



BIMM 194

Genomics, Big Data & Human Health

Barry Grant
UC San Diego

<http://thegrantlab.org/bimm194>

HELLO
my name is

BARRY

bjgrant@ucsd.edu

Introduce Yourself!

Your preferred name,
Place you identify with,
Major area of study/research,
Favorite joke (optional)!

Today's Menu

Course Logistics

Website, ethics, assessment and grading procedure.

Learning Objectives

What you need to learn to succeed in this course.

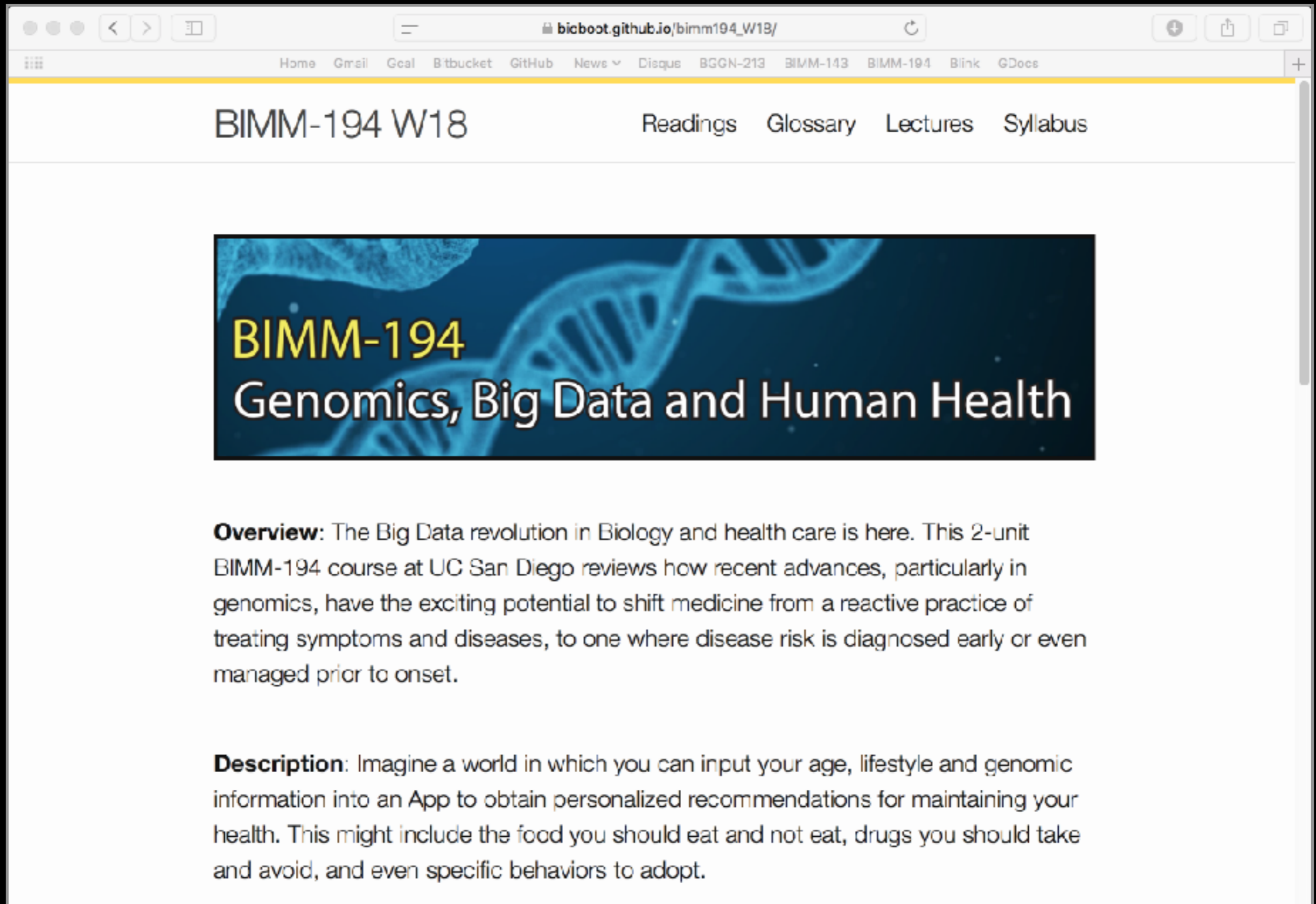
Course Structure

Major class topics and student group presentations.

Human Genome Review

What is a genome? What does the genome do? How is the genome decoded? How do we examine differences and disease mutants?


<http://thegrantlab.org/bimm194/>



Home Gmail Geal Bitbucket GitHub News Disqus BGGN-213 BIMM-143 BIMM-194 Blink GDocs

BIMM-194 W18

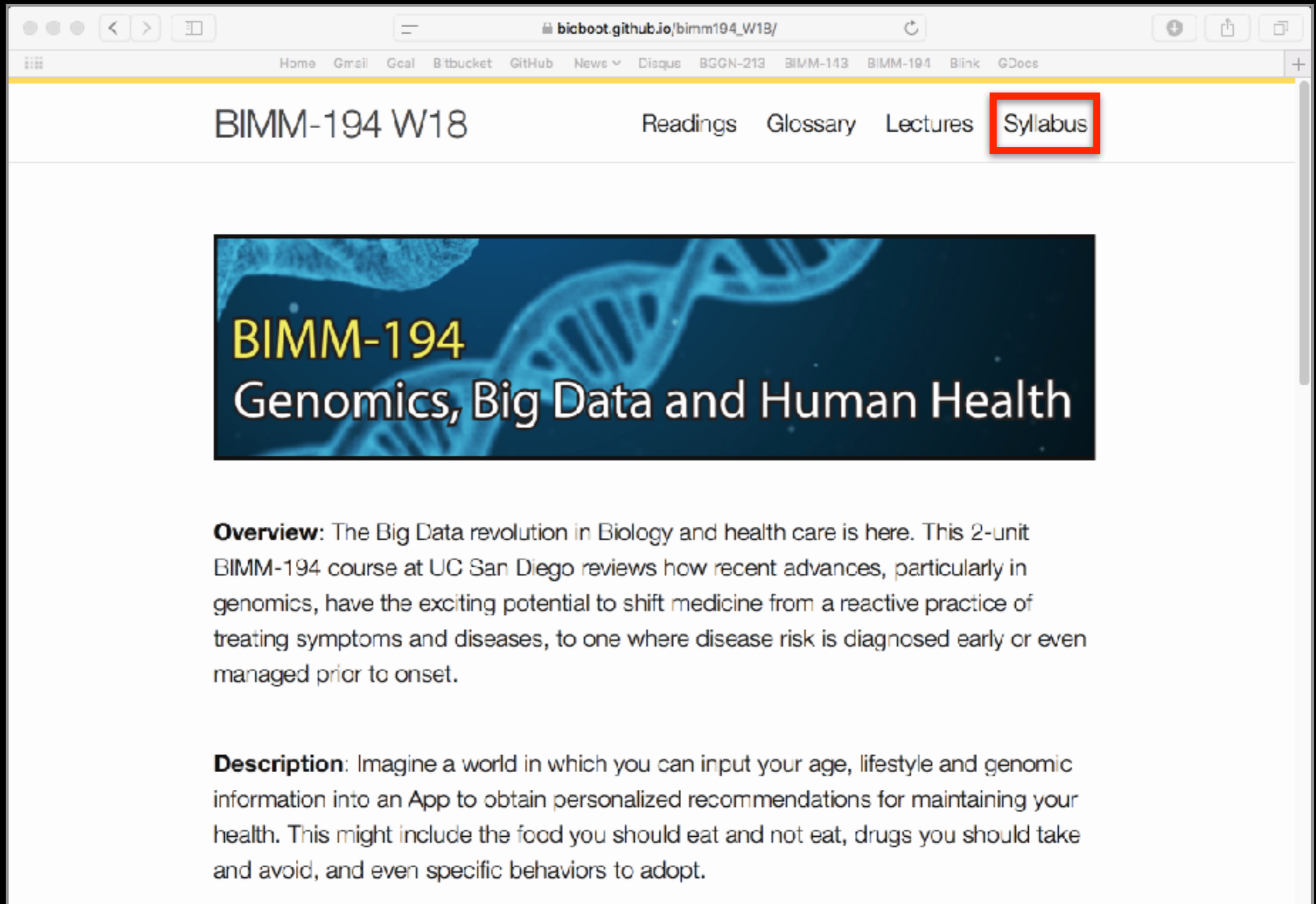
Readings Glossary Lectures Syllabus



Overview: The Big Data revolution in Biology and health care is here. This 2-unit BIMM-194 course at UC San Diego reviews how recent advances, particularly in genomics, have the exciting potential to shift medicine from a reactive practice of treating symptoms and diseases, to one where disease risk is diagnosed early or even managed prior to onset.


Description: Imagine a world in which you can input your age, lifestyle and genomic information into an App to obtain personalized recommendations for maintaining your health. This might include the food you should eat and not eat, drugs you should take and avoid, and even specific behaviors to adopt.

<http://thegrantlab.org/bimm194/>



Home Gmail Geal Bitbucket GitHub News Disqus BGGN-213 BIMM-143 BIMM-194 Blink GDocs

BIMM-194 W18 Readings Glossary Lectures **Syllabus**



BIMM-194
Genomics, Big Data and Human Health

Overview: The Big Data revolution in Biology and health care is here. This 2-unit BIMM-194 course at UC San Diego reviews how recent advances, particularly in genomics, have the exciting potential to shift medicine from a reactive practice of treating symptoms and diseases, to one where disease risk is diagnosed early or even managed prior to onset.

Description: Imagine a world in which you can input your age, lifestyle and genomic information into an App to obtain personalized recommendations for maintaining your health. This might include the food you should eat and not eat, drugs you should take and avoid, and even specific behaviors to adopt.

Grading:

Letter grades (F through A+) will be assigned on the basis of student presentations (50 points), homework and in-class quiz assignments (25 points), contributions to class discussion (15 points), and attendance (10 points). Further details will be given in class.

Note, there is no final exam or mid-term for this course.

Ethics code:

You are encouraged to collaborate with your fellow students. However, all material submitted to the instructor must be your own work.

“Academic Integrity is expected of everyone at UC San Diego. This means that you must be honest, fair, responsible, respectful, and trustworthy in all of your actions. Lying, cheating or any other forms of dishonesty will not be tolerated because they undermine learning and the University’s ability to certify students’ knowledge and abilities. Thus, any attempt to get, or help another get, a grade by cheating, lying or dishonesty will be reported to the Academic Integrity Office and will result sanctions.

Sanctions can include an F in this class and suspension or dismissal from the University. So, think carefully before you act. Before you act, ask yourself the following questions: a) is my action honest, fair, respectful, responsible & trustworthy and, b) is my action authorized by the instructor? If you are unsure, don’t ask a friend—ask your instructor, instructional assistant, or the Academic Integrity Office”.

You can learn more about academic integrity at academicintegrity.ucsd.edu

(Source: UCSD Academic Integrity Office, 2017)

Assessment & Grading:

- Letter grades (F through A+) will be assigned on the basis of:
 - ▶ Student presentations (50 points),
 - ▶ Homework and quiz assignments (25 points),
 - ▶ Contributions to class discussion (15 points),
 - ▶ Attendance (10 points).
- There will be occasional opportunities for extra credit
- There is no final exam or mid-term!

Today's Menu

Course Logistics

Website, ethics, assessment and grading procedure.

Learning Objectives

What you need to learn to succeed in this course.

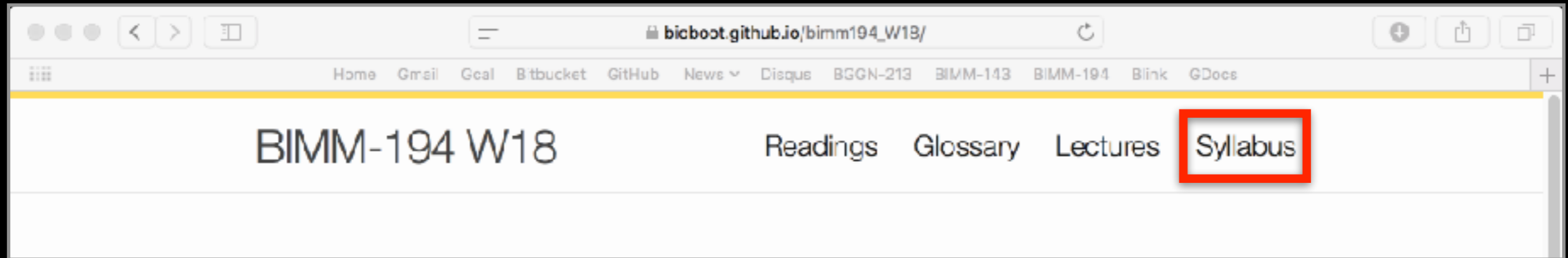
Course Structure

Major class topics and student group presentations.

Human Genome Review

What is a genome? What does the genome do? How is the genome decoded? How do we examine differences and disease mutants?

<http://thegrantlab.org/bimm194/>



Course objectives:

At the end of this course students will be able to:

- Describe human genome structure and how genomes differ between individuals.
- Appreciate and be able to describe in general terms the recent rapid advances in sequencing technologies and understand the process by which genomes are currently sequenced.
- Develop an understanding of how genomics can inform us about disease risks.
- Critically evaluate and summarize primary research literature in the genomics area.
- Discuss major ethical, legal and social implications of advances in genomic technologies.
- Utilize terminology such as gene, genotype, phenotype, variant, variants of unknown significance, traits, multifactorial disease, SNP, genetic test, pharmacogenomics, epigenetics, microbiome, whole genome sequencing and exome sequencing.

At the end of this course students will be able to:

- Describe human genome structure and how genomes differ between individuals.
- Appreciate and be able to describe in general terms the recent rapid advances in sequencing technologies and how genomics can inform us about disease risks.
- Critically evaluate and summarize primary research literature in the genomics area.
- Discuss major ethical, legal and social implications of advances in genomic technologies.
- Utilize terminology such as gene, genotype, phenotype, variant, variants of unknown significance, traits, multifactorial disease, SNP, genetic test, pharmacogenomics, epigenetics, microbiome, whole genome sequencing and exome sequencing.

Today's Menu

Course Logistics

Website, ethics, assessment and grading procedure.

Learning Objectives

What you need to learn to succeed in this course.

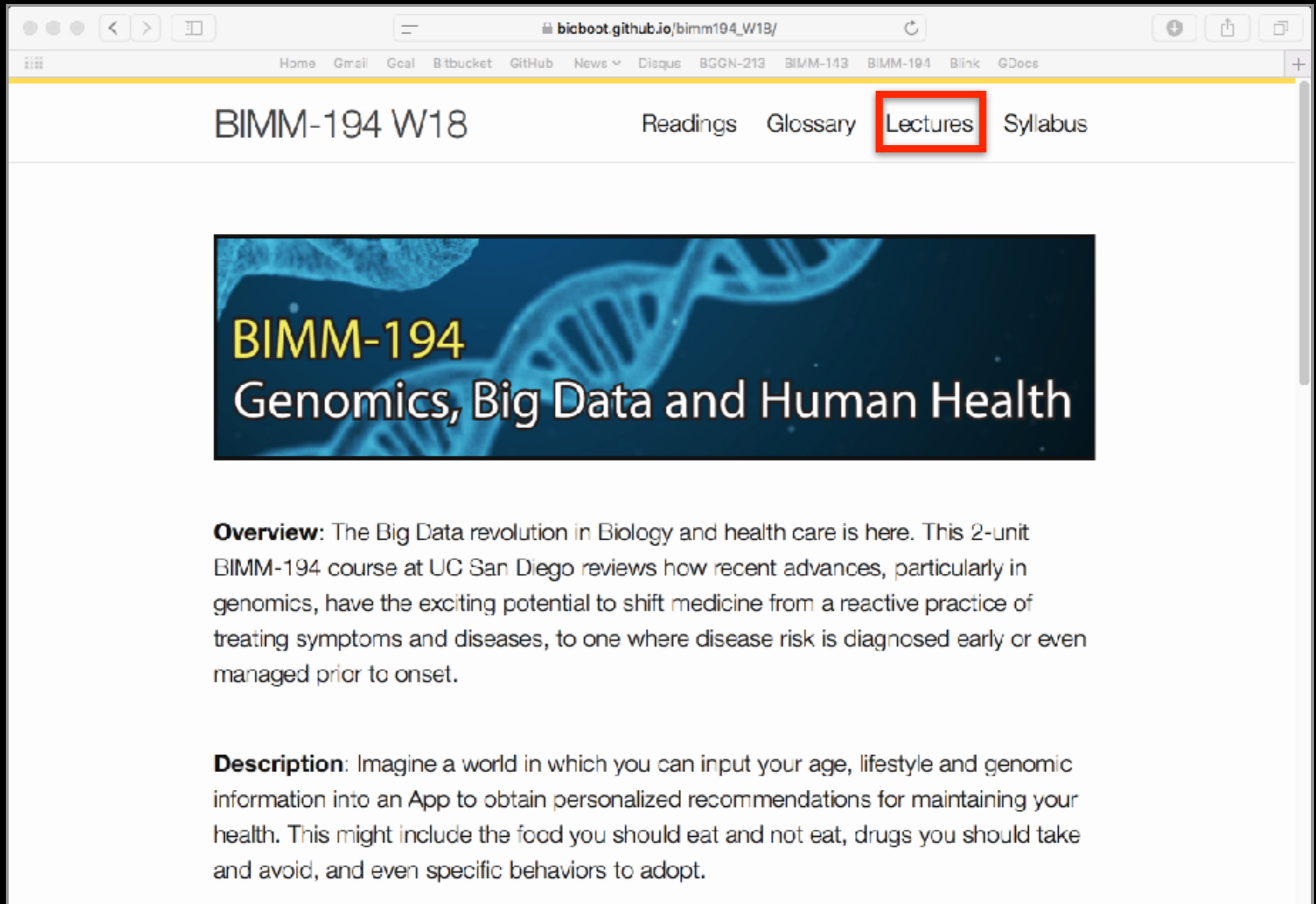
Course Structure

Major class topics and student group presentations.

Human Genome Review


What is a genome? What does the genome do? How is the genome decoded? How do we examine differences and disease mutants?

<http://thegrantlab.org/bimm194/>



Home Gmail Geal Bitbucket GitHub News Disqus BGGN-213 BIMM-143 BIMM-194 Blink GDocs

BIMM-194 W18 Readings Glossary **Lectures** Syllabus



BIMM-194
Genomics, Big Data and Human Health

Overview: The Big Data revolution in Biology and health care is here. This 2-unit BIMM-194 course at UC San Diego reviews how recent advances, particularly in genomics, have the exciting potential to shift medicine from a reactive practice of treating symptoms and diseases, to one where disease risk is diagnosed early or even managed prior to onset.

Description: Imagine a world in which you can input your age, lifestyle and genomic information into an App to obtain personalized recommendations for maintaining your health. This might include the food you should eat and not eat, drugs you should take and avoid, and even specific behaviors to adopt.

<http://thegrantlab.org/bimm194/>

bioboot.github.io/bimm194_W18/lectures/

Home Gmail Geal Bitbucket GitHub News Disqus BGGN-213 BIMM-143 BIMM-194 Blink GDocs

BIMM-194 W18 Readings Glossary **Lectures** Syllabus

Lectures

All Lectures are Friday 2:00-3:20 pm in York Hall 3010 (YH 3010) ([Map](#)). Clicking on the class topics below will take you to corresponding lecture notes, homework assignments, and required reading material.

Course introduction & review of genome fundamentals

1 Fri, 01/12/18 Introduction to the course, Overview of major learning objectives and topic areas. Human genome review: What is DNA? What is a genome? What does the genome do? How do genomes differ between individuals? How is the genome decoded? Exploring what genetic errors are and what causes them.

Genomics and cancer treatment

What is cancer and how does it arise? Example genes implicated in cancer. What has been learned from genome

<http://thegrantlab.org/bimm194/>

bioboot.github.io/bimm194_W18/lectures/

Home Gmail Geal Bitbucket GitHub News Disqus BGGN-213 BIMM-143 BIMM-194 Blink GDoes

BIMM-194 W18 Readings Glossary Lectures Syllabus

Lectures

All Lectures are Friday 2:00-3:20 pm in York Hall 3010 (YH 3010) ([Map](#)). Clicking on the class topics below will take you to corresponding lecture notes, homework assignments, and required reading material.

Course introduction & review of genome fundamentals

1 Fri, 01/12/18 Introduction to the course, Overview of major learning objectives and topic areas. Human genome review: What is DNA? What is a genome? What does the genome do? How do genomes differ between individuals? How is the genome decoded? Exploring what genetic errors are and what causes them.

Genomics and cancer treatment

What is cancer and how does it arise? Example genes implicated in cancer. What has been learned from genome

<http://thegrantlab.org/bimm194/>

bioboot.github.io/bimm194_W18/lectures/

Home Gmail Geal Bitbucket GitHub News Disqus BGGN-213 BIMM-143 BIMM-194 Blink GDocs

BIMM-194 W18

Readings Glossary Lectures Syllabus

Lectures

All Lectures are Friday 2:00-3:20 pm in York Hall 3010 (YH 3010) ([Map](#)). Clicking on the class topics below will take you to the lecture page, which includes assignments, and required reading materials.

1	Fri, 01/12/18	Course introduction and fundamental objectives and DNA? What do genomes decoded? Explain causes them
		Genomics and What is cancer implicated in

BIMM 194
Genomics, Big Data & Human Health
Barry Grant
UC San Diego
<http://thegrantlab.org/bimm194>

<http://thegrantlab.org/bimm194/>

bioooc.github.io/bimm194_W18/readings/

Home Gmail Geal Bitbucket GitHub News Disqus BGGN-213 BIMM-143 BIMM-194 Blink GDocs

BIMM-194 W18 **Readings** Glossary Lectures Syllabus

Readings

Almost daily we hear about the impact of genomics on healthcare and how gene-directed diagnosis and therapies are transforming our understanding of widely divergent fields of biology and medicine.

Here I share with you some of the stories from the last year that I found particularly interesting. These stories exemplify the extent to which genomics is going to change the lives of patients and healthcare professionals.

[Lecture 1 Reading homework assignment](#)

[DNA Snakes and Ladders](#)

[Editing the Embryo](#)

[DIY Crispr: biohacking your own genome](#)

<http://thegrantlab.org/bimm194/>

bioboot.github.io/bimm194_W18/readings/

Home Gmail Geal Bitbucket GitHub News Disqus BGGN-213 BIMM-143 BIMM-194 Blink GDocs

BIMM-194 W18 Readings Glossary Lectures Syllabus

Readings

Almost daily we hear about the impact of genomics on healthcare and how gene-directed diagnosis and therapies are transforming our understanding of widely divergent fields of biology and medicine.

Here I share with you some of the stories from the last year that I found particularly interesting. These stories exemplify the extent to which genomics is going to change the lives of patients and healthcare professionals.

Lecture 1 Reading homework assignment

[DNA Snakes and Ladders](#)

[Editing the Embryo](#)

[DIY Crispr: biohacking your own genome](#)

<http://thegrantlab.org/bimm194/>

Lecture 1 Homework

https://bioboot.github.io/bimm194_W18/

Dr. Barry Grant (bjgrant@ucsd.edu)

Overview:

Almost daily we hear about the impact of genomics on healthcare and how gene-directed diagnosis and therapies are transforming our understanding of widely divergent fields of medicine.

In this document I share with you some of the stories from the last year that I found particularly interesting. These stories exemplify the extent to which genomics is going to change the lives of patients and healthcare professionals.

Homework:

Before next week's class write and email me (bjgrant@ucsd.edu) a paragraph of 250 words or less detailing which of these stories interests you the most and why? Have any other stories about genomics in the press caught your eye recently? Feel free to write about these for bonus points.

1. Editing the Embryo:

Just imagine if you could correct a genetic disease right there in the embryo, before the condition even developed. It may sound like science fiction, but this tantalizing idea edged closer to potential reality over the past few months following ground-breaking work on human embryo genome editing.



In August, a collaboration from the USA and Korea

reported the successful modification of human

<http://thegrantlab.org/bimm194/>

bioboot.github.io/bimm194_W18/lectures/

Home Gmail Geal Bitbucket GitHub News Disque BGGN-213 BMM-143 BMM-194 Blink GDoes

How to read a scientific paper & Introduction to student presentation assignments

4 Fri, 02/02/18 A guide for selecting, reading and understanding peer-reviewed primary research articles, How to obtain a basic understanding of a published science paper and decide whether or not it is a reputable study? How does the described work contribute to advancing the scientific knowledge base or our technical capabilities? Introduction to student presentation assignments.

Student group literature presentations

5 Fri, 02/09/18 Each week 2 student groups of 4 students each will present selected primary literature on recent genomic advances of relevance to biomedical science and health care. Topics may be selected from the following [list](#).

6 Fri, 02/16/18 **Student group literature presentations**

7 Fri, 02/23/18 **Student group literature presentations**

Presentations (25 min):

Based on **YOUR** review of primary literature on recent genomic advances of relevance to biomedical science and health care. Topics can be selected from the provided “Readings” online or address any of the following:

- How useful are genomic approaches to solving mystery genetic diseases?
- How can your genome directly help guide drug treatments for treating disease?
- Can genetic testing be used to predict intelligence or sports performance?
- Can genetic testing and genome editing be useful for choosing healthier embryos and producing designer babies?
- How will increased understating of epigenetics impact health care?
- How does the microbiome affect health and can it be rationally altered to improve health?
- Will having my genome sequenced affect my family members?
- Who has the right to know your genetic test results?

Today's Menu

Course Logistics

Website, ethics, assessment and grading procedure.

Learning Objectives

What you need to learn to succeed in this course.

Course Structure

Major class topics and student group presentations.

Human Genome Review

What is a genome? What does the genome do? How is the genome decoded? How do we examine differences and disease mutants?

GENOME REVIEW

▶ **What is a Genome?**

- Genome sequencing and the Human genome project

▶ **What can we do with a Genome?**

- Comparative genomics

▶ **Modern Genome Sequencing**

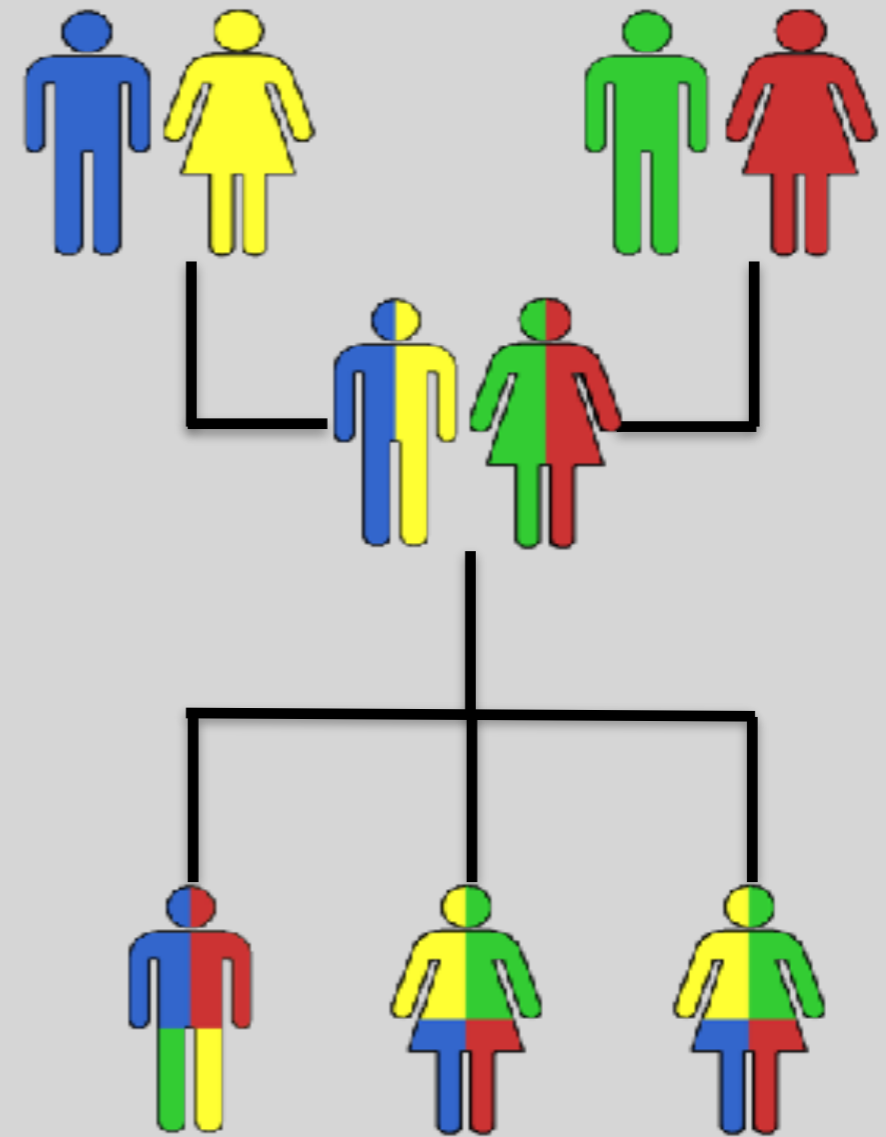
- 1st, 2nd and 3rd generation sequencing
- RNA-Sequencing and discovering variation

Genetics and Genomics

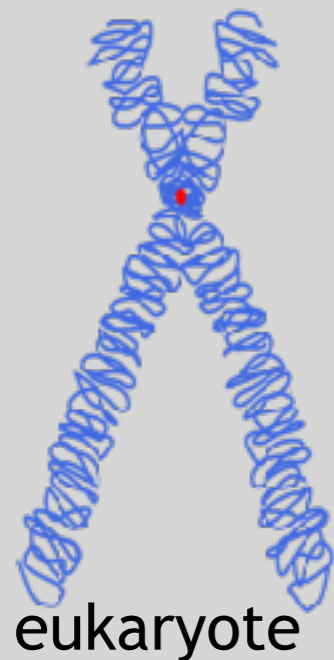
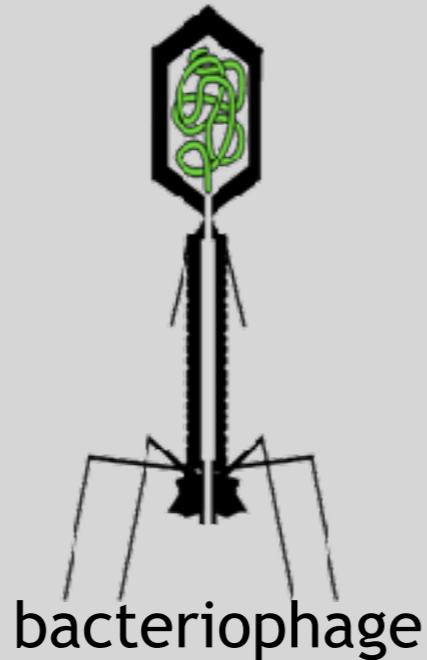
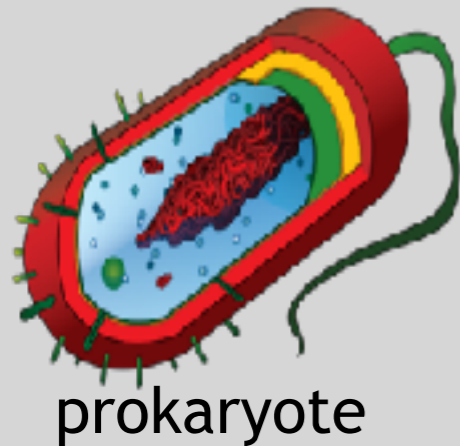
- **Genetics** is primarily the study of individual genes, mutations within those genes, and their inheritance patterns in order to understand specific traits.
- **Genomics** expands upon classical genetics and considers aspects of the entire genome, typically using computer aided approaches.

What is a Genome?

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

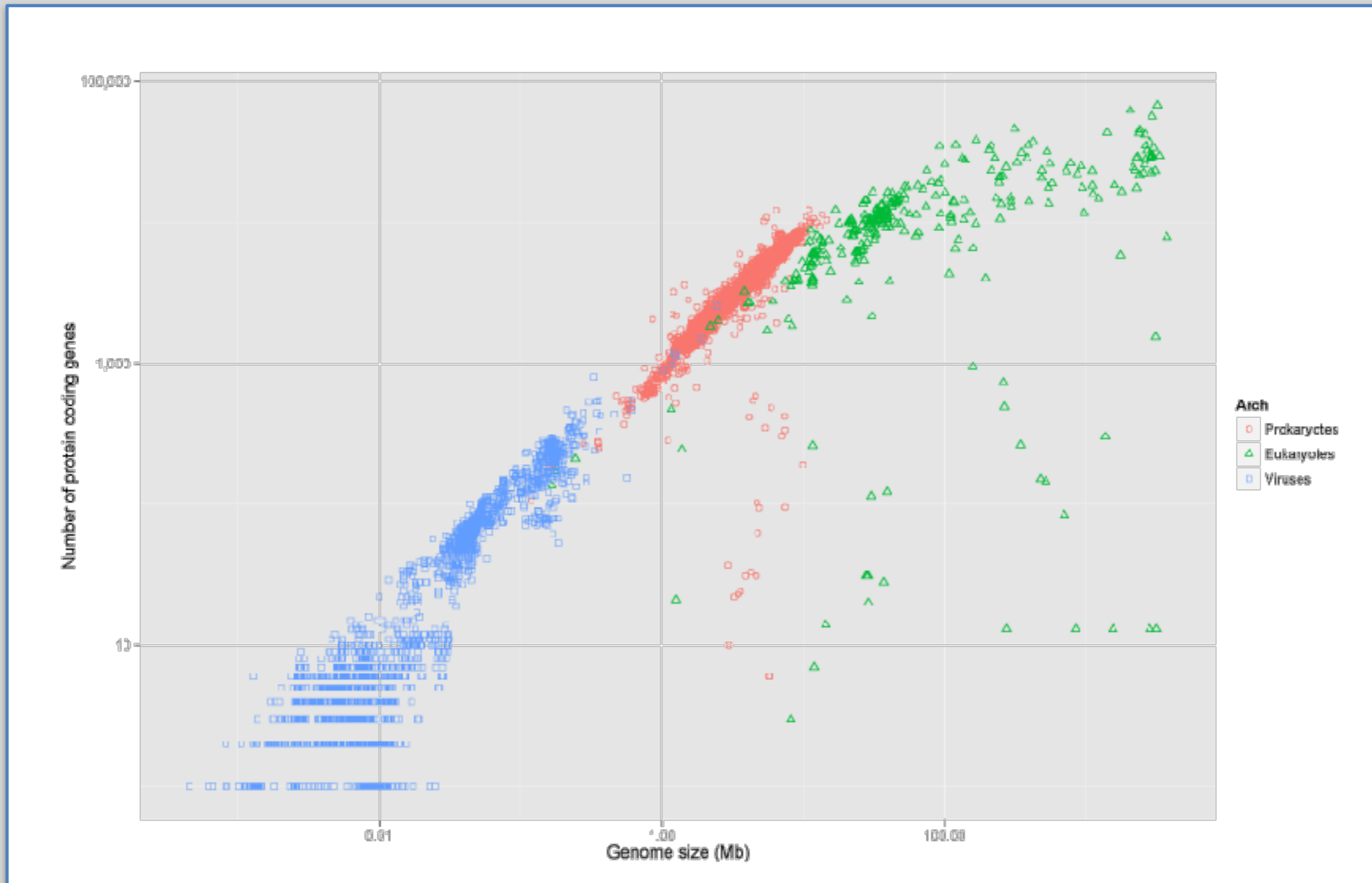


Genomes come in many shapes



- 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)

Genomes come in many sizes



Genome Databases

NCBI Genome:

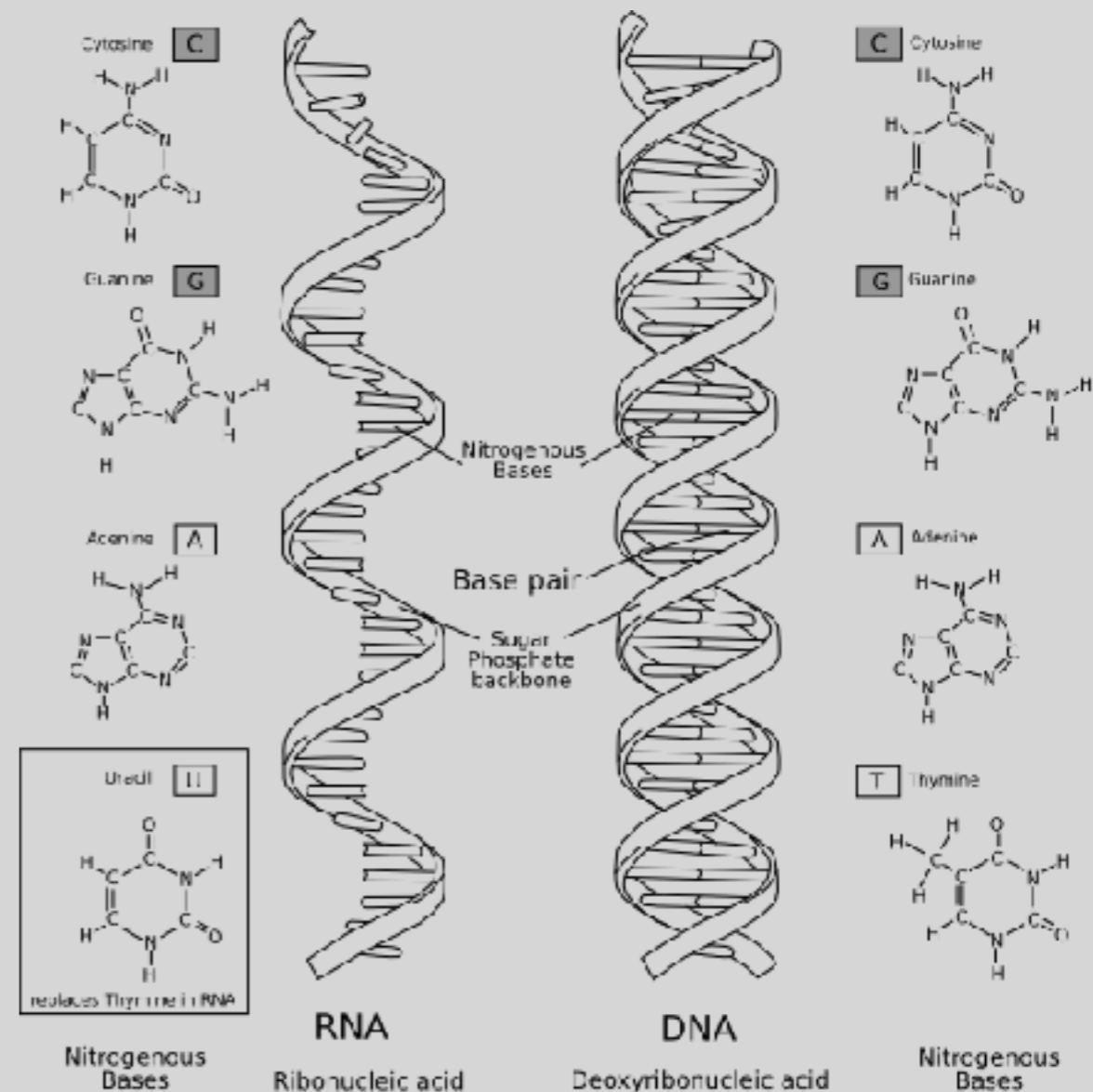
<http://www.ncbi.nlm.nih.gov/genome>

The screenshot shows the NCBI Genome website. At the top, there is a navigation bar with 'NCBI Resources How To' and a 'Sign in to NCBI' link. Below this is a search bar with 'Genome' selected in a dropdown menu, a search button, and links for 'Limits' and 'Advanced'. A banner image shows chromosomes with the title 'Genome' and a description: 'This resource organizes information on genomes including sequences, maps, chromosomes, assemblies and annotations'. The main content area is divided into several sections: 'Using Genome' (with links for Help, Browse by Organism, Download/FTP, Download FAQ, and Submit a genome), 'Genome Tools' (with links for BLAST the Human Genome, Microbial Nucleotide BLAST, and TaxPlot), 'Custom resources' (with links for Human Genome, Microbes, Organelles, Viruses, and Prokaryotic reference genomes), 'Genome Annotation and Analysis' (with links for Eukaryotic Genome Annotation, Prokaryotic Genome Annotation, and PASC), 'Other Resources' (with links for Assembly, BioProject, BioSample, Map Viewer, and Protein Clusters), and 'External Resources' (with links for GOLD, Ensembl Genome Browser, Bacteria Genomes at Sanger, and Large-Scale Genome Sequencing). At the bottom, there is a footer with a navigation menu: '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, Uniprot, SNP, Gene, Protein, PubGen), 'FEATURED' (Genetic Testing Registry, PubMed Health, GenBank, Reference Sequences, Gene Expression Omnibus, Map Viewer, Human Genome, Mouse Genome, Influenza Virus, PrimerBLAST, Sequence Read Archive), and 'NCBI INFORMATION' (About NCBI, Research at NCBI, NCBI News, NCBI FTP Site, NCBI on Facebook, NCBI on Twitter, NCBI on YouTube). The footer also includes copyright information, a disclaimer, and logos for the National Center for Biotechnology Information, the U.S. National Library of Medicine, and the USA.gov logo.

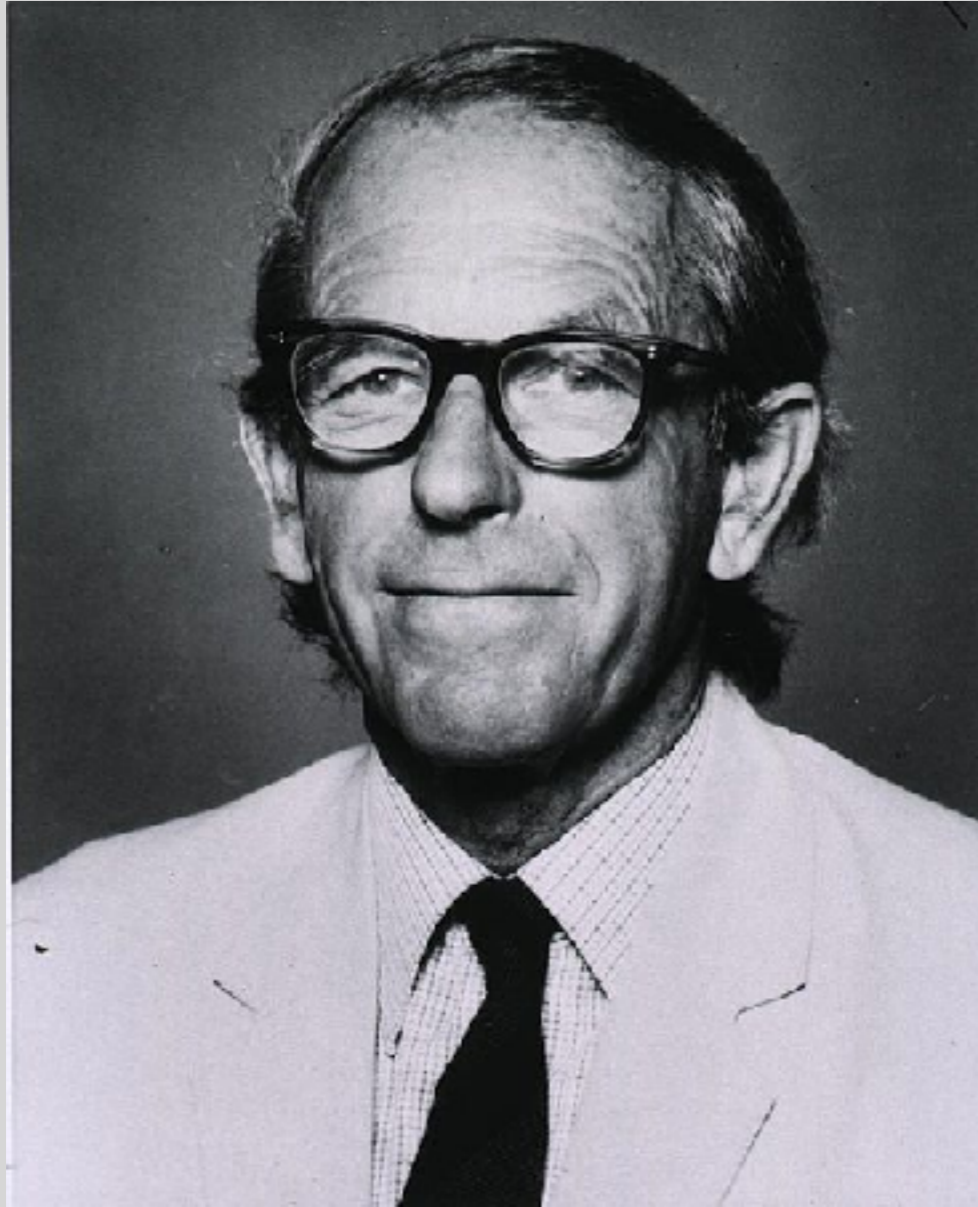
Characteristics of Genomes

- 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

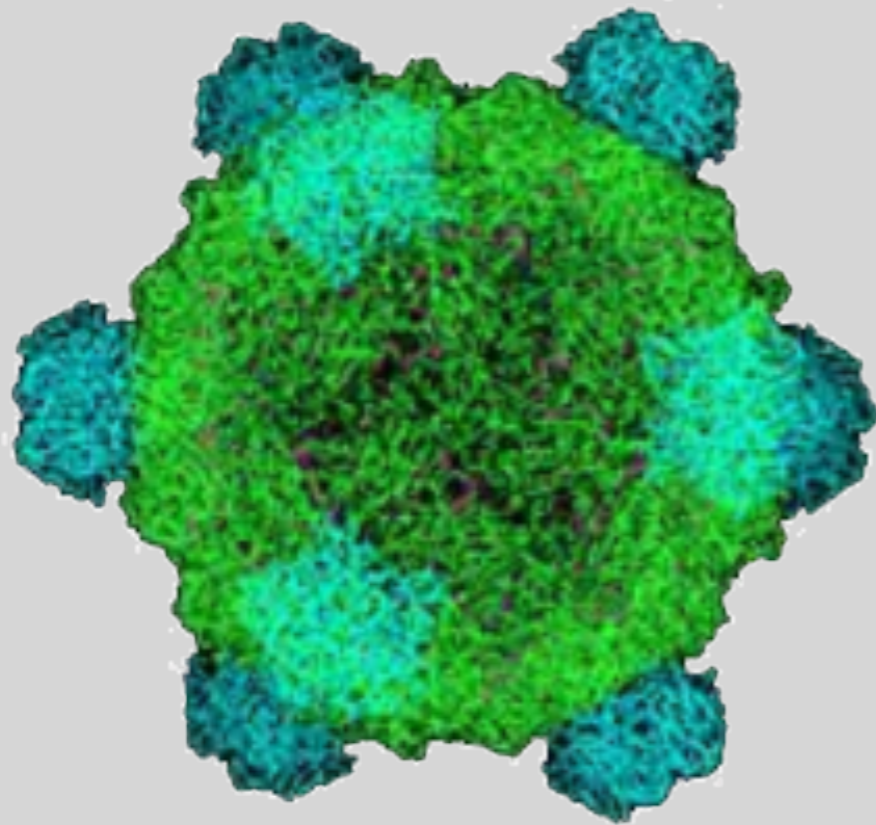


Early Genome Sequencing



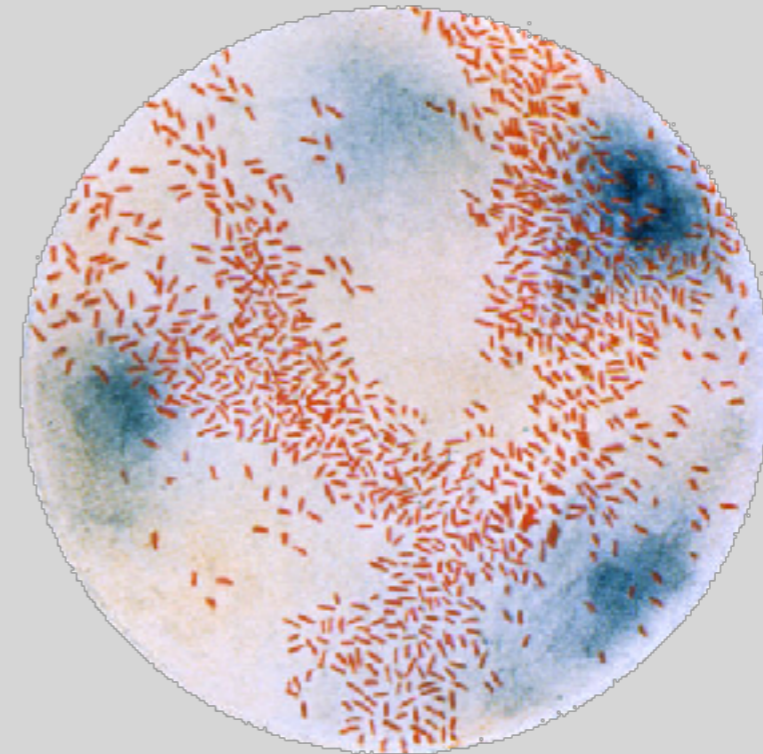
- Chain-termination “Sanger” sequencing was developed in 1977 by Frederick Sanger, colloquially referred to as the “Father of Genomics”
- Sequence reads were typically 750-1000 base pairs in length with an error rate of $\sim 1 / 10000$ bases

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

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

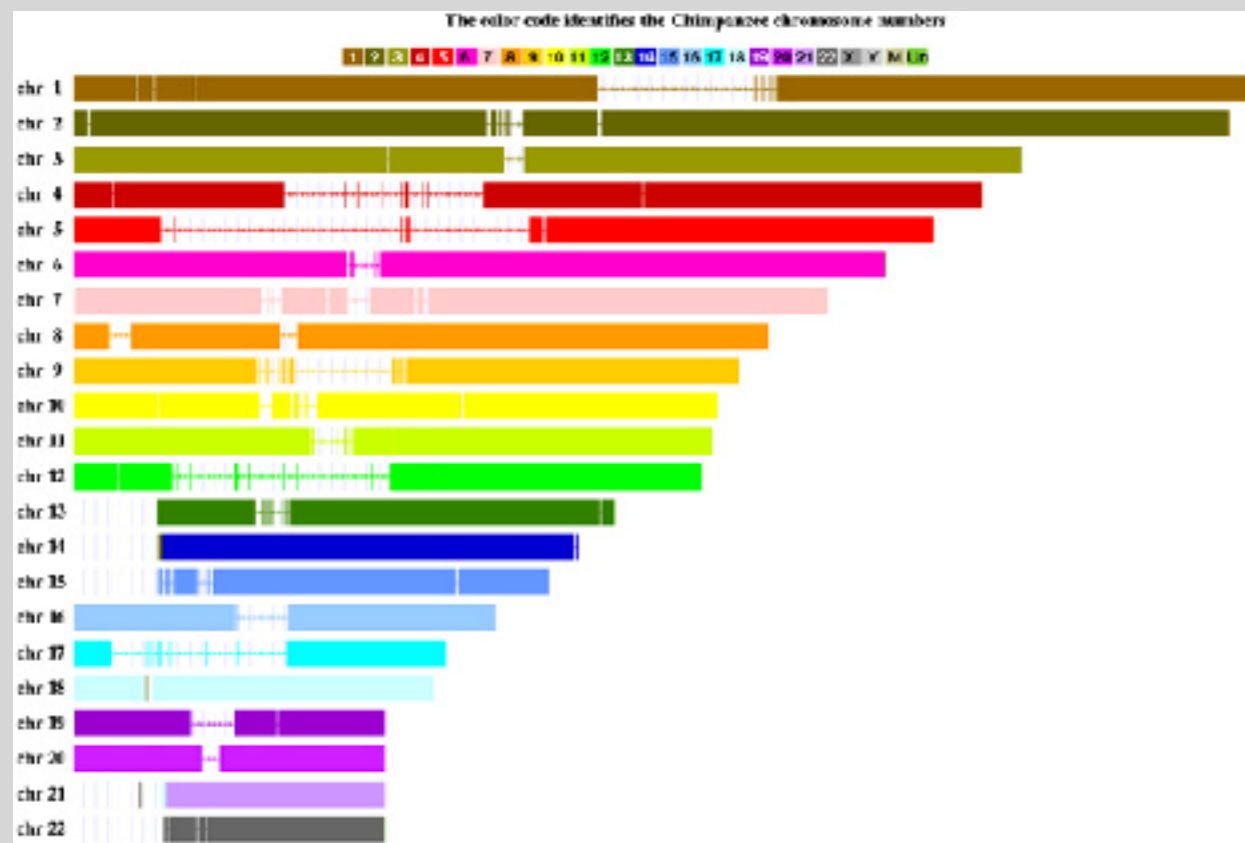


What can we do with a Genome?

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

Comparative Genomics

~6-7 million years

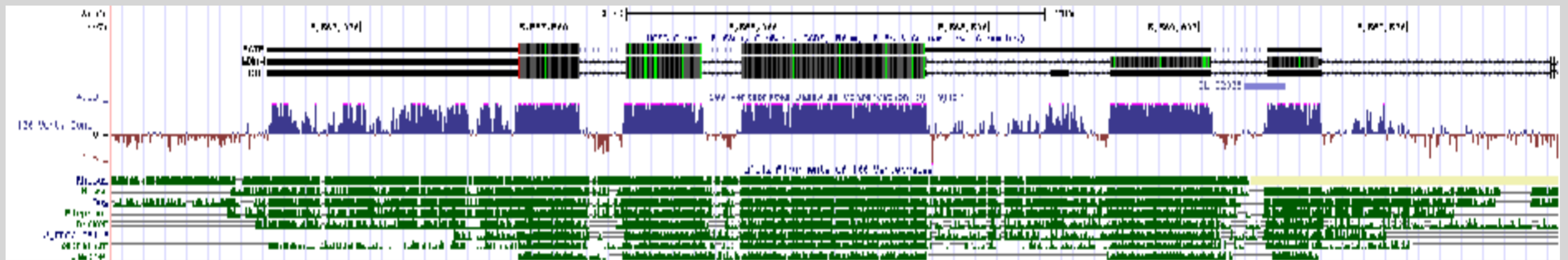


~60-70 million years



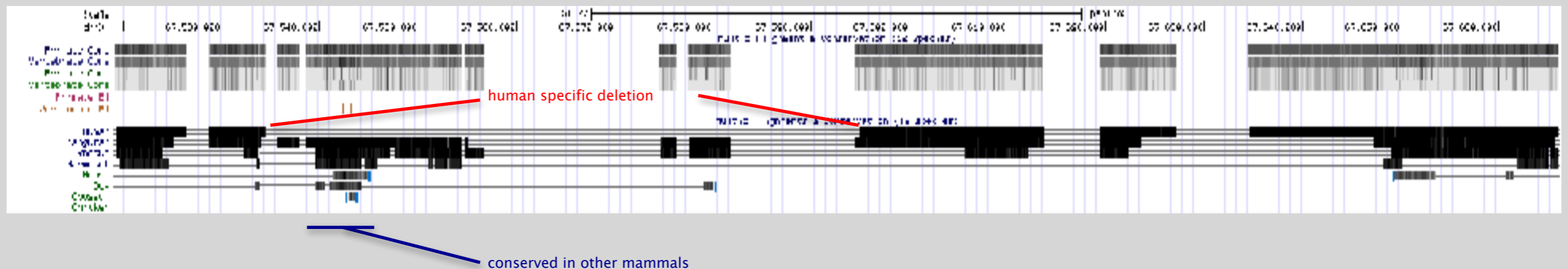
Conservation Suggests Function

- 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



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)



When we look at a persons genome we often look for specific changes (mutations, insertions and deletions) to known genes.

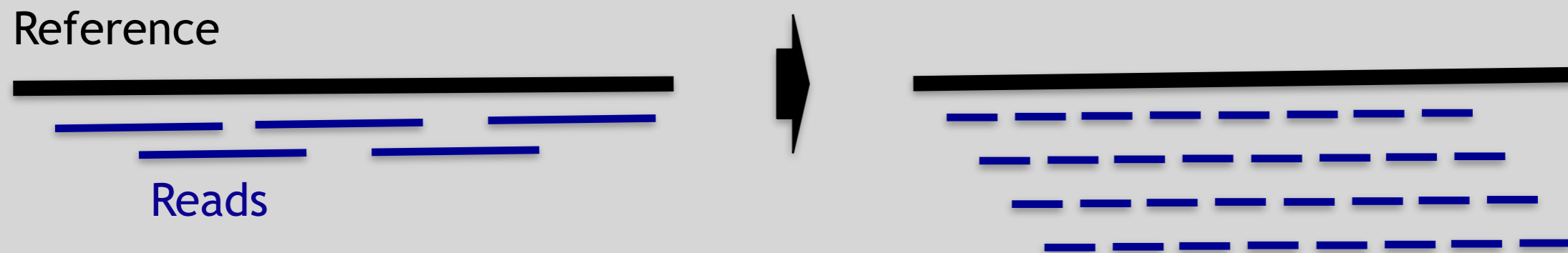
We then sort these genes into distinct piles much like sorting dirty laundry.



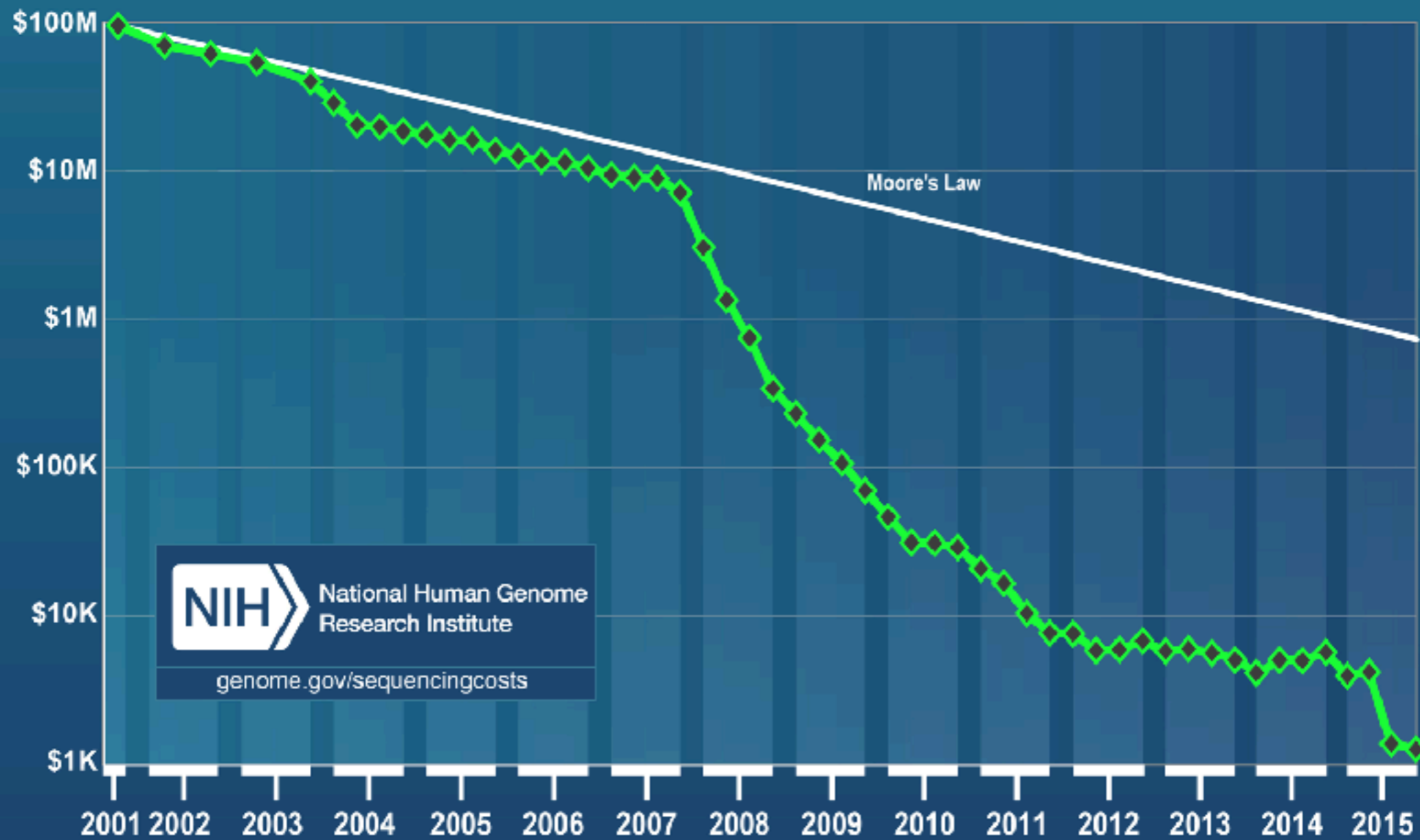
- Into the **1st basket** go the relatively few “**actionable**” disease causing mutant genes. These are the precious few that existing treatments can potentially address (i.e. there are known therapies for treating their associated illness). E.g. BRACA1 and breast cancer. **These are like precious laundry items for the dry-cleaners.**
- The **2nd basket** contains the “**unactionable**” gene mutations, these are the *wish they were actionable* set where medicine has little to offer. E.g. mutations linked to familial Alzheimer disease for which there is no treatment yet. **Many people don't want to know about these.**
- The **3rd basket** contains the “**potentially actionable**” genes. E.g. gene variants yielding adverse reactions to drugs like abacavir or clopidogrel that you have never heard of but may be prescribed in 10 years time. If you do take these drugs you will have serious immune reactions or die in agony from bleeding to death respectively. **These are like the items you don't use but keep in the back of the closet just in case you need them some day.**
- The **4th basket** contain “**other**” mutant genes that don't directly affect your health but may be significant for your siblings and kids. E.g. the cystic fibrosis gene where you have one good copy and one bad copy (1/23). **This is like laundry that is someone else's and not yours but you want to do it anyway.**

Modern Genome Sequencing

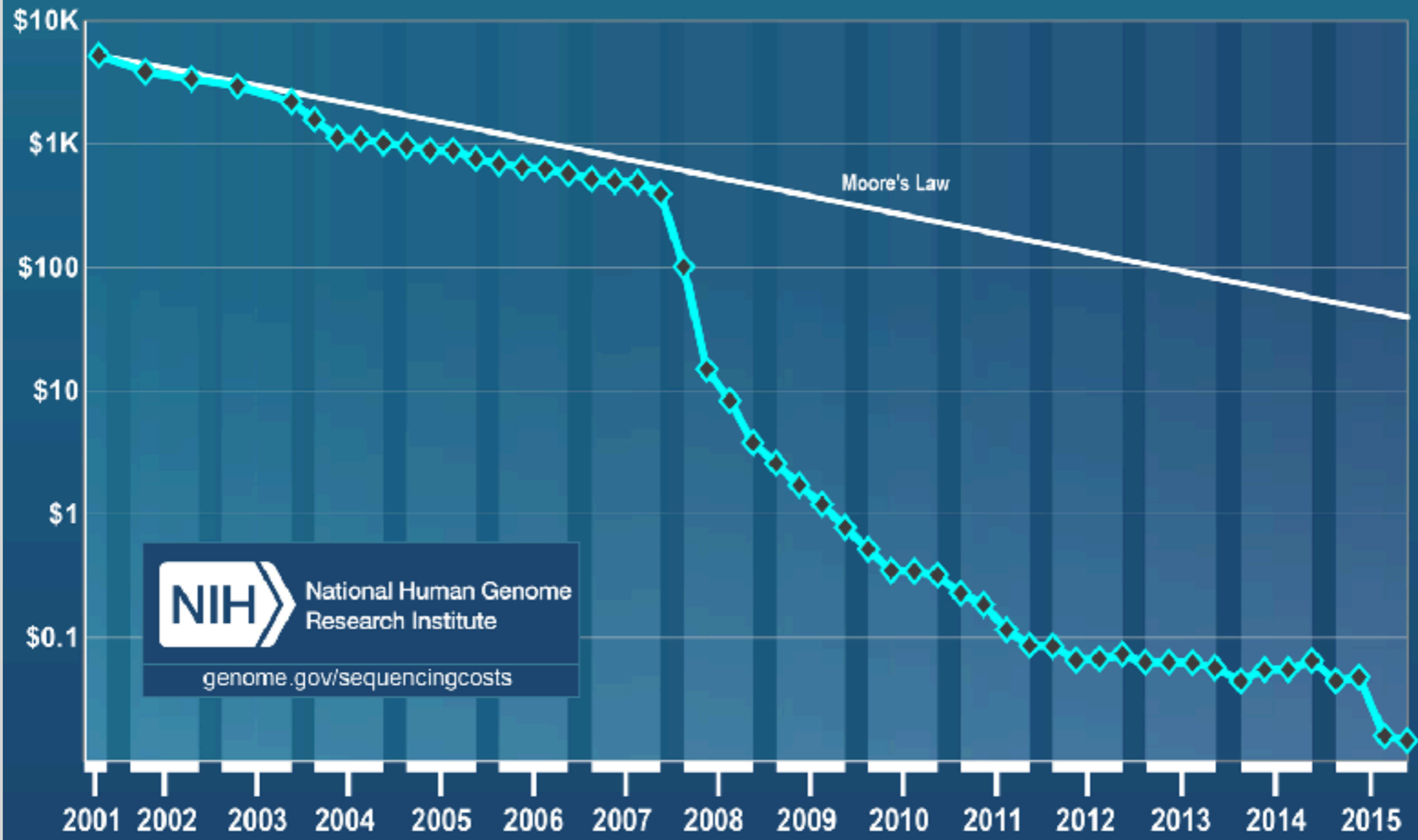
- Next Generation Sequencing (NGS) technologies have resulted in a paradigm shift from long reads at low coverage to short reads at high coverage
- This provides numerous opportunities for new and expanded genomic applications



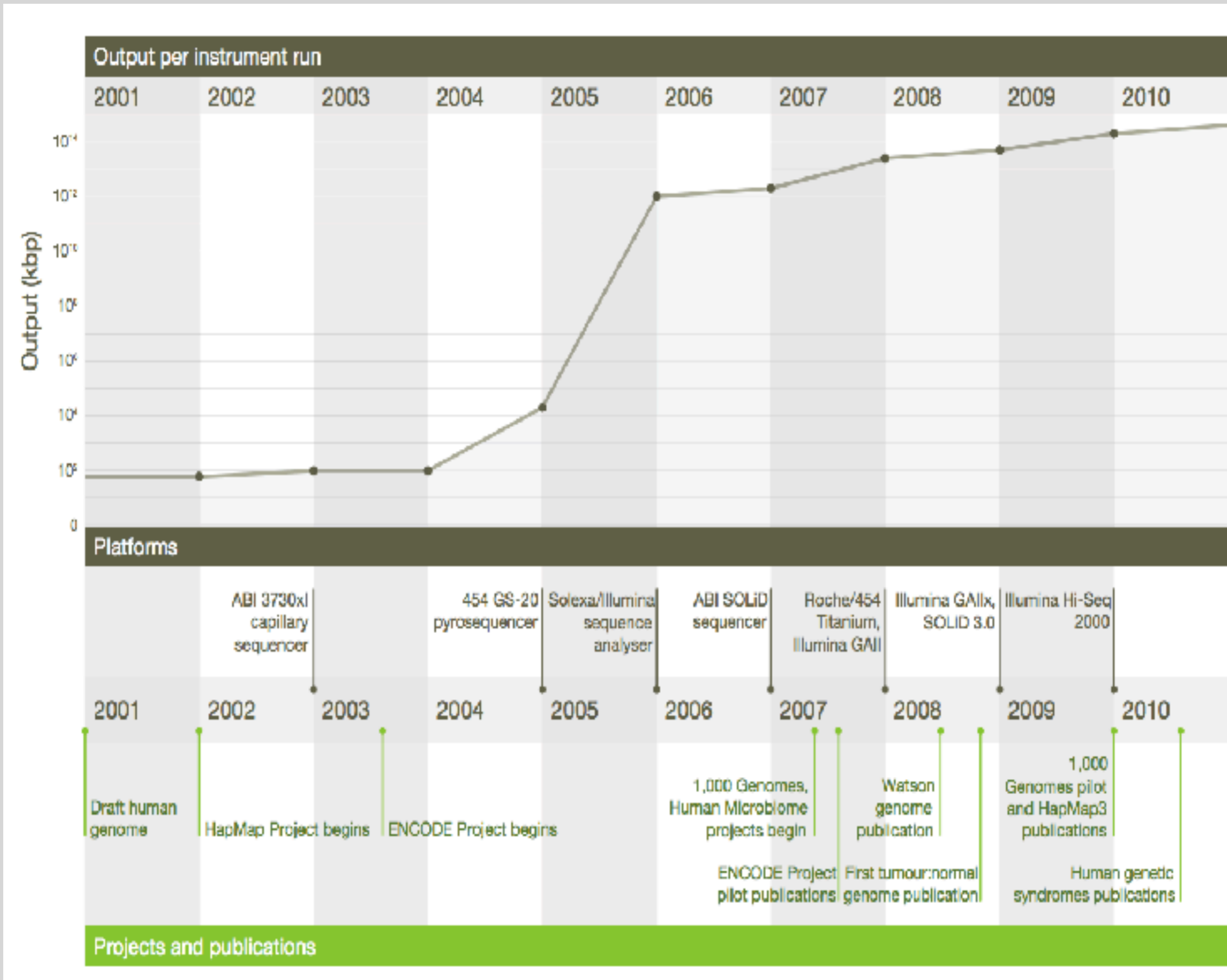
Cost per Genome



Cost per Raw Megabase of DNA Sequence



Timeline of Sequencing Capacity



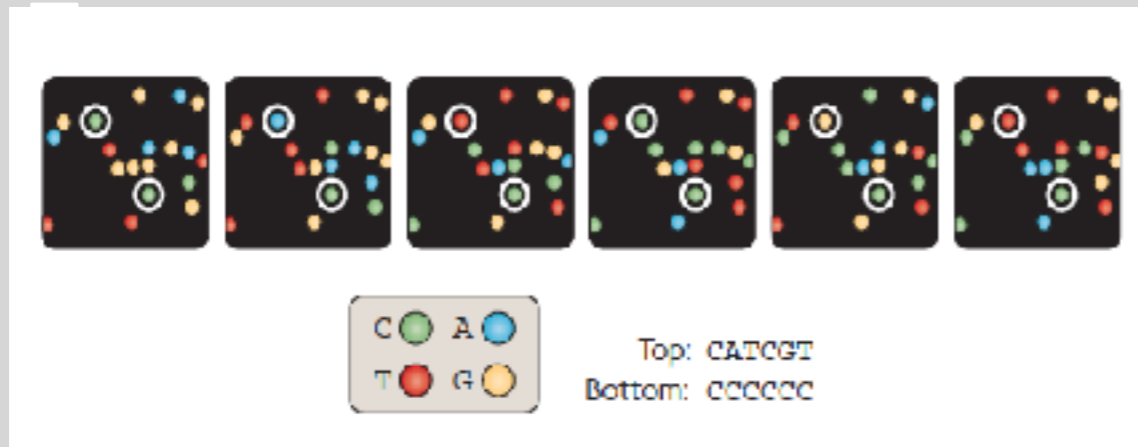
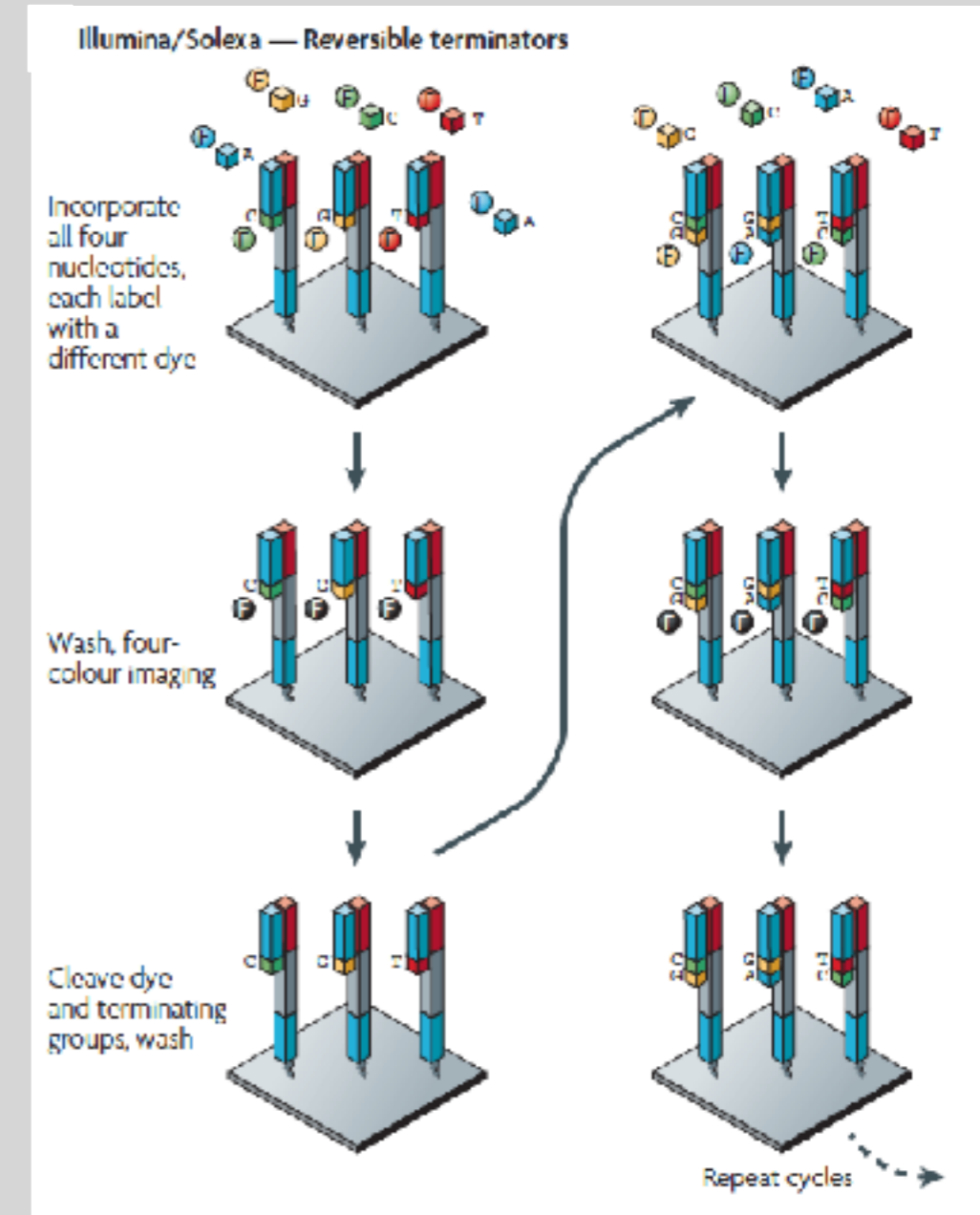
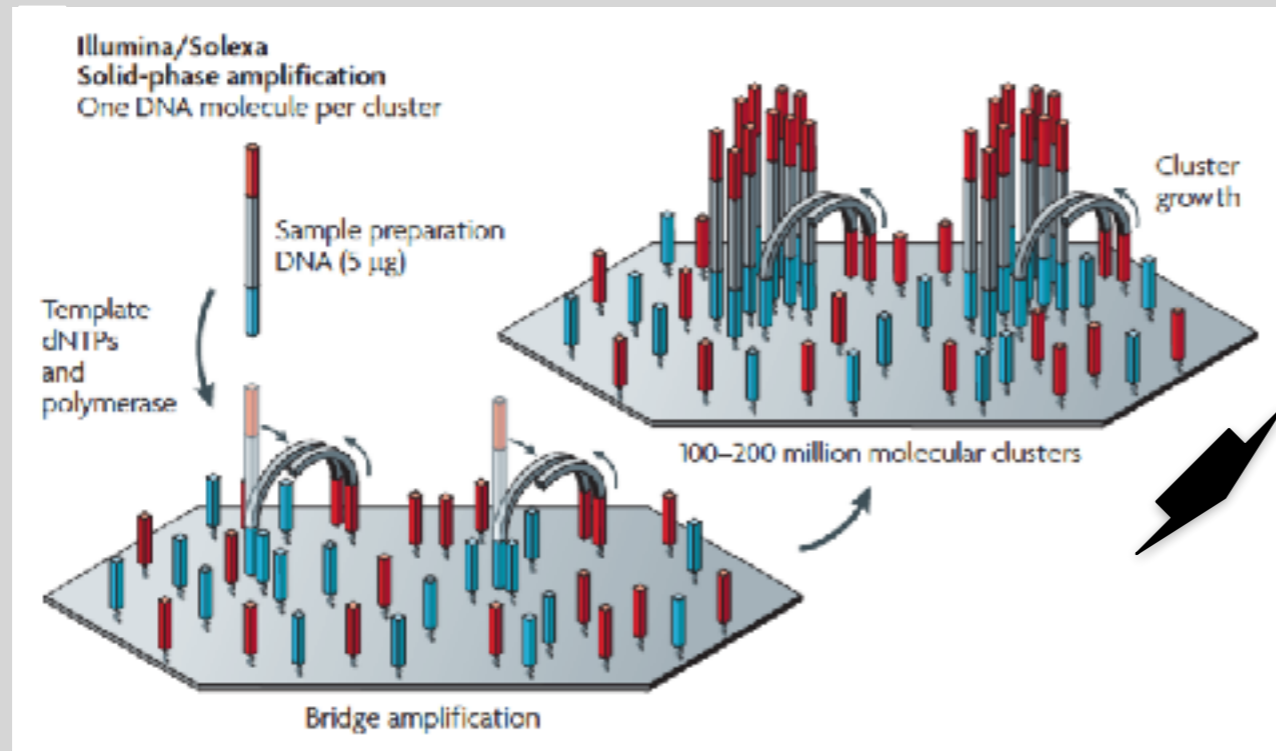
DNA Sequencing Concepts

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

Modern NGS Sequencing Platforms

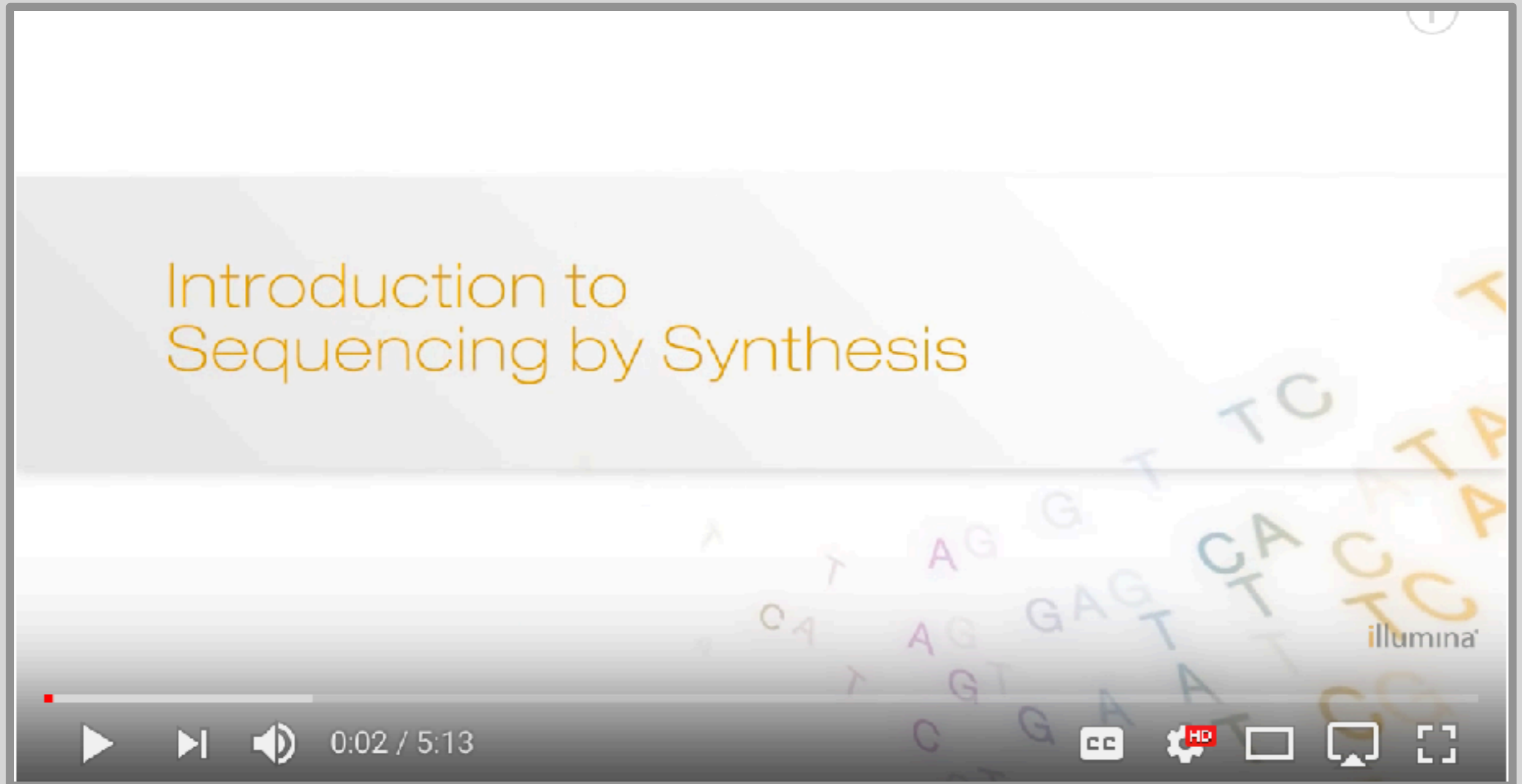
	Roche/454	Life Technologies SOLiD	Illumina Hi-Seq 2000
Library amplification method	emPCR* on bead surface	emPCR* on bead surface	Enzymatic amplification on glass surface
Sequencing method	Polymerase-mediated incorporation of unlabelled nucleotides	Ligase-mediated addition of 2-base encoded fluorescent oligonucleotides	Polymerase-mediated incorporation of end-blocked fluorescent nucleotides
Detection method	Light emitted from secondary reactions initiated by release of PPi	Fluorescent emission from ligated dye-labelled oligonucleotides	Fluorescent emission from incorporated dye-labelled nucleotides
Post incorporation method	NA (unlabelled nucleotides are added in base-specific fashion, followed by detection)	Chemical cleavage removes fluorescent dye and 3' end of oligonucleotide	Chemical cleavage of fluorescent dye and 3' blocking group
Error model	Substitution errors rare, insertion/deletion errors at homopolymers	End of read substitution errors	End of read substitution errors
Read length (fragment/paired end)	400 bp/variable length mate pairs	75 bp/50+25 bp	150 bp/100+100 bp

Illumina - Reversible terminators



(other sequencing platforms summarized at end of slide set)

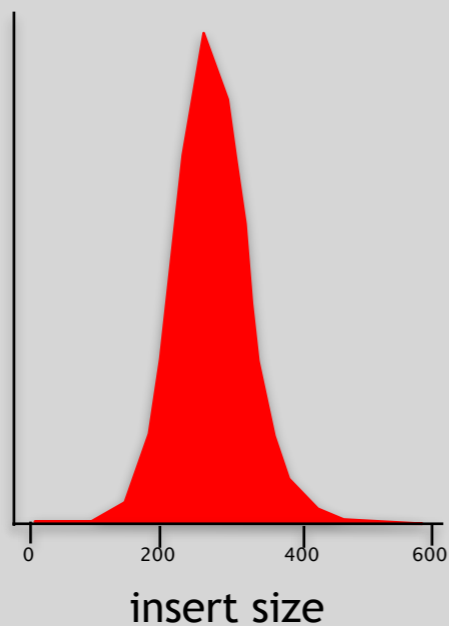
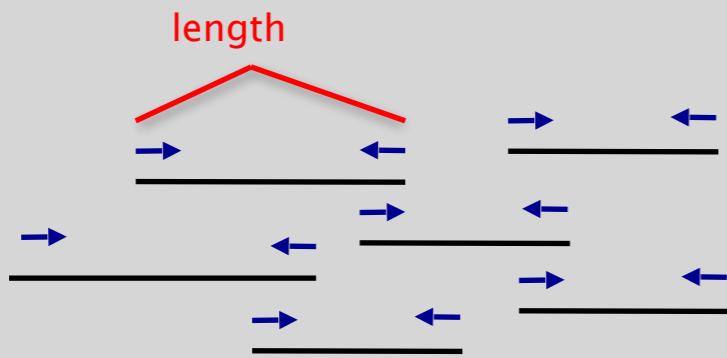
Illumina Sequencing - Video



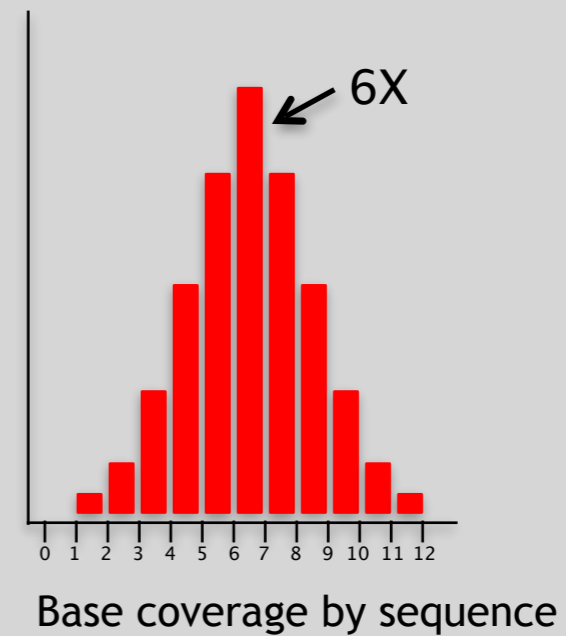
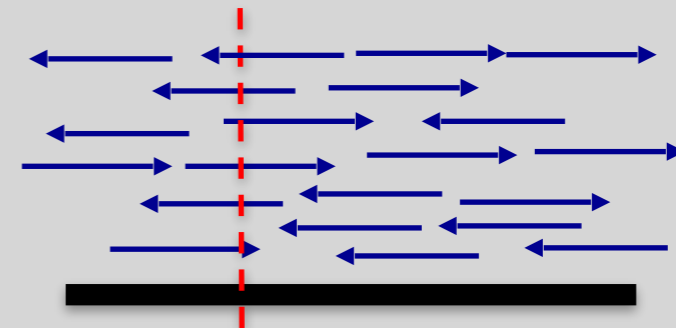
https://www.youtube.com/watch?src_vid=womKfikWlxM&v=fCd6B5HRaZ8

NGS Sequencing Terminology

Insert Size



Sequence Coverage



Summary: “Generations” of DNA Sequencing

	First generation	Second generation ^a	Third generation ^a
Fundamental technology	Size-separation of specifically end-labeled DNA fragments, produced by SBS or degradation	Wash-and-scan SBS	SBS, by degradation, or direct physical inspection of the DNA molecule
Resolution	Averaged across many copies of the DNA molecule being sequenced	Averaged across many copies of the DNA molecule being sequenced	Single-molecule resolution
Current raw read accuracy	High	High	Moderate
Current read length	Moderate (800–1000 bp)	Short, generally much shorter than Sanger sequencing	Long, 1000 bp and longer in commercial systems
Current throughput	Low	High	Moderate
Current cost	High cost per base Low cost per run	Low cost per base High cost per run	Low-to-moderate cost per base Low cost per run
RNA-sequencing method	cDNA sequencing	cDNA sequencing	Direct RNA sequencing and cDNA sequencing
Time from start of sequencing reaction to result	Hours	Days	Hours
Sample preparation	Moderately complex, PCR amplification not required	Complex, PCR amplification required	Ranges from complex to very simple depending on technology
Data analysis	Routine	Complex because of large data volumes and because short reads complicate assembly and alignment algorithms	Complex because of large data volumes and because technologies yield new types of information and new signal processing challenges
Primary results	Base calls with quality values	Base calls with quality values	Base calls with quality values, potentially other base information such as kinetics

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>

Page [Discussion](#) [Read](#) [View source](#) [View history](#)

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:

[CSV](#)
[JSON](#)

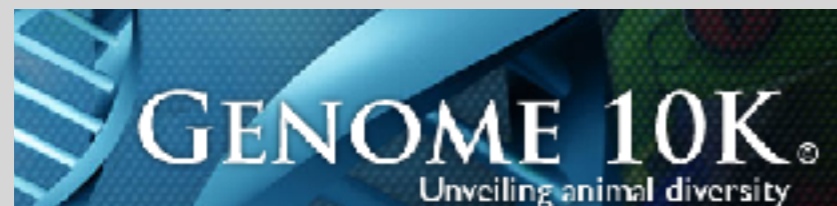
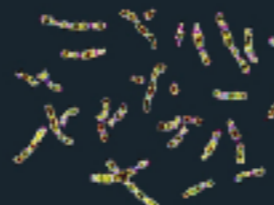
Name	Summary	Bio Tags	Meth Tags	Features	Language	Licence	OS
4peaks	Allows viewing sequencing trace files, motif searching, trimming, BLAST and exporting sequences.	Sequencing	Sequence analysis			Freeware	Mac OS X
AB Large Indel Tool	Identifies deviations in clone insert size that indicate intra-chromosomal 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
Artariter Removal	Removes artariter fragments from raw short read	General	Artariter 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

1000 Genomes
A Deep Catalog of Human Genetic Variation



The Cancer Genome Atlas  Understanding genomics to improve cancer care

The 100,000 Genomes Project

Genomics England & Partners



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

“Variety’s the very spice of life”

-William Cowper, 1785

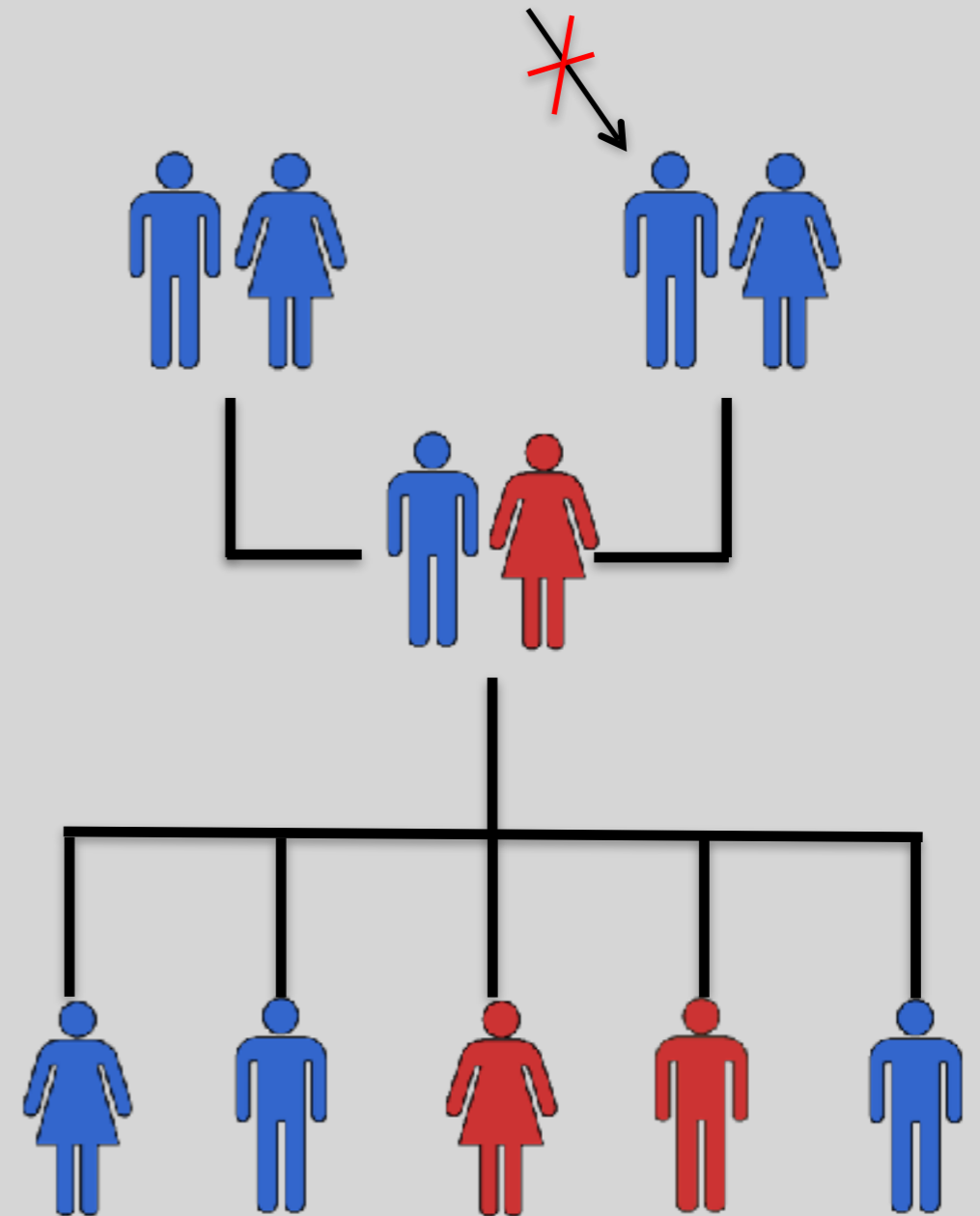
“Variation is the spice of life”

-Kruglyak & Nickerson, 2001

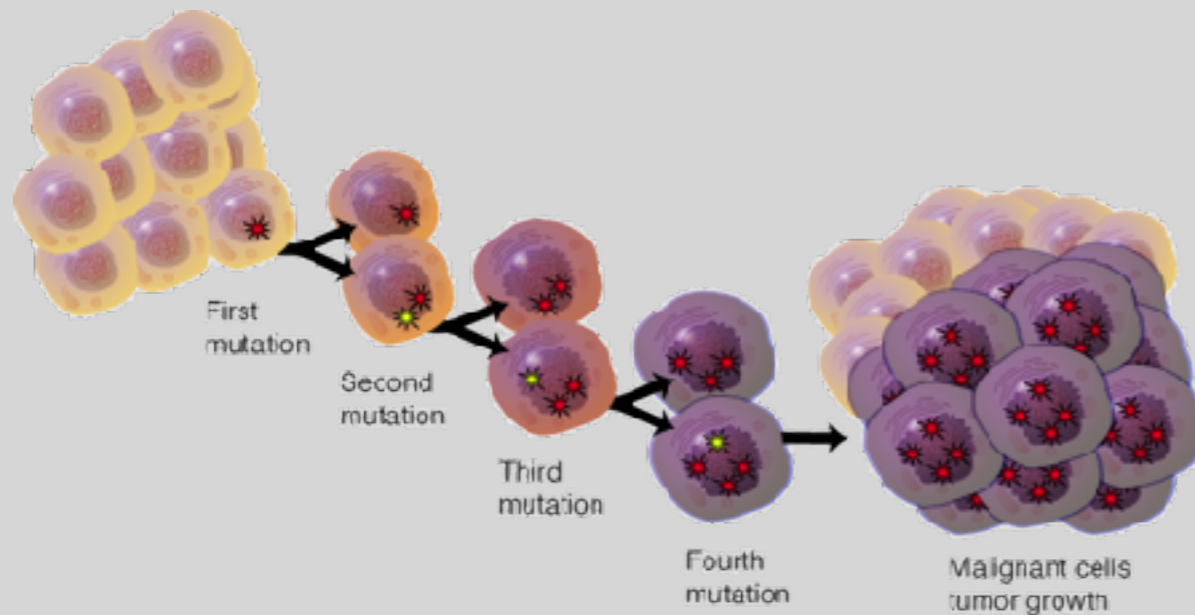
- While the sequencing of the human genome was a great milestone, the DNA from a single person is not representative of the millions of potential differences that can occur between individuals
- These unknown genetic variants could be the cause of many phenotypes such as differing morphology, susceptibility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

Germline Variation

- Mutations in the germline are passed along to offspring and are present in the DNA over every cell
- In animals, these typically occur in meiosis during gamete differentiation



Somatic Variation



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

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 simply 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.”

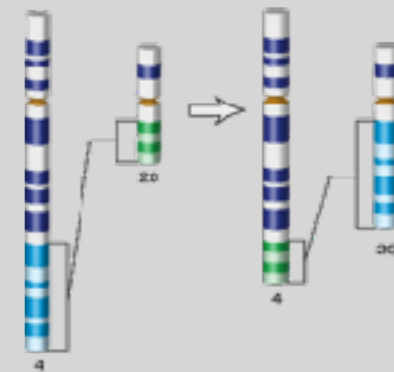
-Motoo Kimura

Types of Genomic Variation

- **Single Nucleotide Polymorphisms (SNPs)** - mutations of one nucleotide to another
- **Insertion/Deletion Polymorphisms (INDELs)** - small mutations removing or adding one or more nucleotides at a particular locus
- **Structural Variation (SVs)**
- medium to large sized rearrangements of chromosomal DNA

AATCTGAGGCAT
AATCTCAGGCAT

AATCTGAAGGCAT
AATCT--AGGCAT



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!

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

ATCCTGATTCGGTGAACGTTATCGACGATCCGATCGA
ATCCTGATTCGGTGAACGTTATCGACGATCCGATCGA
CGGTGAACGTTATCGACGATCCGATCGAACTGTCAGC
GGTGAACGTTATCGACGTTCCGATCGAACTGTCAGCG
TGAACGTTATCGACGTTCCGATCGAACTGTCATCGGC
TGAACGTTATCGACGTTCCGATCGAACTGTCAGCGGC
TGAACGTTATCGACGTTCCGATCGAACTGTCAGCGGC
GTTATCGACGATCCGATCGAACTGTCAGCGGCAAGCT
TTATCGACGATCCGATCGAACTGTCAGCGGCAAGCT

sequencing error or genetic variant?

reference genome

ATCCTGATTCGGTGAACGTTATCGACGATCCGATCGAACTGTCAGCGGCAAGCTGATCGATCGATCGATGCTAGTG

TTATCGACGATCCGATCGAACTGTCAGCGGCAAGCT
TCGACGATCCGATCGAACTGTCAGCGGCAAGCTGAT
ATCCGATCGAACTGTCAGCGGCAAGCTGATCG CGAT
TCCGAGCGAACTGTCAGCGGCAAGCTGATCG CGATC
TCCGATCGAACTGTCAGCGGCAAGCTGATCGATCGA
GATCGAACTGTCAGCGGCAAGCTGATCG CGATCGA
AACTGTCAGCGGCAAGCTGATCG CGATCGATGCTA
TGTCAGCGGCAAGCTGATCGATCGATCGATGCTAG
TCAGCGGCAAGCTGATCGATCGATCGATGCTAGTG

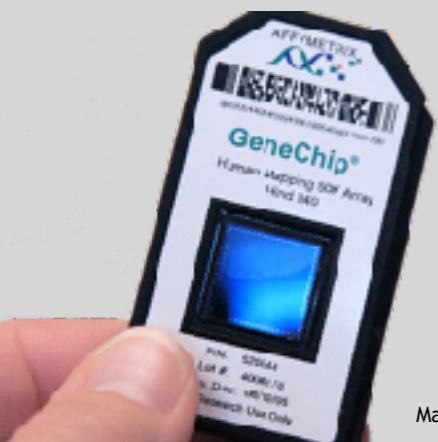
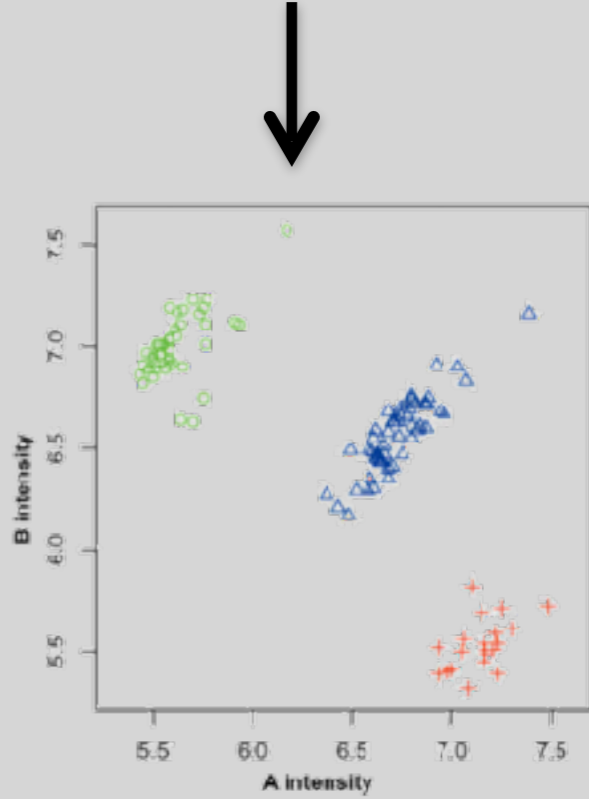
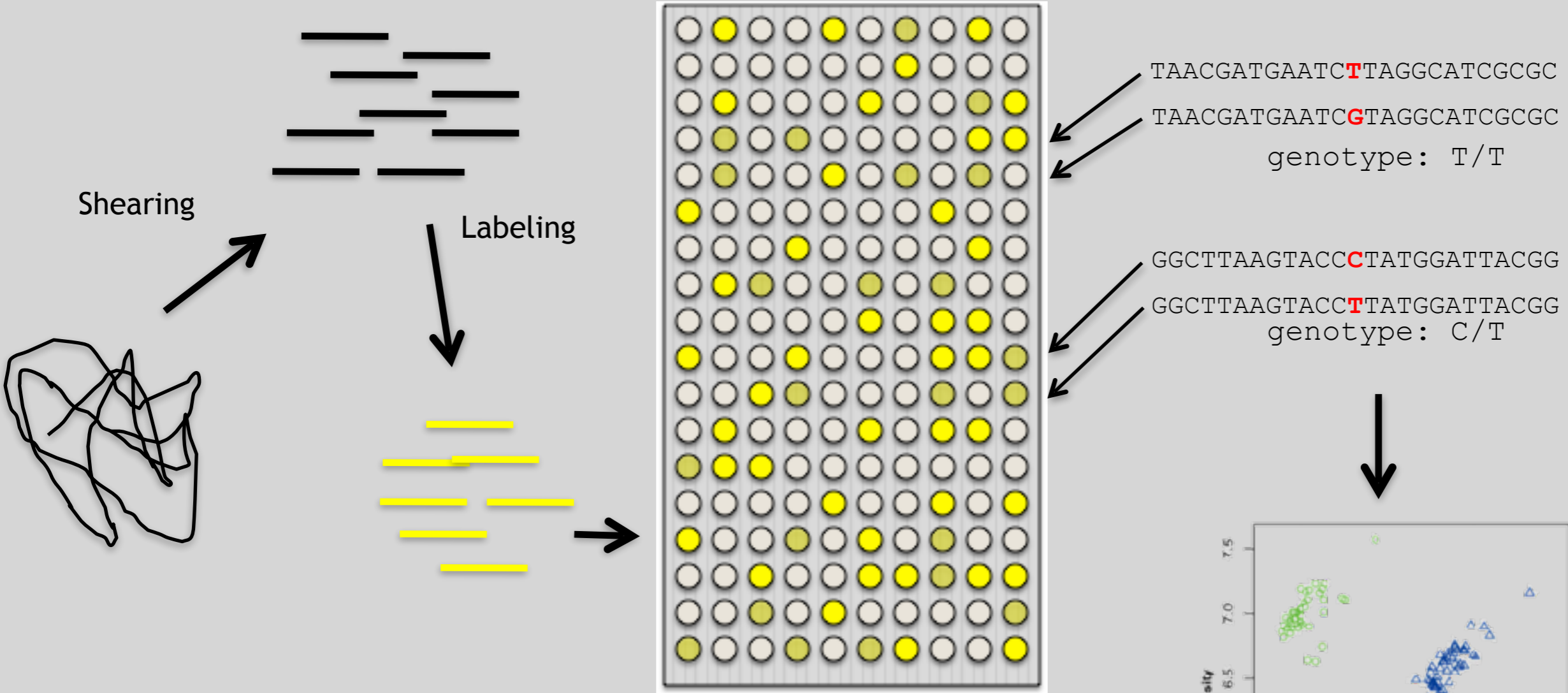
sequencing error or genetic variant?

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



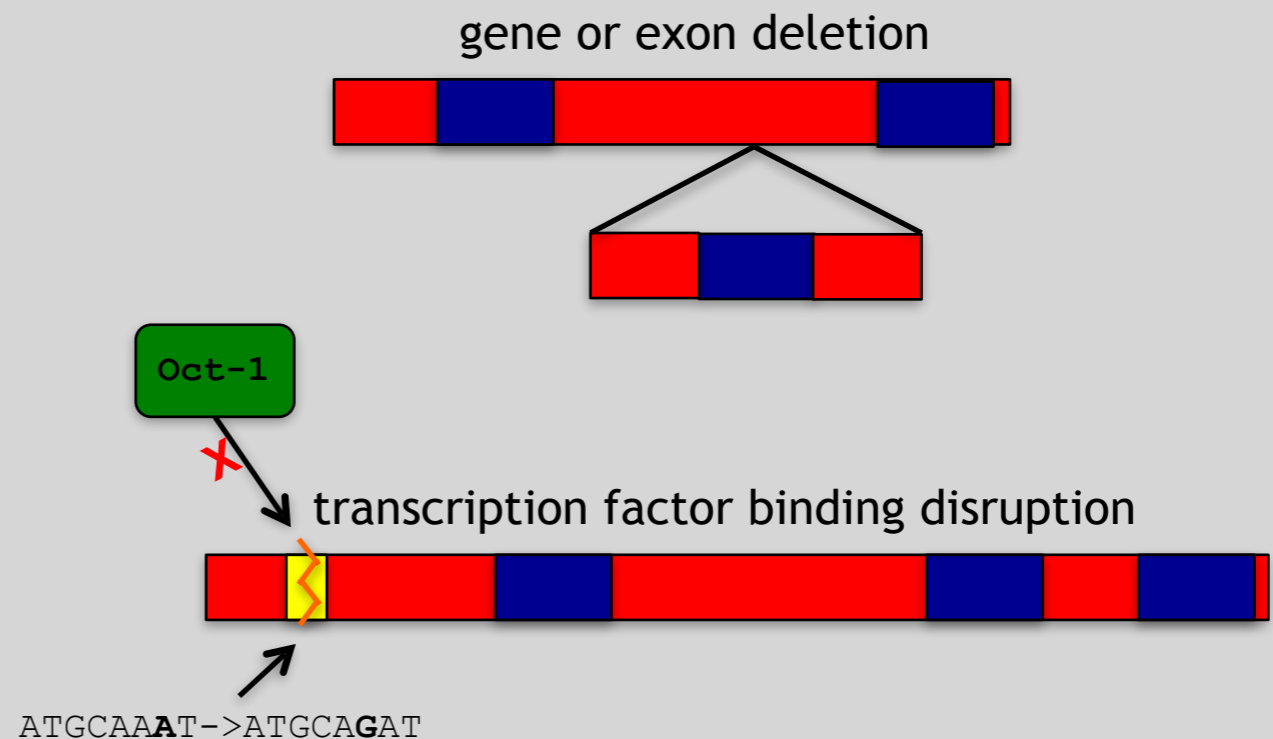
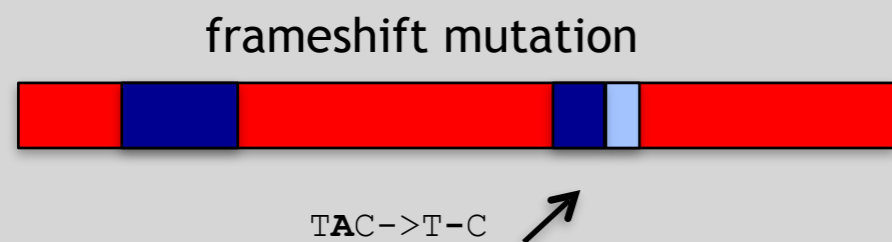
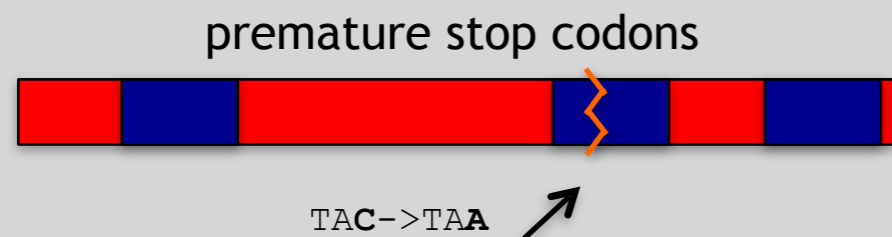
Maggie Bartlett, Courtesy: [National Human Genome Research Institute](http://www.nhgri.nih.gov).

Discovering Variation: SVs

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

Impact of Genetic Variation

There are numerous ways genetic variation can exhibit functional effects



Geuvadis Consortium

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

gEUVADIS | CONTACT US



[HOME](#) [Project](#) [Partners](#) [News & Events](#) [Publications](#) [Resources](#) [Related Projects](#) [PRIVATE](#)

Login

Email Address

Password

Remember Me

[Sign In](#)

Logout

[Click here to Logout](#)

Search

Related Events

[ESGI Symposium on Functional Genomics and Metabolism Research](#)
21st and 22nd March 2013

[Discussing whole genome sequencing in medical practice](#)
November 7th 2012

[From Genetic Discovery to Future Health](#)
November 15th, 2012

[International Congress of Human Genetics 2011](#)
11-15.10.2011

[The Genomics of Common Diseases 2011](#)
30.08 - 02.09 2011

[4th Paris Workshop on Genomic Epidemiology](#)
May 30, 31 & June 1, 2011

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

[GEUVADIS RNA sequencing project for 1000 Genomes samples](#)



Welcome !

[Welcome to the GEUVADIS website](#)

We are committed to gaining insights into the **human genome** and its role in **health and medicine** by sharing data, experience and expertise in **high-throughput sequencing**.

The purpose of this website is to keep you up to date with the project, and to help you find accessible information about genomics and personalised medicine.

Funded by the European Commission (FP7, HEALTH), GEUVADIS brings together 17 partners including academic institutes and private companies from 7 different countries.

Upcoming Geuvadis Events

Genomic Medicine in the Mediterranean
Inaugural conference
Hersonissos, Crete, Greece
October 2-5, 2013

GENOMIC MEDICINE IN THE MEDITERRANEAN (GM²)
INAUGURAL CONFERENCE
OCTOBER 2-5, 2013
HERONISSOS, CRETE, GREECE

[more...](#)

Latest News

[Transcriptome and genome sequencing uncovers functional variation in humans](#)
15.09.2013

[Check-out our 10 GEUVADIS publications](#)
17.09.12

[Results from a GEUVADIS study presented at the Genomes Network Conference in London](#)
01.05.2012

[Study Says Predictive Whole-Genome Sequencing Is Probably Not Very Useful](#)
08.05.2012

[Sequencing projects bring age-old wisdom to genomics](#)
07.11.2011

[The new data, new format, new goals and new sponsor of the Anthon Genomics X PRIZE Competition](#)
27.10.2011

[We updated our project and project-related publication list. Feel free to have a look!](#)
02.09.2011

[Listen to our podcast!](#)
27.07.2011

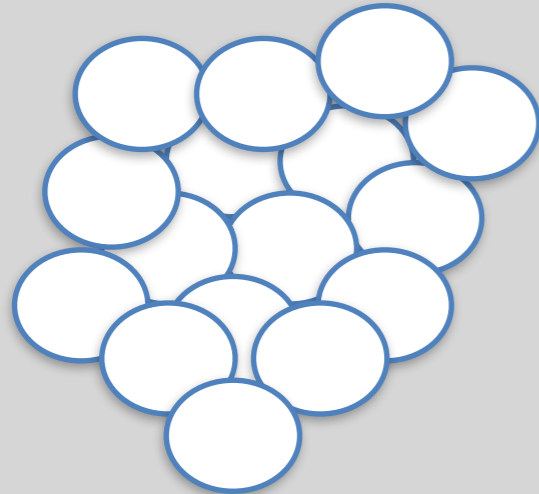
[We are now on facebook!](#)
05.07.2011

[A framework for variation discovery and genotyping using next-generation DNA sequencing data](#)
10.04.2011

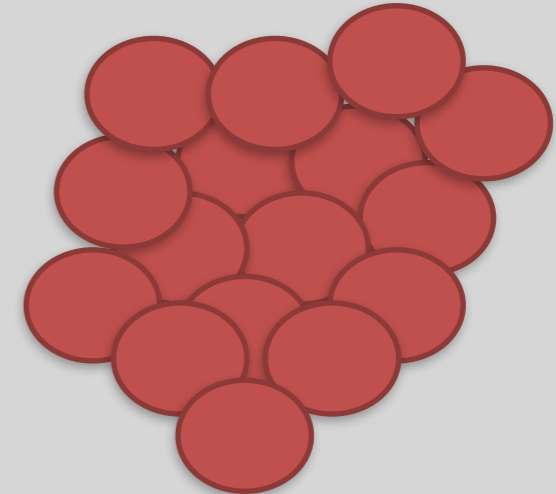
RNA Sequencing

The absolute basics

Normal Cells

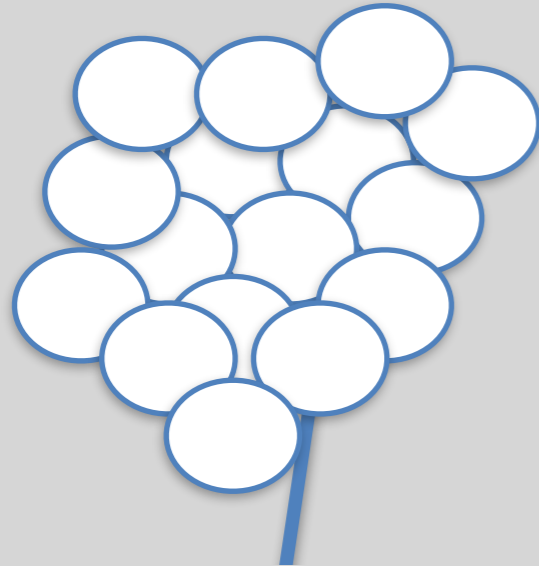


Mutated Cells

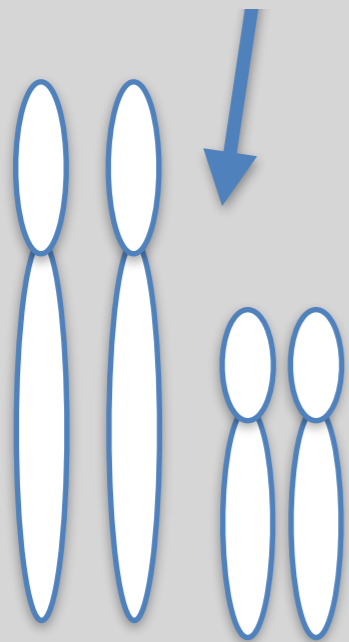


- The **mutated cells** behave differently than the **normal cells**
- We want to know what genetic mechanism is causing the difference
- One way to address this is to examine differences in gene expression via RNA sequencing...

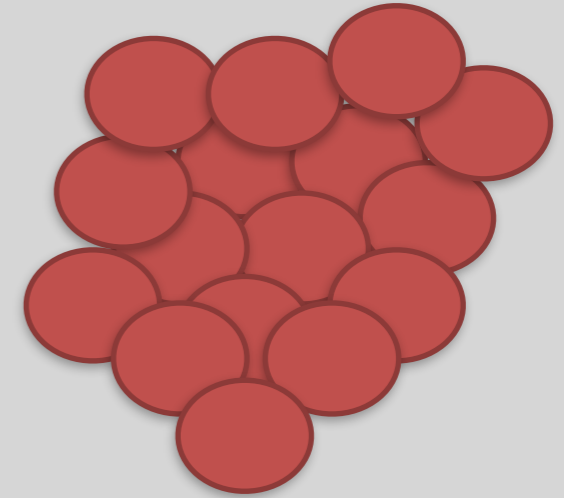
Normal Cells



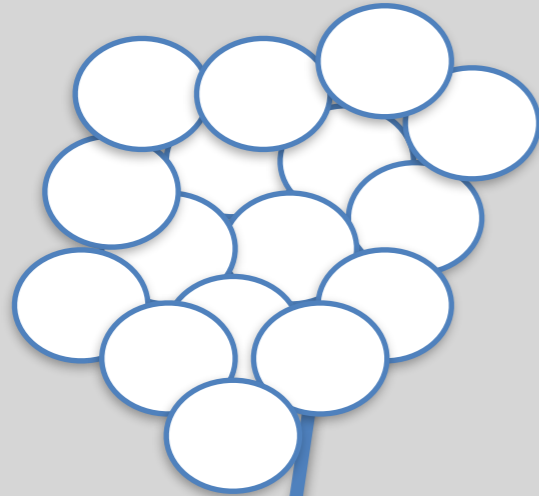
Each cell has a bunch of chromosomes



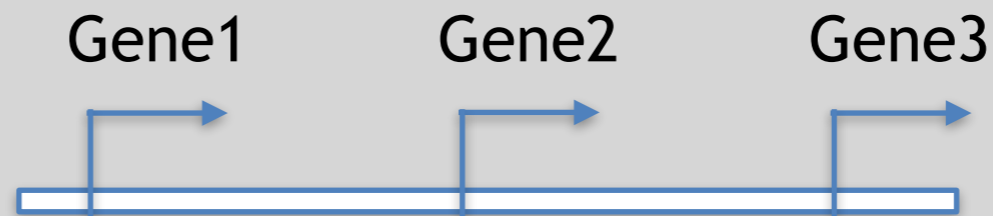
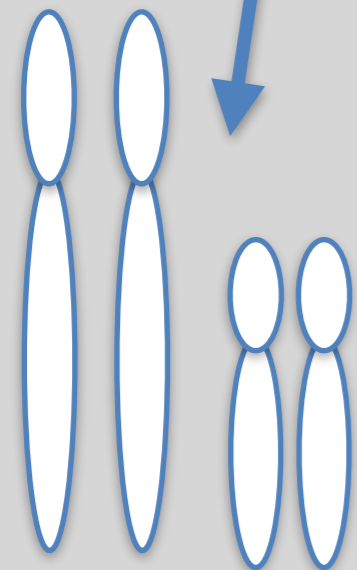
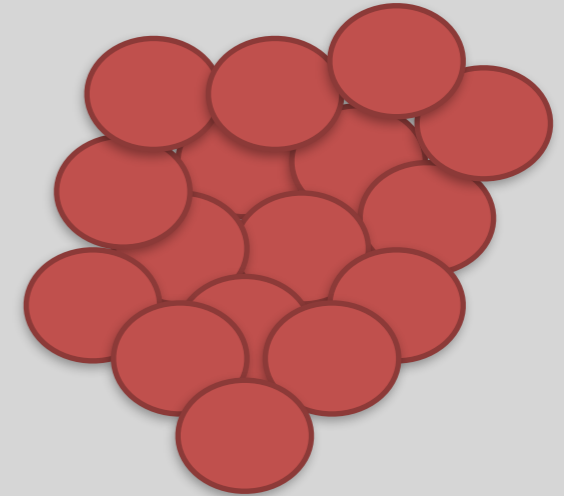
Mutated Cells



Normal Cells

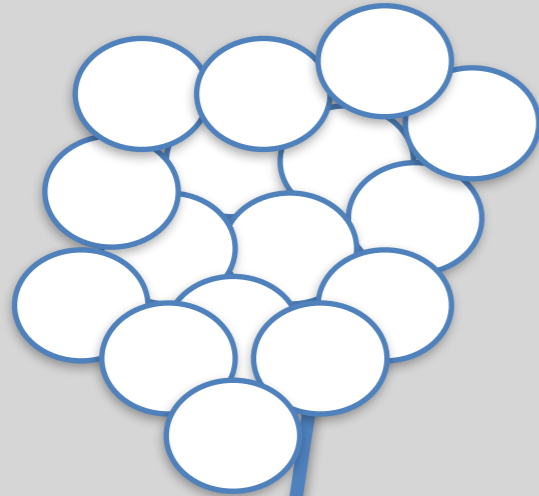


Mutated Cells

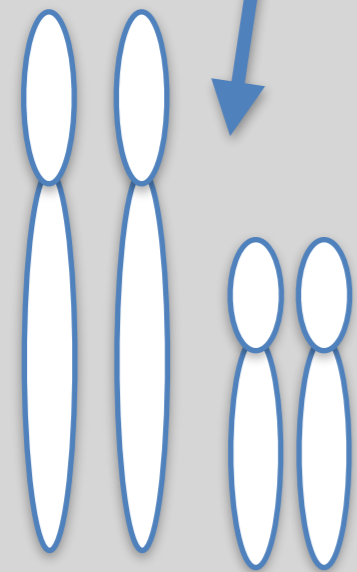
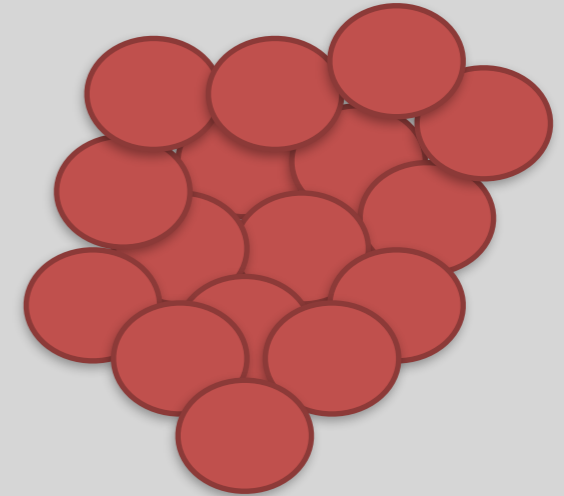


Each chromosome has
a bunch of genes

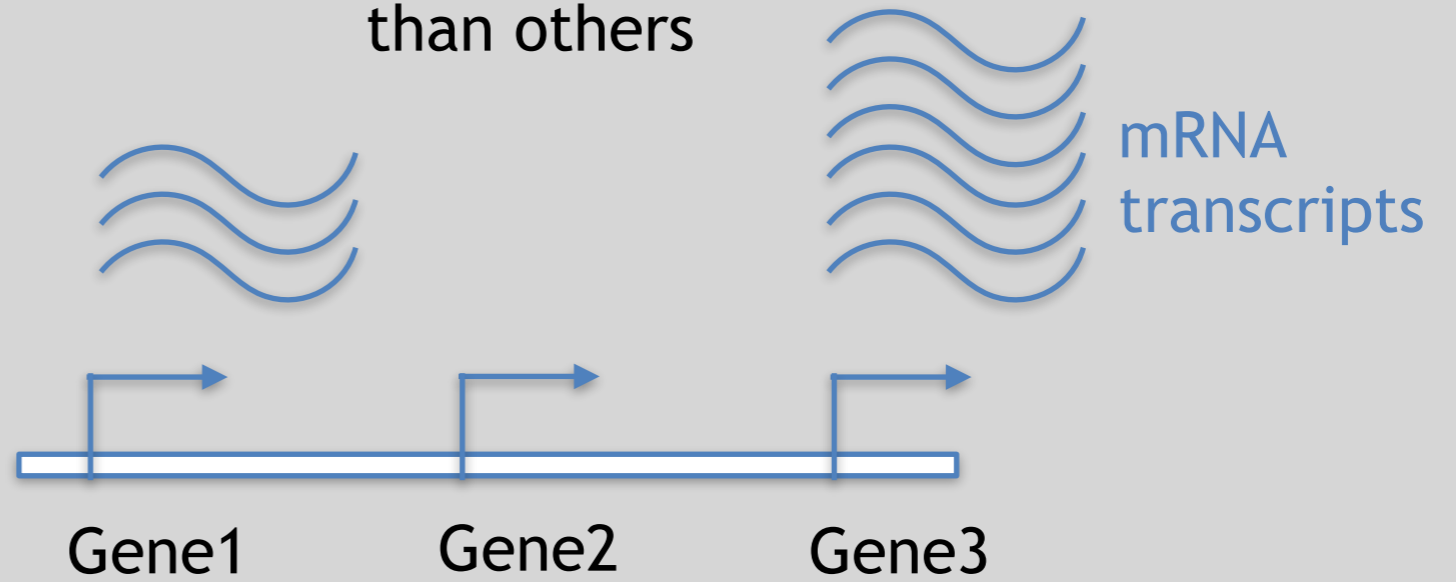
Normal Cells



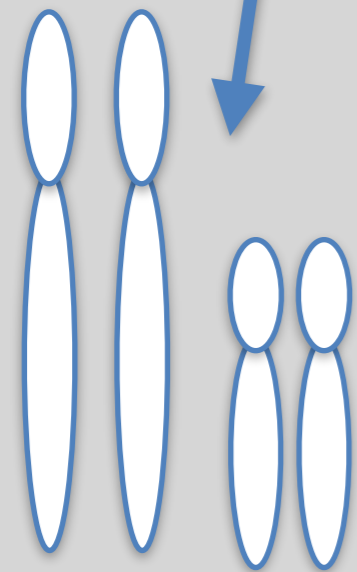
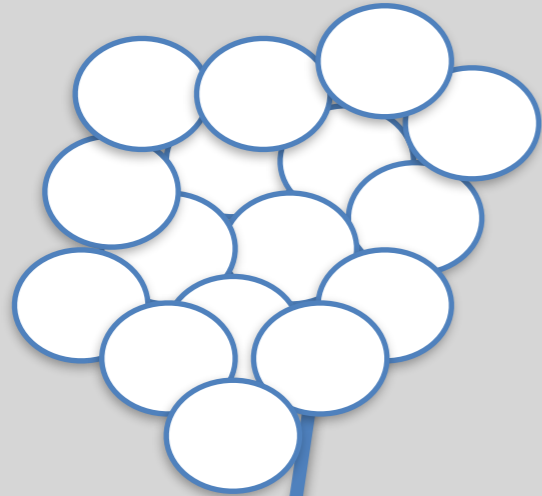
Mutated Cells



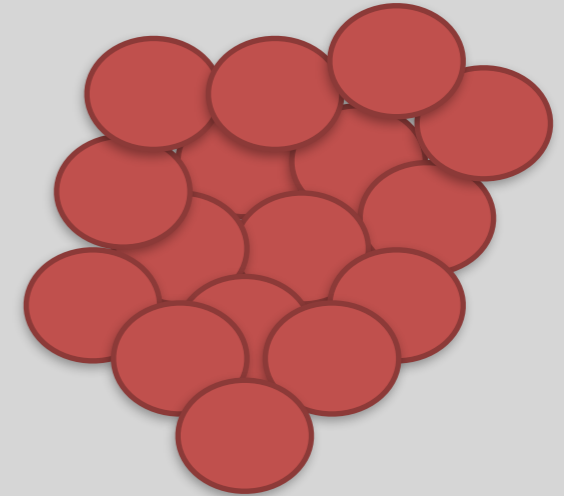
Some genes are active more than others



Normal Cells

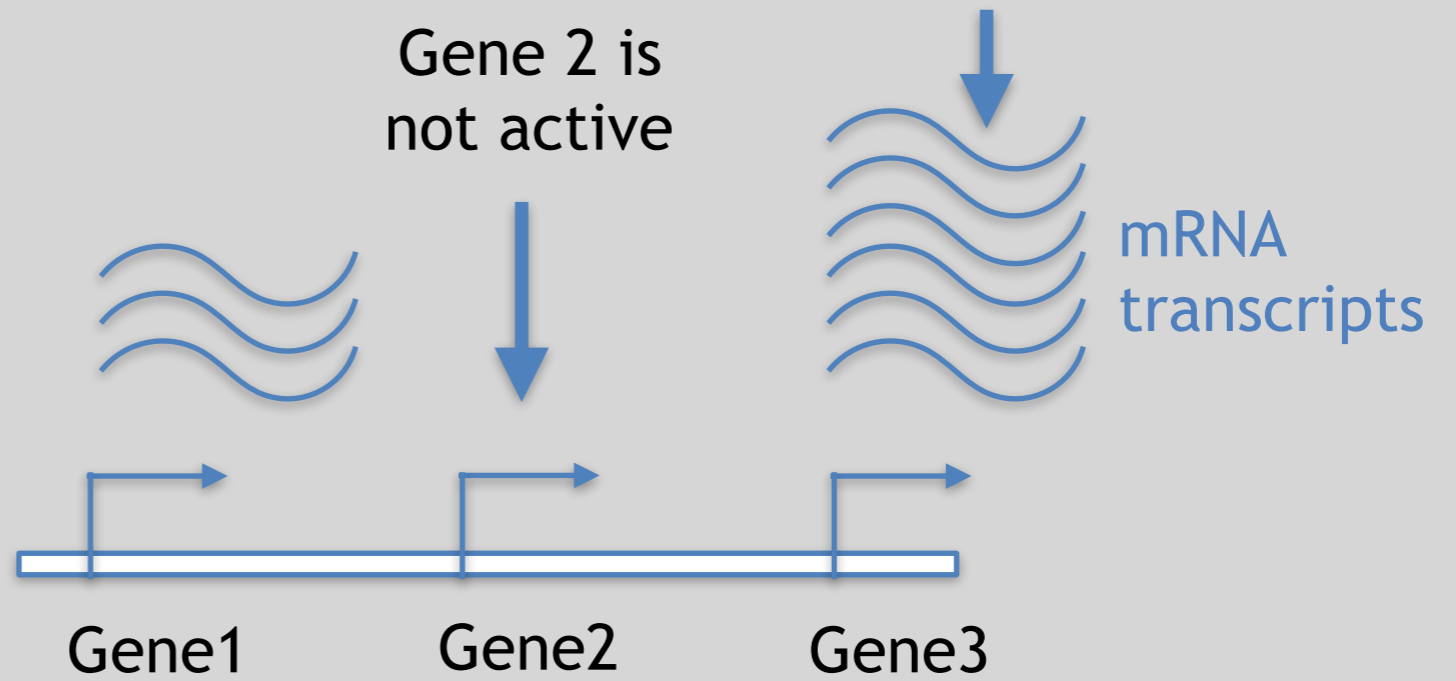


Mutated Cells

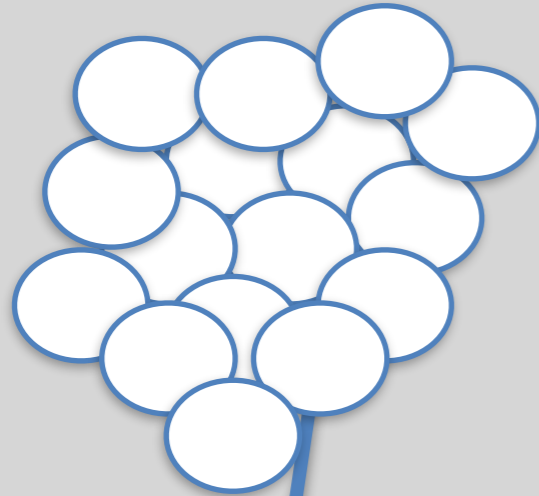


Gene 3 is the most active

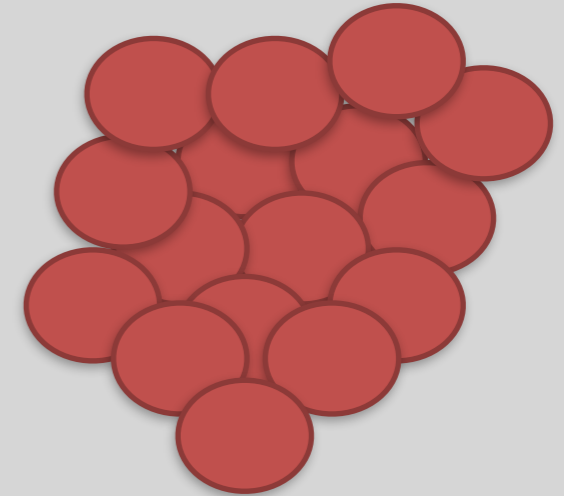
Gene 2 is not active



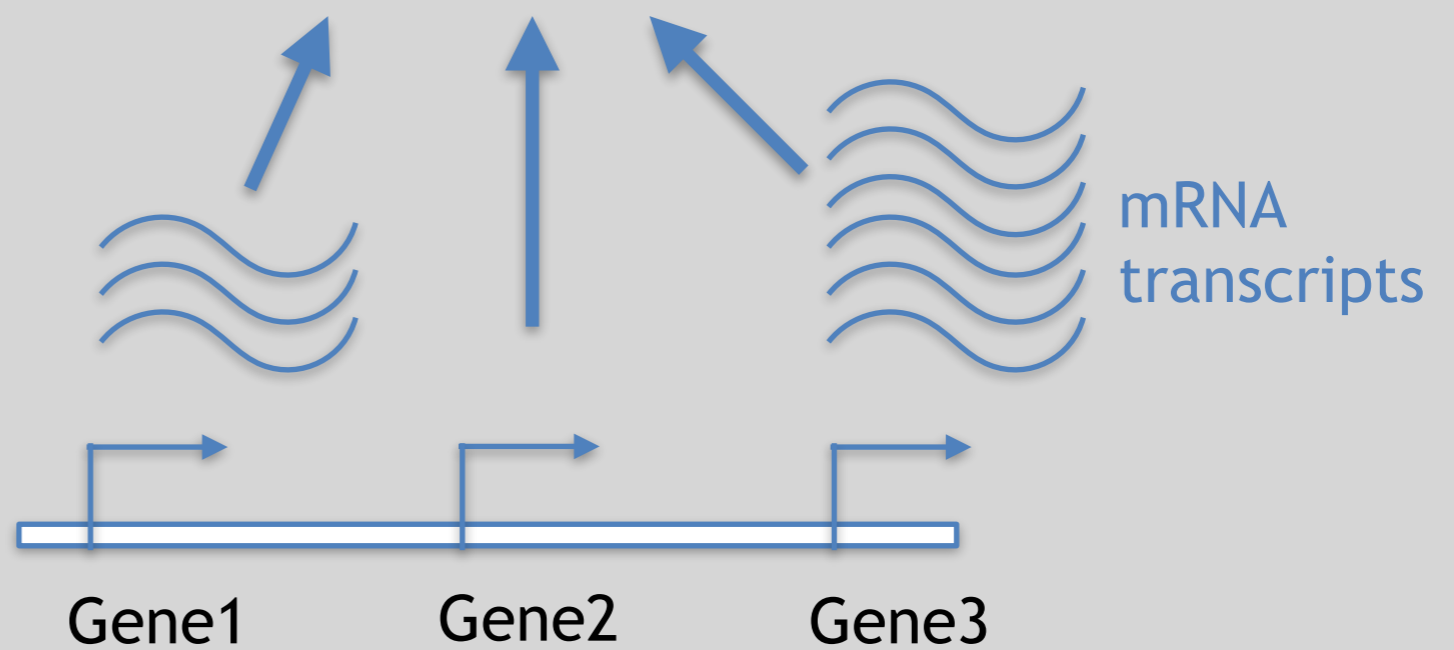
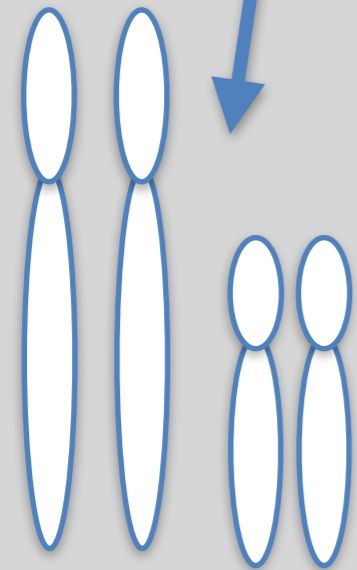
Normal Cells



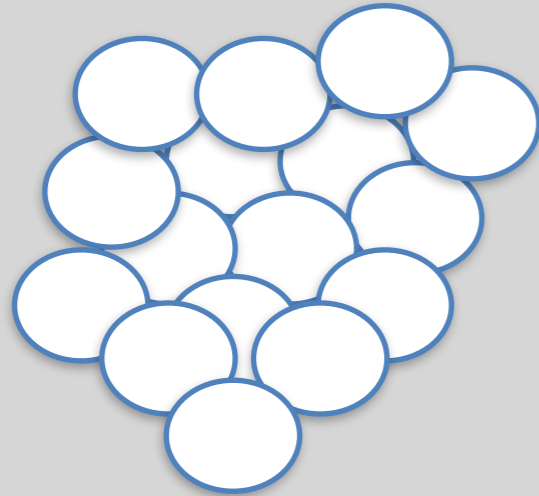
Mutated Cells



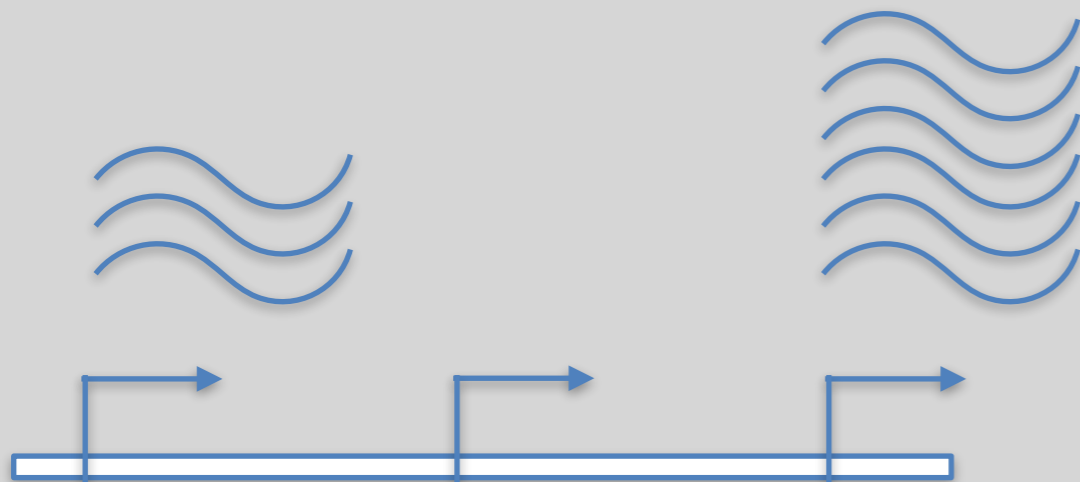
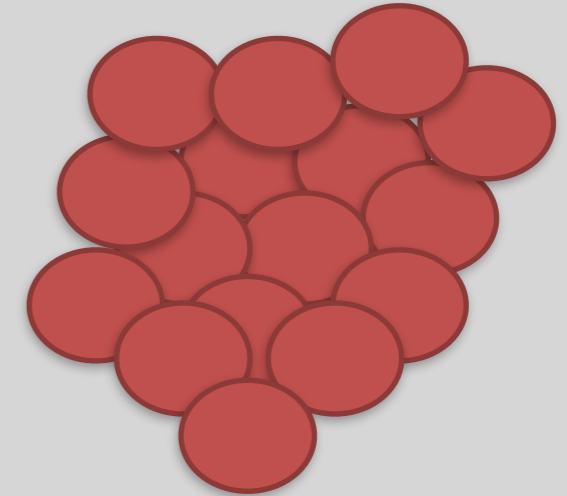
HTS tells us which genes are active, and how much they are transcribed!



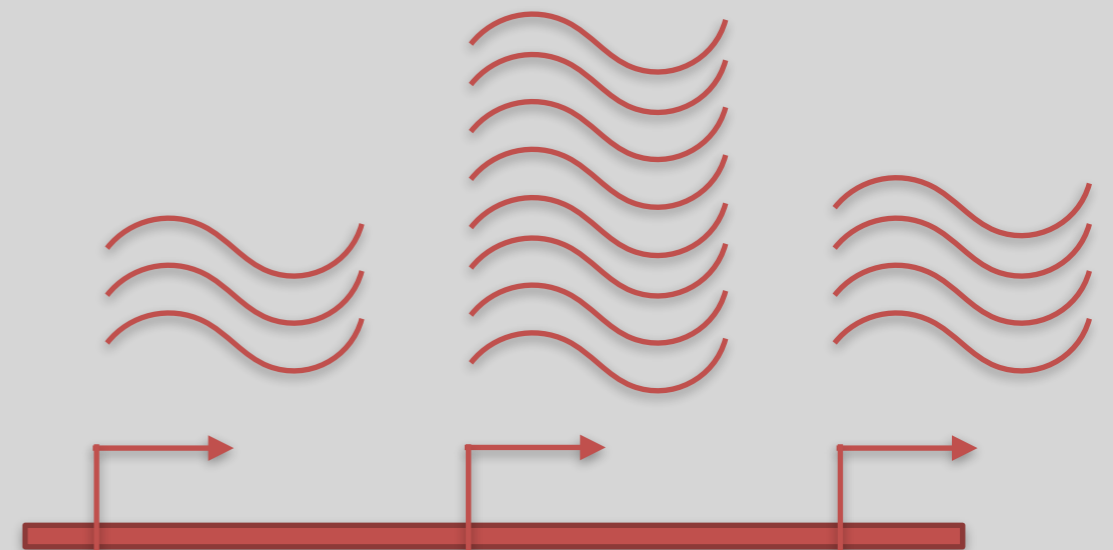
Normal Cells



Mutated Cells

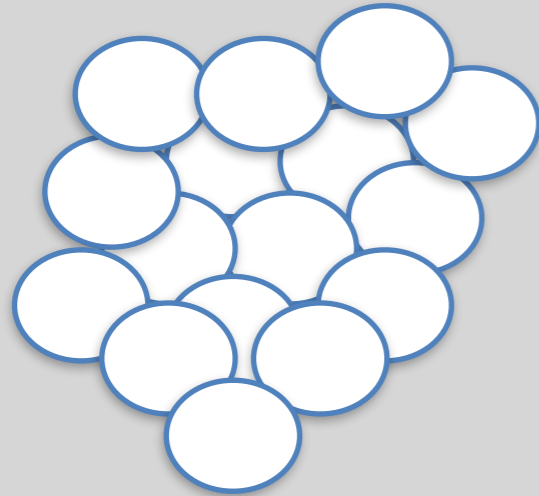


We use RNA-Seq to measure gene expression in normal cells ...

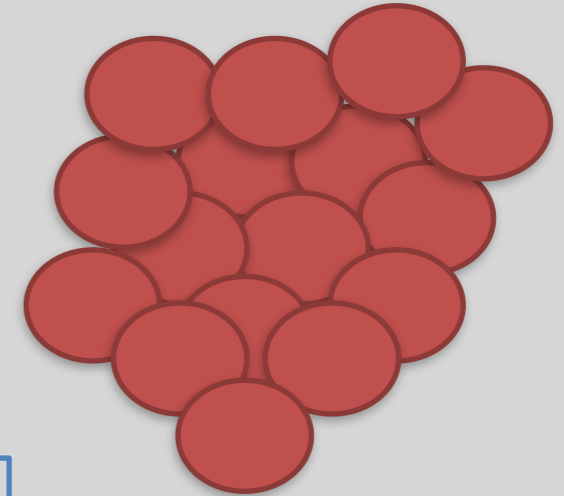


... then use it to measure gene expression in mutated cells

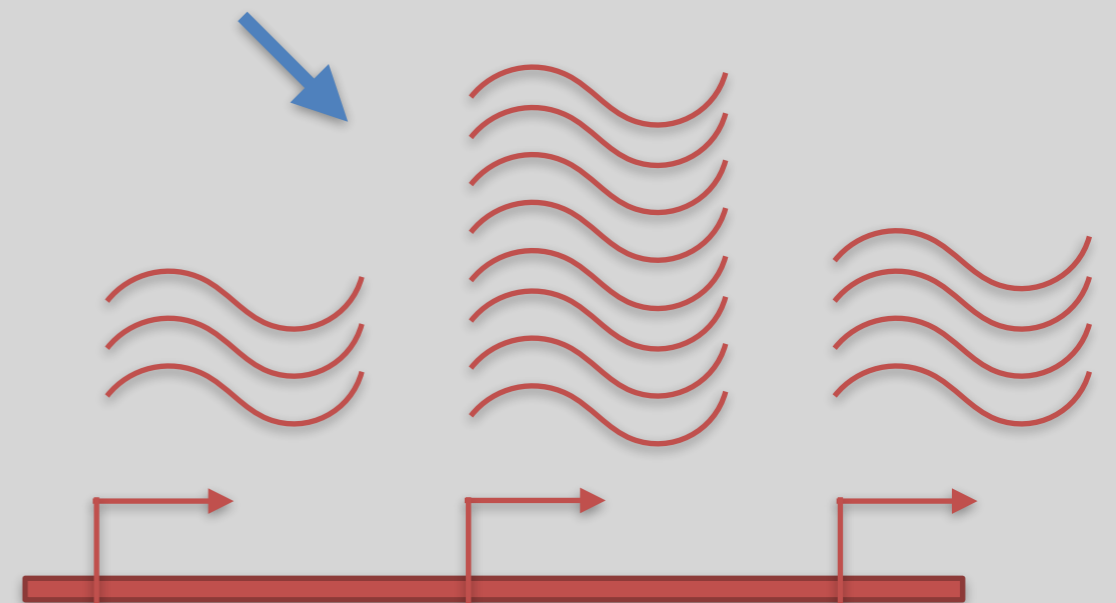
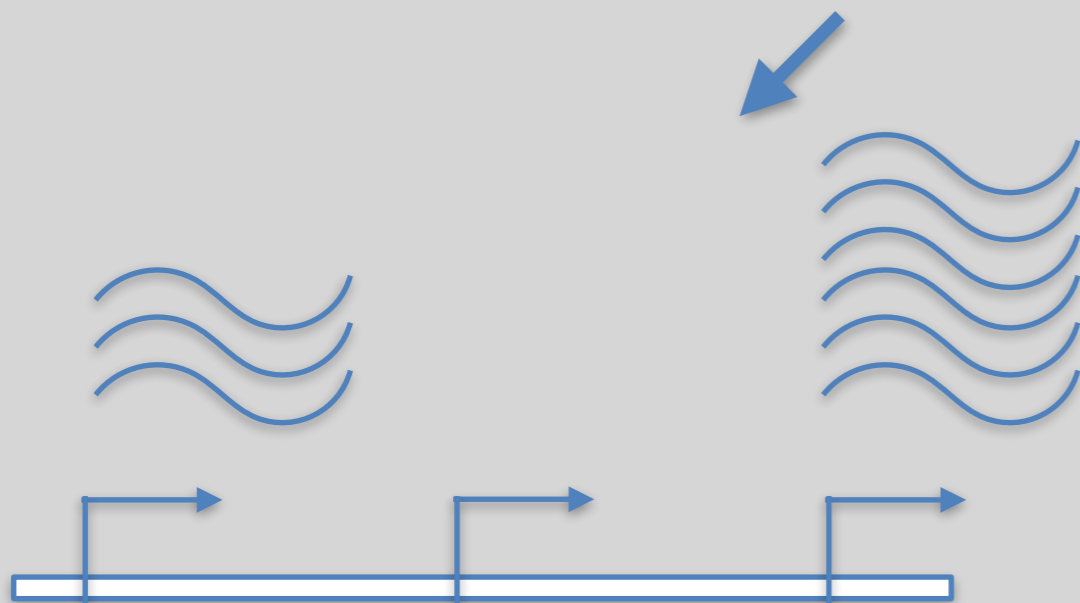
Normal Cells



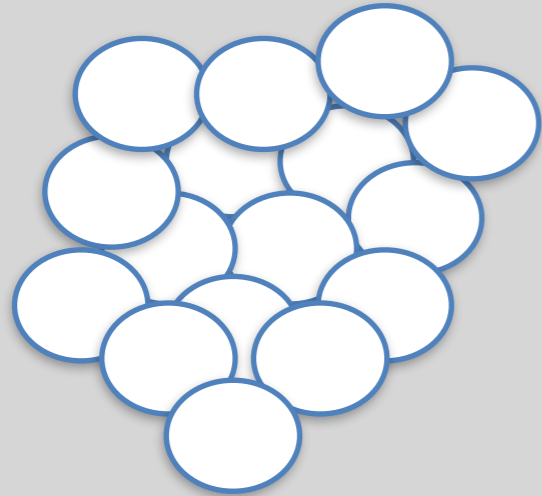
Mutated Cells



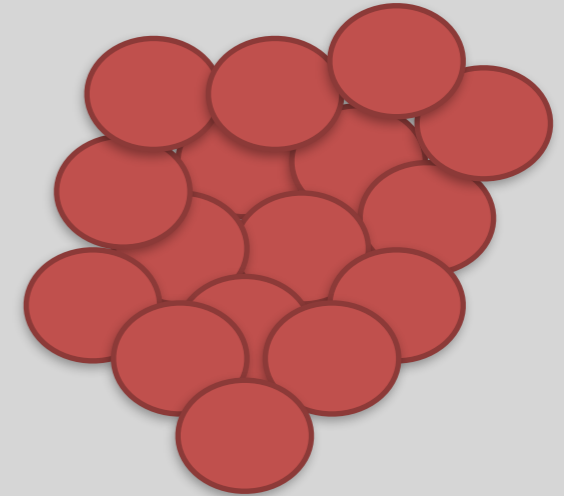
Then we can compare the two cell types to figure out what is different in the mutated cells!



Normal Cells

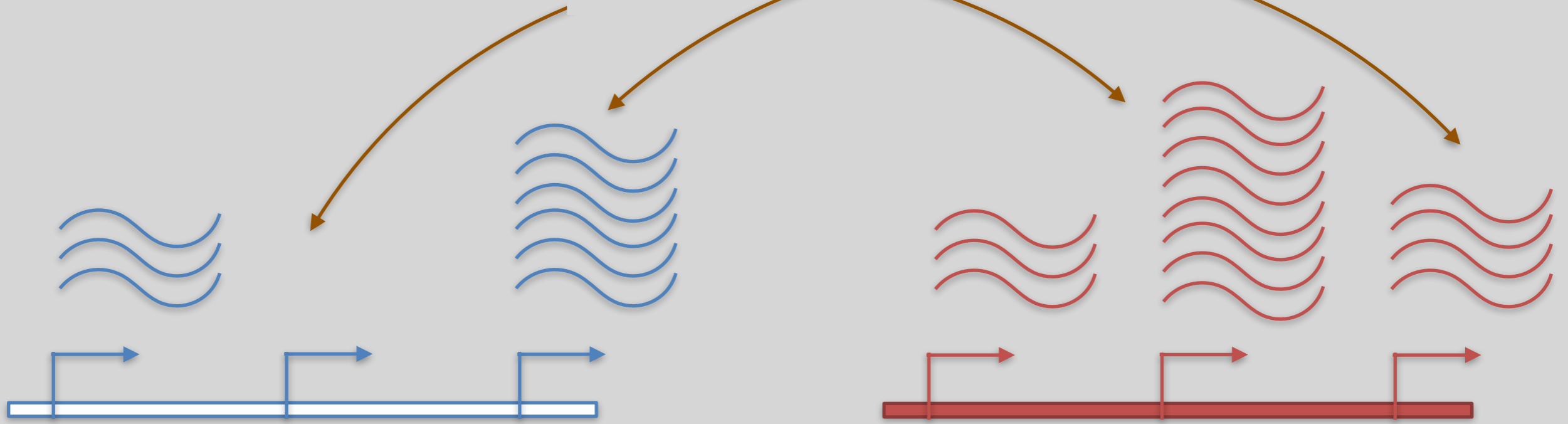


Mutated Cells



Gene2

Gene3



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

3 Main Steps for RNA-Seq:

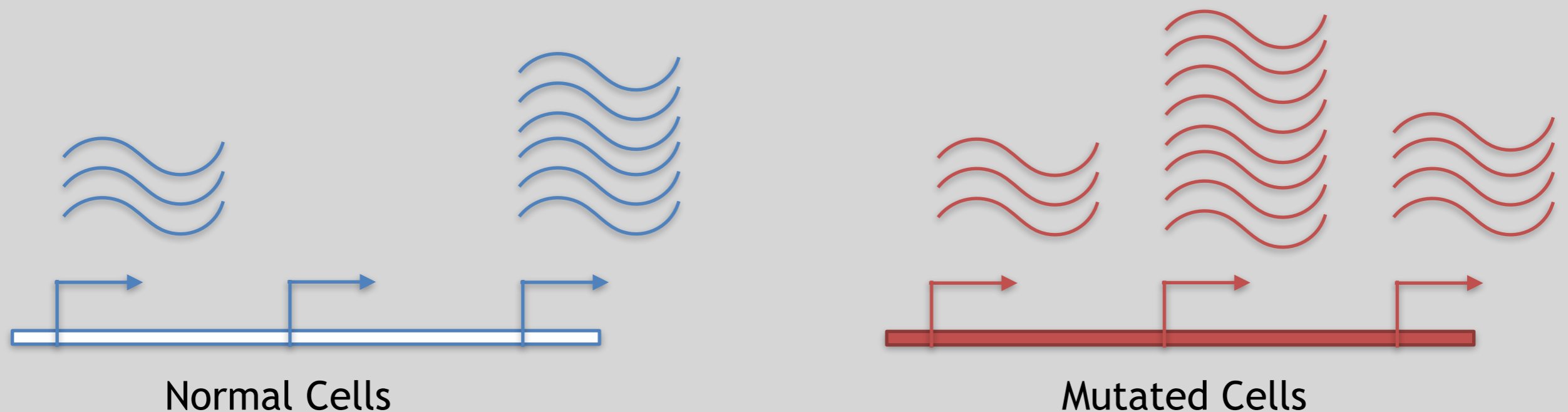
- 1) Prepare a sequencing library**
(RNA to cDNA conversion via reverse transcription)
- 2) Sequence**
(Using the same technologies as DNA sequencing)
- 3) Data analysis**
(Often the major bottleneck to overall success!)

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:

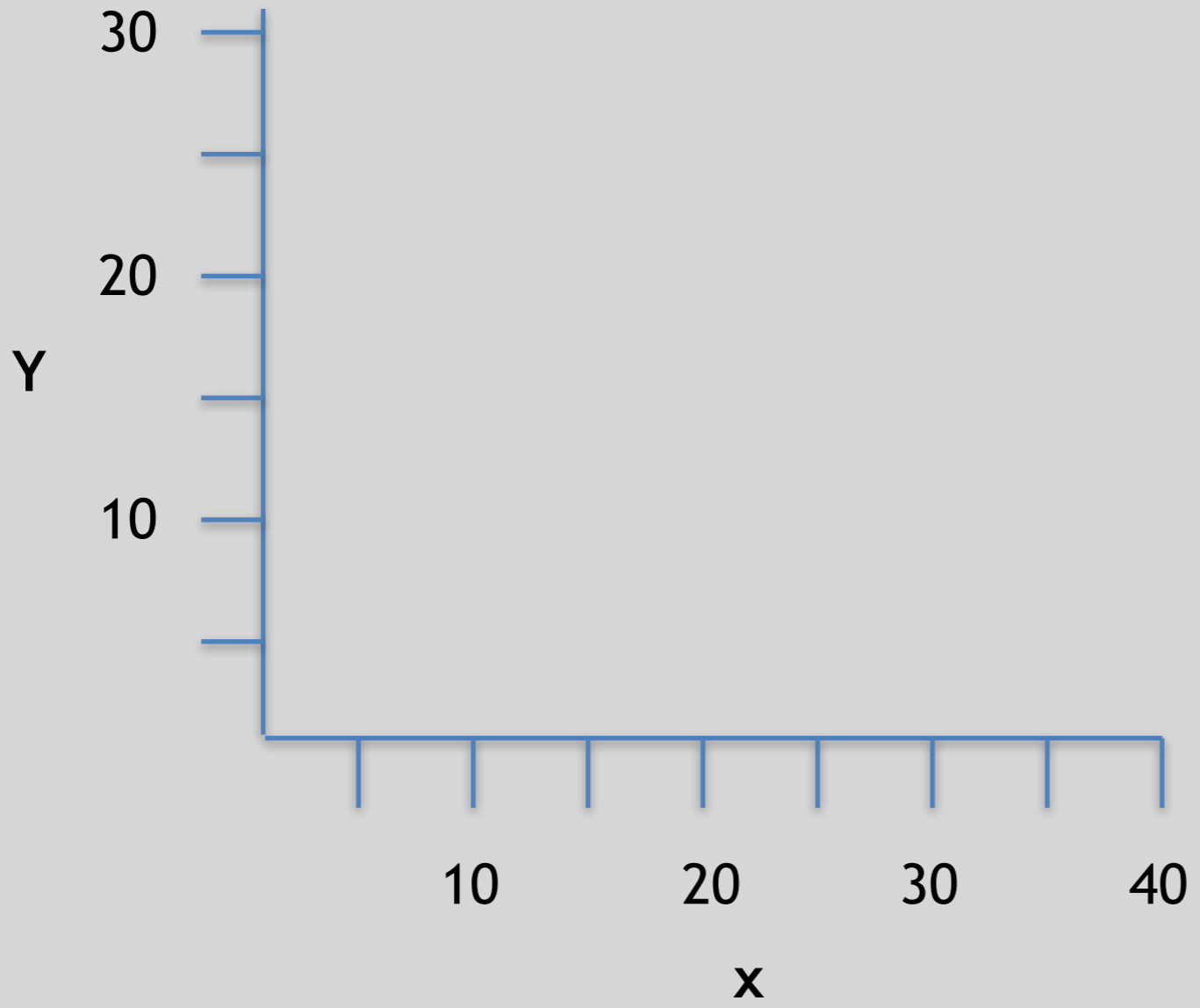
PLOT THE DATA!!

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

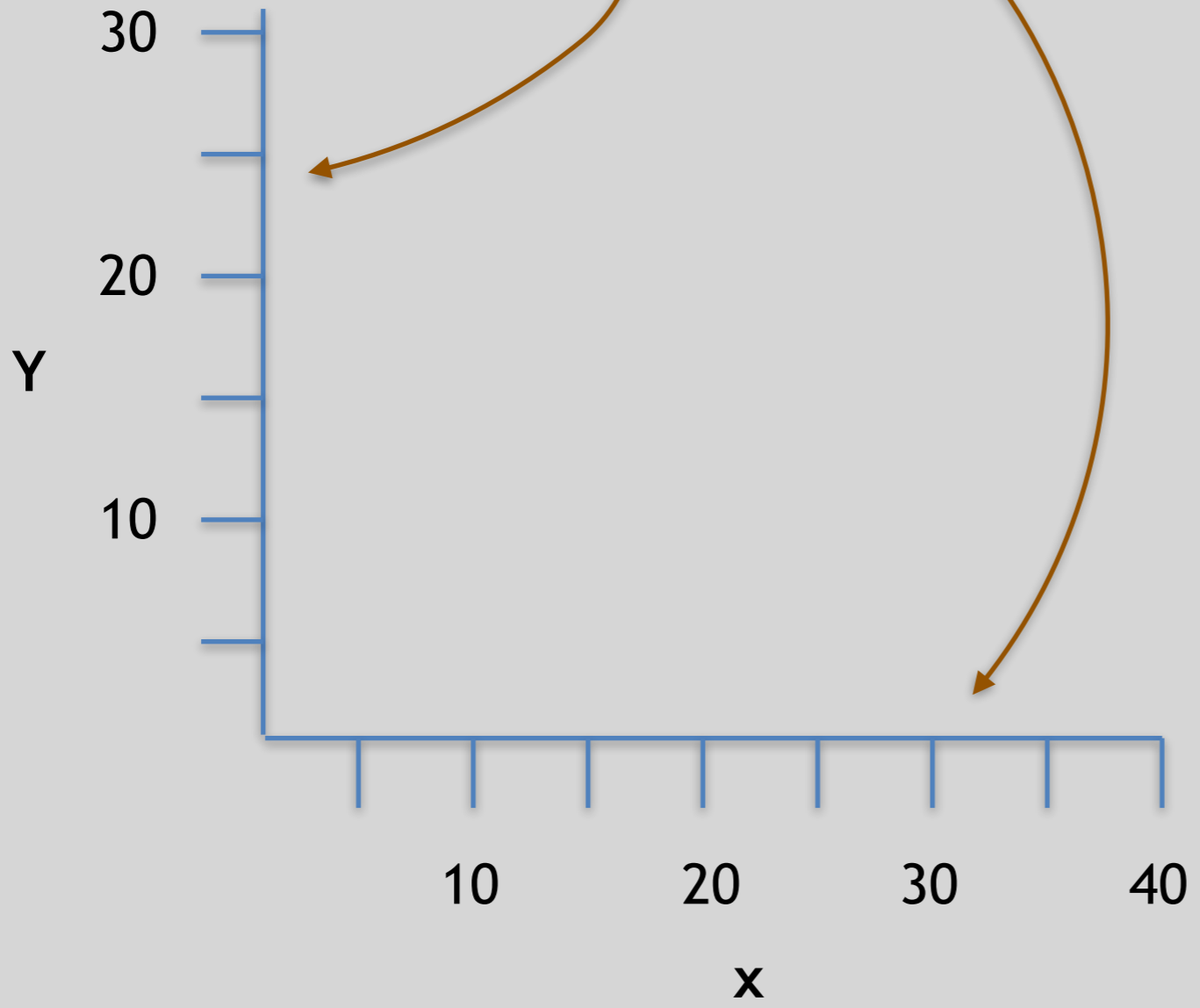
Gene	WT-1	WT-2	WT-3
x	30	5	13
y	24	10	18

Just replace the gene names with “x” and “y” and plot!

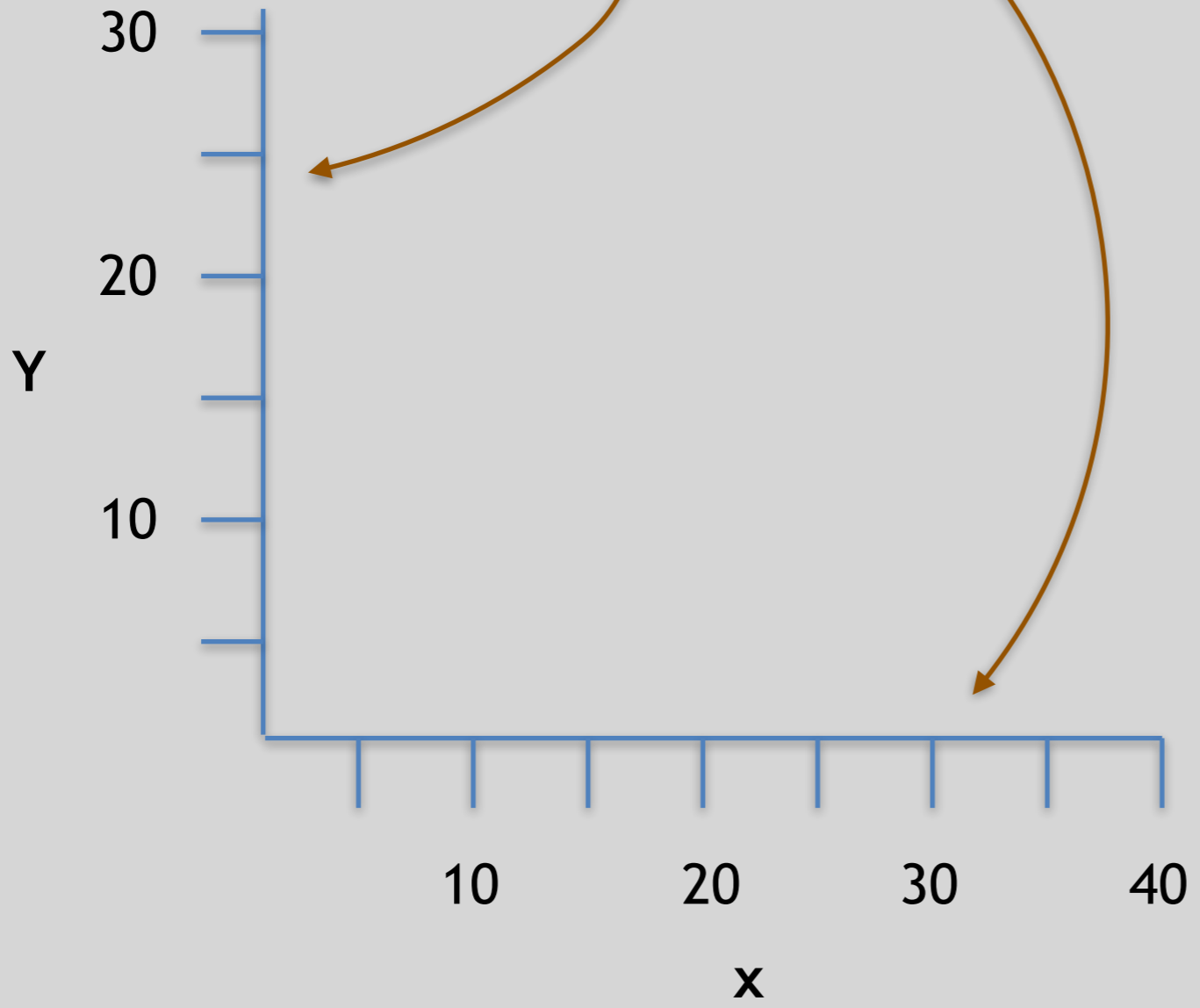
	sample-1	sample-2	sample-3
x	30	5	13
y	24	10	18



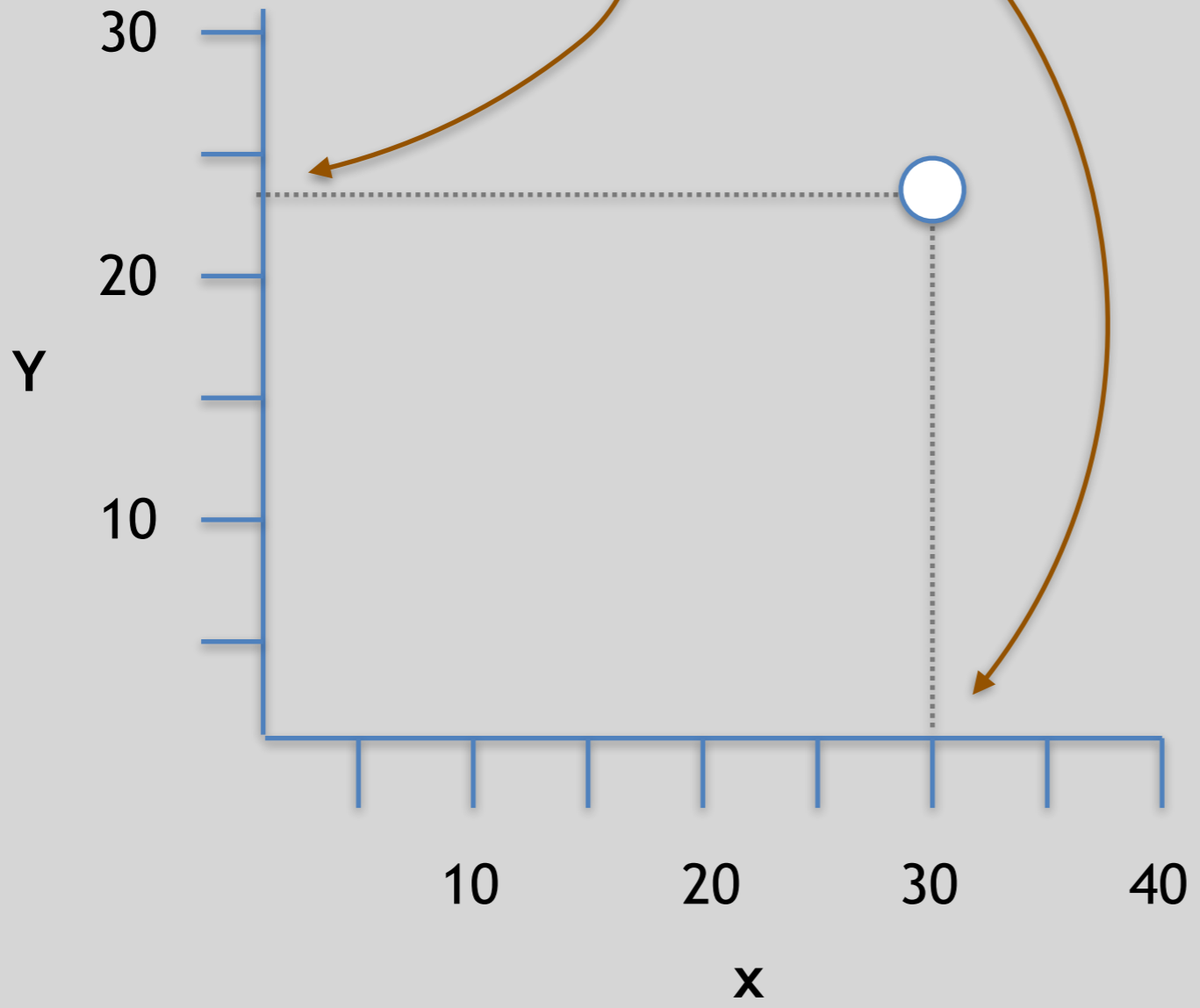
	sample-1	sample-2	sample-3
x	30	5	13
y	24	10	18



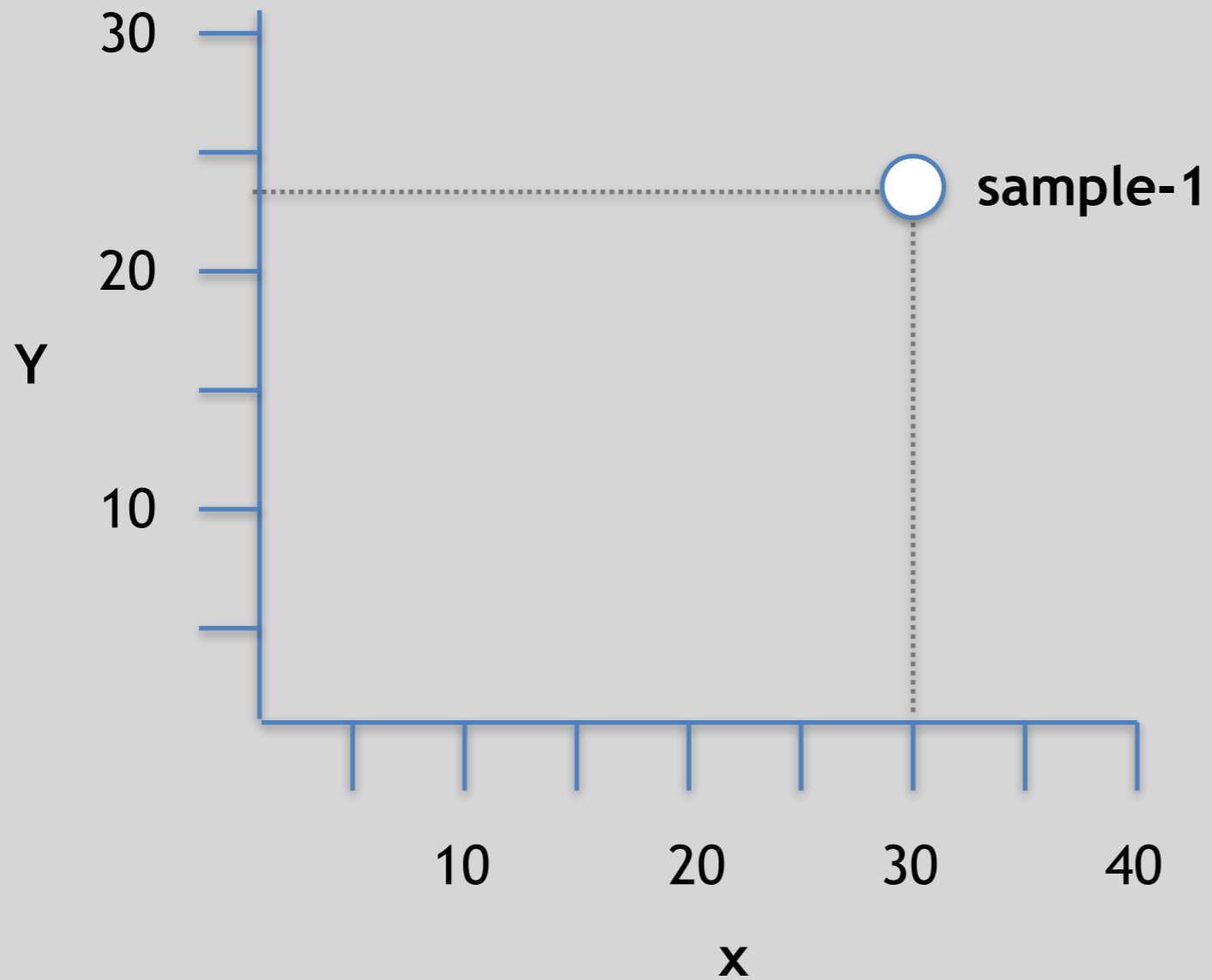
	sample-1	sample-2	sample-3
x	30	5	13
y	24	10	18



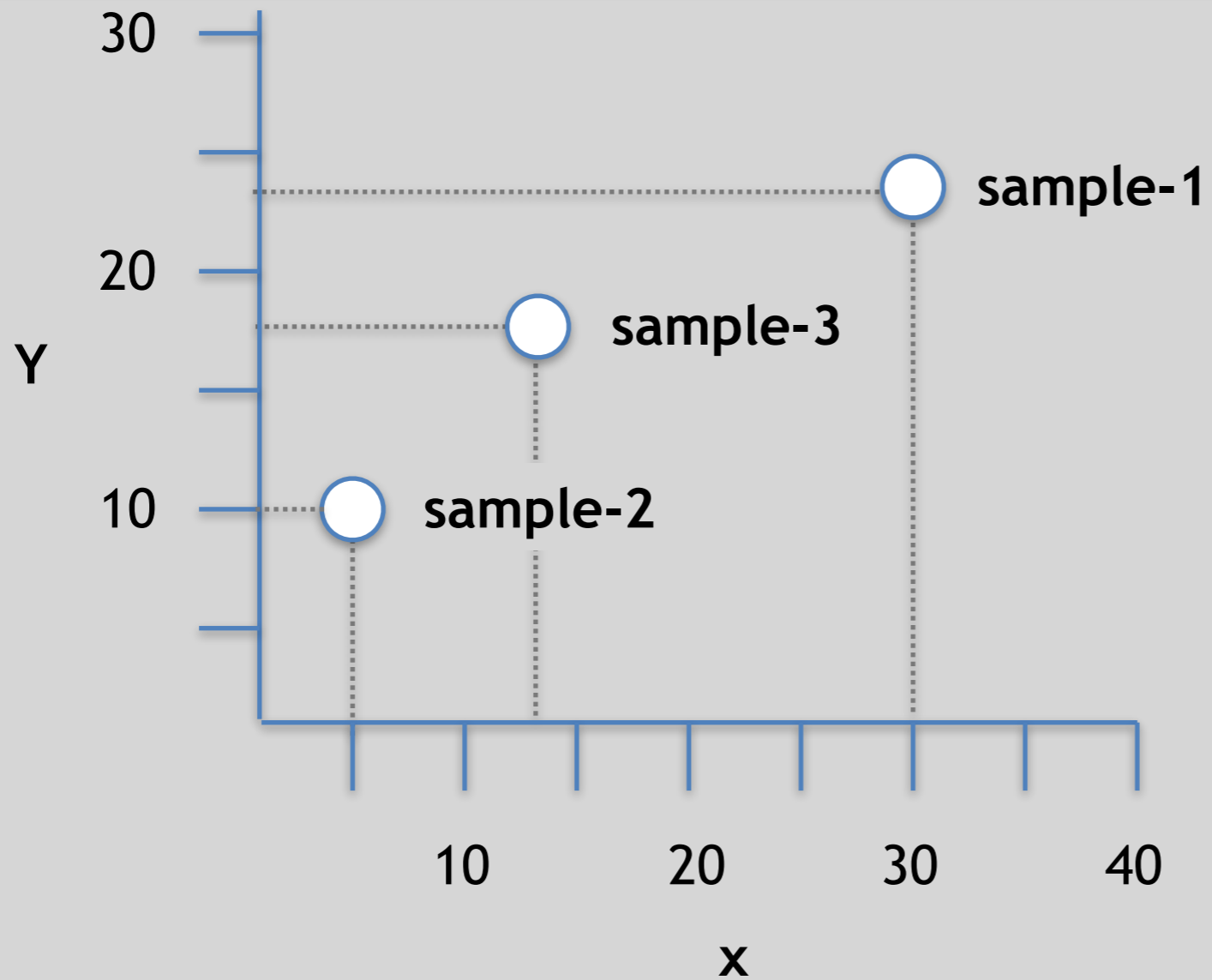
	sample-1	sample-2	sample-3
x	30	5	13
y	24	10	18



	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



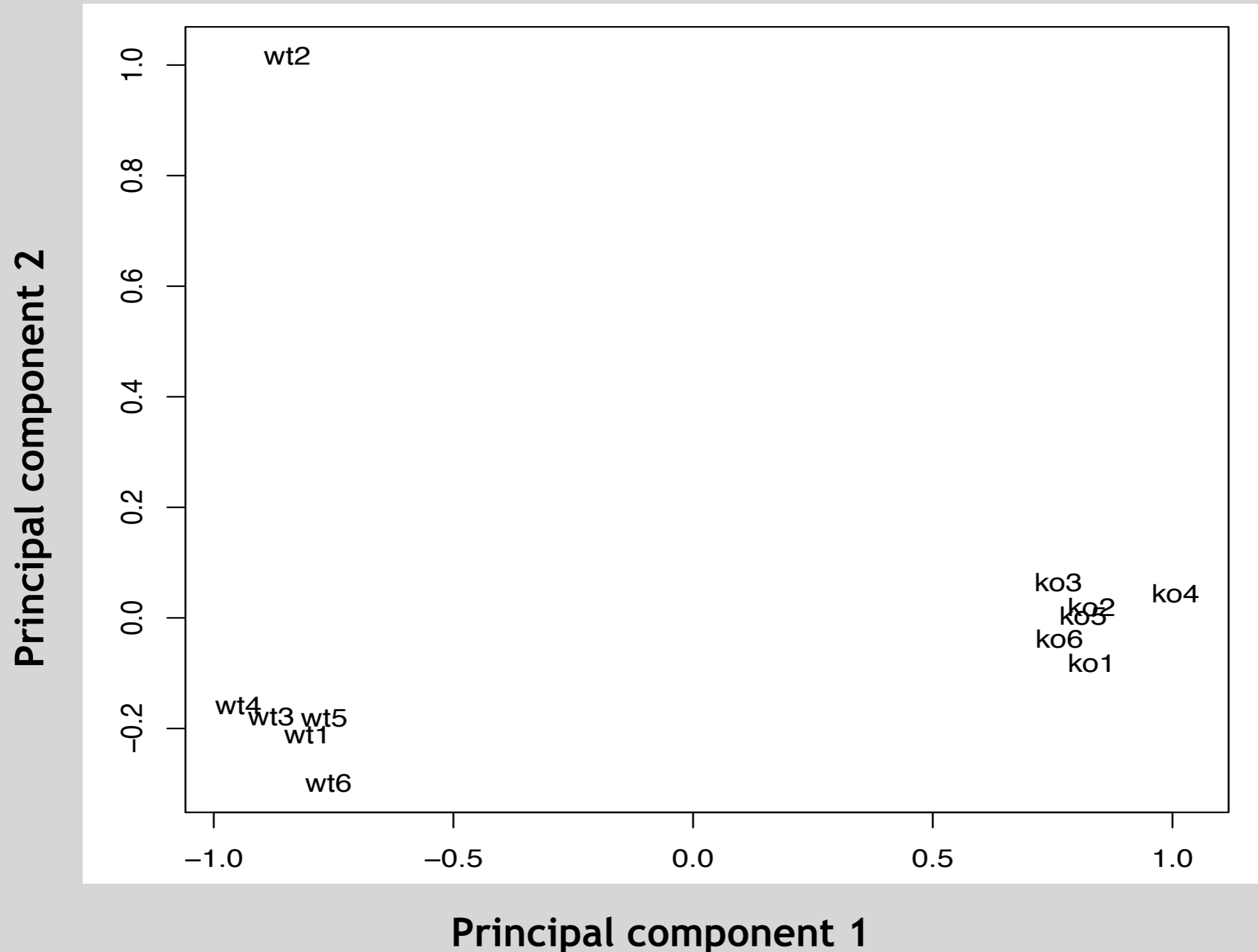
But we have 20,000 genes...

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

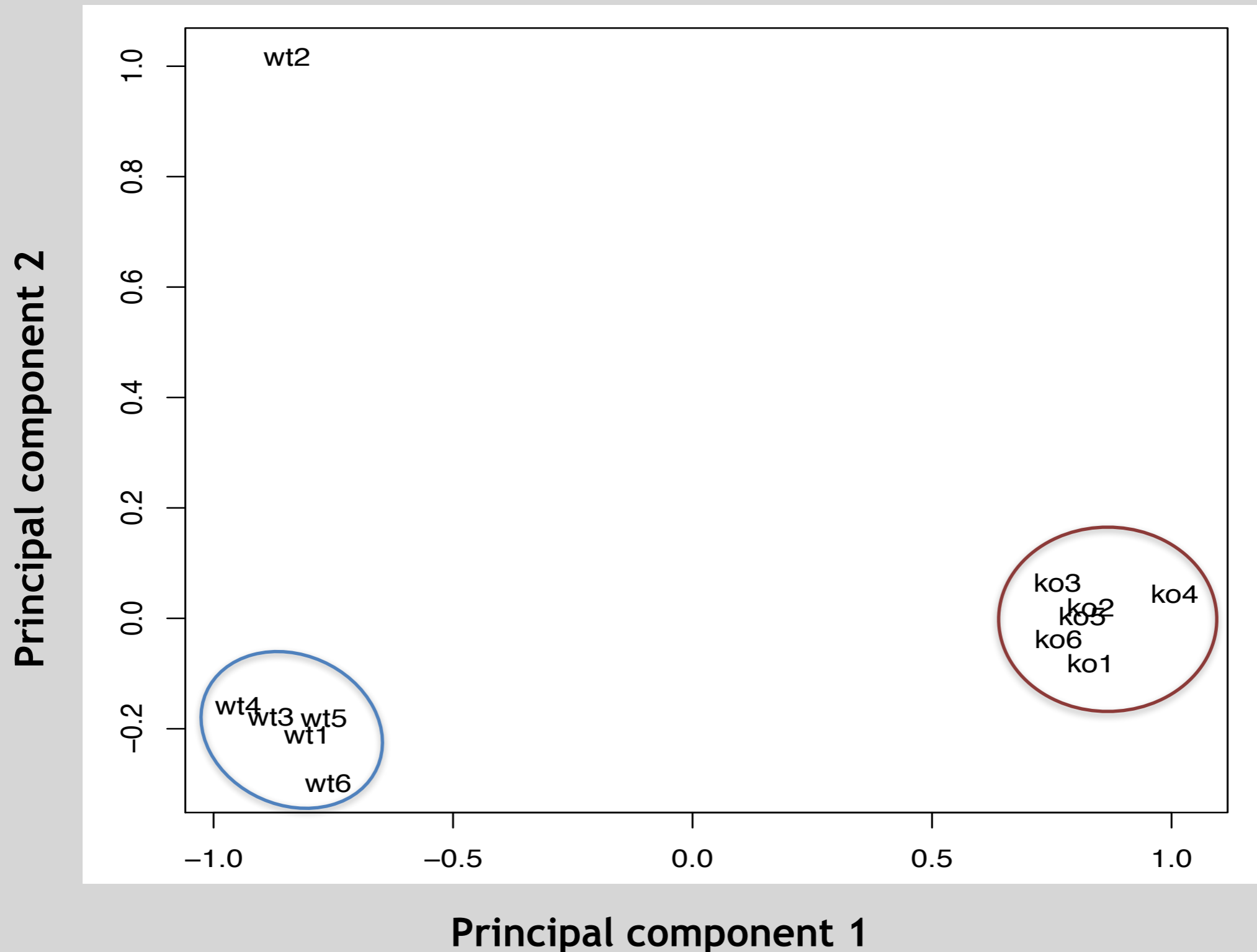
So we use PCA (principal component analysis) or something like it to plot this data.

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.

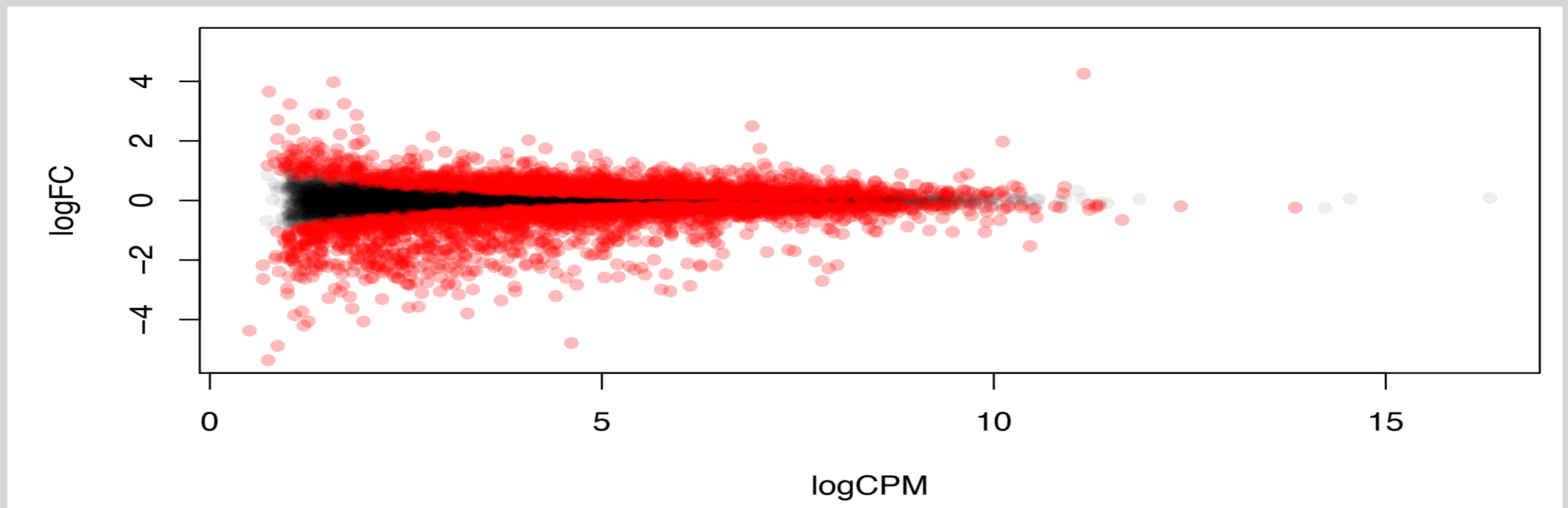


Plotting the data:

- (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.

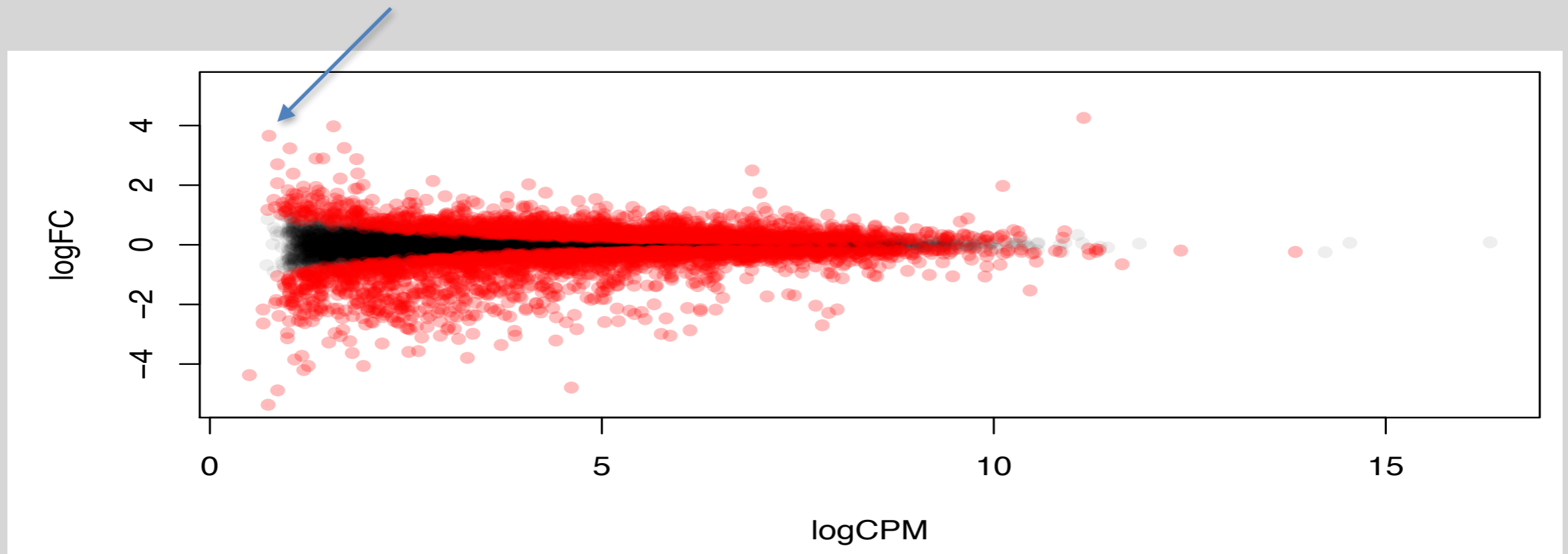
Step 2: Identify differentially expressed genes between the “normal” and “mutant” samples

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

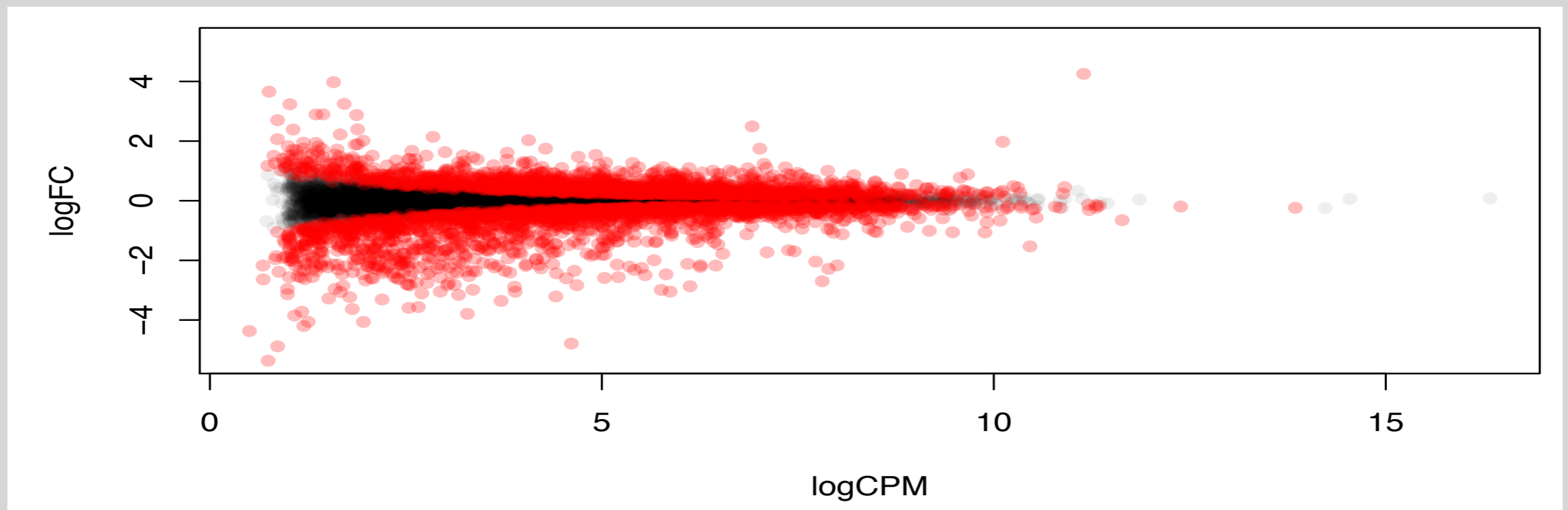


Step 2: Identify differentially expressed genes between the “normal” and “mutant” samples

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



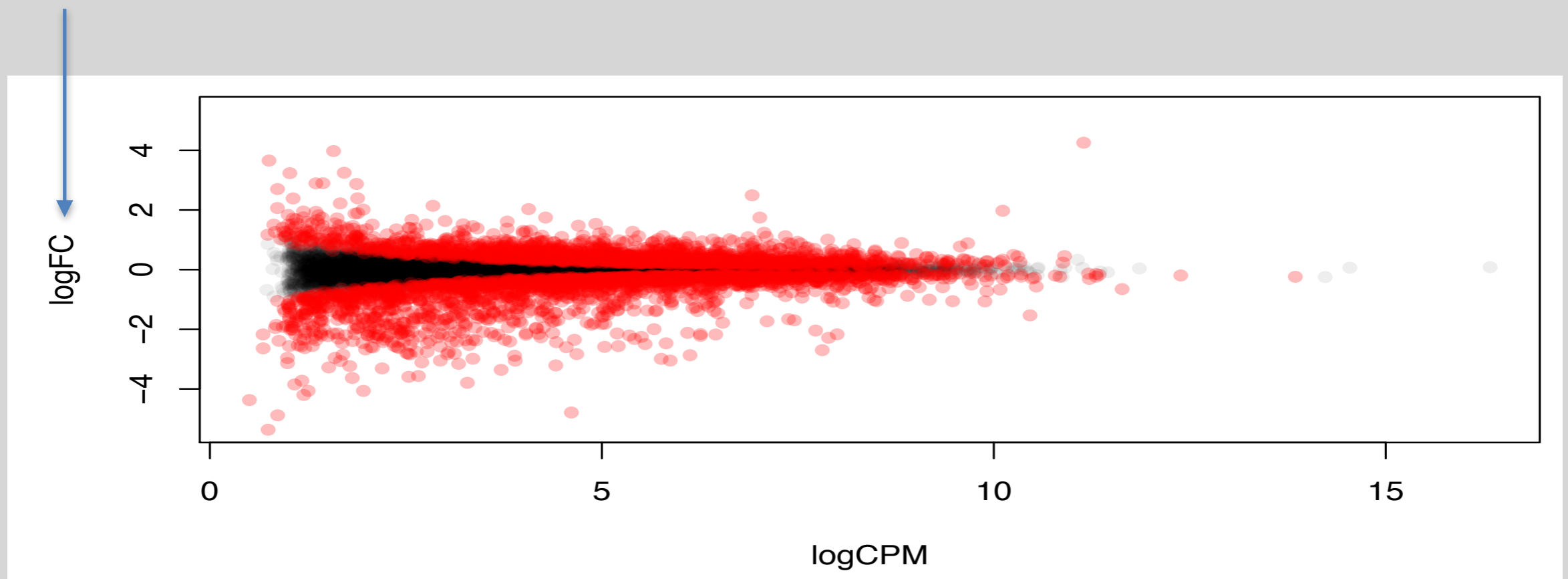
Step 2: Identify differentially expressed genes between the “normal” and “mutant” samples



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

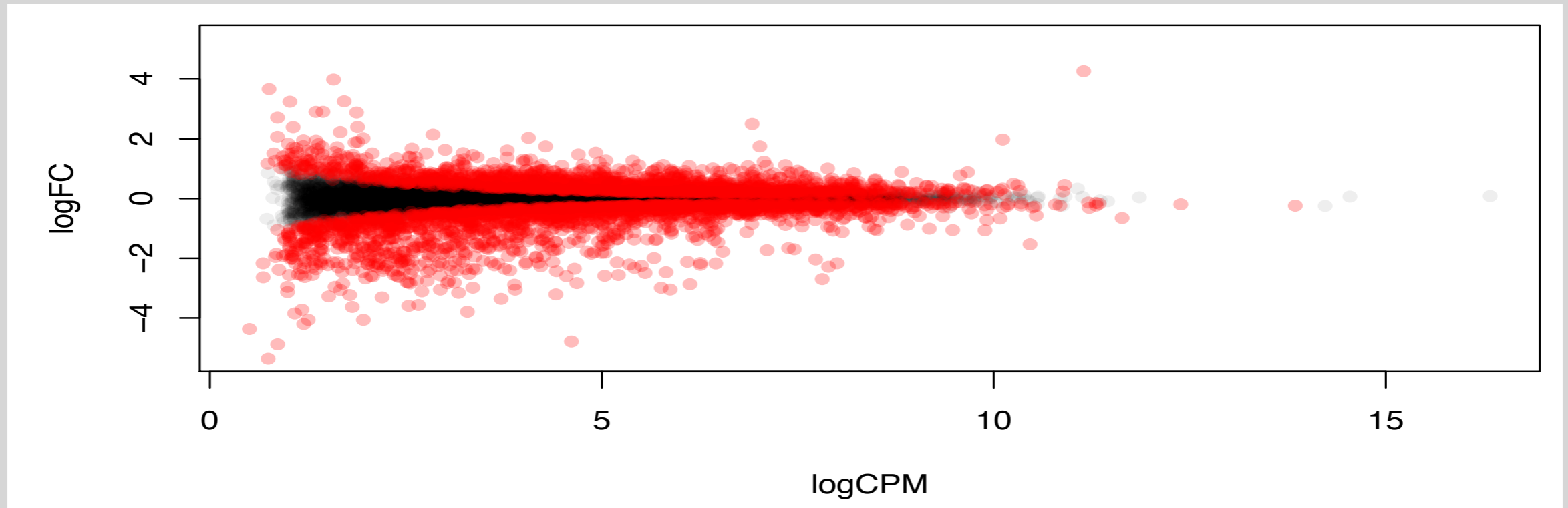
Step 2: Identify differentially expressed genes between the “normal” and “mutant” samples

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>

NCBI Resources How To

remiba@ncbi My NCBI Sign Out

SRA

ERA

Advanced

Search

Help

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 Helicoscope®, Complete Genomics®, and Pacific Biosciences SMRT®.

Getting Started

- [Understanding and Using SRA](#)
- [How to Submit](#)
- [Learn 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 <ul style="list-style-type: none">NCBI EducationNCBI Help ManualNCBI HandbookTraining & Tutorials	RESOURCES <ul style="list-style-type: none">Chemicals & BiologicsData & SoftwareDNA & RNADomains & StructuresGenes & ExpressionGenetics & MedicineGenomes & MapsHistologyLipidsProteinsSequence AnalysisTaxonomyTraining & TutorialsVariation	POPULAR <ul style="list-style-type: none">PubMedBioRxivPubMed CentralPubMed HealthBLASTNucleotideGenomeSNPGeneProteinPubChem	FEATURED <ul style="list-style-type: none">Genetic Testing RegistryPubMed HealthGenBankReference SequencesGene Expression OmnibusMap ViewerHuman GenomeMammal GenomeInfluenza VirusPrimer-BLASTSequence Read Archive	NCBI INFORMATION <ul style="list-style-type: none">About NCBIResearch at NCBINCBI NewsNCBI FTP SiteNCBI on FacebookNCBI on TwitterNCBI on YouTube
---	---	---	---	--

Copyright | Disclaimer | Privacy | Browsers | Accessibility | Contact

National Center for Biotechnology Information, U.S. National Library of Medicine
8900 Rockville Pike, Bethesda MD, 20894 USA

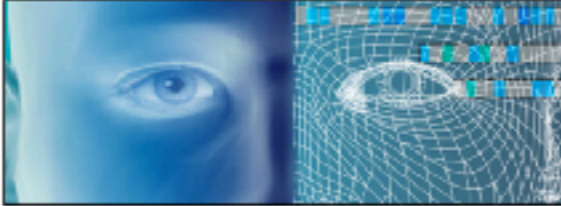
Protected Data - dbGaP

NCBI Database of Genotypes and Phenotypes (dbGaP):

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

NCBI Resources How To remilla@ncbi My NCBI Sign Out

dbGaP dbGaP Search Limits Advanced Help



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 Interactor](#)

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 [collection's page](#).

Study	Embargo Release	Details	Participants	Type Of Study	Links	Platform
phs00739.v1.p1 Comparative Analysis of Primary and Metastatic Colorectal Cancer	Version 1: 2015-01-29	V D A B	4	Cohort	Links	HiSeq 2500
phs00848.v1.p1 Autosomal recessive TTP2 mutations cause a new human immunodeficiency	Version 1: 2015-12-16	V D A B	3	Case-Control	Links	Genome Analyzer IX
phs00842.v1.p1 PodGF2	Version 1: passed embargo	V D A B	1572	Multicenter, Prospective, Observational, Cohort	Links	HumanOmni2.5-Quad
phs00807.v25.p9 Framingham Cohort	Version 1-25: passed embargo Version 23: 2016-04-26 Version 24: 2016-09-28 Version 25: 2016-12-03	V D A B	16173	Longitudinal	Links	HiGeneFocused50K_Affy Mapping50K_Map Mapping50K_Sy Mapping50K_HiSeq40 Mapping50K_Iba240
phs00825.v1.p1 Whole Genome Sequencing of HUE953 and HUE854	Version 1: passed embargo	V D A B	2	Control Set	Links	HiSeq 2500 HiSeq 2500

[List Top Level Studies](#)

You are here: NCBI > Genetics & Medicine > Database of Genotypes and Phenotypes (dbGaP) [Write to the Help Desk](#)

Summary

Course Logistics

Website, ethics, assessment and grading procedure.

Learning Objectives

What you need to learn to succeed in this course.

Course Structure

Major class topics and student group presentations.

Human Genome Review

What is a genome? What does the genome do? How is the genome decoded? How do we examine differences and disease mutants?