BGGN 213

Genome Informatics I

Lecture 13

Barry Grant

UC San Diego

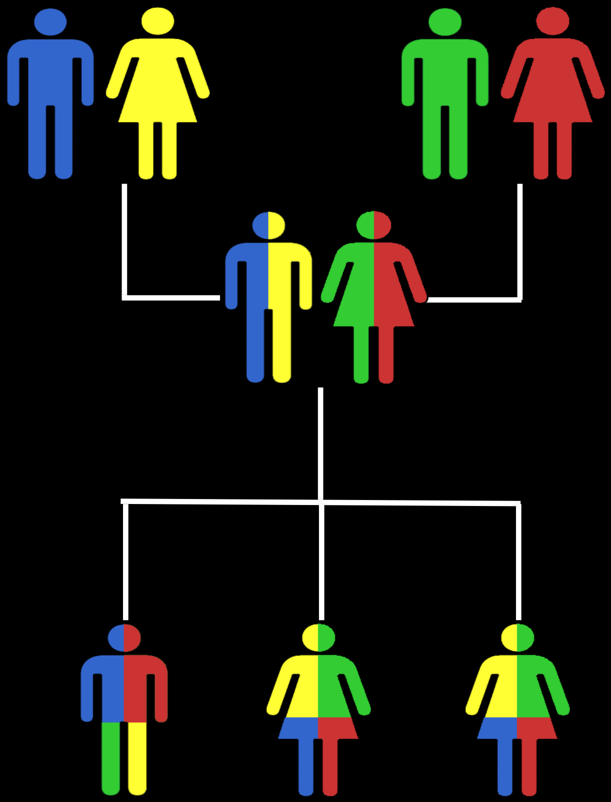
http://thegrantlab.org/bggn213

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!

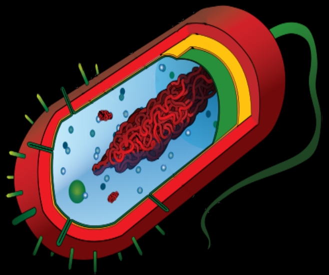
- **Genetics** is primarily the study of *individual genes*, mutations within those genes, and their inheritance patterns in order to understand specific traits.

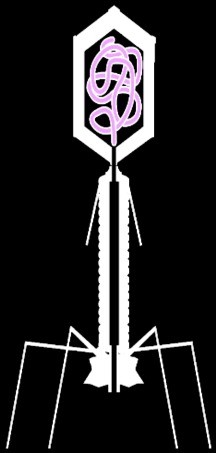
- **Genomics** expands upon classical genetics and considers aspects of the *entire genome*, typically using computer aided approaches.

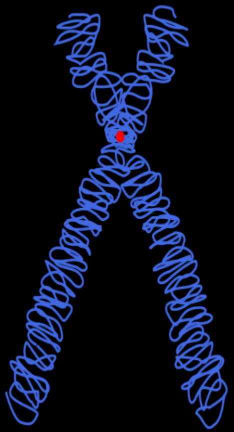
Side note!

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

Side note!

CHROMOSOMES CLOSE-UP

Chromosomes consist largely of double-helical DNA. Cells package the DNA into the nucleus by wrapping it around “spools” composed of histone proteins. The DNA-protein combination is known as chromatin. (Each color represents one chromosome.)

Under a microscope, a Eukaryotic cell's genome (i.e. collection of chromosomes) resembles a chaotic jumble of noodles. The looping is not random however and appears to play a role in controlling gene regulation.

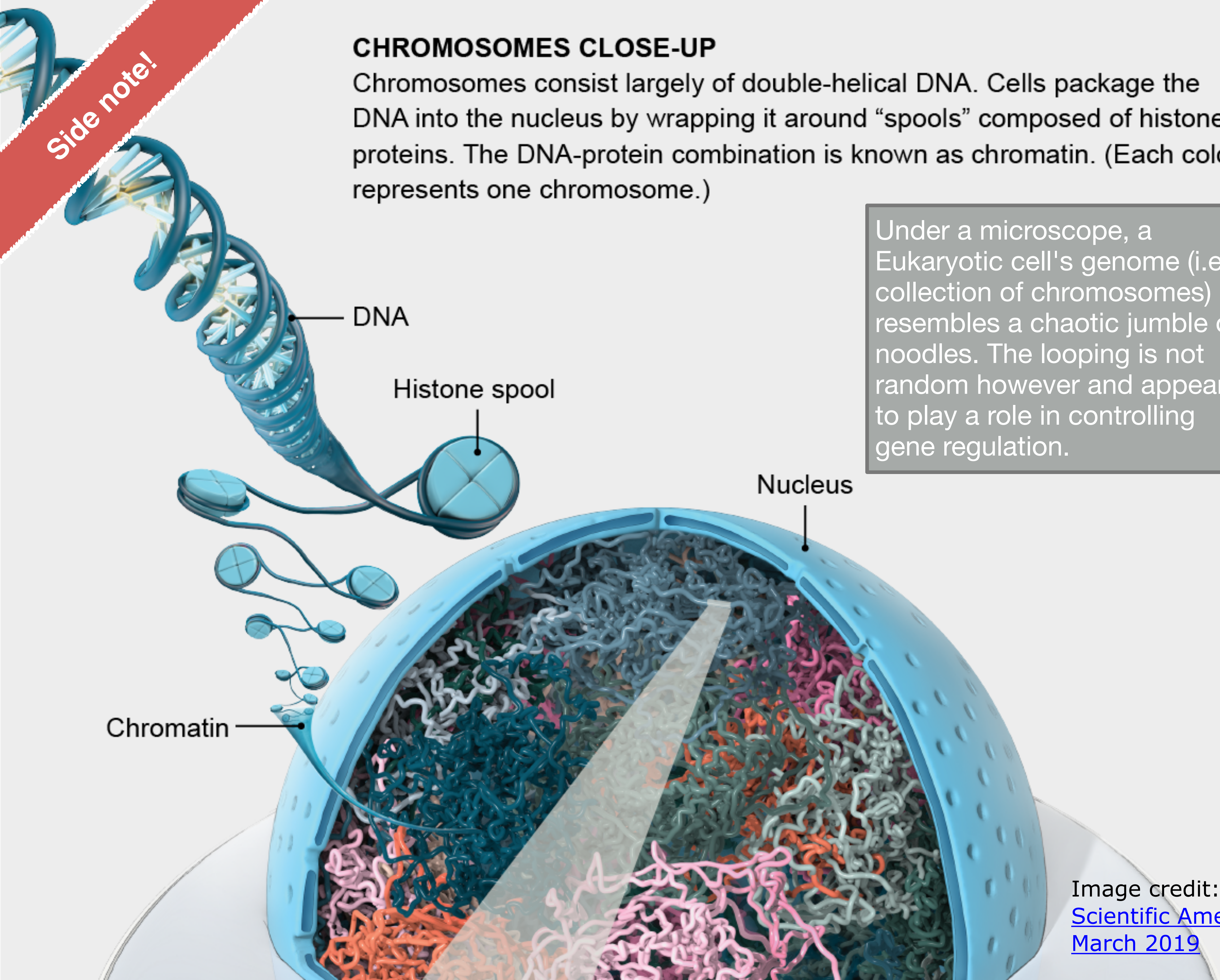
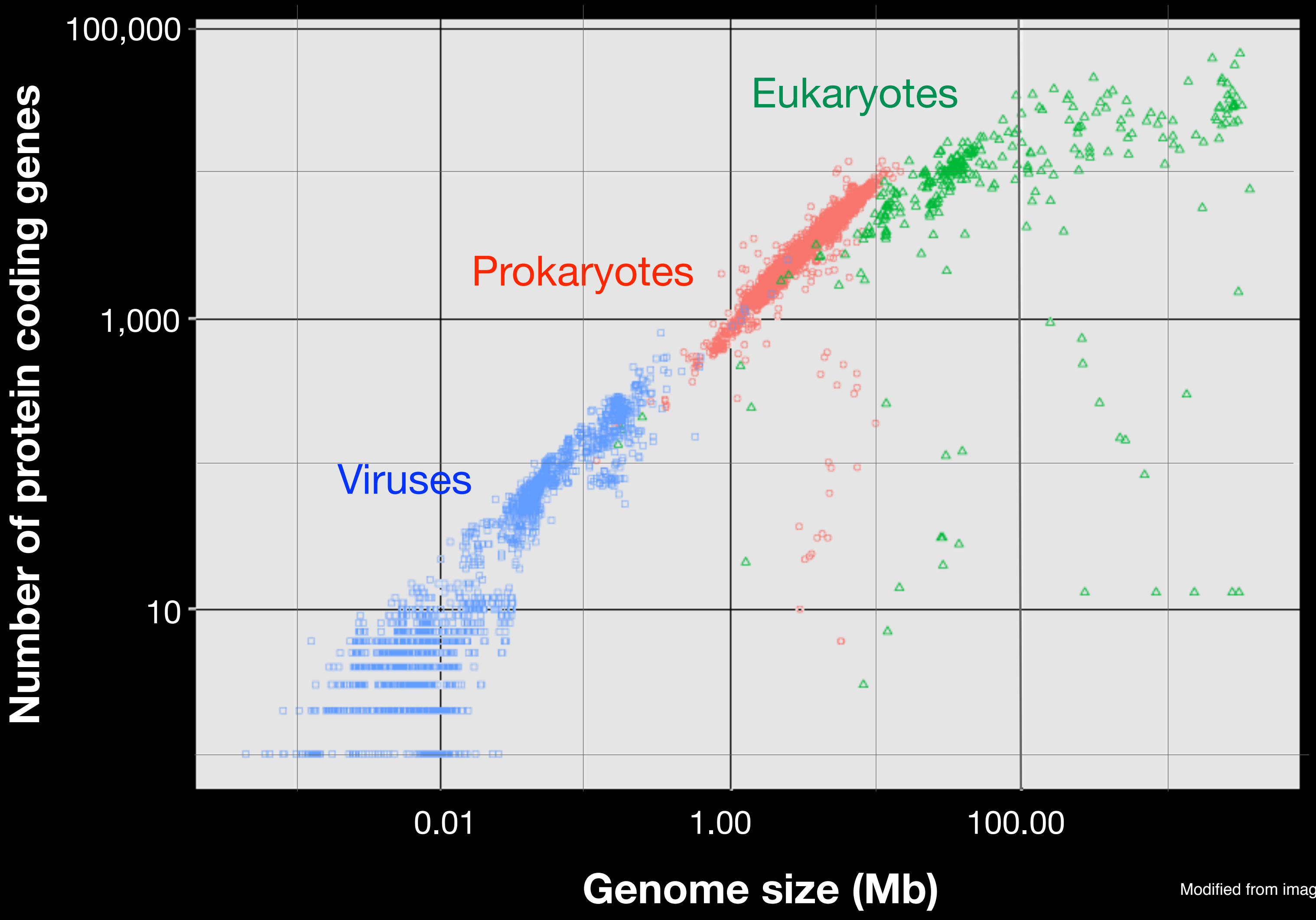


Image credit:
Scientific American March 2019

Genomes come in many sizes

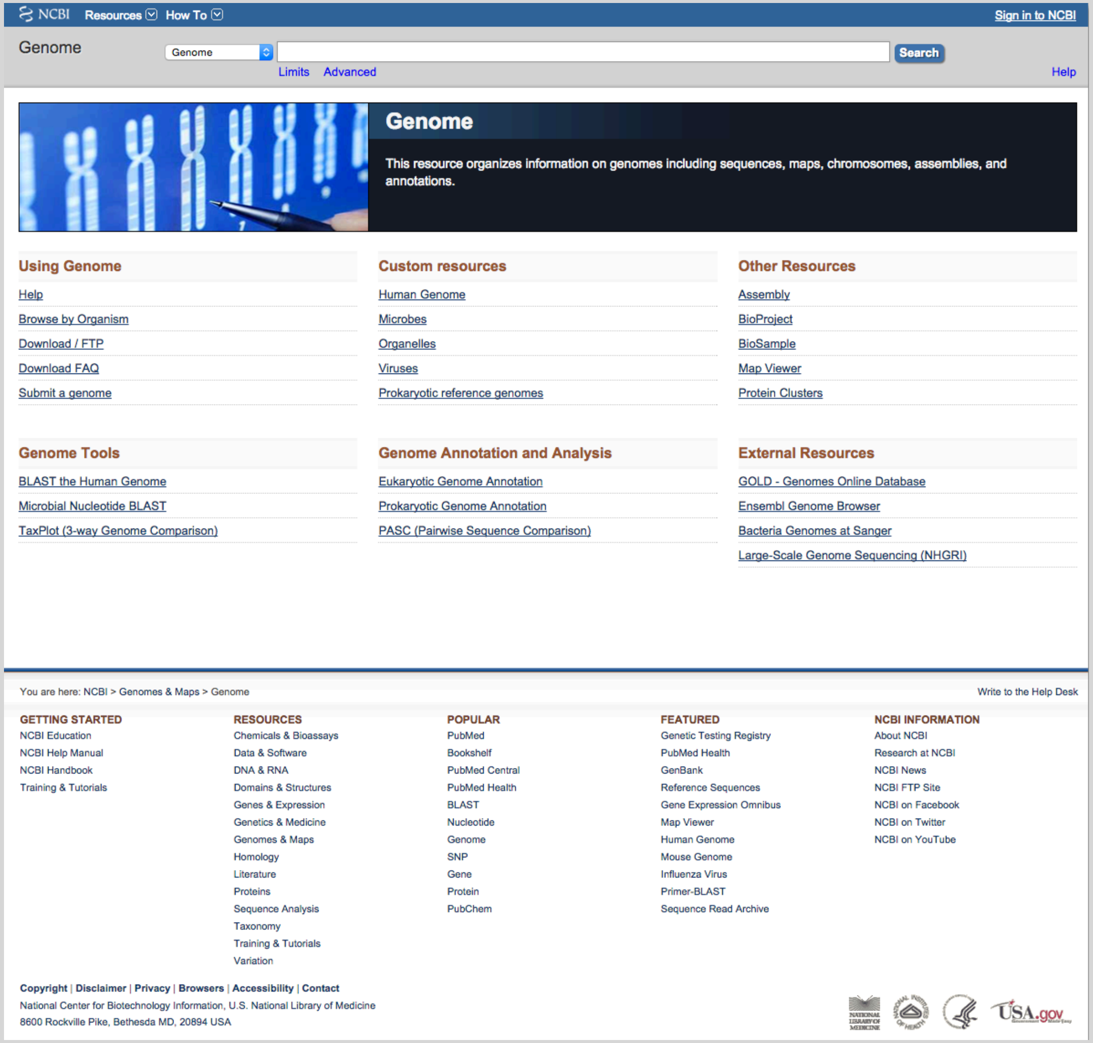


Modified from image by Estevezj / CC BY-SA

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 Comparison)

Genome Annotation and Analysis

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

External Resources

- GOLD - Genomes Online Database
- Ensembl Genome Browser
- Bacteria Genomes at Sanger
- Large-Scale Genome Sequencing (NHGRI)

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

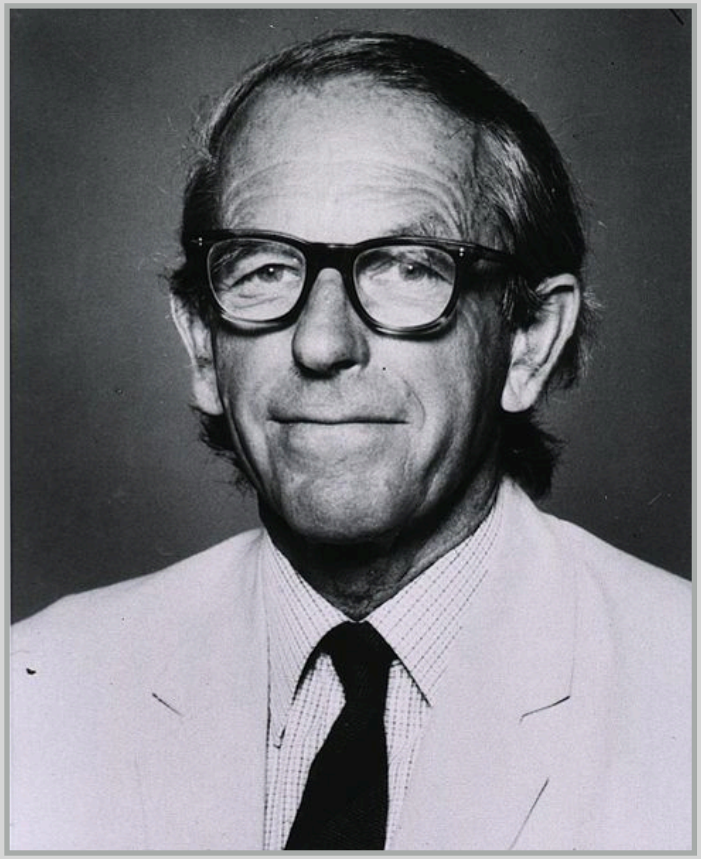
POPULAR
PubMed
Bookshelf
PubMed Central
PubMed Health
BLAST
Nucleotide
Genome
SNP
Gene
Protein
PubChem

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

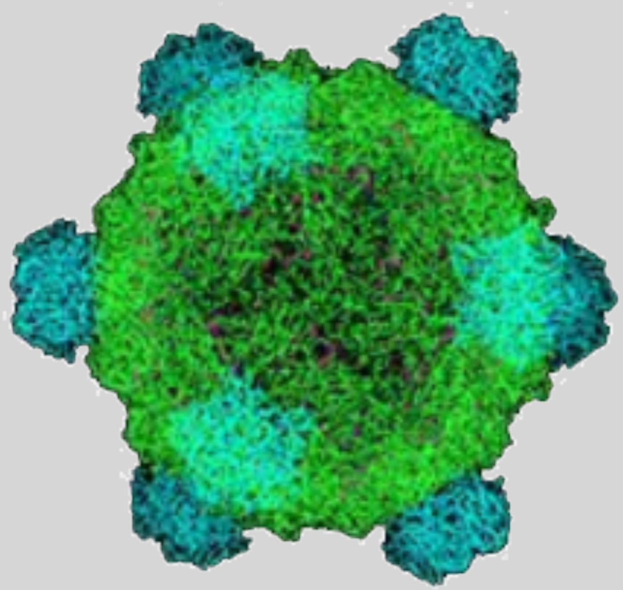


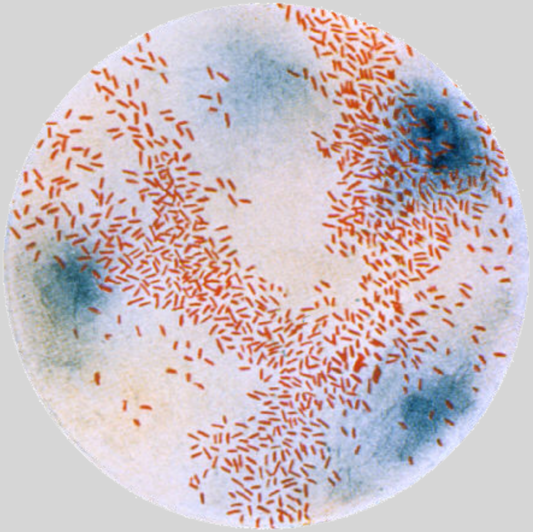
http://en.wikipedia.org/wiki/Frederick_Sanger

- 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

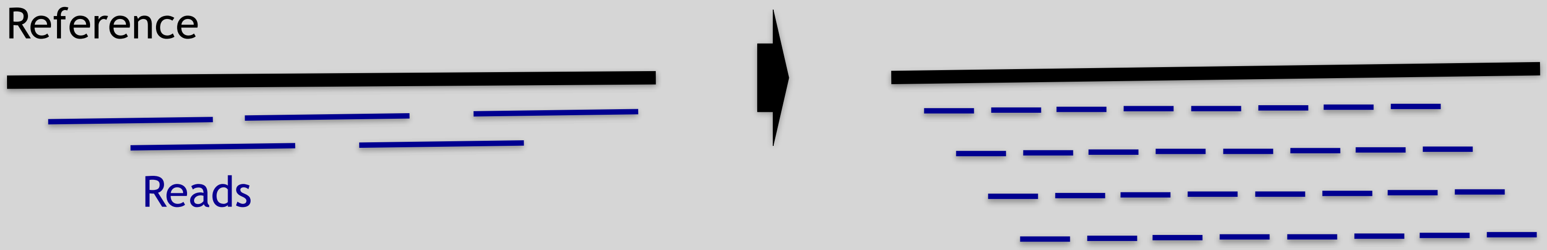
- The Human Genome Project (HGP) was an international, public consortium that began in 1990
  - Initiated by James Watson
  - Primarily led by Francis Collins
  - Eventual Cost: $2.7 Billion
- Celera Genomics was a private corporation that started in 1998
  - Headed by Craig Venter
  - Eventual Cost: $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.

HHMI

Modern Genome Sequencing

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



Rapid progress of genome sequencing

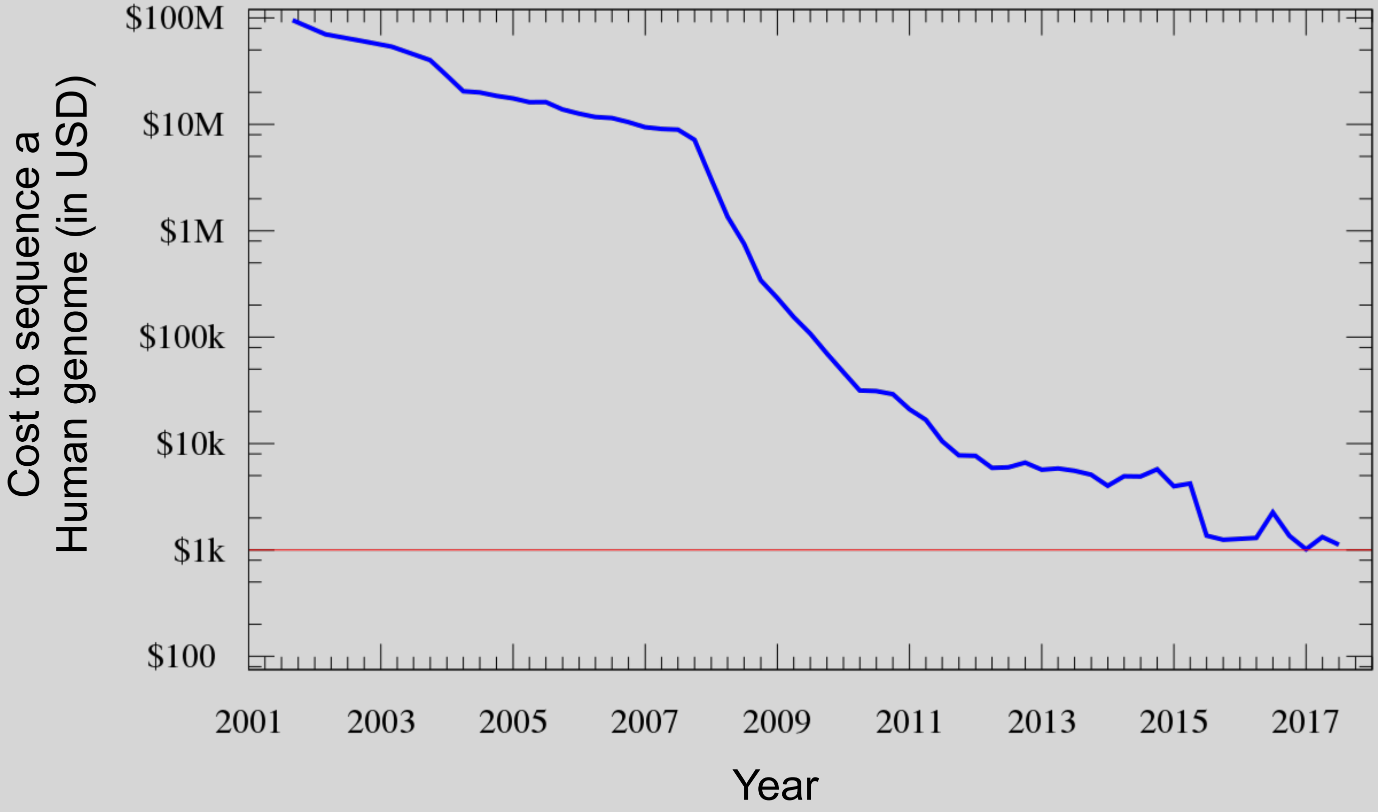
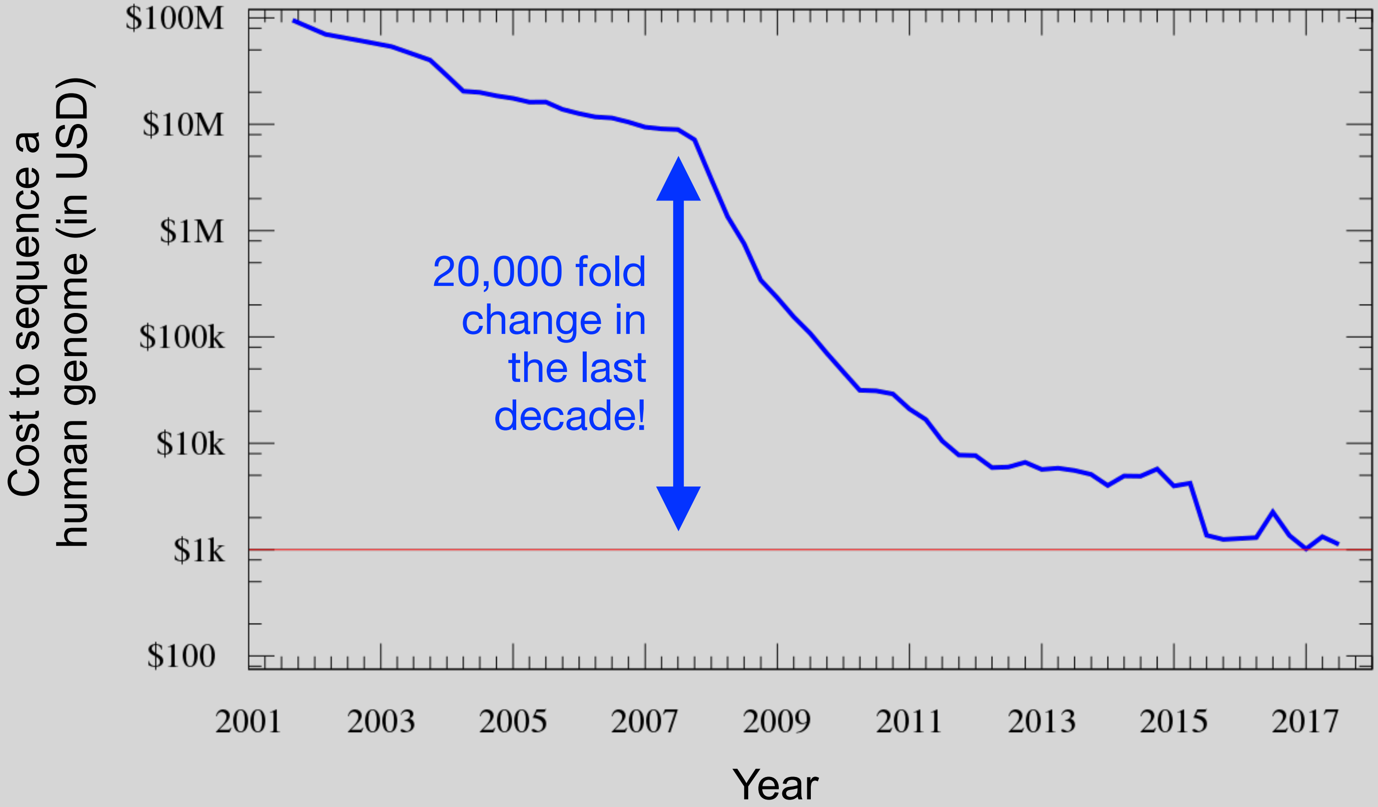


Image source: https://en.wikipedia.org/wiki/Carlson_curve

Rapid progress of genome sequencing

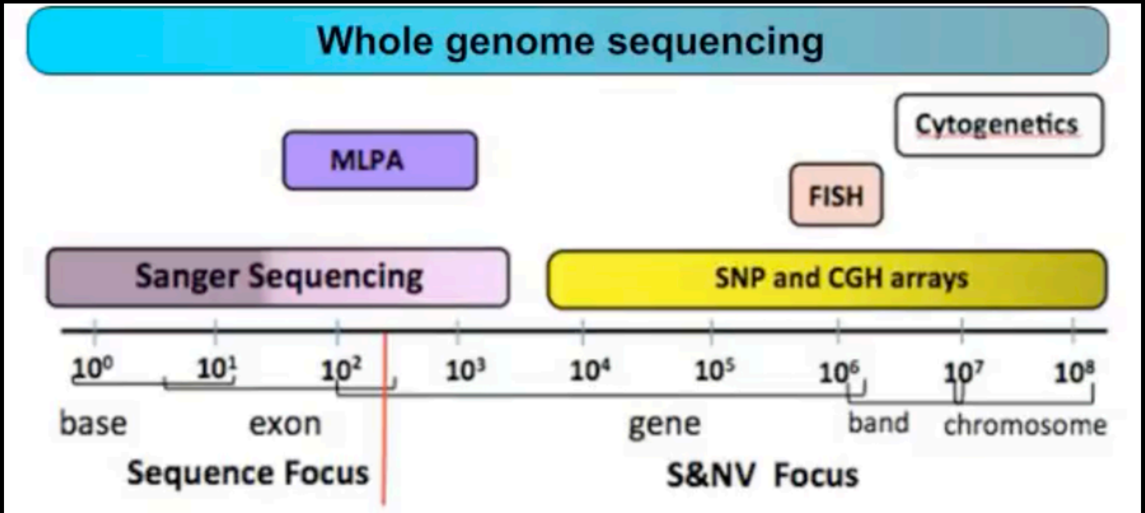


20,000 fold change in the last decade!

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 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

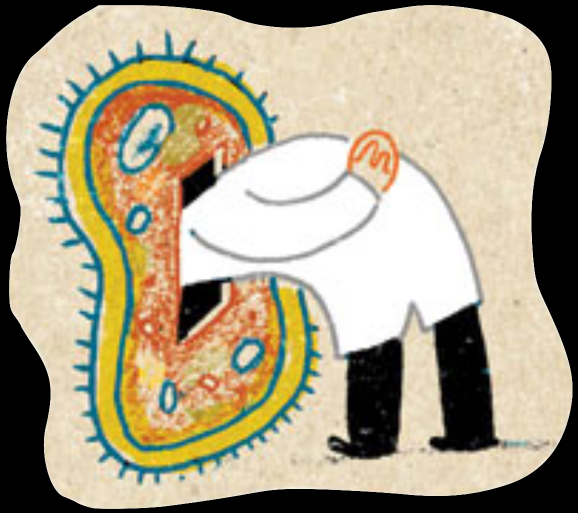
- Identify changes in the genomes of tumors that drive cancer progression

- 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

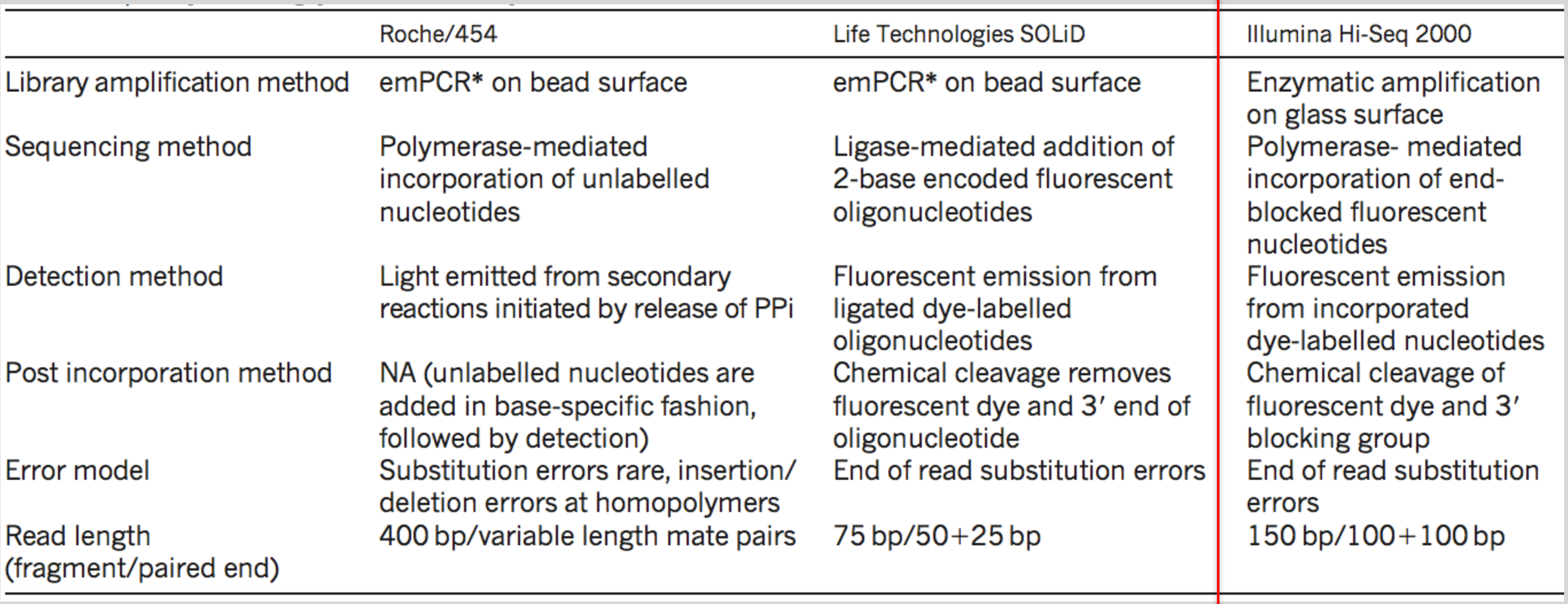
WGS = whole genome sequencing, WXS = whole exome sequencing
RRBS = reduced representation bisulfite sequencing, WGBS = whole genome bisulfite sequencing

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



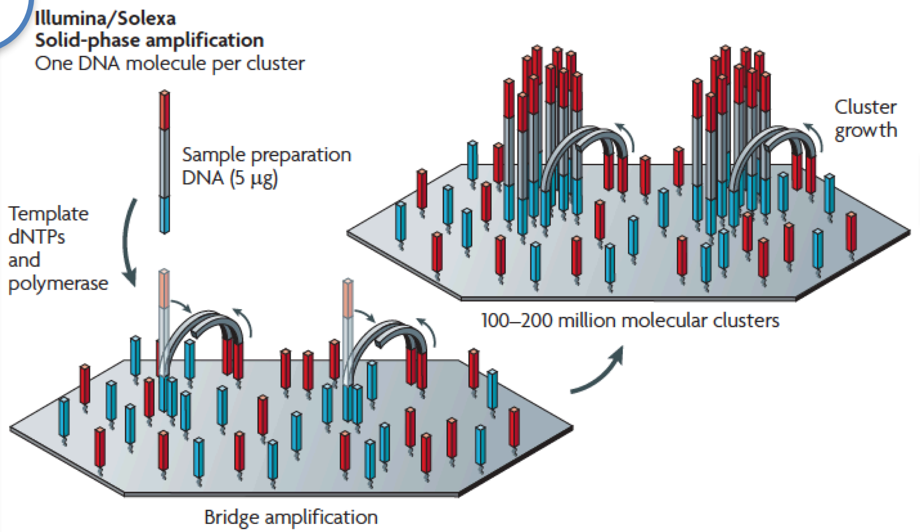
	Roche/454	Life Technologies SOLiD	Illumina Hi-Seq 2000
Library amplification method	emPCR* on bead surface	emPCR* on bead surface	Enzymatic amplification on glass surface
Sequencing method	Polymerase-mediated incorporation of unlabelled nucleotides	Ligase-mediated addition of 2-base encoded fluorescent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

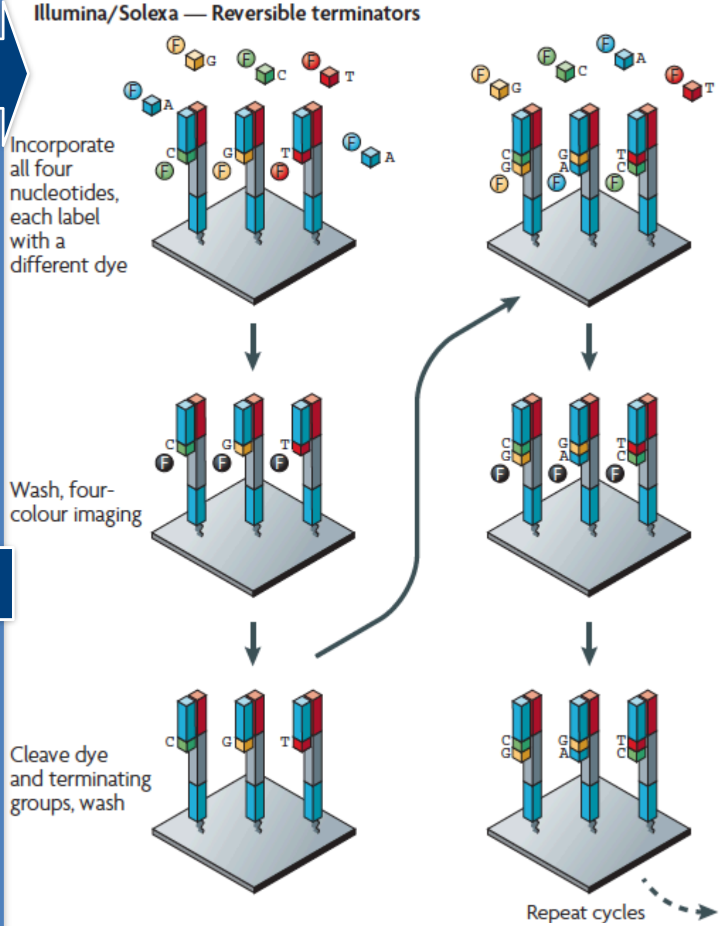
1

Enzymatic amplification on glass surface


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

2

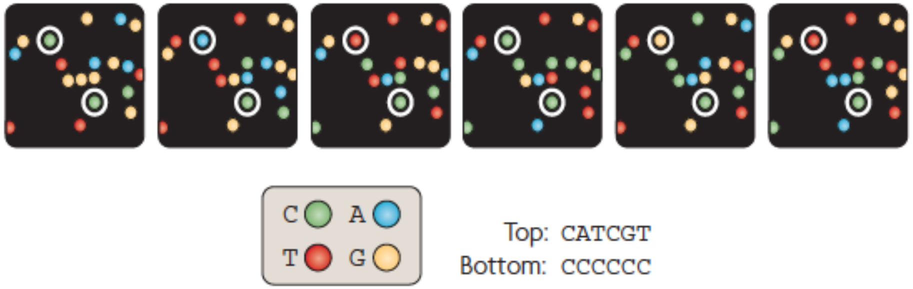
Polymerase-mediated incorporation of end blocked fluorescent nucleotides


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
F
A
G
C
T

Cleave dye & blocking group, repeat…

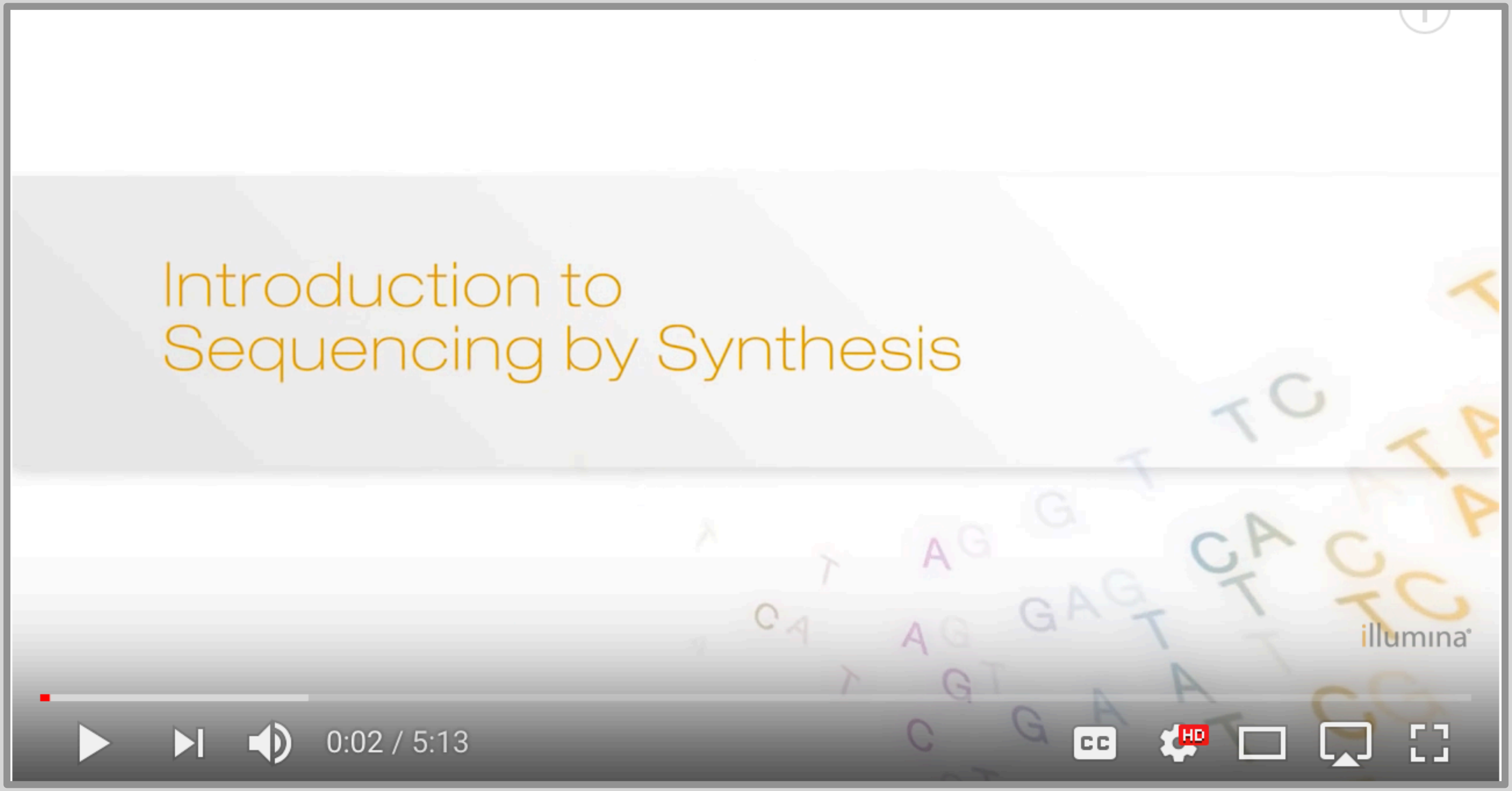
3

Fluorescent emission from incorporated dye-labeled nucleotides



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

Illumina Sequencing - Video

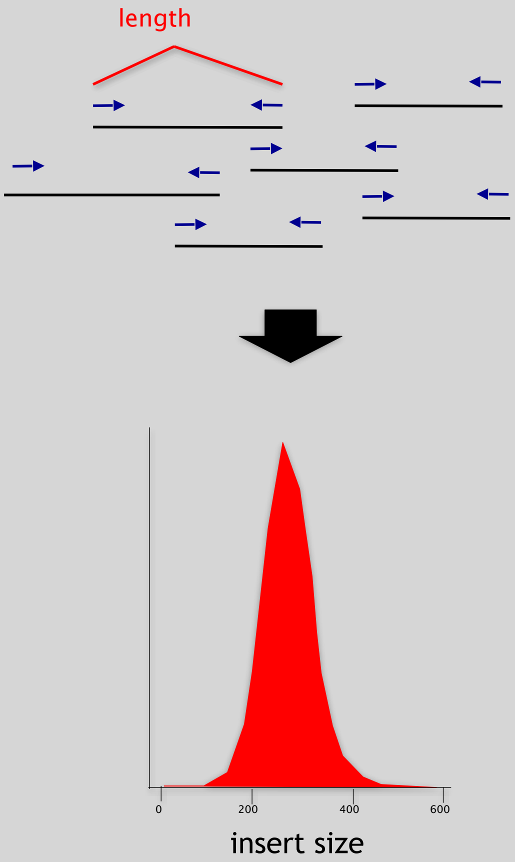


Introduction to Sequencing by Synthesis

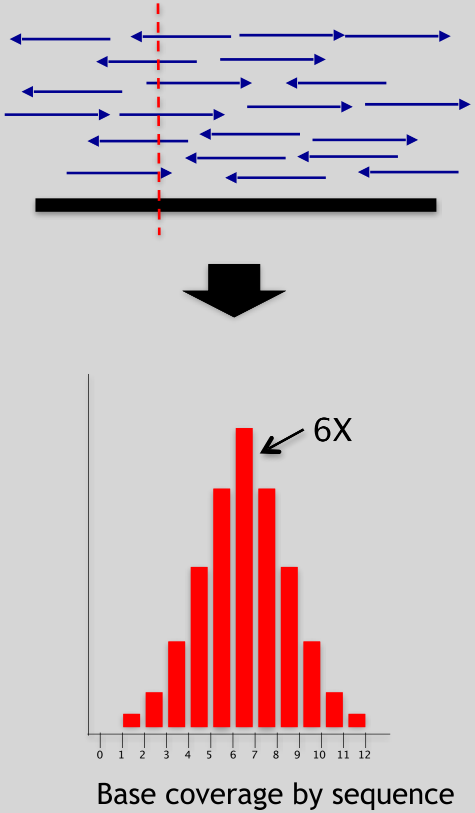
https://www.youtube.com/watch?src_vid=womKfikWlxM&v=fCd6B5HRaZ8

NGS Sequencing Terminology

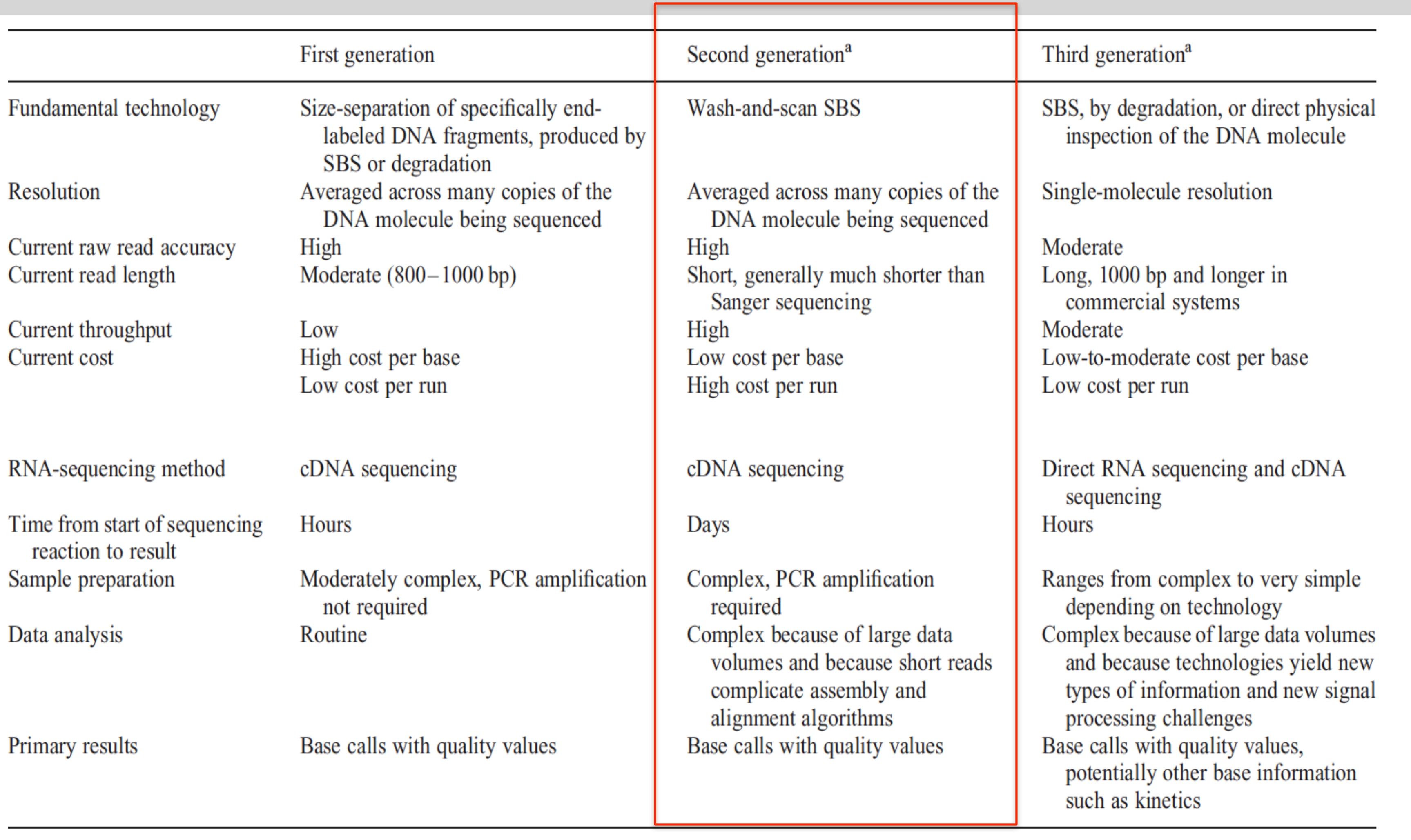
Insert Size



Sequence Coverage



Summary: “Generations” of DNA Sequencing



	First generation	Second generation[a]	Third generation[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:

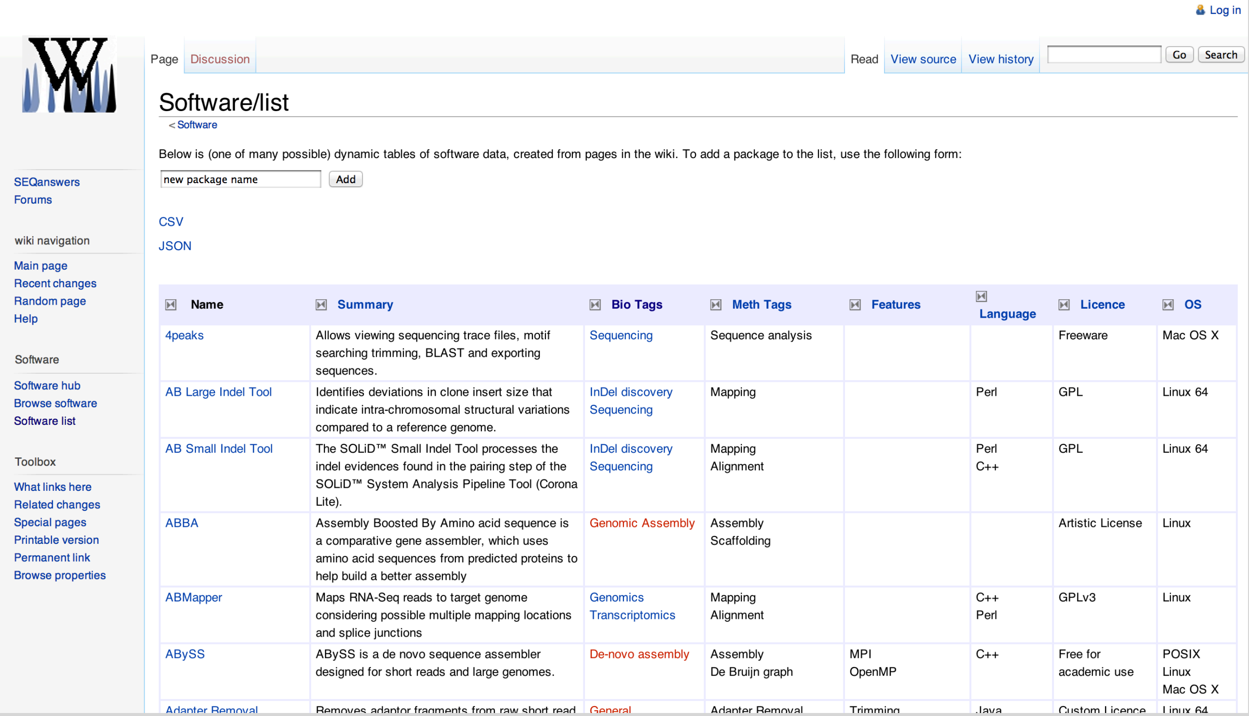
- For example this new nanopore sequencing method was just published!
  https://www.nature.com/articles/nmeth.4577

- "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

Side-Note:

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



Log in

Page | Discussion | Read | View source | View history | Go | Search

SEQanswers
Forums

wiki navigation

- Main page
- Recent changes
- Random page
- Help

Software

- Software hub
- Browse software
- Software list

Toolbox

- What links here
- Related changes
- Special pages
- Printable version
- Permanent link
- Browse properties

Software/list

< Software

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

new package name | Add

CSV

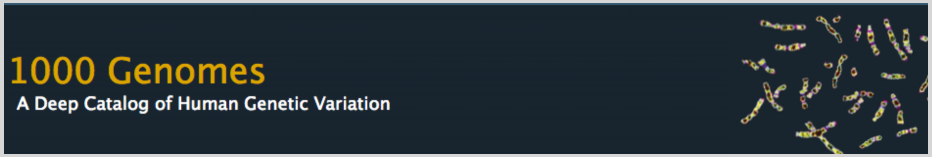
JSON

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

What can we do with all this sequence information?

Population Scale Analysis

We can now begin to assess genetic differences on a very large scale, both as naturally occurring variation in human and non-human populations as well somatically within tumors



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

“Variety’s the very spice of life”

–William Cowper, 1785

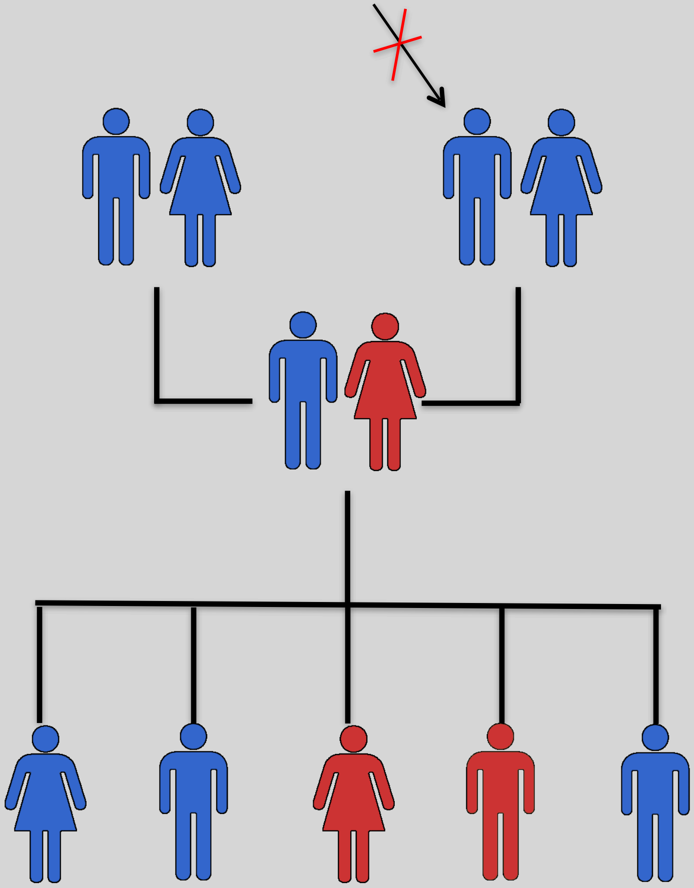
“Variation is the spice of life”

–Kruglyak & Nickerson, 2001

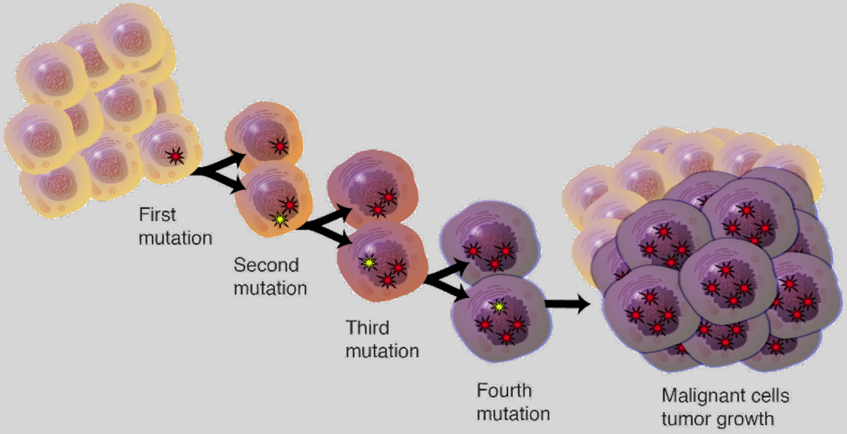
- While the sequencing of the human genome was a great milestone, the DNA from a single person is not representative of the millions of potential differences that can occur between individuals
- These unknown genetic variants could be the cause of many phenotypes such as differing morphology, susceptibility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

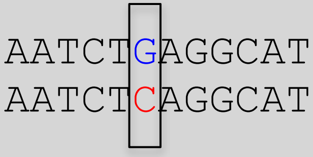
- Can occur during mitosis or from the environment itself

- Are an integral part in tumor progression and evolution

Darryl Leja, Courtesy: National Human Genome Research Institute.

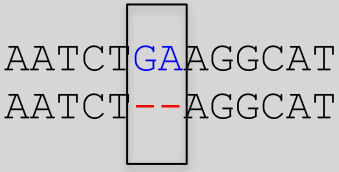
Types of Genomic Variation

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



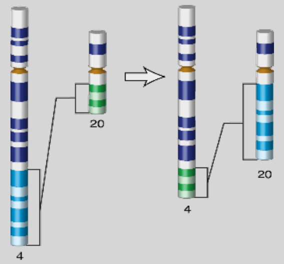
AATCTGAGGCAT
AATCTCAGGCAT

- **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

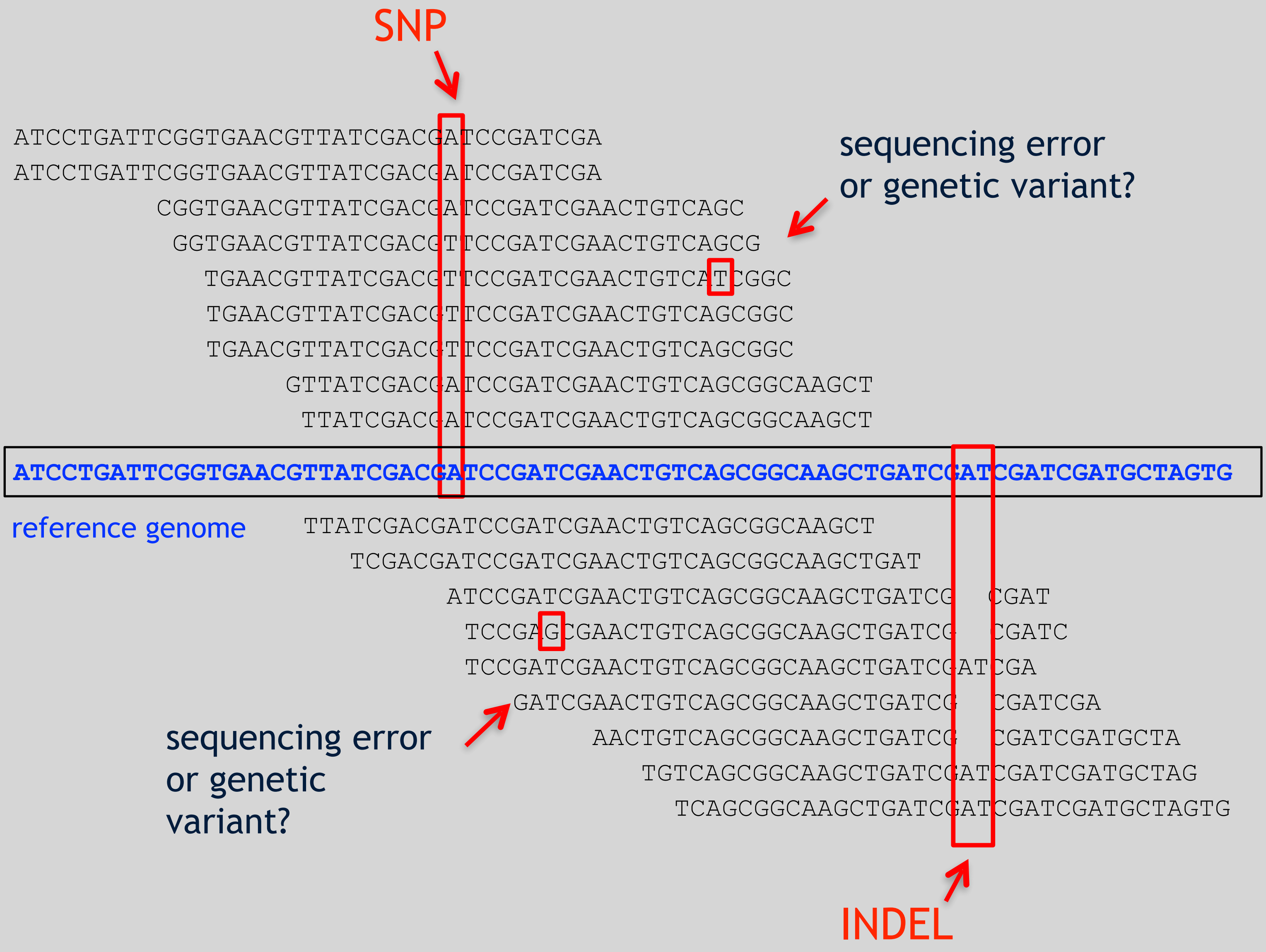
The average number of genetic differences in the germline between two random humans can be broken down as follows:

- 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



SNP

```
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?

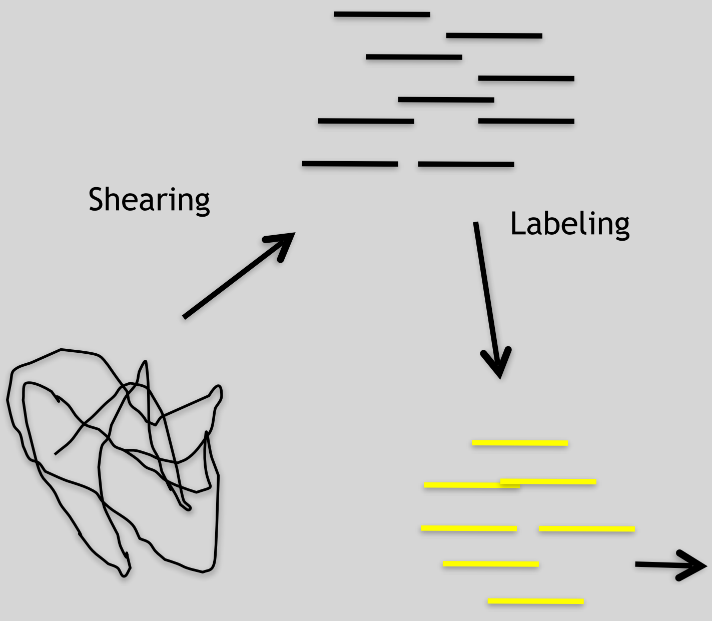
sequencing error or genetic variant?

INDEL

Genotyping Small Variants

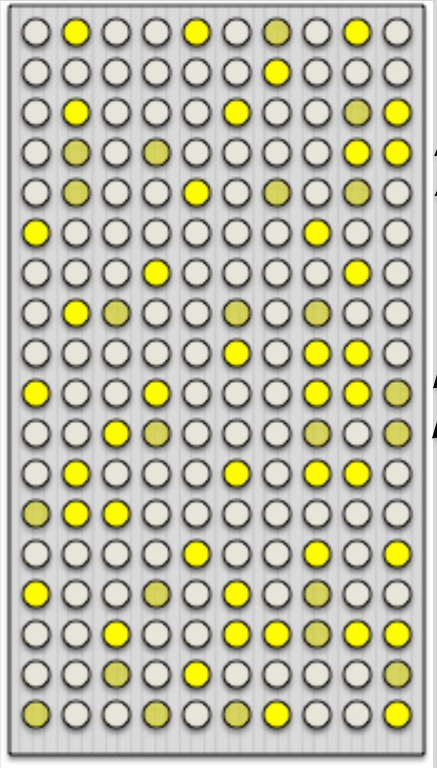
- Once discovered, oligonucleotide probes can be generated with each individual allele of a variant of interest
- A large number can then be assessed simultaneously on microarrays to detect which combination of alleles is present in a sample

SNP Microarrays



Shearing

Labeling



TAACGATGAATC**T**TAGGCATCGCGC

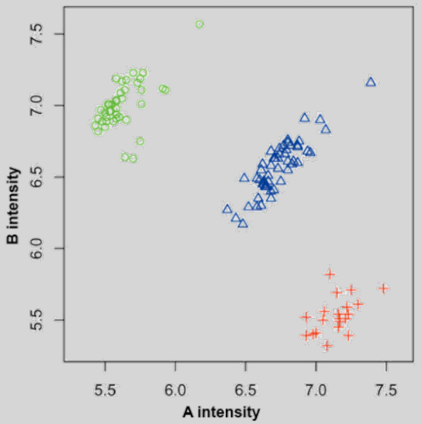
TAACGATGAATC**G**TAGGCATCGCGC

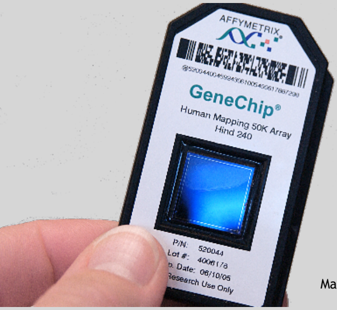
genotype: T/T

GGCTTAAGTACCC**C**TATGGATTACGG

GGCTTAAGTACCC**T**TATGGATTACGG

genotype: C/T



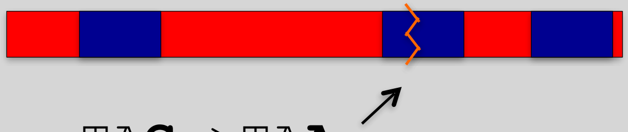


Maggie Bartlett, Courtesy: National Human Genome Research Institute.

Impact of Genetic Variation

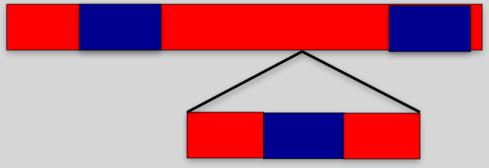
There are numerous ways genetic variation can exhibit functional effects

Premature stop codons

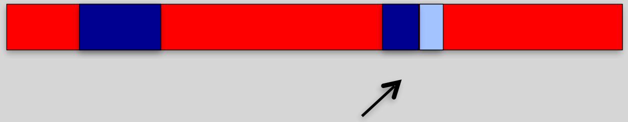


TA**C**->TA**A**

Gene or exon deletion

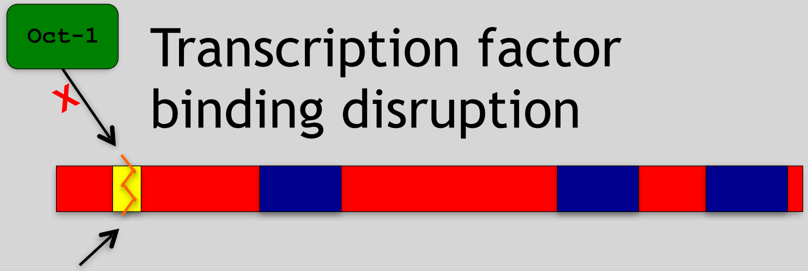


Frameshift mutation



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

Transcription factor binding disruption



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

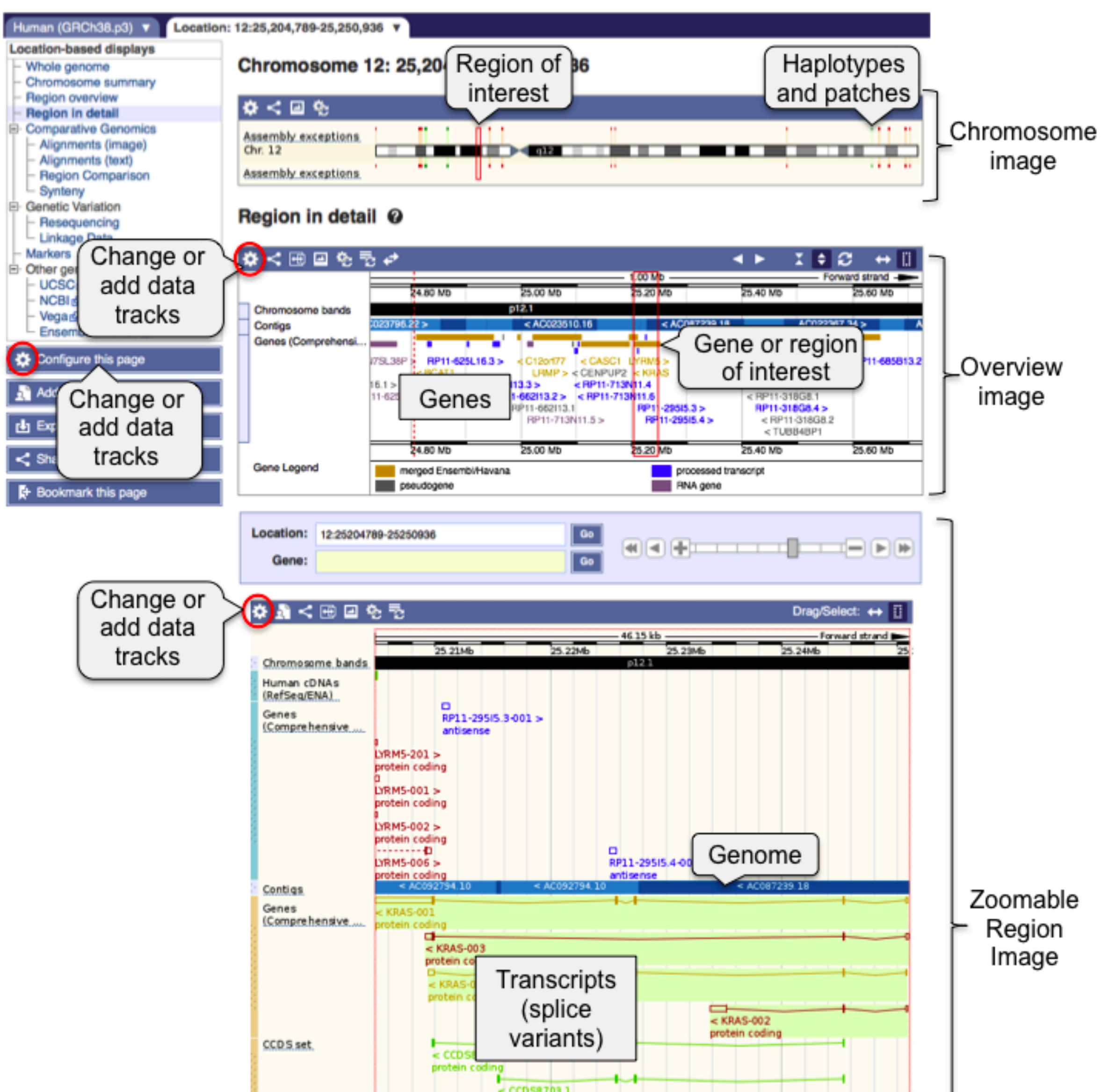
Do it Yourself!

Hand-on time!

https://bioboot.github.io/bggn213_W19/lectures/#13

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

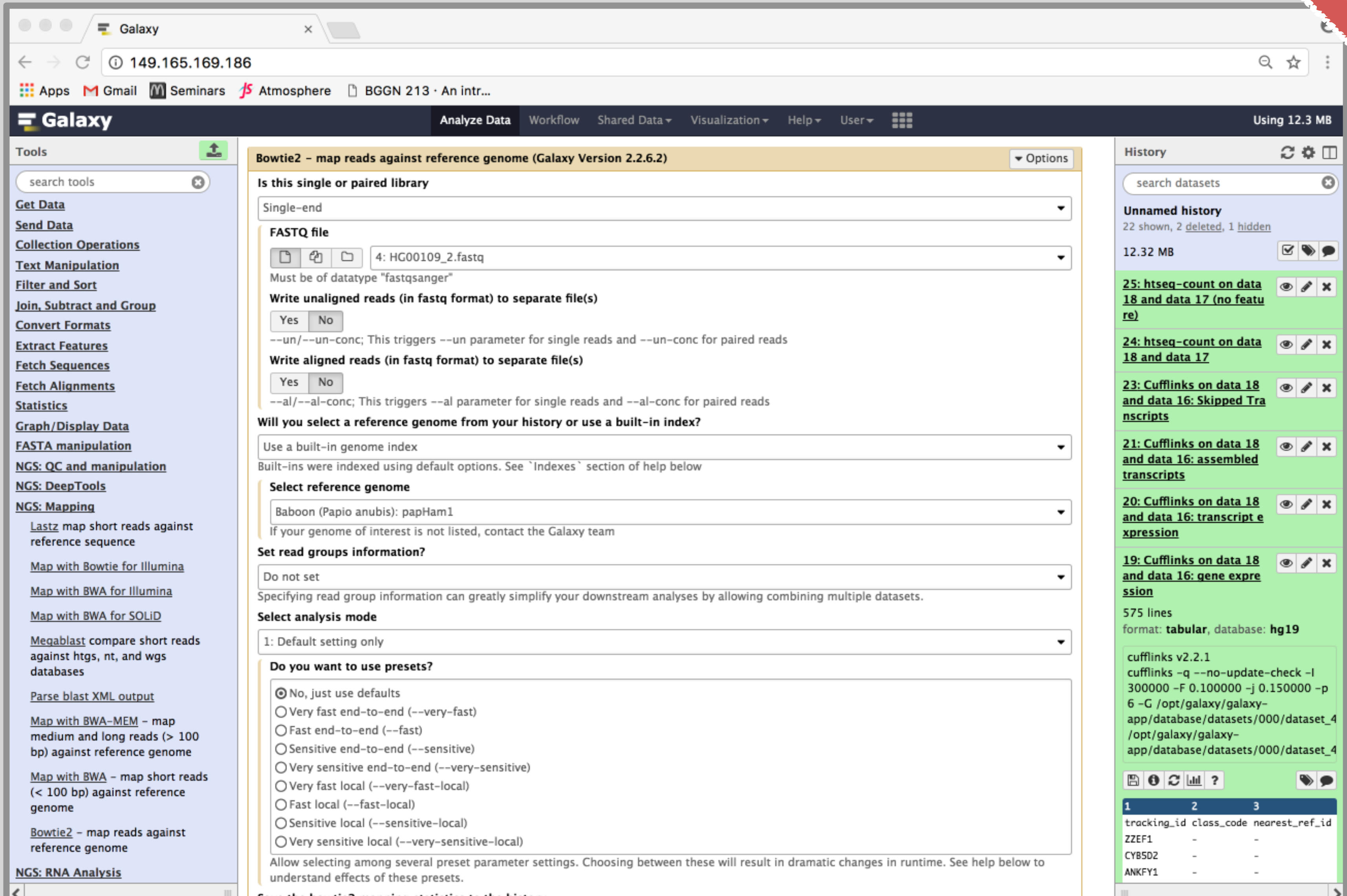
http://uswest.ensembl.org/Help/View?id=140



Access a jetstream galaxy instance!

Use assigned IP address

Do it Yourself!



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 read group 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
22 shown, 2 deleted, 1 hidden
12.32 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	-	-

Raw data usually in **FASTQ format**

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

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…

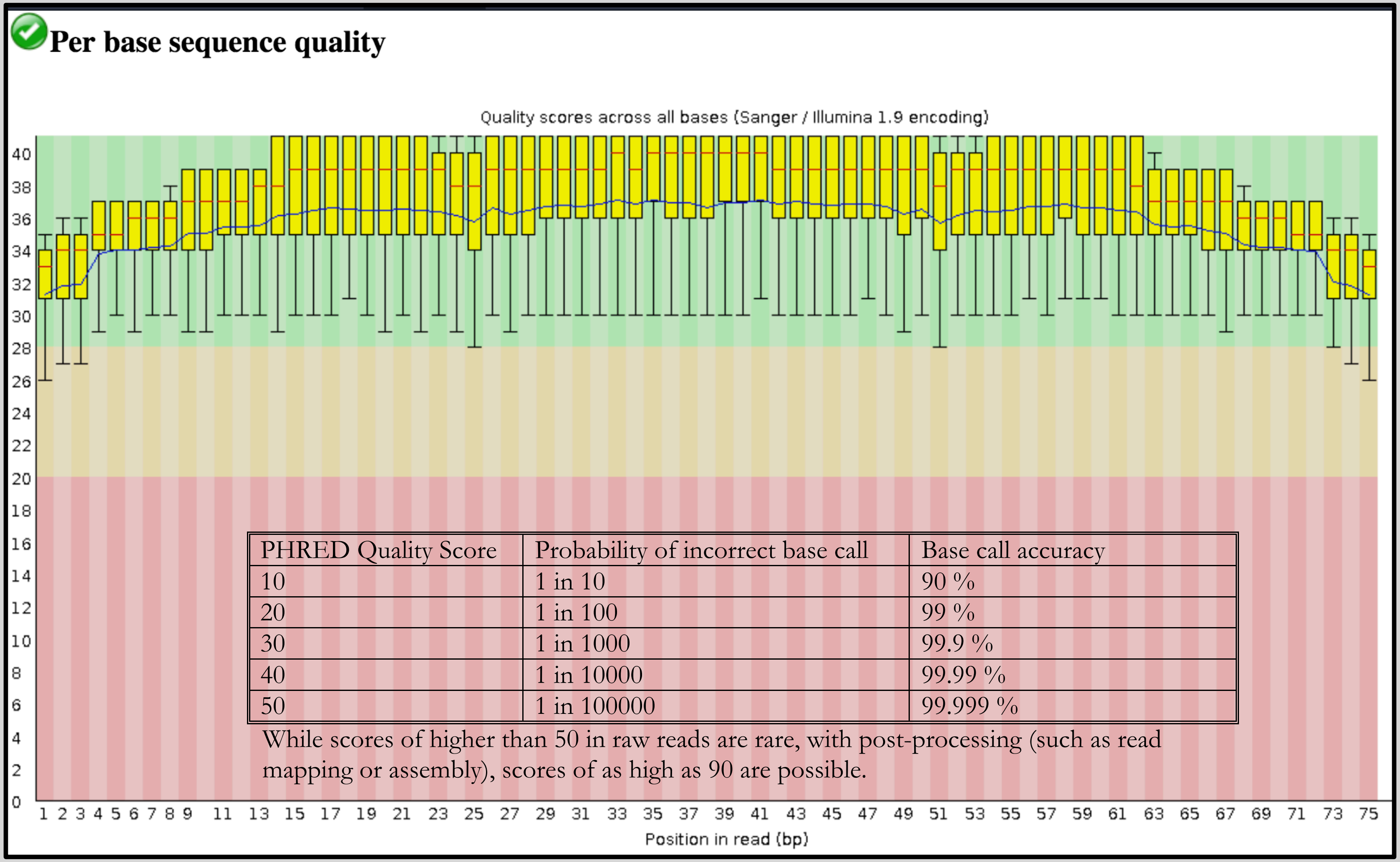
Interpreting Base Qualities in R

		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



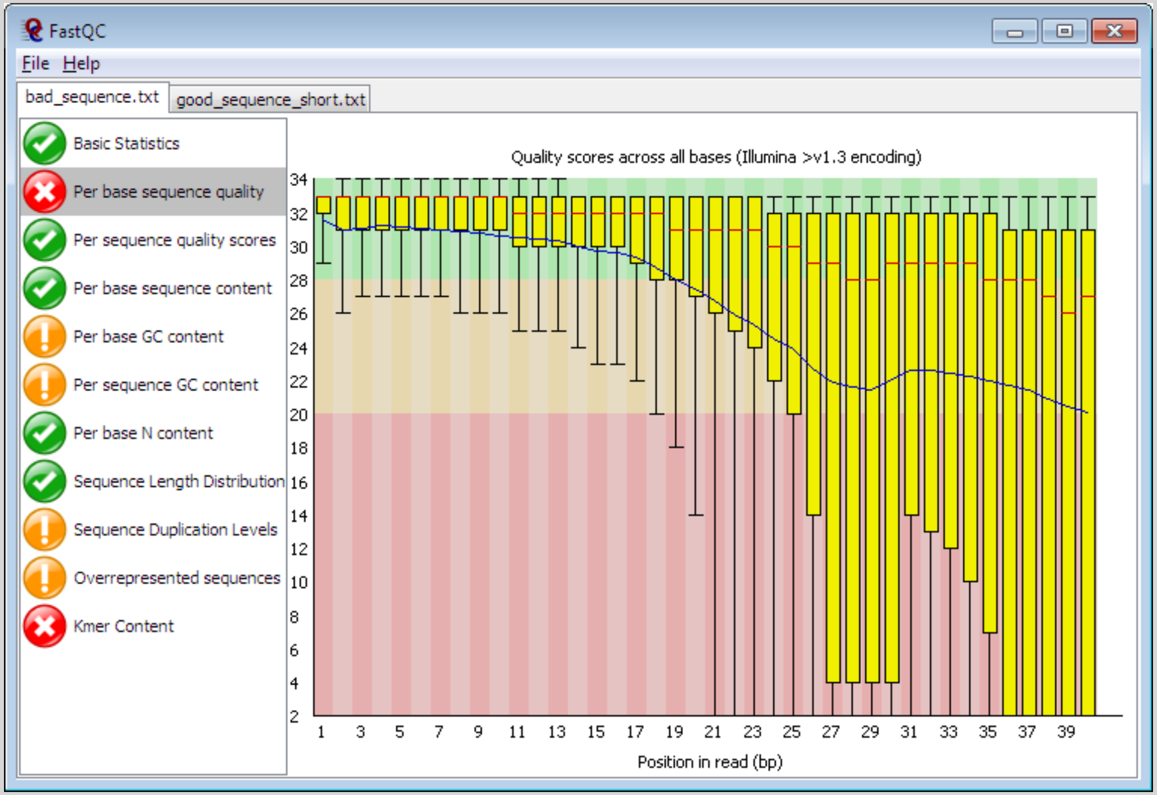
PHRED Quality Score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90 %
20	1 in 100	99 %
30	1 in 1000	99.9 %
40	1 in 10000	99.99 %
50	1 in 100000	99.999 %

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/


FastQC
File Help
bad_sequence.txt
good_sequence_short.txt
✓ Basic Statistics
✗ Per base sequence quality
✓ Per sequence quality scores
✓ Per base sequence content
! Per base GC content
! Per sequence GC content
✓ Per base N content
✓ Sequence Length Distribution
! Sequence Duplication Levels
! Overrepresented sequences
✗ Kmer Content
Quality scores across all bases (Illumina >v1.3 encoding)
34 32 30 28 26 24 22 20 18 16 14 12 10 8 6 4 2
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35 37 39
Position in read (bp)

Sequence Alignment

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

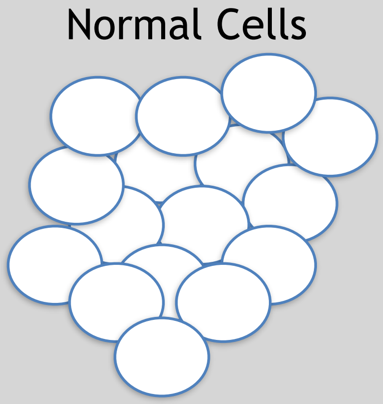
BWA
Bowtie
SOAP2
Novoalign
mr/mrsFast
Eland
Blat
Bfast

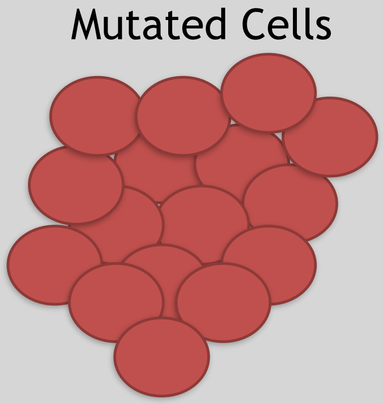
BarraCUDA
CASHx
GSNAP
Mosiak
Stampy
SHRiMP
SeqMap
SLIDER

RMAP
SSAHA
etc

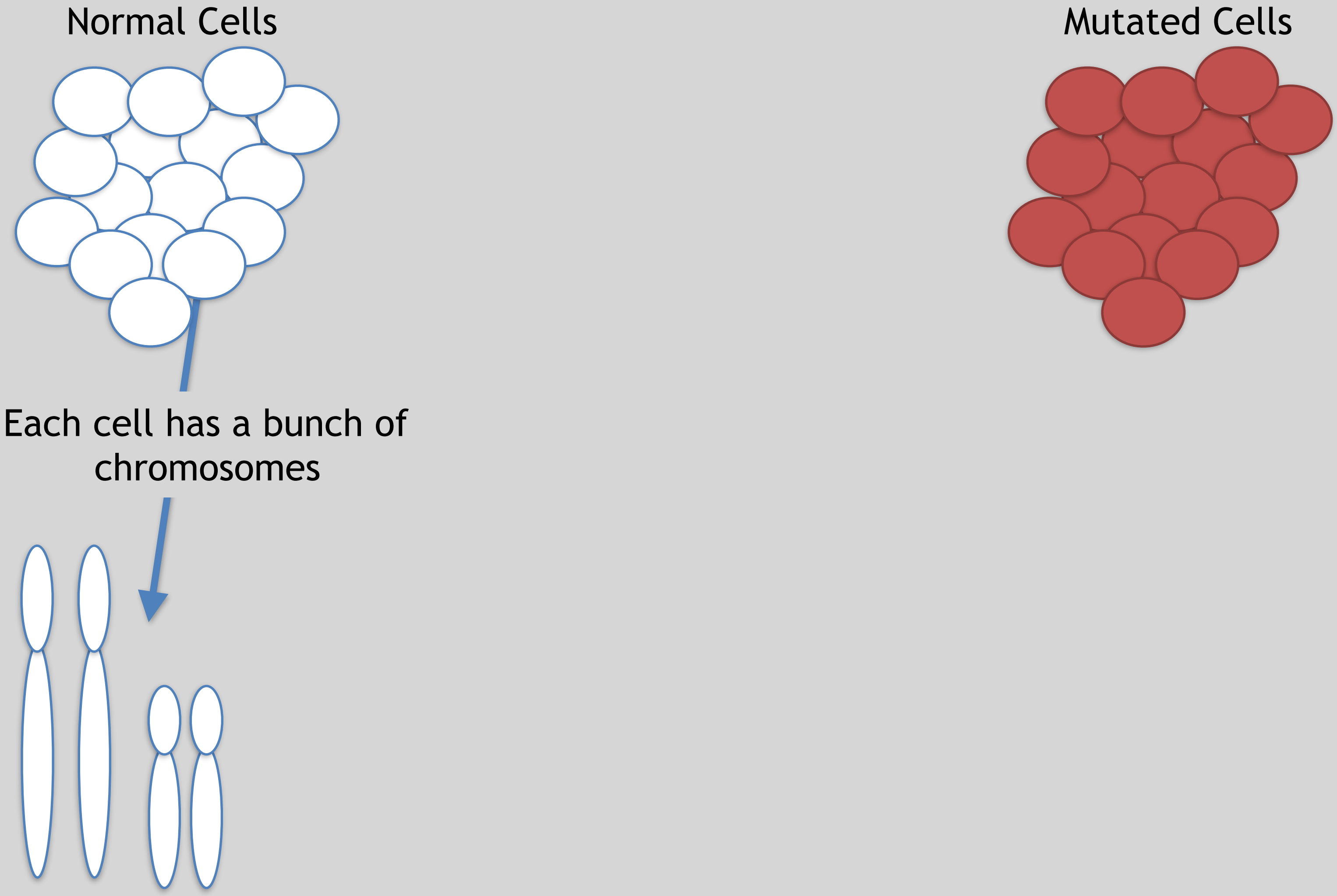
RNA Sequencing

The absolute basics





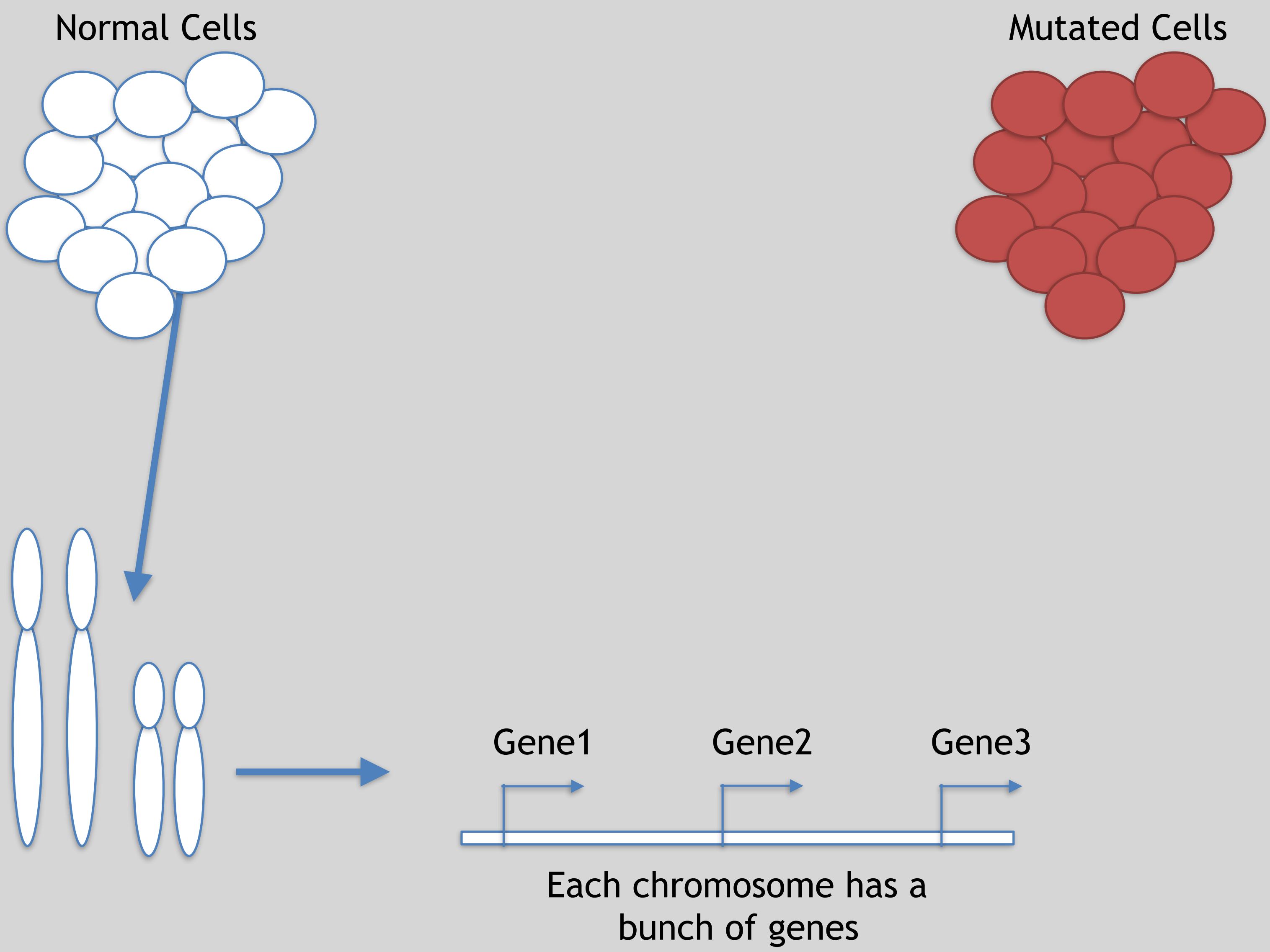
- The mutated cells behave differently than the normal cells
- We want to know what genetic mechanism is causing the difference
- One way to address this is to examine differences in gene expression via RNA sequencing…



Normal Cells

Mutated Cells

Each cell has a bunch of chromosomes



Normal Cells

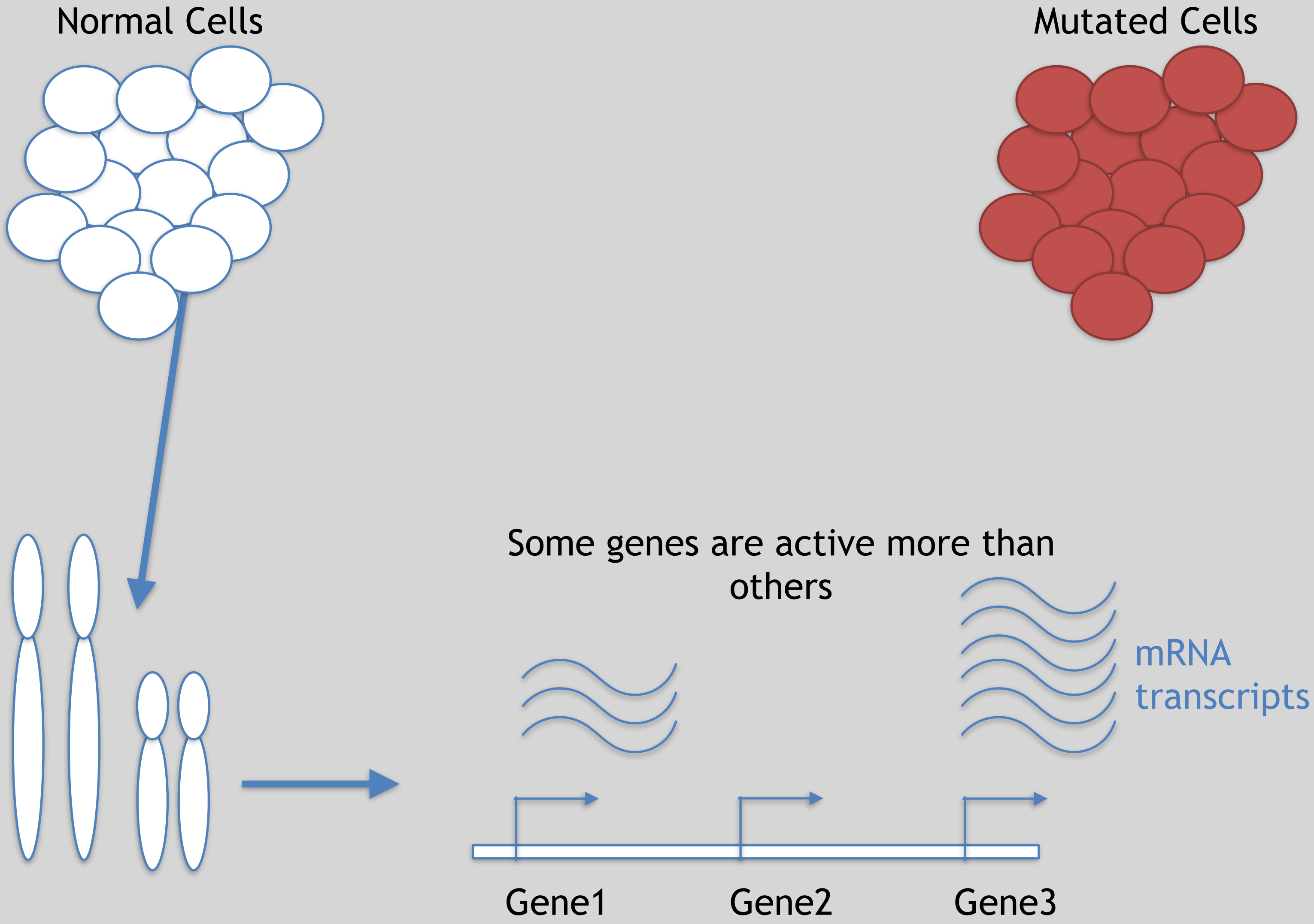
Mutated Cells

Gene1

Gene2

Gene3

Each chromosome has a bunch of genes



Normal Cells

Mutated Cells

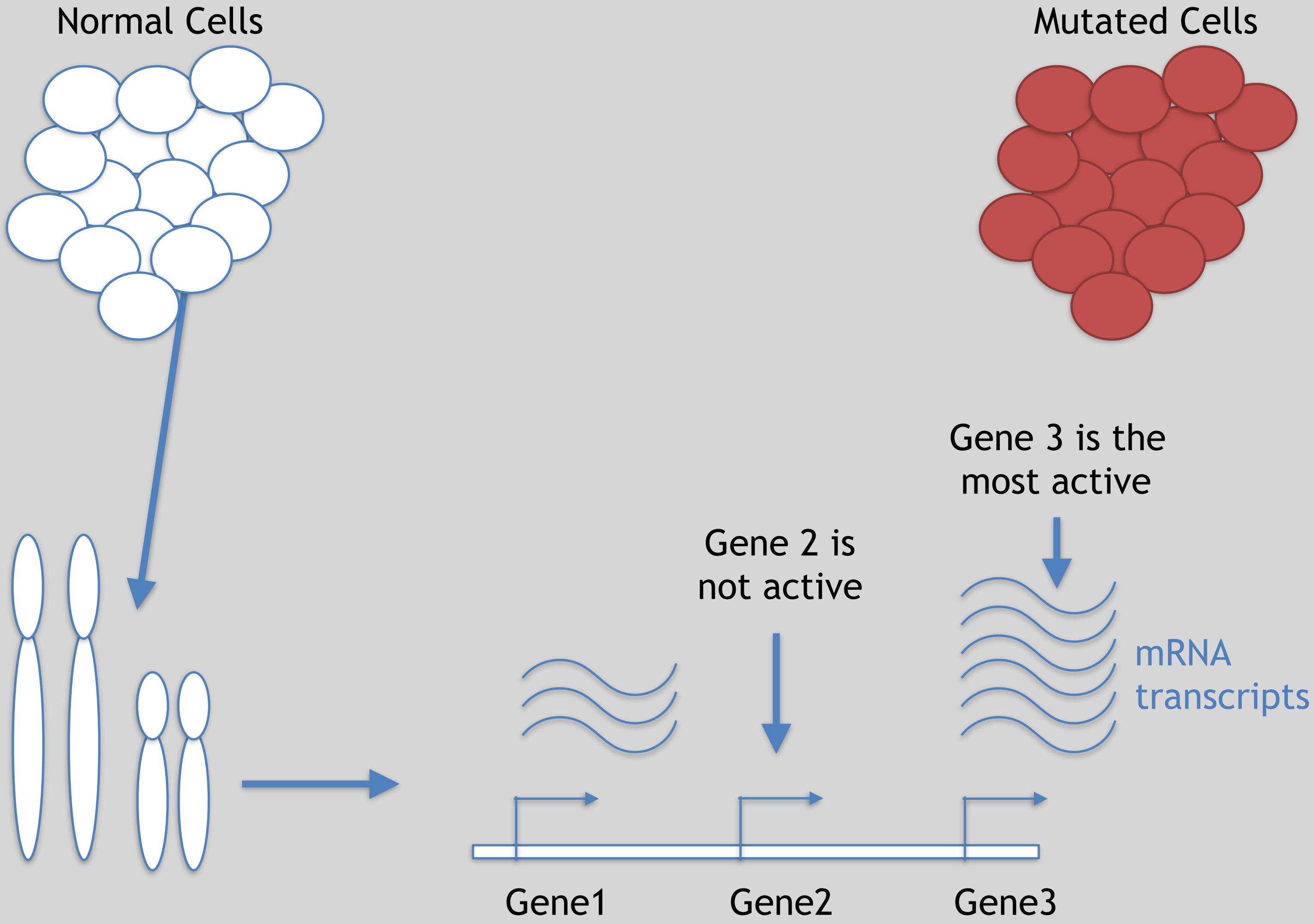
Some genes are active more than others

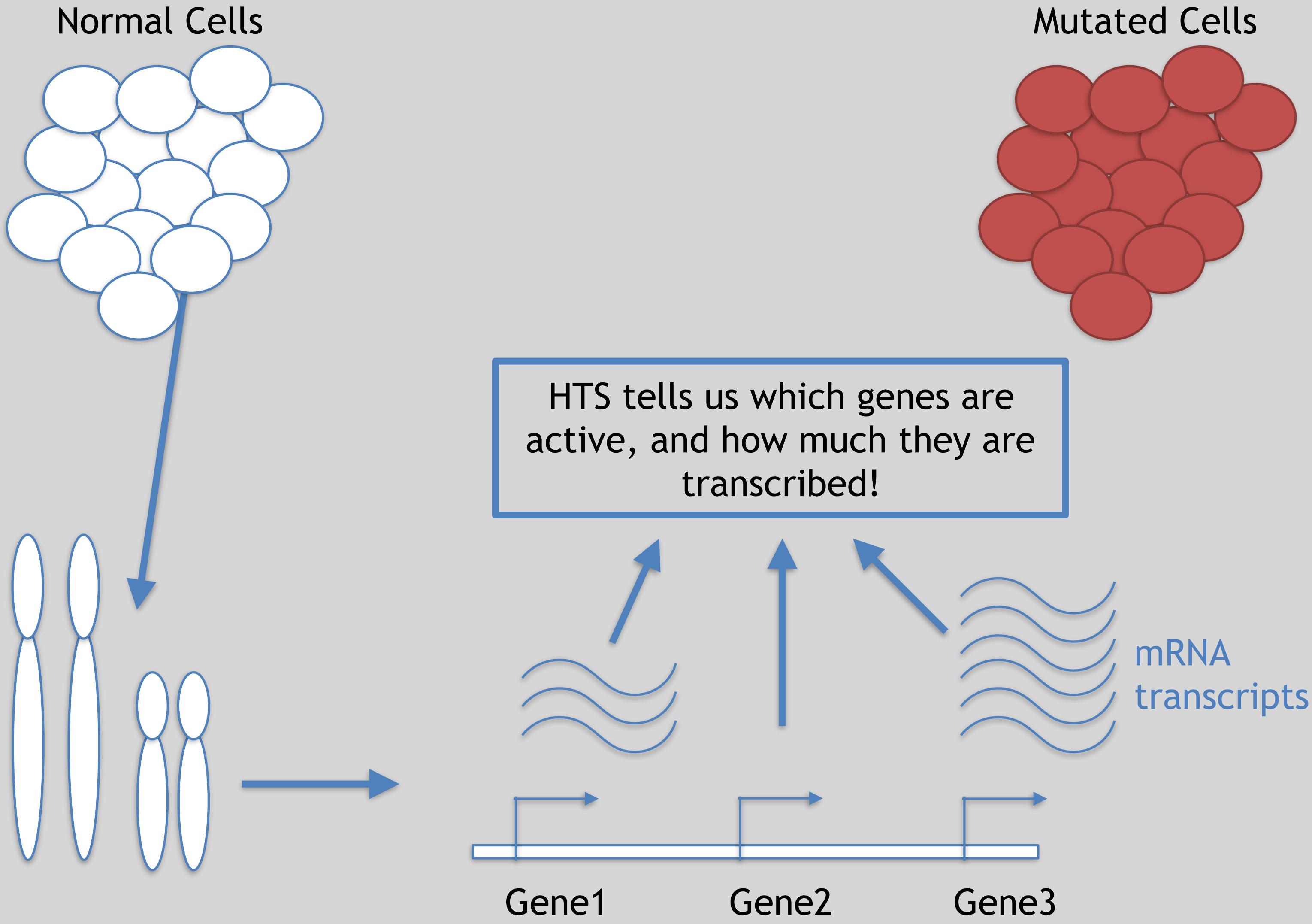
mRNA transcripts

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

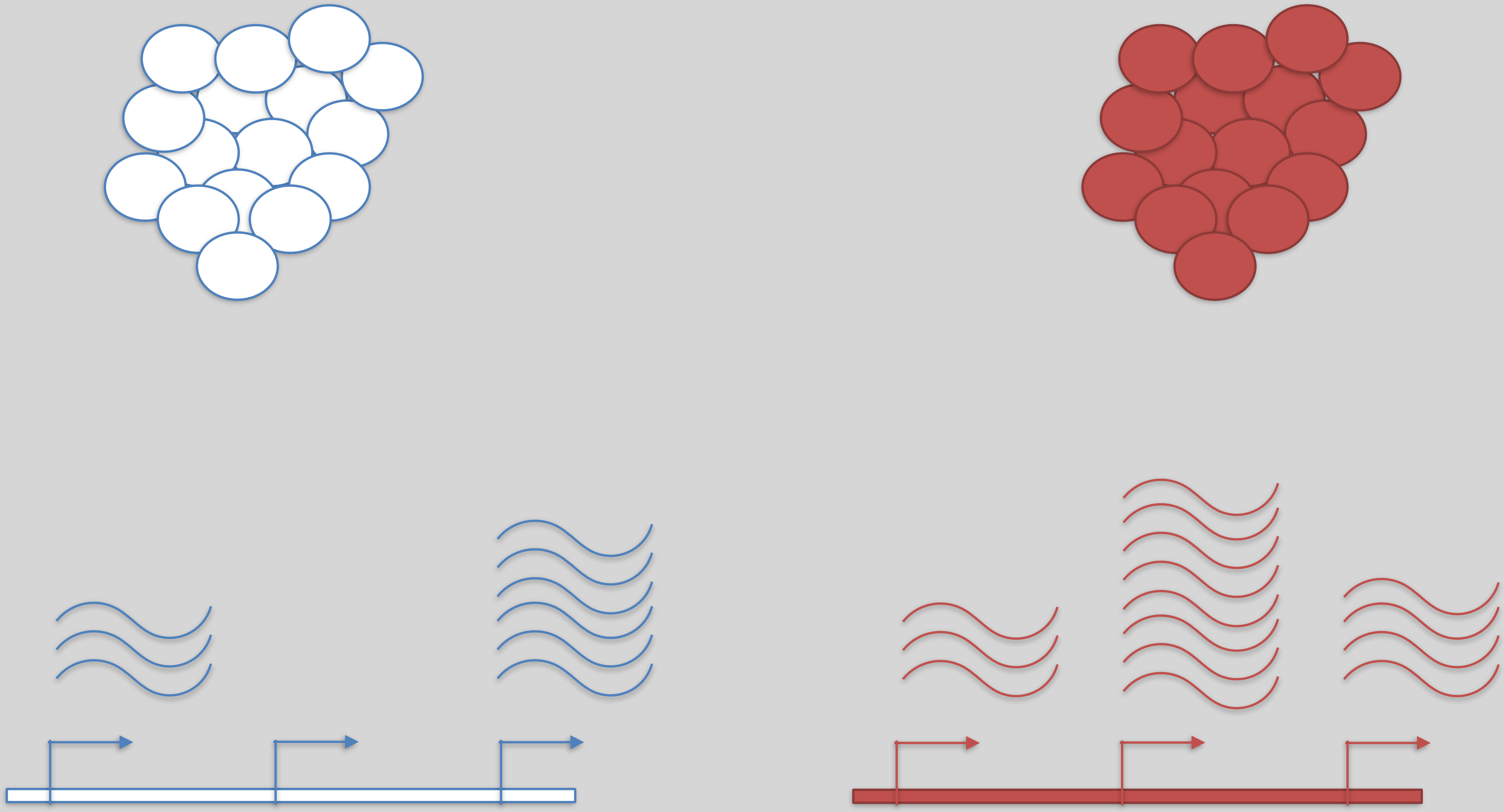
Gene1

Gene2

Gene3

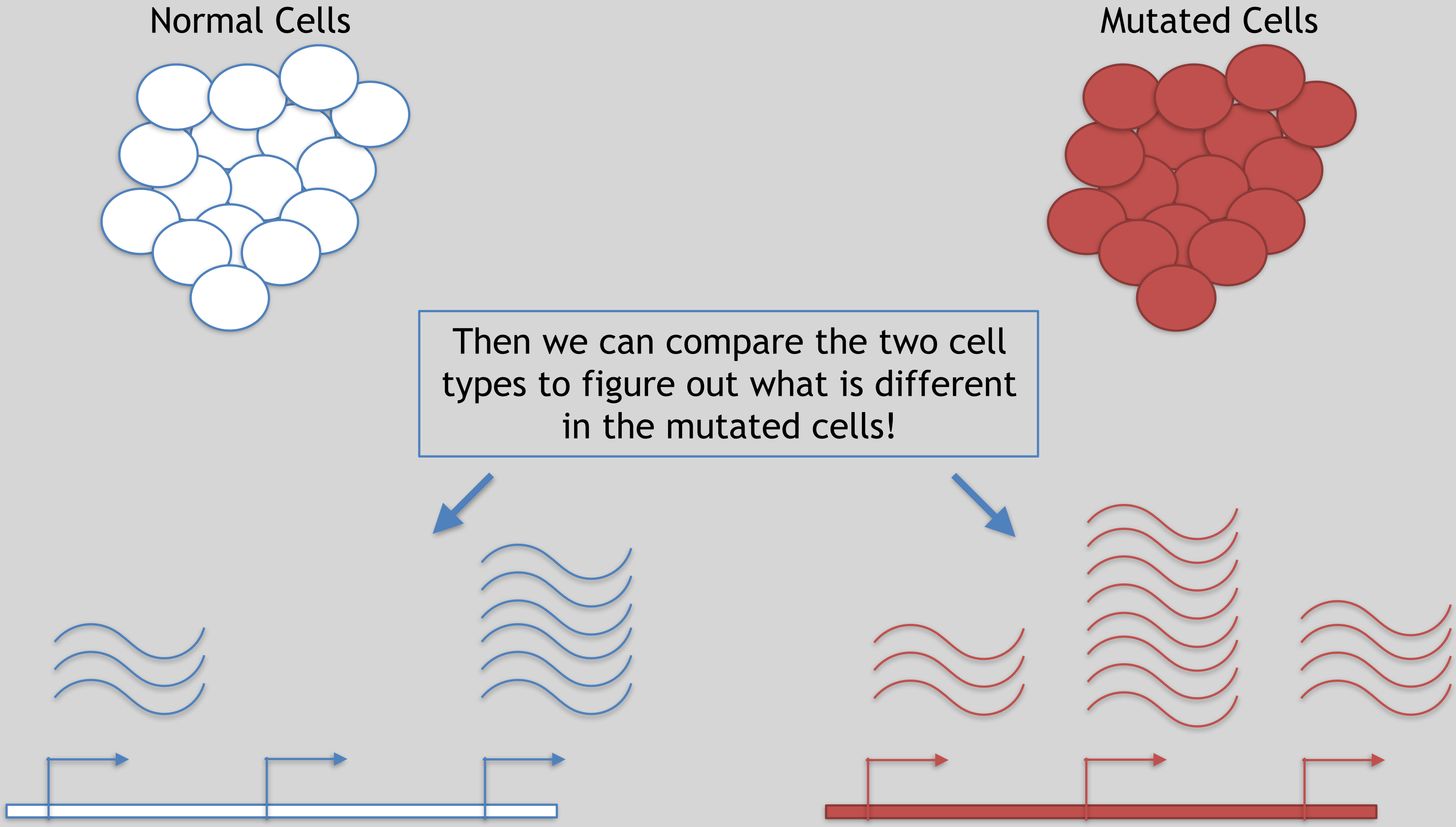
Normal Cells

Mutated Cells



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

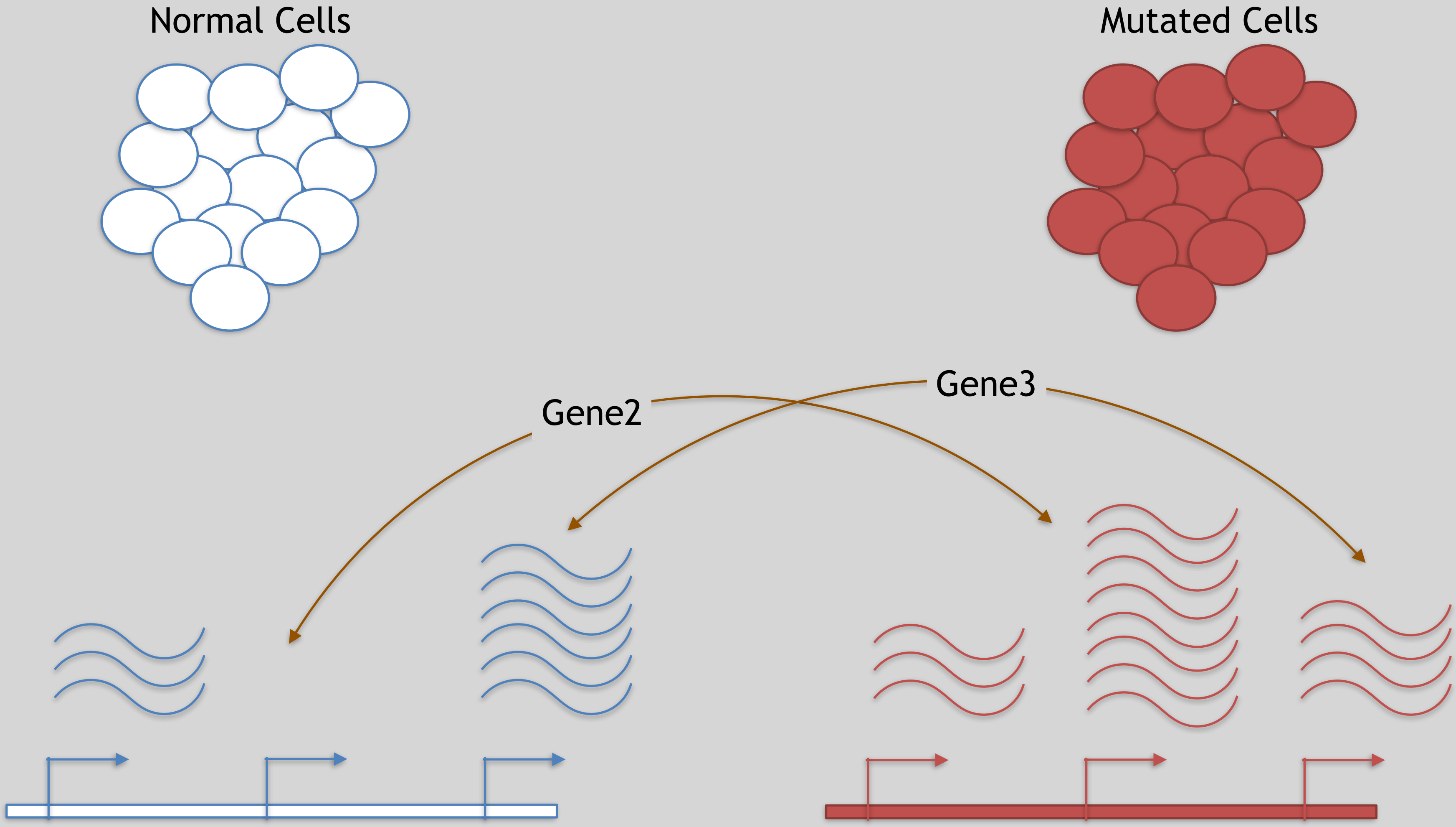
… then use it to measure gene expression in mutated cells



Normal Cells

Mutated Cells

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



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**

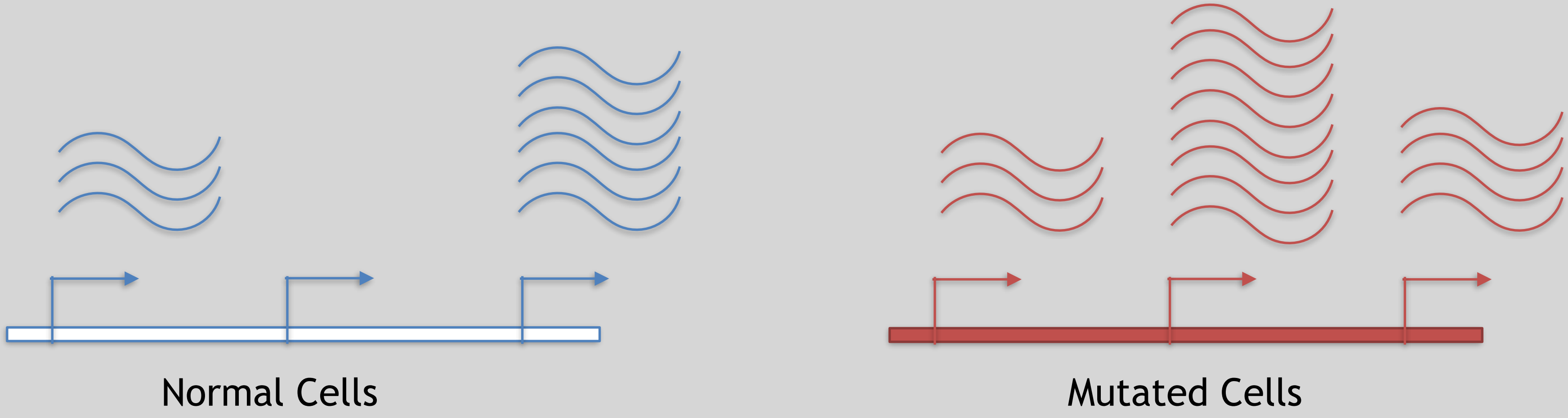
   (Often the major bottleneck to overall success!)

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_W19/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
Bowtie
SOAP2
Novoalign
mr/mrsFast
Eland
Blat
Bfast
BarraCUDA
CASHx
GSNAP
Mosiak
Stampy
SHRiMP
SeqMap
SLIDER
RMAP
SSAHA
etc

Reference

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

Reference

Example SAM File

- 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

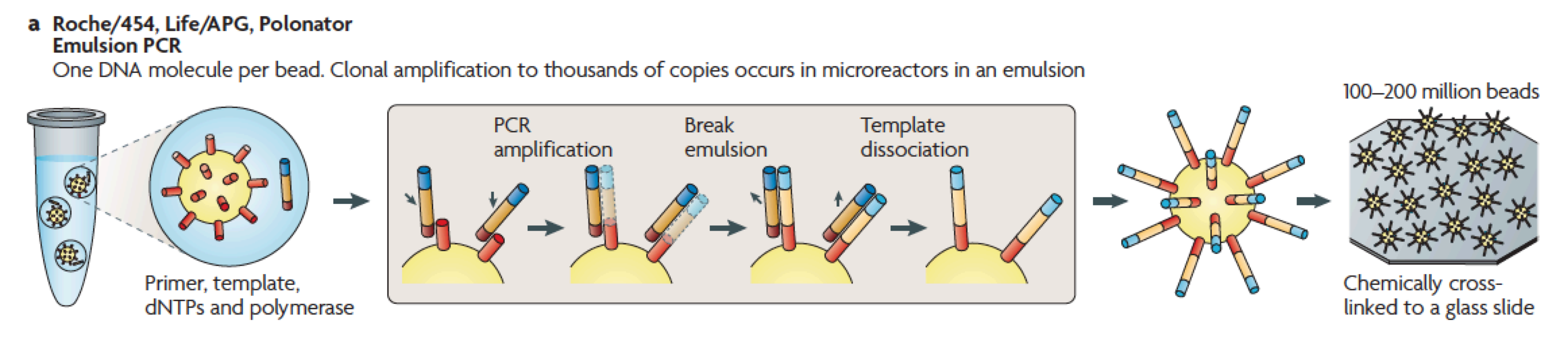
Reference

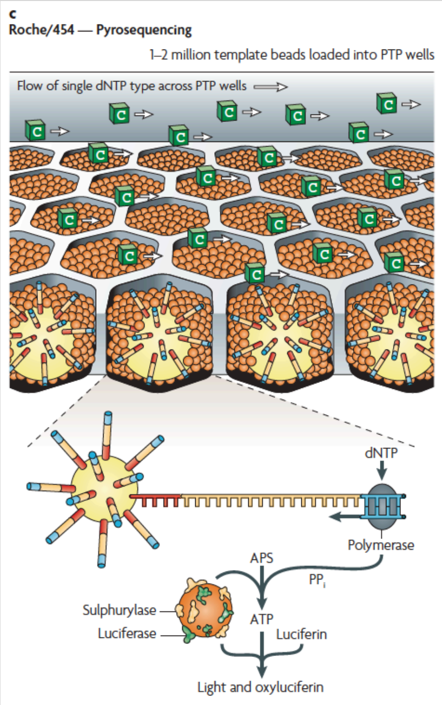
- **Samtools** is a common toolkit for analyzing and manipulating files in SAM/BAM format
  - http://samtools.sourceforge.net/
- **Picard** is a another set of utilities that can used to manipulate and modify SAM files
  - http://picard.sourceforge.net/
- These can be used for viewing, parsing, sorting,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

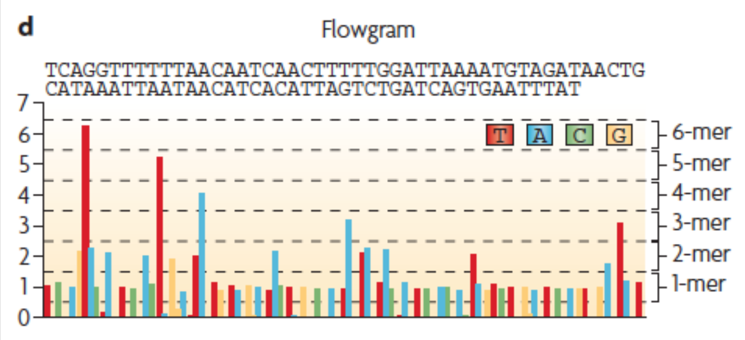
Reference

Additional Reference Slides on Sequencing Methods

Roche 454 - Pyrosequencing

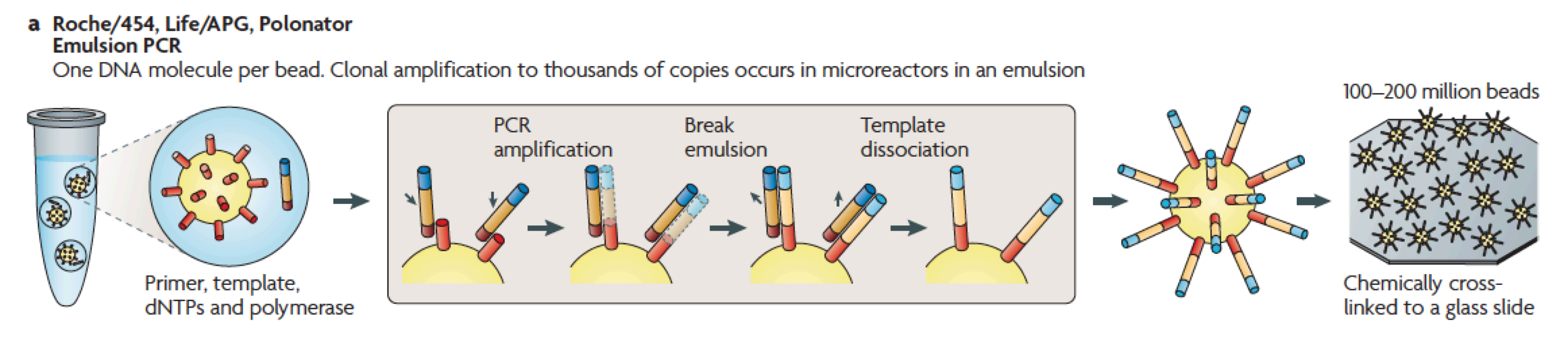

a
Roche/454, Life/APG, Polonator
Emulsion PCR
One DNA molecule per bead. Clonal amplification to thousands of copies occurs in microreactors in an emulsion
Primer, template, dNTPs and polymerase
PCR amplification
Break emulsion
Template dissociation
100–200 million beads
Chemically cross-linked to a glass slide

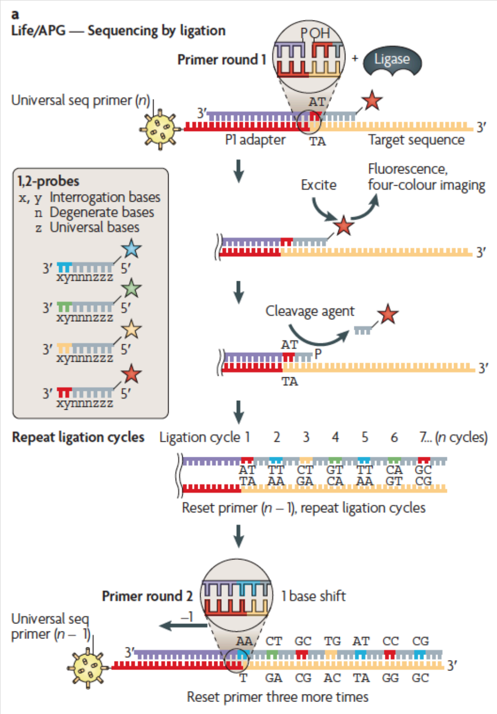

c
Roche/454 — Pyrosequencing
1–2 million template beads loaded into PTP wells
Flow of single dNTP type across PTP wells
C
dNTP
Polymerase
APS
PP$_i$
Sulphurylase
Luciferase
ATP
Luciferin
Light and oxyluciferin

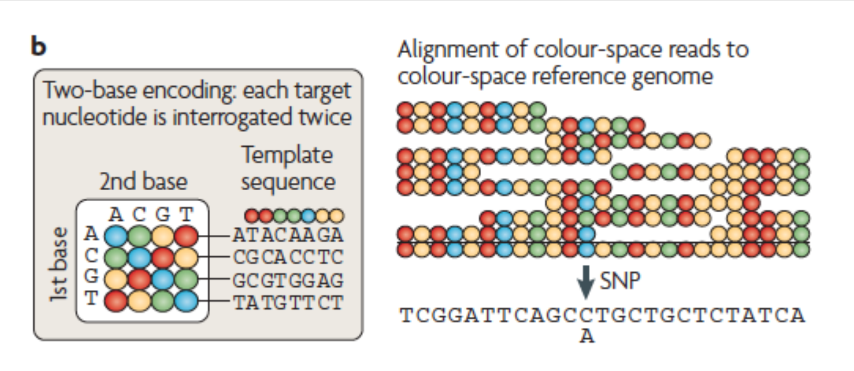


Metzker, ML (2010), *Nat. Rev. Genet*, 11, pp. 31-46

Life Technologies SOLiD – Sequence by Ligation



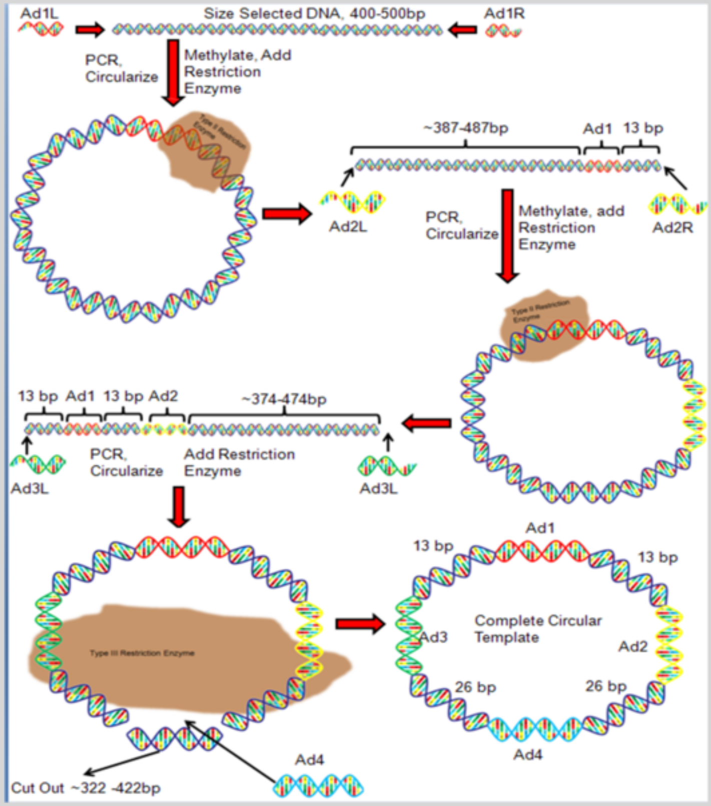



b
Two-base encoding: each target nucleotide is interrogated twice
2nd base
A C G T
1st base
A
C
G
T
Template
sequence
ATACAAGA
CGCACCTC
GCGTGGAG
TATGTTCT
Alignment of colour-space reads to colour-space reference genome
SNP
TCGGATTCAGCCTGCTGCTCTATCA
A

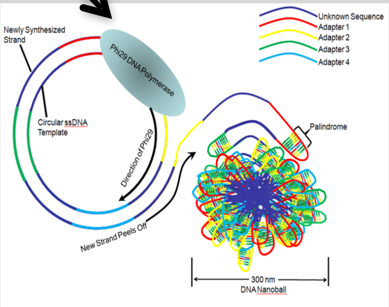
Metzker, ML (2010), *Nat. Rev. Genet*, 11, pp. 31-46

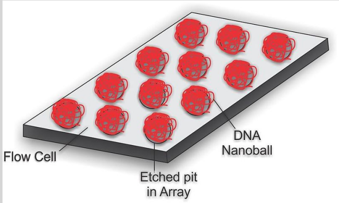
Complete Genomics – Nanoball Sequencing

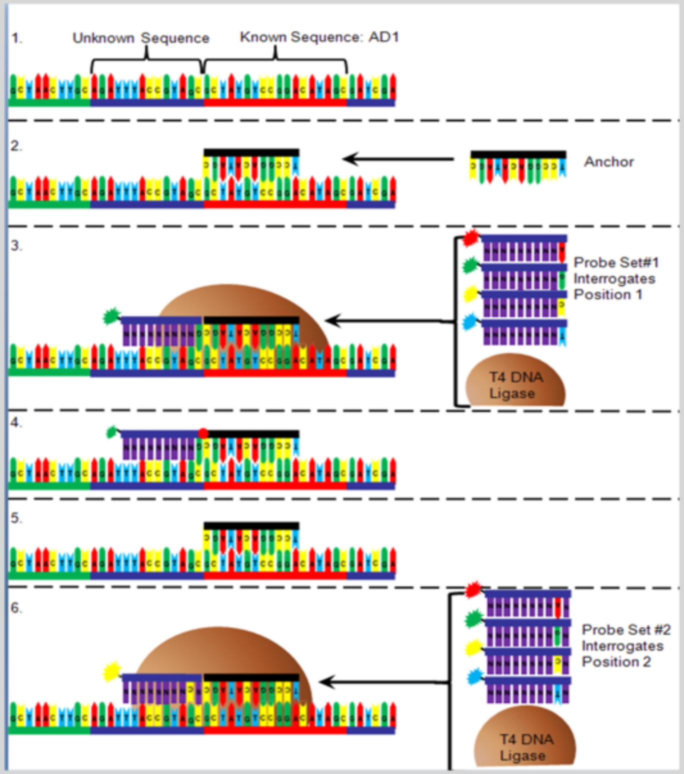
Has proofreading ability!



Ad1L
Size Selected DNA, 400-500bp
Ad1R
PCR, Circularize
Methylate, Add Restriction Enzyme
Type II Restriction Enzyme
~387-487bp
Ad1
13 bp
Ad2L
PCR, Circularize
Methylate, add Restriction Enzyme
Ad2R
Type II Restriction Enzyme
13 bp
Ad1
13 bp
Ad2
~374-474bp
PCR, Circularize
Add Restriction Enzyme
Ad3L
Ad3L
Type III Restriction Enzyme
Cut Out ~322 -422bp
Ad4
Ad1
13 bp
13 bp
Complete Circular Template
Ad3
Ad2
26 bp
26 bp
Ad4





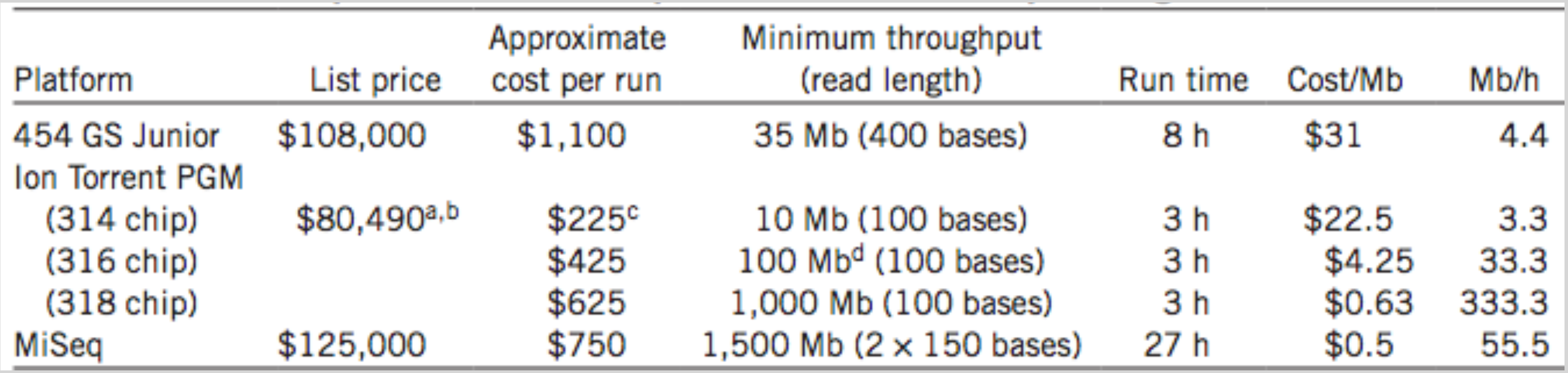


Niedringhaus, TP et al (2011), *Analytical Chem.*, 83, pp. 4327-4341

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

“Benchtop” Sequencers

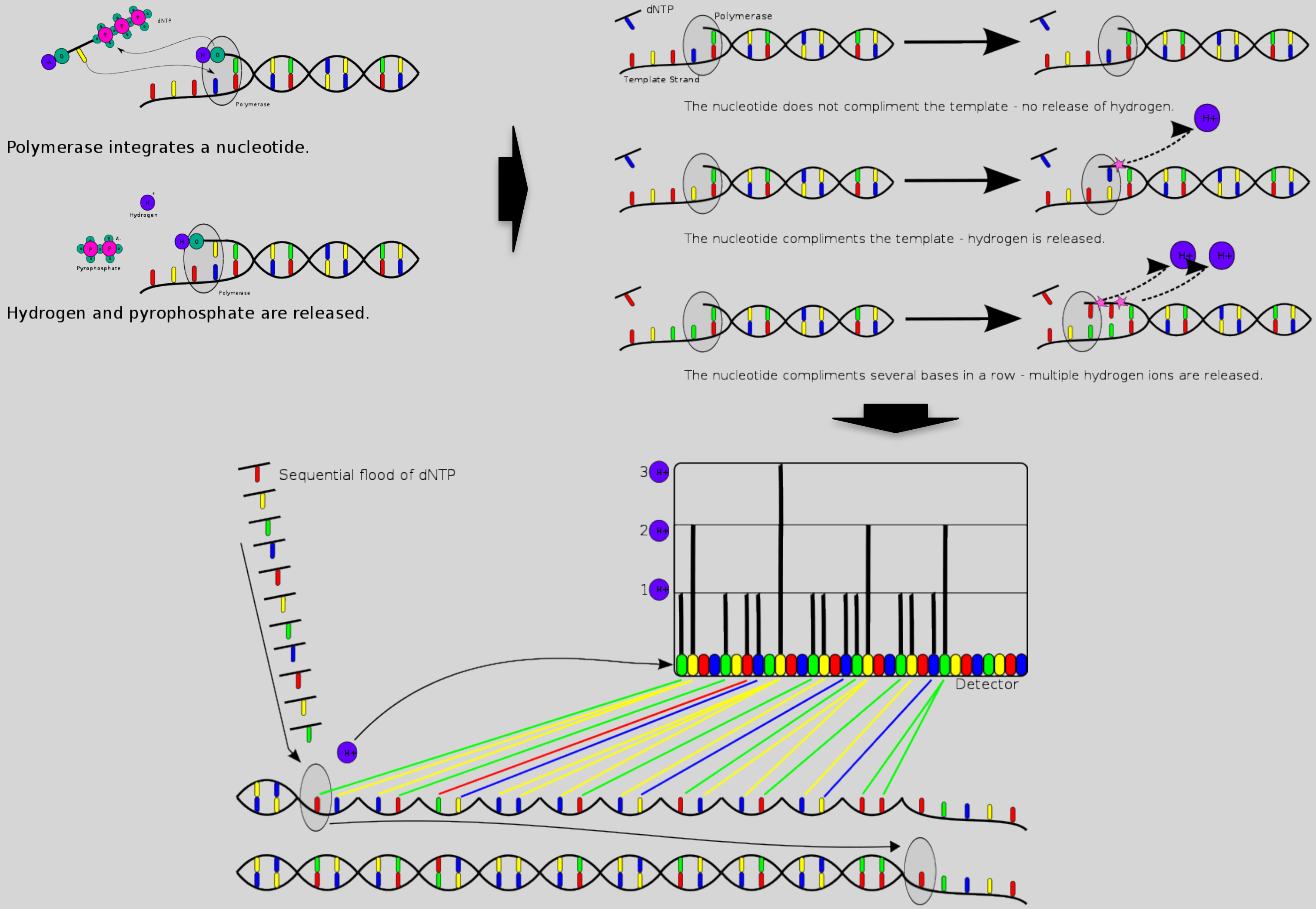
- Lower cost, lower throughput alternative for smaller scale projects
- Currently three significant platforms
  - Roche 454 GS Junior
  - Life Technology Ion Torrent
    - Personal Genome Machine (PGM)
    - Proton
  - Illumina MiSeq



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.

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.

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