



University of Michigan  
Medical School

---

# Genome Informatics

**Ryan E. Mills, Ph.D.**

Department of Computational Medicine & Bioinformatics

Department of Human Genetics

University of Michigan Medical School

Ann Arbor, MI, USA

---



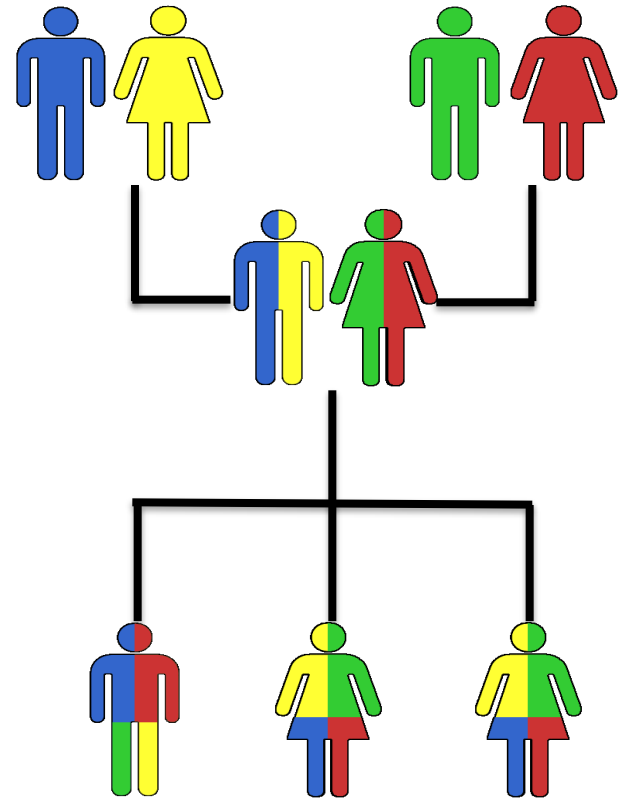
- Genetics is primarily the study of individual genes, mutations within those genes, and their inheritance patterns in order to understand specific traits.
  - Genomics expands upon classical genetics and considers aspects of the entire genome, typically using computer aided approaches.
-



University of Michigan  
Medical School

# What is a Genome?

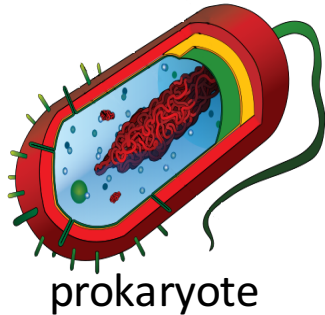
The total genetic material of an organism by which individual traits are encoded, controlled, and ultimately passed on to future generations



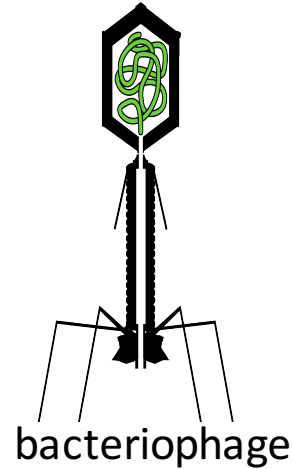


University of Michigan  
Medical School

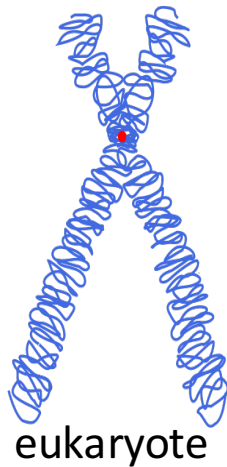
# Genomes come in many shapes



prokaryote



bacteriophage



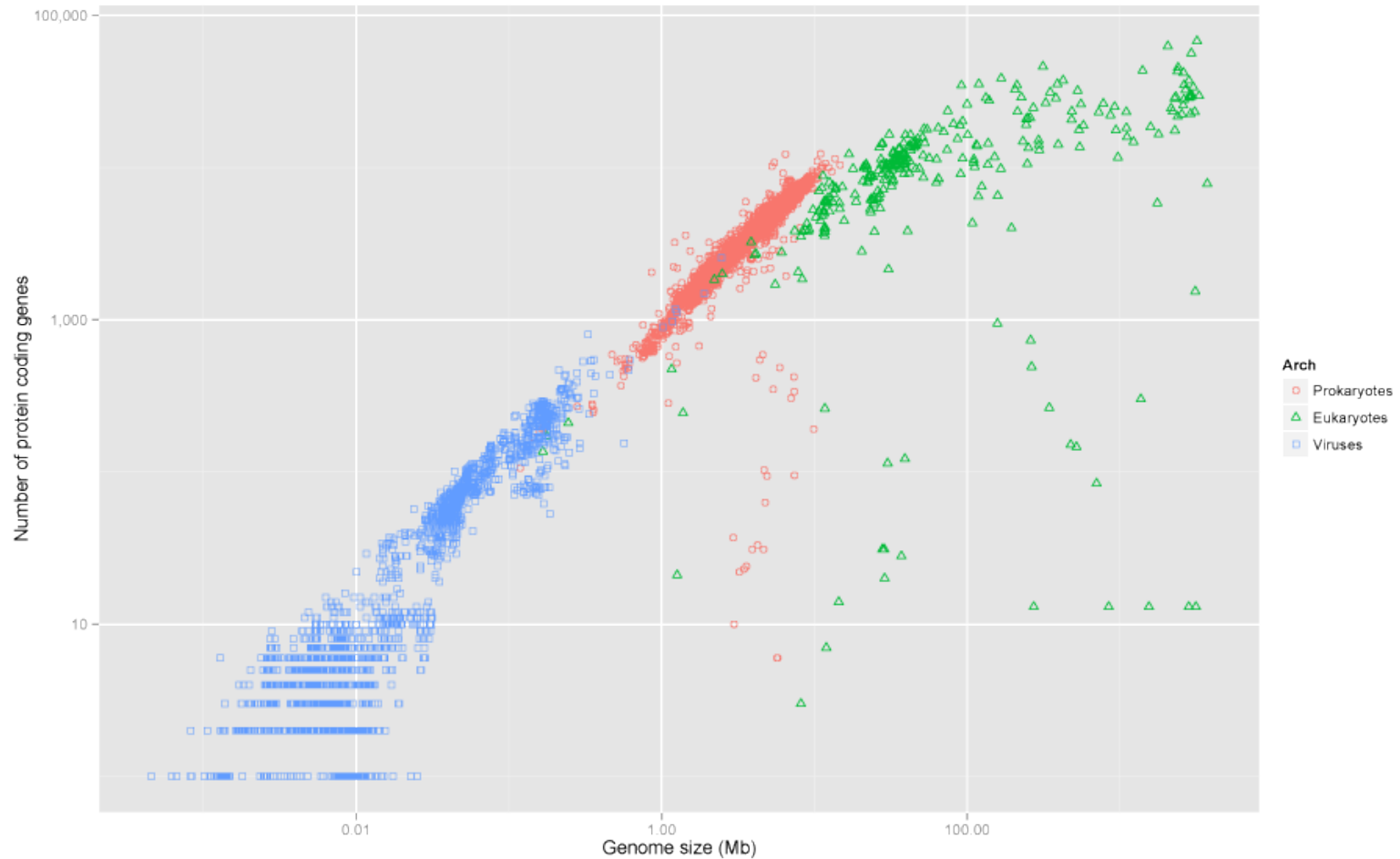
eukaryote

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





# Genomes come in many sizes






## NCBI Genome:

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

NCBI Resources How To Sign in to NCBI

Genome  Search Limits Advanced Help



### Genome

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

#### Using Genome

- [Help](#)
- [Browse by Organism](#)
- [Download / FTP](#)
- [Download FAQ](#)
- [Submit a genome](#)

#### Custom resources

- [Human Genome](#)
- [Microbes](#)
- [Organelles](#)
- [Viruses](#)
- [Prokaryotic reference genomes](#)

#### Other Resources

- [Assembly](#)
- [BioProject](#)
- [BioSample](#)
- [Map Viewer](#)
- [Protein Clusters](#)

#### Genome Tools

- [BLAST the Human Genome](#)
- [Microbial Nucleotide BLAST](#)
- [TaxPlot \(3-way Genome Comparison\)](#)

#### Genome Annotation and Analysis

- [Eukaryotic Genome Annotation](#)
- [Prokaryotic Genome Annotation](#)
- [PASC \(Pairwise Sequence Comparison\)](#)

#### External Resources

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

You are here: NCBI > Genomes & Maps > Genome Write to the Help Desk

#### GETTING STARTED

- [NCBI Education](#)
- [NCBI Help Manual](#)
- [NCBI Handbook](#)
- [Training & Tutorials](#)

#### RESOURCES

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

#### POPULAR

- [PubMed](#)
- [Bookshelf](#)
- [PubMed Central](#)
- [PubMed Health](#)
- [BLAST](#)
- [Nucleotide](#)
- [Genome](#)
- [SNP](#)
- [Gene](#)
- [Protein](#)
- [PubChem](#)

#### FEATURED


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

#### NCBI INFORMATION

- [About NCBI](#)
- [Research at NCBI](#)
- [NCBI News](#)
- [NCBI FTP Site](#)
- [NCBI on Facebook](#)
- [NCBI on Twitter](#)
- [NCBI on YouTube](#)

Copyright | Disclaimer | Privacy | Browsers | Accessibility | Contact

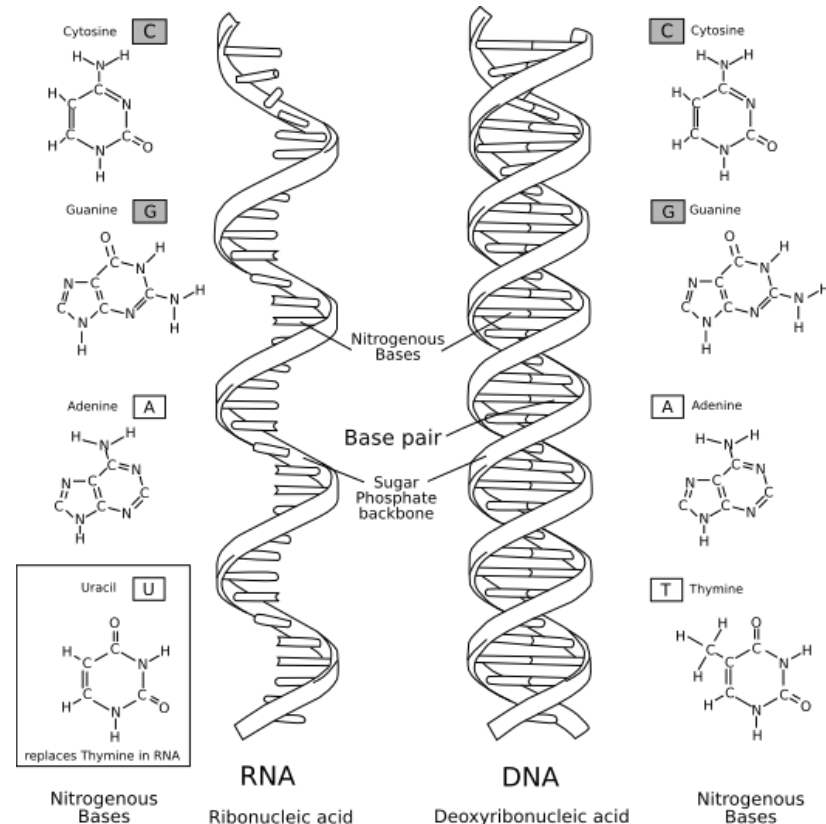
National Center for Biotechnology Information, U.S. National Library of Medicine  
8600 Rockville Pike, Bethesda MD, 20894 USA



# 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:

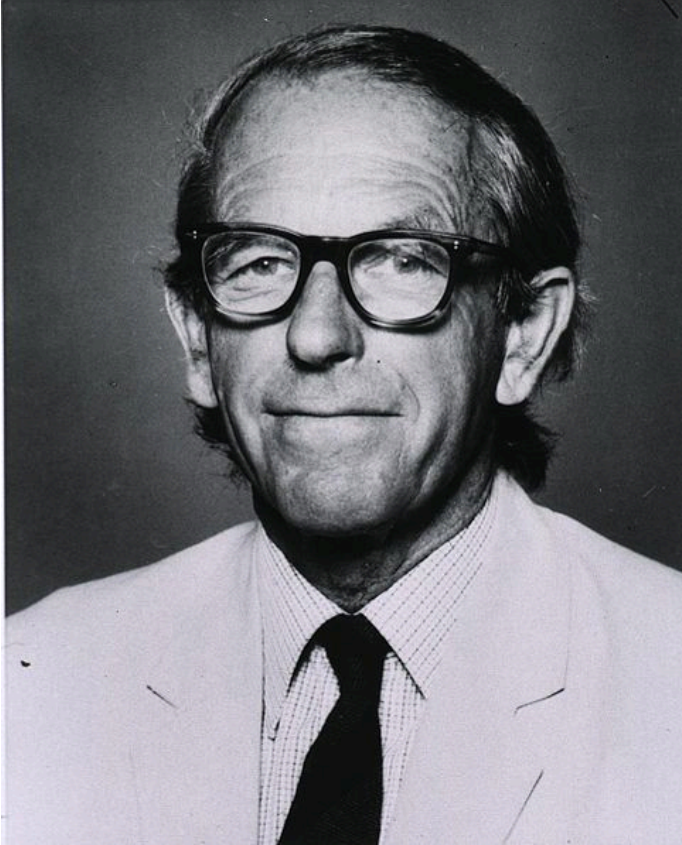
*GGACTTCAGGCAACTGCAACTACCTTAGGA*





University of Michigan  
Medical School

# Early Genome Sequencing

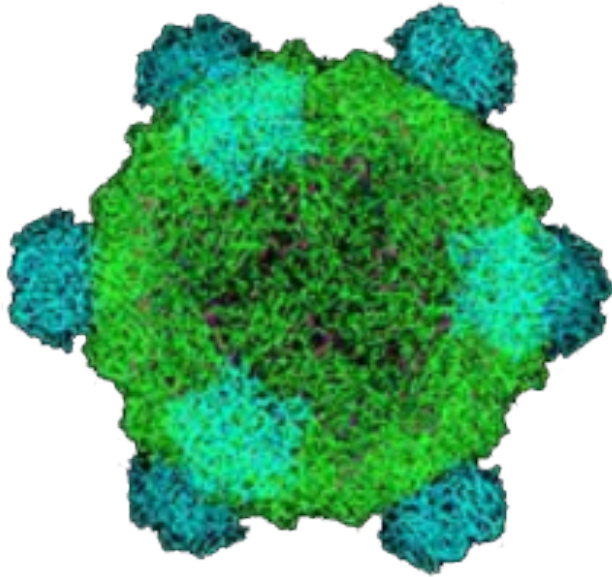


- Chain-termination “Sanger” sequencing was developed in 1977 by Frederick Sanger, colloquially referred to as the “Father of Genomics”
- Sequence reads were typically 750-1000 base pairs in length with an error rate of  $\sim 1 / 10000$  bases



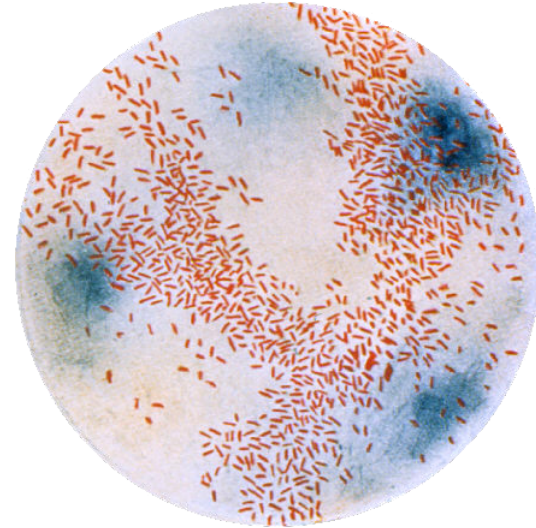
University of Michigan  
Medical School

# The First Sequenced Genomes



Bacteriophage  $\phi$ -X174

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



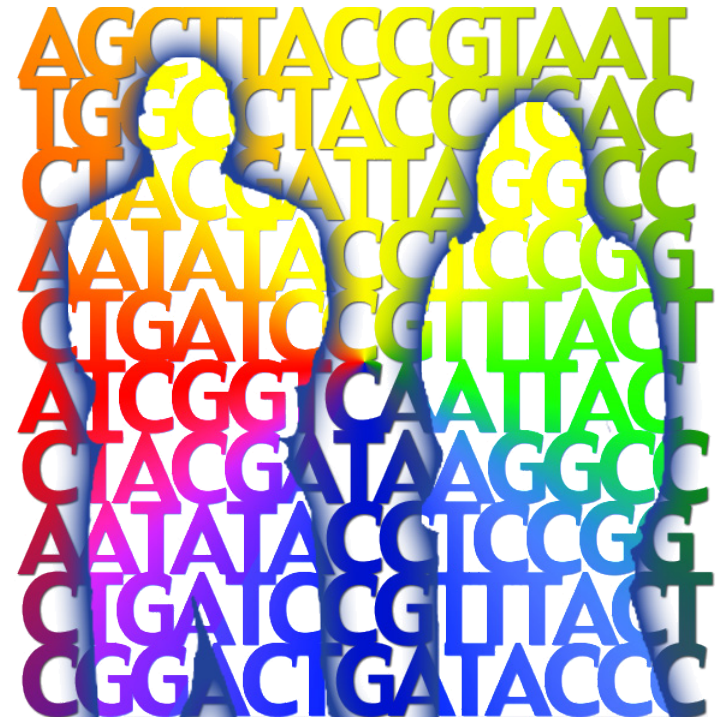
Haemophilus influenzae

- 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?

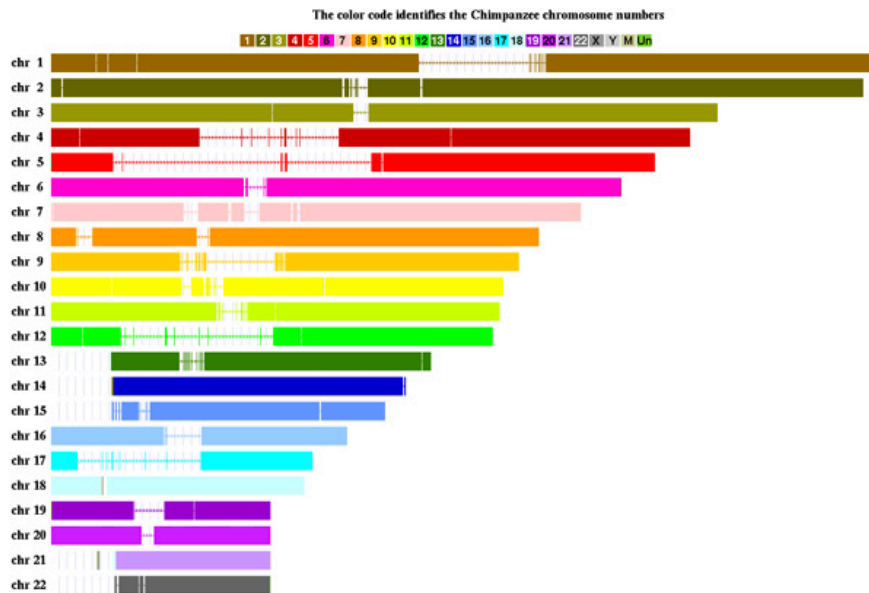
---

- 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 *edit* genomes, to add, remove, or modify genes and other regions for adjusting individual traits
-

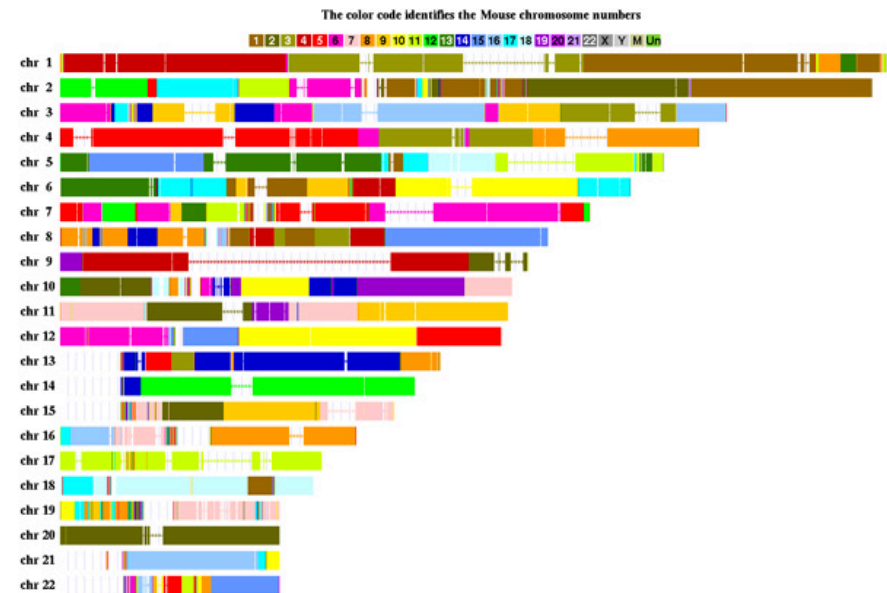




~6-7 million years



~60-70 million years

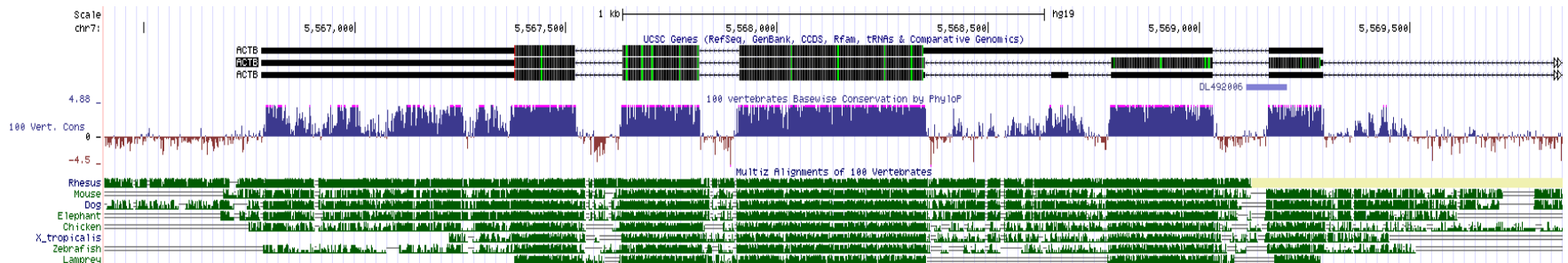






# Conservation Suggests Function

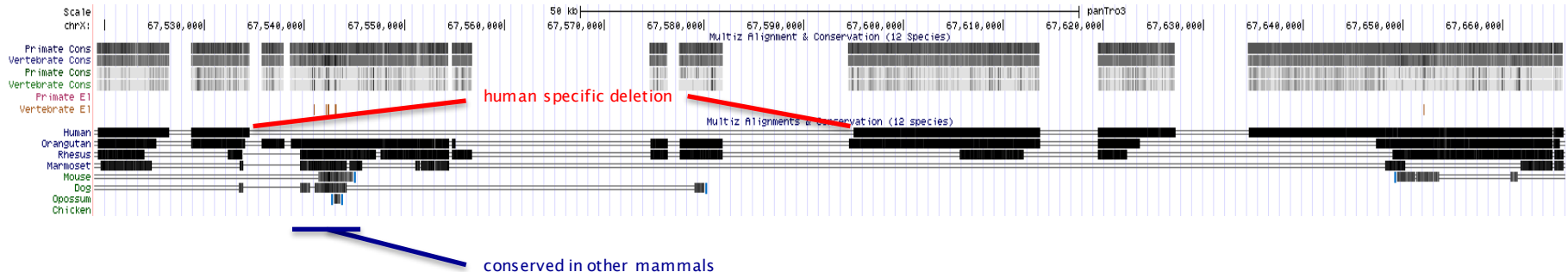
- Functional regions of the genome tend to mutate slower than nonfunctional regions due to selective pressures
- Comparing genomes can therefore indicate segments of high similarity that have remained conserved across species as candidate genes or regulatory regions





# 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)





# Gene Modeling and Prediction

Genomic features such as codon usage patterns can be modeled to identify novel genic regions

## Genic Regions:

ATGACGGTCTGACTGACCATTTCAT...  
ATGCGGTATGCCTGGAAGCCAGCC...  
ATGCTATGGCAAGCCATTGGAGAG...  
ATGCCGTTGTTATATGAGCAGATA...

$P(\text{AAA}) = 0.001$   
 $P(\text{AAC}) = 0.020$   
 $P(\text{AAG}) = 0.025$   
 $P(\text{AAT}) = 0.003$   
.  
.  
 $P(\text{TTT}) = 0.022$



## Intergenic Regions:

TTAGATTAAAGCCAGGACTGAACG...  
TAATGAATGGAACCATACAGAACG...  
GACTTAGCCCCGAATTTATAGATAC...  
TAACTGAATCGATTTAGCCAGGAA...

$P(\text{AAA}) = 0.020$   
 $P(\text{AAC}) = 0.001$   
 $P(\text{AAG}) = 0.003$   
 $P(\text{AAT}) = 0.040$   
.  
.  
 $P(\text{TTT}) = 0.030$





## GeneMark:

<http://exon.gatech.edu/GeneMark/>

### GeneMark

A family of gene prediction programs developed at  
**Georgia Institute of Technology**, Atlanta, Georgia, USA.

What's New: A new algorithm,  
**BRAKER1**, an RNA-seq based  
eukaryotic genome annotation pipeline --  
using  
GeneMark-ET and AUGUSTUS



#### Gene Prediction in Bacteria, Archaea, Metagenomes and Metatranscriptomes



Novel genomic sequences can be analyzed either by the self-training program **GeneMarkS** (sequences longer than 50 kb) or by **GeneMark.hmm with Heuristic models**. For many species pre-trained model parameters are ready and available through the **GeneMark.hmm** page. Metagenomic sequences can be analyzed by **MetaGeneMark**, the program optimized for speed.

#### Gene Prediction in Eukaryotes



Novel genomes can be analyzed by the program **GeneMark-ES** utilizing unsupervised training. Note that GeneMark-ES has a special mode for analyzing fungal genomes. Recently, we have developed a semi-supervised version of GeneMark-ES, called GeneMark-ET that uses RNA-Seq reads to improve training. For several species pre-trained model parameters are ready and available through the **GeneMark.hmm** page.

#### Gene Prediction in Transcripts



Sets of assembled eukaryotic transcripts can be analyzed by the modified **GeneMarkS** algorithm (the set should be large enough to permit self-training). A single transcript can be analyzed by a special version of **GeneMark.hmm with Heuristic models**. A new advanced algorithm GeneMarkS-T was developed recently (manuscript sent to publisher); The GeneMarkS-T software (beta version) is available for [download](#).

#### Gene Prediction in Viruses, Phages and Plasmids



Sequences of viruses, phages or plasmids can be analyzed either by the **GeneMark.hmm with Heuristic models** (if the sequence is shorter than 50 kb) or by the self-training program **GeneMarkS**.

All the software programs mentioned here are available for download and local installation.

The software of GeneMark line is a part of genome annotation pipelines at NCBI, JGI, Broad Institute as well as the following software packages:

- **QUAST**: quality assessment tool for genome assemblies  
-- using GeneMarkS
- **MetAMOS**: a modular and open source metagenomic assembly and analysis  
-- using MetaGeneMark
- **MAKER2**: a eukaryotic genome annotation pipeline  
-- using GeneMark-ES (along with SNAP and AUGUSTUS)
- **BRAKER1**: an RNA-seq based eukaryotic genome annotation pipeline  
-- using GeneMark-ET and AUGUSTUS

For more information see [Background](#) and [Publications](#).

#### Borodovsky Group Group news

##### Gene Prediction Programs

- GeneMark
- GeneMark.hmm
- GeneMarkS
- Heuristic models
- MetaGeneMark
- Mirror site at NCBI
- GeneMarkS+
- BRAKER1

##### Information

- Publications
- Selected Citations
- Background
- FAQ
- Contact

##### Downloads

- Programs
- Prebuild species models

##### Other Programs

- UnSplicer
- GeneTack
- Frame-by-Frame
- IPSSP

##### In silico Biology International Conferences

- 2013
- 2011
- 2009
- 2007
- 2005
- 2003
- 2001
- 1999
- 1997

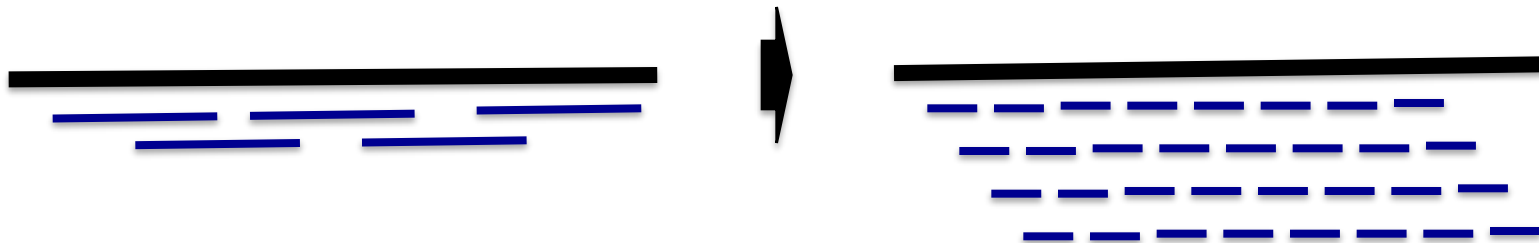
##### Bioinformatics Studies at Georgia Tech

- MS Program
- PhD Program
- Center for Bioinformatics and Computational Biology
- Lectures



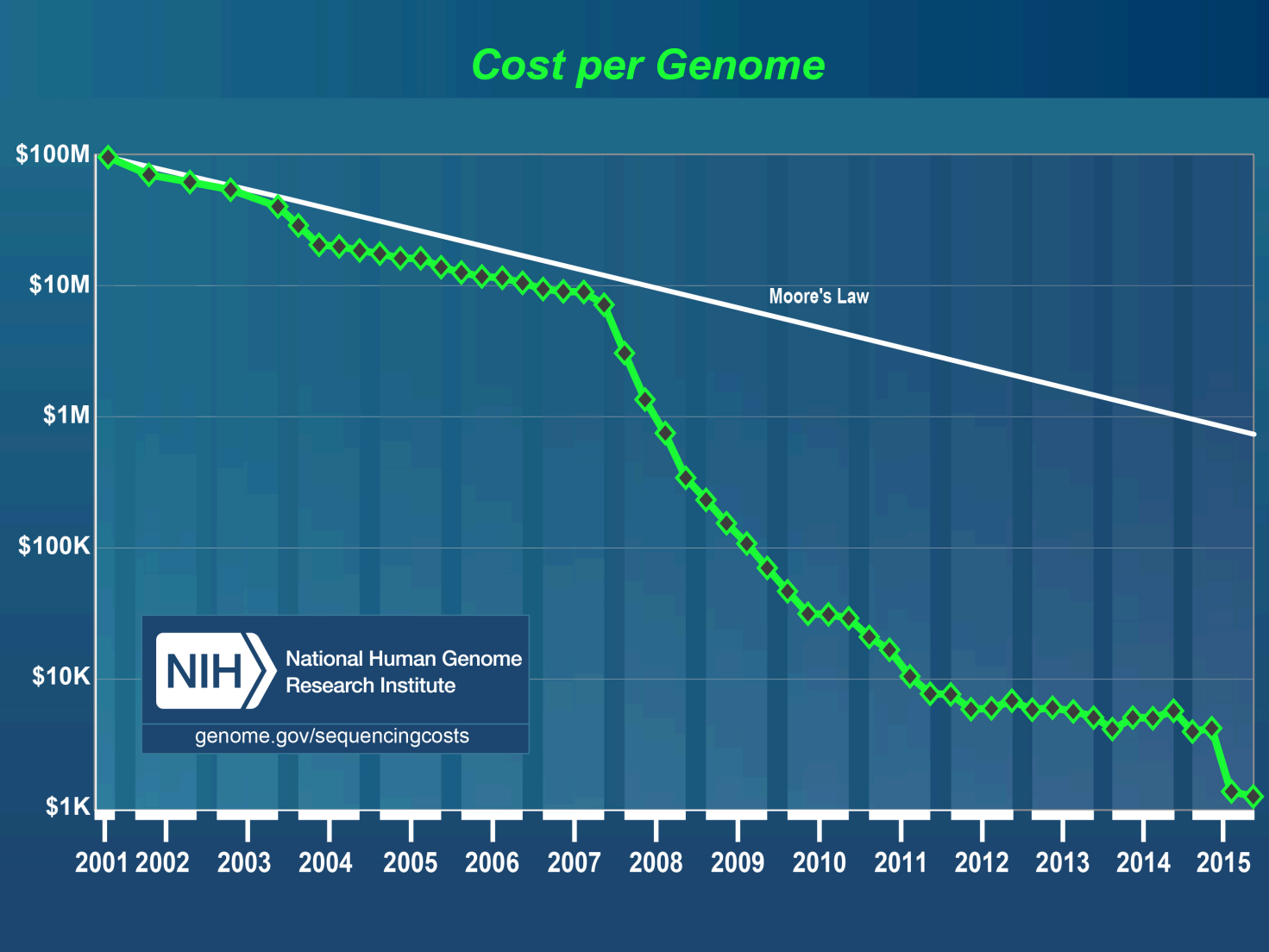
# Modern Genome Sequencing

- 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



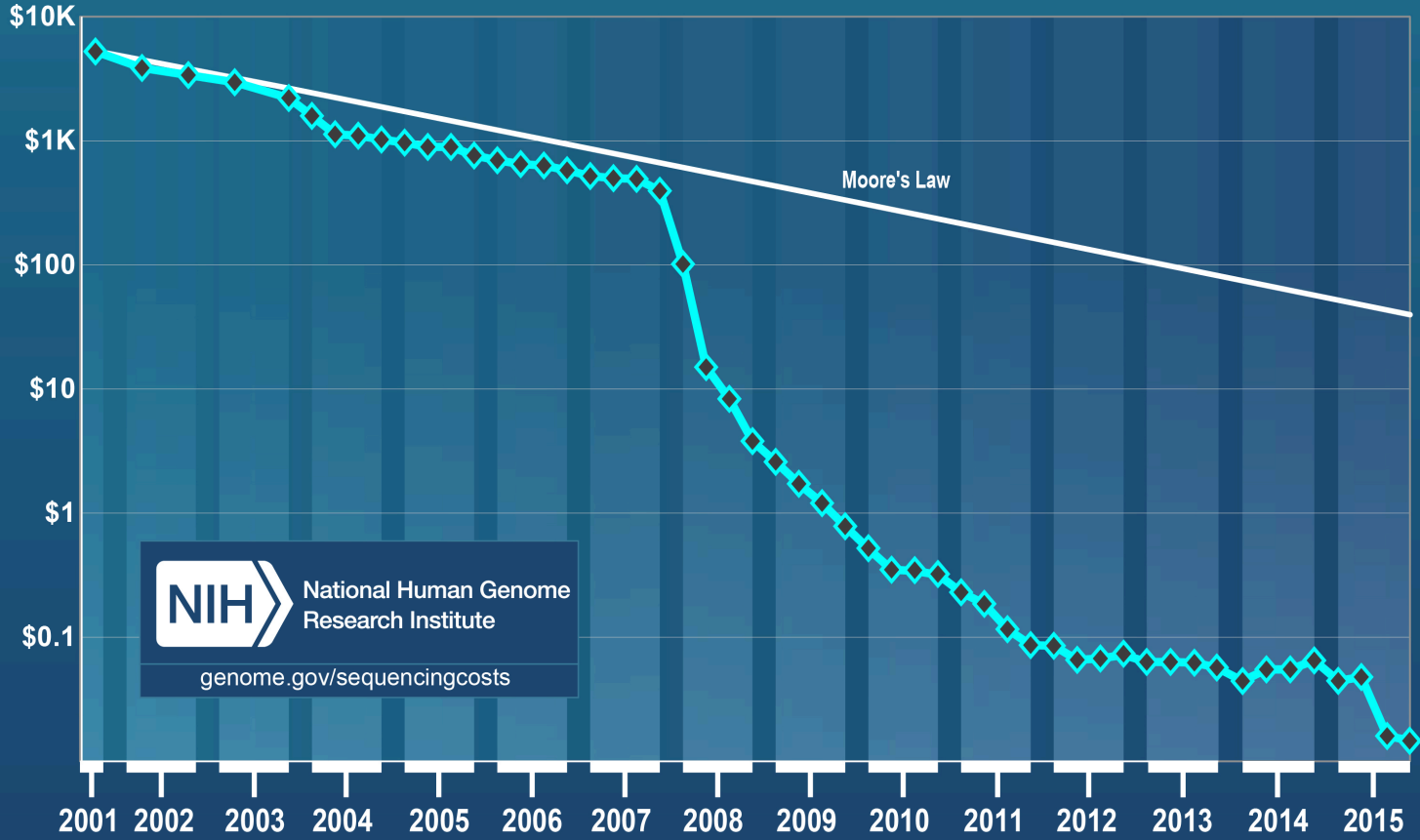


University of Michigan  
Medical School



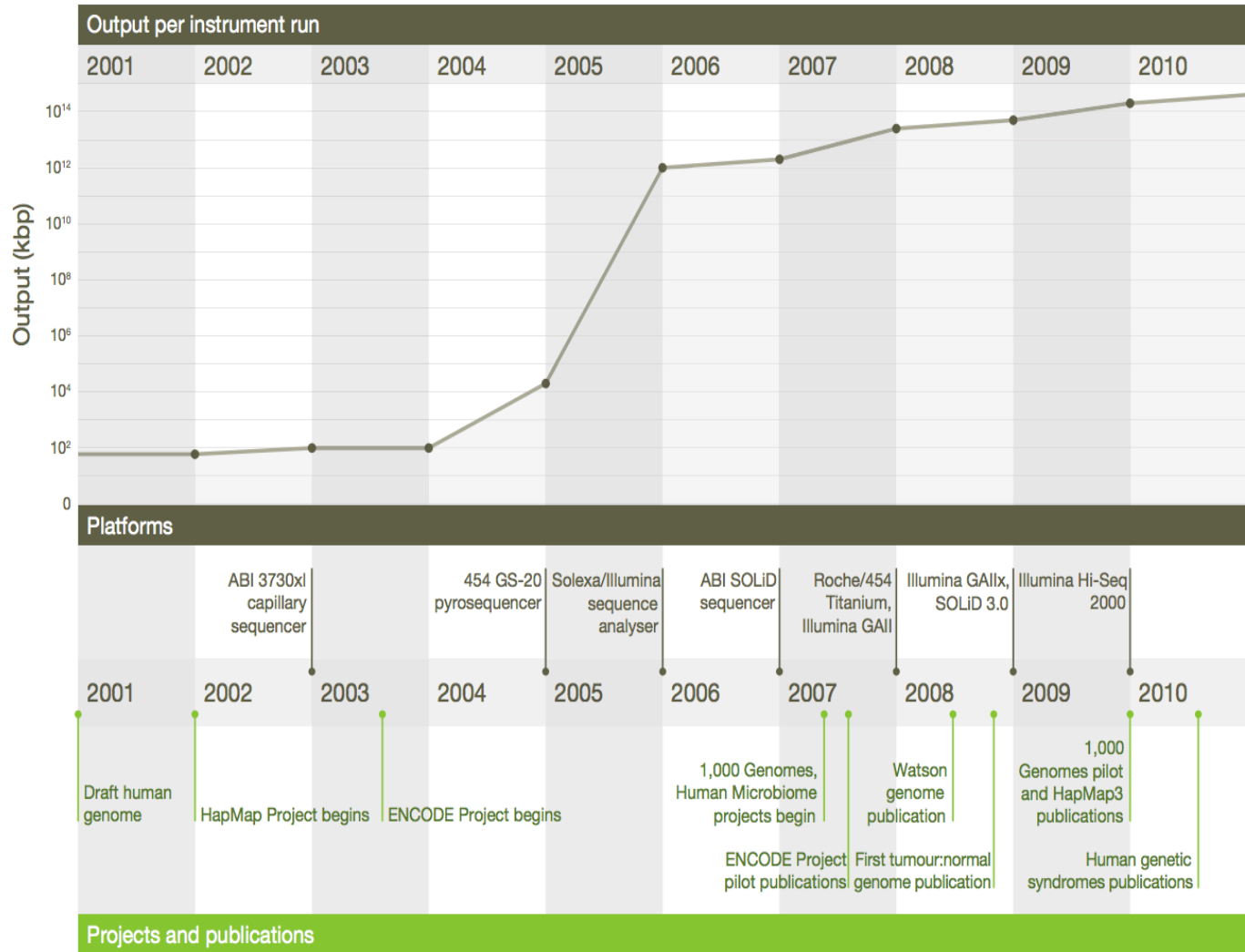


# Cost per Raw Megabase of DNA Sequence





# 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)



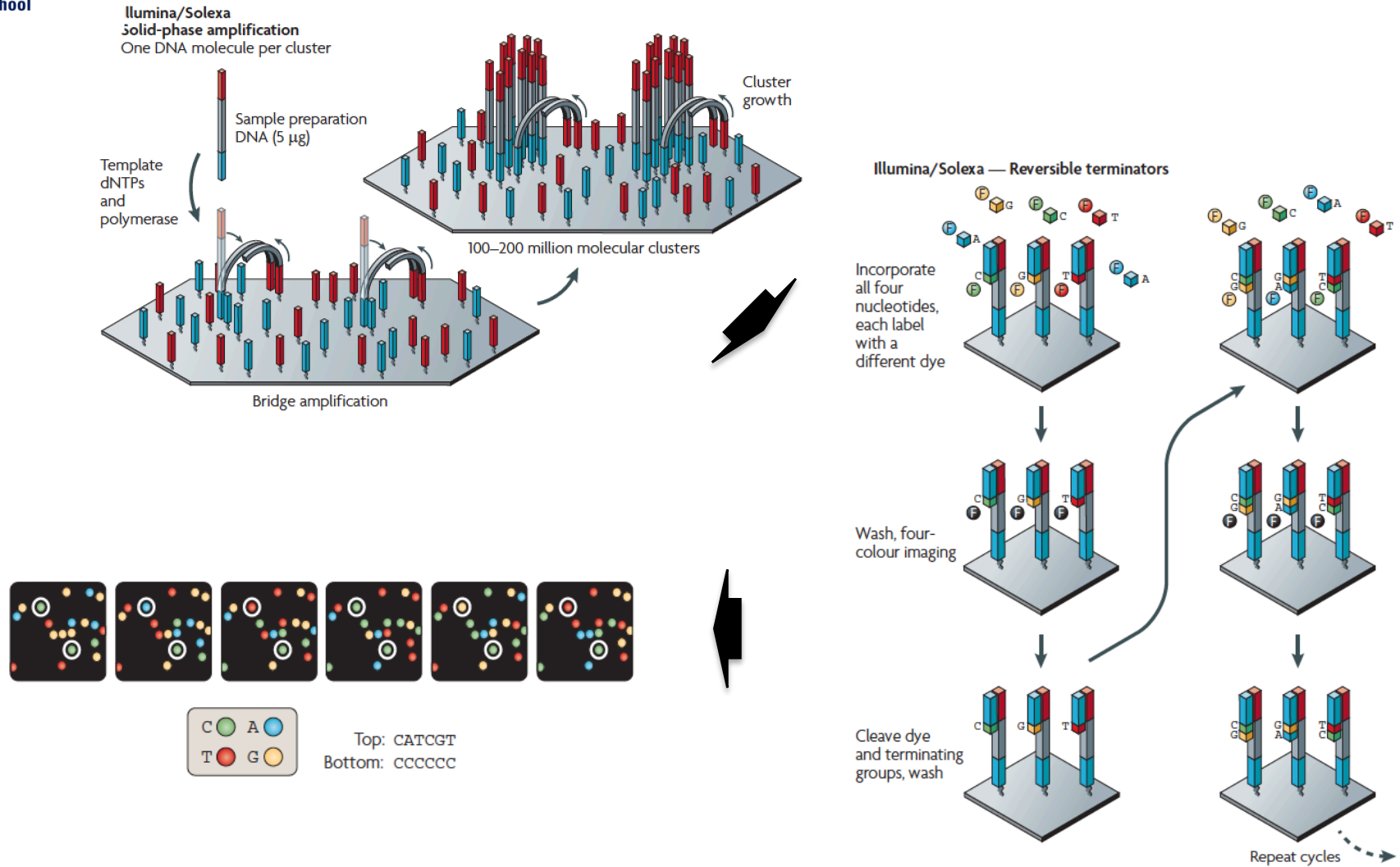
# Modern NGS Sequencing Platforms

|                                   | 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  |



University of Michigan  
Medical School

# Illumina – Reversible terminators



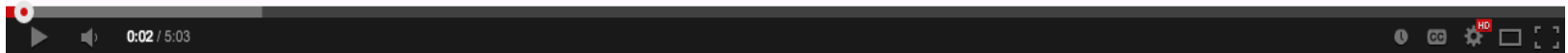
(other sequencing platforms summarized at end of slide set)



University of Michigan  
Medical School

# Illumina Sequencing - Video

illumina®

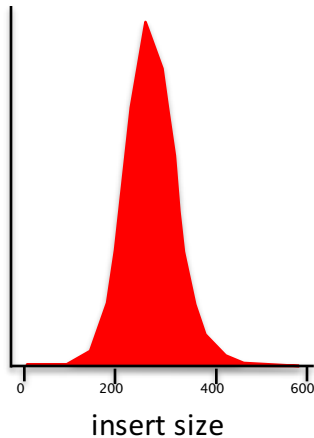
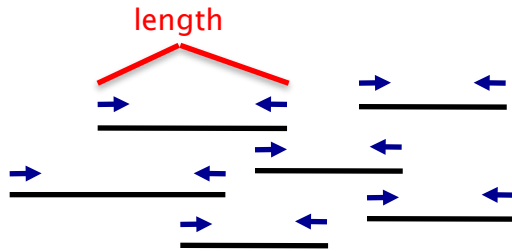




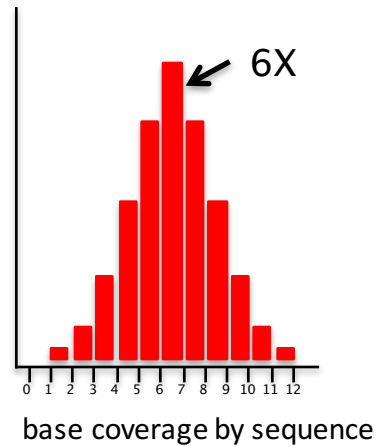
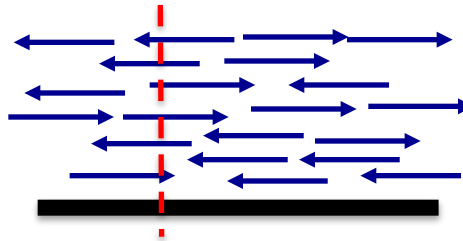
University of Michigan  
Medical School

# NGS Sequencing Terminology

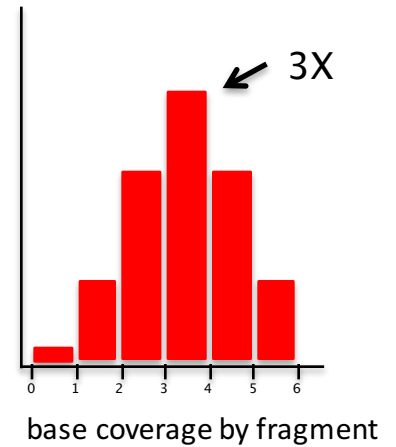
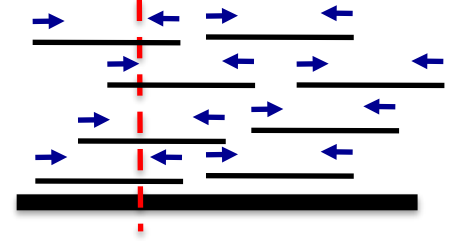
Insert Size



Sequence Coverage



Physical Coverage





# Third Generation Sequencing

---

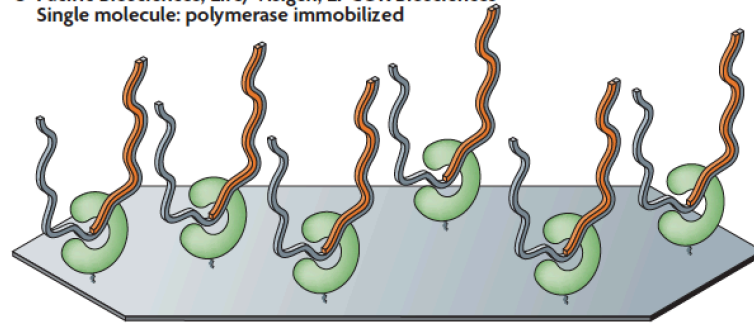
- Currently in transition / development
- Hard to define what “3<sup>rd</sup>” generation means
- Typical characteristics:
  - Long (1,000bp+) sequence reads
  - Single molecule (no amplification step)
  - Often associated with nanopore technology
    - But not necessarily!



University of Michigan  
Medical School

# Pacific Biosystems – Real Time Sequencing

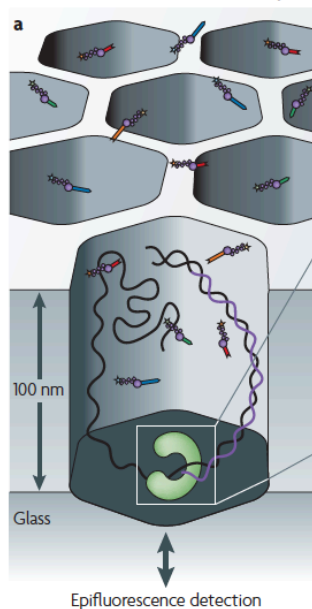
e Pacific Biosciences, Life/Visigen, LI-COR Biosciences  
Single molecule: polymerase immobilized



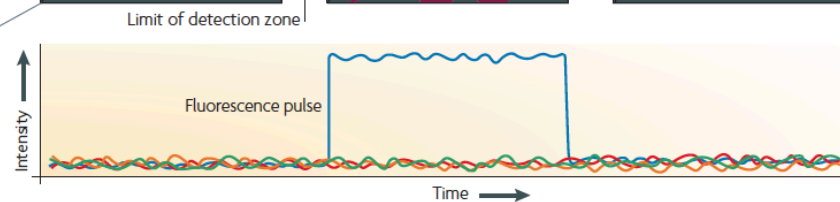
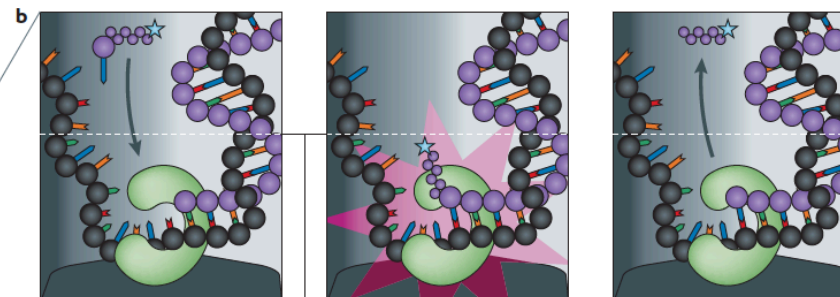
Thousands of primed, single-molecule templates



Pacific Biosciences — Real-time sequencing



Phospholinked hexaphosphate nucleotides



zero mode  
waveguides

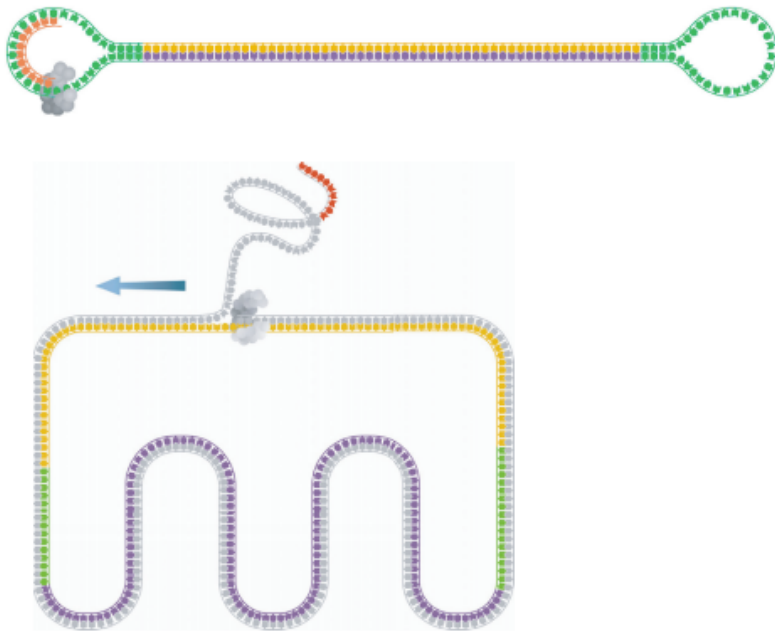




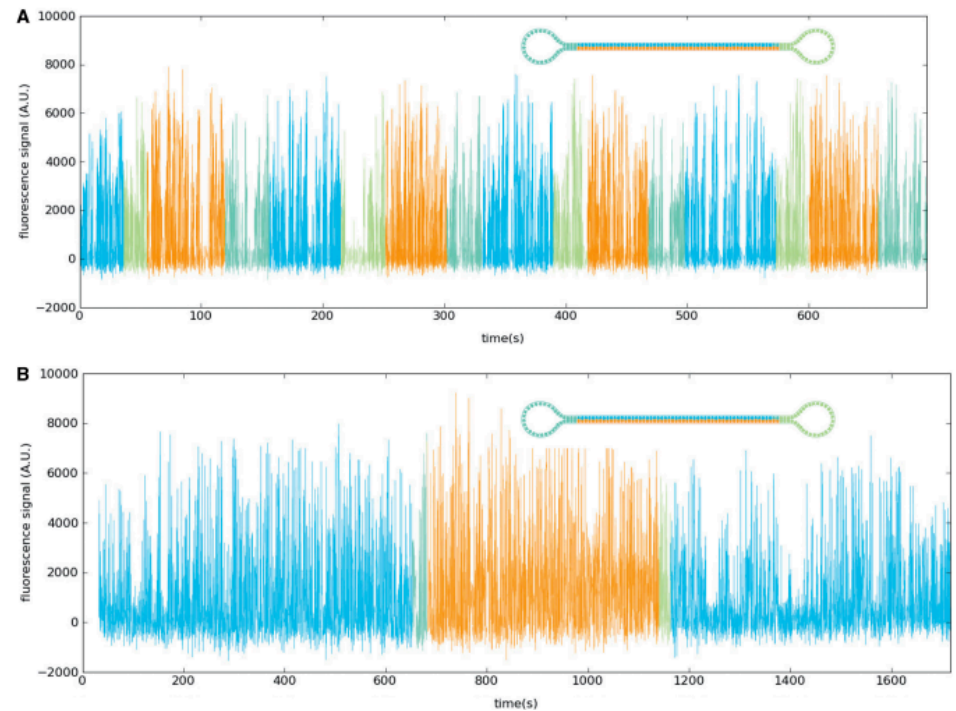
University of Michigan  
Medical School

# Pacific Biosystems – Circular Consensus

SMRTbell template



Subread Consensus Sequencing








# Summary: Generations of DNA Sequencing

|  | First generation  | Second generation <sup>a</sup>   | Third generation <sup>a</sup>  |
|--|---|--|--|
| 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<br>Low cost per run  | Low cost per base<br>High cost per run   | Low-to-moderate cost per base<br>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  |



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



SEQanswers  
Forums

wiki navigation  
Main page  
Recent changes  
Random page  
Help

Software  
Software hub  
Browse software  
Software list

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

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 |
|--|--|--|--|--|--|---|--|
| <a href="#">4peaks</a>   | Allows viewing sequencing trace files, motif searching trimming, BLAST and exporting sequences.  | <a href="#">Sequencing</a>   | Sequence analysis  |  |  | Freeware  | Mac OS X   |
| <a href="#">AB Large Indel Tool</a>  | Identifies deviations in clone insert size that indicate intra-chromosomal structural variations compared to a reference genome.                                 | <a href="#">InDel discovery Sequencing</a>   | Mapping  |  | Perl   | GPL   | Linux 64   |
| <a href="#">AB Small Indel Tool</a>  | The SOLiD™ Small Indel Tool processes the indel evidences found in the pairing step of the SOLiD™ System Analysis Pipeline Tool (Corona Lite).                   | <a href="#">InDel discovery Sequencing</a>   | Mapping Alignment  |  | Perl<br>C++  | GPL   | Linux 64   |
| <a href="#">ABBA</a>   | 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 | <a href="#">Genomic Assembly</a>   | Assembly Scaffolding   |  |  | Artistic License  | Linux  |
| <a href="#">ABMapper</a>   | Maps RNA-Seq reads to target genome considering possible multiple mapping locations and splice junctions   | <a href="#">Genomics Transcriptomics</a>   | Mapping Alignment  |  | C++<br>Perl  | GPLv3   | Linux  |
| <a href="#">ABYSS</a>  | ABYSS is a de novo sequence assembler designed for short reads and large genomes.  | <a href="#">De-novo assembly</a>   | Assembly<br>De Bruijn graph  | MPI<br>OpenMP  | C++  | Free for academic use   | POSIX<br>Linux<br>Mac OS X   |
| <a href="#">Arlanter Removal</a>   | Removes arlanter fragments from raw short read   | <a href="#">General</a>  | Arlanter Removal   | Trimming   | Java   | Custom Licence  | Linux 64   |



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

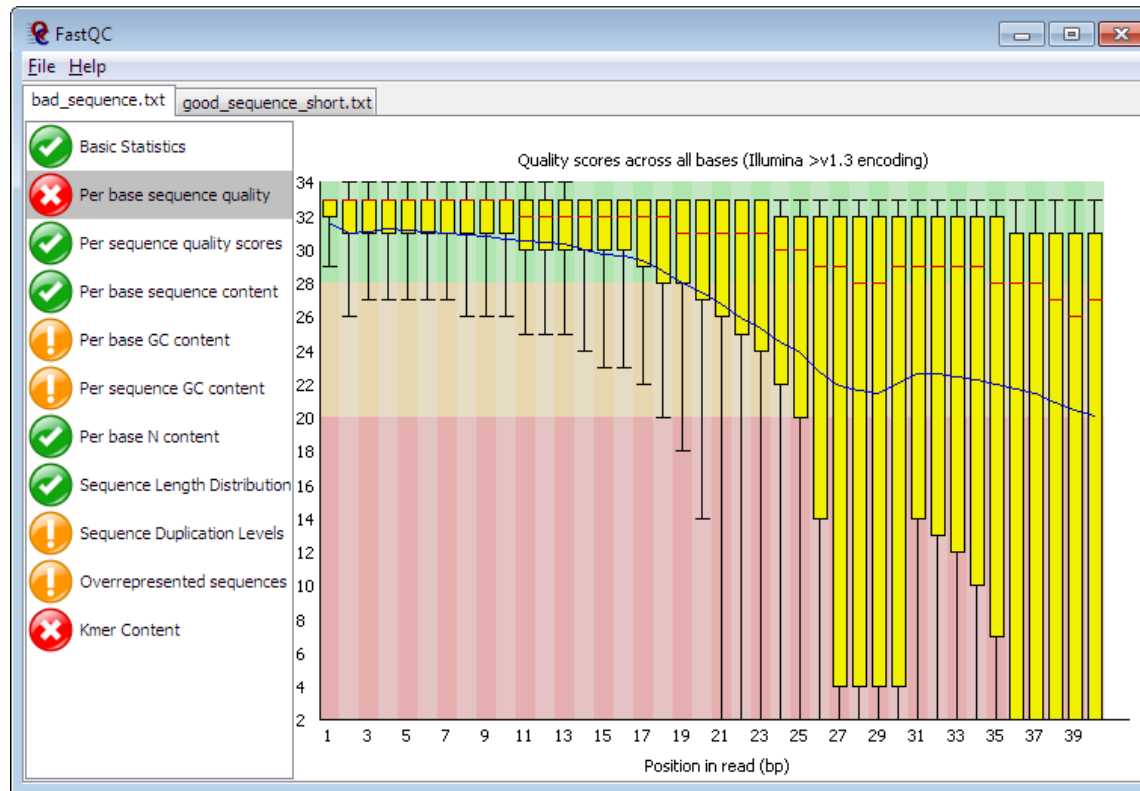




- Quality checks of raw sequence data are *very* important
  - Common problems can include:
    - Sample mix-up
    - Sample contamination
    - Machine interruption
    - DNA quality
  - It is crucial that investigators examine their sequences upon first receipt before any downstream analysis is conducted
-

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

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





# Sequence Alignment

---

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

BWA

Bowtie

SOAP2

Novoalign

mr/mrsFast

Eland

Blat

Bfast

BarraCUDA

CASHx

GSNAP

Mosiak

Stampy

SHRiMP

SeqMap

SLIDER

RMAP

SSAHA

etc



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



## Header section

## Alignment section

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





- 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)
-



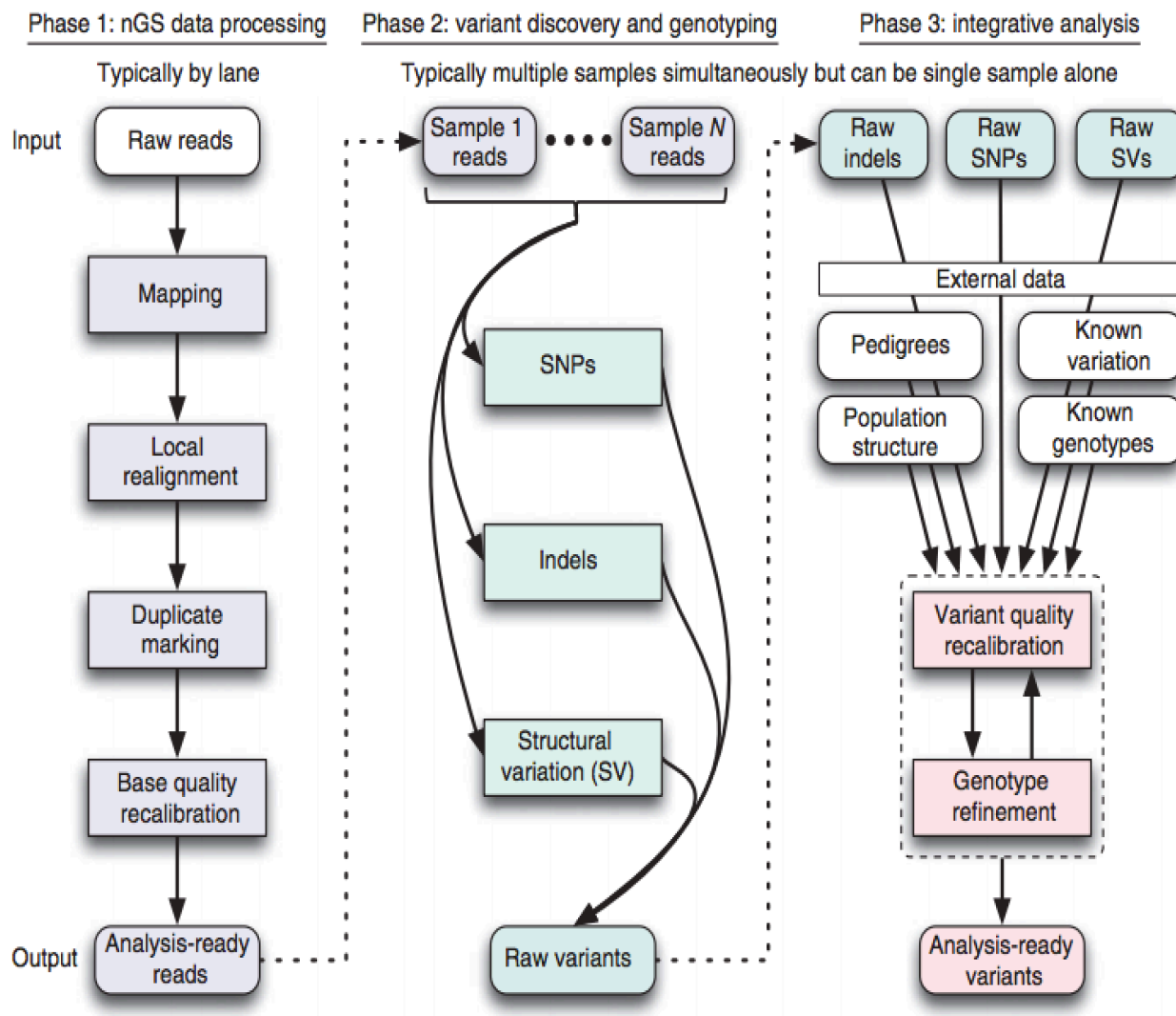
- A lot of research has been conducted to improve and optimize sequence alignments
  - However, genomic sequences are very complex and by their very nature can preclude the ability to accurately determine where a sequence read originated
  - New tools and approaches have been developed to help address these shortcomings and improve our overall ability to interpret the alignments
-



# Genome Analysis Toolkit (GATK)

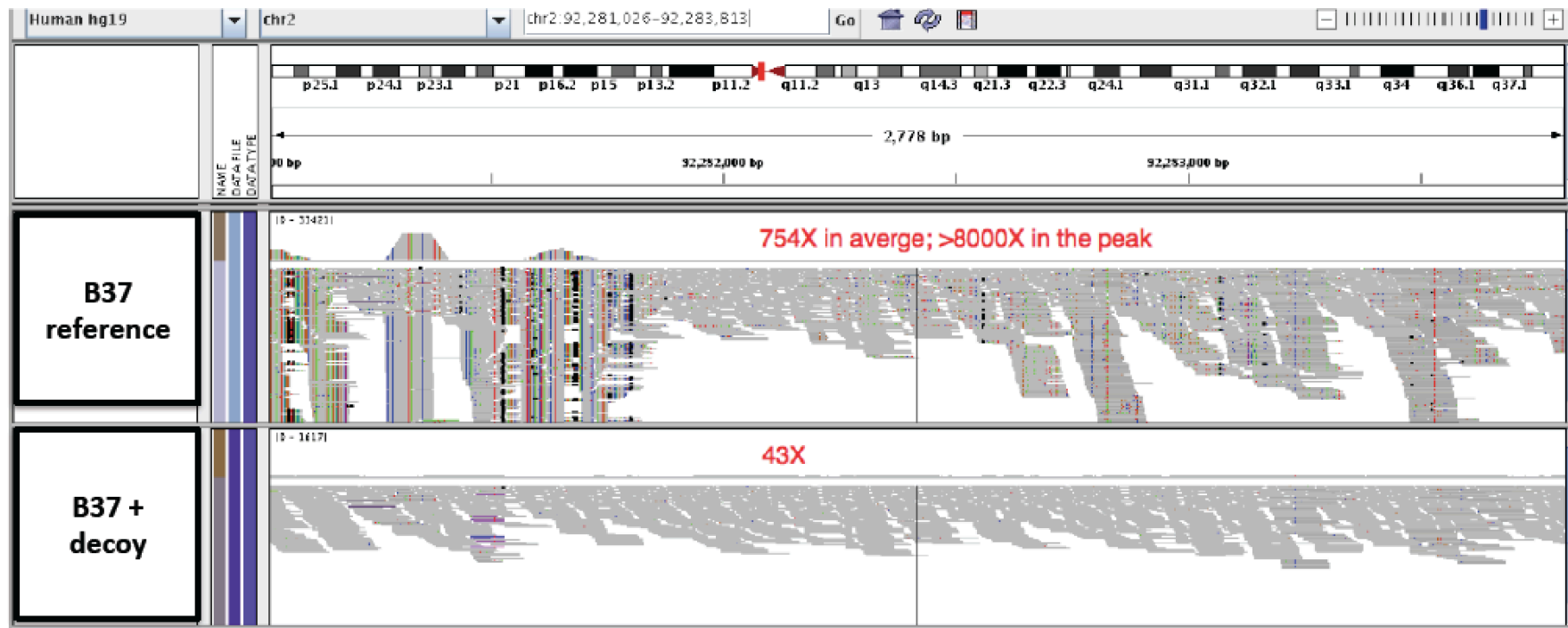
---

- Developed in part to aid in the analysis of 1000 Genomes Project data
- Includes many tools for manipulating, filtering, and utilizing next generation sequence data
- <http://www.broadinstitute.org/gatk/>





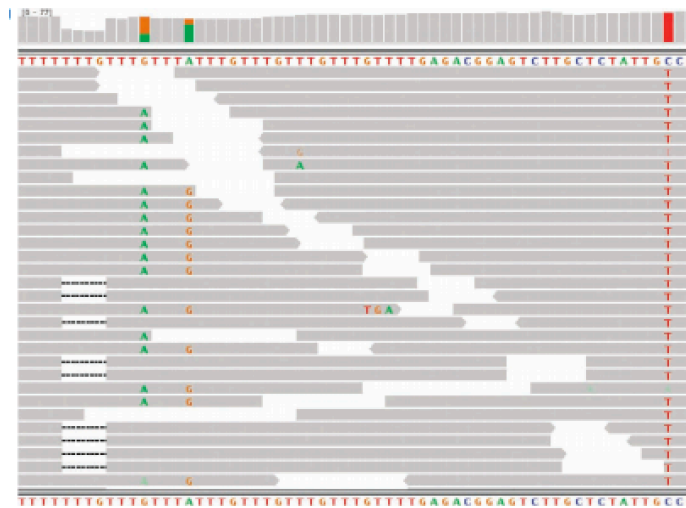
One approach is to allow reads to map to a “decoy” alignment of extra-chromosomal or unassembled sequences





# Realignment around INDELs

- Insertions and deletions in samples can cause misalignments, resulting in false variant detection
- By identifying regions with known INDELs or reads which may have INDEL characteristics and performing multiple sequence alignments, these alignments can be rescued



HiSeq data, raw BWA alignments



HiSeq data, after MSA



# Marking/Removing Duplicate Sequences

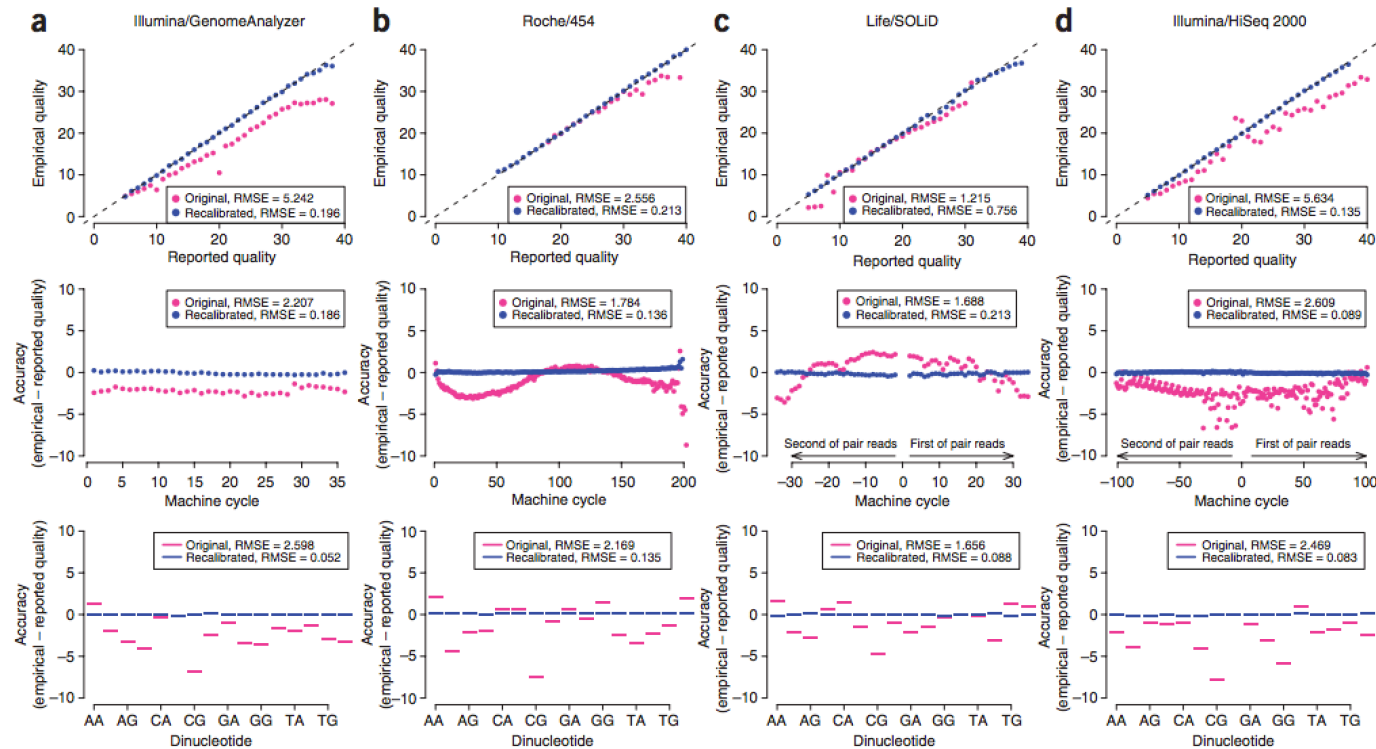
---

- Sequence biases can arise from PCR amplification effects during the construction of the library
  - There can also be optical duplicates which occur when sequences from one cluster are accidentally identified as arising as well from adjacent clusters
  - Both Picard (MarkDuplicates) and Samtools (rmdup) have utilities for addressing one or both of these issues
-



# Base Recalibration

- Provides empirically accurate base quality scores for each base in every read
- Also corrects for error covariates like machine cycle and dinucleotide content







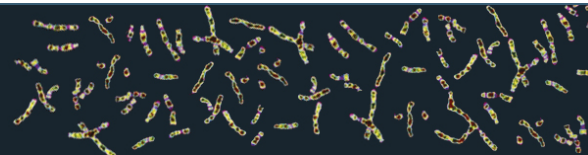
University of Michigan  
Medical School

# 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



“Variety’s the very spice of life”

–William Cowper, 1785

“Variation is the spice of life”

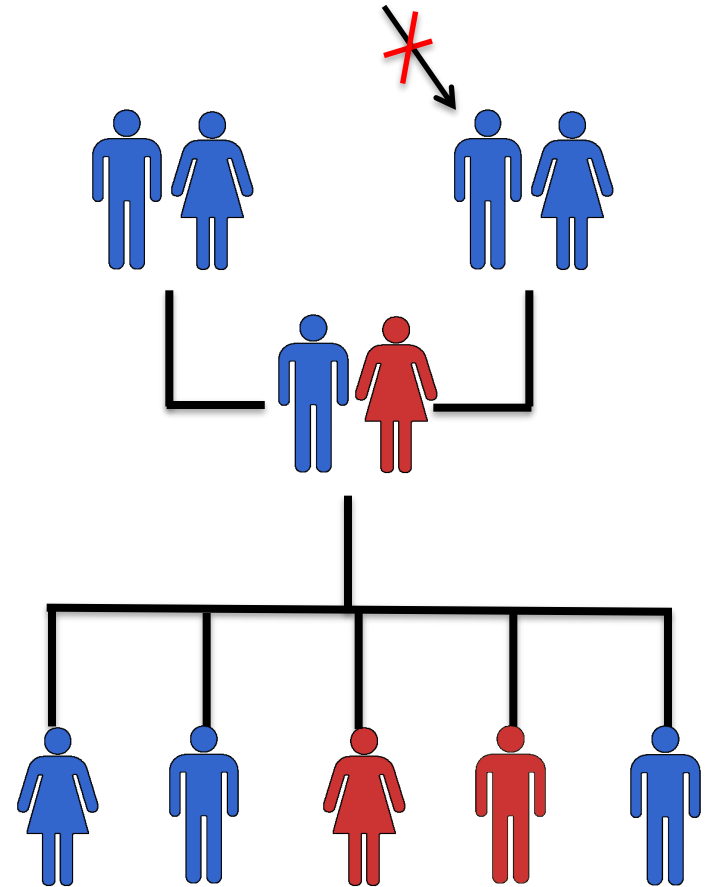
–Kruglyak & Nickerson, 2001

- 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.
-



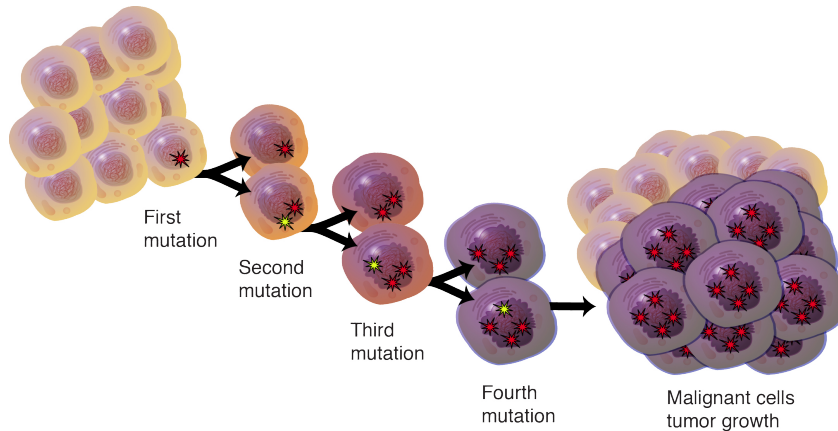
# Germline Variation

- 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 simple those alleles whose frequencies within a species happen to be in a low-frequency range  $(0, q)$ , whereas polymorphic alleles are those whose frequencies happen to be in the higher-frequency range  $(q, 1-q)$ , where I arbitrarily take  $q = 0.01$ . Both represent a phase of molecular evolution.”*


*-Motoo Kimura*



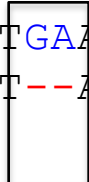
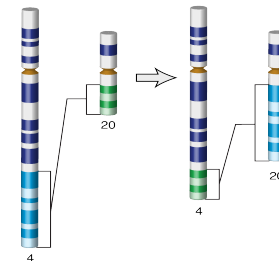
# Types of Genomic Variation

- 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

A vertical rectangle highlights the single nucleotide difference between the two DNA sequences: a 'G' in the top sequence and a 'C' in the bottom sequence.

AATCTGAAGGCAT  
AATCT--AGGCAT

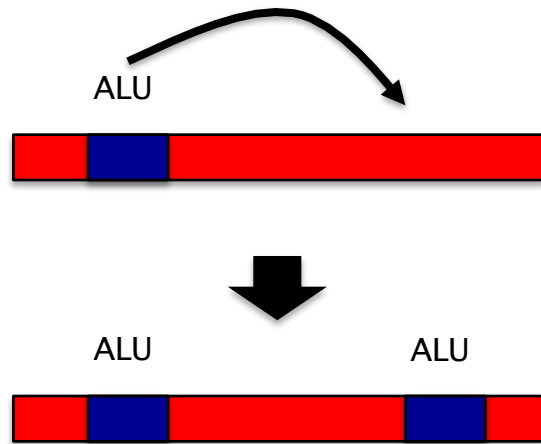
A vertical rectangle highlights the difference between the two DNA sequences: the top sequence has 'GA' and the bottom sequence has two dashes '--'.



University of Michigan  
Medical School

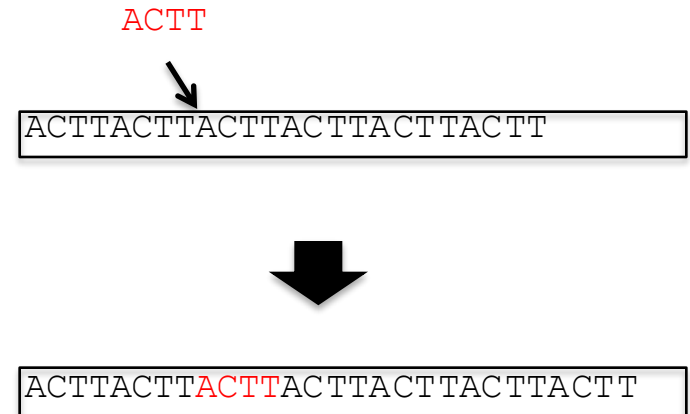
# Variant Subtypes: Repetitive Elements

## Mobile Elements / Retrotransposons



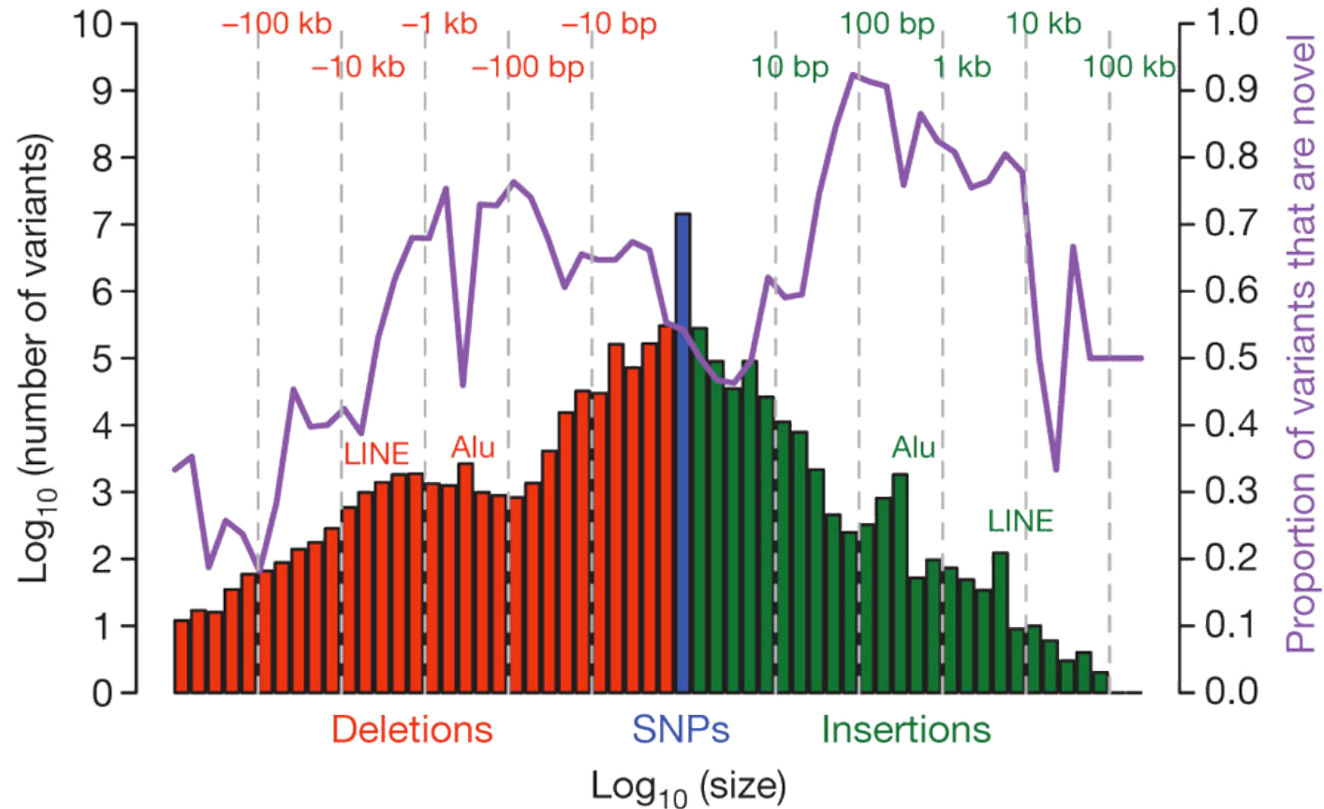
(in humans, primarily ALU, LINE, and SVA)

## Repeat Expansions





# Variant Length Distribution







# 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

ATCCTGATTCCGGTGAACGTTATCGACGATCCGATCGA  
ATCCTGATTCCGGTGAACGTTATCGACGATCCGATCGA  
CGGTGAACGTTATCGACGATCCGATCGAACTGTCAGC  
GGTGAACGTTATCGACGTTCCGATCGAACTGTCAGCG  
TGAACGTTATCGACGTTCCGATCGAACTGTCATCGGC  
TGAACGTTATCGACGTTCCGATCGAACTGTCAGCGGC  
TGAACGTTATCGACGTTCCGATCGAACTGTCAGCGGC  
GTTATCGACGATCCGATCGAACTGTCAGCGGCAAGCT  
TTATCGACGATCCGATCGAACTGTCAGCGGCAAGCT

sequencing error  
or genetic variant?

**ATCCTGATTCCGGTGAACGTTATCGACGATCCGATCGAACTGTCAGCGGCAAGCTGATCGATCGATCGATGCTAGTG**

reference genome

TTATCGACGATCCGATCGAACTGTCAGCGGCAAGCT  
TCGACGATCCGATCGAACTGTCAGCGGCAAGCTGAT  
ATCCGATCGAACTGTCAGCGGCAAGCTGATCGCGAT  
TCCGAGCGAACTGTCAGCGGCAAGCTGATCGCGATC  
TCCGATCGAACTGTCAGCGGCAAGCTGATCGATCGA  
GATCGAACTGTCAGCGGCAAGCTGATCGCGATCGA  
AACTGTCAGCGGCAAGCTGATCGCGATCGATGCTA  
TGTCAGCGGCAAGCTGATCGATCGATCGATGCTAG  
TCAGCGGCAAGCTGATCGATCGATCGATGCTAGTG

sequencing error  
or genetic variant?

INDEL



# Genotyping Small Variants

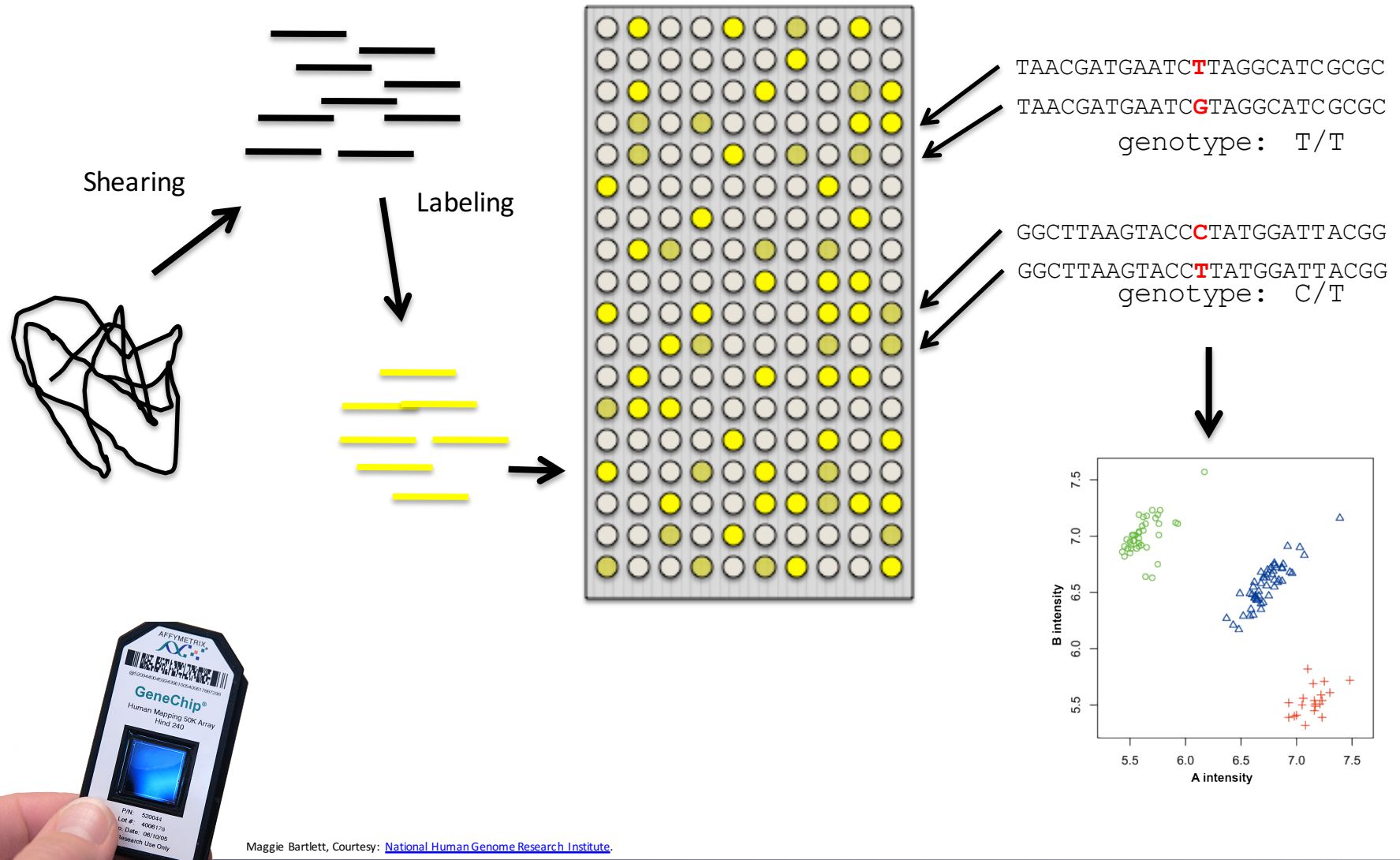
---

- 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



University of Michigan  
Medical School

# SNP Microarrays



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



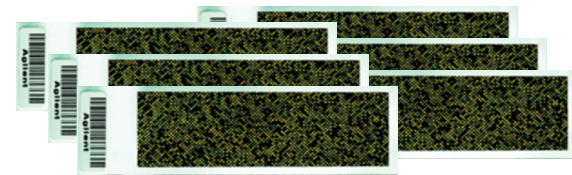
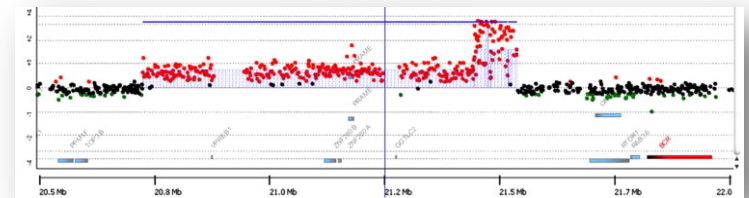
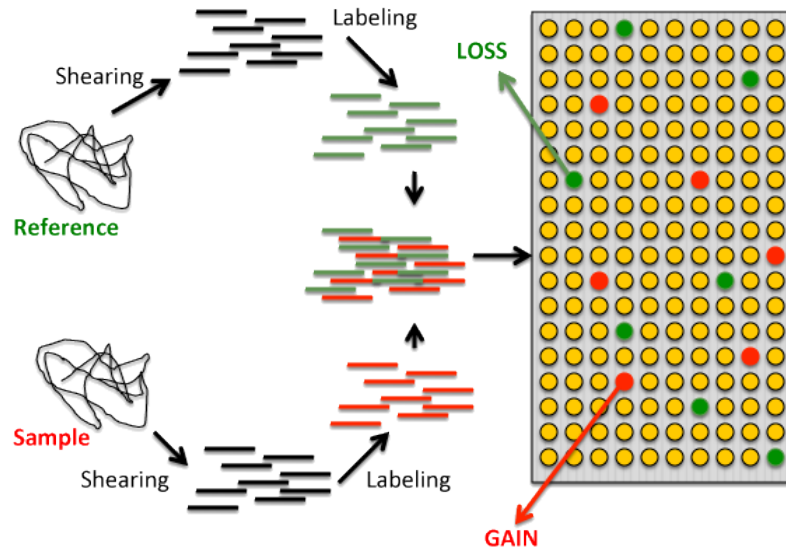
- Structural variants can be discovered by both sequence and microarray approaches
  - Microarrays can only detect genomic imbalances, specifically copy number variants (CNVs)
  - Sequence based approaches can, in principle, identify all types of structural rearrangements
-



University of Michigan  
Medical School

# Microarray-based CNV Discovery

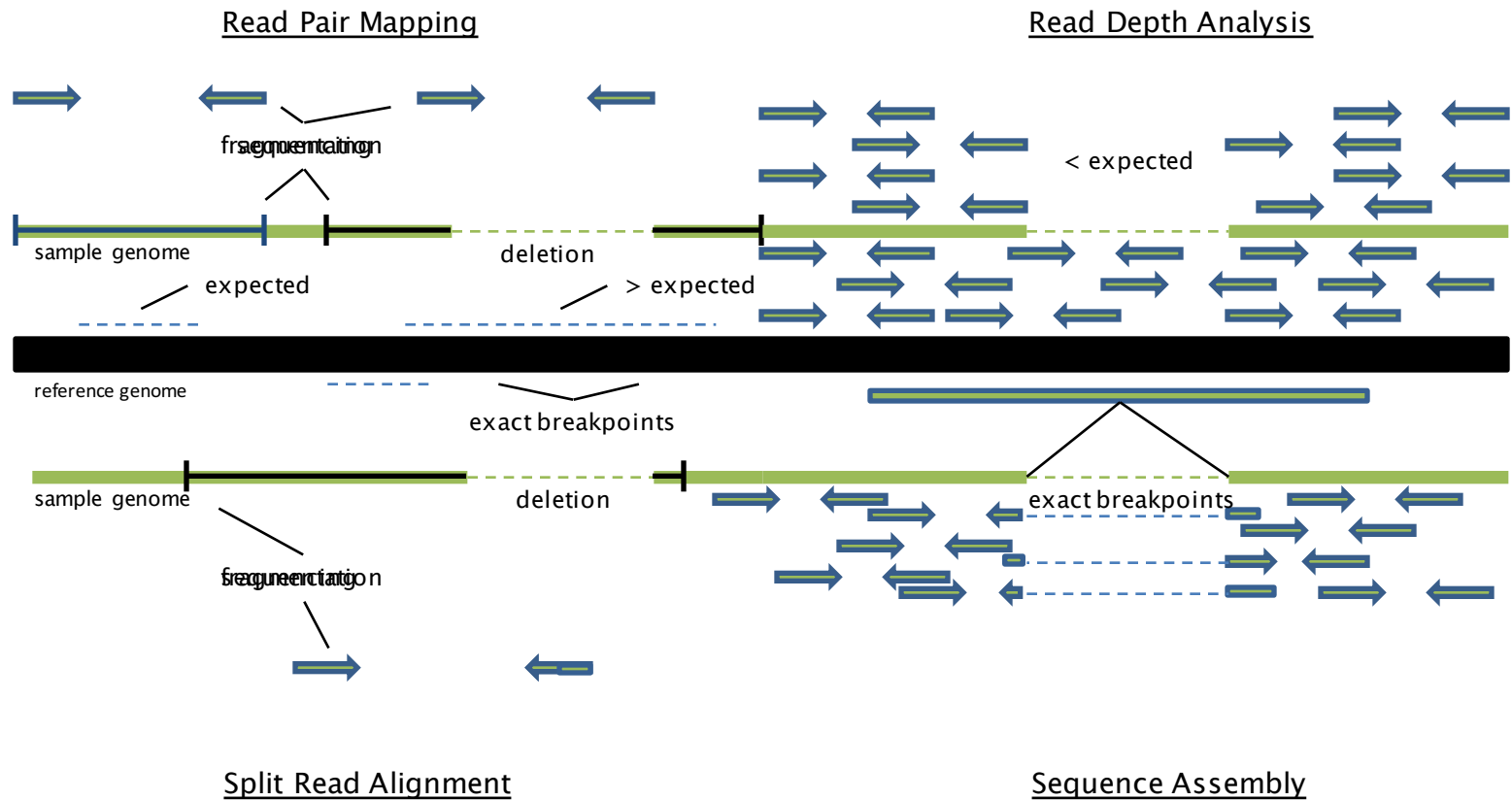
## Comparative Genomic Hybridization (CGH)





University of Michigan  
Medical School

# Sequenced-based SV Discovery







# Variant Databases and Formats

---

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



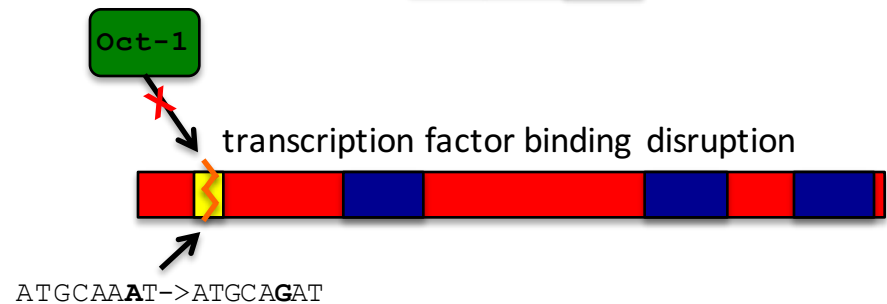
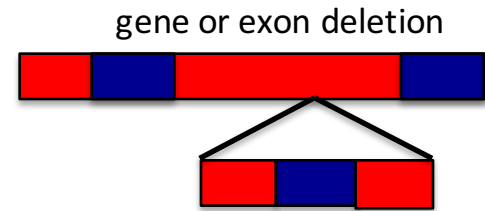
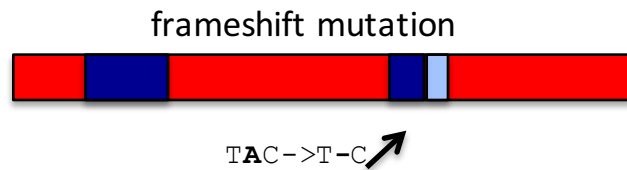
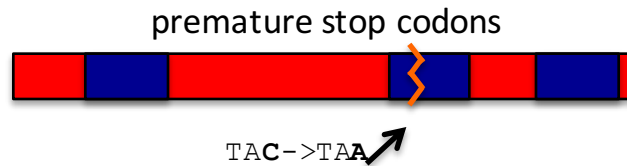
# VCF Format Example

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



# Impact of Genetic Variation

There are numerous ways genetic variation can exhibit functional effects





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



# Variant Annotation Classes

## High Impact

- exon\_deleted
- frame\_shift
- splice\_acceptor
- splice\_donor
- start\_loss
- stop\_gain
- stop\_loss
- non\_synonymous\_start
- transcript\_codon\_change

## Medium Impact

- non\_syn\_coding
- inframe\_codon\_gain
- inframe\_codon\_loss
- inframe\_codon\_change
- codon\_change\_del
- codon\_change\_ins
- UTR\_5\_del
- UTR\_3\_del
- other\_splice\_variant
- mature\_miRNA
- regulatory\_region
- TF\_binding\_site
- regulatory\_region\_ablation
- regulatory\_region\_amplification
- TFBS\_ablation
- TFBS\_amplification

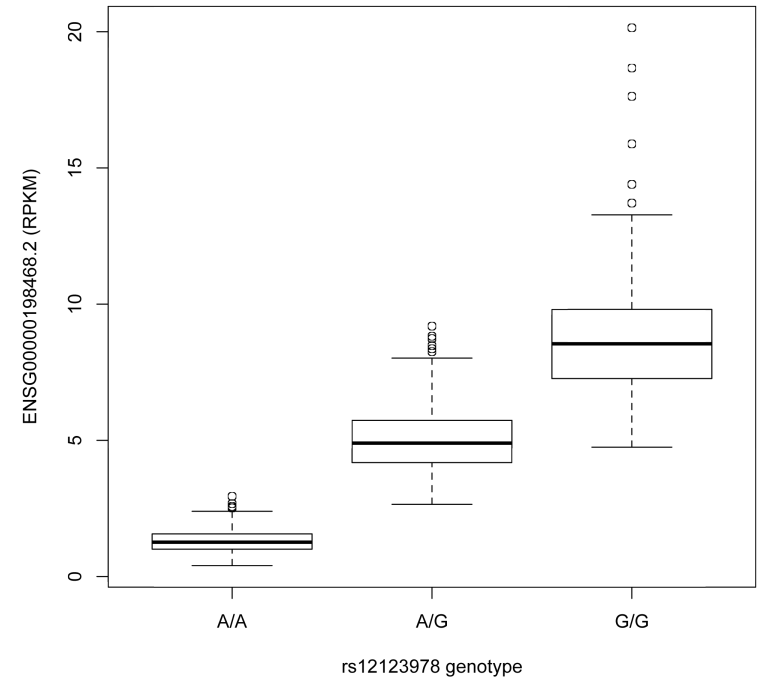
## Low Impact

- synonymous\_stop
- synonymous\_coding
- UTR\_5\_prime
- UTR\_3\_prime
- intron
- CDS
- upstream
- downstream
- intergenic
- intragenic
- gene
- transcript
- exon
- start\_gain
- synonymous\_start
- intron\_conserved
- nc\_transcript
- NMD\_transcript
- transcript\_codon\_change
- incomplete\_terminal\_codon
- nc\_exon
- transcript\_ablation
- transcript\_amplification
- feature\_elongation
- feature\_truncation



# Variation and Gene Expression

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





University of Michigan  
Medical School

# 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

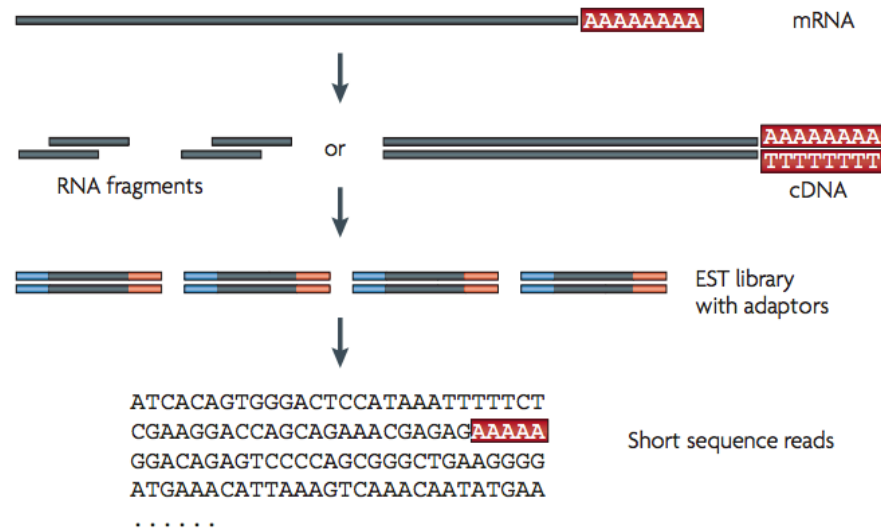
**GEUVADIS Genetic European Variation in Health and Disease, A European Medical Sequencing Consortium**  
  
**GEUVADIS RNA sequencing project for 1000 Genomes samples**  
  
  
  
**Welcome !**  
  
[Welcome to the GEUVADIS website](#)  
  
We are committed to gaining insights into the **human genome** 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 (GM<sup>2</sup>)**  
**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 Archon 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



- Uses same technologies as DNA sequencing
- Primary difference is in library preparation
  - RNA converted to cDNA through reverse transcription

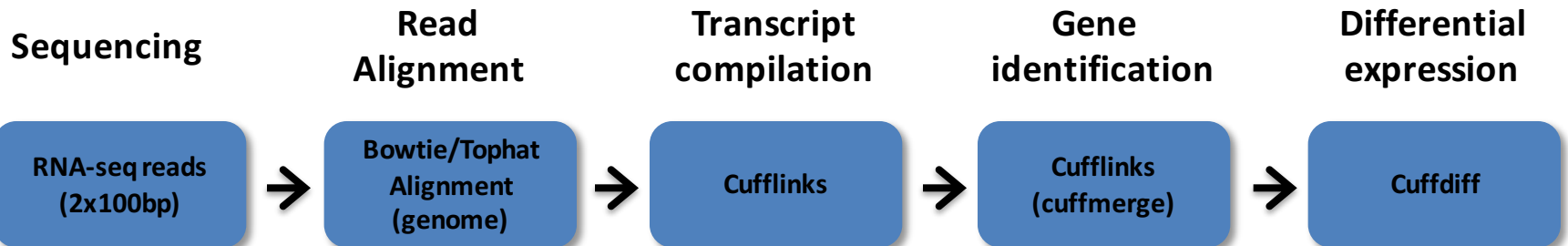
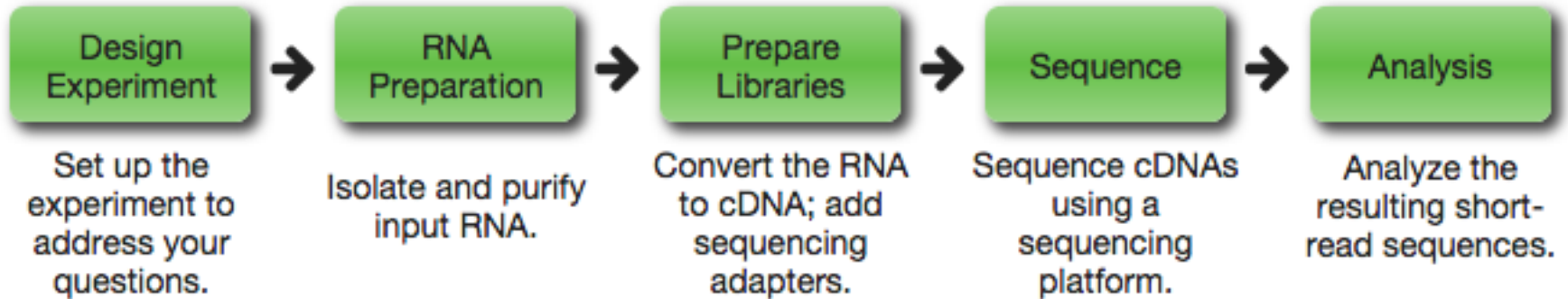






# RNA-Seq Overview

## Sample Preparation



## Analysis Pipeline (Tophat)



# Types of RNA-Seq Libraries

---

- poly(A) capture – utilizes oligo(dT) to prime off of mature mRNA
    - Won't amplify non-coding mRNA (e.g. lincRNAs, miRNAs, etc)
    - Has enrichment biases between 3' and 5' ends due to RT drop-offs
    - Won't work with fragmented RNA
  - random hexamer priming – primes at random positions along the transcript
    - Will work with fragmented or degraded RNA (e.g. FFPE samples)
    - Removes positional biases of poly(A) capture
    - Requires some type of rRNA removal (e.g. Ribo-Zero) to address its overabundance
-



University of Michigan  
Medical School

# DNA- and RNA-Seq Databases

NCBI Short Read Archive (SRA):

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

NCBI

Resources

How To

remills@orcid

My NCBI

Sign Out


SRA

SRA

Search

Advanced

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

### Getting Started

- [Understanding and Using SRA](#)
- [How to Submit](#)
- [Login to Submit](#)
- [Download Guide](#)

### Tools and Software

- [Download SRA Toolkit](#)
- [SRA Toolkit Documentation](#)
- [SRA-BLAST](#)
- [SRA Run Browser](#)
- [SRA Run Selector](#)

### Related Resources

- [dbGaP Home](#)
- [Trace Archive Home](#)
- [BioSample](#)
- [GenBank Home](#)

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

Write to the Help Desk

#### GETTING STARTED

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

#### RESOURCES

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

#### POPULAR

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

#### FEATURED


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

#### NCBI INFORMATION

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

Copyright | Disclaimer | Privacy | Browsers | Accessibility | Contact

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





University of Michigan  
Medical School

# Protected Data - dbGaP

## NCBI Database of Genotypes and Phenotypes (dbGaP):

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

NCBI

Resources

How To

remills@orcid

My NCBI

Sign Out

dbGaP

dbGaP

Limits

Advanced

Search

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 Integrator](#)

#### Important Links

- [Summary Statistics](#)
- [dbGaP RSS Feed](#)
- [Code of Conduct](#)
- [Security Procedures](#)
- [Contact Us](#)

#### Latest Studies

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

| Study  | Embargo Release   | Details   | Participants | Type Of Study                                   | Links                 | Platform   |
|--|---|---|--------------|---|-----------------------|--|
| <a href="#">phs000790.v1.p1</a><br>Comparative Analysis of Primary and Metastatic Colorectal Cancer      | Version 1: 2015-01-29   | <a href="#">V</a> <a href="#">D</a> <a href="#">A</a> <a href="#">S</a> | 4            | Cohort  | <a href="#">Links</a> | HiSeq 2000   |
| <a href="#">phs000848.v1.p1</a><br>Autosomal recessive TPP2 mutations cause a new human immunodeficiency | Version 1: 2015-12-16   | <a href="#">V</a> <a href="#">D</a> <a href="#">A</a> <a href="#">S</a> | 3            | Case-Control                                    | <a href="#">Links</a> | Genome Analyzer IIX  |
| <a href="#">phs000842.v1.p1</a><br>PedIGFR   | Version 1: passed embargo   | <a href="#">V</a> <a href="#">D</a> <a href="#">A</a> <a href="#">S</a> | 1572         | Multicenter, Prospective, Observational, Cohort | <a href="#">Links</a> | HumanOmni2.5-Quad  |
| <a href="#">phs000007.v25.p9</a><br>Framingham Cohort  | Versions 1-22: passed embargo<br>Version 23: 2015-04-25<br>Version 24: 2015-09-25<br>Version 25: 2015-12-23 | <a href="#">V</a> <a href="#">D</a> <a href="#">A</a> <a href="#">S</a> | 15173        | Longitudinal                                    | <a href="#">Links</a> | HuGeneFocused50K_Affy<br>Mapping250K_Nap<br>Mapping250K_Sty<br>Mapping50K_Hind240<br>Mapping50K_Xba240 |
| <a href="#">phs000825.v1.p1</a><br>Whole Genome Sequencing of HUES63 and HUES64                          | Version 1: passed embargo   | <a href="#">V</a> <a href="#">D</a> <a href="#">A</a> <a href="#">S</a> | 2            | Control Set                                     | <a href="#">Links</a> | HiSeq 2000<br>HiSeq 2500   |

[List Top Level Studies](#)

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

Write to the Help Desk



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



Galaxy

Analyze Data Workflow Shared Data Visualization Cloud Help User

Using 0%

Tools

search tools

[Get Data](#)  
[Send Data](#)  
[ENCODE Tools](#)  
[Lift-Over](#)  
[Text Manipulation](#)  
[Convert Formats](#)  
[FASTA manipulation](#)  
[Filter and Sort](#)  
[Join, Subtract and Group](#)  
[Extract Features](#)  
[Fetch Sequences](#)  
[Fetch Alignments](#)  
[Get Genomic Scores](#)  
[Operate on Genomic Intervals](#)  
[Statistics](#)  
[Graph/Display Data](#)  
[Regional Variation](#)  
[Multiple regression](#)  
[Multivariate Analysis](#)  
[Evolution](#)  
[Motif Tools](#)  
[Multiple Alignments](#)  
[Metagenomic analyses](#)  
[Phenotype Association](#)  
[Genome Diversity](#)  
[EMBOSS](#)  
  
[NGS TOOLBOX BETA](#)  
[NGS: QC and manipulation](#)  
[NGS: Mapping](#)  
[NGS: SAM Tools](#)

Built-ins were indexed using default options

Select a reference genome:

Arabidopsis lyrata: Araly1

if your genome of interest is not listed – contact Galaxy team

Is this library mate-paired?:

Single-end

FASTQ file:

Must have ASCII encoded quality scores

Bowtie settings to use:

Commonly used

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

Suppress the header in the output SAM file:

☐

Bowtie produces SAM with several lines of header information by default

Execute

What it does

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

Know what you are doing

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

Input formats

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

A Note on Built-in Reference Genomes

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

History

0 bytes

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



# Additional Slides for Reference

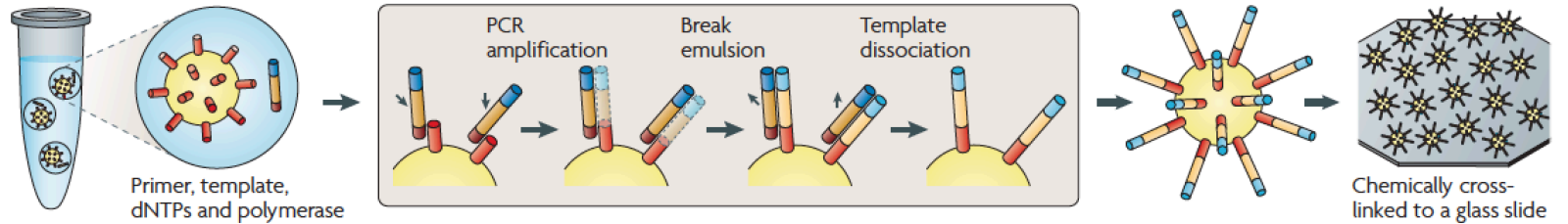
---

University of Michigan  
Medical School

# Roche 454 - Pyrosequencing

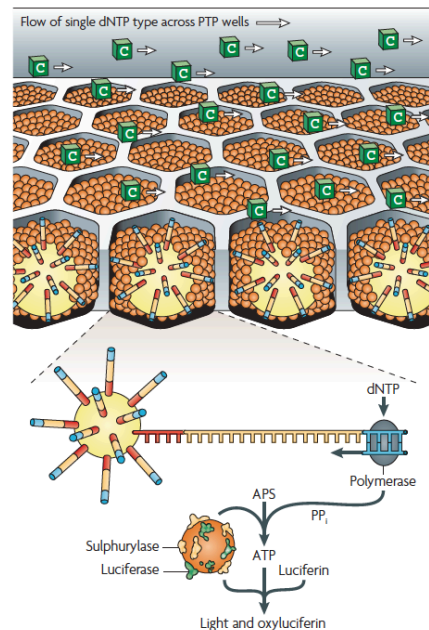
## a Roche/454, Life/APG, Polonator Emulsion PCR

One DNA molecule per bead. Clonal amplification to thousands of copies occurs in microreactors in an emulsion



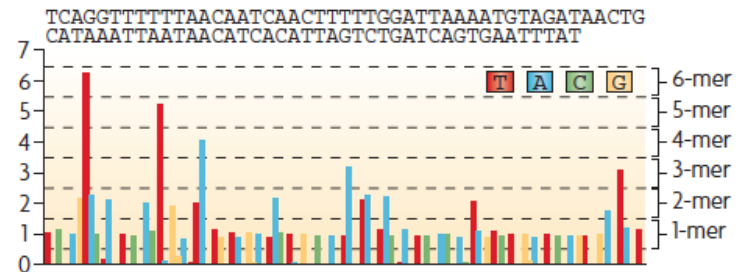
## c Roche/454 — Pyrosequencing

1–2 million template beads loaded into PTP wells



## d

Flowgram





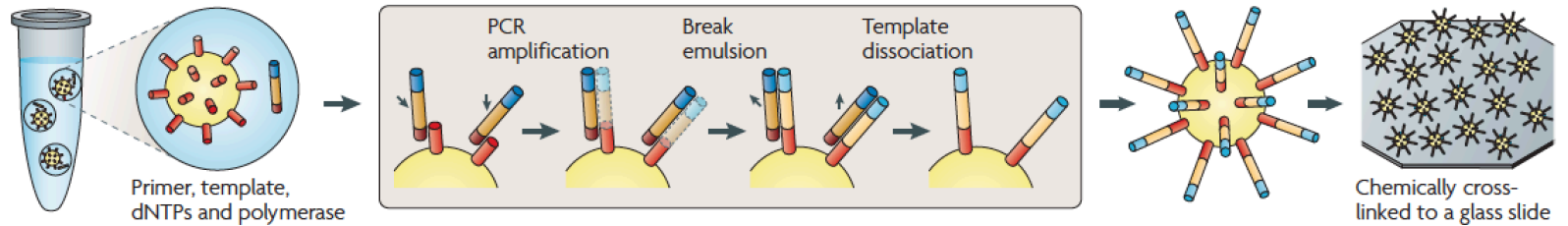


University of Michigan  
Medical School

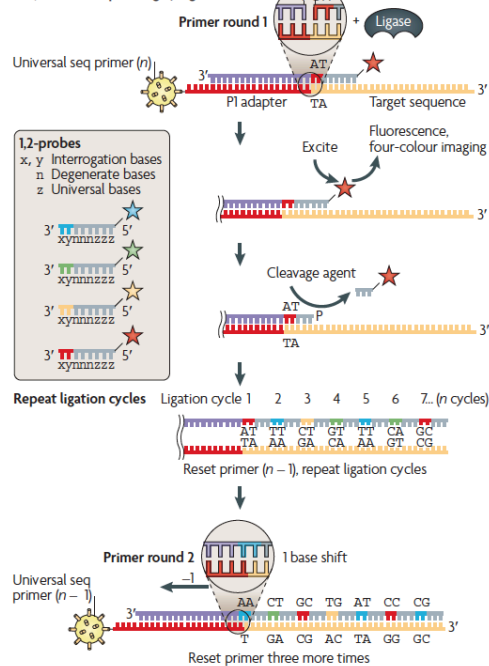
# Life Technologies SOLiD – Sequence by Ligation

## 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

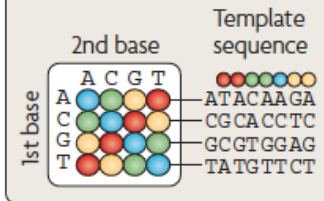


## a Life/APG — Sequencing by ligation

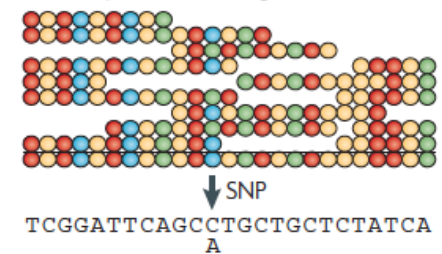


## b

Two-base encoding: each target nucleotide is interrogated twice



Alignment of colour-space reads to colour-space reference genome

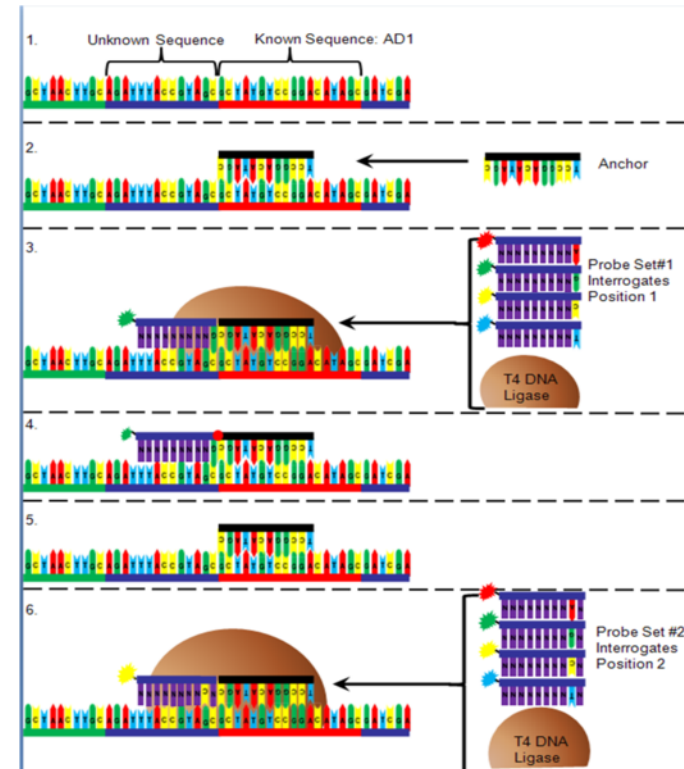
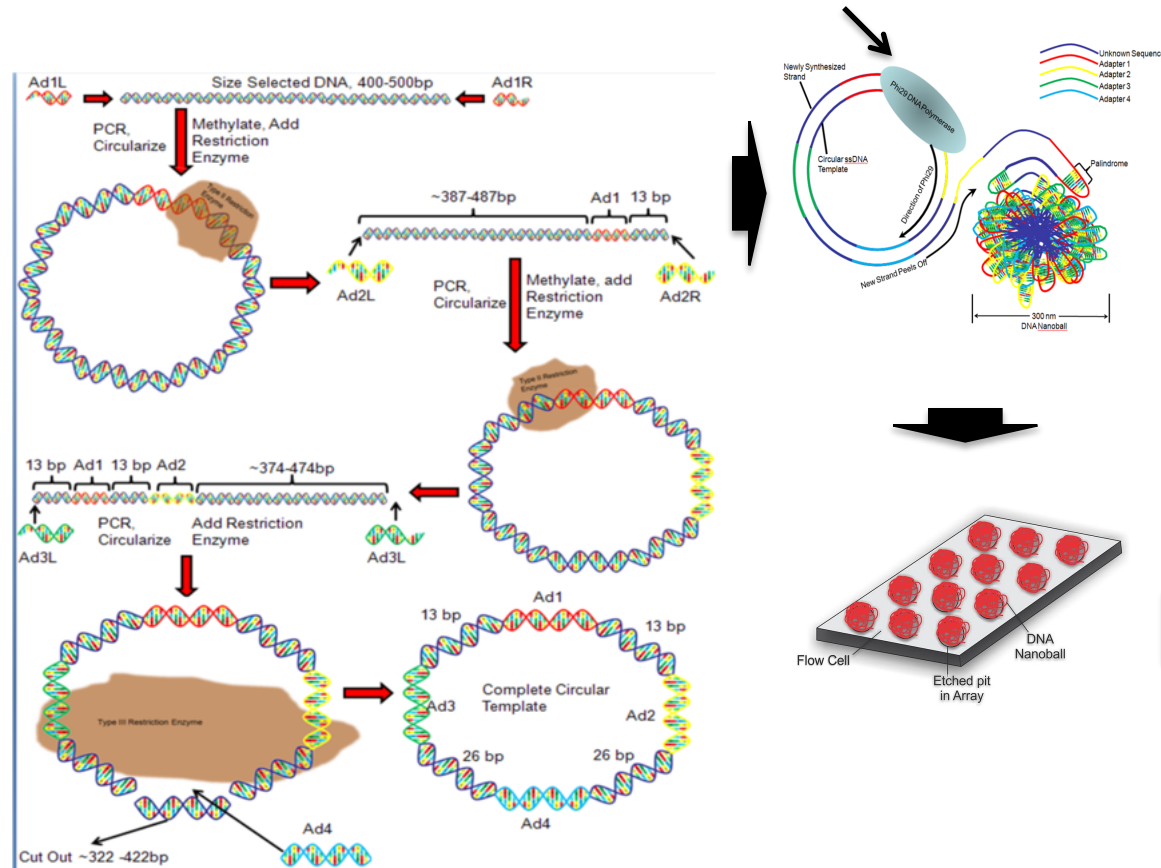




University of Michigan  
Medical School

# Complete Genomics – Nanoball Sequencing

Has proofreading ability!





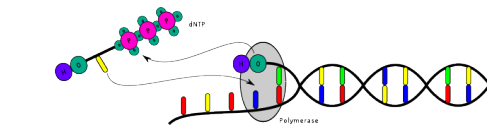
# “Benchtop” Sequencers

- 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

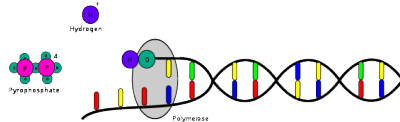
| Platform        | List price              | Approximate cost per run | Minimum throughput (read length) | Run time | Cost/Mb | Mb/h  |
|-----------------|-------------------------|--------------------------|----------------------------------|----------|---------|-------|
| 454 GS Junior   | \$108,000               | \$1,100                  | 35 Mb (400 bases)                | 8 h      | \$31    | 4.4   |
| Ion Torrent PGM |                         |                          |                                  |          |         |       |
| (314 chip)      | \$80,490 <sup>a,b</sup> | \$225 <sup>c</sup>       | 10 Mb (100 bases)                | 3 h      | \$22.5  | 3.3   |
| (316 chip)      |                         | \$425                    | 100 Mb <sup>d</sup> (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  |



# PGM - Ion Semiconductor Sequencing



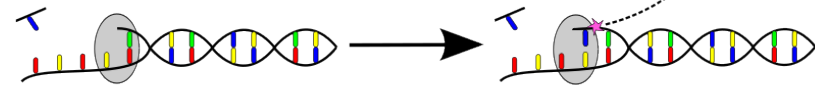
Polymerase integrates a nucleotide.



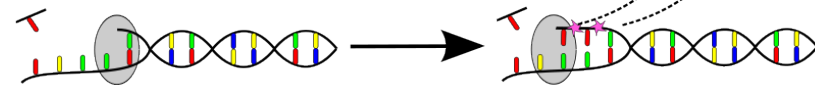
Hydrogen and pyrophosphate are released.



The nucleotide does not compliment the template - no release of hydrogen.



The nucleotide compliments the template - hydrogen is released.



The nucleotide compliments several bases in a row - multiple hydrogen ions are released.

