BGGN 213

Genome Informatics

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?**
  - Comparative genomics
- **Modern Genome Sequencing**
  - 1st, 2nd and 3rd generation sequencing
- **Workflow for NGS**
  - RNA-Sequencing and Ddiscovering variation

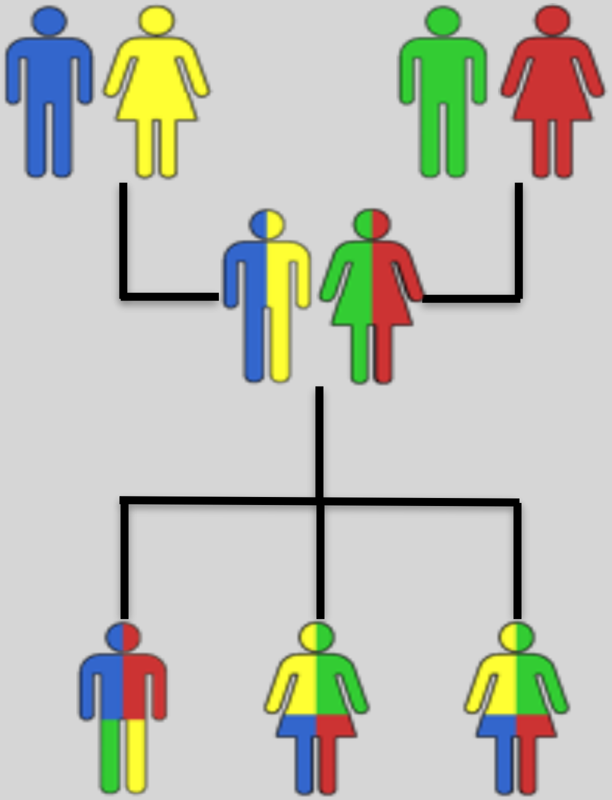
Genetics and Genomics

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

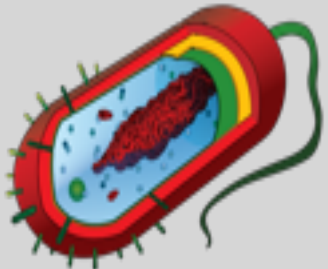
- **Genomics** expands upon classical genetics and considers aspects of the entire genome, typically using computer aided approaches.

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

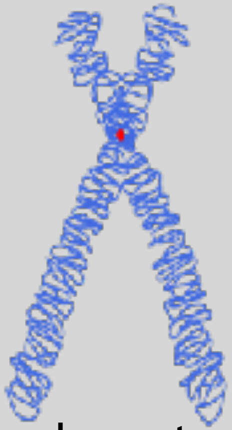


Genomes come in many shapes



prokaryote

bacteriophage



eukaryote

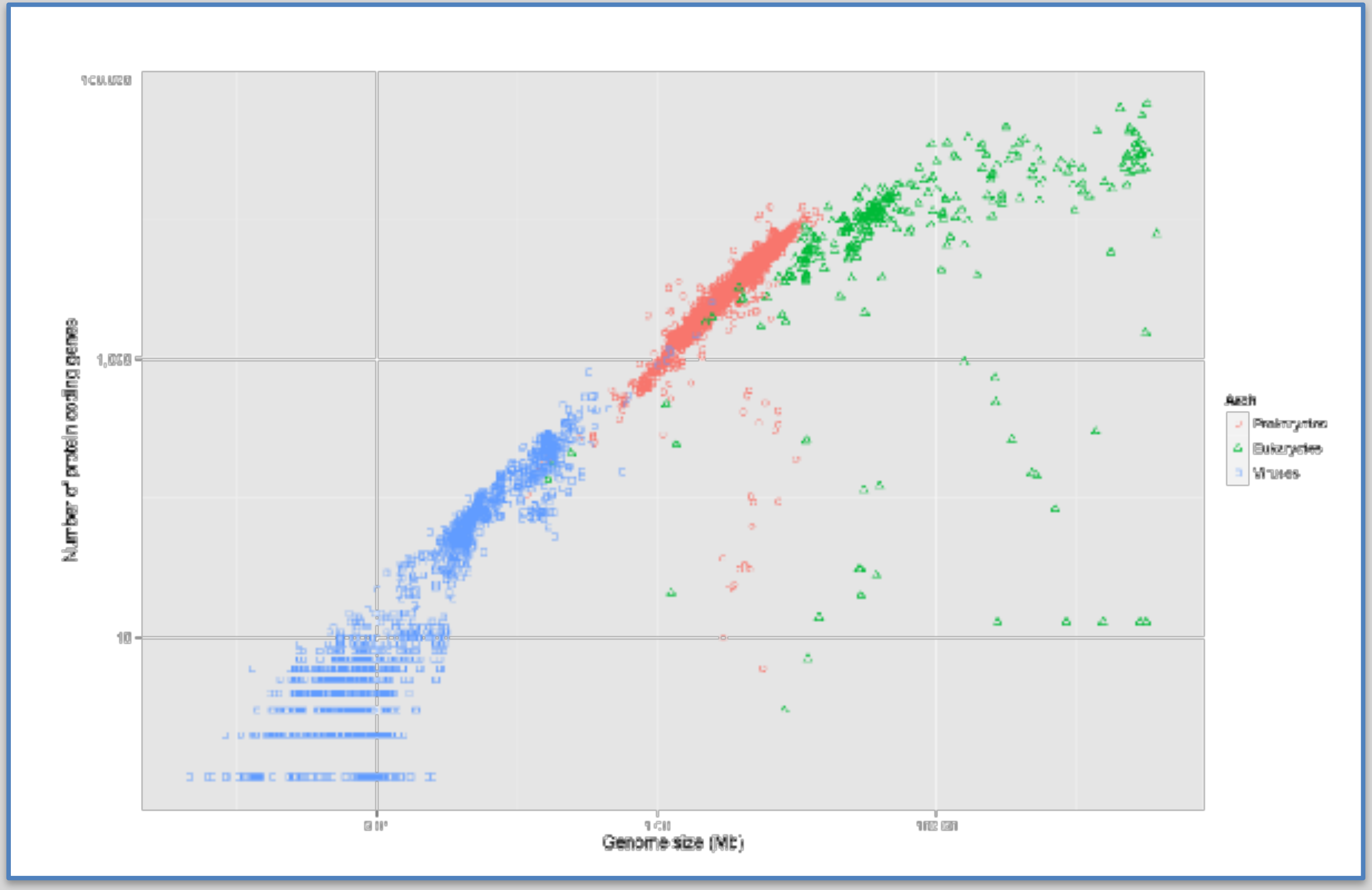
- Primarily DNA, but can be RNA in the case of some viruses

- 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

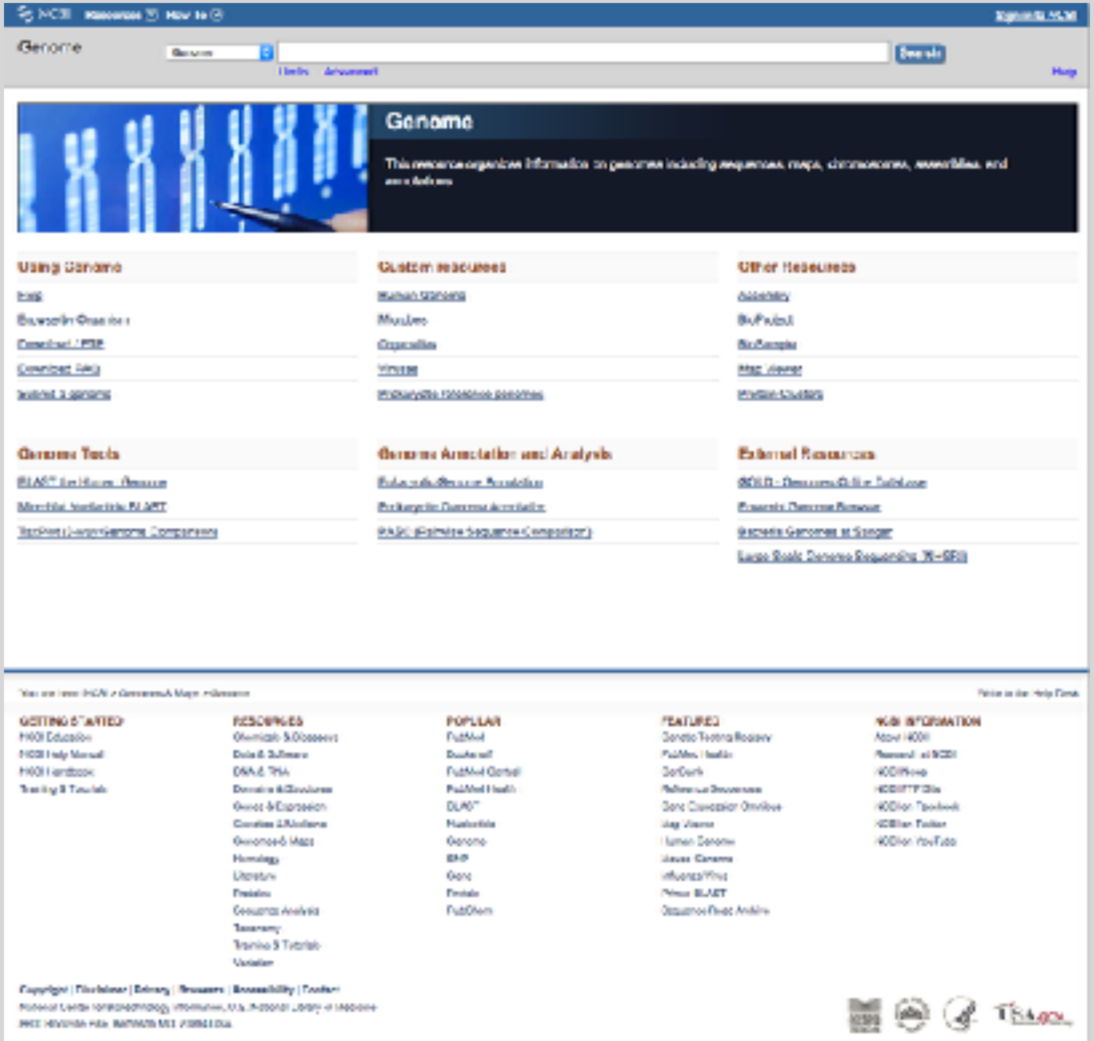


Modified from image by Estevezj / CC BY-SA

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 (Large 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
- Bacterial 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

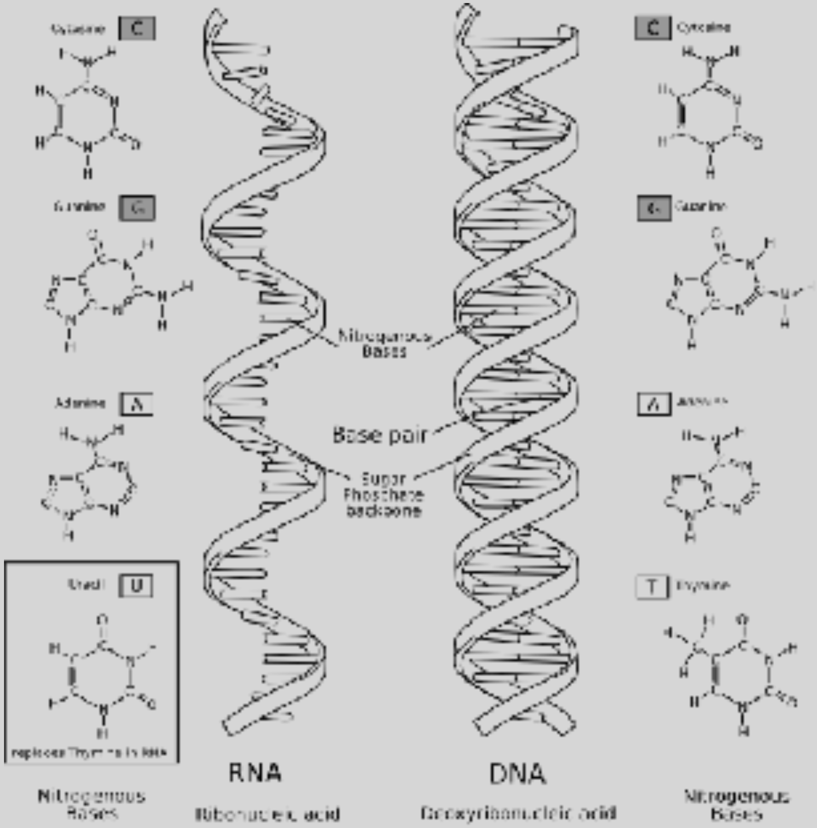
Copyright | Disclaimer | Privacy | Browsers | Accessibility | Contact

National Center for Biotechnology Information, U.S. National Library of Medicine
8600 Rockville Pike, Bethesda MD, 20894 USA

Characteristics of Genomes

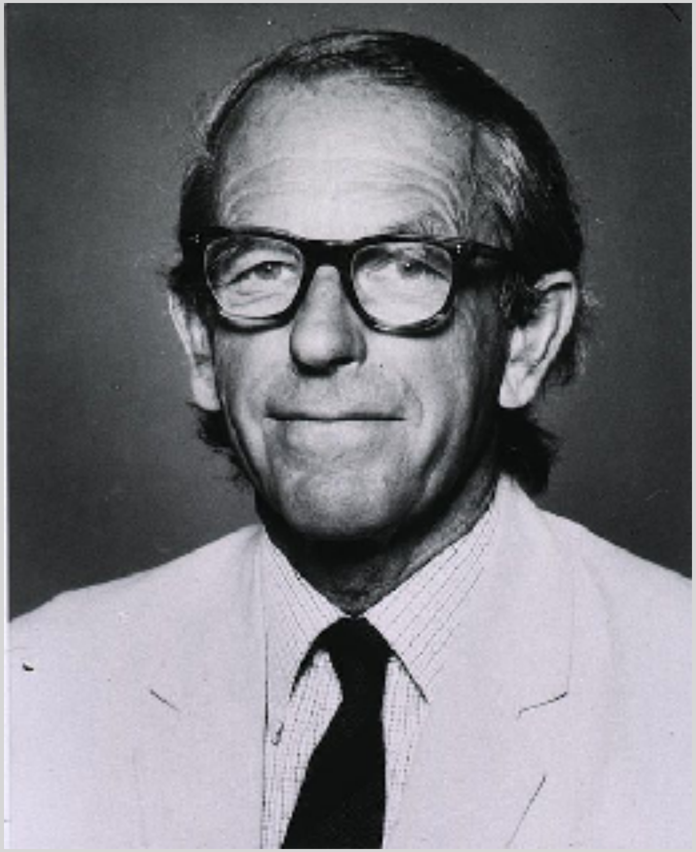
- All genomes are made up of nucleic acids
  - DNA and RNA: Adenine (A), Cytosine (C), Guanine (G)
  - DNA Only: Thymine (T)
  - RNA Only: Uracil (U)
- Typically (but not always), DNA genomes are double stranded (double helix) while RNA genomes are single stranded
- Genomes are described as long sequences of nucleic acids, for example:

  GGACTTCAGGCAACTGCAACTACCTTAGGA



Darryl Leja, Courtesy: National Human Genome Research Institute.

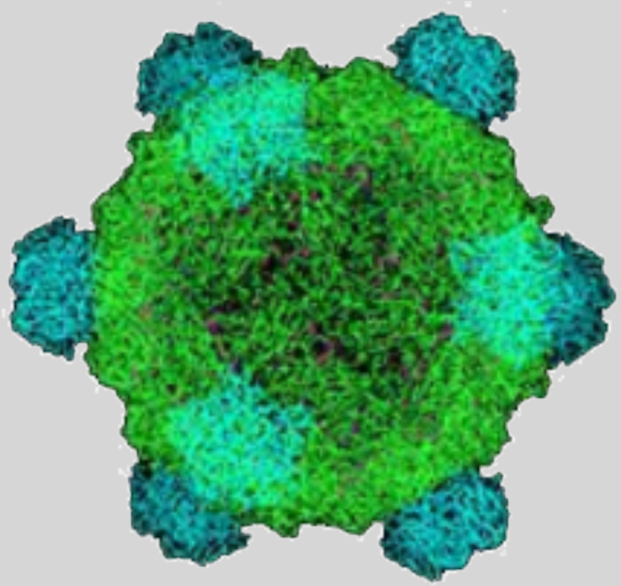
Early Genome Sequencing



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

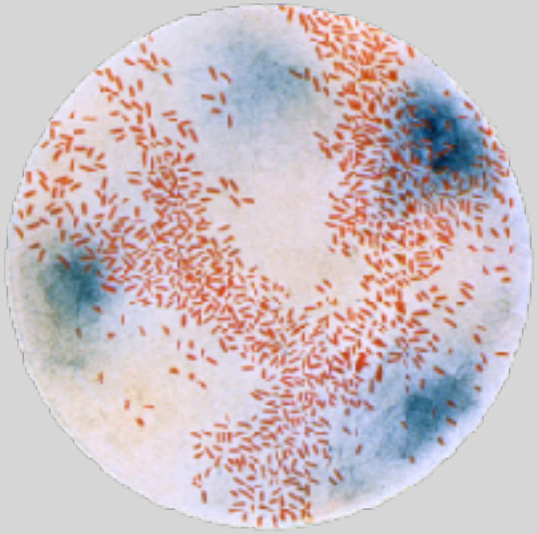
http://en.wikipedia.org/wiki/Frederick_Sanger

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

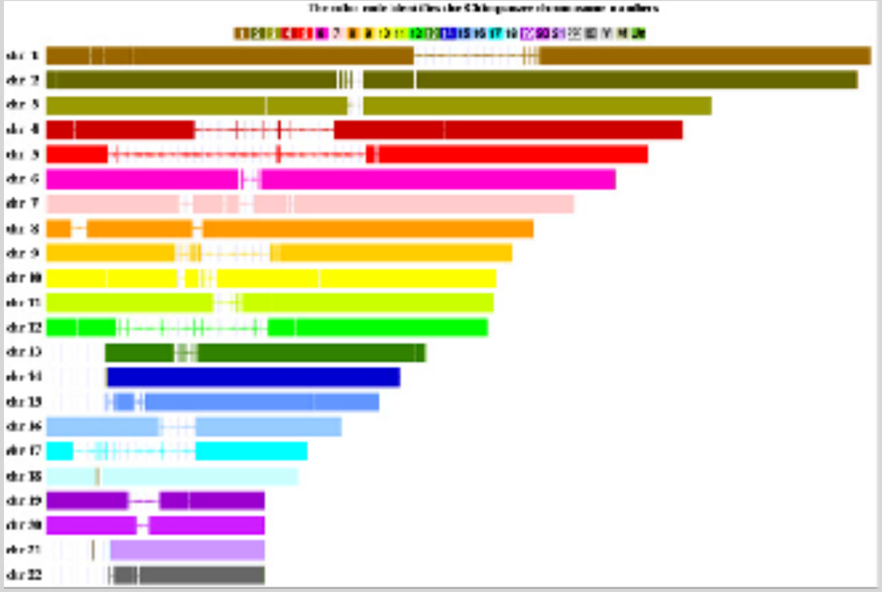
Jane Ades, Courtesy: National Human Genome Research Institute.

What can we do with a Genome?

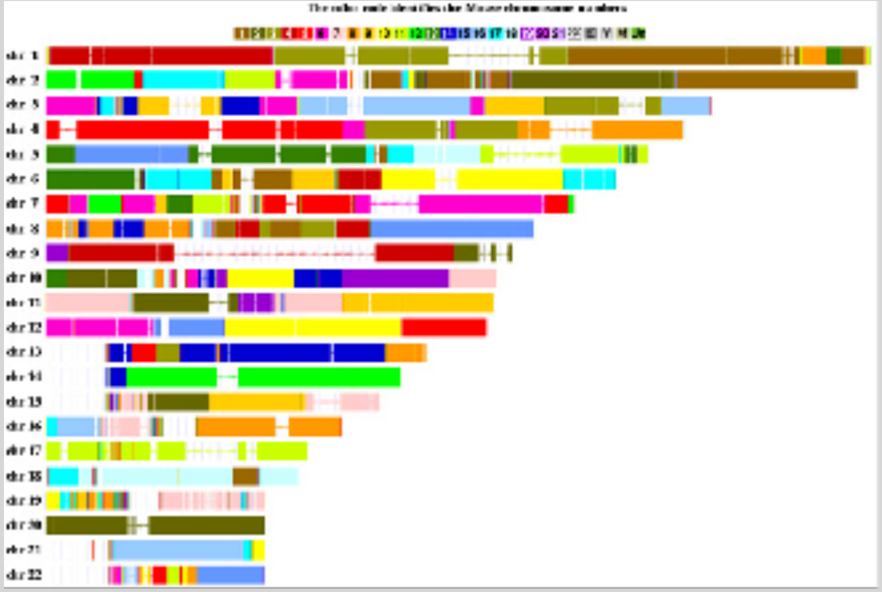
- We can *compare* genomes, both within and between species, to identify regions of variation and of conservation
- We can *model* genomes, to find interesting patterns reflecting functional characteristics
- We can *edit* genomes, to add, remove, or modify genes and other regions for adjusting individual traits

Comparative Genomics

~6-7 million years



~60-70 million years



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

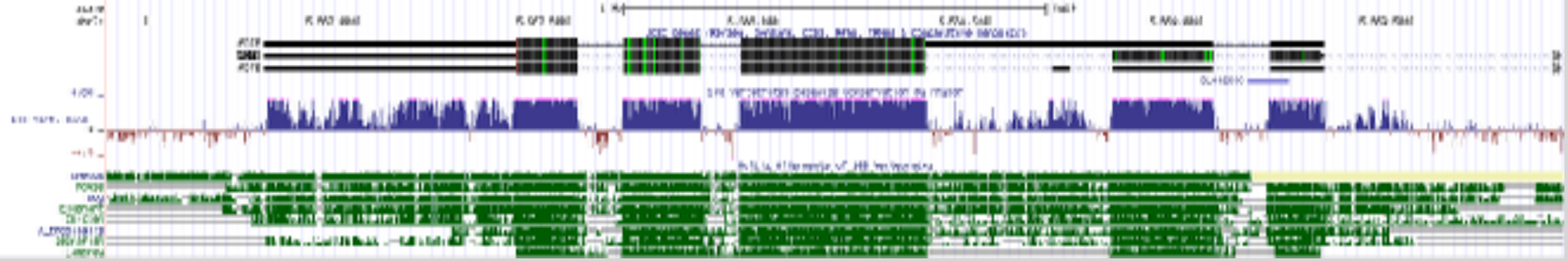
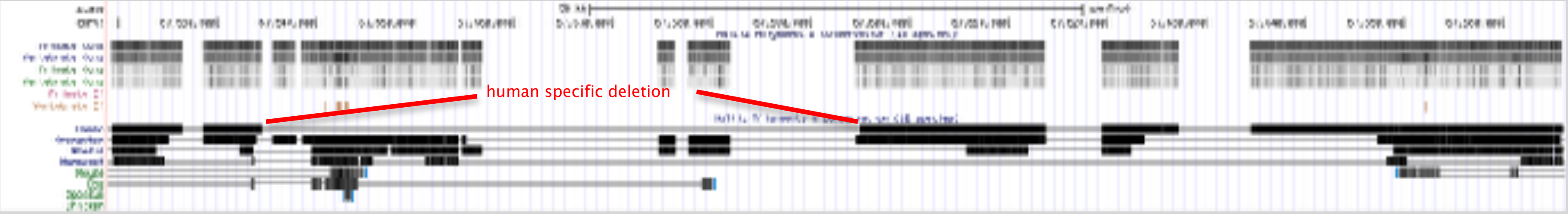


figure generated from: http://genome.ucsc.edu/

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)



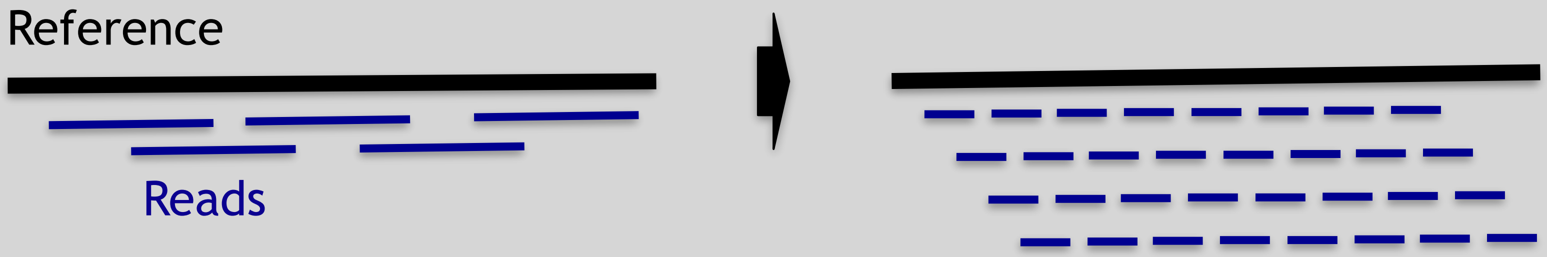
human specific deletion

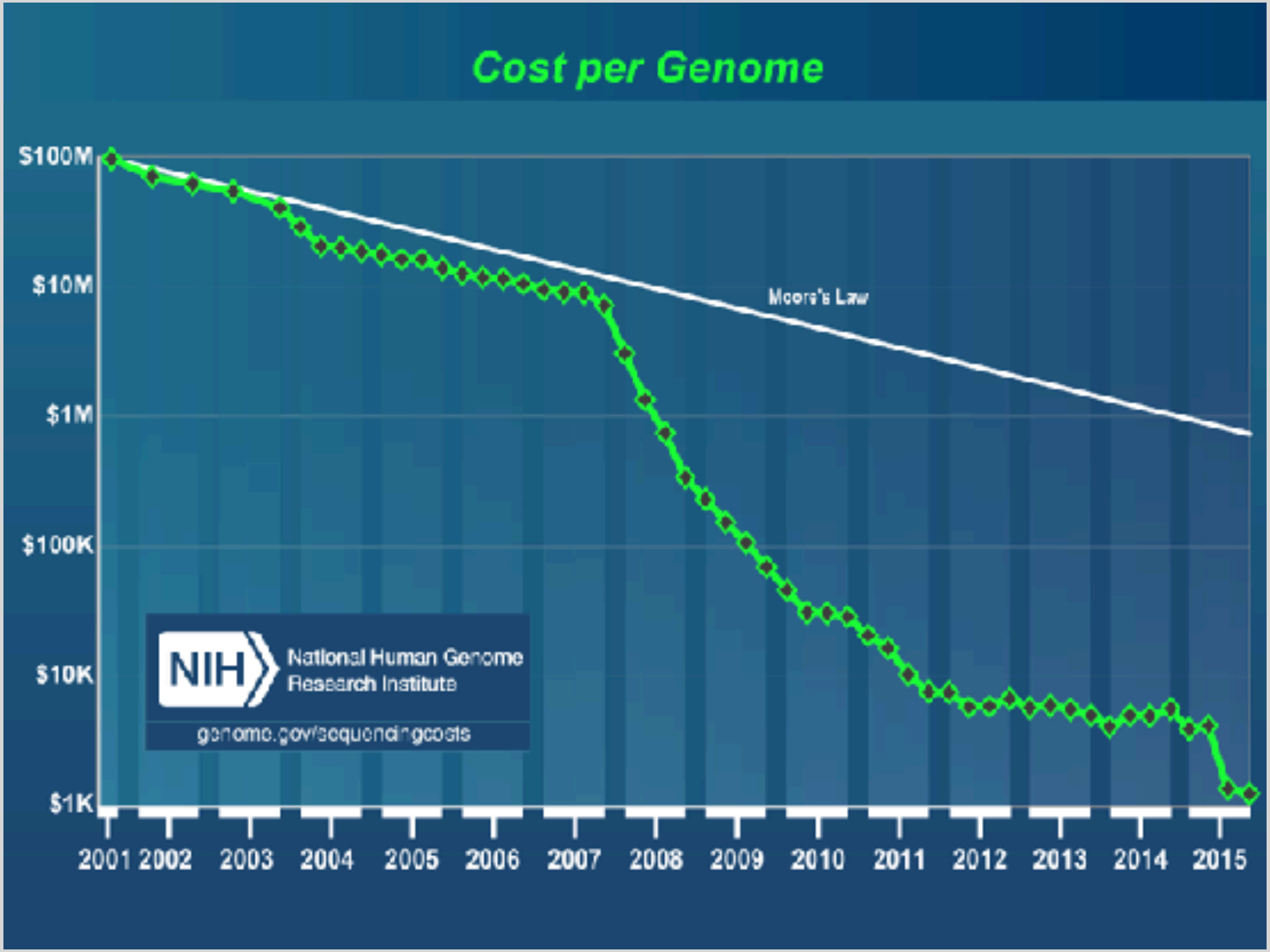
conserved in other mammals

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

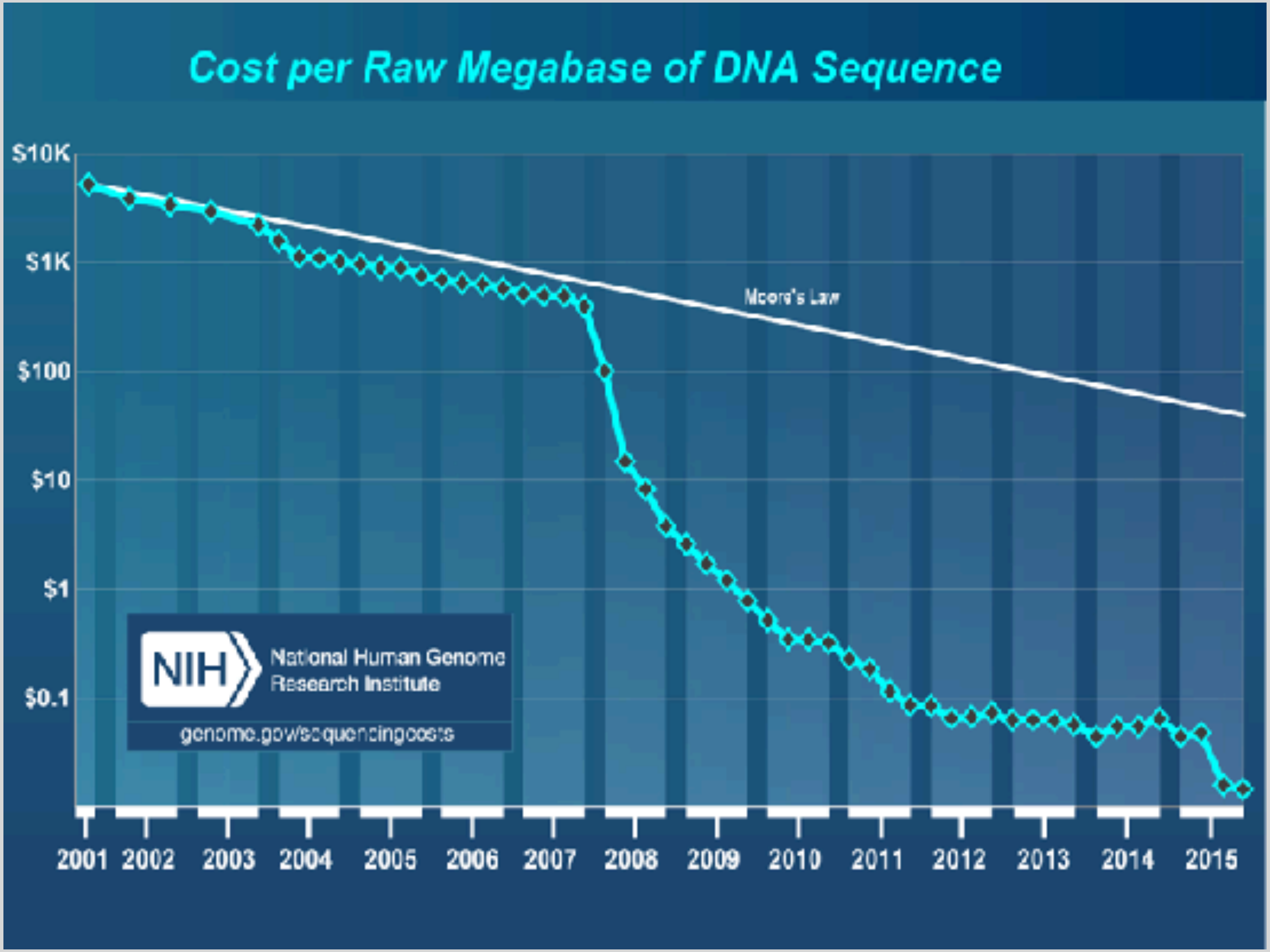




Cost per Genome

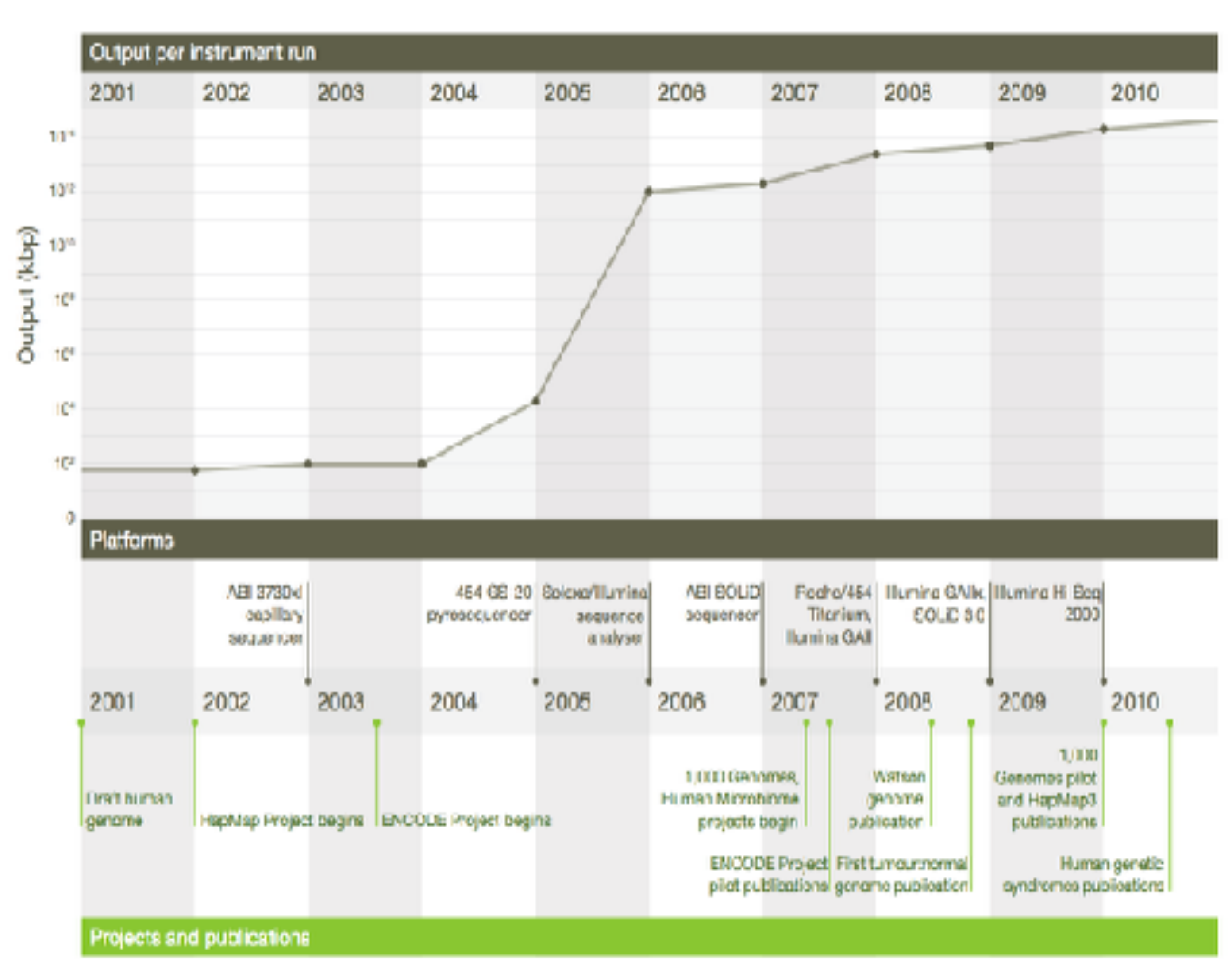
National Human Genome Research Institute

genome.gov/sequencingcosts



Cost per Raw Megabase of DNA Sequence

Timeline of Sequencing Capacity

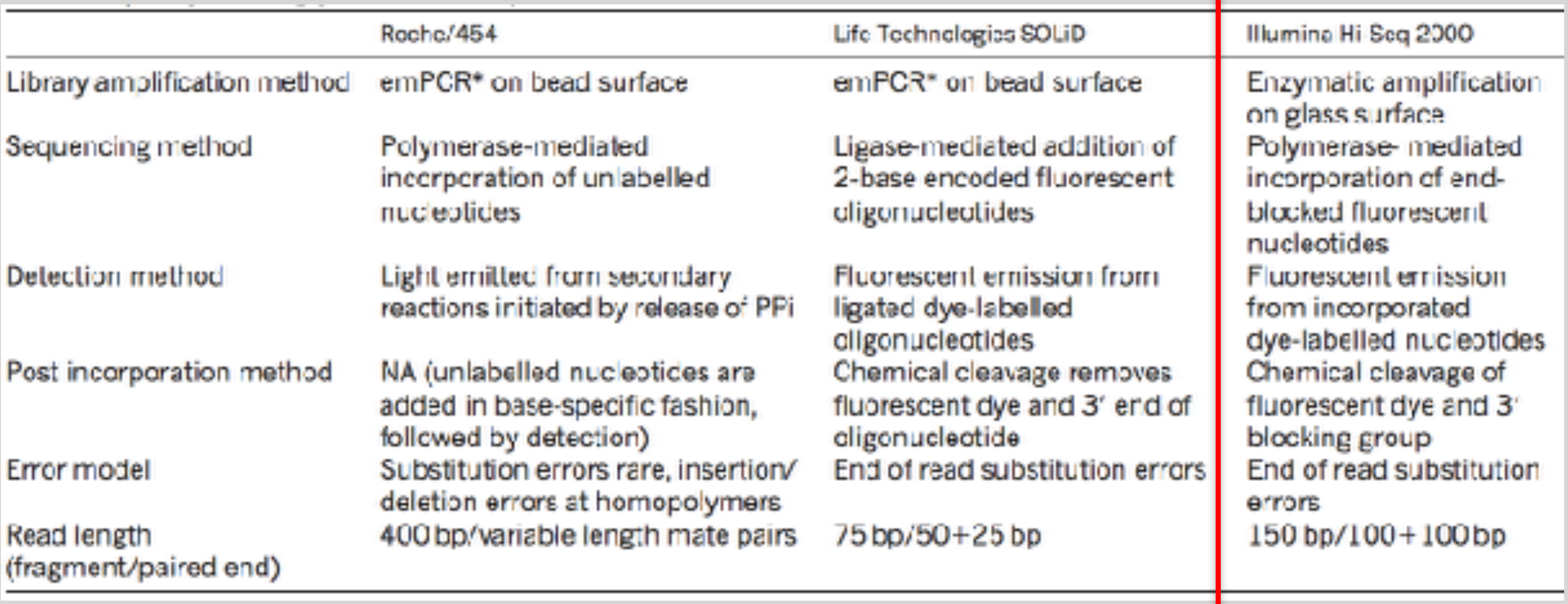


Mardis, ER (2011), Nature, 470, pp. 198-203

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)

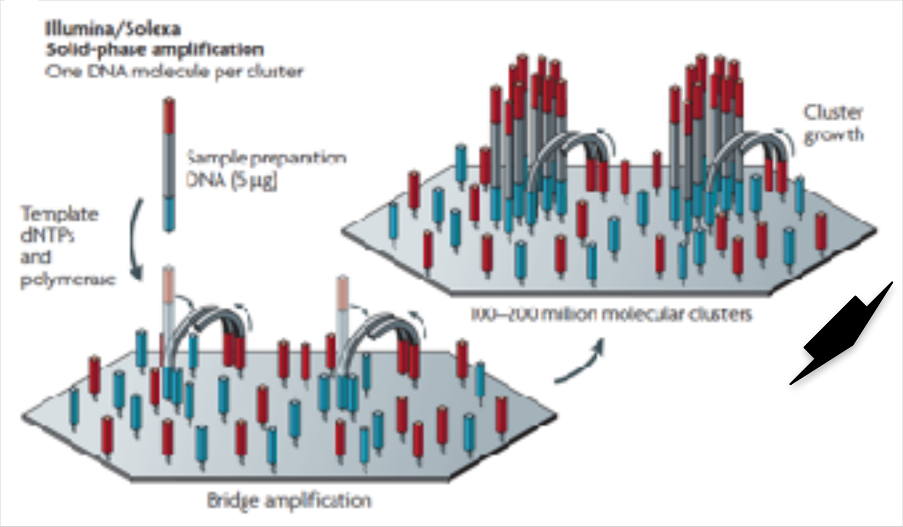
Modern NGS Sequencing Platforms

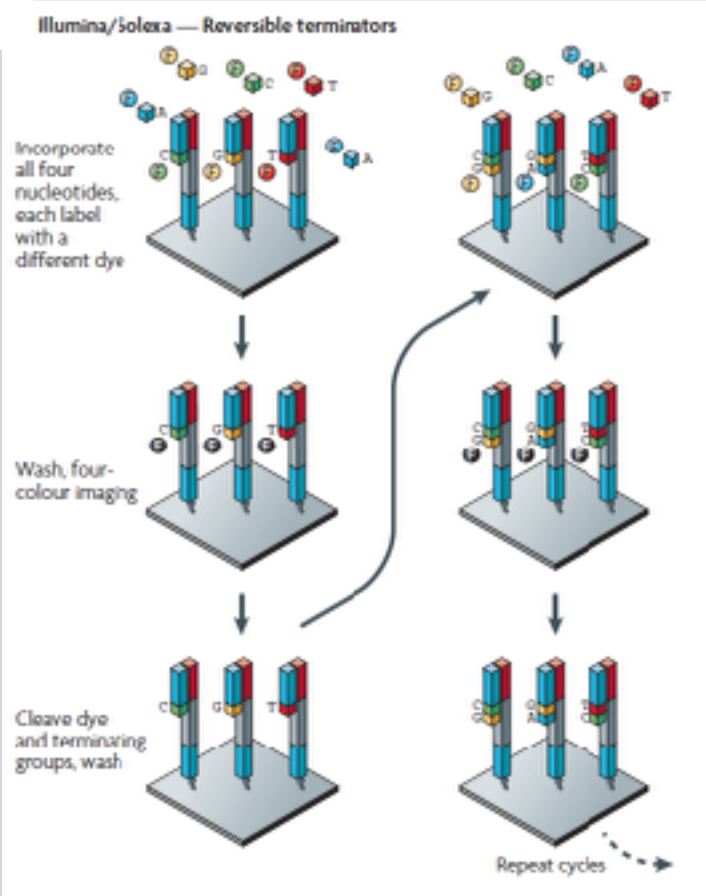


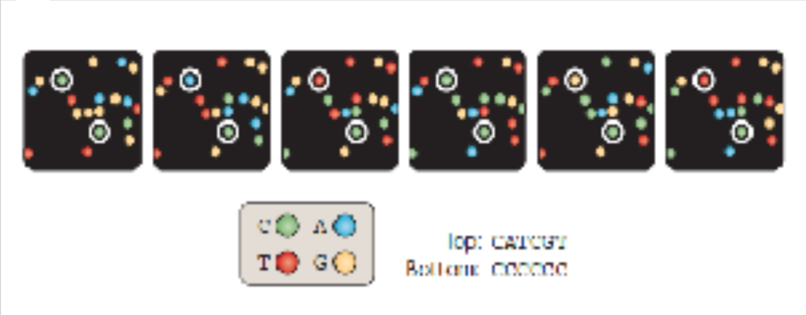
	Roche/454	Life Technologies SOLiD	Illumina Hi Seq 2000
Library amplification method	emPCR* on bead surface	emPCR* on bead surface	Enzymatic amplification on glass surface
Sequencing method	Polymerase-mediated incorporation of unlabelled nucleotides	Ligase-mediated addition of 2-base encoded fluorescent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



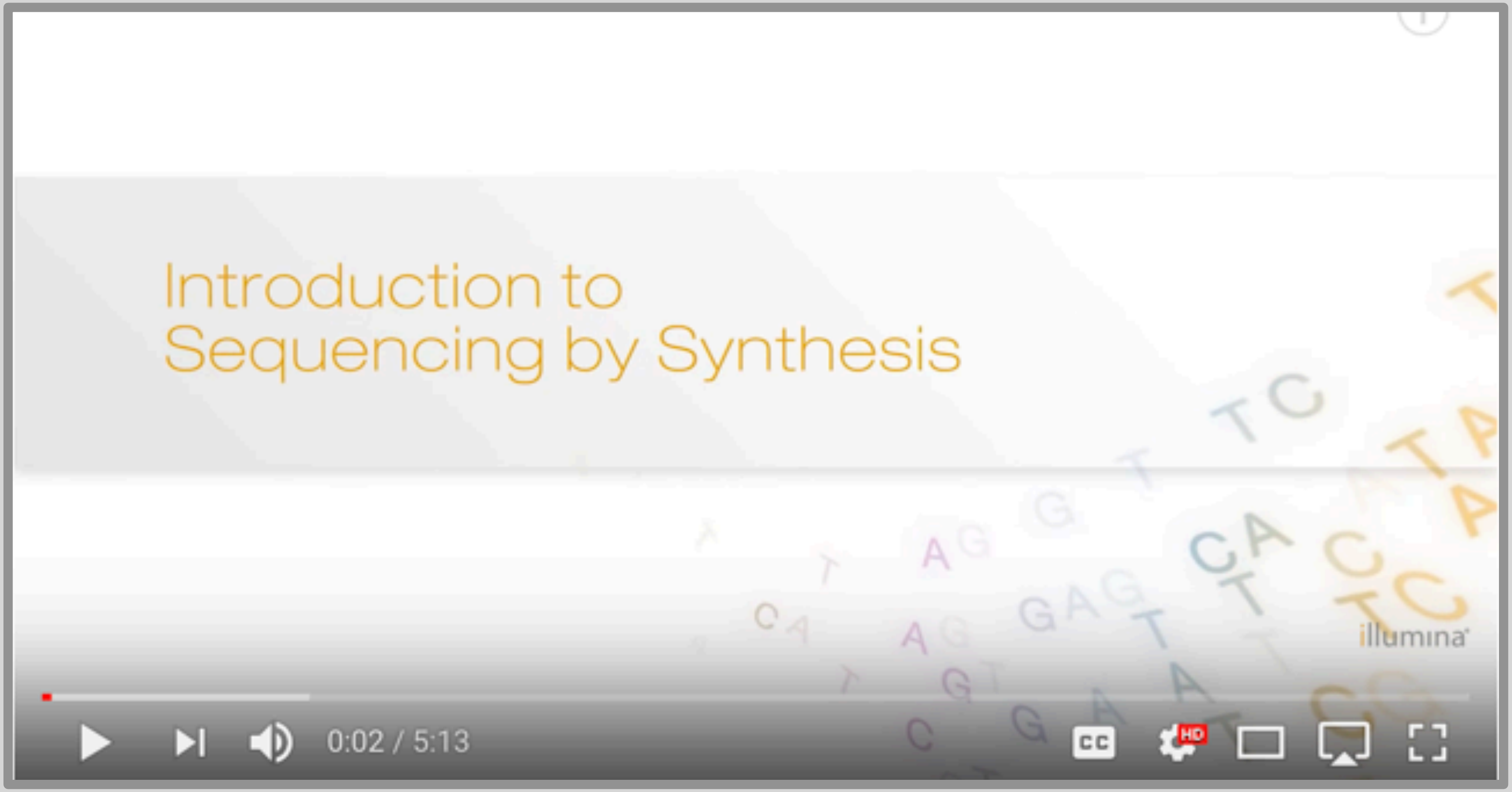




(other sequencing platforms summarized at end of slide set)

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

Illumina Sequencing - Video

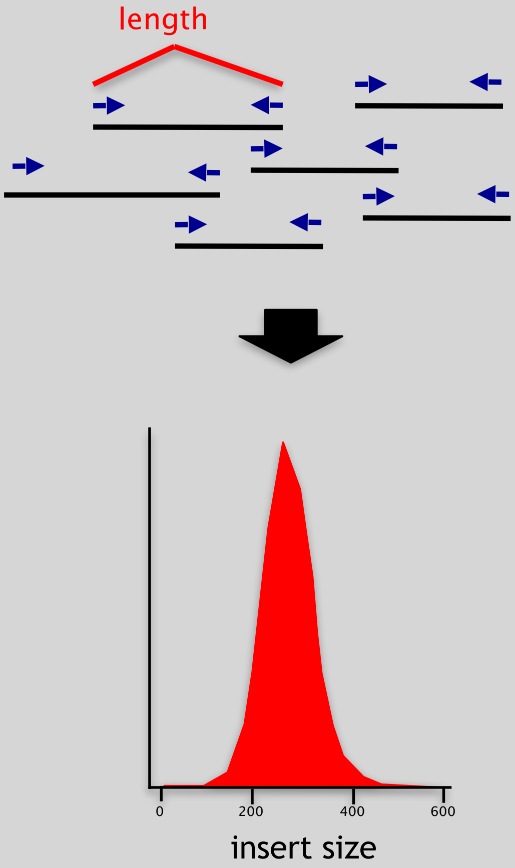


Introduction to Sequencing by Synthesis

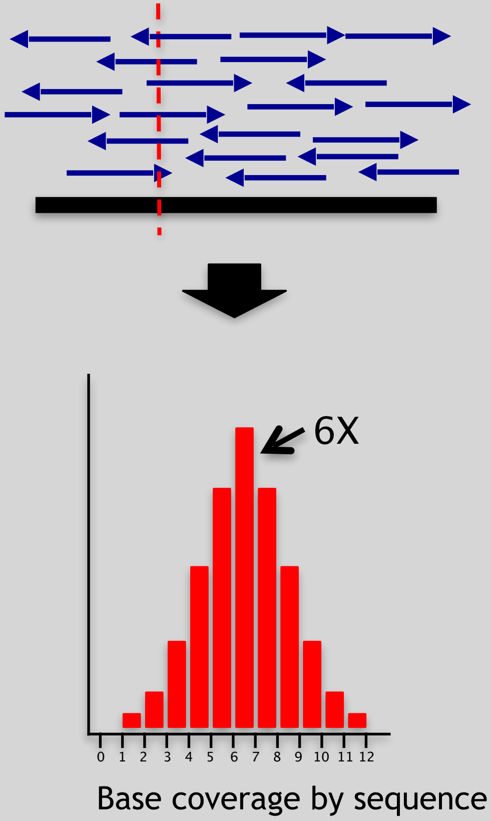
https://www.youtube.com/watch?src_vid=womKfikWlxM&v=fCd6B5HRaZ8

NGS Sequencing Terminology

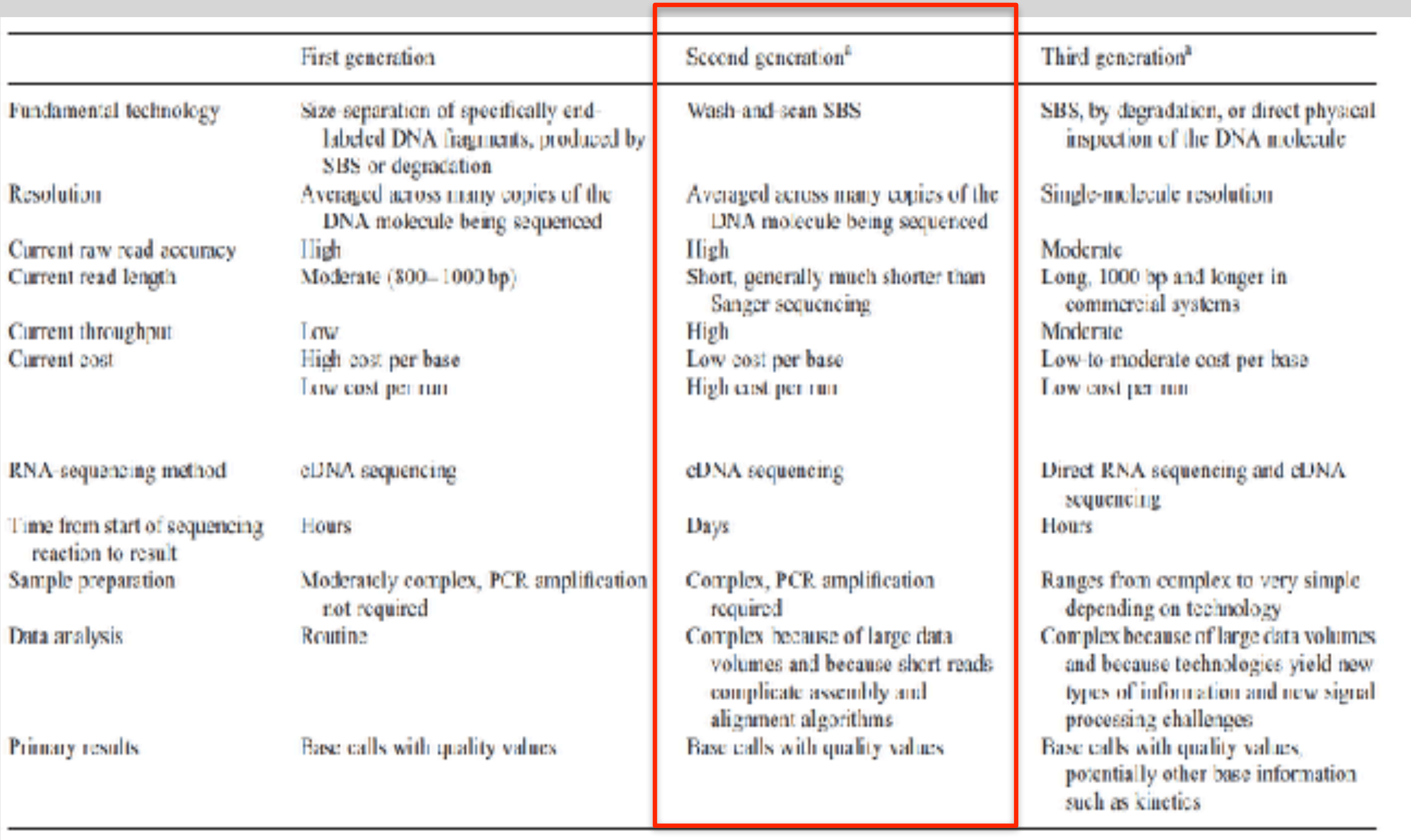
Insert Size



Sequence Coverage



Summary: “Generations” of DNA Sequencing



	First generation	Second generation[a]	Third generation[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

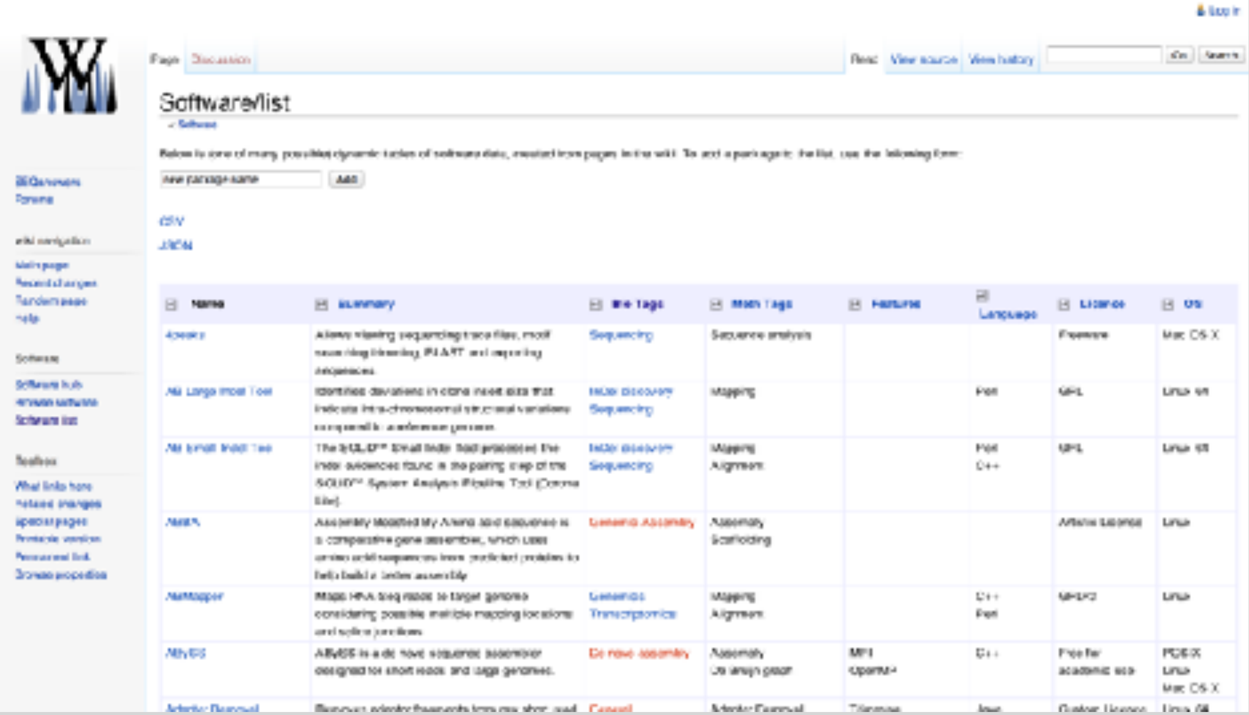
Schadt, EE et al (2010), *Hum. Mol. Biol.*, 19(RI2), pp. R227-R240

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!

SeqAnswers Wiki

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



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	Method 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 sequences 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 OpenMPI	C++	Free for academic use	POSIX Linux Mac OS X
AdapterRemoval	Removes adaptor fragments from raw short read	General	Adaptor Removal	Trimming	Java	Custom License	Linux 64

Raw data usually in **FASTQ format**

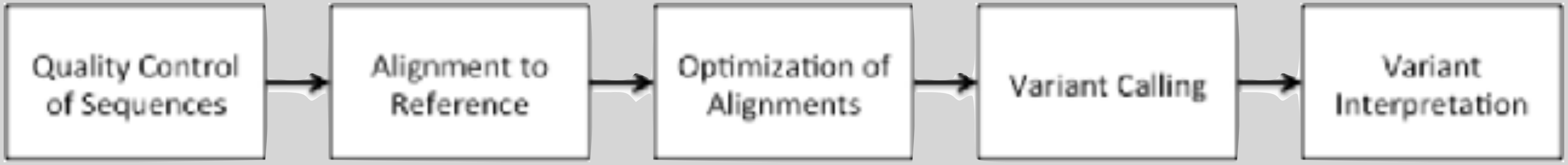
```
@NS500177:196:HFTTTAFXX:1:11101:10916:1458 2:N:0:CGCGGCTG
ACACGACGATGAGGTGACAGTCACGGAGGATAAGATCAATGCCCTCATTAAAGCAGCCGGTGTAA
+
AAAAAEEEEEEEEEEE//AEEEAEEEEEEEEEEE/EE/<<EE/AAEEAEE///EEEEAEEEAEA<
```

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

Generic Workflow for NGS

- There are many different ways to analyze sequences generated from NGS, depending on the specific question you are investigating
- For the analysis of genomic sequence data, a typical (if generic) approach is as follows



Quality Control (QC)

- Quality checks of raw sequence data are *very* important

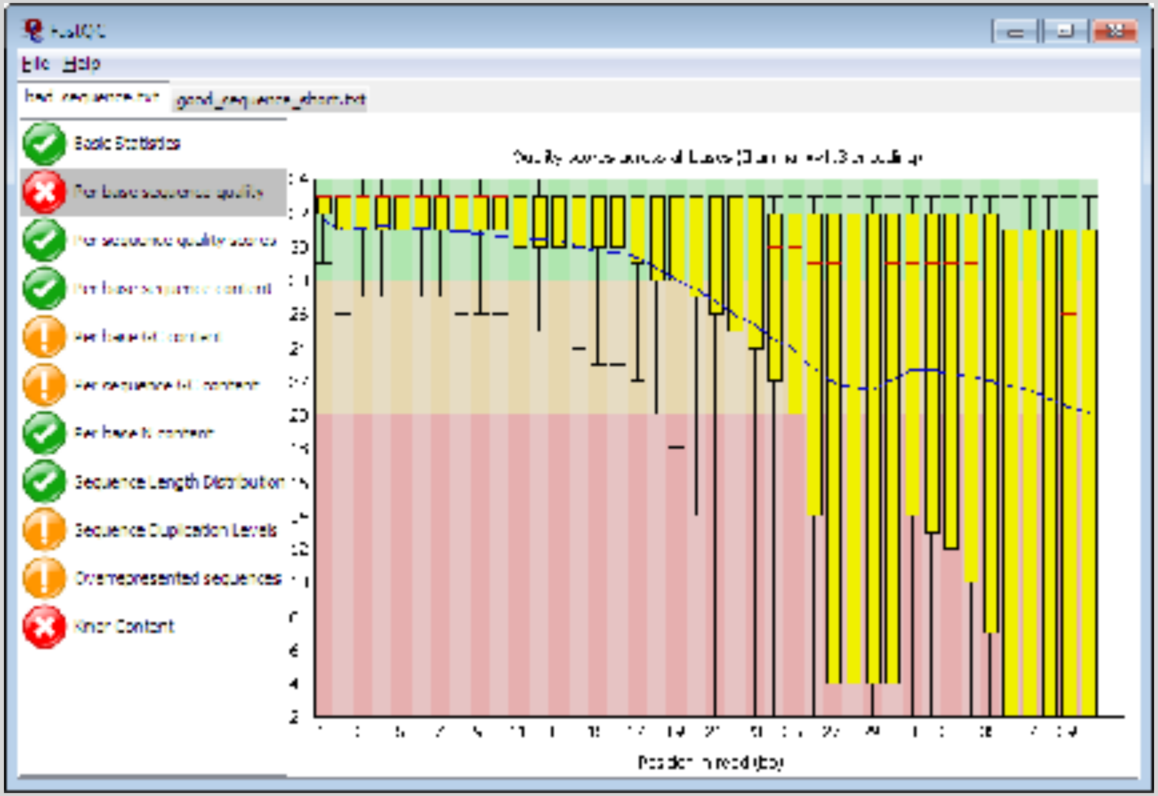
- Common problems can include:
  - Sample mix-up
  - Sample contamination
  - Machine interruption
  - DNA quality

- It is crucial that investigators examine their sequences upon first receipt before any downstream analysis is conducted

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 1.3 encoding)
34
32
30
28
26
24
22
20
18
16
14
12
10
8
6
4
2
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

SAM Format

- Sequence Alignment/Map (**SAM**) format is the almost-universal sequence alignment format for NGS
  - binary version is BAM
- It consists of a header section (lines start with ‘@’) and an alignment section
- The official specification can be found here:
  - http://samtools.sourceforge.net/SAM1.pdf

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 Utilities

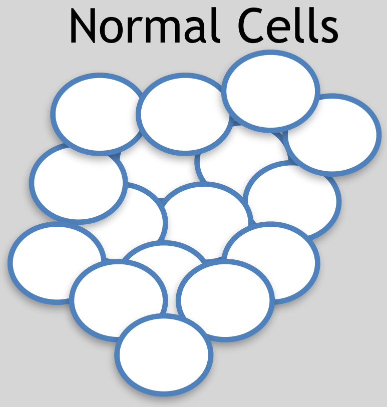
- **Samtools** is a common toolkit for analyzing and manipulating files in SAM/ BAM format
  - http://samtools.sourceforge.net/
- **Picard** is a another set of utilities that can used to manipulate and modify SAM files
  - http://picard.sourceforge.net/
- These can be used for viewing, parsing, sorting,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

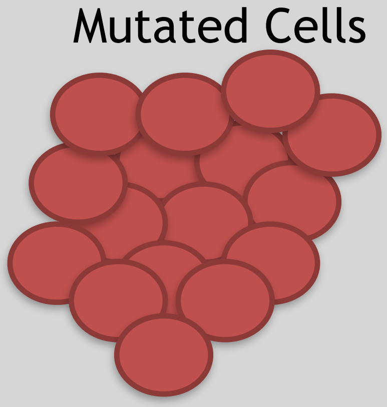
Genome Analysis Toolkit (**GATK**)

- 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/

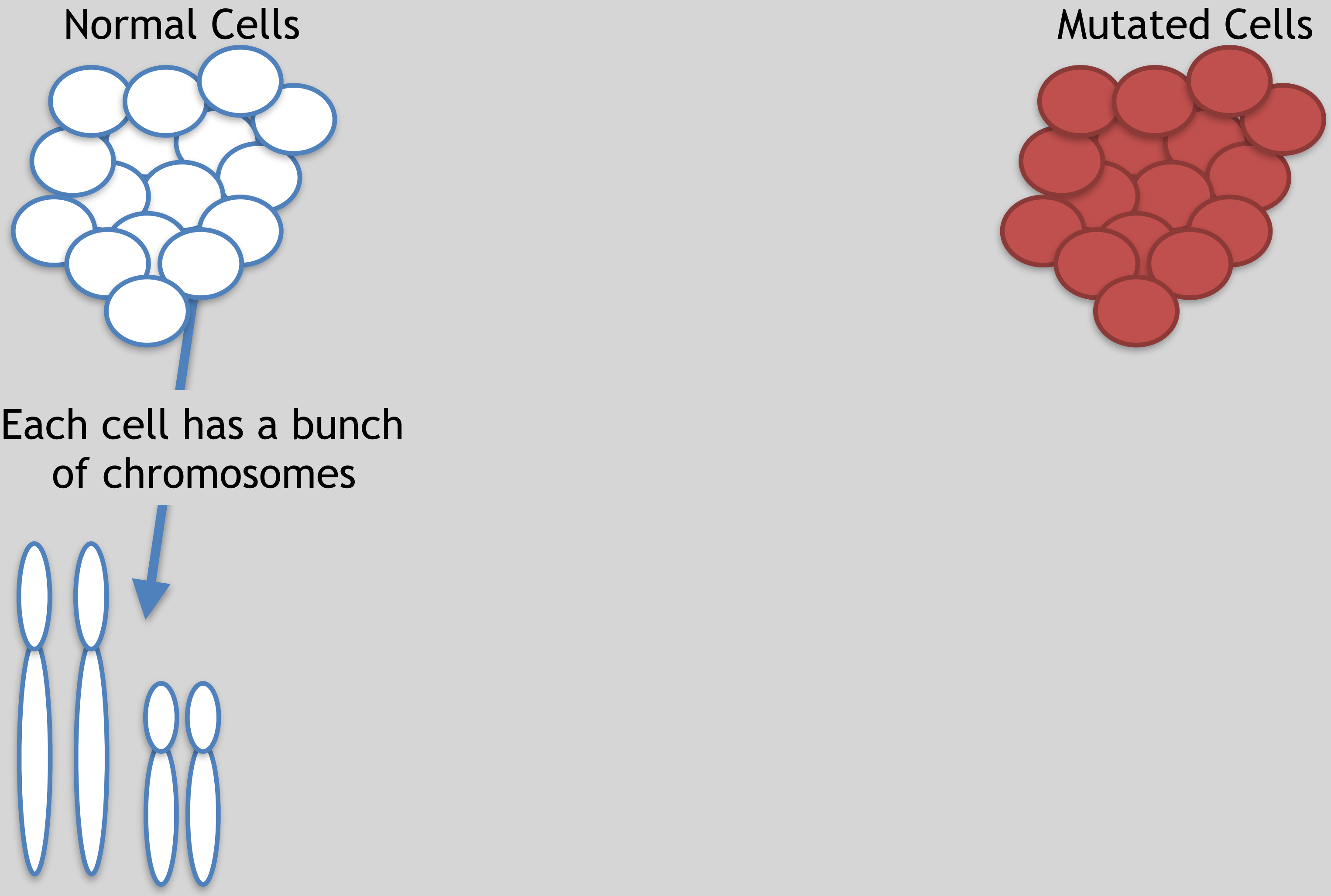
RNA Sequencing

The absolute basics





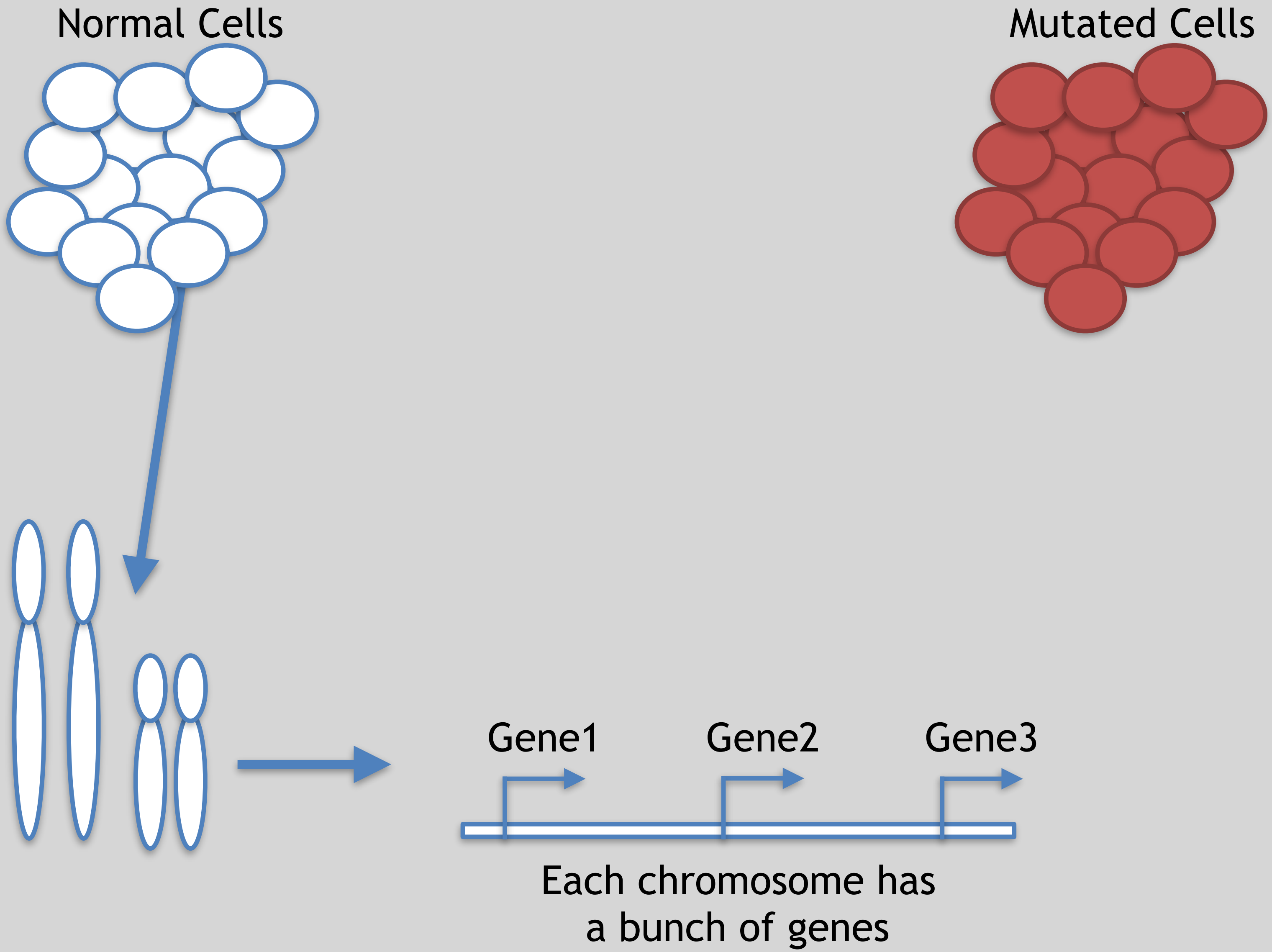
- The mutated cells behave differently than the normal cells
- We want to know what genetic mechanism is causing the difference
- One way to address this is to examine differences in gene expression via RNA sequencing…



Normal Cells

Mutated Cells

Each cell has a bunch of chromosomes



Normal Cells

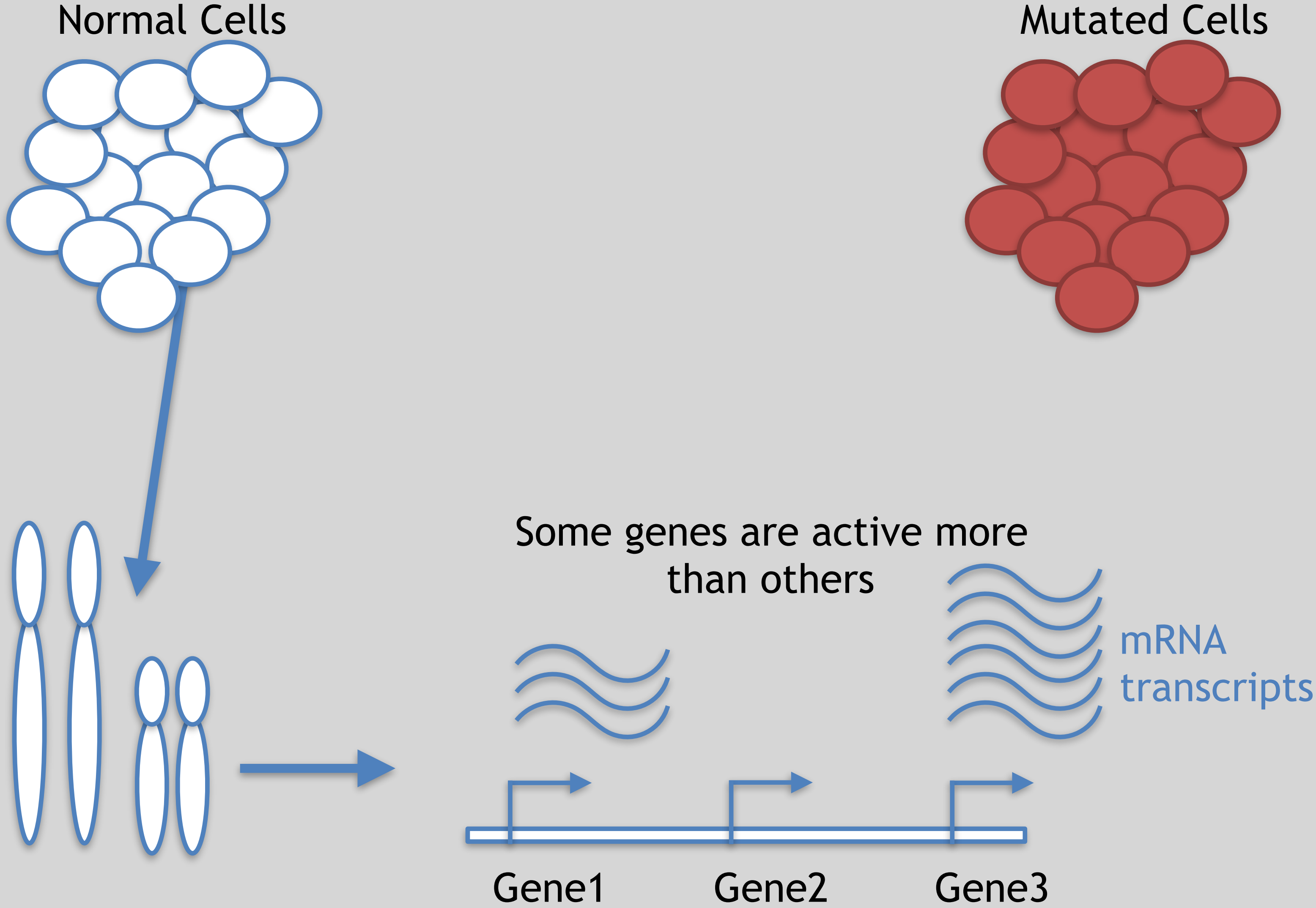
Mutated Cells

Gene1

Gene2

Gene3

Each chromosome has
a bunch of genes



Normal Cells

Mutated Cells

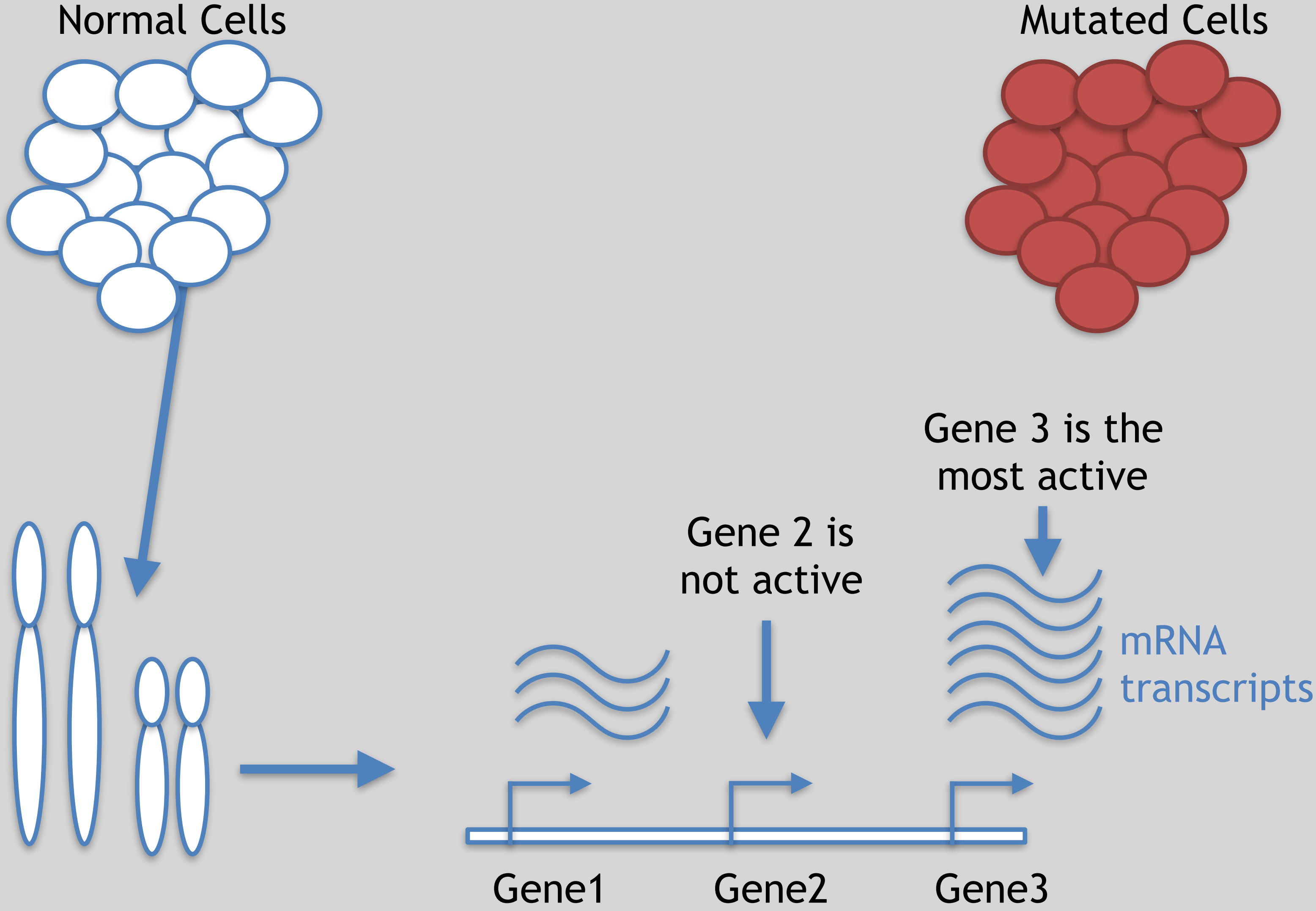
Some genes are active more than others

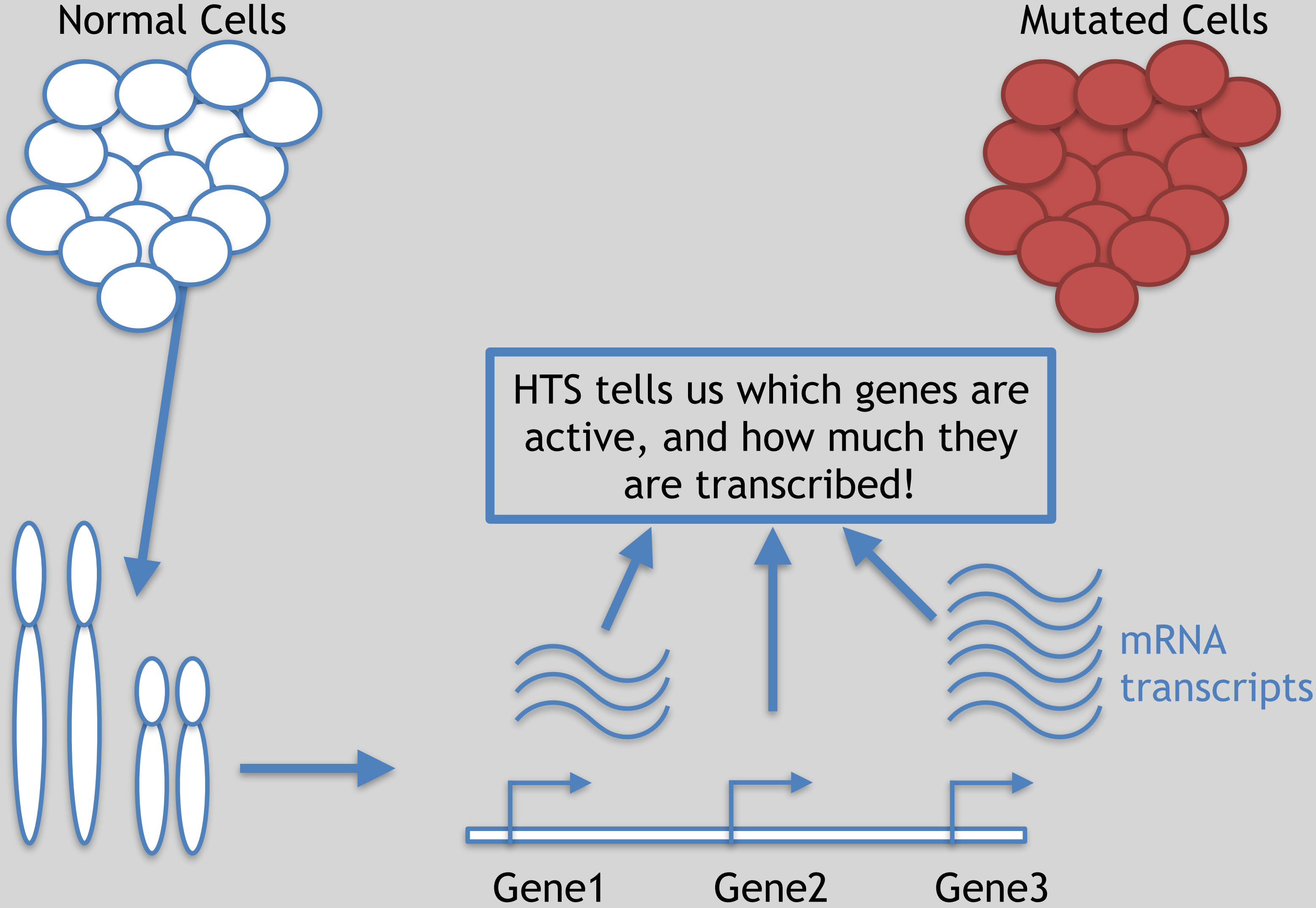
mRNA transcripts

Gene1

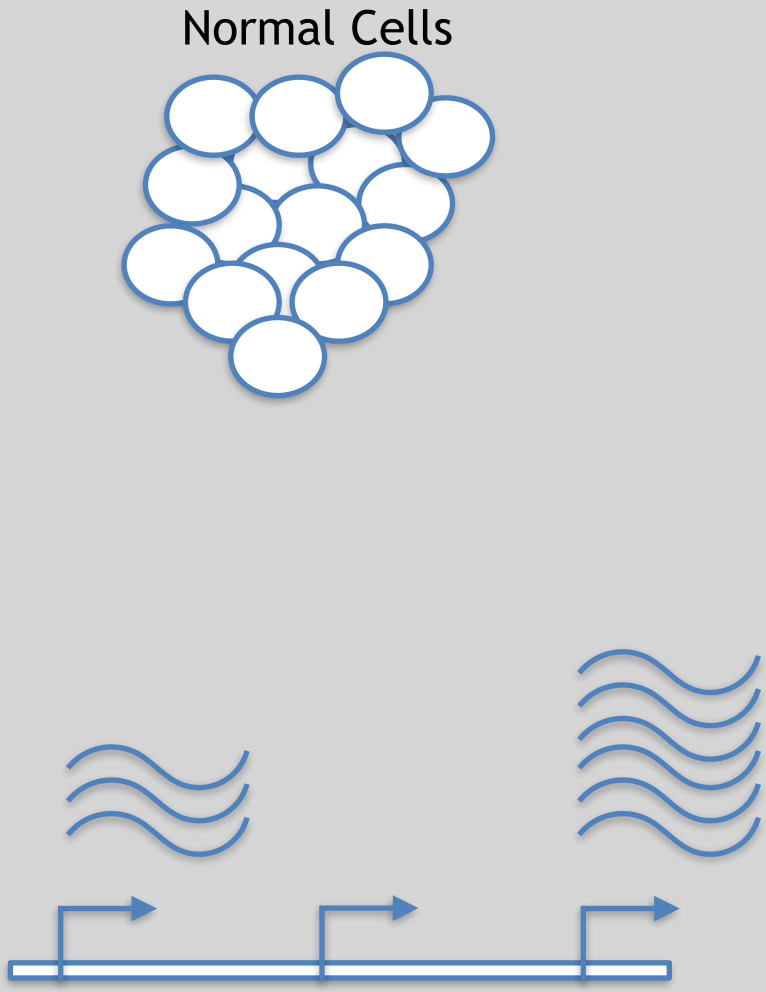
Gene2

Gene3

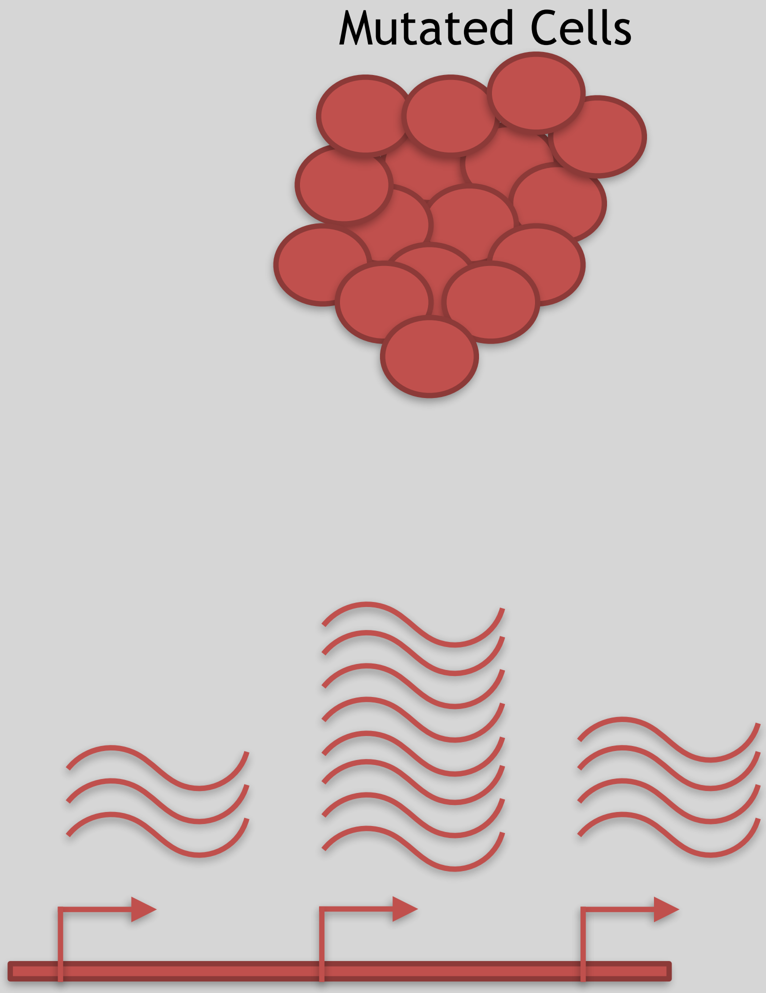

Normal Cells
Mutated Cells
Gene 3 is the most active
Gene 2 is not active
mRNA transcripts
Gene1
Gene2
Gene3



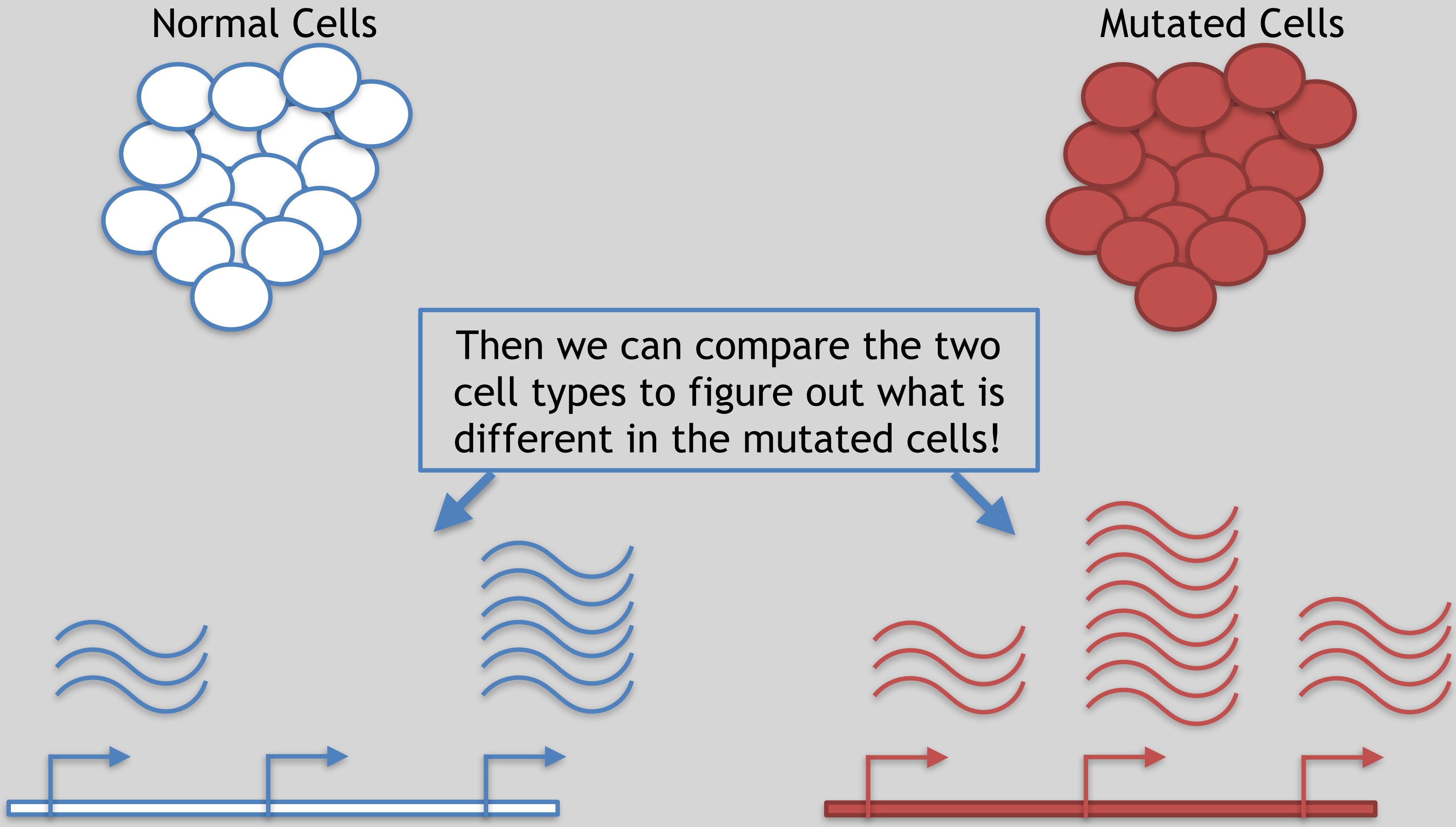
HTS tells us which genes are active, and how much they are transcribed!



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



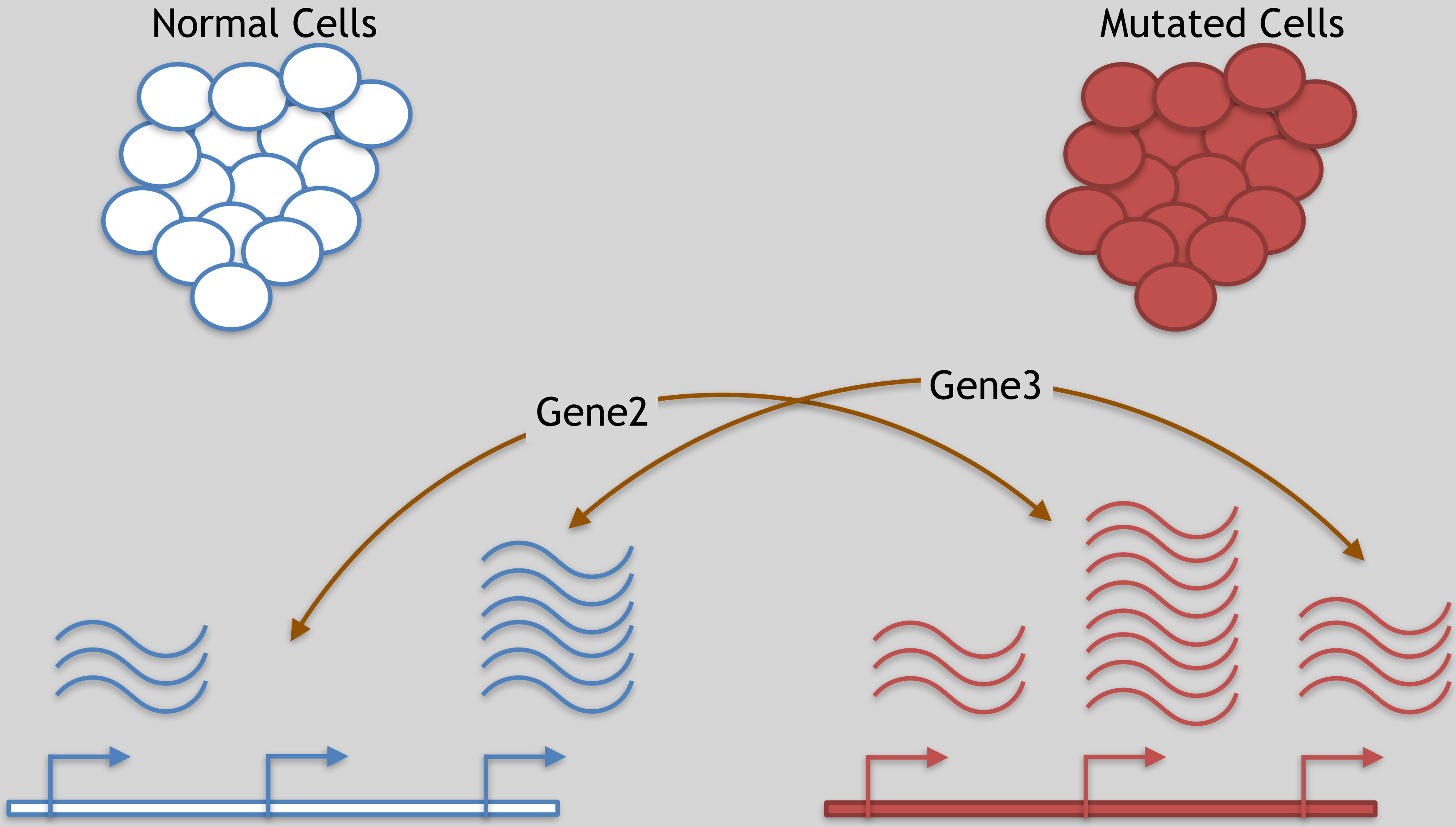
… them use it to measure gene expression in mutated cells



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**

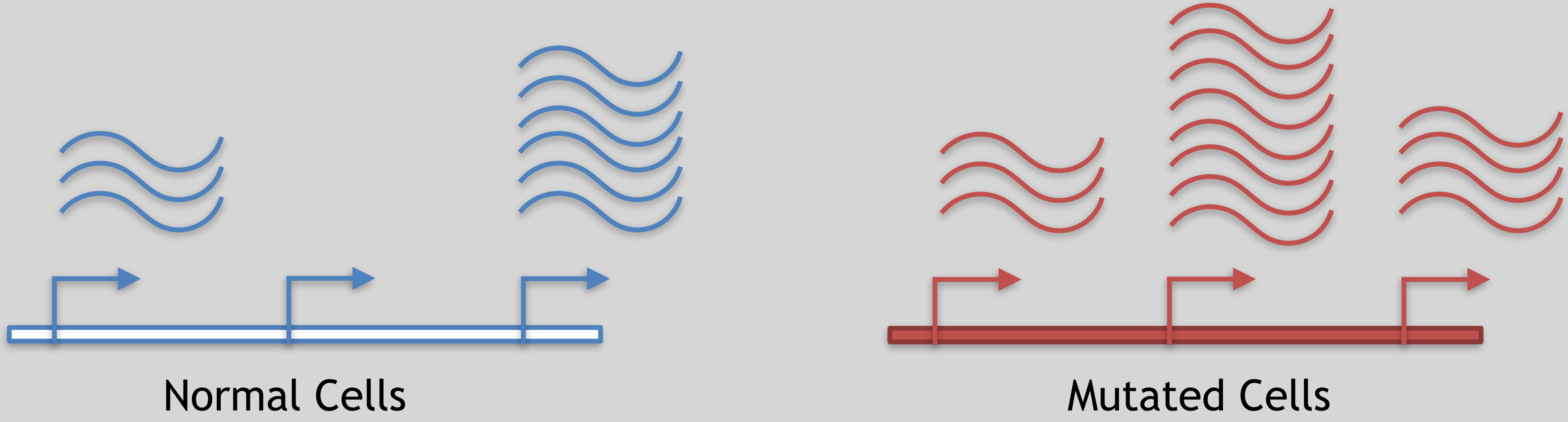
   (Often the major bottleneck to overall success!)

We will discuss each of these steps in detail (particularly the 3rd) next day!

Lets skip ahead 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 and **normalized** to arrive at our data matrix



Step 1 in any analysis is always the same:

Step 1 in any analysis is always the same:

PLOT THE DATA!!

Step 1 in any analysis is always the same:

PLOT THE DATA!!

- If there were only two genes, then plotting the data would be easy

Gene	WT-1	WT-2	WT-3
A1BG	30	5	13
AS1	24	10	18

Step 1 in any analysis is always the same:

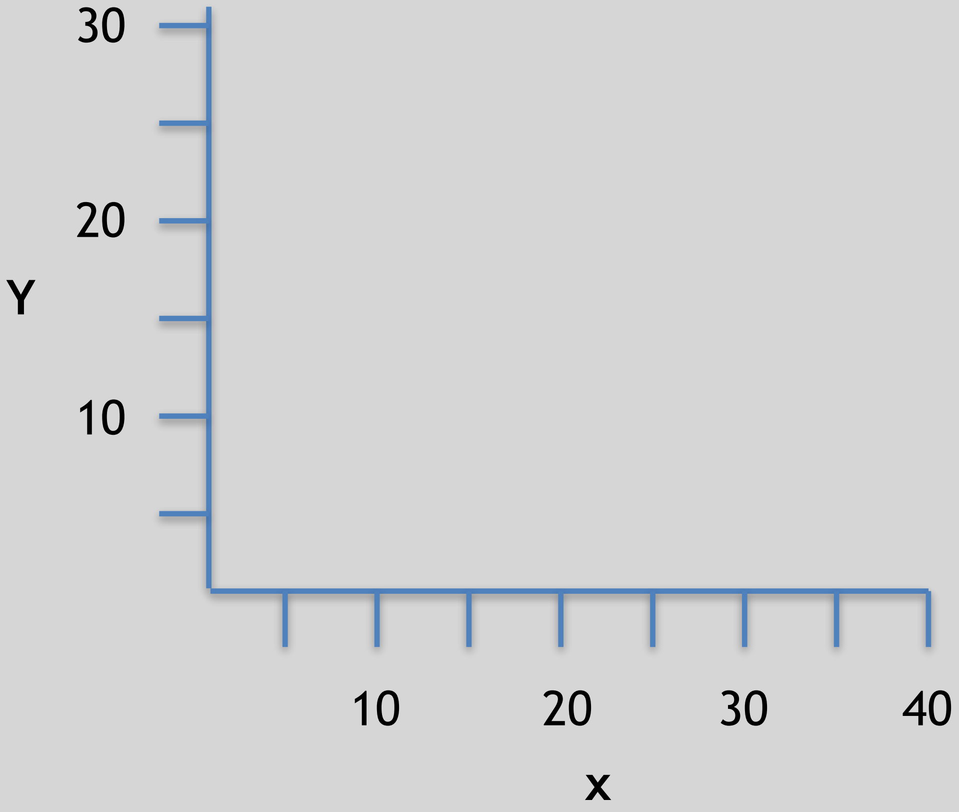
PLOT THE DATA!!

- If there were only two genes, then plotting the data would be easy

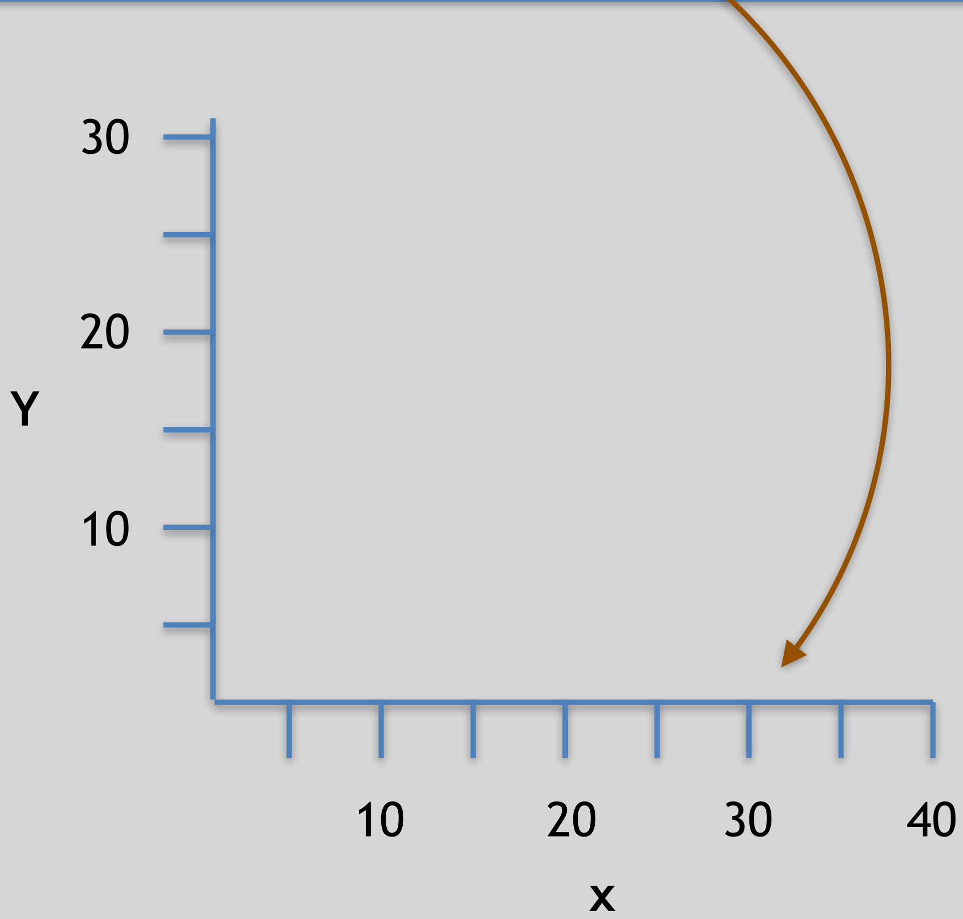
Gene	WT-1	WT-2	WT-3
x	30	5	13
y	24	10	18

Just replace the gene names with "x" and "y" and plot!

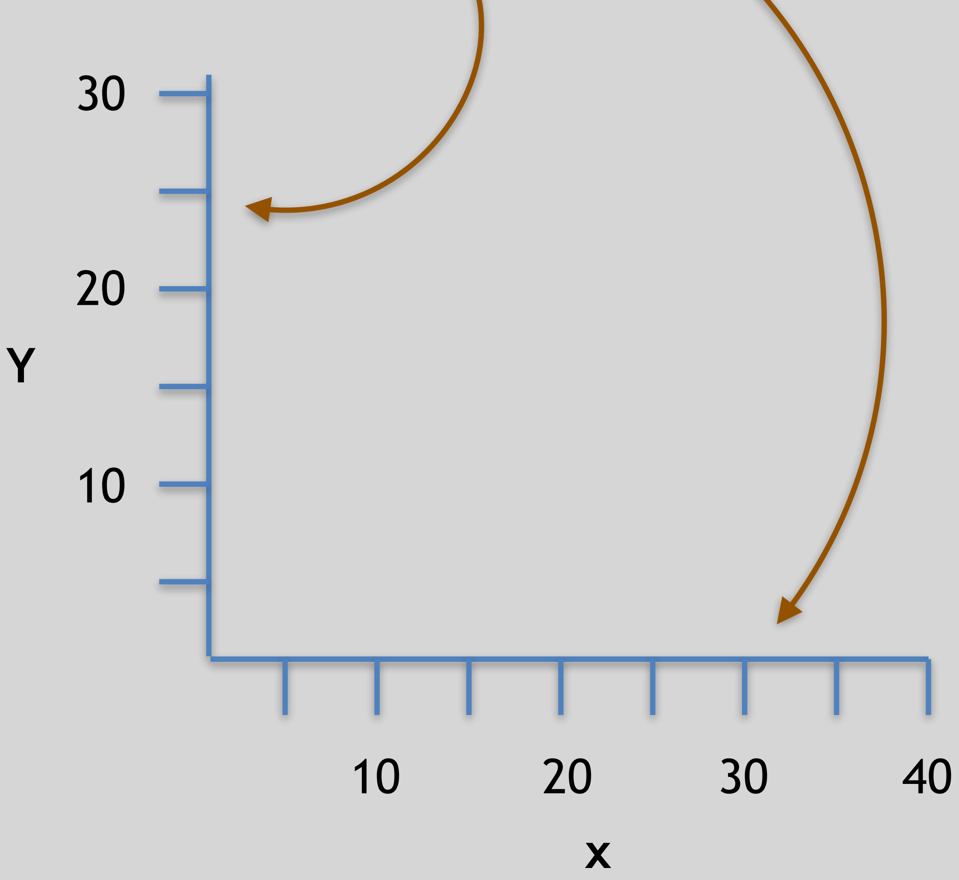
	sample-1	sample-2	sample-3
x	30	5	13
y	24	10	18



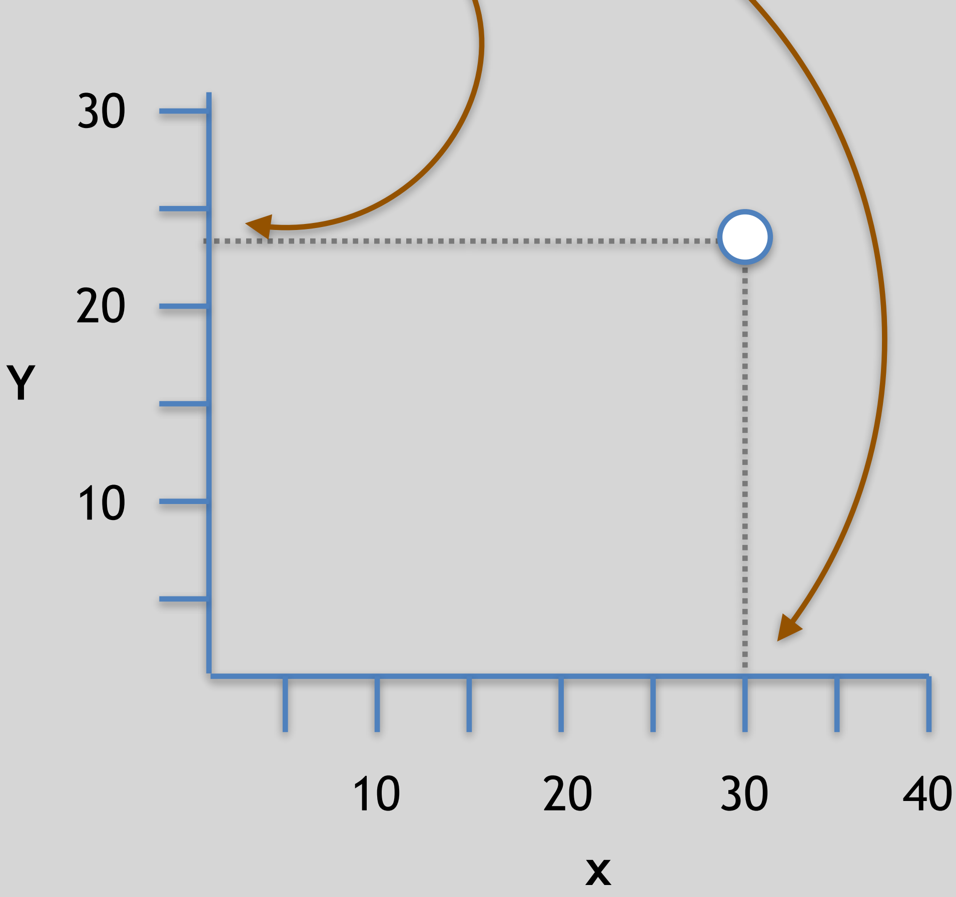
	sample-1	sample-2	sample-3
x	30	5	13
y	24	10	18



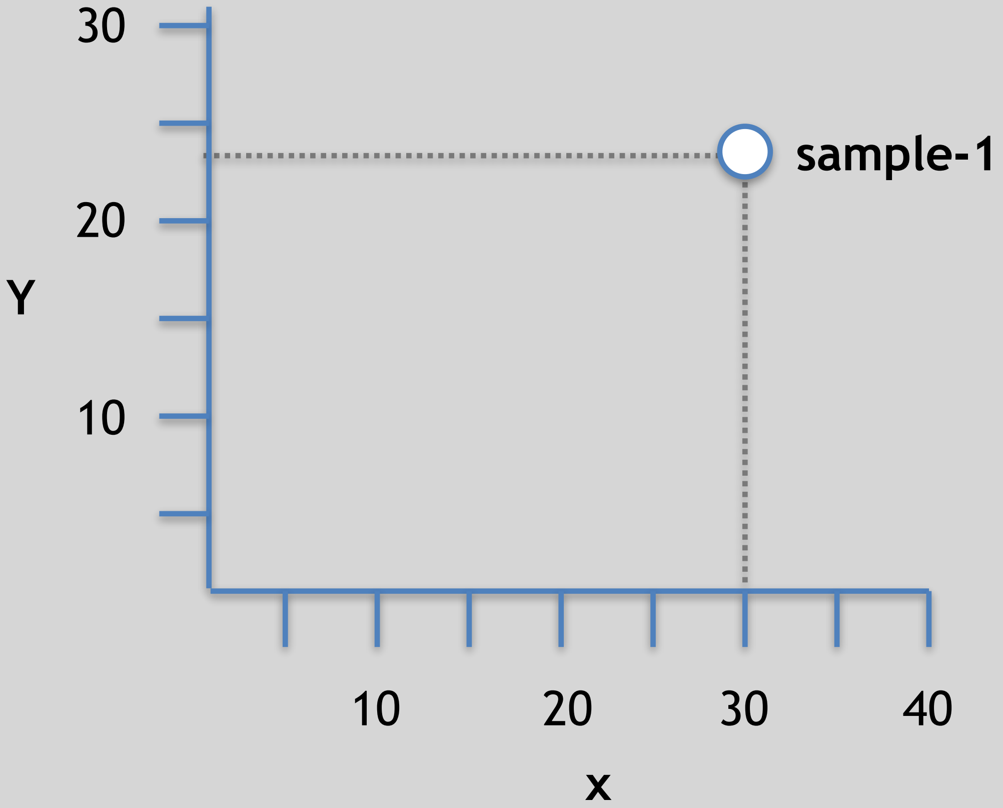
	sample-1	sample-2	sample-3
x	30	5	13
y	24	10	18



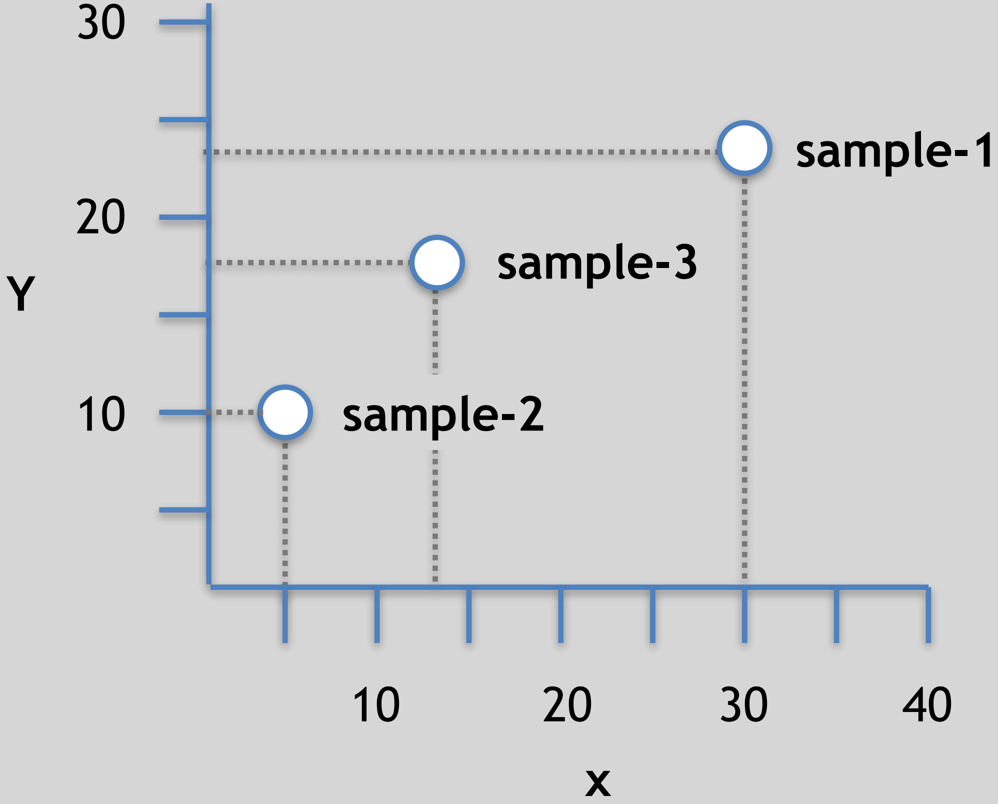
	sample-1	sample-2	sample-3
x	30	5	13
y	24	10	18



	sample-1	sample-2	sample-3
x	30	5	13
y	24	10	18



	sample-1	sample-2	sample-3
x	30	5	13
y	24	10	18



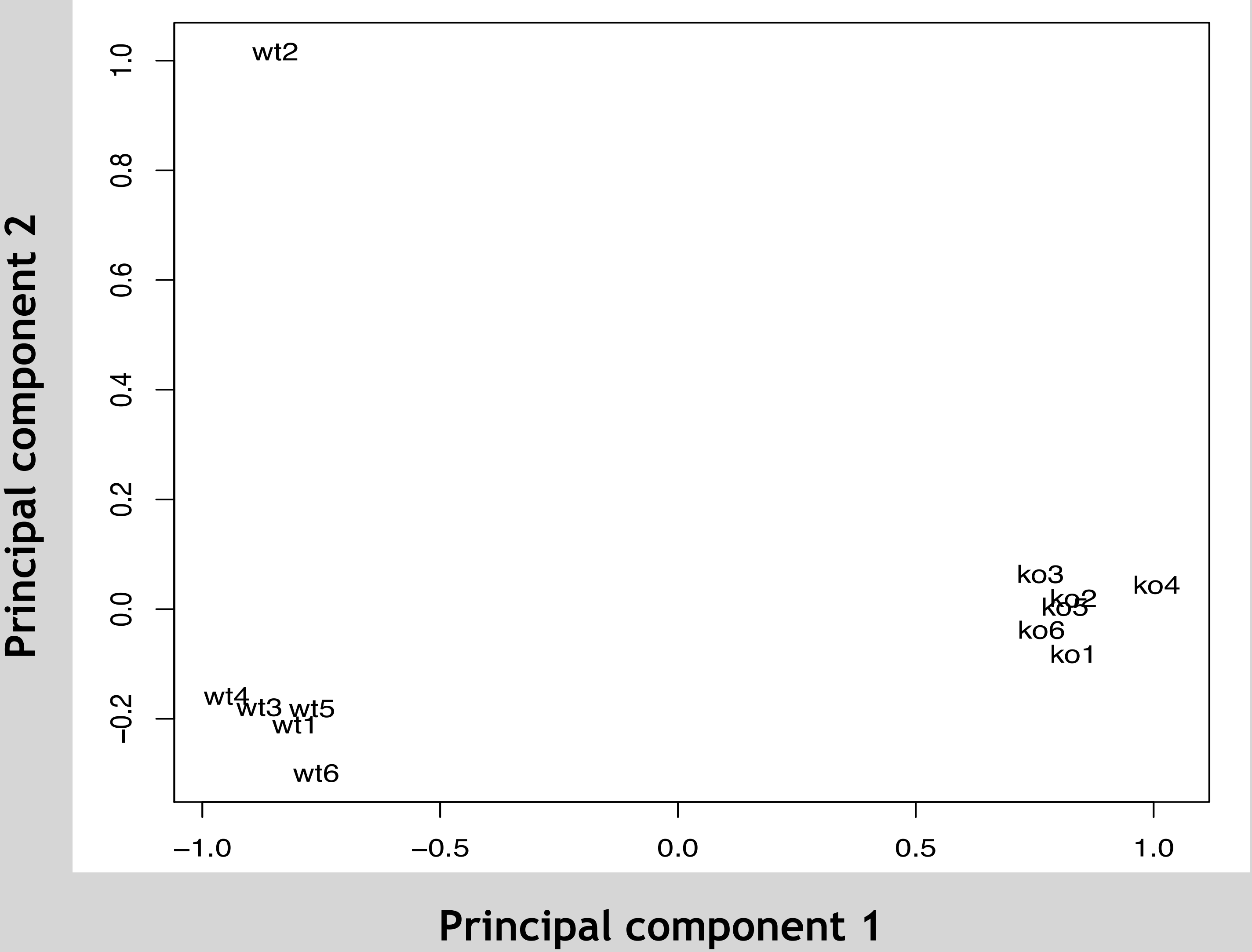
But we have 20,000 genes…

So we would need a graph with 20,000 axes to plot the data!

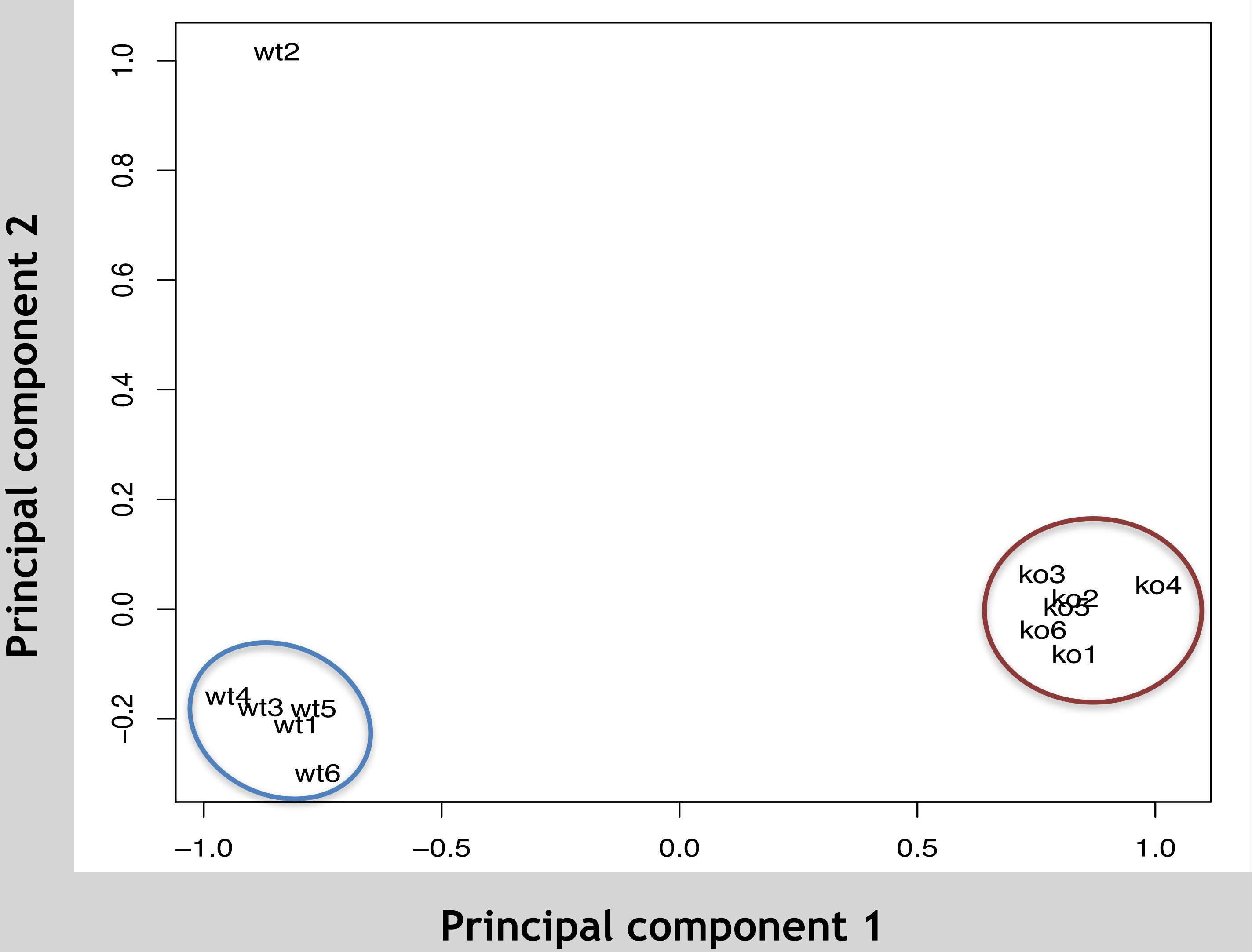
So we use PCA (principal component analysis) or something like it to plot this data.

PCA reduces the number of axes you need to display the important aspects of the data.

This is a PCA plot from a real RNA-seq experiment done on neural cells. The “wt” samples are “normal”. The “ko” samples are samples that were mutated.



This is a PCA plot from a real RNA-seq experiment done on neural cells. The “wt” samples are “normal”. The “ko” samples are samples that were mutated.



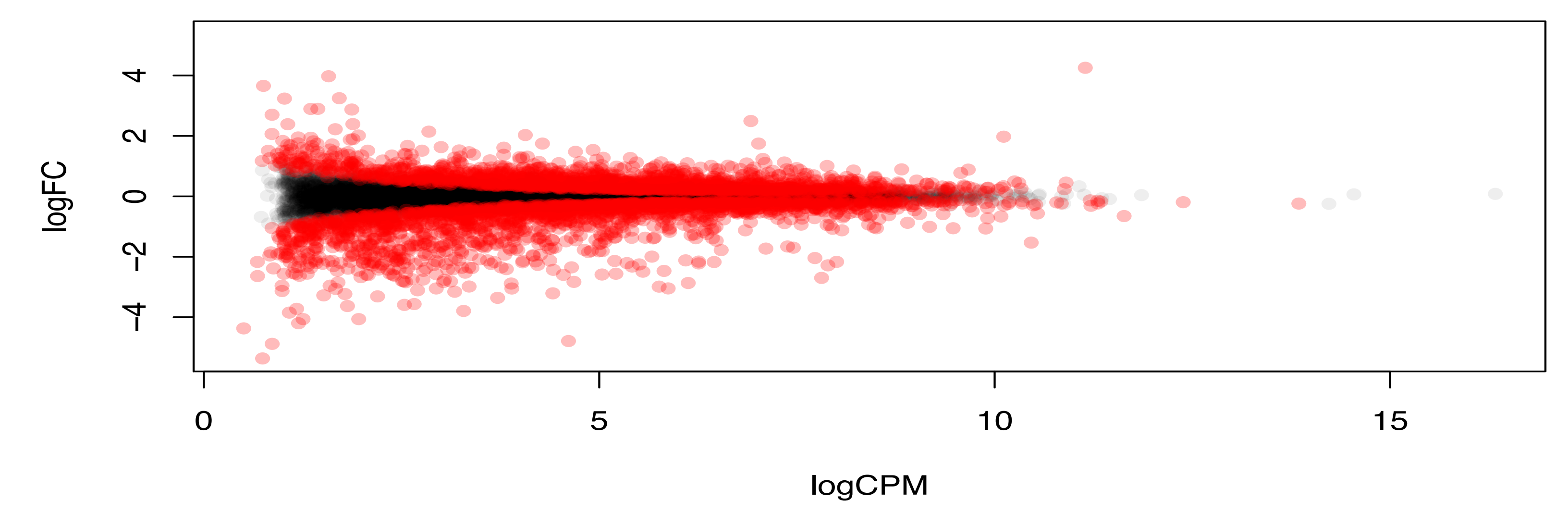
Plotting the data:

(**1**) Tells us if we can expect to find some interesting differences

(**2**) Tells us if we should exclude some samples from any down stream analysis.

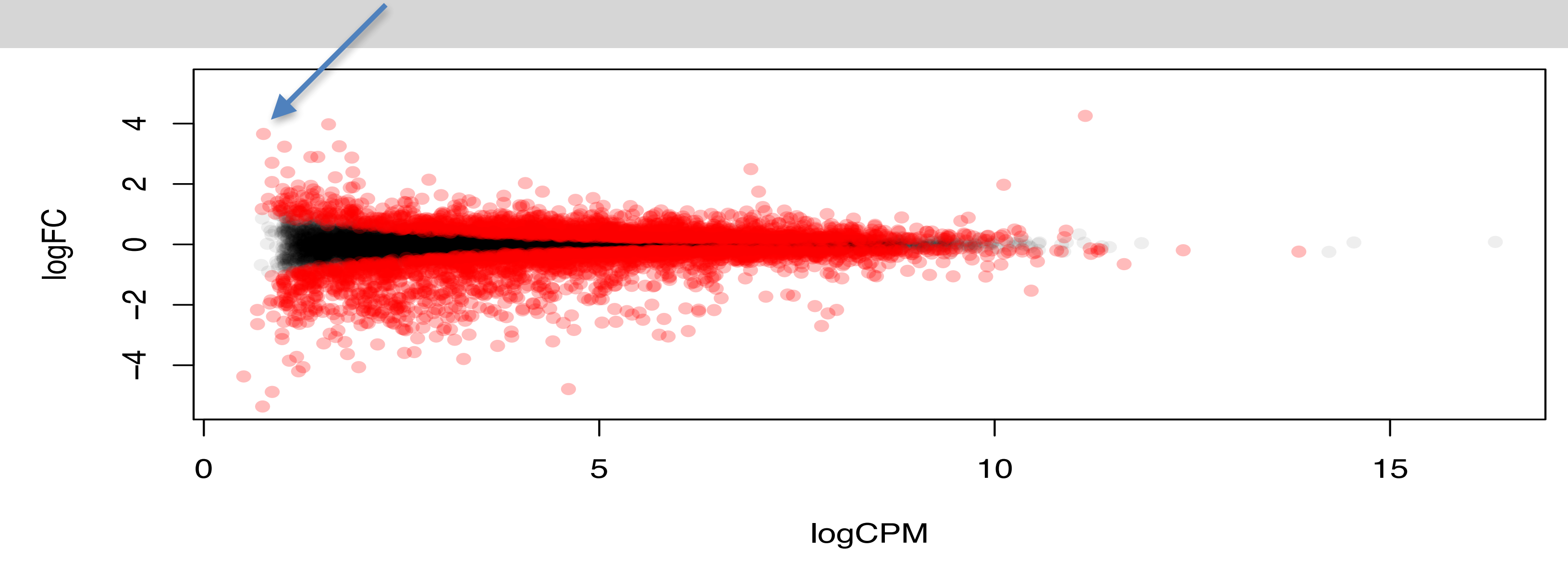
Step 2: Identify differentially expressed genes between the "normal" and "mutant" samples

This is typically done using R with either the **edgeR** or **DESeq2** packages and the results are generally displayed using graphs like this one

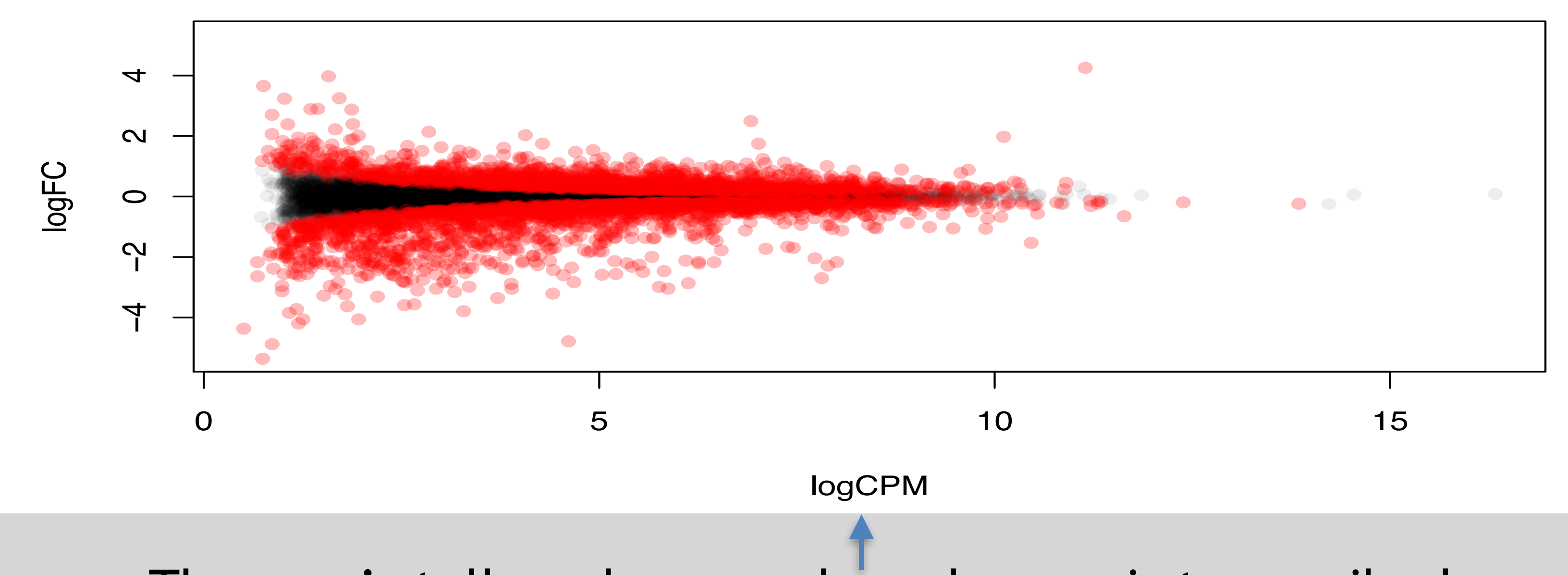


Step 2: Identify differentially expressed genes between the “normal” and “mutant” samples

A Red dot is a gene that is different between “normal” and “mutant” samples (black dots are the same).



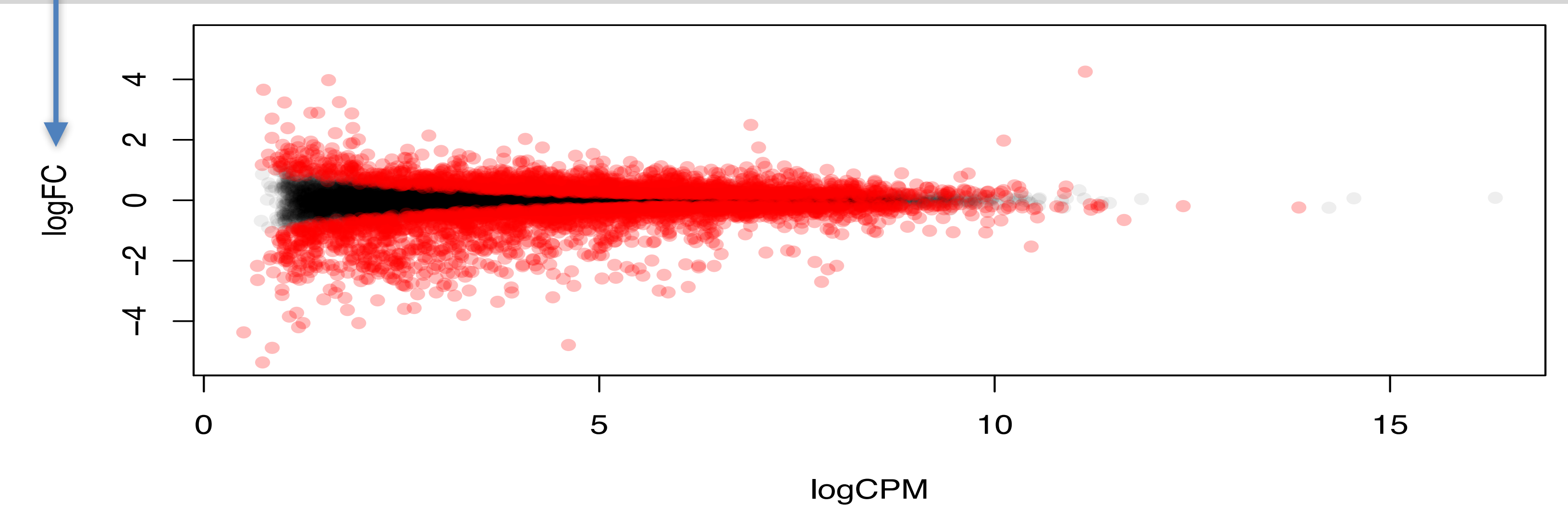
Step 2: Identify differentially expressed genes between the “normal” and “mutant” samples



The **x axis** tells us how much each gene is transcribed (CPM stands for Counts Per Million)

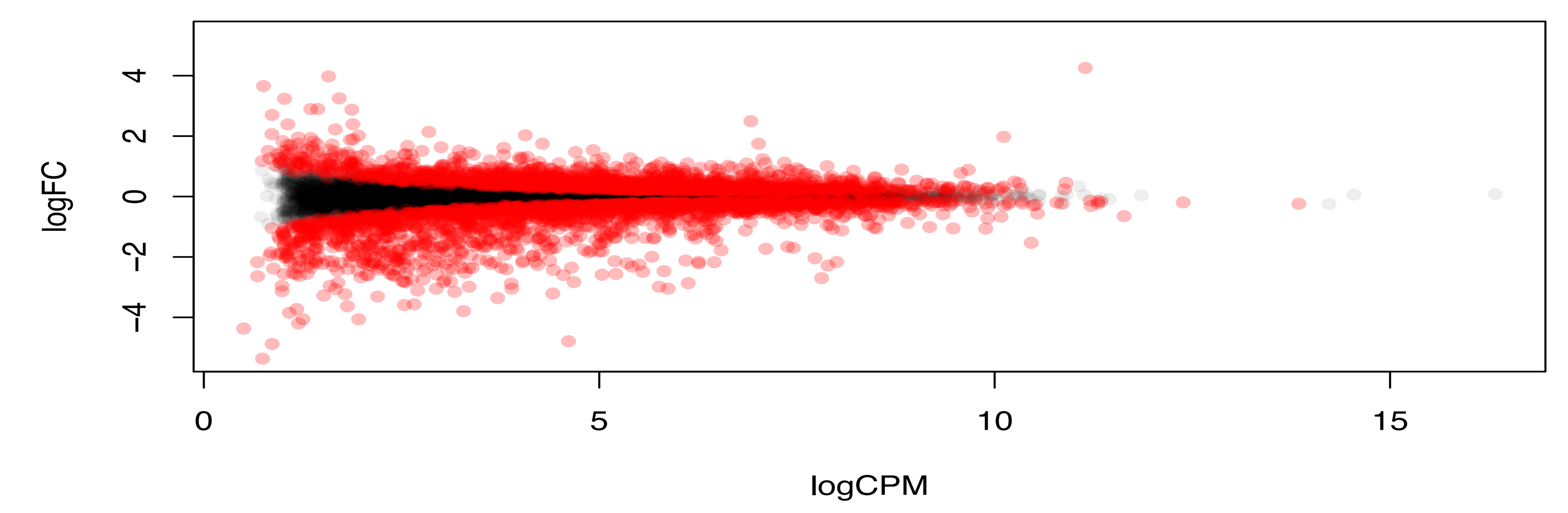
Step 2: Identify differentially expressed genes between the “normal” and “mutant” samples

The **y axis** tells you how big the relative difference is between “normal” and “mutant” (FC stands for Fold change)



The **x axis** tells us how much each gene is transcribed (CPM stands for Counts Per Million)

Step 3 and beyond: We've identified interesting genes, now what?

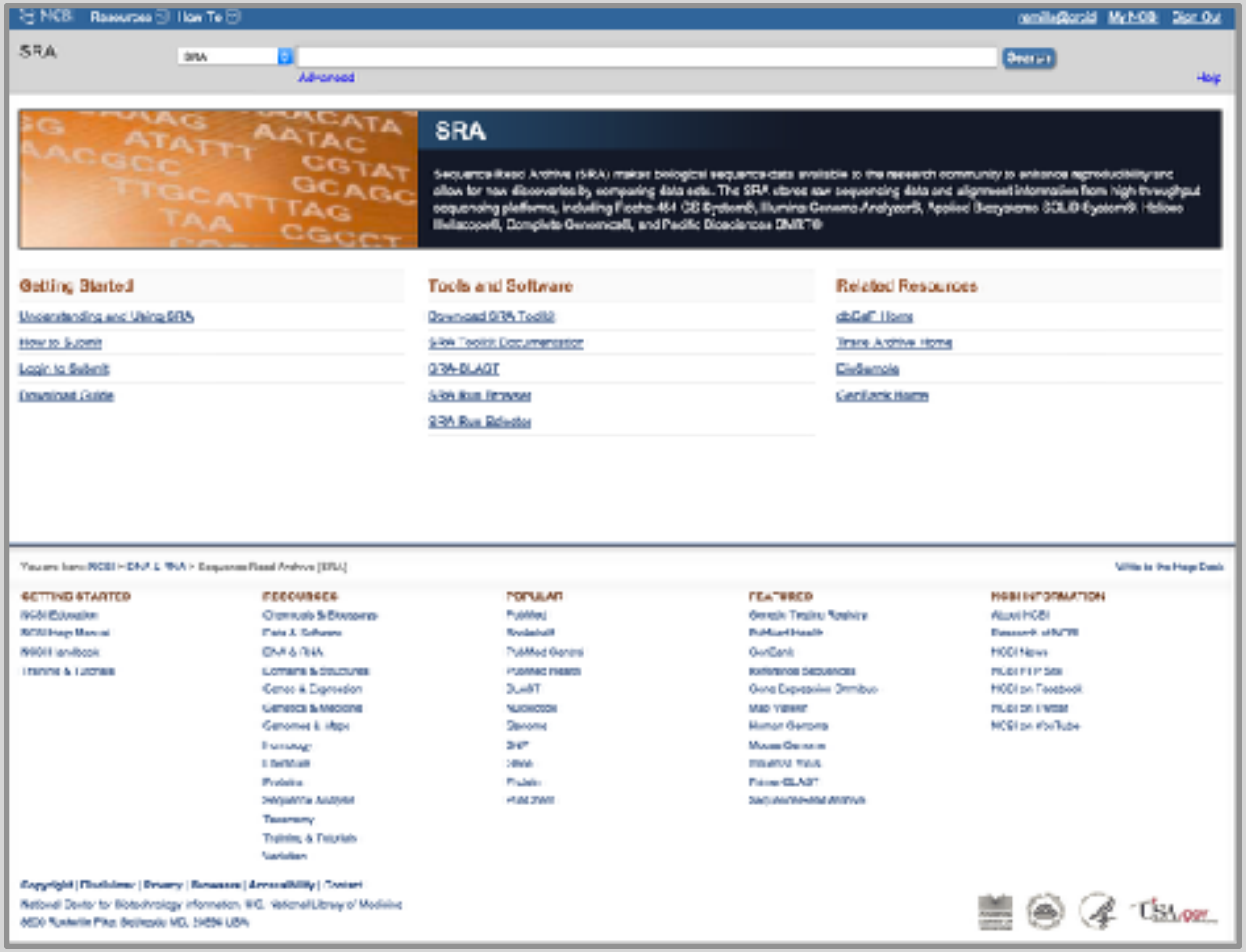


1. If you know what you're looking for, you can see if the experiment validated your hypothesis.

2. If you don't know what you're looking for, you can see if certain pathways are enriched in either the normal or mutant gene sets.

DNA- and RNA-Seq Databases

NCBI Short Read Archive (SRA):
http://www.ncbi.nlm.nih.gov/sra



SRA

Sequence Read Archive (SRA) makes biological sequence data available to the research community to enhance reproducibility and allow for new discoveries by comparing data sets. The SRA stores raw sequencing data and alignment information from high-throughput sequencing platforms, including Roche 454 GS System®, Illumina Genome Analyzer®, Applied Biosystems SOLiD System®, Helicos Heliscope®, Complete Genomics®, and Pacific Biosciences SMRT®

Getting Started

- Understanding and Using SRA
- How to Submit
- Login to Submit
- Download Guide

Tools and Software

- Download SRA Toolkit
- SRA Toolkit Documentation
- SRA-BLAST
- SRA Run Browser
- SRA Run Selector

Related Resources

- dbGaP Home
- Trace Archive Home
- BioSample
- GenBank Home

You are here: NCBI > DNA & RNA > Sequence Read Archive (SRA)

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

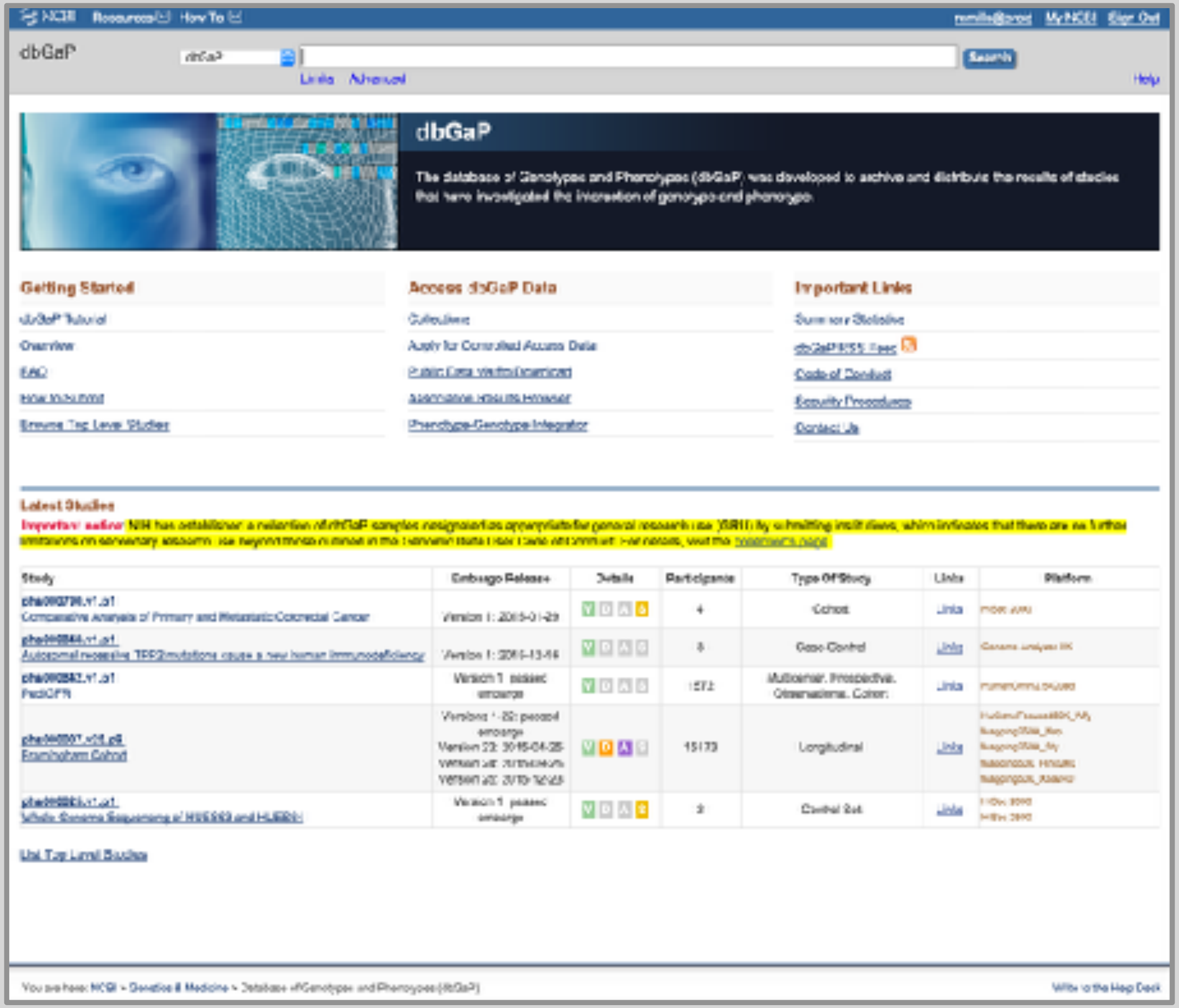
- About NCBI
- Research at NCBI
- NCBI News
- NCBI FTP Site
- NCBI on Facebook
- NCBI on Twitter
- NCBI on YouTube

Copyright | Disclaimer | Privacy | Browsers | Accessibility | Contact

National Center for Biotechnology Information, U.S. National Library of Medicine
8600 Rockville Pike, Bethesda MD, 20894 USA

Protected Data - dbGaP

NCBI Database of Genotypes and Phenotypes (dbGaP): http://www.ncbi.nlm.nih.gov/sra



dbGaP

The database of Genotypes and Phenotypes (dbGaP) was developed to archive and distribute the results of studies that have investigated the interaction of genotype and phenotype.

Getting Started

- dbGaP Tutorial
- Overview
- FAQ
- How to Submit
- Browse Top Level Studies

Access dbGaP Data

- Collections
- Apply for Controlled Access Data
- Public Data via ftp Download
- Association Results Browser
- Phenotype-Genotype Integrator

Important Links

- Summary Statistics
- dbGaP RSS Feed
- Code of Conduct
- Security Procedures
- Contact Us

Latest Studies

Important notice: NIH has established a collection of dbGaP samples designated as appropriate for general research use (GRU) by submitting institutions, which indicates that there are no further limitations on secondary research use beyond those outlined in the Genomic Data User Code of Conduct. For details, visit the [illegible] page.

Study	Embargo Release	Details	Participants	Type Of Study	Links	Platform
phs000790.v1.p1 Comparative Analysis of Primary and Metastatic Colorectal Cancer	Version 1: 2015-01-29	V D A S	4	Cohort	Links	[illegible]
phs000844.v1.p1 Autosomal recessive TPP2 mutations cause a new human immunodeficiency	Version 1: 2016-13-16	V D A S	6	Case-Control	Links	Genome Analyzer IIX
phs000842.v1.p1 PediCFR	Version 1: passed embargo	V D A S	1572	Multicenter, Prospective, Observational, Cohort	Links	[illegible]
phs000007.v25.p9 Framingham Cohort	Versions 1-22: passed embargo Version 23: 2015-04-25 Version 24: [illegible] Version 25: 2015-12-23	V D A S	15173	Longitudinal	Links	[illegible] [illegible] [illegible] [illegible] [illegible]
phs000[illegible].v1.p1 Whole Genome Sequencing of HUES[illegible] and HUES[illegible]	Version 1: passed embargo	V D A S	2	Control Set	Links	[illegible] [illegible]

List Top Level Studies

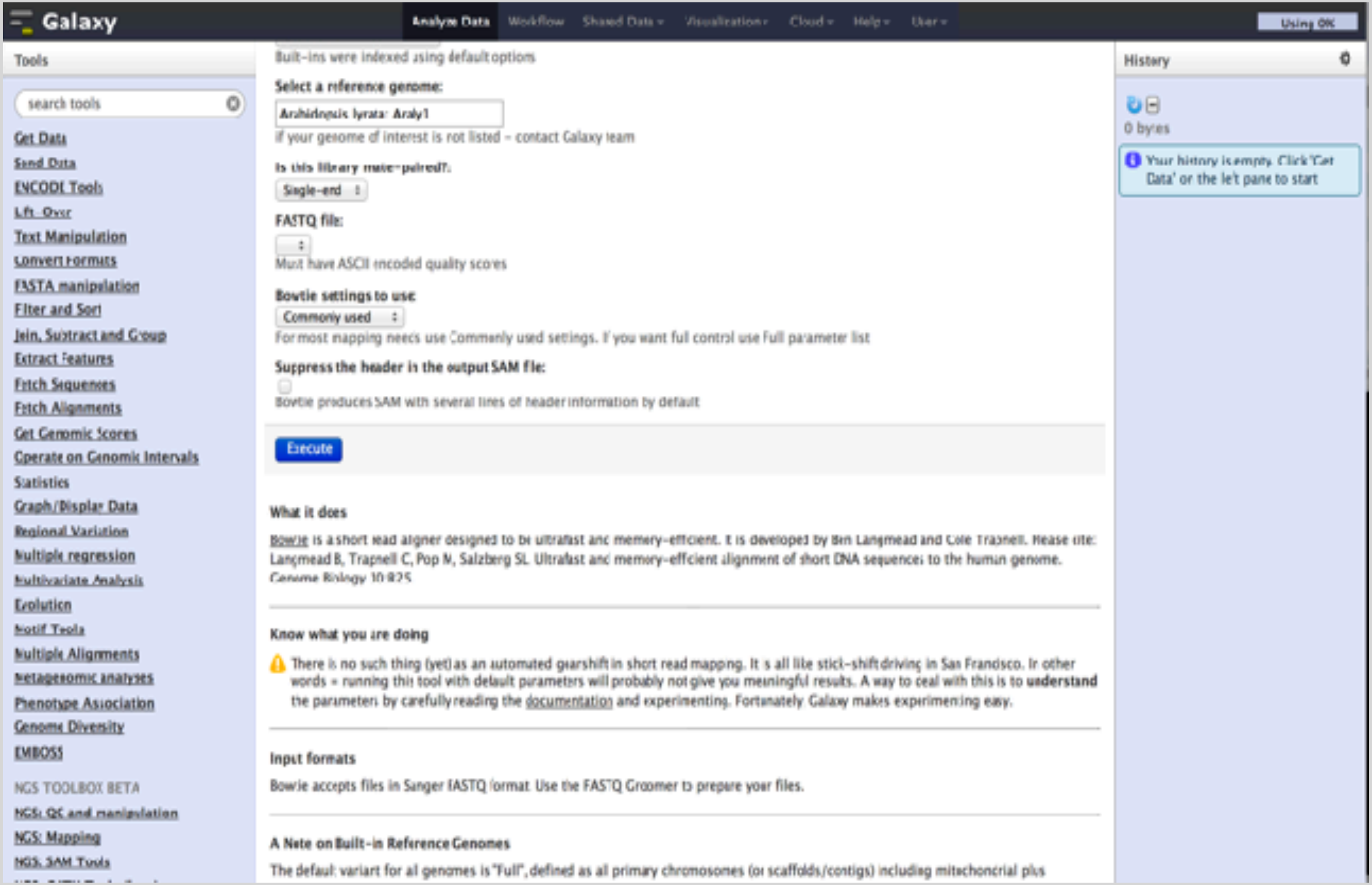
You are here: NCBI > Genetics & Medicine > Database of Genotypes and Phenotypes (dbGaP)

Write to the Help Desk

Today we will use **Galaxy**

- Galaxy is a useful web-based application for the manipulation of NGS data sets
  - https://main.g2.bx.psu.edu/
- It contains many common analysis utilities and provides a somewhat standardized approach to analyzing NGS data
- However, it requires the uploading of data to their server, which typically precludes its application to protected data sets (e.g. human samples) - Or you have to build your own server
- You are also limited to only those tools which have been incorporated into their system

Galaxy Website



Built-ins were indexed using default options

Select a reference genome:
Arabidopsis lyrata: Araly1
If your genome of interest is not listed - contact Galaxy team

Is this library mate-paired?:
Single-end

FASTQ file:

Must have ASCII encoded quality scores

Bowtie settings to use:
Commonly used
For most mapping needs use Commonly used settings. If you want full control use Full parameter list

Suppress the header in the output SAM file:
□
Bowtie produces SAM with several lines of header information by default

What it does

Bowtie is a short read aligner designed to be ultrafast and memory-efficient. It is developed by Ben Langmead and Cole Trapnell. Please cite: Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biology 10:R25

Know what you are doing

⚠ There is no such thing (yet) as an automated gearshift in short read mapping. It's all like stick-shift driving in San Francisco. In other words = running this tool with default parameters will probably not give you meaningful results. A way to deal with this is to **understand** the parameters by carefully reading the documentation and experimenting. Fortunately Galaxy makes experimenting easy.

Input formats

Bowtie accepts files in Sanger FASTQ format. Use the FASTQ Groomer to prepare your files.

A Note on Built-in Reference Genomes

The default variant for all genomes is "Full", defined as all primary chromosomes (or scaffolds/contigs) including mitochondrial plus

Your history is empty. Click 'Get Data' on the left pane to start

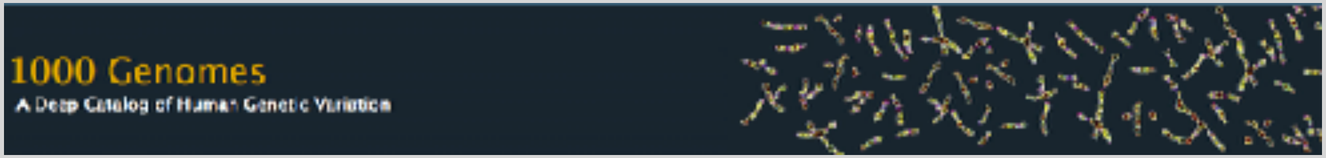
Hands-on Time!

https://bioboot.github.io/bggn213_f17/class-material/lecture14-BGGN213_F17.pdf

Additional Slides follow for Reference

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



“Variety’s the very spice of life”

–William Cowper, 1785

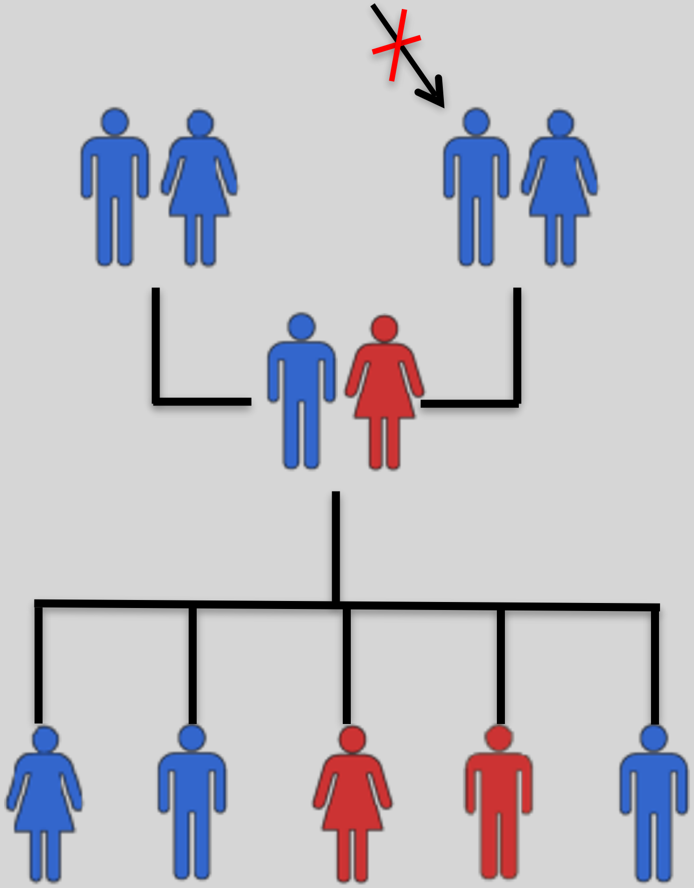
“Variation is the spice of life”

–Kruglyak & Nickerson, 2001

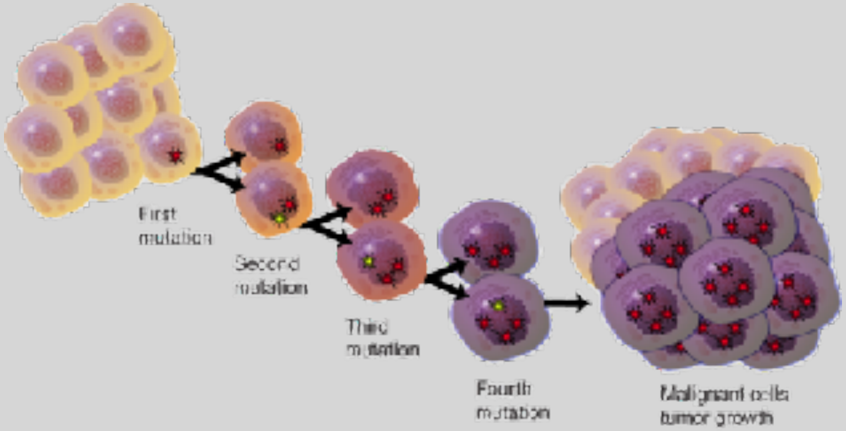
- While the sequencing of the human genome was a great milestone, the DNA from a single person is not representative of the millions of potential differences that can occur between individuals
- These unknown genetic variants could be the cause of many phenotypes such as differing morphology, susceptibility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

Mutation vs Polymorphism

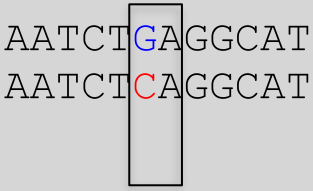
- A mutation must persist to some extent within a population to be considered polymorphic
  - >1% frequency is often used
- Germline mutations that are not polymorphic are considered rare variants

> *"From the standpoint of the neutral theory, the rare variant alleles are simple those alleles whose frequencies within a species happen to be in a low-frequency range (0,q), whereas polymorphic alleles are those whose frequencies happen to be in the higher-frequency range (q, 1-q), where I arbitrarily take q = 0.01. Both represent a phase of molecular evolution."*
>
> *-Motoo Kimura*

Kimura M (1983) Mol. Biol. Evol., 1(1), pp. 84-93

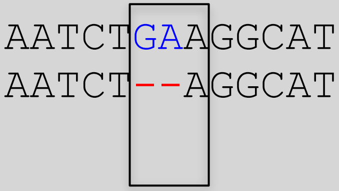
Types of Genomic Variation

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



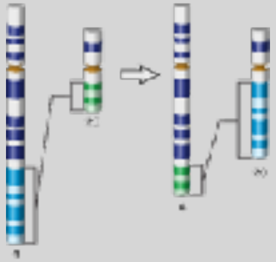
```
AATCTGAGGCAT
AATCTCAGGCAT
```

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



```
AATCTGAAGGCAT
AATCT--AGGCAT
```

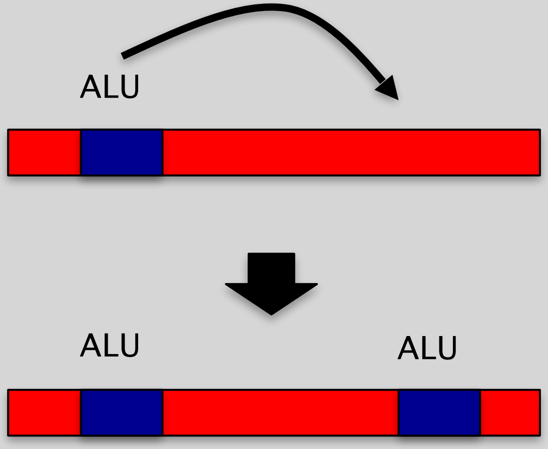
- Structural Variation (SVs) – medium to large sized rearrangements of chromosomal DNA



Darryl Leja, Courtesy: National Human Genome Research Institute.

Variant Subtypes: Repetitive Elements

Mobile Elements / Retrotransposons



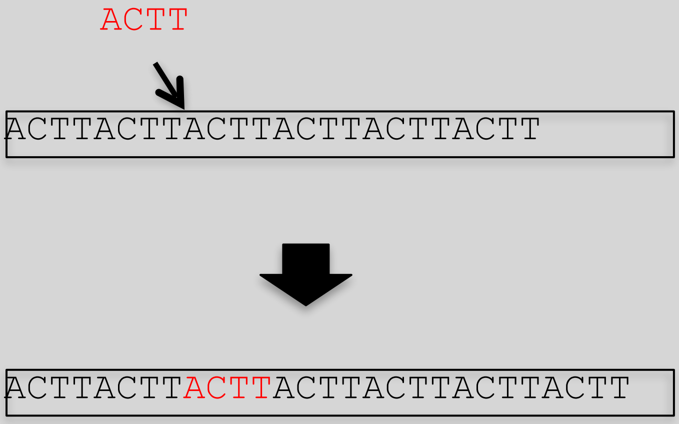
ALU

ALU

ALU

(in humans, primarily ALU, LINE, and SVA)

Repeat Expansions

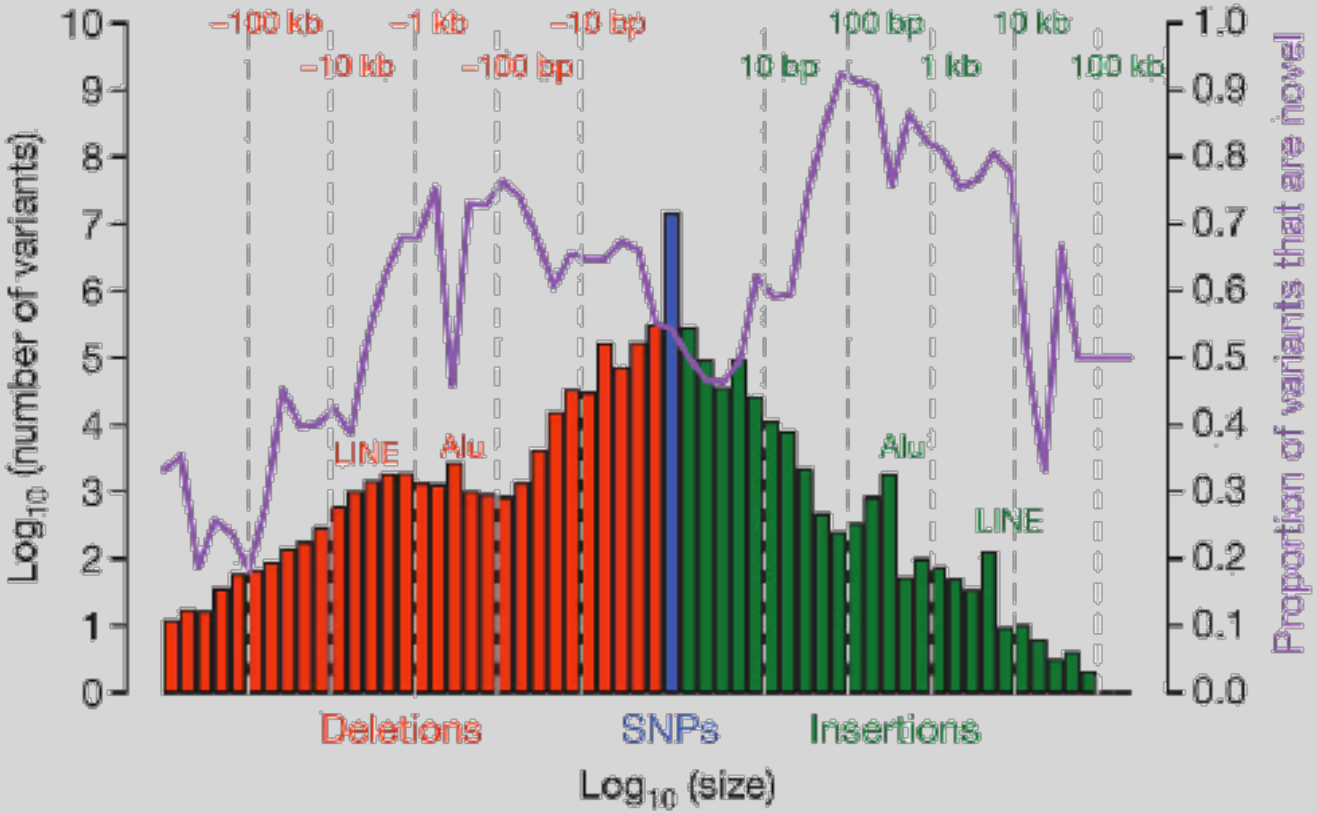


ACTT

ACTTACTTACTTACTTACTTACTT

ACTTACTTACTTACTTACTTACTTACTT

Variant Length Distribution



1000 Genomes Project, Nature, 2010

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

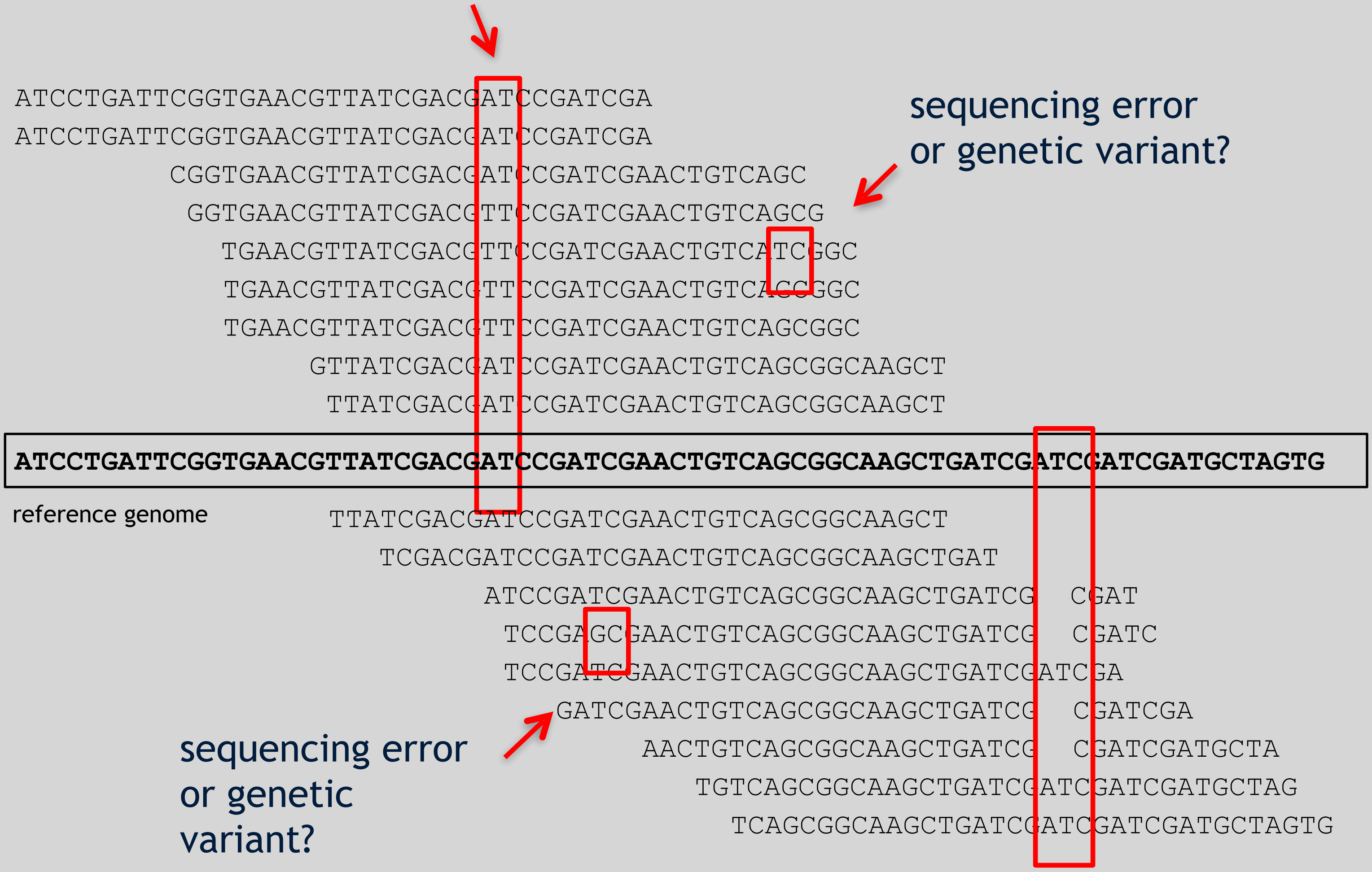
Numbers change depending on ancestry!

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

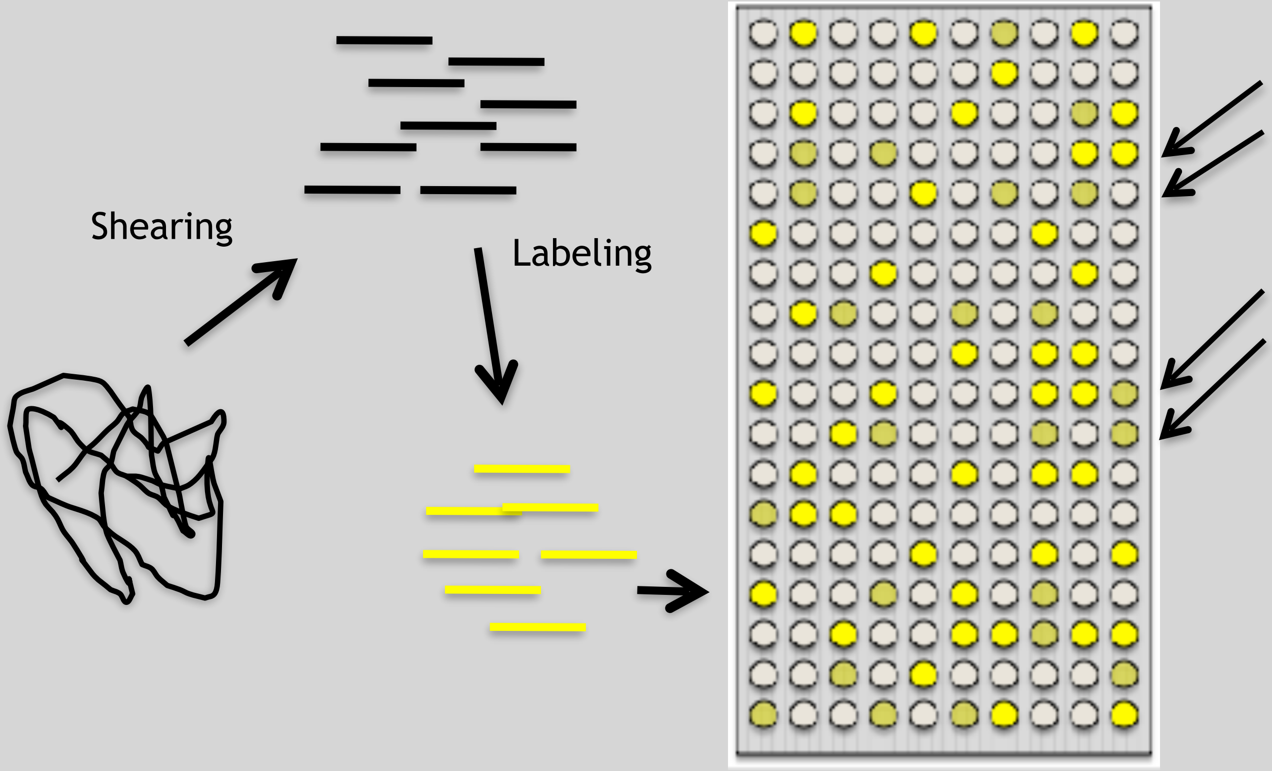
sequencing error or genetic variant?

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



TAACGATGAATCTTAGGCATCGCGC

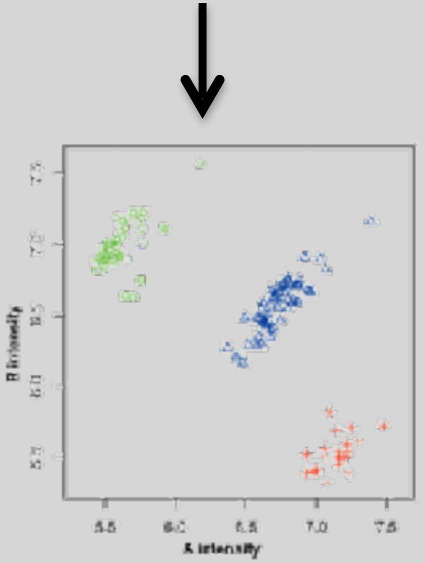
TAACGATGAATCGTAGGCATCGCGC

genotype: T/T

GGCTTAAGTACCCTATGGATTACGG

GGCTTAAGTACCTTATGGATTACGG

genotype: C/T



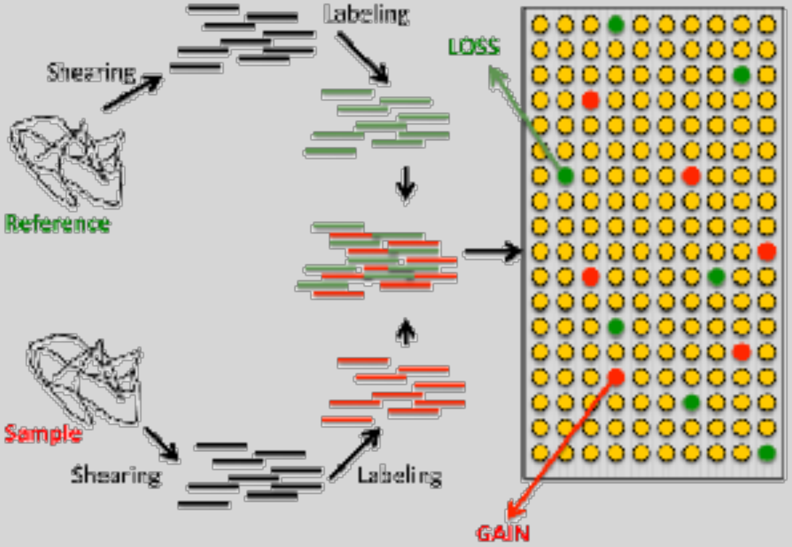
Maggie Bartlett, Courtesy: National Human Genome Research Institute.

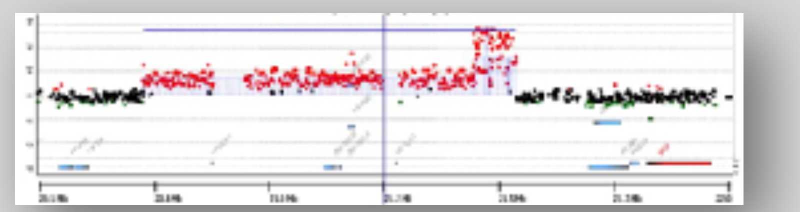
Discovering Variation: SVs

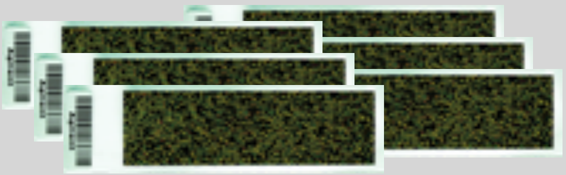
- Structural variants can be discovered by both sequence and microarray approaches
- Microarrays can only detect genomic imbalances, specifically copy number variants (CNVs)
- Sequence based approaches can, in principle, identify all types of structural rearrangements

Microarray-based CNV Discovery

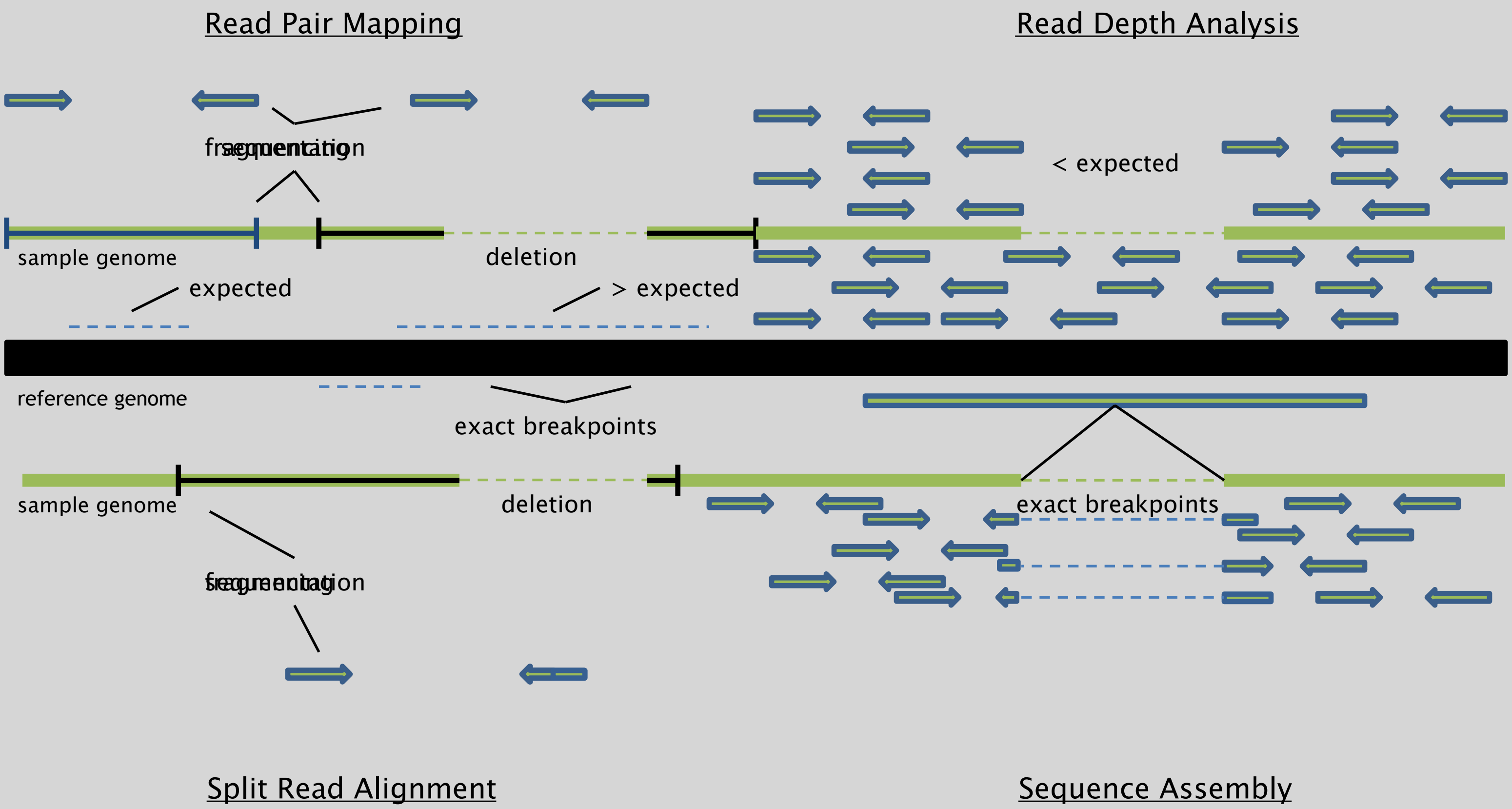
Comparative Genomic Hybridization (CGH)







Sequenced-based SV Discovery



Variant Databases and Formats

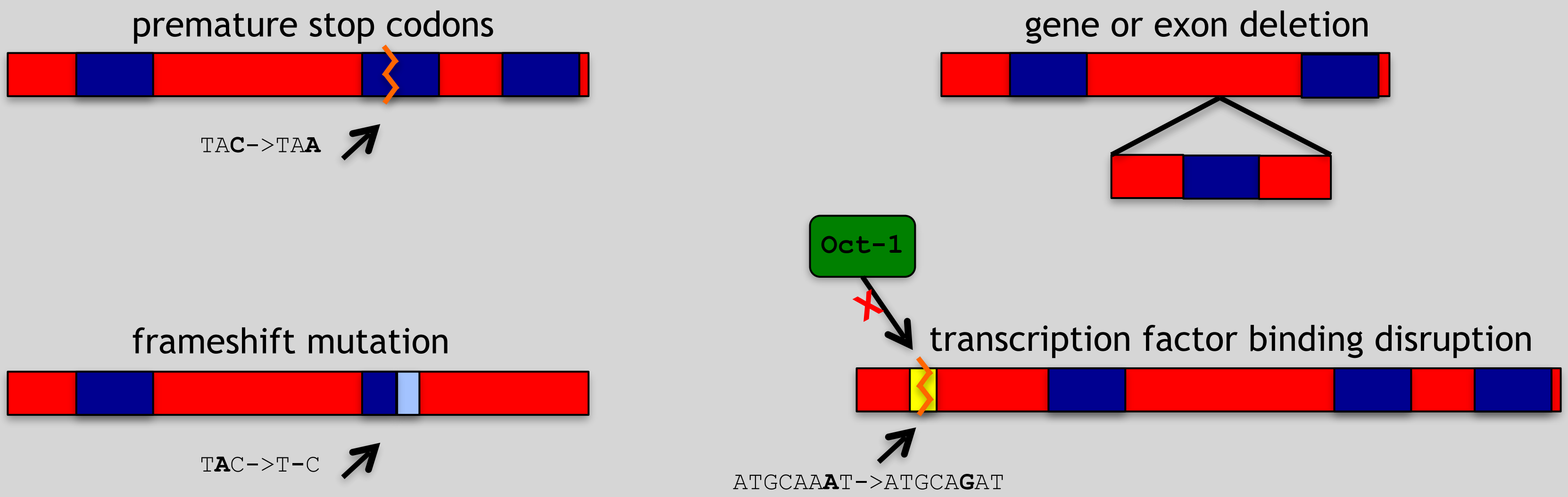
- dbSNP – repository for SNP and small INDELs
  - http://www.ncbi.nlm.nih.gov/SNP/
- VCF – variant call format for reporting variation
  - https://github.com/samtools/hts-specs

VCF Format Example

```
##fileformat=VCFv4.2
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta
##contig=<ID=20,length=62435964,assembly=B36,md5=f126cdf8a6e0c7f379d618ff66beb2da,species="Homo sapiens",taxonomy=x>
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data”>
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth”>
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency”>
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele”>
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129”>
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership”>
##FILTER=<ID=q10,Description="Quality below 10”>
##FILTER=<ID=s50,Description="Less than 50% of samples have data”>
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype”>
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality”>
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth”>
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality”>
#CHROM POS     ID         REF   ALT     QUAL   FILTER   INFO                               FORMAT       NA00001         NA00002         NA00003
20     14370   rs6054257  G     A       29     PASS     NS=3;DP=14;AF=0.5;DB;H2            GT:GQ:DP:HQ  0|0:48:1:51,51  1|0:48:8:51,51  1/1:43:5:.,.
20     17330   .          T     A       3      q10      NS=3;DP=11;AF=0.017                GT:GQ:DP:HQ  0|0:49:3:58,50  0|1:3:5:65,3    0/0:41:3
20     1110696 rs6040355  A     G,T     67     PASS     NS=2;DP=10;AF=0.333,0.667;AA=T;DB  GT:GQ:DP:HQ  1|2:21:6:23,27  2|1:2:0:18,2    2/2:35:4
20     1230237 .          T     .       47     PASS     NS=3;DP=13;AA=T                    GT:GQ:DP:HQ  0|0:54:7:56,60  0|0:48:4:51,51  0/0:61:2
21     1234567 microsat1  GTC   G,GTCT  50     PASS     NS=3;DP=9;AA=G                     GT:GQ:DP     0/1:35:4        0/2:17:2        1/1:40:3
```

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

Oct-1

transcription factor binding disruption

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

Variant Annotation

- Variants are *annotated* based on their potential functional impact
- For variants falling inside genes, there are a number of software packages that can be used to quickly determine which may have a functional role (missense/ nonsense mutations, splice site disruption, etc)
- A few examples are:
  - ANNOVAR (http://www.openbioinformatics.org/annovar/)
  - VAAST (http://www.yandell-lab.org/software/vaast.html)
  - VEP (http://http://grch37.ensembl.org/Homo_sapiens/Tools/VEP)
  - SeattleSeq (http://snp.gs.washington.edu/SeattleSeqAnnotation134/)
  - snpEff (http://snpeff.sourceforge.net/)

Variant Annotation Classes

High Impact

- exon_deleted
- frame_shift
- splice_acceptor
- splice_donor
- start_loss
- stop_gain
- stop_loss
- non_synonymous_start
- transcript_codon_change

Medium Impact

- non_syn_coding
- inframe_codon_gain
- inframe_codon_loss
- inframe_codon_change
- codon_change_del
- codon_change_ins
- UTR_5_del
- UTR_3_del
- other_splice_variant
- mature_miRNA
- regulatory_region
- TF_binding_site
- regulatory_region_ablation
- regulatory_region_amplification
- TFBS_ablation
- TFBS_amplification

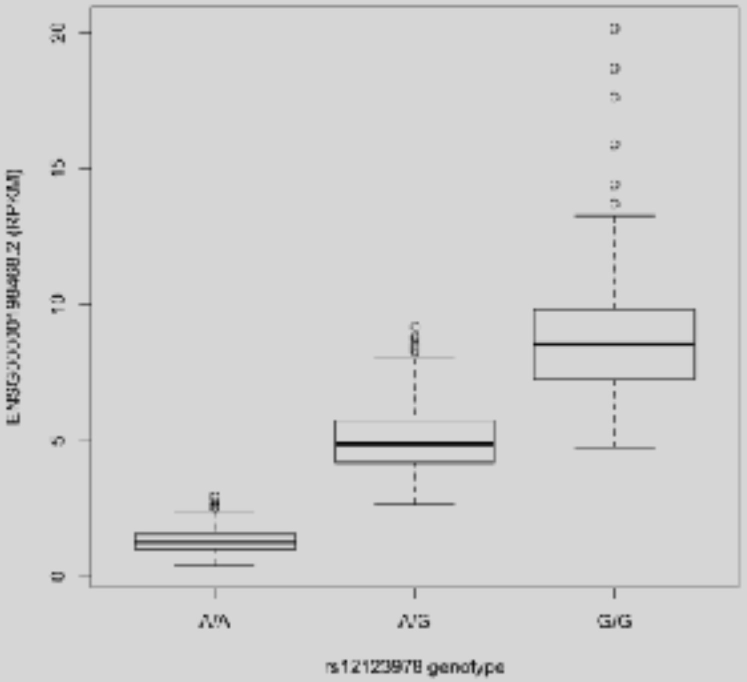
Low Impact

- synonymous_stop
- synonymous_coding
- UTR_5_prime
- UTR_3_prime
- intron
- CDS
- upstream
- downstream
- intergenic
- intragenic
- gene
- transcript
- exon
- start_gain
- synonymous_start
- intron_conserved
- nc_transcript
- NMD_transcript
- transcript_codon_change
- incomplete_terminal_codon
- nc_exon
- transcript_ablation
- transcript_amplification
- feature elongation
- feature truncation

GEMINI, http://gemini.readthedocs.org/

Variation and Gene Expression

- Expression quantitative trait loci (eQTLs) are regions of the genome that are associated with expression levels of genes
- These regions can be nearby (cis) or far away (trans) from the genes that they affect
- Genetic variants in eQTL regions are typically responsible through changes to regulatory elements



Data generated from http://www.geuvadis.org/

Geuvadis Consortium

http://www.geuvadis.org/web/geuvadis

GEUVADIS CONSORTIUM

HOME | Project | Partners | News & Events | Publications | Resources | Related Projects | PRIVATE

GEUVADIS Genetic European Variation in Health and Disease, A European Medical Sequencing Consortium

GEUVADIS RNA sequencing project for 1000 Genomes samples

Login

Email Address @org.co

Password

Remember Me

Sign In

Logout

Click here to Logout

Search

Search...

Related Events

[illegible]

Welcome!

Welcome to the GEUVADIS website

[illegible]

Funded by the European Commission (FP7, HEALTH), GEUVADIS brings together 17 partners [illegible] from 7 different countries.

Upcoming Geuvadis Events

Genomic Medicine in the Mediterranean

Inaugural conference

[illegible]

Genomic Medicine in the Mediterranean (GM²) International Conference October 2-5, 2013 [illegible] Crete, Greece

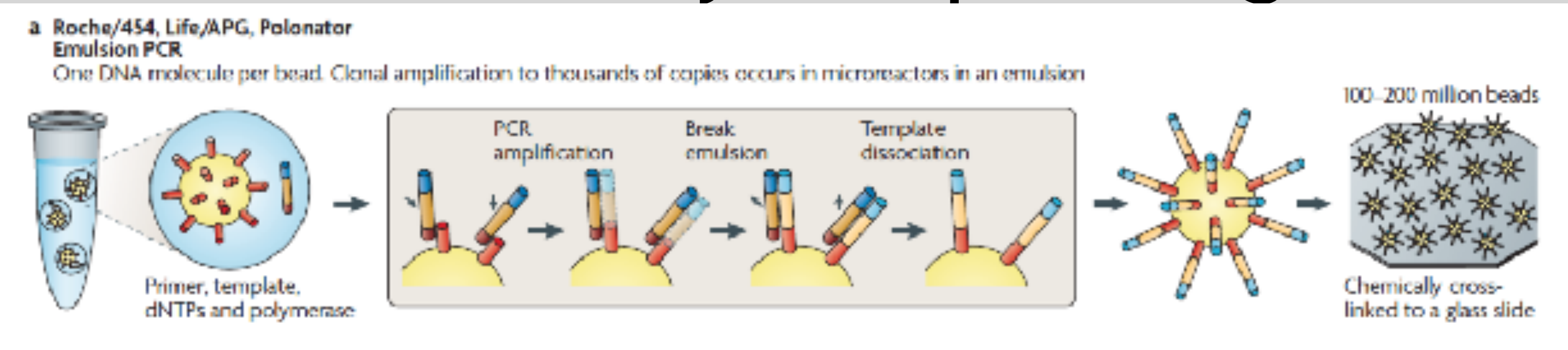
more...

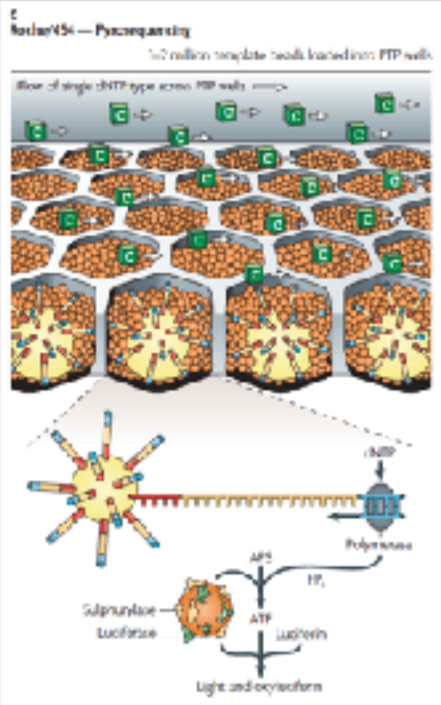
Latest News

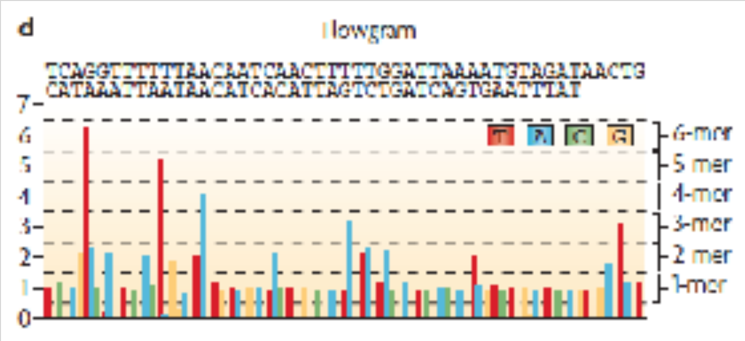
[illegible]

Additional Reference Slides on Sequencing Methods

Roche 454 - Pyrosequencing

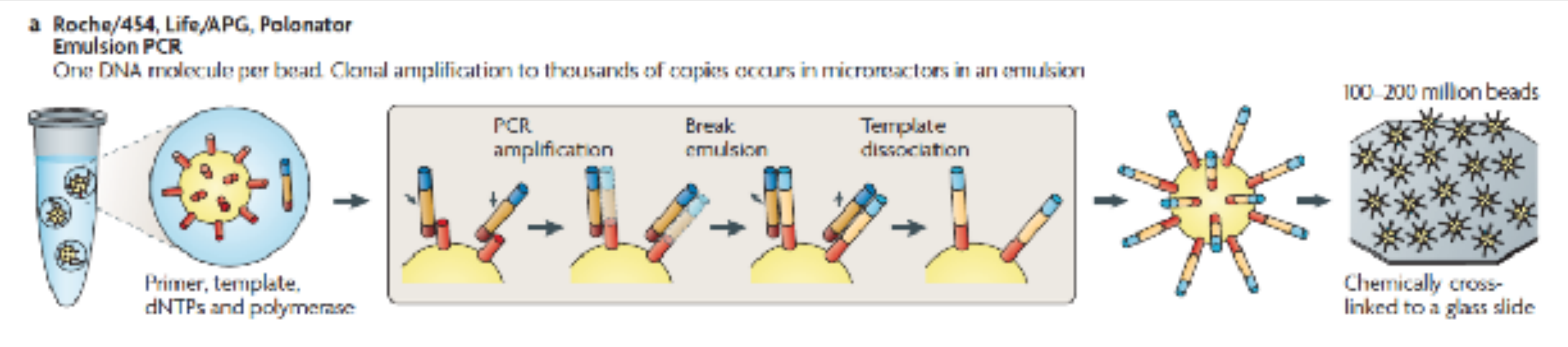

a Roche/454, Life/APG, Polonator
Emulsion PCR
One DNA molecule per bead. Clonal amplification to thousands of copies occurs in microreactors in an emulsion
Primer, template, dNTPs and polymerase
PCR amplification
Break emulsion
Template dissociation
100–200 million beads
Chemically cross-linked to a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

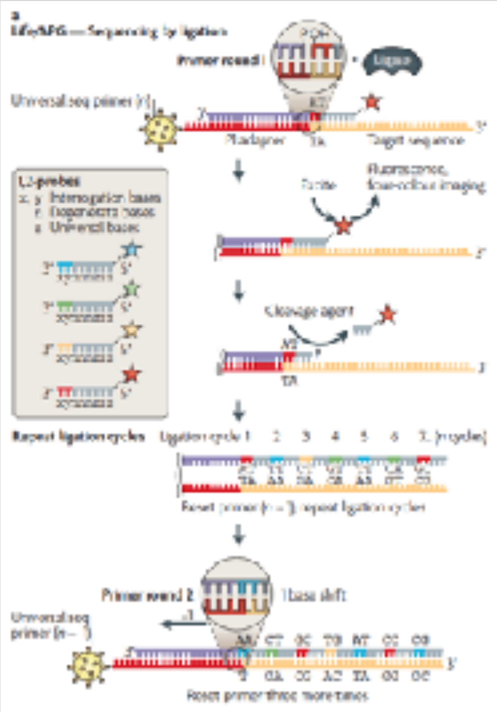

c
Roche/454 — Pyrosequencing
1–2 million template beads loaded into PTP wells
Flow of single dNTP type across PTP wells
C
dNTP
Polymerase
APS
PPi
Sulphurylase
Luciferase
ATP
Luciferin
Light and oxyluciferin

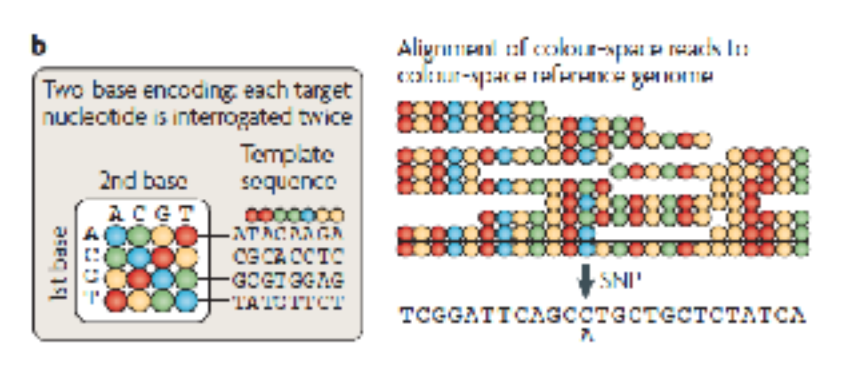


Metzker, ML (2010), *Nat. Rev. Genet*, 11, pp. 31-46

Life Technologies SOLiD – Sequence by Ligation

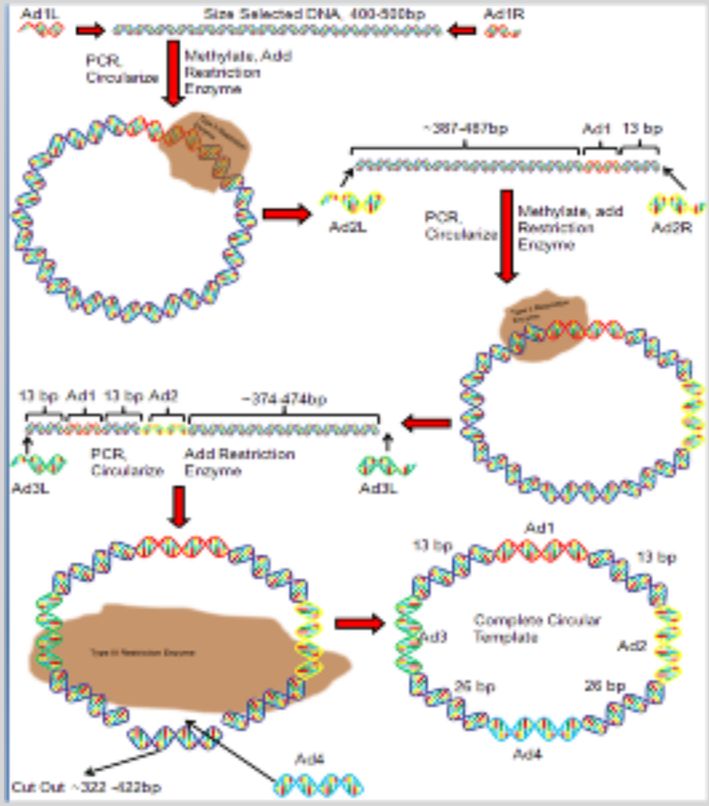




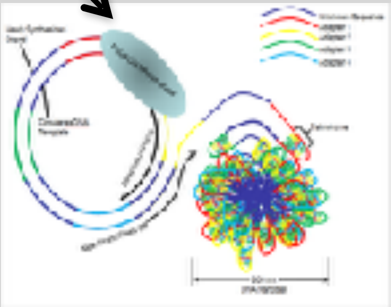

b
Two base encoding: each target
nucleotide is interrogated twice
2nd base
Template
sequence
A C G T
1st base
A
C
G
T
AT AC AA GA
CG CA CC TC
GC GT GG AG
TA TC TT CT
Alignment of colour-space reads to
colour-space reference genome
SNP
TCGGATTCAGCCTGCTGCTCTATCA
A

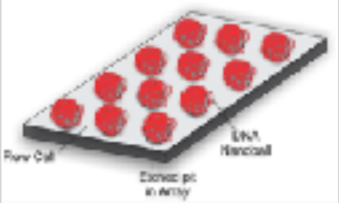
Metzker, ML (2010), *Nat. Rev. Genet*, 11, pp. 31-46

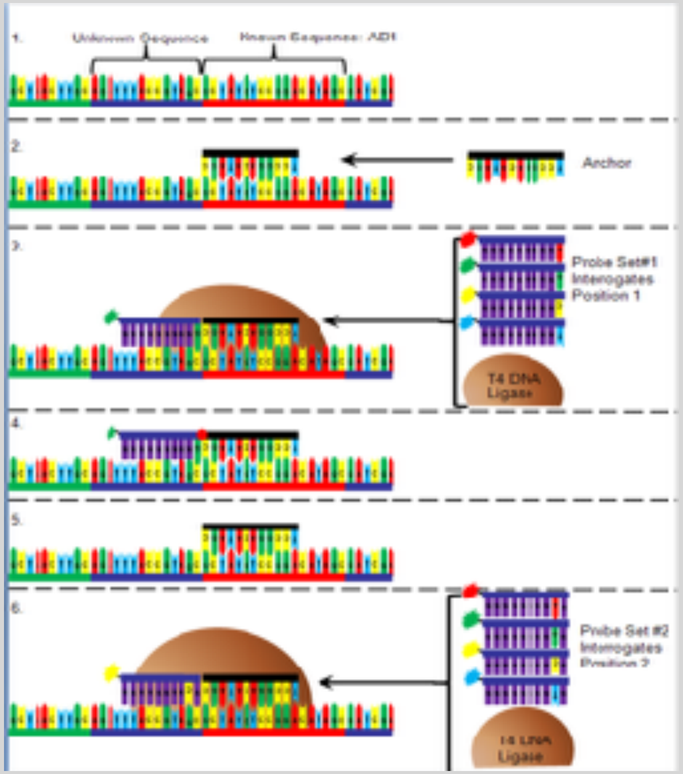
Complete Genomics – Nanoball Sequencing



Has proofreading ability!





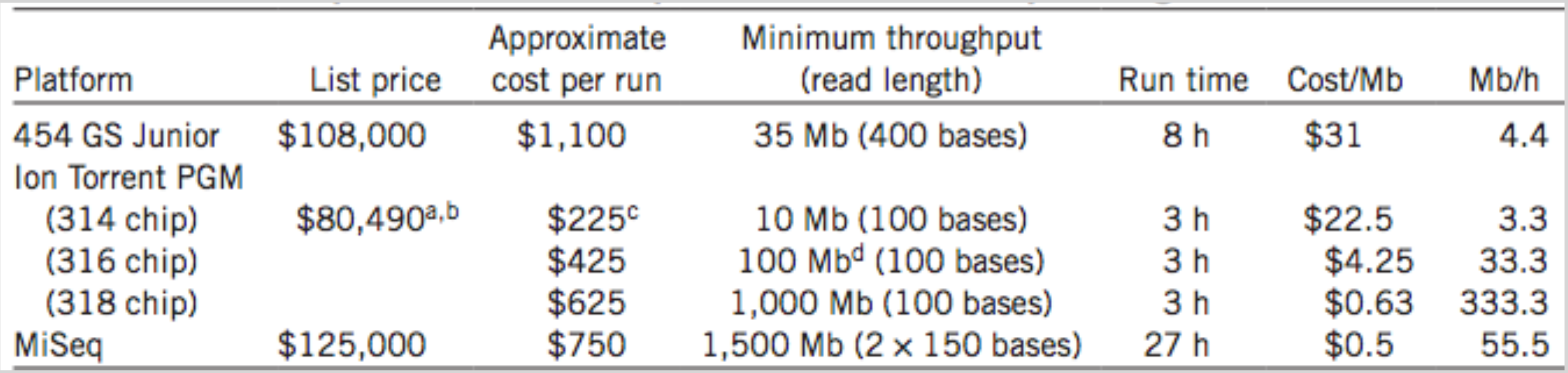


Niedringhaus, TP et al (2011), *Analytical Chem.*, 83, pp. 4327-4341

Wikipedia, “DNA Nanoball Sequencing”, September 26, 2012

“Benchtop” Sequencers

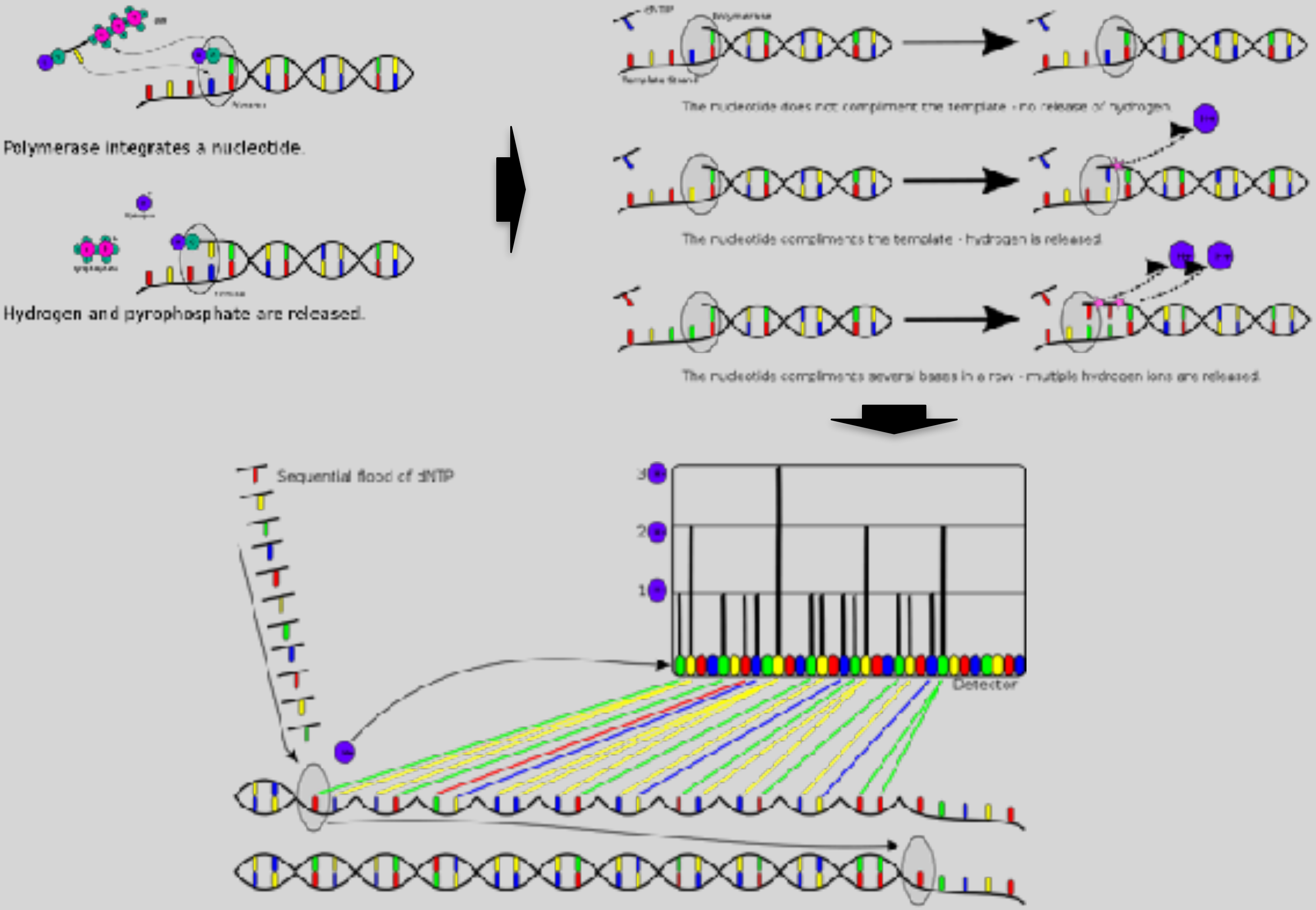
- Lower cost, lower throughput alternative for smaller scale projects
- Currently three significant platforms
  - Roche 454 GS Junior
  - Life Technology Ion Torrent
    - Personal Genome Machine (PGM)
    - Proton
  - Illumina MiSeq



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