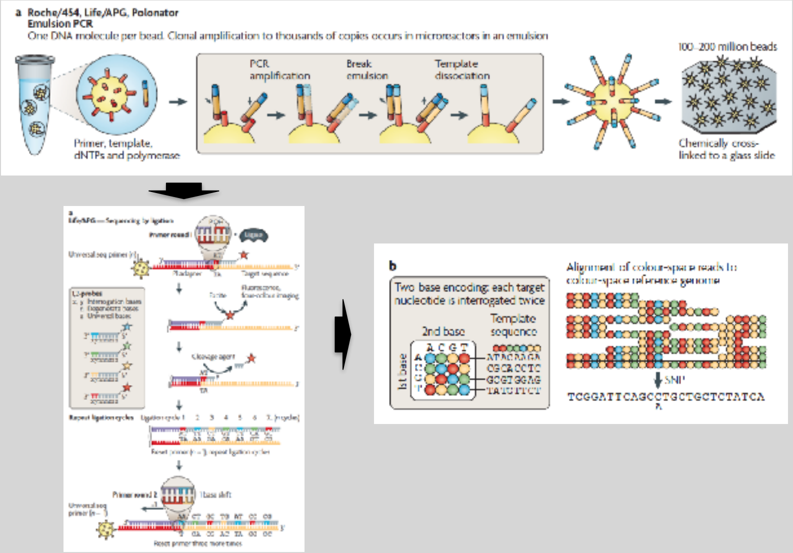
Additional Reference Slides on Sequencing Methods

Roche 454 - Pyrosequencing


a Roche/454, Life/APG, Polonator
Emulsion PCR
One DNA molecule per bead. Clonal amplification to thousands of copies occurs in microreactors in an emulsion
Primer, template, dNTPs and polymerase
PCR amplification
Break emulsion
Template dissociation
100–200 million beads
Chemically cross-linked to a glass slide
c
Roche/454 — Pyrosequencing
1–2 million template beads loaded into PTP wells
Flow of single dNTP type across PTP wells
dNTP
Polymerase
APS
PPi
Sulphurylase
Luciferase
ATP
Luciferin
Light and oxyluciferin
d
Flowgram
TCAGGTTTTTTAACAATCAACTTTTTGGATTAAAATGTAGATAACTG
CATAAATTAATAACATCACATTAGTCTGATCAGTGAATTTAT
7
6
5
4
3
2
1
0
T
A
C
G
6-mer
5 mer
4-mer
3-mer
2 mer
1-mer

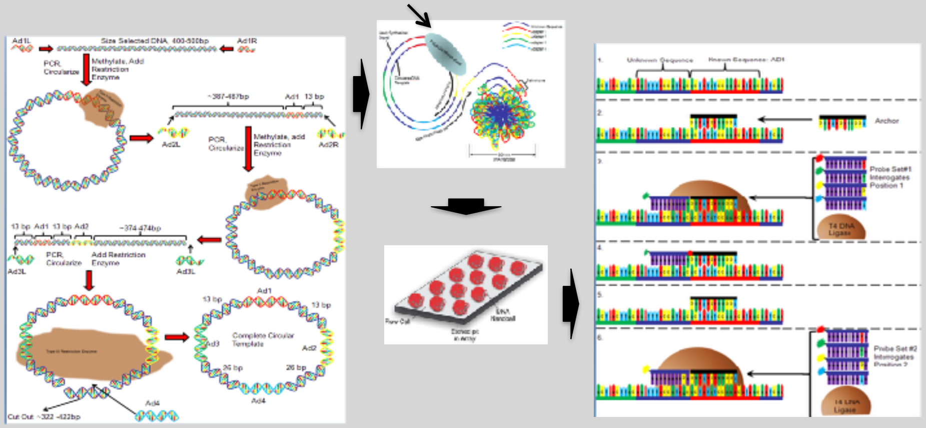
Metzker, ML (2010), *Nat. Rev. Genet*, 11, pp. 31-46

Life Technologies SOLiD – Sequence by Ligation


a Roche/454, Life/APG, Polonator
Emulsion PCR
One DNA molecule per bead. Clonal amplification to thousands of copies occurs in microreactors in an emulsion
Primer, template, dNTPs and polymerase
PCR amplification
Break emulsion
Template dissociation
100–200 million beads
Chemically cross-linked to a glass slide
Life/APG — Sequencing by ligation
Primer round 1
Ligase
Universal seq primer
Ligation cycle 1 2 3 4 5 6 7...
Primer round 2
b
Two base encoding: each target nucleotide is interrogated twice
2nd base
A C G T
1st base
A C G T
Template sequence
AT AC AA GA
CG CA CC TC
GC GT GG AG
TA TC TT CT
Alignment of colour-space reads to colour-space reference genome
SNP
TCGGATTCAGCCTGCTGCTCTATCA
A

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

Complete Genomics – Nanoball Sequencing



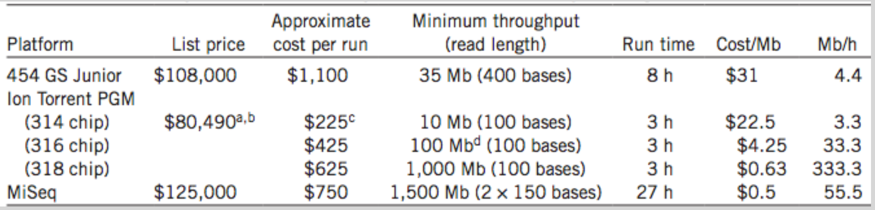
Has proofreading ability!

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

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

"Benchtop" Sequencers

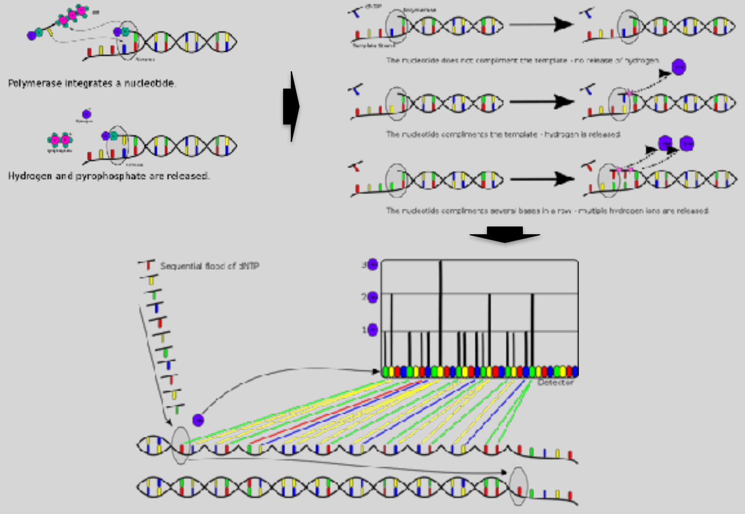
- Lower cost, lower throughput alternative for smaller scale projects
- Currently three significant platforms
  - Roche 454 GS Junior
  - Life Technology Ion Torrent
    - Personal Genome Machine (PGM)
    - Proton
  - Illumina MiSeq



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


Polymerase integrates a nucleotide.
Hydrogen and pyrophosphate are released.
dNTP
Polymerase
Template Strand
The nucleotide does not compliment the template - no release of hydrogen
The nucleotide compliments the template - hydrogen is released
The nucleotide compliments several bases in a row - multiple hydrogen ions are released.
Sequential flood of dNTP
3
2
1
Detector

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