BGGN 213

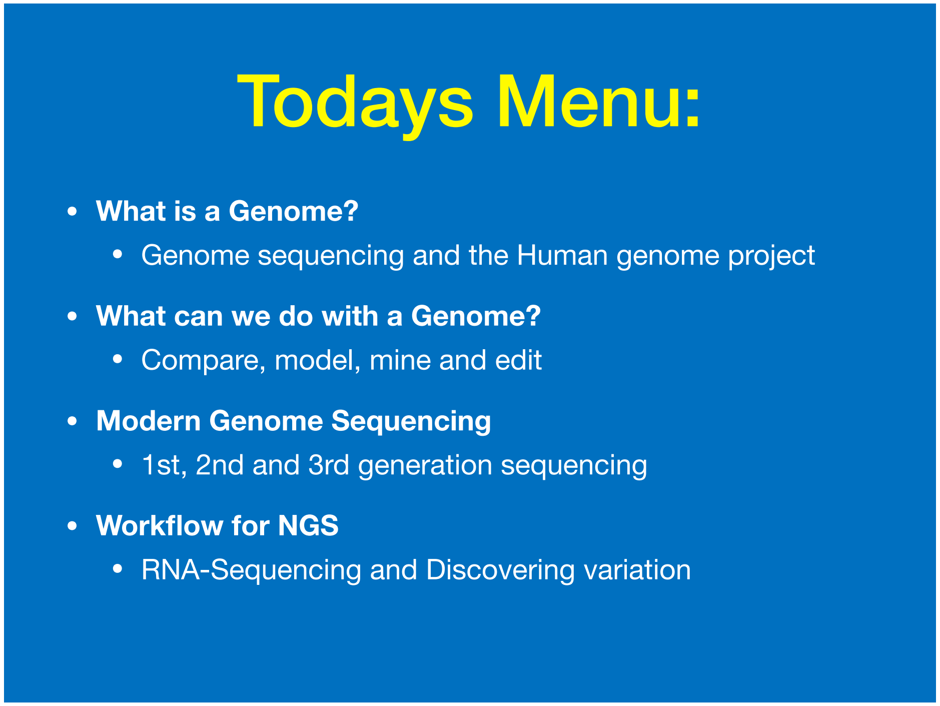
Genome Informatics I

Lecture 14

Barry Grant

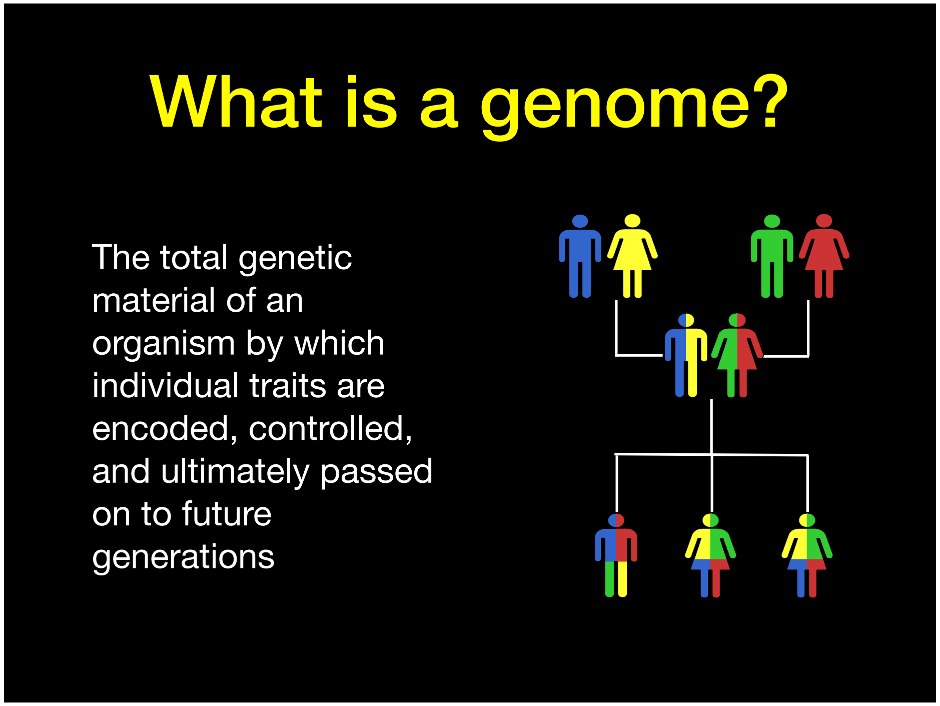
UC San Diego

http://thegrantlab.org/bggn213



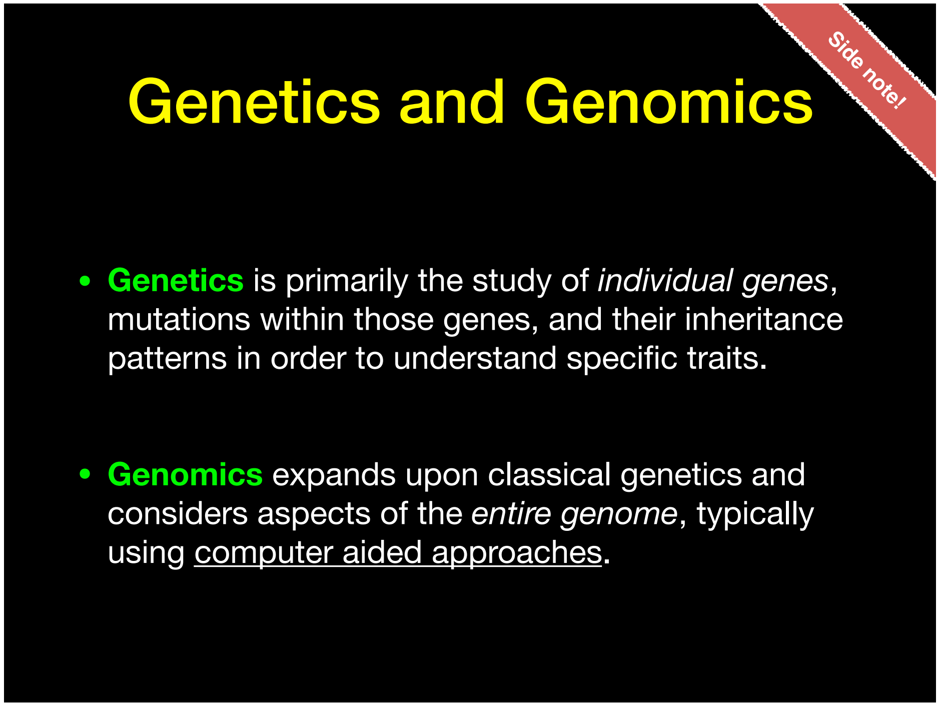
Todays Menu:

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



What is a genome?

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



Side note!

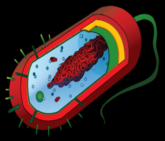
Genetics and Genomics

- **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.

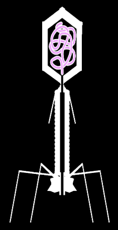
Genomes come in many shapes

Side note!

- Primarily DNA, but can be RNA in the case of some viruses
- Some genomes are circular, others linear
- Can be organized into discrete units (chromosomes) or freestanding molecules (plasmids)

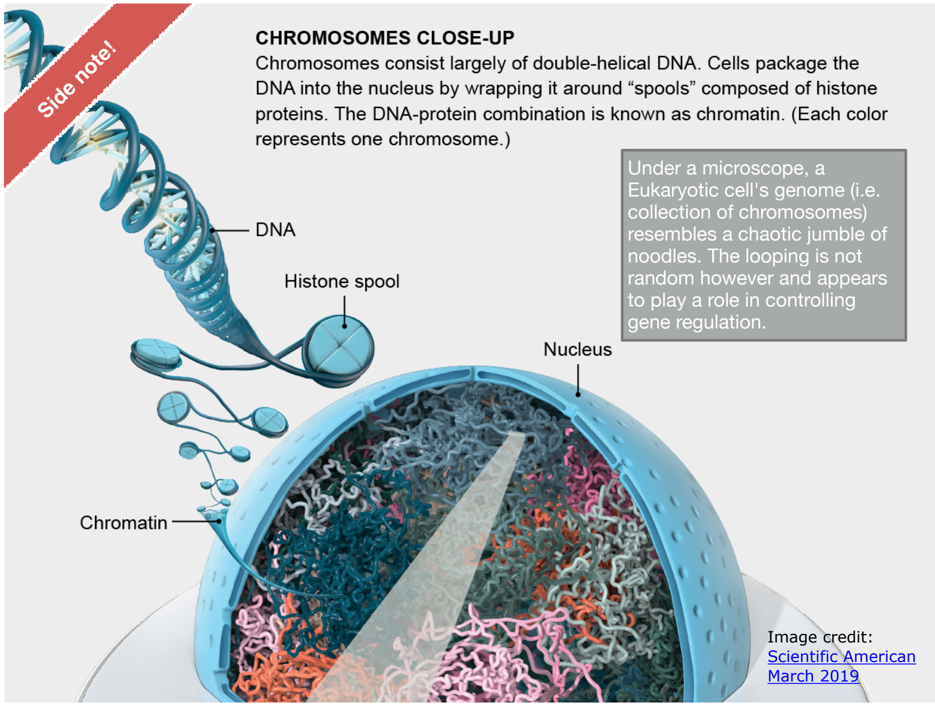


Prokaryote



Bacteriophage

Eukaryote



Side note!

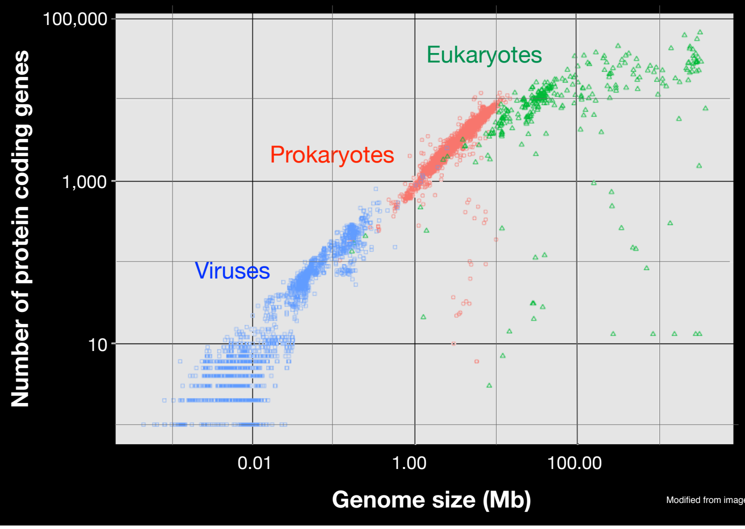
CHROMOSOMES CLOSE-UP

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

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

Image credit: Scientific American March 2019

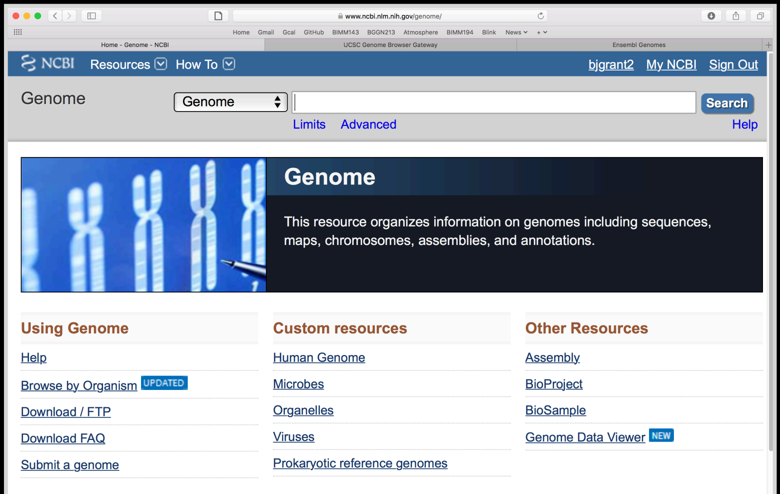
Genomes come in many sizes



Modified from image by Estevezj / CC BY-SA

Genome Databases

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



www.ncbi.nlm.nih.gov/genome/
Home Gmail Gcal GitHub BIMM143 BGGN213 Atmosphere BIMM194 Blink News +
Home - Genome - NCBI
UCSC Genome Browser Gateway
Ensembl Genomes
NCBI
Resources
How To
bjgrant2
My NCBI
Sign Out
Search
Limits
Advanced
Help

Genome

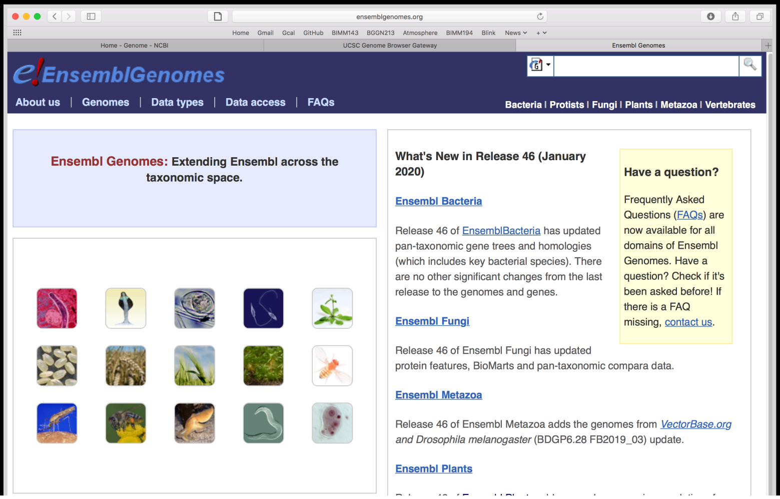
Genome

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

Using Genome	Custom resources	Other Resources
Help	Human Genome	Assembly
Browse by Organism UPDATED	Microbes	BioProject
Download / FTP	Organelles	BioSample
Download FAQ	Viruses	Genome Data Viewer NEW
Submit a genome	Prokaryotic reference genomes	

Genome Databases

(EBI) Ensemble Genomes:
http://ensemblgenomes.org


ensemblgenomes.org
Home
Gmail
Gcal
GitHub
BIMM143
BGGN213
Atmosphere
BIMM194
Blink
News
Home - Genome - NCBI
UCSC Genome Browser Gateway
Ensembl Genomes
e!EnsemblGenomes
About us | Genomes | Data types | Data access | FAQs
Bacteria | Protists | Fungi | Plants | Metazoa | Vertebrates

Ensembl Genomes: Extending Ensembl across the taxonomic space.

What's New in Release 46 (January 2020)

Ensembl Bacteria

Release 46 of EnsemblBacteria has updated pan-taxonomic gene trees and homologies (which includes key bacterial species). There are no other significant changes from the last release to the genomes and genes.

Ensembl Fungi

Release 46 of Ensembl Fungi has updated protein features, BioMarts and pan-taxonomic compara data.

Ensembl Metazoa

Release 46 of Ensembl Metazoa adds the genomes from *VectorBase.org and Drosophila melanogaster* (BDGP6.28 FB2019_03) update.

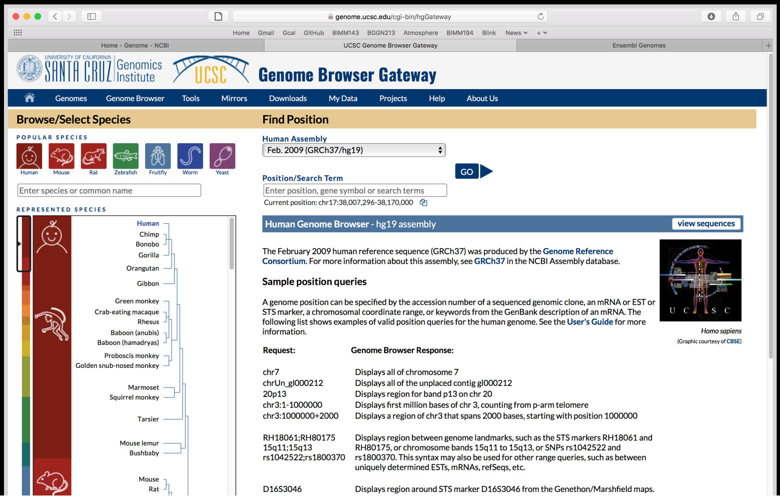
Ensembl Plants

Have a question?

Frequently Asked Questions (FAQs) are now available for all domains of Ensembl Genomes. Have a question? Check if it's been asked before! If there is a FAQ missing, contact us.

Genome Databases

UCSC Genome Browser Gateway:
https://genome.ucsc.edu/



Browse/Select Species

Find Position

Human Assembly
Feb. 2009 (GRCh37/hg19)

Position/Search Term
Current position: chr17:38,007,296-38,170,000

Human Genome Browser - hg19 assembly

The February 2009 human reference sequence (GRCh37) was produced by the Genome Reference Consortium. For more information about this assembly, see GRCh37 in the NCBI Assembly database.

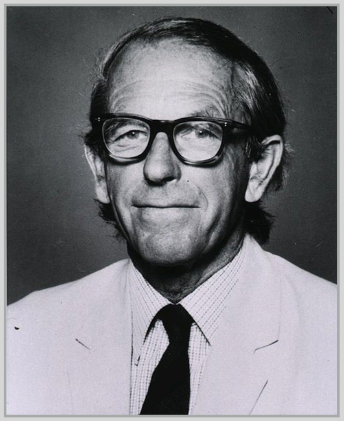
Homo sapiens
(Graphic courtesy of CBSE)

Sample position queries

A genome position can be specified by the accession number of a sequenced genomic clone, an mRNA or EST or STS marker, a chromosomal coordinate range, or keywords from the GenBank description of an mRNA. The following list shows examples of valid position queries for the human genome. See the User's Guide for more information.

Request:	Genome Browser Response:
chr7	Displays all of chromosome 7
chrUn_gl000212	Displays all of the unplaced contig gl000212
20p13	Displays region for band p13 on chr 20
chr3:1-1000000	Displays first million bases of chr 3, counting from p-arm telomere
chr3:1000000+2000	Displays a region of chr3 that spans 2000 bases, starting with position 1000000
RH18061;RH80175 15q11;15q13 rs1042522;rs1800370	Displays region between genome landmarks, such as the STS markers RH18061 and RH80175, or chromosome bands 15q11 to 15q13, or SNPs rs1042522 and rs1800370. This syntax may also be used for other range queries, such as between uniquely determined ESTs, mRNAs, refSeqs, etc.
D16S3046	Displays region around STS marker D16S3046 from the Genethon/Marshfield maps.

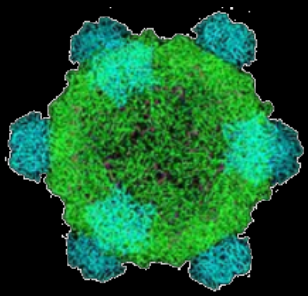
Early Genome Sequencing



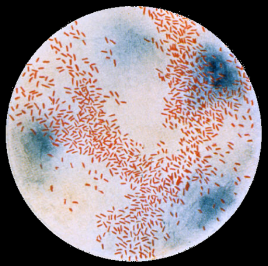
http://en.wikipedia.org/wiki/Frederick_Sanger

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

The First Sequenced Genomes



Bacteriophage φ-X174
- Completed in 1977
- 5,386 base pairs, ssDNA
- 11 genes



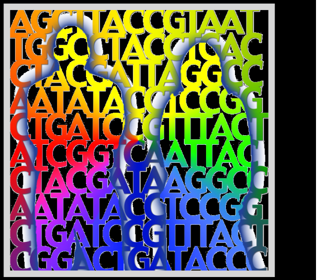
Haemophilus influenzae
- Completed in 1995
- 1,830,140 base pairs, dsDNA
- 1,740 genes

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
  - ~20,400 coding (& ~24,000 non-coding) genes*

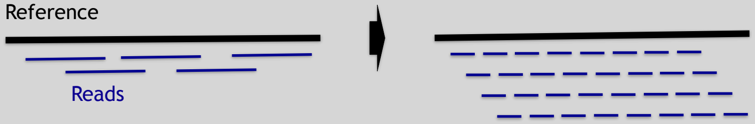


*Latest numbers < link >

DeCode Genetics INC.

Modern Genome Sequencing

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



Rapid progress of genome sequencing

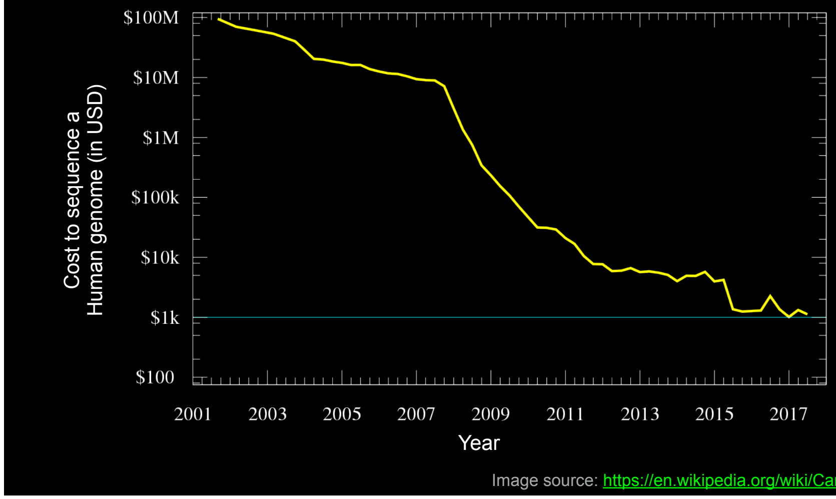


Image source: https://en.wikipedia.org/wiki/Carlson_curve

Rapid progress of genome sequencing

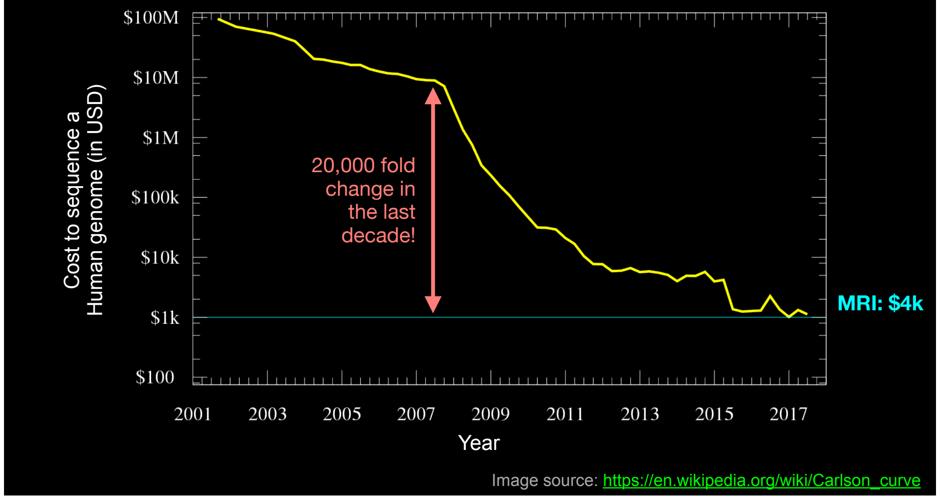


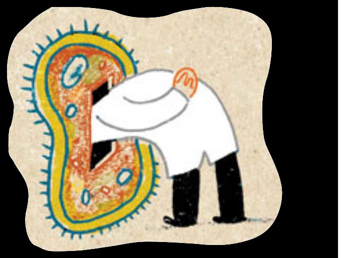
Image source: https://en.wikipedia.org/wiki/Carlson_curve

Major impact areas for genomic medicine

- Cancer: Identification of driver mutations and drugable variants, Molecular stratification to guide and monitor treatment, Identification of tumor specific variants for personalized immunotherapy approaches (precision medicine).
- Genetic disease diagnose: Rare, inherited and so-called 'mystery' disease diagnose.
- Health management: Predisposition testing for complex diseases (e.g. cardiac disease, diabetes and others), optimization and avoidance of adverse drug reactions.
- Health data analytics: Incorporating genomic data with additional health data for improved healthcare delivery.
- Prenatal testing, transplant rejection, pathogen detection, microbiome etc.

Goals of Cancer Genome Research

- Identify changes in the genomes of tumors that drive cancer progression
- Identify new targets for therapy
- Select drugs based on the genomics of the tumor
- Provide early cancer detection and treatment response monitoring
- Utilize cancer specific mutations to derive neoantigen immunotherapy approaches



What can go wrong in cancer genomes?

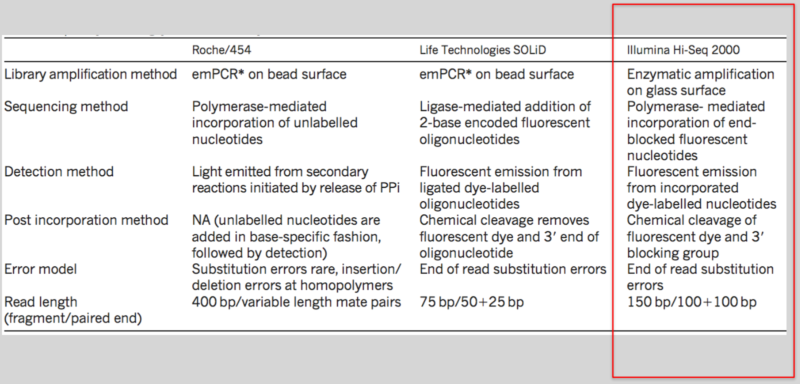
Type of change	Some common technology to study changes
DNA mutations	WGS, WXS
DNA structural variations	WGS
Copy number variation (CNV)	CGH array, SNP array, WGS
DNA methylation	Methylation array, RRBS, WGBS
mRNA expression changes	mRNA expression array, RNA-seq
miRNA expression changes	miRNA expression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

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

DNA Sequencing Concepts

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

Modern NGS Sequencing Platforms

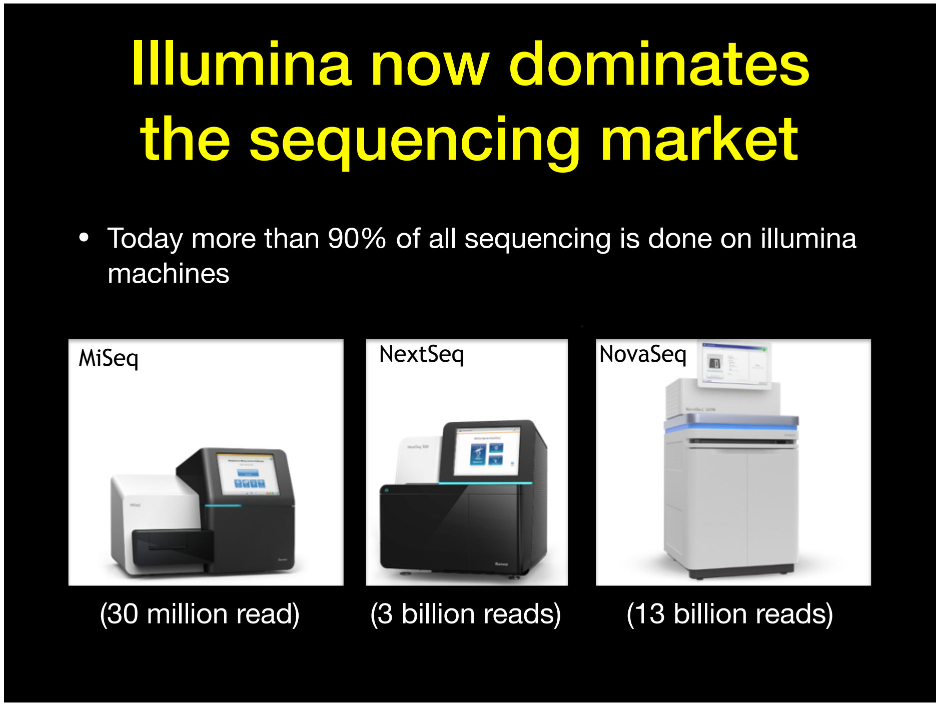


	Roche/454	Life Technologies SOLiD	Illumina Hi-Seq 2000
Library amplification method	emPCR* on bead surface	emPCR* on bead surface	Enzymatic amplification on glass surface
Sequencing method	Polymerase-mediated incorporation of unlabelled nucleotides	Ligase-mediated addition of 2-base encoded fluorescent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

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

Illumina now dominates the sequencing market

- Today more than 90% of all sequencing is done on illumina machines
- Generating millions to billions of reads per run (machine dependent)
- High fidelity (>99.9% accuracy for short ~300 bp reads)
- $1,000 per human genome in 48 hours*



Illumina now dominates the sequencing market

- Today more than 90% of all sequencing is done on illumina machines

MiSeq

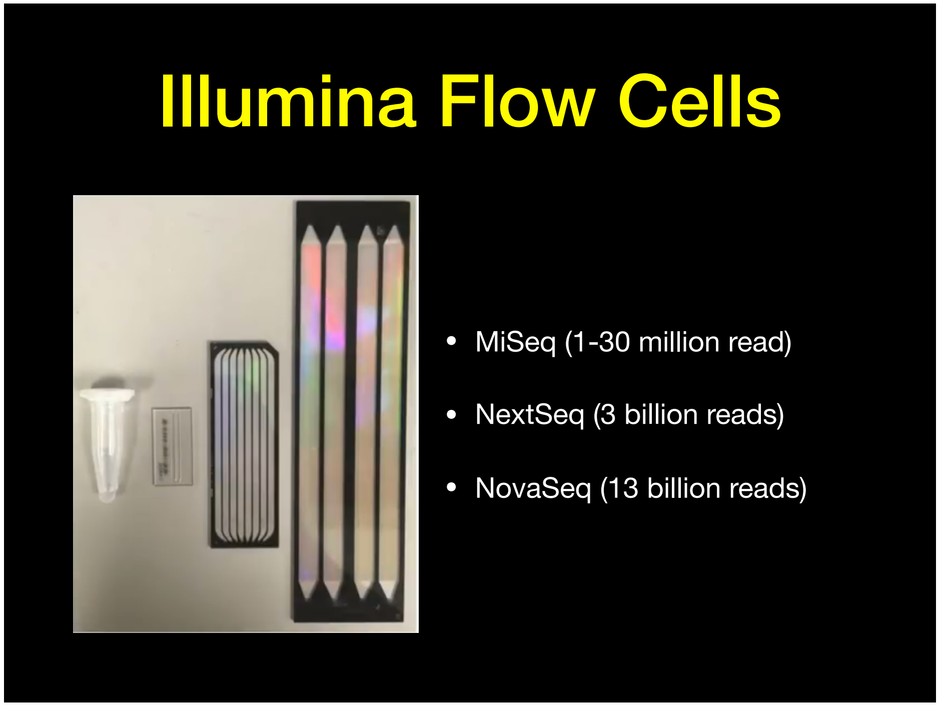
(30 million read)

NextSeq

(3 billion reads)

NovaSeq

(13 billion reads)

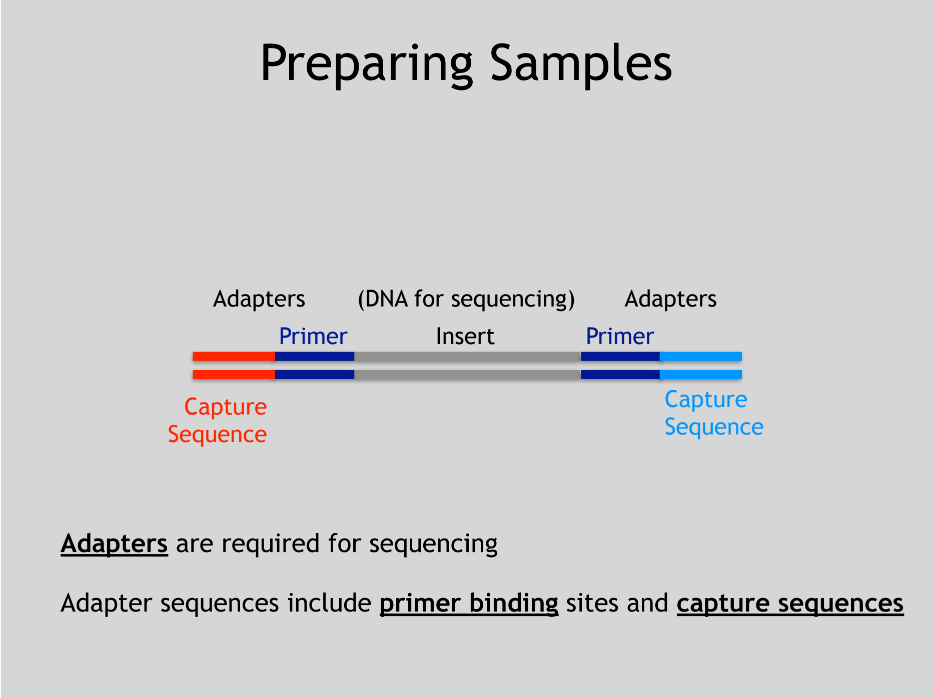


Illumina Flow Cells

- MiSeq (1-30 million read)
- NextSeq (3 billion reads)
- NovaSeq (13 billion reads)

Preparing Samples

(DNA for sequencing)
Insert

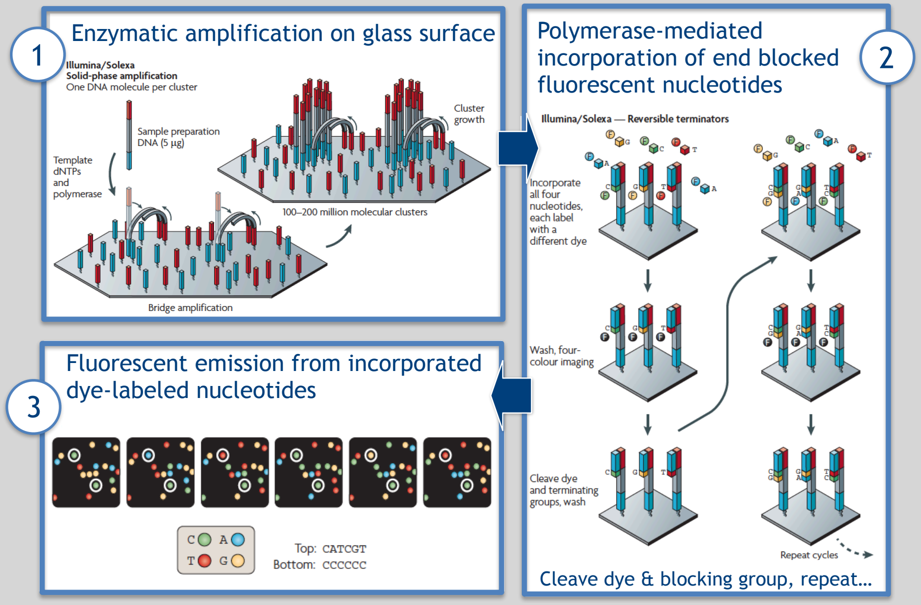


Preparing Samples

Adapters are required for sequencing

Adapter sequences include **primer binding** sites and **capture sequences**

Illumina – Reversible terminators



1 Enzymatic amplification on glass surface

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

2 Polymerase-mediated incorporation of end blocked fluorescent nucleotides

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

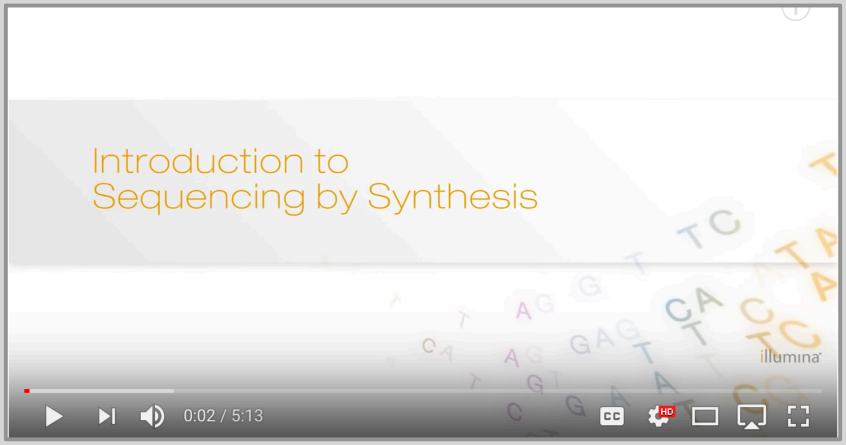
Cleave dye & blocking group, repeat...

3 Fluorescent emission from incorporated dye-labeled nucleotides

C
A
T
G
Top: CATCGT
Bottom: CCCCCC

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

Illumina Sequencing - Video

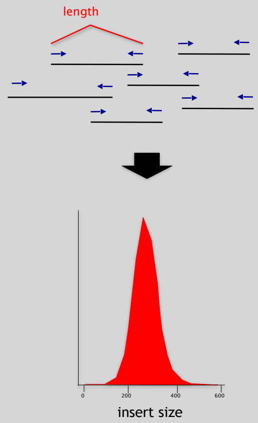


Introduction to Sequencing by Synthesis

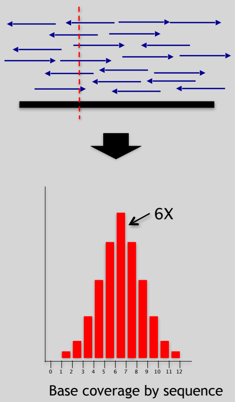
https://www.youtube.com/watch?src_vid=womKfikWlxM&v=fCd6B5HRaZ8

NGS Sequencing Terminology

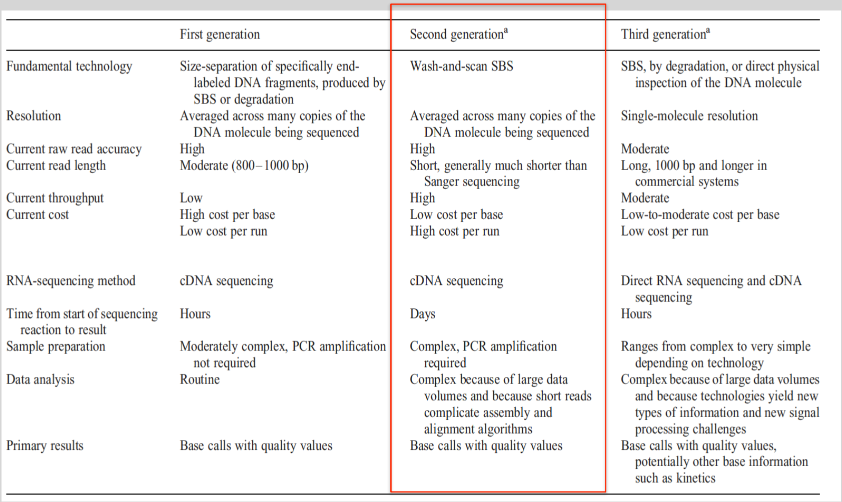
Insert Size



Sequence Coverage



Terminology: "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 sequence reads (1,000bp+)
  - Single molecule (no PCR amplification step required)
  - Often associated with "nanopore technology" (e.g. *Oxford Nanopore's* MinION USB sequencer)
  - Note that other approaches are being developed...

The first direct RNA sequencing by nanopore

Side-Note:

- For example this new nanopore direct RNA-sequencing method was published last year:
  https://www.nature.com/articles/nmeth.4577
- *"Sequencing the RNA in a biological sample can unlock a wealth of information, including the identity of bacteria and viruses, the nuances of alternative splicing or the transcriptional state of organisms. However, current methods have limitations due to short read lengths and reverse transcription or amplification biases. Here we demonstrate nanopore direct RNA-seq, a highly parallel, real-time, single-molecule method that circumvents reverse transcription or amplification steps.”*

What can we do with all this sequence information?

Population Scale Analysis

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

1000 Genomes
A Deep Catalog of Human Genetic Variation
GENOME 10K
Unveiling animal diversity
The Cancer Genome Atlas
Understanding genomics to improve cancer care
The 100,000 Genomes Project
Genomics England & Partners

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

"Variety's the very spice of life"

–William Cowper, 1785

"Variation is the spice of life"

–Kruglyak & Nickerson, 2001

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

Types of Genomic Variation

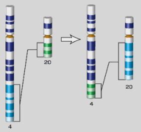
- **Single Nucleotide Polymorphisms** (SNPs) – mutations of one nucleotide to another

```
AATCTGAGGCAT
AATCTCAGGCAT
```

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

```
AATCTGAAGGCAT
AATCT--AGGCAT
```

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



Darryl Leja, Courtesy: National Human Genome Research Institute.

Differences Between Individuals

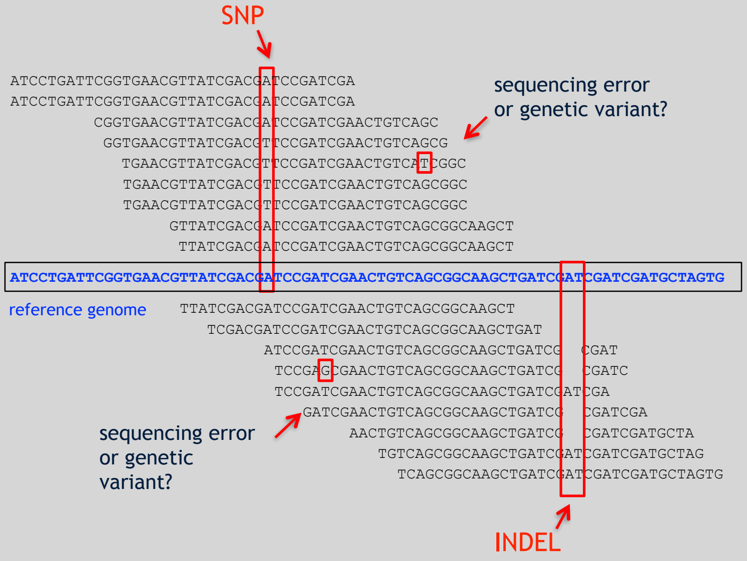
The average number of genetic differences in the germline between two random humans can be broken down as follows:

- 3,600,000 single nucleotide differences
- 344,000 small insertion and deletions
- 1,000 larger deletion and duplications

Numbers change depending on ancestry!

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

Discovering Variation: SNPs and INDELs



SNP

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

reference genome

sequencing error or genetic variant?

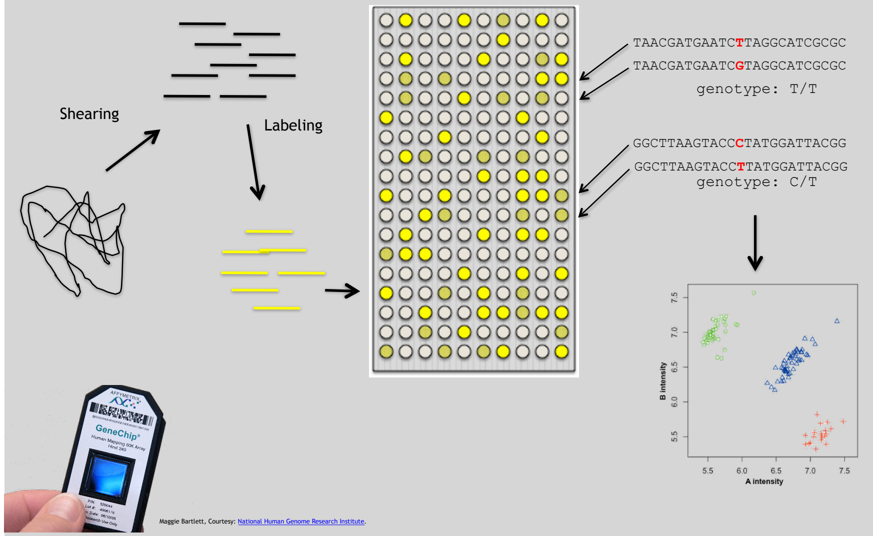
sequencing error or genetic variant?

INDEL

Genotyping Small Variants

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

SNP Microarrays



Shearing

Labeling

TAACGATGAATC**T**TAGGCATCGCGC

TAACGATGAATC**G**TAGGCATCGCGC

genotype: T/T

GGCTTAAGTACC**C**TATGGATTACGG

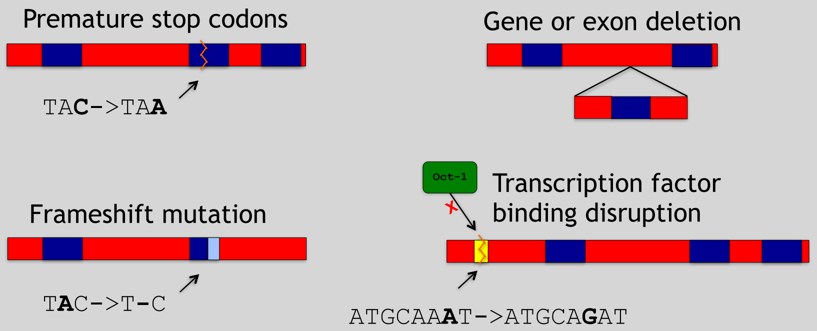
GGCTTAAGTACC**T**TATGGATTACGG

genotype: C/T

Maggie Bartlett, Courtesy: National Human Genome Research Institute.

Impact of Genetic Variation

There are numerous ways genetic variation can exhibit functional effects



Premature stop codons

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

Gene or exon deletion

Frameshift mutation

TAC->T-C

Transcription factor binding disruption

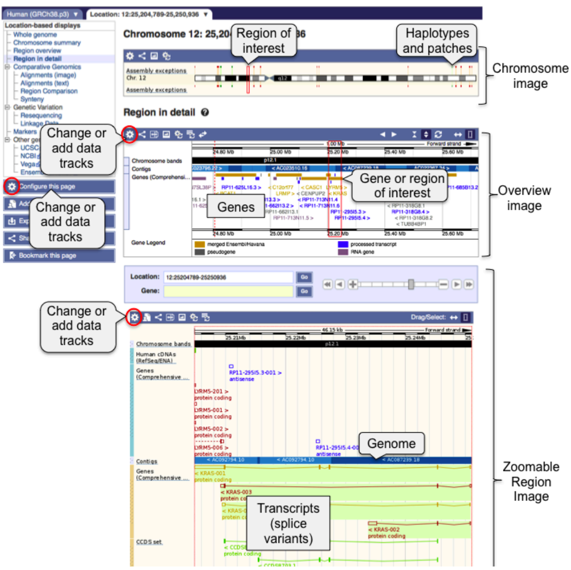
ATGCAA**A**T->ATGCA**G**AT

Do it Yourself!

Hand-on time!

Sections **1** to **3** please (up to running Read Alignment)

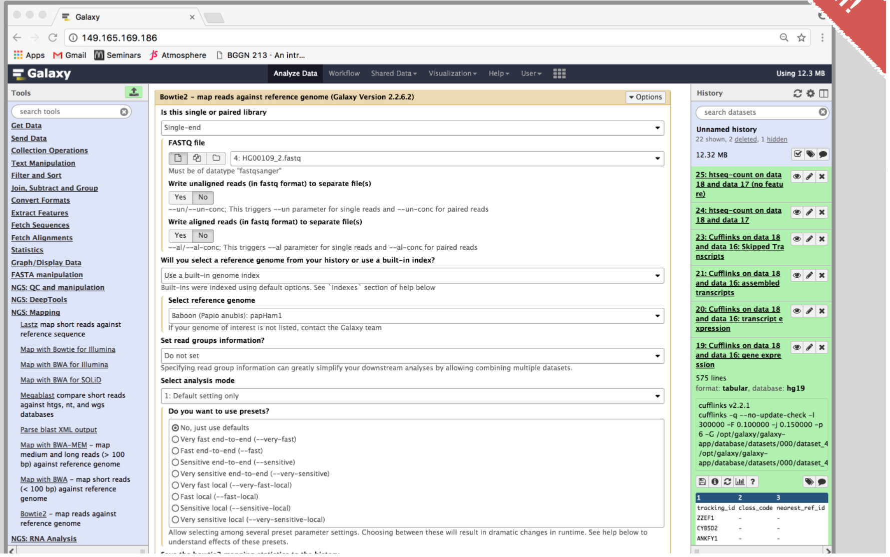
See IP address on website for **your** Galaxy server http://uswest.ensembl.org/Help/View?id=140



Access a jetstream galaxy instance!

Use assigned IP address

Do it Yourself!


Galaxy

149.165.169.186

Apps Gmail Seminars Atmosphere BGGN 213 · An intr...

Galaxy | Analyze Data | Workflow | Shared Data | Visualization | Help | User | Using 12.3 MB

Tools

search tools

Get Data
Send Data
Collection Operations
Text Manipulation
Filter and Sort
Join, Subtract and Group
Convert Formats
Extract Features
Fetch Sequences
Fetch Alignments
Statistics
Graph/Display Data
FASTA manipulation
NGS: QC and manipulation
NGS: DeepTools
NGS: Mapping
- Lastz map short reads against reference sequence
- Map with Bowtie for Illumina
- Map with BWA for Illumina
- Map with BWA for SOLiD
- Megablast compare short reads against htgs, nt, and wgs databases
- Parse blast XML output
- Map with BWA-MEM – map medium and long reads (> 100 bp) against reference genome
- Map with BWA – map short reads (< 100 bp) against reference genome
- Bowtie2 – map reads against reference genome

NGS: RNA Analysis

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

Is this single or paired library
Single-end

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

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

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

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

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

Set read groups information?
Do not set
Specifying read group information can greatly simplify your downstream analyses by allowing combining multiple datasets.

Select analysis mode
1: Default setting only

Do you want to use presets?
● No, just use defaults
○ Very fast end-to-end (--very-fast)
○ Fast end-to-end (--fast)
○ Sensitive end-to-end (--sensitive)
○ Very sensitive end-to-end (--very-sensitive)
○ Very fast local (--very-fast-local)
○ Fast local (--fast-local)
○ Sensitive local (--sensitive-local)
○ Very sensitive local (--very-sensitive-local)

Allow selecting among several preset parameter settings. Choosing between these will result in dramatic changes in runtime. See help below to understand effects of these presets.

History

search datasets

Unnamed history
22 shown, 2 deleted, 1 hidden
12.32 MB

- 25: htseq-count on data 18 and data 17 (no feature)
- 24: htseq-count on data 18 and data 17
- 23: Cufflinks on data 18 and data 16: Skipped Transcripts
- 21: Cufflinks on data 18 and data 16: assembled transcripts
- 20: Cufflinks on data 18 and data 16: transcript expression
- 19: Cufflinks on data 18 and data 16: gene expression

575 lines
format: **tabular**, database: **hg19**

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

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

```
>Identifier1 (comment)
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
>Identifier2 (comment)
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XX
```

>Identifier1 (comment)
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
>Identifier2 (comment)
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XX

```
@Identifier1 (comment)
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
+
QQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQ
@Identifier2 (comment)
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
+
QQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQ
```

Raw data usually in **FASTQ format**

```
@NS500177:196:HFTTTAFXX:1:11101:10916:1458 2:N:0:CGCGGCTG ①
ACACGACGATGAGGTGACAGTCACGGAGGATAAGATCAATGCCCTCATTAAAGCAGCCGGTGTAA ②
+ ③
AAAAAEEEEEEEEEEE//AEEEAEEEEEEEEEEE/EE/<<EE/AAEEAEE///EEEEAEEEAEA< ④
```

Each sequencing "read" consists of 4 lines of data :

① The first line (which always starts with '@') is a unique ID for the sequence that follows

② The second line contains the bases called for the sequenced fragment

③ The third line is always a "+" character

④ The forth line contains the quality scores for each base in the sequenced fragment (these are ASCII encoded...)

ASCII Encoded Base Qualities

```
@NS500177:196:HFTTTAFXX:1:11101:10916:1458 2:N:0:CGCGGCTG
ACACGACGATGAGGTGACAGTCACGGAGGATAAGATCAATGCCCTCATTAAAGCAGCCGGTGTAA
+
AAAAAEEEEEEEEEEE//AEEEAEEEEEEEEEEE/EE/<<EE/AAEEAEE///EEEEAEEEAEA< ④
```

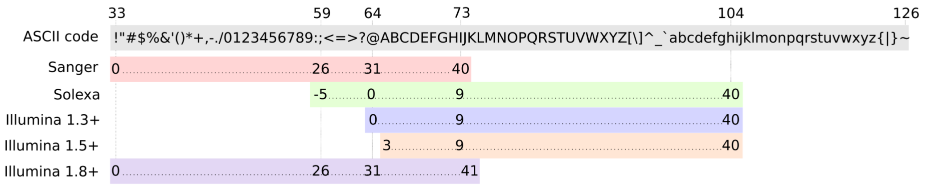
- Each sequence base has a corresponding numeric quality score encoded by a single ASCII character typically on the 4th line (see ④ above)
- ASCII characters represent integers between 0 and 127
- Printable ASCII characters range from 33 to 126
- Unfortunately there are 3 quality score formats that you may come across...

Interpreting Base Qualities in R

		ASCII Range	Offset	Score Range
Sanger, Illumina (Ver > 1.8)	**fastqsanger**	**33-126**	**33**	**0-93**
Solexa, Ilumina (Ver < 1.3)	fastqsolexa	59-126	64	5-62
Illumina (Ver 1.3 -1.7)	fastqillumina	64-126	64	0-62

```
> library(seqinr)
> library(gtools)
> phred <- asc( s2c("DDDDCDEDCDDDDBBDDDCC@") ) - 33
> phred
## D  D  D  D  C  D  E  D  C  D  D  D  D  B  B  D  D  D  C  C  @
## 35 35 35 35 34 35 36 35 34 35 35 35 35 33 33 35 35 35 34 34 31

> prob <- 10**(-phred/10)
```

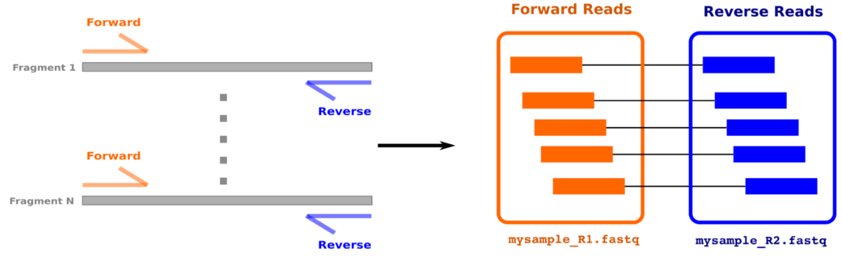
Interpreting Base Qualities in R



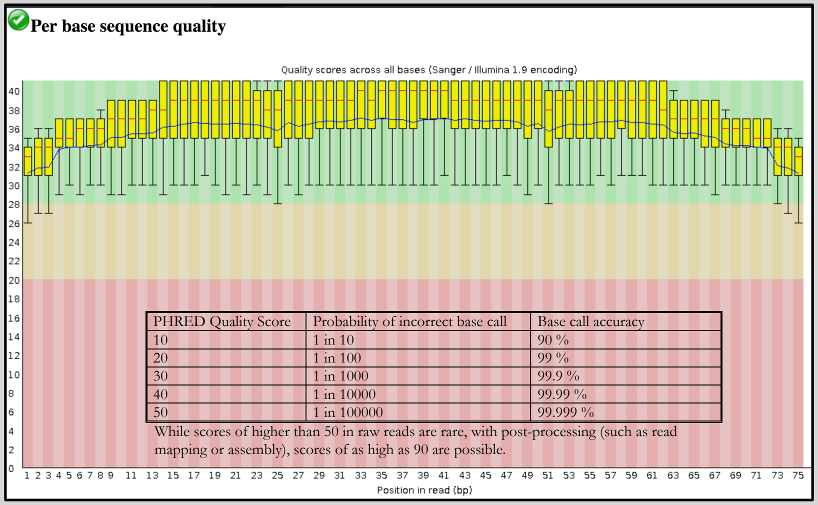
```
> library(seqinr)
> library(gtools)
> phred <- asc( s2c("DDDDCDEDCDDDDBBDDDCC@") ) - 33
> phred
## D  D  D  D  C  D  E  D  C  D  D  D  D  B  B  D  D  D  C  C  @
## 35 35 35 35 34 35 36 35 34 35 35 35 35 33 33 35 35 35 34 34 31

> prob <- 10**(-phred/10)
```

Paired-end FASTQ files

- Sequencer produces two FASTQ files:
  - **Forward** reads (usually **_1** or **_R1** in file name)
  - **Reverse** reads (usually **_2** or **_R2** in file name)



FastQC Report



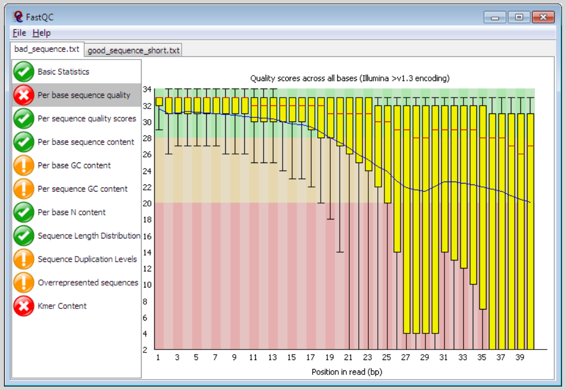
PHRED Quality Score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90 %
20	1 in 100	99 %
30	1 in 1000	99.9 %
40	1 in 10000	99.99 %
50	1 in 100000	99.999 %

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

FASTQC

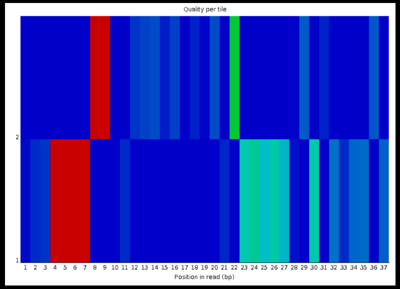
FASTQC is one approach which provides a visual interpretation of the raw sequence reads

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



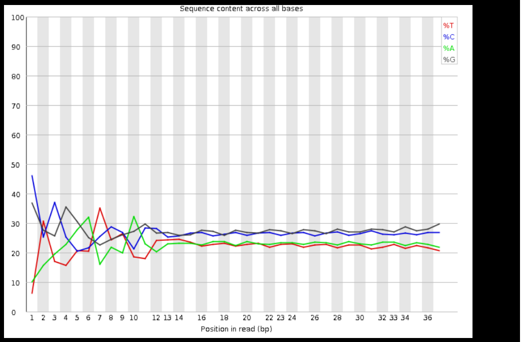
Per Tile Quality shows shows the deviation from the average quality for each tile

- In Illumina libraries the sequence identifier encodes the flowcell tile from which each read came.
- "Hot" colors indicate that a tile had worse quality reads than other tiles for that base
- Suggesting transient problems such as bubbles going through the flowcell, smudges or debris inside the flowcell lane.



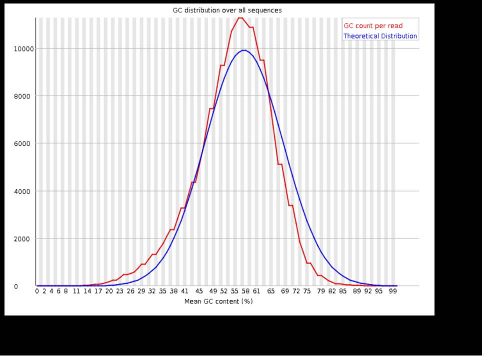
Per-base sequence content highlights the proportion of each base in each position

- In a random library there would be little to no difference between the different bases of a sequence run.
- Note that some types of libraries (e.g. RNA-Seq) will nearly always produce biased sequence composition at the start of the read.



GC content should follow a normal distribution

- An unusually shaped distribution could indicate a contaminated library or some other kinds of biased subset (frequent in metagenomic data sets).
- Sharp peaks on an otherwise smooth distribution are normally the result of a specific contaminant (e.g. adapter dimers)

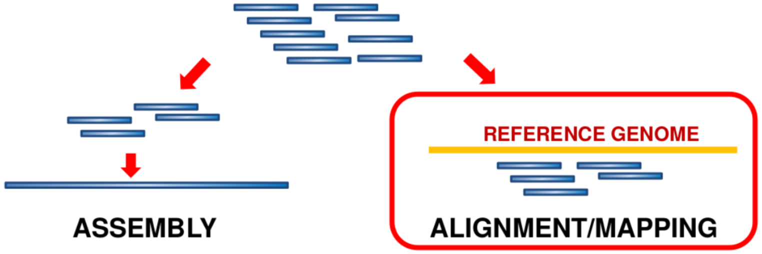


Increasing the quality of sequences

- **Filtering of sequences (i.e. removing sequences):**
  - with small mean quality score
  - with too many N bases
  - based on their GC content
- **Cutting/Trimming sequences from low quality score parts** (i.e the tails/ends of reads)
- Re-run your sequencing job

What is mapping?

- Short reads must be combined into longer fragments
- **Mapping:** use a reference genome as a guide
- **De-novo assembly:** without reference genome



Sequence Alignment

- Once sequence quality has been assessed, the next step is to **align/map** 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

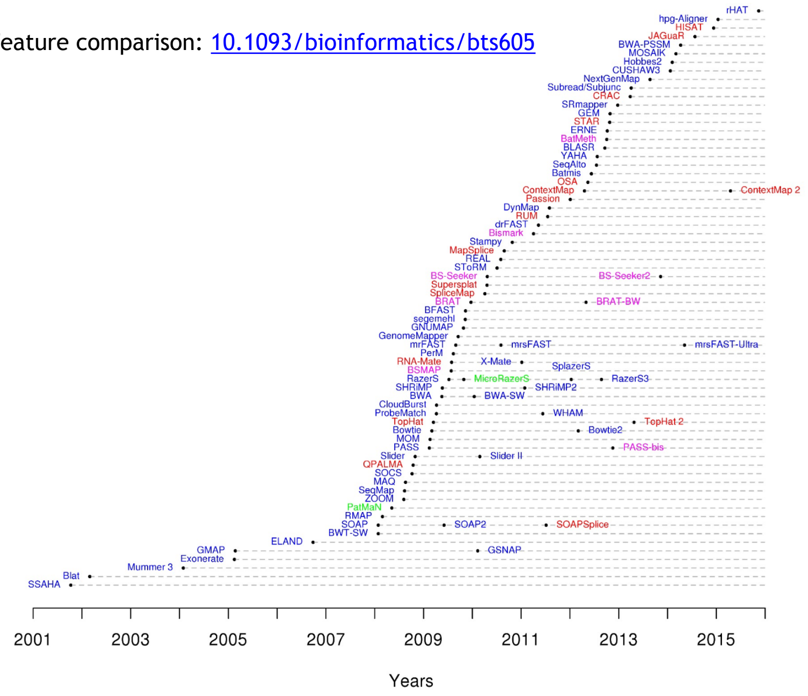
BWA
Bowtie2
SOAP2
Novoalign
mr/mrsFast
Eland
Blat
Bfast

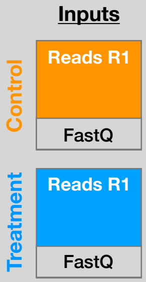
BarraCUDA
CASHx
GSNAP
Mosiak
Stampy
SHRiMP
SeqMap
SLIDER

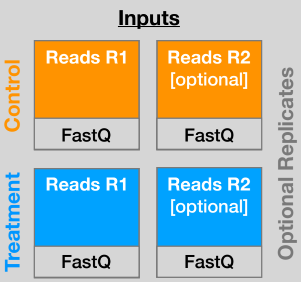
RMAP
SSAHA
etc

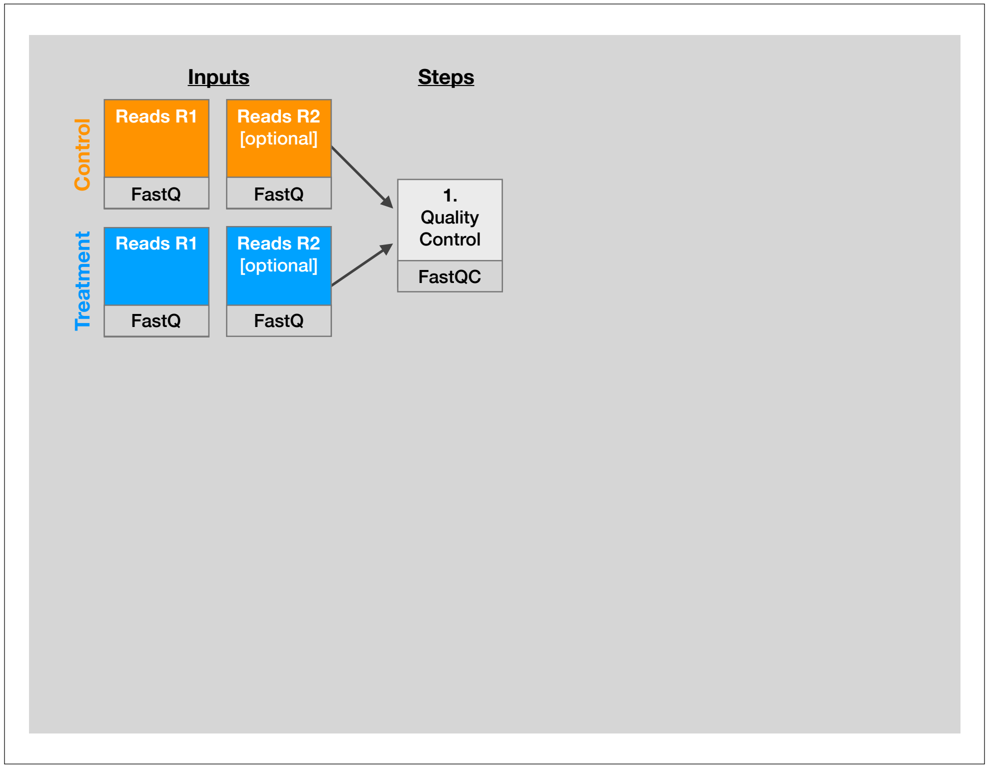
Feature comparison: 10.1093/bioinformatics/bts605

Feature comparison: 10.1093/bioinformatics/bts605







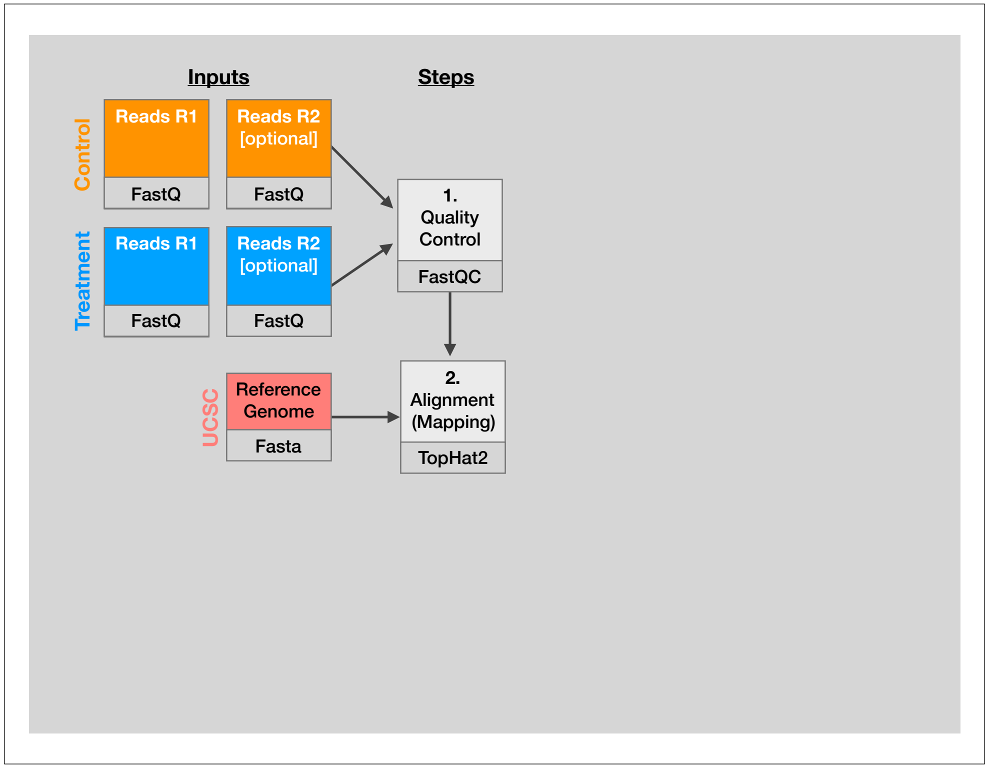


Inputs **Steps**

Control: Reads R1 — FastQ; Reads R2 [optional] — FastQ

Treatment: Reads R1 — FastQ; Reads R2 [optional] — FastQ

1. Quality Control — FastQC



Inputs **Steps**

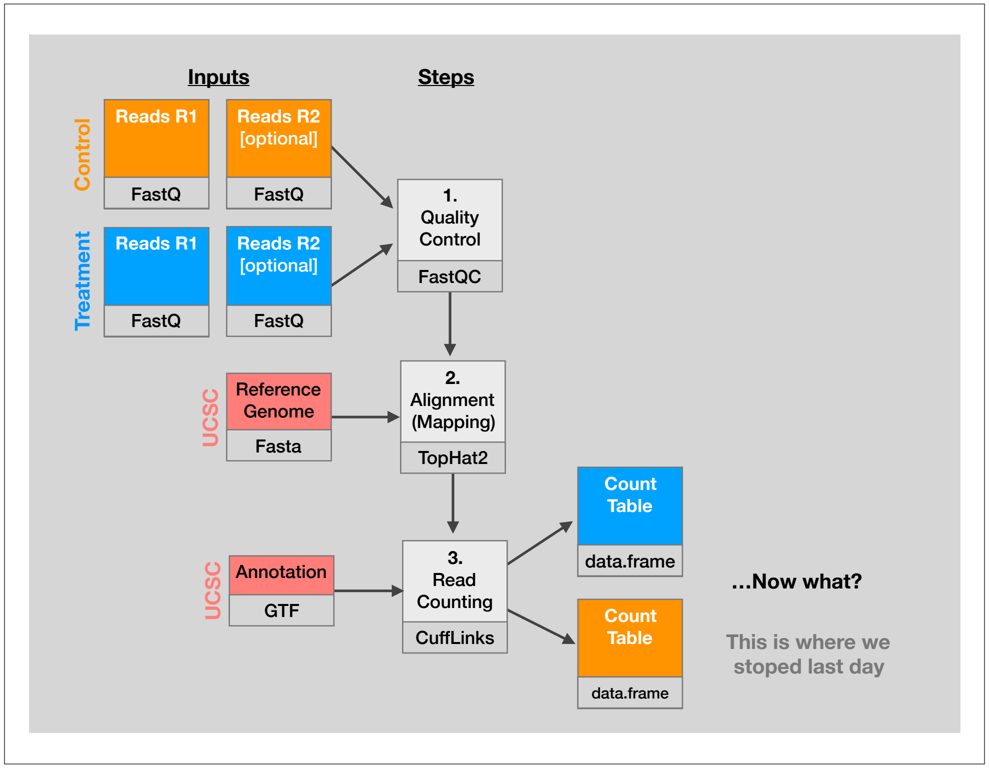
Control: Reads R1 — FastQ; Reads R2 [optional] — FastQ

Treatment: Reads R1 — FastQ; Reads R2 [optional] — FastQ

1. Quality Control — FastQC

UCSC: Reference Genome — Fasta

2. Alignment (Mapping) — TopHat2



Inputs **Steps**

Control: Reads R1 — FastQ; Reads R2 [optional] — FastQ

Treatment: Reads R1 — FastQ; Reads R2 [optional] — FastQ

1. Quality Control — FastQC

UCSC: Reference Genome — Fasta

2. Alignment (Mapping) — TopHat2

UCSC: Annotation — GTF

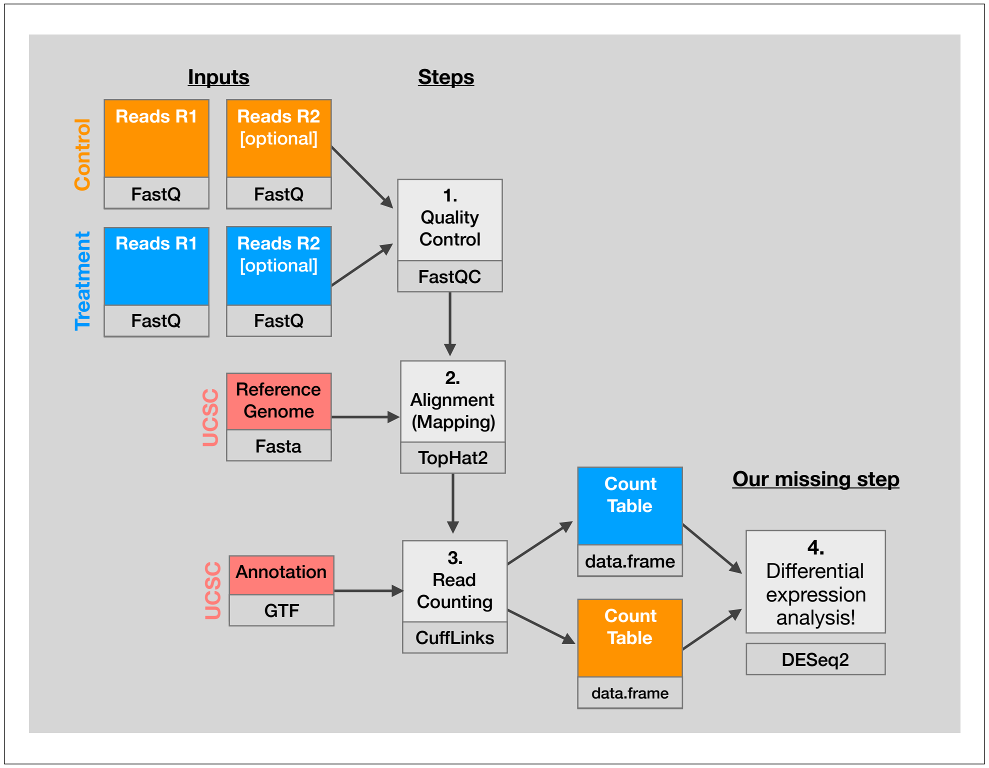
3. Read Counting — CuffLinks

Count Table — data.frame

Count Table — data.frame

...Now what?

This is where we stoped last day



Inputs **Steps**

Control: Reads R1 — FastQ; Reads R2 [optional] — FastQ

Treatment: Reads R1 — FastQ; Reads R2 [optional] — FastQ

1. Quality Control — FastQC

UCSC: Reference Genome — Fasta

2. Alignment (Mapping) — TopHat2

UCSC: Annotation — GTF

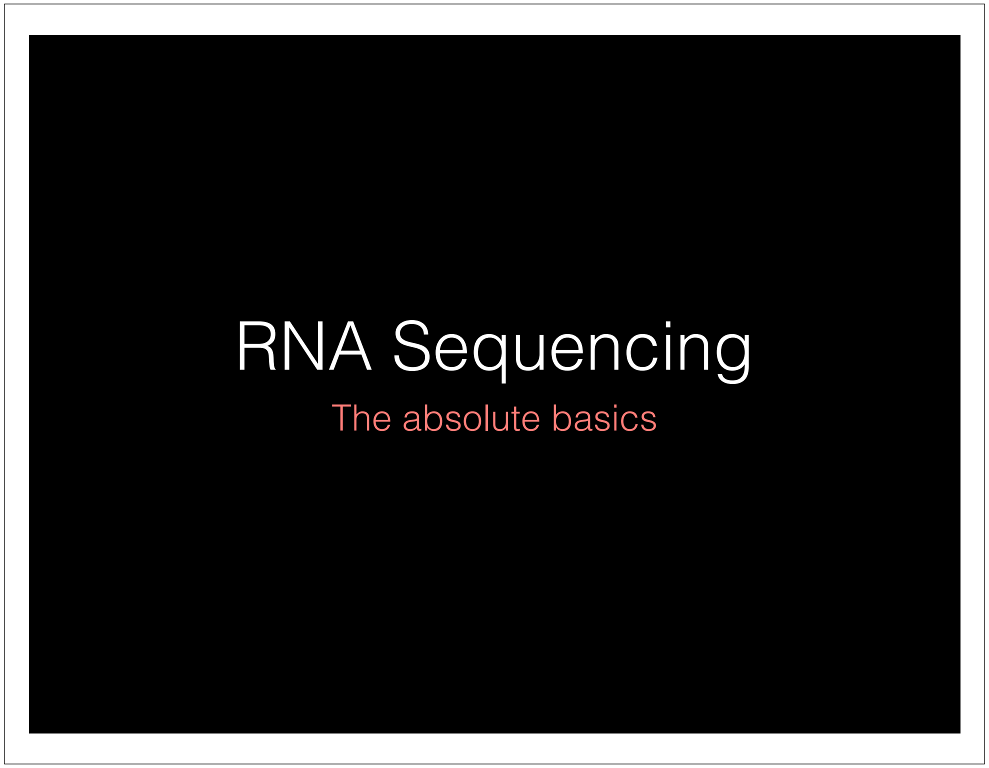
3. Read Counting — CuffLinks

Count Table — data.frame

Count Table — data.frame

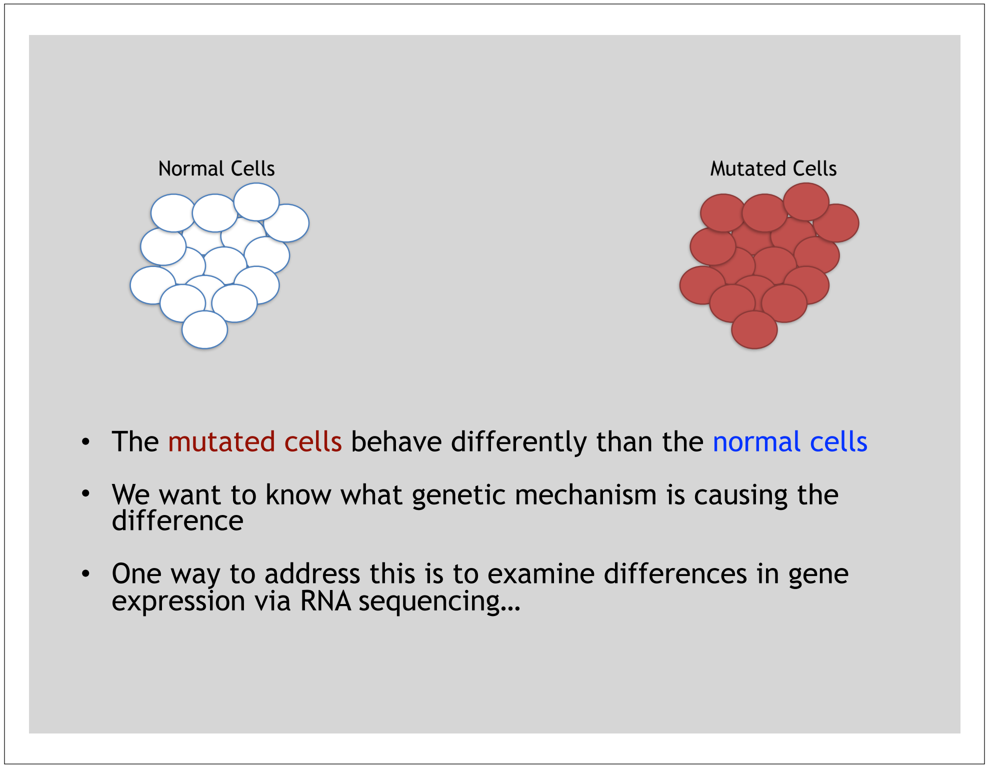
Our missing step

4. Differential expression analysis! — DESeq2



RNA Sequencing

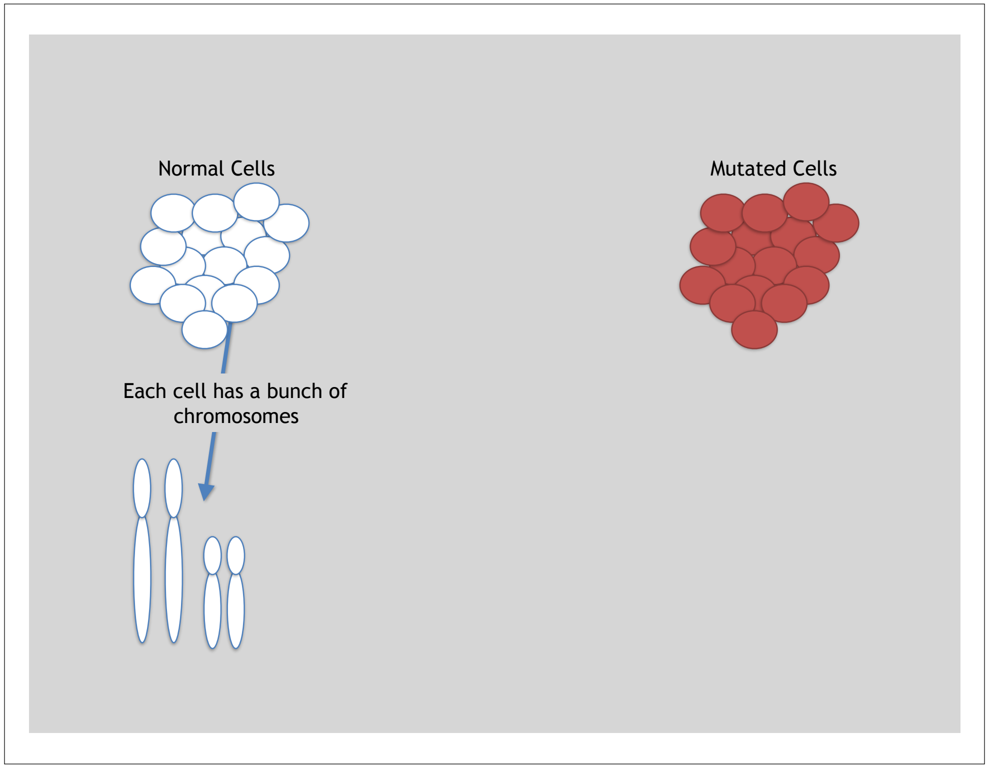
The absolute basics



Normal Cells

Mutated Cells

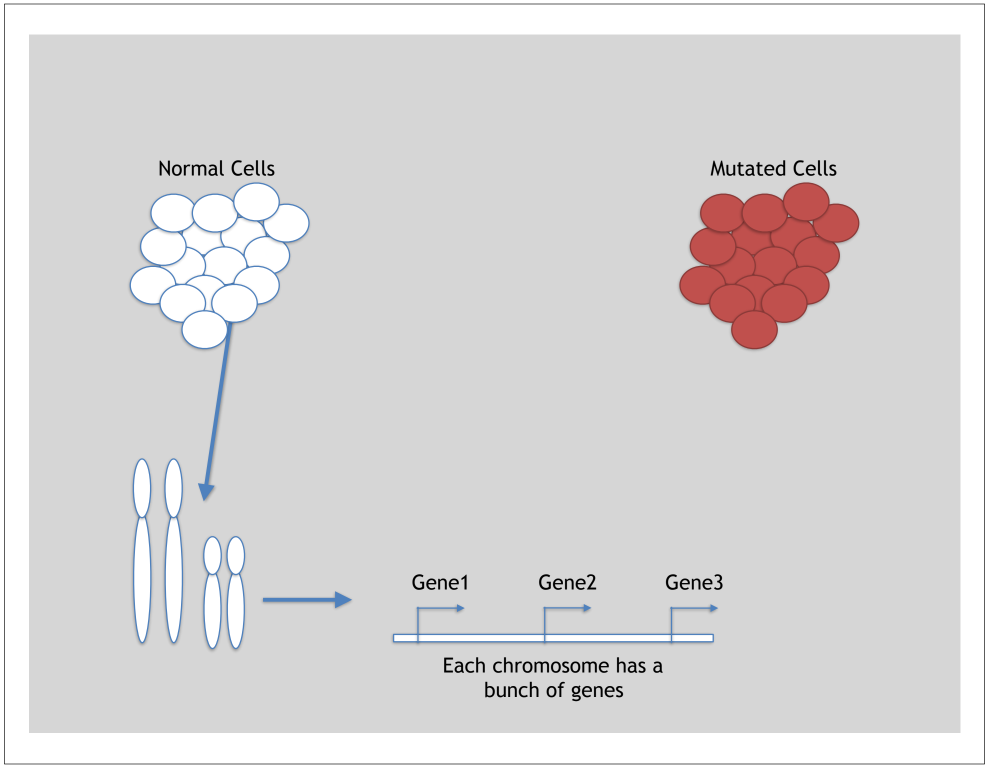
- The mutated cells behave differently than the normal cells
- We want to know what genetic mechanism is causing the difference
- One way to address this is to examine differences in gene expression via RNA sequencing…



Normal Cells

Mutated Cells

Each cell has a bunch of chromosomes



Normal Cells

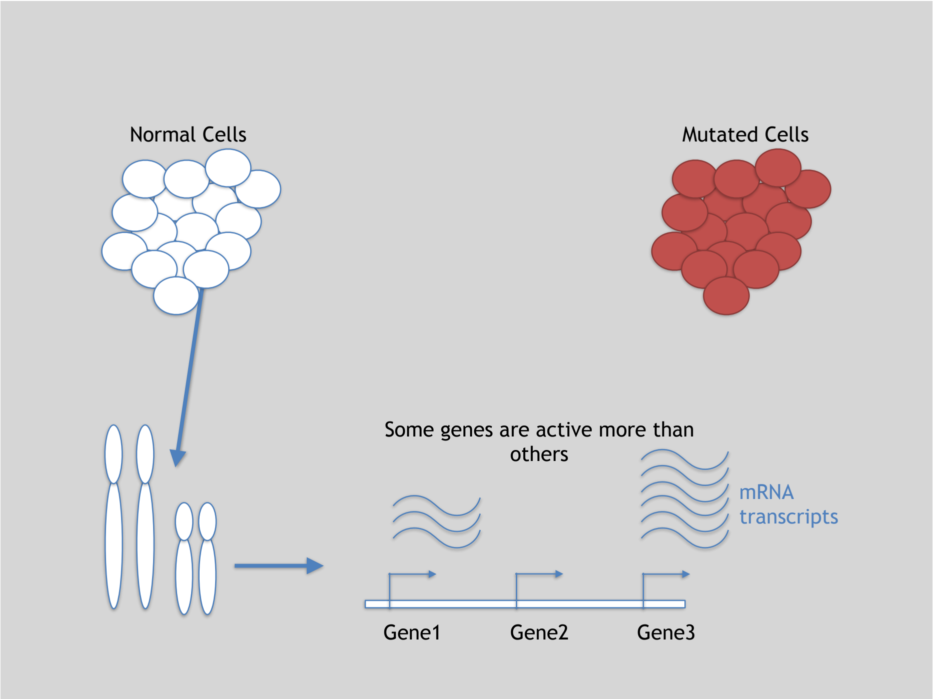
Mutated Cells

Gene1

Gene2

Gene3

Each chromosome has a bunch of genes



Normal Cells

Mutated Cells

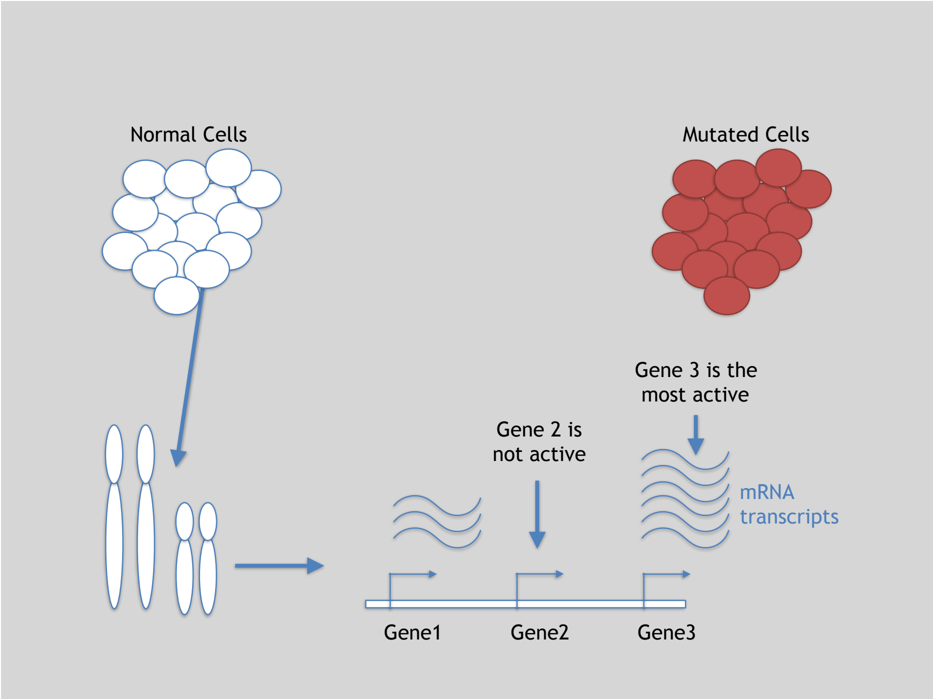
Some genes are active more than others

mRNA transcripts

Gene1

Gene2

Gene3



Normal Cells

Mutated Cells

Gene 3 is the most active

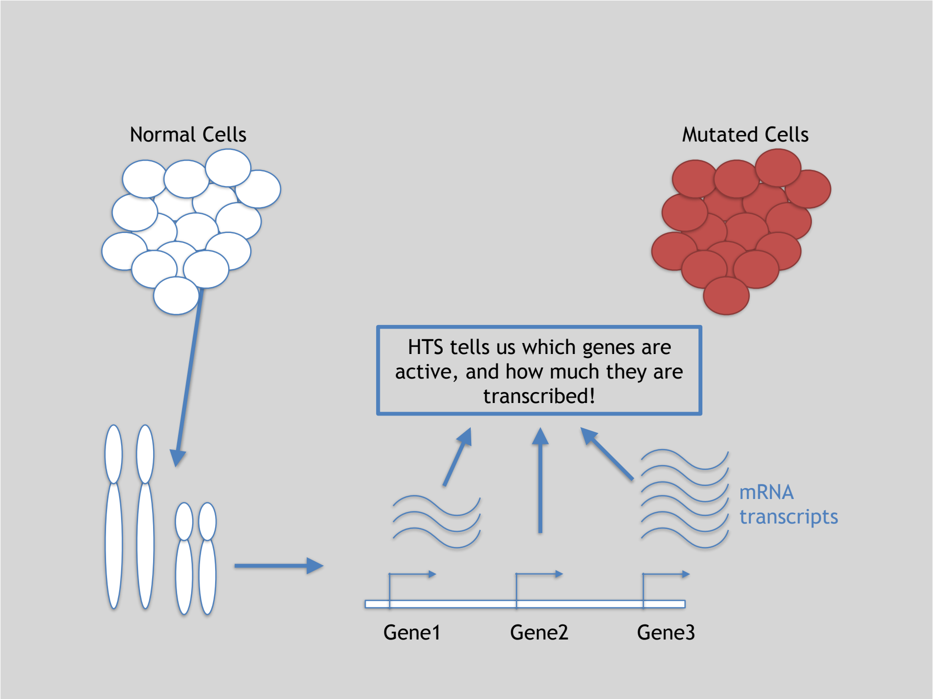
Gene 2 is not active

mRNA transcripts

Gene1

Gene2

Gene3



Normal Cells

Mutated Cells

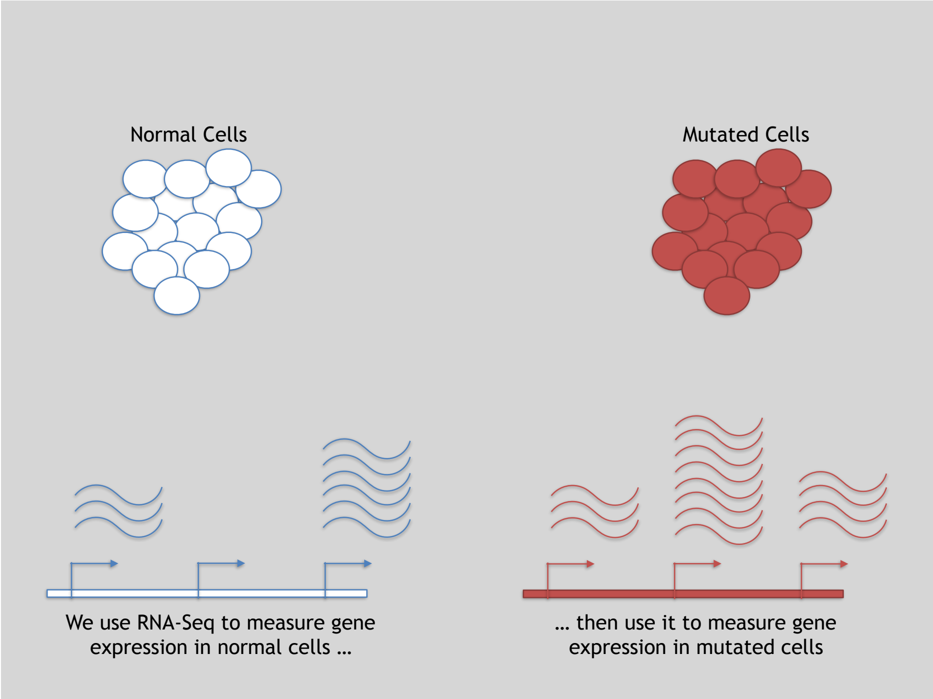
HTS tells us which genes are active, and how much they are transcribed!

mRNA transcripts

Gene1

Gene2

Gene3

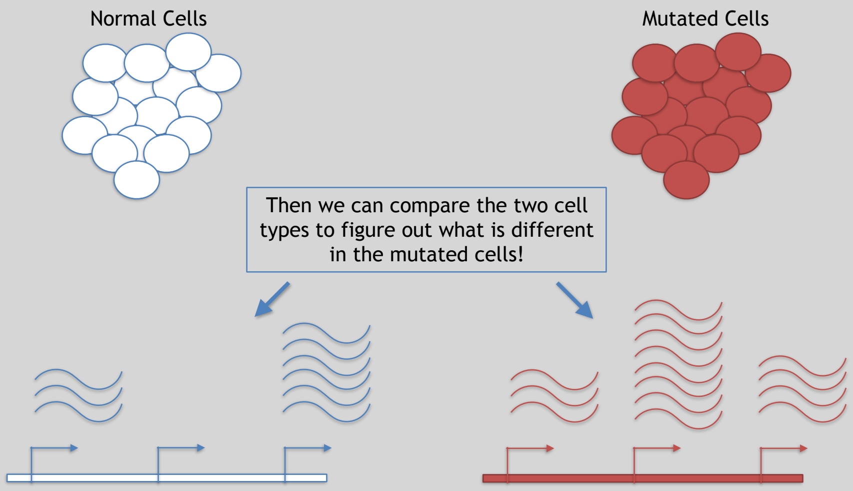


Normal Cells

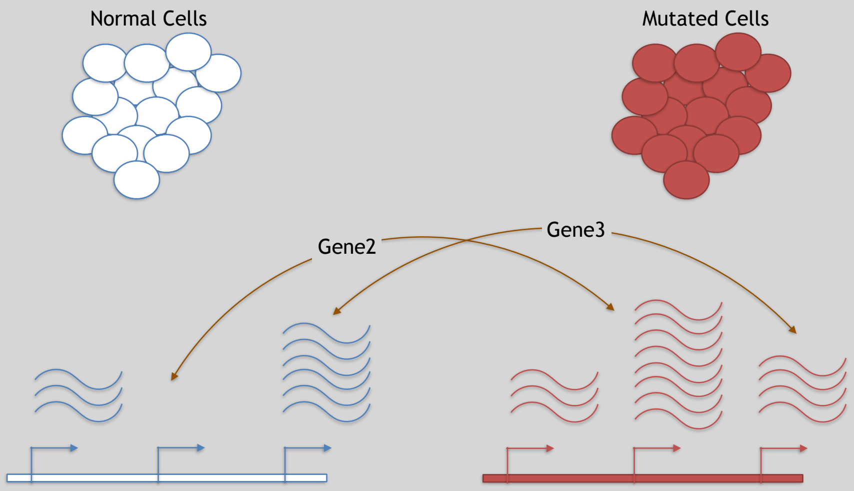
Mutated Cells

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

… then use it to measure gene expression in mutated cells



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



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

3 Main Steps for RNA-Seq:

1) **Prepare a sequencing library**
   (RNA to cDNA conversion via reverse transcription)

2) **Sequence**
   (Using the same technologies as DNA sequencing)

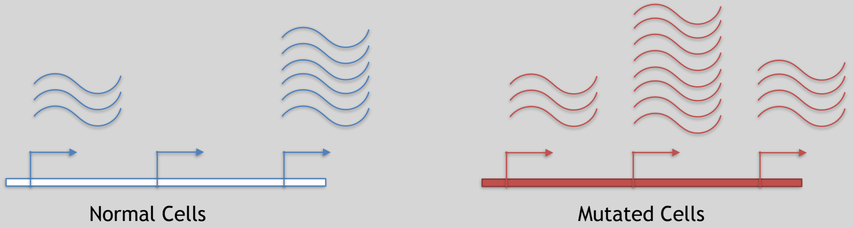
3) **Data analysis**
   (Often the major bottleneck to overall success!)

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

Today we will get start of step 3!

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

We **sequenced**, **aligned**, **counted** the reads per gene in each sample to arrive at our data matrix



Do it Yourself!

Hand-on time!

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

Feedback:
[Muddy Point Assessment]

Reference

Additional Reference Slides
on SAM/BAM Format and Sequencing Methods

Reference

SAM Format

- **S**equence **A**lignment/**M**ap (**SAM**) format is the almost-universal sequence alignment format for NGS
  - binary version is BAM
- It consists of a header section (lines start with '@') and an alignment section
- The official specification can be found here:
  - http://samtools.sourceforge.net/SAM1.pdf

Example SAM File

Reference

- Because SAM files are plain text (unlike their binary counterpart, BAM), we can take a peek at a few lines of the header with head, See:

https://bioboot.github.io/bimm143_F18/class-material/sam_format/

Header section

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

Alignment section

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

SAM header section

Reference

- Header lines contain vital metadata about the reference sequences, read and sample information, and (optionally) processing steps and comments.
- Each header line begins with an **@**, followed by a two-letter code that distinguishes the different type of metadata records in the header.
- Following this two-letter code are tab-delimited key-value pairs in the format **KEY:VALUE** (the SAM format specification names these tags and values).

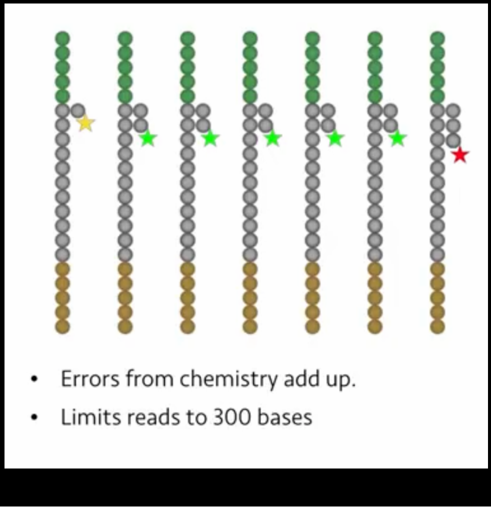
https://bioboot.github.io/bimm143_F18/class-material/sam_format/

SAM Utilities

Reference

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

Length limits for Illumina Sequencing

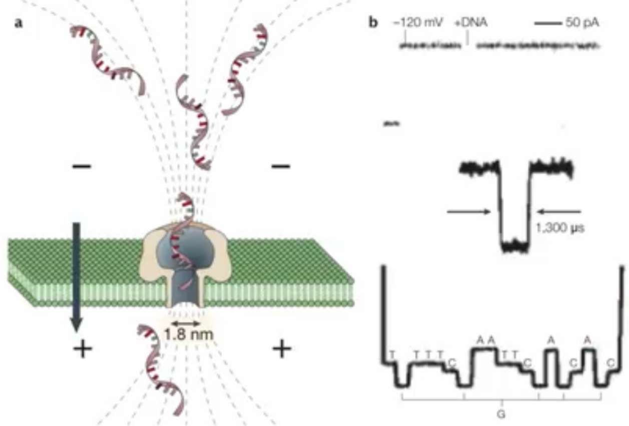


- Errors from chemistry add up.
- Limits reads to 300 bases

Reference

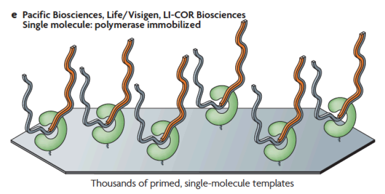
Additional Reference Slides on Sequencing Methods

Oxford Nanopore

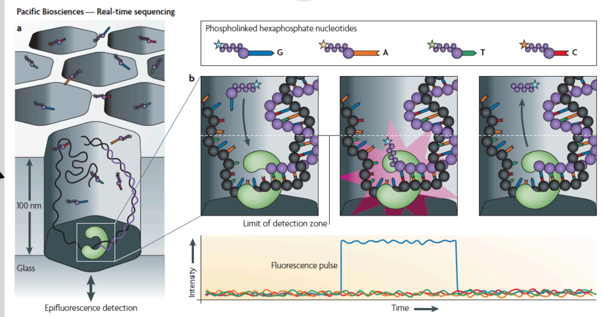


Nature Reviews Drug Discovery **volume 1**, 77-84 (2002)

Pacific Biosystems – Real Time Sequencing


e Pacific Biosciences, Life/Visigen, LI-COR Biosciences
Single molecule: polymerase immobilized
Thousands of primed, single-molecule templates

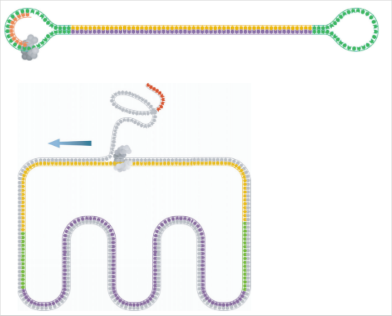
zero mode waveguides


Pacific Biosciences — Real-time sequencing
a
100 nm
Glass
Epifluorescence detection
Phospholinked hexaphosphate nucleotides
G
A
T
C
b
Limit of detection zone
Intensity
Fluorescence pulse
Time

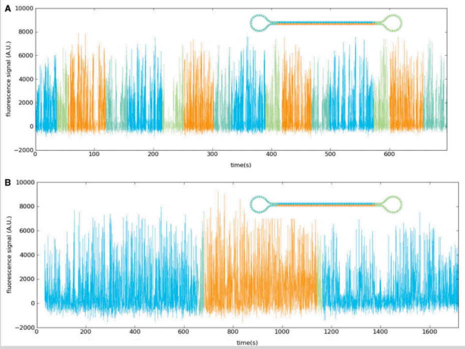
Metzker, ML (2010), *Nat. Rev. Genet*, 11, pp. 31-46

Pacific Biosystems – Circular Consensus

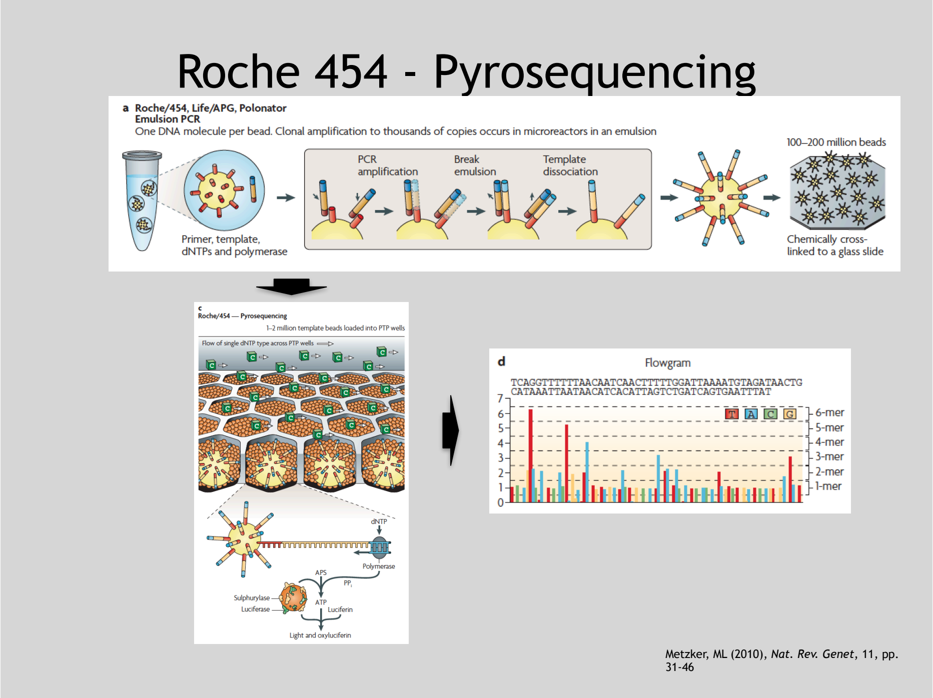
SMRTbell template



Subread Consensus Sequencing



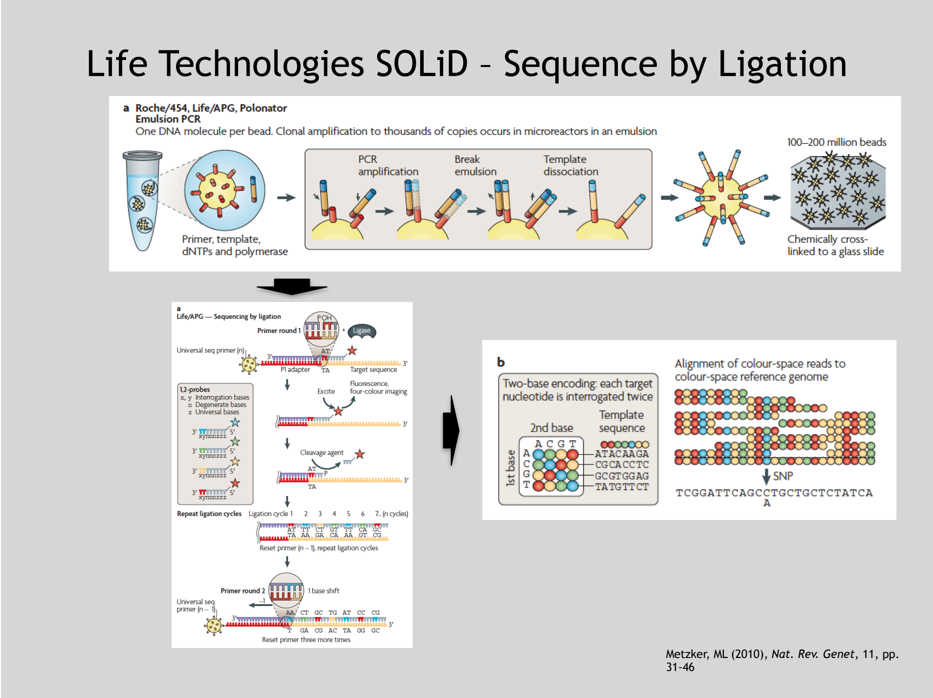
Travers, KJ et al (2010), *Nucl. Acids. Res.*, 38(15) pp. e159



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
100–200 million beads
PCR amplification
Break emulsion
Template dissociation
Primer, template, dNTPs and polymerase
Chemically cross-linked to a glass slide
c
Roche/454 — Pyrosequencing
1–2 million template beads loaded into PTP wells
Flow of single dNTP type across PTP wells
C
dNTP
Polymerase
APS
PP$_i$
ATP
Sulphurylase
Luciferase
Luciferin
Light and oxyluciferin
d
Flowgram
TCAGGTTTTTTAACAATCAACTTTTTGGATTAAAATGTAGATAACTG
CATAAATTAATAACATCACATTAGTCTGATCAGTGAATTTAT
7
6
5
4
3
2
1
0
T
A
C
G
6-mer
5-mer
4-mer
3-mer
2-mer
1-mer

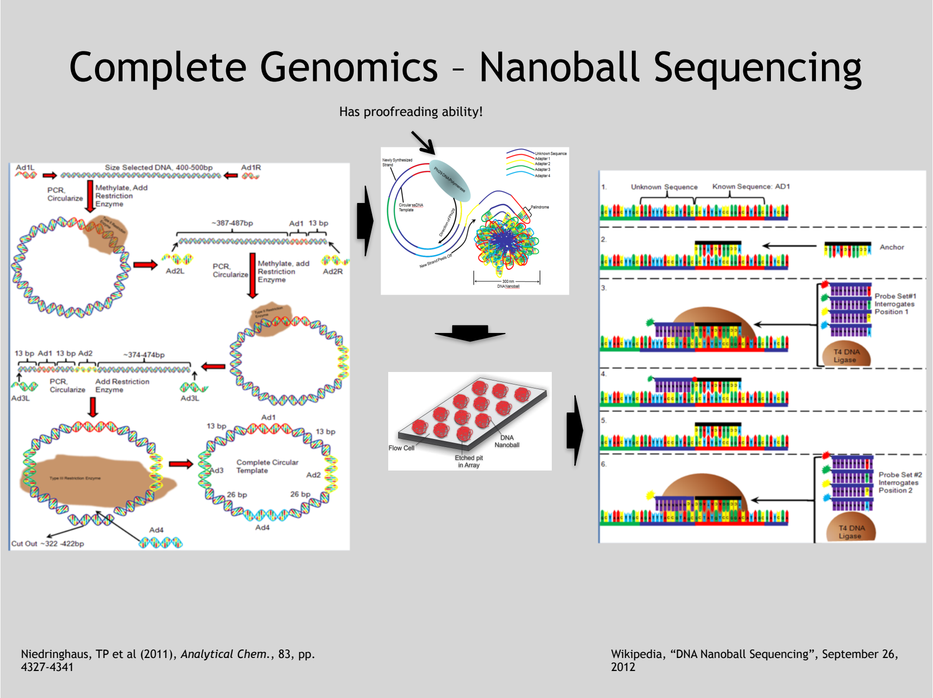
Metzker, ML (2010), *Nat. Rev. Genet*, 11, pp. 31-46



Life Technologies SOLiD – Sequence by Ligation

a Roche/454, Life/APG, Polonator
Emulsion PCR
One DNA molecule per bead. Clonal amplification to thousands of copies occurs in microreactors in an emulsion
100–200 million beads
PCR amplification
Break emulsion
Template dissociation
Primer, template, dNTPs and polymerase
Chemically cross-linked to a glass slide
a
Life/APG — Sequencing by ligation
Primer round 1
POH
Ligase
Universal seq primer (n)
3′
AT
TA
P1 adapter
Target sequence
3′
1,2-probes
x, y Interrogation bases
n Degenerate bases
z Universal bases
3′ xynnnzzz 5′
Excite
Fluorescence, four-colour imaging
Cleavage agent
P
Repeat ligation cycles
Ligation cycle 1 2 3 4 5 6 7... (n cycles)
AT TT CT GT TT CA GC
TA AA GA CA AA GT CG
Reset primer (n – 1), repeat ligation cycles
Primer round 2
1 base shift
Universal seq primer (n – 1)
–1
AA CT GC TG AT CC CG
T GA CG AC TA GG GC
Reset primer three more times
b
Two-base encoding: each target nucleotide is interrogated twice
2nd base
A C G T
1st base
A
C
G
T
Template sequence
ATACAAGA
CGCACCTC
GCGTGGAG
TATGTTCT
Alignment of colour-space reads to colour-space reference genome
SNP
TCGGATTCAGCCTGCTGCTCTATCA
A

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

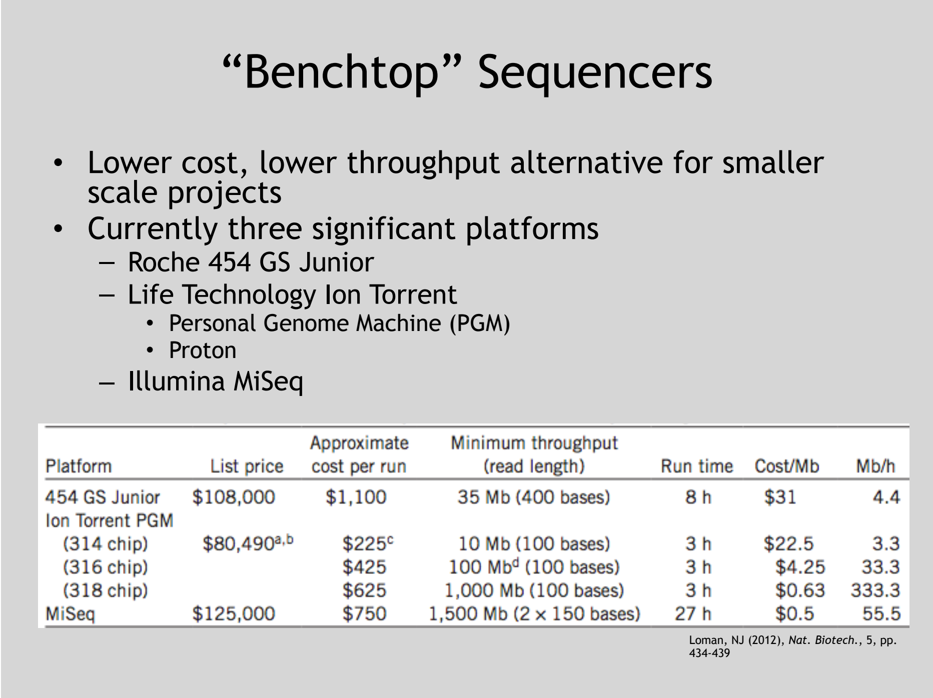


Complete Genomics – Nanoball Sequencing

Has proofreading ability!

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

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



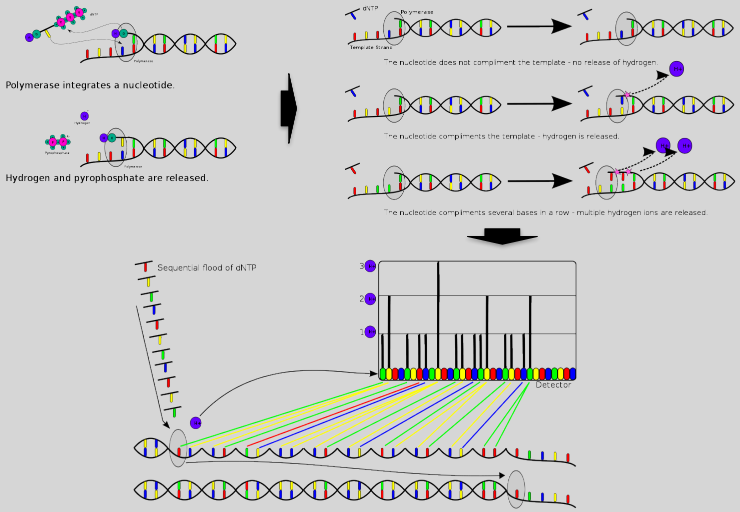
"Benchtop" Sequencers

- 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


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

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

Normalization

- Normalization is required to make comparisons in gene expression – Between 2+ genes in one sample – Between genes in 2+ samples
- Genes will have more reads mapped in sample with high coverage than with low read coverage – 2x depth ≈ 2x expression
- Longer genes will have more reads mapped than shorter genes – 2x length ≈ 2x more reads

Normalization: RPKM, FPKM and TPM

- **N.B.** Some tools for differential expression analysis such as edgeR and DESeq2 want raw read counts - i.e. non normalized input!
- However, often for your manuscripts and reports you will want to report normalized counts - e.g. plots of Log(FoldChange) vs Transcripts Per Million (or TPM)
- RPKM, FPKM and TPM all aim to normalize for sequencing depth and gene length.
- RPKM was made for single-end RNA-seq and stands for Reads per :
  - Count up the total reads in a sample and divide that number by 1,000,000 – this is our “per million” scaling factor.
  - Divide the read counts by the “per million” scaling factor. This normalizes for sequencing depth, giving you reads per million (RPM)
  - Divide the RPM values by the length of the gene, in kilobases. This gives you RPKM.

- FPKM was made for paired-end RNA-seq
- With paired-end RNA-seq, two reads can correspond to a single fragment
- The only difference between RPKM and FPKM is that FPKM takes into account that two reads can map to one fragment (and so it doesn’t count this fragment twice).

- TPM is very similar to RPKM and FPKM. The only difference is the order of operations. Here’s how you calculate TPM:
  - Divide the read counts by the length of each gene in kilobases. This gives you reads per kilobase (RPK).
  - Count up all the RPK values in a sample and divide this number by 1,000,000. This is your “per million” scaling factor.
  - Divide the RPK values by the “per million” scaling factor. This gives you TPM.
- So you see, when calculating TPM, the only difference is that you normalize for gene length first, and then normalize for sequencing depth second. However, the effects of this difference are quite profound.

- When you use TPM, the sum of all TPMs in each sample are the same. This makes it easier to compare the proportion of reads that mapped to a gene in each sample. In contrast, with RPKM and FPKM, the sum of the normalized reads in each sample may be different, and this makes it harder to compare samples directly.