Bioinformatics: Genomics Part I

Matt Simon
Dept. of Molecular Biophysics & Biochemistry
Chemical Biology Institute
January 14, 2015

What is genomics?

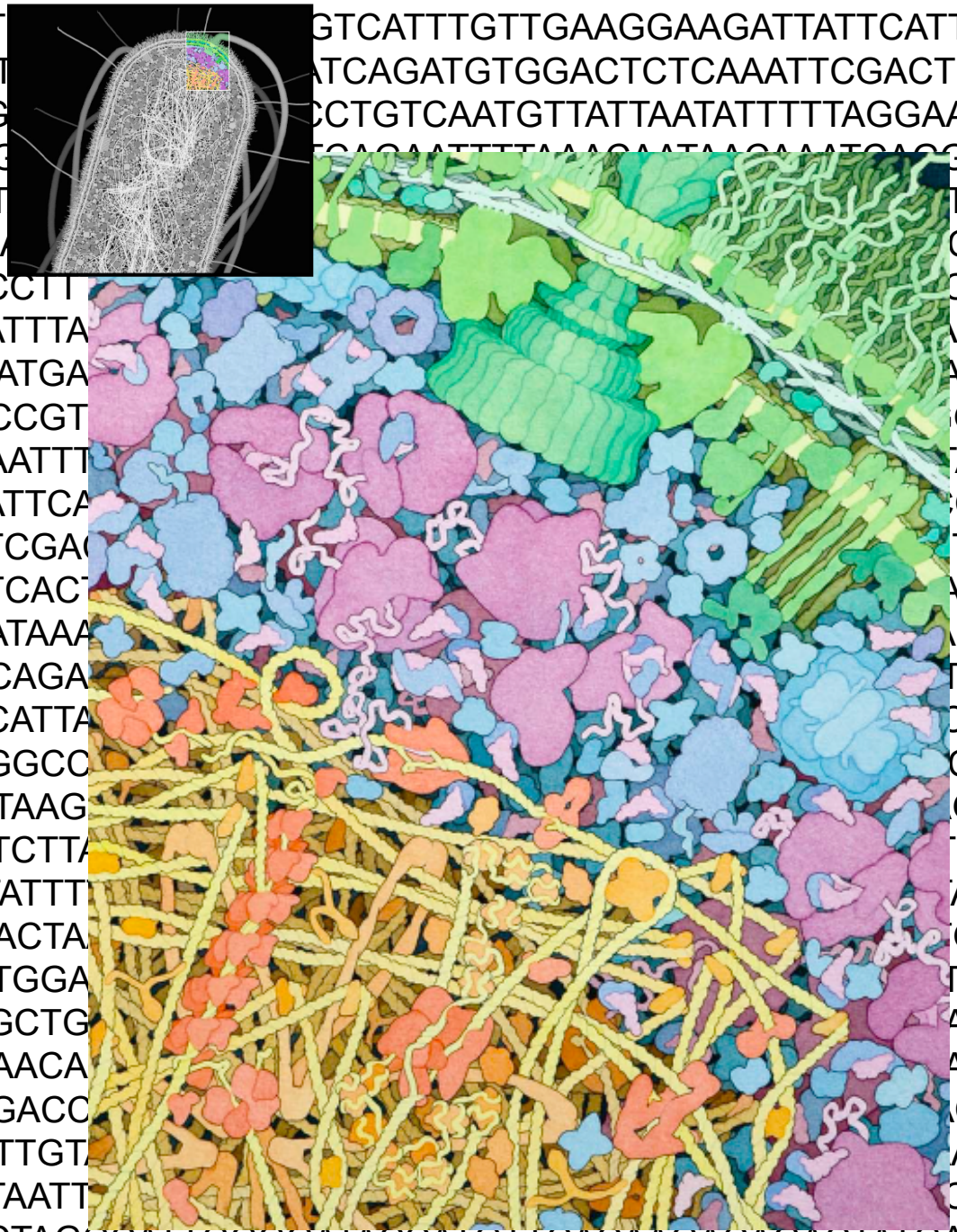
1. The **global** study of how biological **information** is encoded in genome sequence

   Genes
   Regulatory sequences
   Genetic variation

2. How this information is **read out** to produce distinct **biological outcomes**

   Gene expression and regulation
   Cellular identity, differentiation and development
   Phenotypic variation among individuals and species

In practice, many experiments that involve **deep sequencing** are considered genomics.



Overview

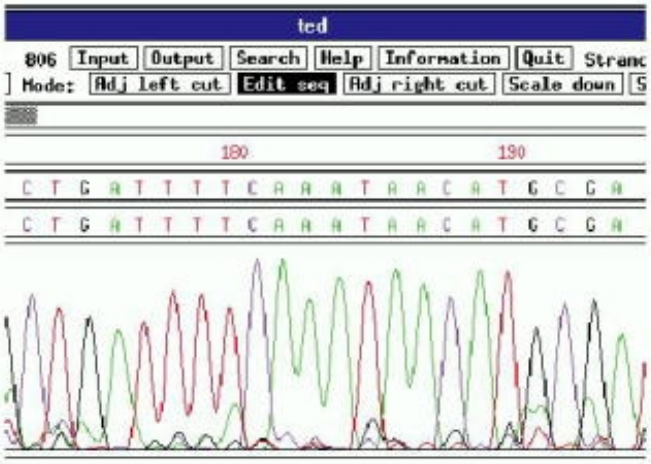
- Genomics I (today's lecture): Focus on sequencing technology and genomes.
- Genomics II: (Friday's lecture): Focus on applications of sequencing technology.

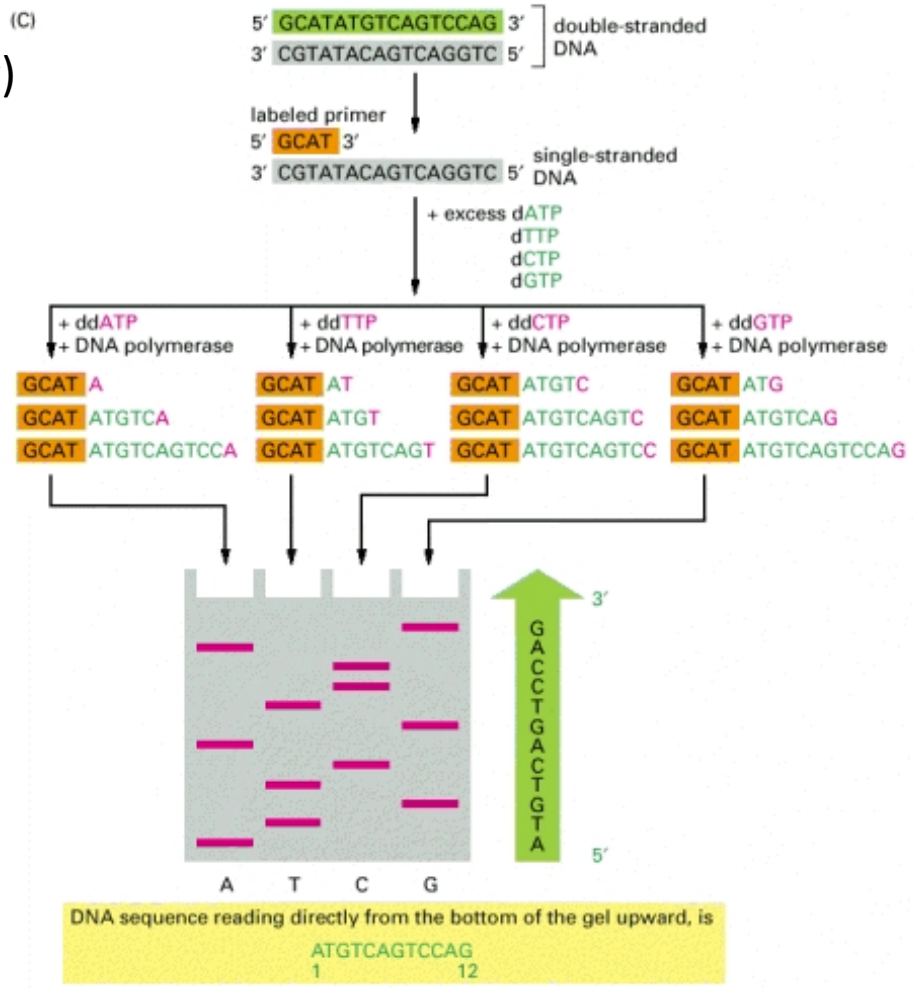
Credit: Jim Noonan for many of the slides

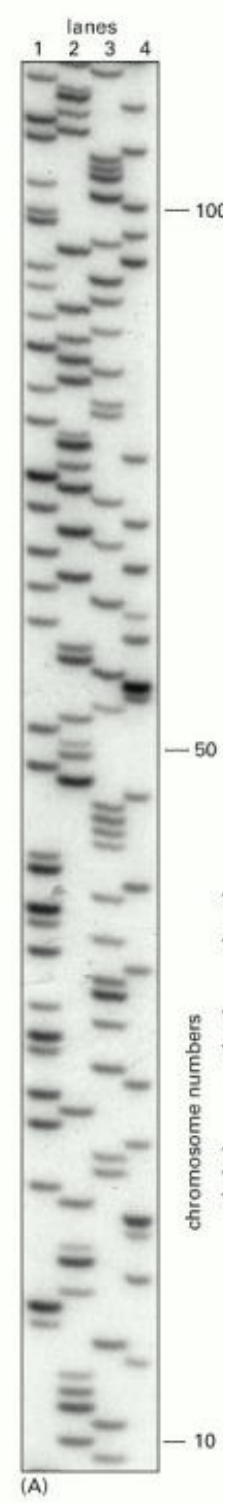
What is sequencing?

1. Yesterday (First generation sequencing)

    a. Maxam-Gilbert Sequencing
    b. Sangar Sequencing







Metrics for evaluating sequencing technology

- **Throughput:**
  - Number of high quality bases per unit time
  - Number of independent samples run in parallel
  - Difficulty of sample preparation
- **Yield**
  - Number of useful reads per sample
  - Read length
- **Cost**
  - Per run cost
  - Per base cost
  - Equipment
  - Reagents
  - Labor
  - Analysis

What is sequencing?

1. Yesterday (First generation sequencing)

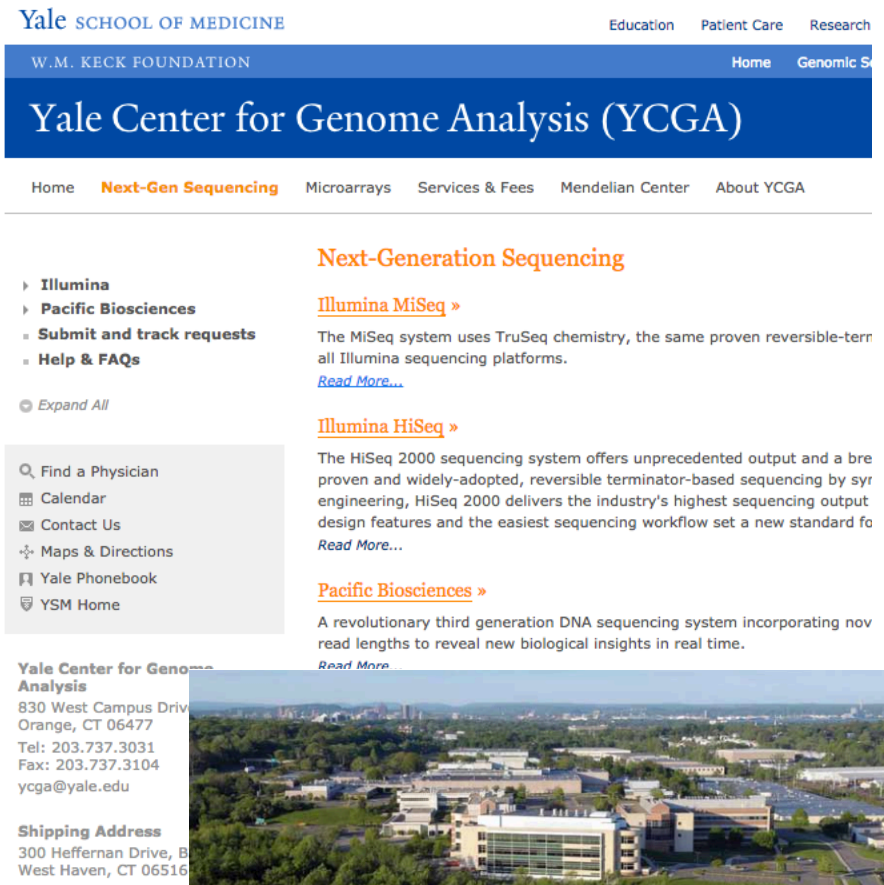
    a. Maxam-Gilbert Sequencing
    b. Sangar Sequencing

2. Today (Second generation sequencing)

    a. Illumina Sequencing
    b. Ion Torrent
    c. Pacific Bioscience Sequencing (3rd-ish)

3. Tomorrow (Third generation sequencing)

    a. Nanopore based
    b. Transistor based
    c. FRET based



The technology will change, but your need to critically understand the input and output will not.

The steps of sequencing experiments

1. Sample preparation

    a. Isolation
    b. Library construction

2. **Sequencing**

    a. Flow cell loading
    b. Cluster generation
    c. Sequencing
    d. Processing image files
    e. De-multiplexing samples

3. Data analysis

    a. Read filtering
    b. Alignment to a genome
    c. Diverse analyses

What is the output from an Illumina sequencing experiment?

One read (fastq format)

```
@HWI-ST1239:178:H0KPNADXX:2:1101:3120:1979 1:N:0:TGACCA
NCTGTAGGCTGCGTAGCCTCCCTGCAGGGTAAGTGGGAGGAGAGAGAGCAGAGGGACTTAGTGGGGCTCCCCAGGG
+
#1=DDFFFHHHHHIJIJJJIJJJJJIJJJ?FHIDGIJ=GIHGIIIHGIJIHEHIHHGFFFFEEEDDDDDDDDDDDD
```

1. Read identifier
2. **Sequence**
3. Quality score identifier “+”
4. Quality score

What is the output from an Illumina sequencing experiment?

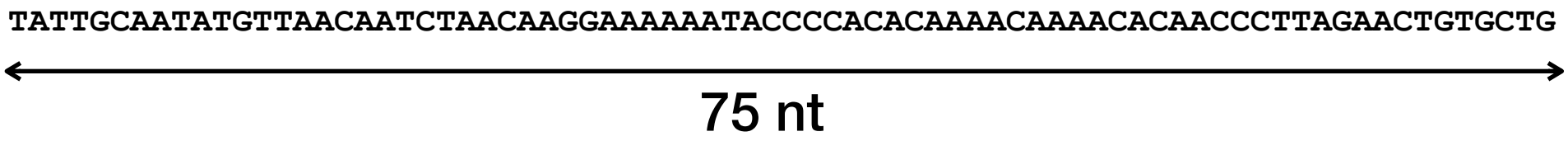
Many reads…

```
@HWI-D00306:498:HBB89ADXX:1:1101:1180:1882 1:N:0:CGATGT
NCATCACTTTCTGCACCAGCCATGACGTCAATCTTCGTCCGAACCCCAAACTCGAGATCGGAAGAGCACACGTCTG
+
#11BBDDDFFDFBFFFIIIIIIIIIIIIIFEGIIIIIFIGAGIIFIII=FEEEEEFFFDDD=@9A@BBBBB=?BB<
@HWI-D00306:498:HBB89ADXX:1:1101:1167:1902 1:N:0:CGATGT
TATTGCAATATGTTAACAATCTAACAAGGAAAAAATACCCCACACAAAACAAAACACAACCCTTAGAACTGTGCTG
+
B@@FFDFFHFHHHJJIJIGIIJJJJJIJJHFIJJJJJIJJJEHHJJIJJJJIIJJJJJJGHHHHFBDFFFE>CEEC
@HWI-D00306:498:HBB89ADXX:1:1101:1190:1928 1:N:0:CGATGT
ACCAAGCCACAATAAGTTAGTGTTTCCATAGTACATGCTGAGTTATTTGATCCCGTATCTATACACTGCTACTGTC
+
@<@DDDDD8CDDDGE?2<AFFBCCEEHEIEGHIIEGEIDD@CDGFFFEFIDGCFCDABFG>FBFGFGIEIFFFDDD
@HWI-D00306:498:HBB89ADXX:1:1101:1157:1931 1:N:0:CGATGT
CTGAGATTCTTTGCCATAGTCCTTAACCACTACGCAACTGCAACCAACCACCTTCCGTGGTTTGCCCTCTCGATCG
+
CCCFFFFFHHHHHHIJJIIJJJIIGHHIJGGJIGIJJJJJJJIJIIIJJJIIJJJJIIJGJJHCHFBDFFFDDECB
```

Generally ~ 300,000,000 reads/sequencing lane

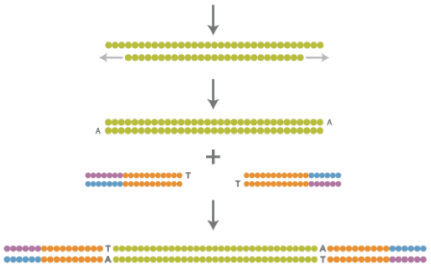
Note: This is for an Illumina HiSeq 2500 with current chemistry, but this number changes

How long are the reads?



While there are other technologies that can give longer read lengths, Illumina reads are generally 50 nt - 250 nt

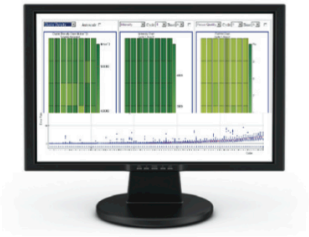
Where do these reads come from?



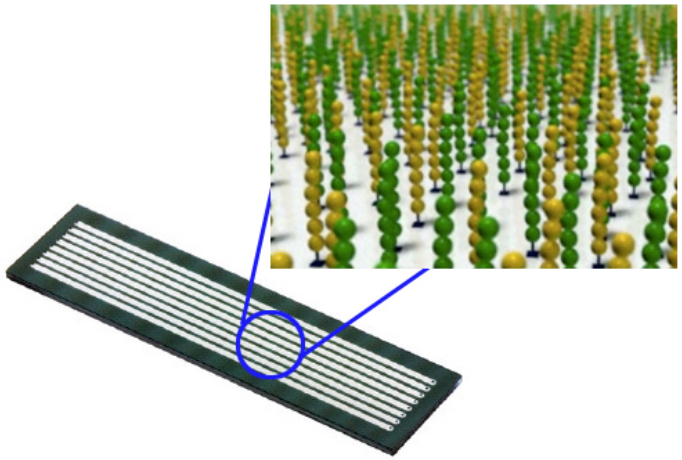
Library Preparation
~2 h [15 min hands-on (Nextera)]
< 6 h [< 3 h hands-on (TruSeq)]

Cluster Generation
~5 h (<10 min hands-on)

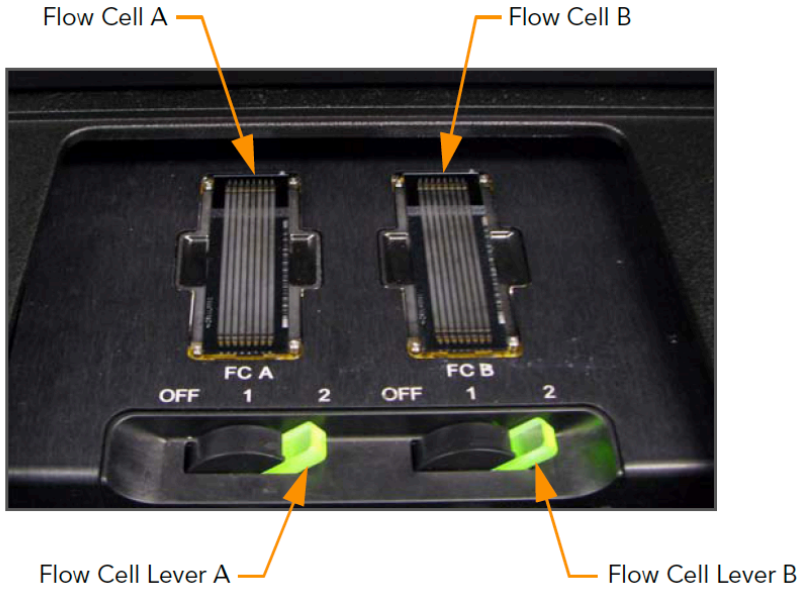
Sequencing by Synthesis
~1.5 to 11 days



CASAVA
2 days (30 min hands-on)



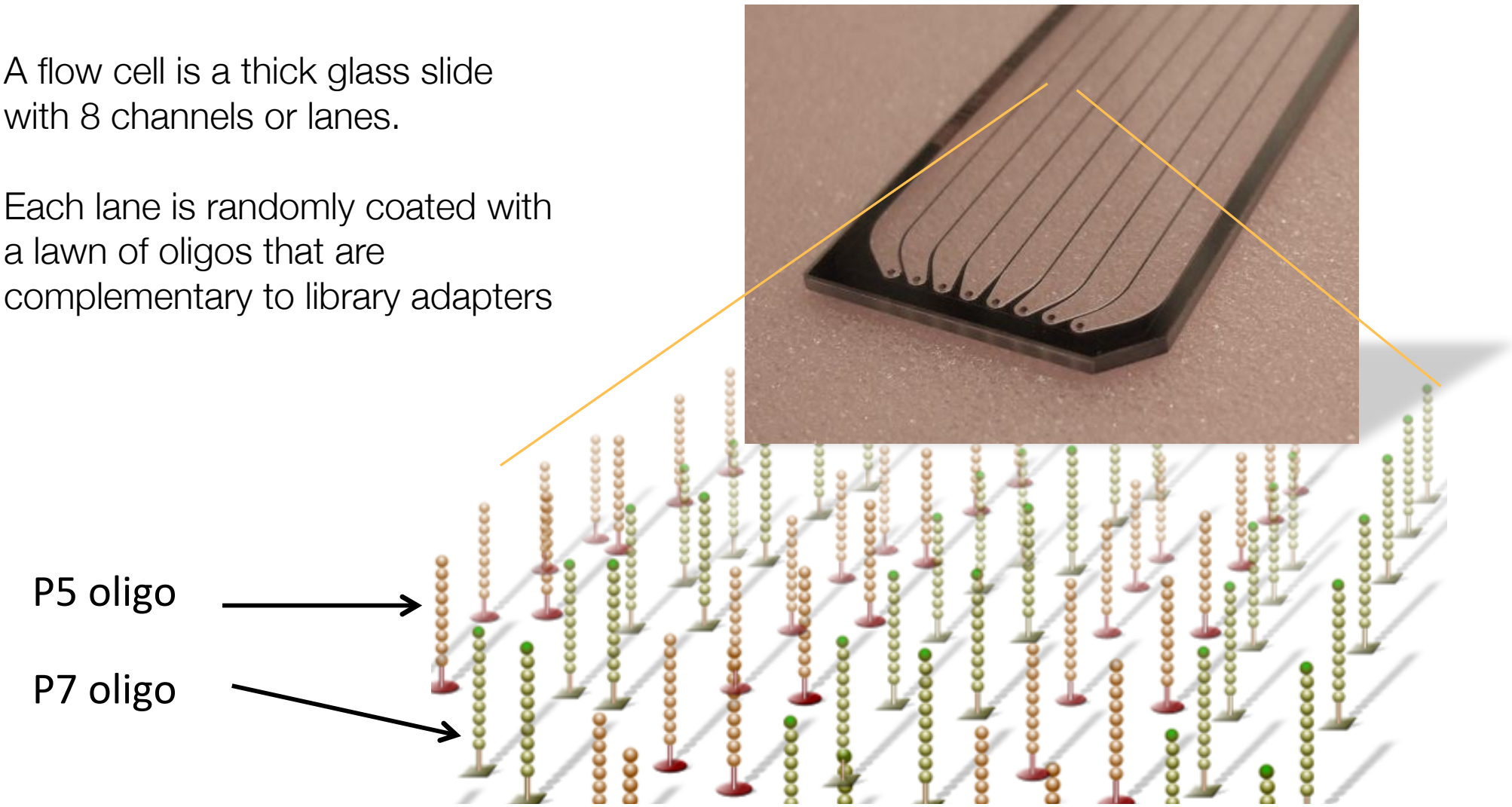
Flow cell



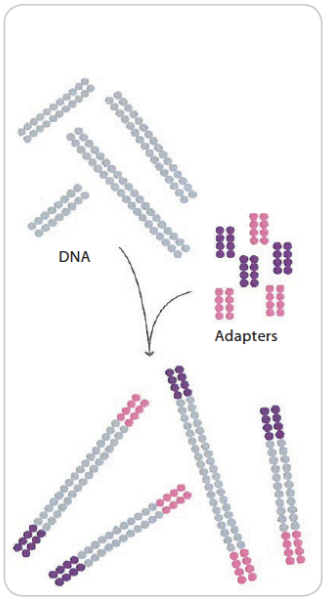
What is a flow cell?

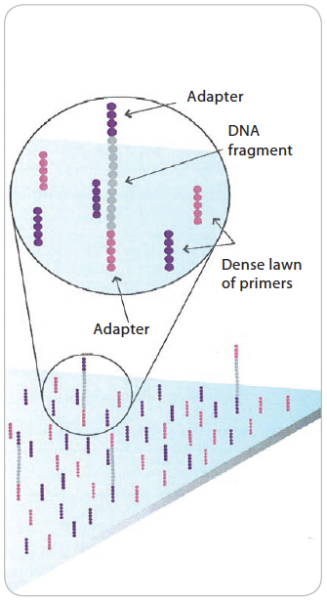
A flow cell is a thick glass slide with 8 channels or lanes.

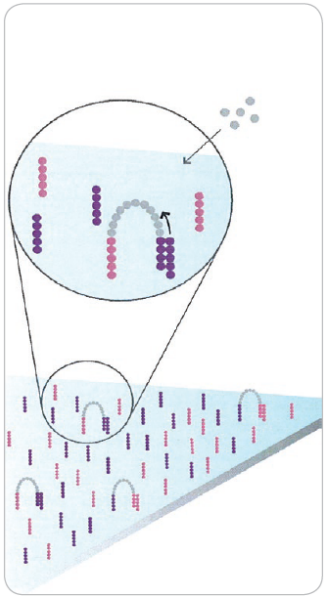
Each lane is randomly coated with a lawn of oligos that are complementary to library adapters

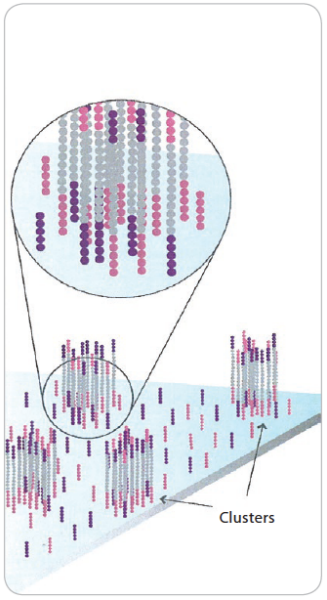


Cluster PCR
on flow cell
(8 lanes)



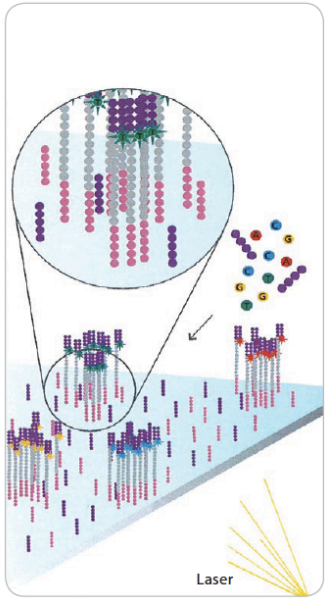






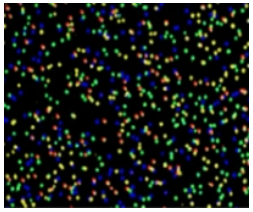
Attach to flow cell ‘bridge PCR’ Cluster generation

Sequencing
by synthesis
with reversible
dye terminators



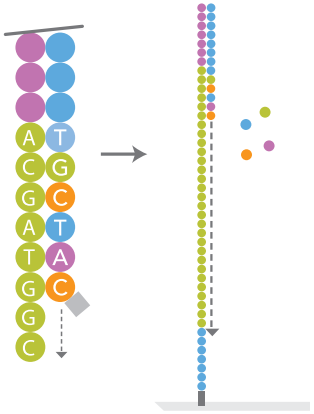
Add base

Scan flow cell



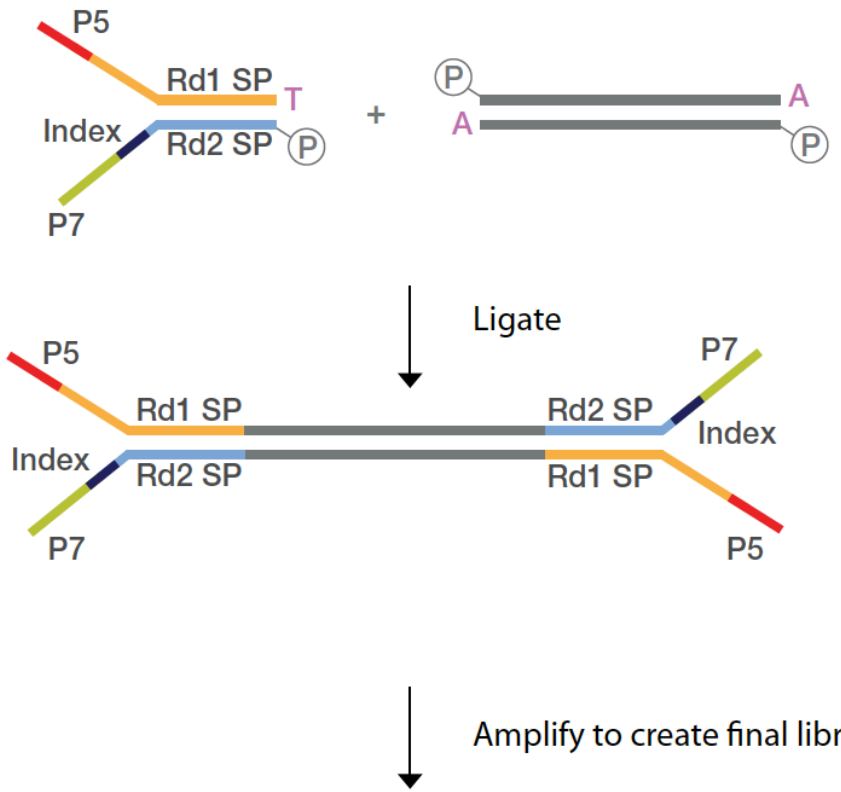
1 cycle

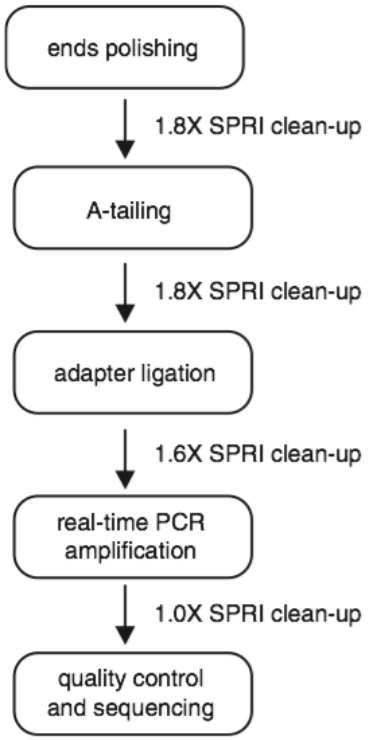
Reverse
termination
Add next base

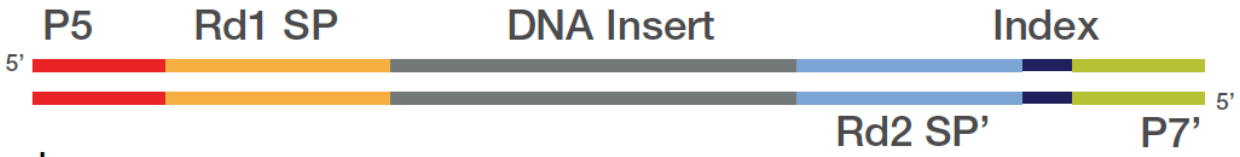


How do you make a sequencing library?

Index = unique sequence
key to identify library







12 samples per lane

Potential sources of bias:

1. Selective PCR amplification (issue of duplicates).
2. Size selection.
3. Enzyme specificities.

Challenging but possible to analyze pg quantities of DNA. (In humans, ~6 pg DNA/cell).

What is the output from an Illumina sequencing experiment?

Paired read (fastq format)

```
@HWI-ST1239:178:H0KPNADXX:2:1101:3120:1979 1:N:0:TGACCA
NCTGTAGGCTGCGTAGCCTCCCTGCAGGGTAAGTGGGAGGAGAGAGAGCAGAGGGACTTAGTGGGGCTCCCCAGGG
+
#1=DDFFFHHHHHIJIJJJIJJJJJIJJJ?FHIDGIJ=GIHGIIIHGIJIHEHIHHGFFFFEEEDDDDDDDDDDD

@HWI-ST1239:178:H0KPNADXX:2:1101:3120:1979 2:N:0:TGACCA
NNACCTAGCCATCTGCAGTCCTCGGTCCTGTGTTAGACCAGAACTAGGTGCCCAGGCCAGGTACCACCTAATCCTT
+
##4<@@@@@@@@@@?@@@?@@??????@@??@??????????????????>???????????@>???@@?@@??????
```

1. Read identifier
    a. Instrument
    b. Flow cell
    c. Read ID
    d. Coordinates
    e. Which read from a paired end sample
    f. Which index for multiplexed read
2. Quality score identifier “+”
3. Quality score

What limits the insert size and read length?

One read (fastq format)

```
@HWI-D00306:498:HBB89ADXX:1:1101:1180:1882 1:N:0:CGATGT
NCATCACTTTCTGCACCAGCCATGACGTCAATCTTCGTCCGAACCCCAAACTCGAGATCGGAAGAGCACACGTCTG
+
#11BBDDDFFDFBFFFIIIIIIIIIIIIIIFEGIIIIIFIGAGIIFIII=FEEEEEFFFDDD=@9A@BBBBB=?BB<
```

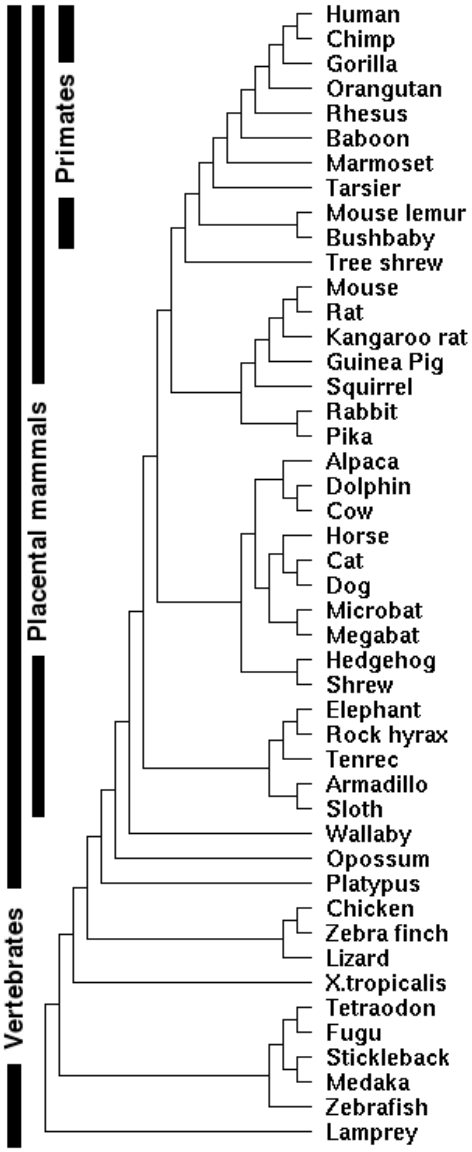
- For each single end read: Incomplete incorporation of bases.

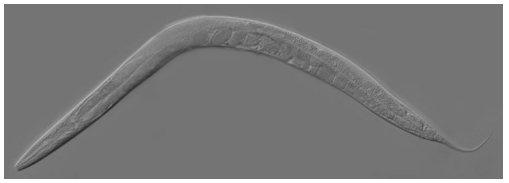
- For the size of the insert (especially for paired end analysis): Ability to get consistent clusters.

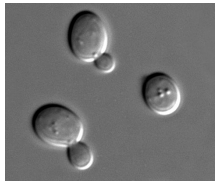
What do I do with my sequencing reads?

Source: Slate via Noonan

Many reference genomes are available





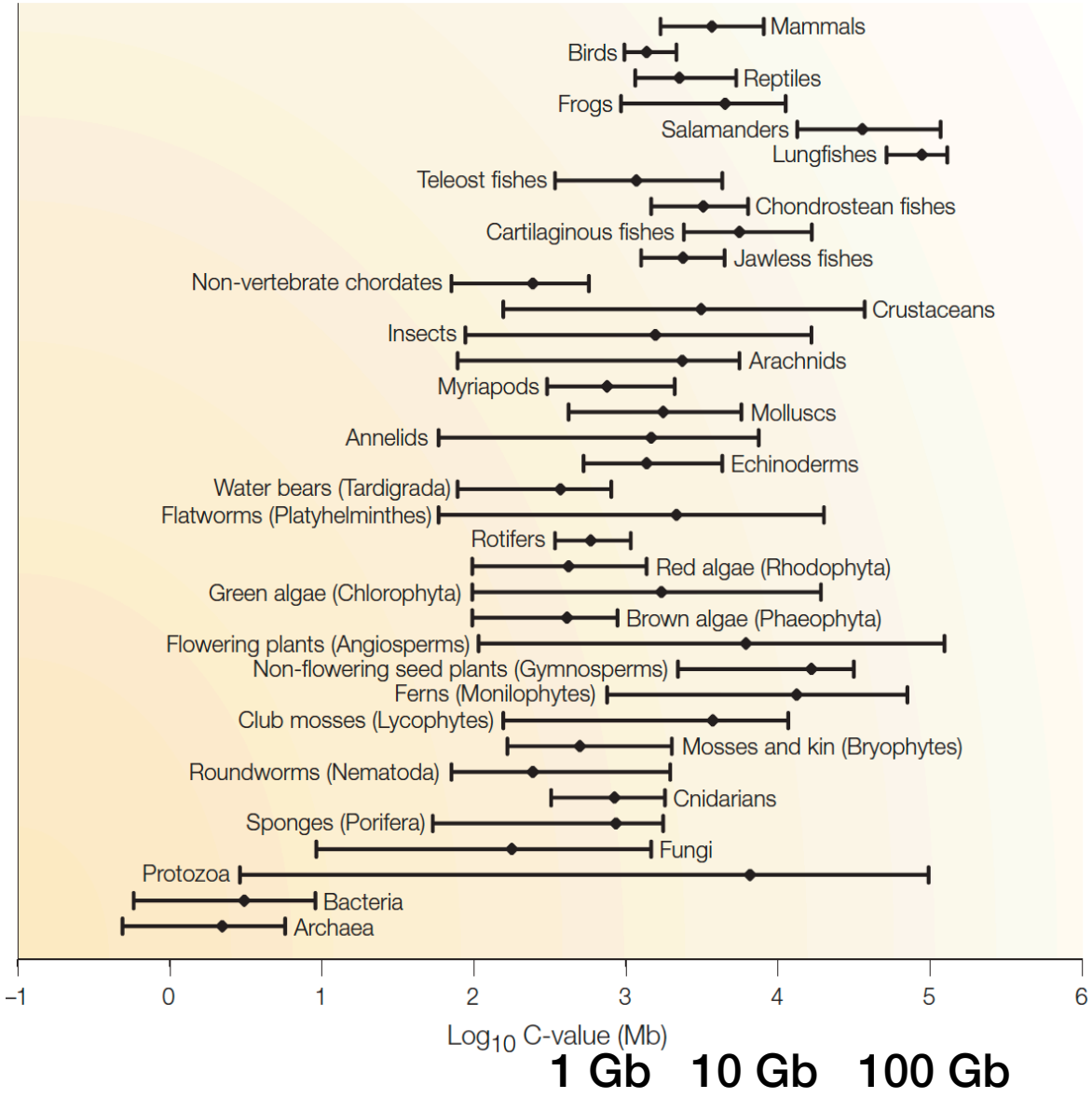


There is a wide range of genome sizes.

kb = 1000 bp
Mb = $1x10^6$ bp
Gb = $1x10^9$ bp
Tb = $1x10^{12}$ bp

Human haploid genome ~ 3 Gb

75 nt x $3x10^8$ reads/lane is about the right scale, but the amount of **coverage** necessary depends on application.



Sequencing of the human genome

Victory declared **2003**

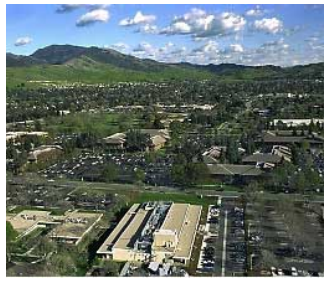
HGSC HUMAN GENOME SEQUENCING CENTER

National Human Genome Research Institute

JGI DOE JOINT GENOME INSTITUTE US DEPARTMENT OF ENERGY OFFICE OF SCIENCE

wellcome trust sanger institute

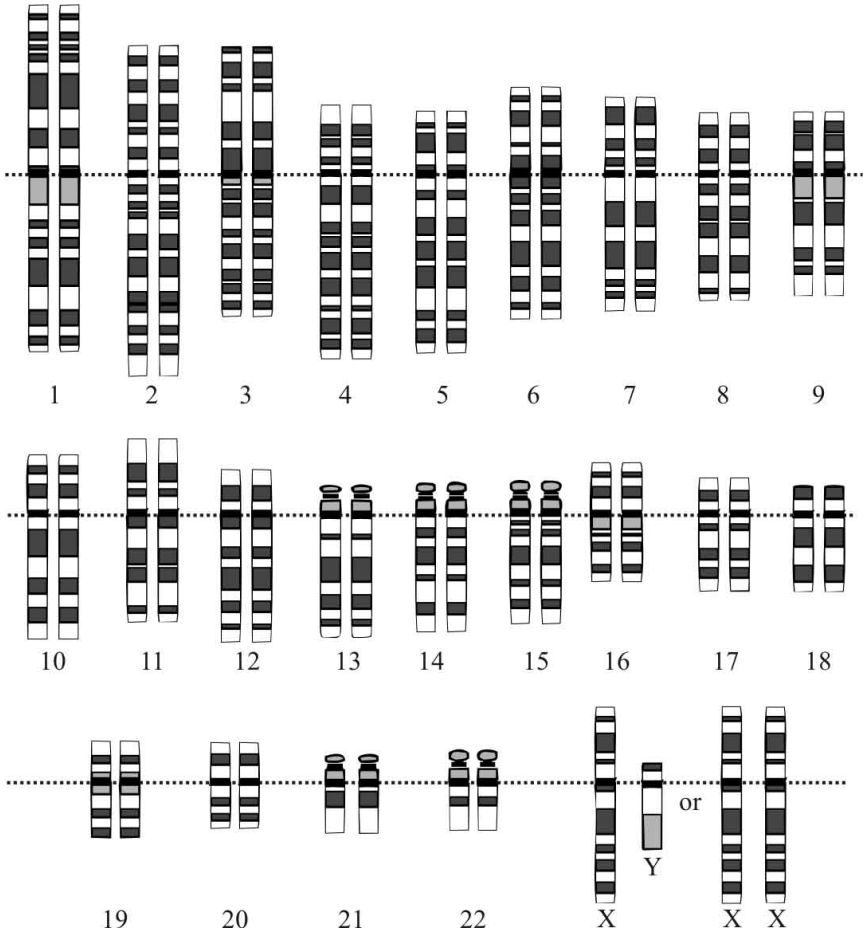
- Industrialization of Sanger sequencing, library construction, sample preparation, analysis, etc.
- $3 billion total cost
- 1 Gb/month at largest centers (2005)
- YCGA = 9.6 Tb per month (2011)



Assembling a genome from short reads

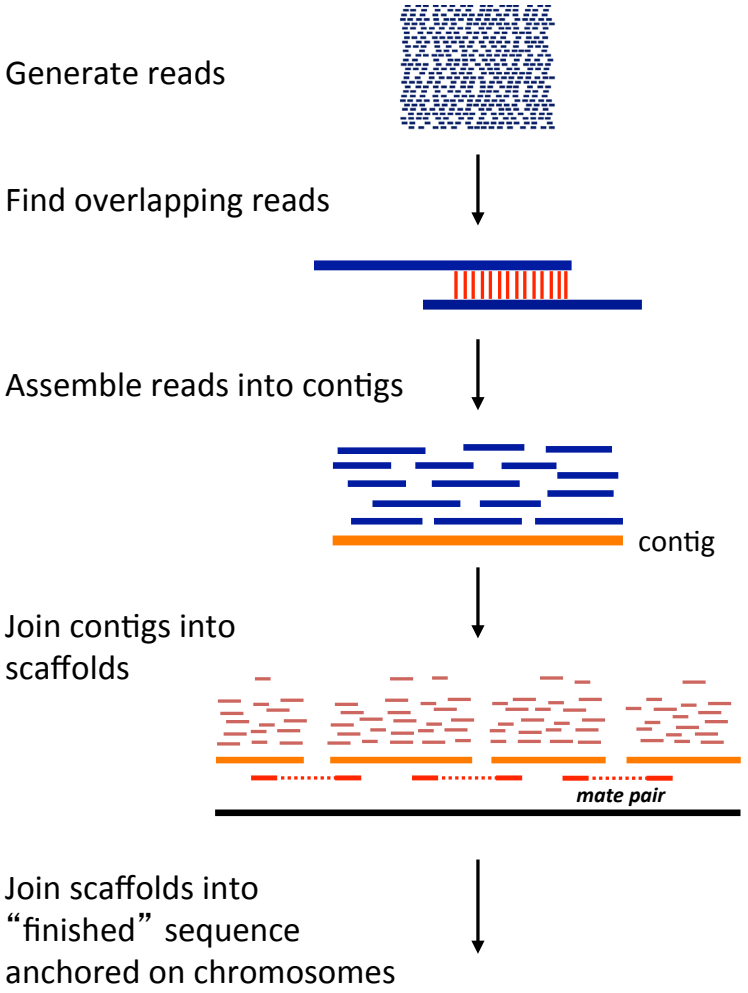
>>10^9 sequencing reads

36 bp - 1 kb



3 Gb

How to assemble a genome



Assembly quality criteria:

Accuracy: number of errors
(Human << 1/100,000 bp)

Contiguity: number of gaps
(Human: est. 357)

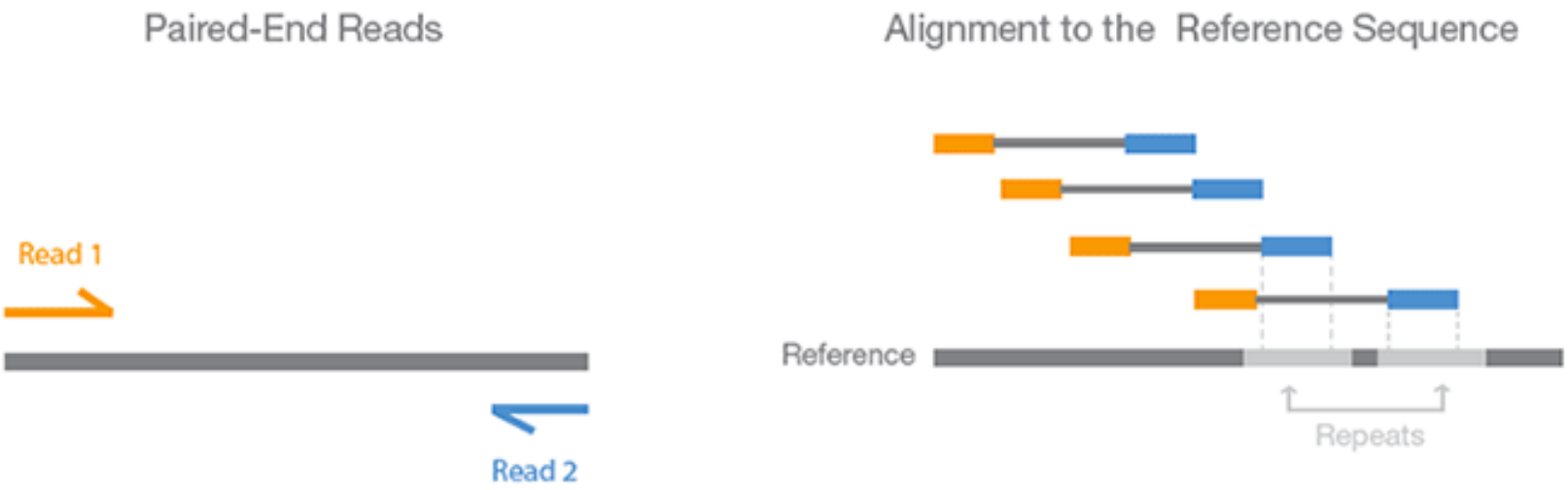
Coverage:
Average number of reads representing a particular position in the assembly

Human, Mouse, Rat: > 20x
Chimpanzee: ~6x
Squirrel: ~2x

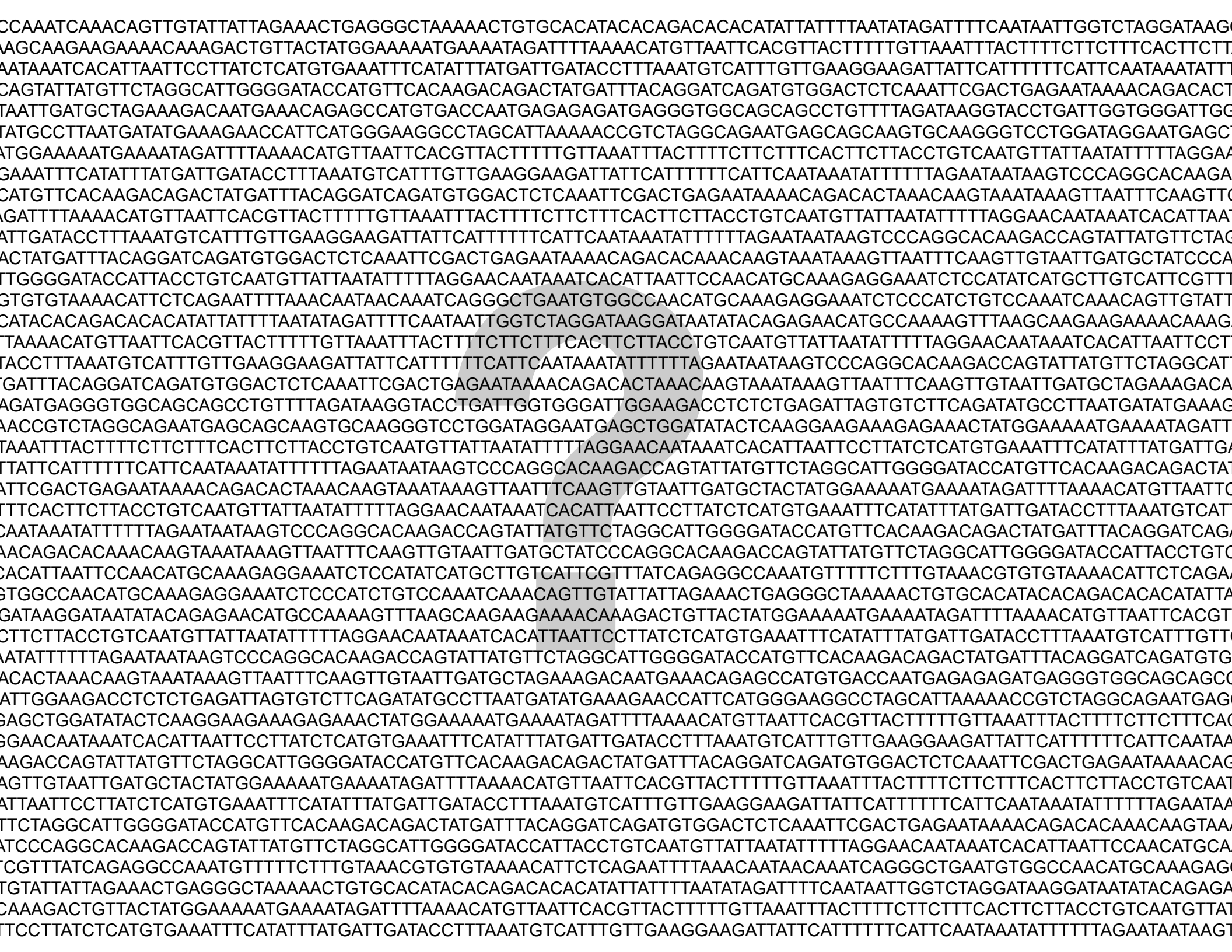
Scaffold_0: 12,865,123 – 12,965-110

Chr5: 133,876,119 – 134,876,119

The importance of paired end reads



- Increase coverage of the insert.
- Particularly helpful when one read maps to multiple places in the genome.



What types of annotation do we have/want?

~3 billion bp

ACAATAAATCACATTAATTCCTTATCTCATGTGAAATTTCATATTTATGATTG
ATACCTTTAAATGTCATTTGTTGAAGGAAGATTATTCATTTTTTCATTCAAT
AAATATTTTTTAGAATAATAAGTCCCAGGCACAAGACCAGTATTATGTTCT
AGGCATTGGGGATACCATGTTCACAAGACAGACTATGATTTACAGGATC
AGATGTGGACTCTCAAATTCGACTGAGAATAAAACAGACACTAAACAAG
TAAATAAAGTTAATTTCAAGTTGTAATTGATGCTAGAAAGACAATGAAACA
GAGCCATGTGACCAATGAGAGAGATGAGGGTGGCAGCAGCCTGTTTTA
GATAAGGTACCTGATTGGTGGGATTGGAAGACCTCTCTGAGATTAGTGT
CTTCAGATATGCCTTAATGATATGAAAGAACCATTCATGGGAAGGCCTAG
CATTAAAAACCGTCTAGGCAGAATGAGCAGCAAGTGCAAGGGTCCTGG
ATAGGAATGAGCTGGATATACTCAAGGAAGAAAGAGAAACTATGGAAAA
ATGAAAATAGATTTTAAAACATGTTAATTCACGTTACTTTTTGTTAAATTTA
CTTTTCTTCTTTCACTTCTTACCTGTCAATGTTATTAATATTTTTAGGAACA
ATAAATCACATTAATTCCTTATCTCATGTGAAATTTCATATTTATGATTGATA
CCTTTAAATGTCATTTGTTGAAGGAAGATTATTCATTTTTTCATTCAATAAA
TATTTTTTAGAATAATAAGTCCCAGGCACAAGACCAGTATTATGTTCTAGG
CATTGGGGATACCATGTTCACAAGACAGACTATGATTTACAGGATCAGAT
GTGGACTCTCAAATTCGACTGAGAATAAAACAGACACTAAACAAGTAAAT
AAAGTTAATTTCAAGTTGTAATTGATGCTACTATGGAAAAATGAAAATAGA
TTTTAAAACATGTTAATTCACGTTACTTTTTGTTAAATTTACTTTTCTTCTTT
CACTTCTTACCTGTCAATGTTATTAATATTTTTAGGAACAATAAATCACATT
AATTCCTTATCTCATGTGAAATTTCATATTTATGATTGATACCTTTAAATGT
CATTTGTTGAAGGAAGATTATTCATTTTTTCATTCAATAAATATTTTTTAGA
ATAATAAGTCCCAGGCACAAGACCAGTATTATGTTCTAGGCATTGGGGAT
ACCATGTTCACAAGACAGACTATGATTTACAGGATCAGATGTGGACTCTC
AAATTCGACTGAGAATAAAACAGACACAAACAAGTAAATAAAGTTAATTT
CAAGTTGTAATTGATGCTATCCCAGGCACAAGACCA....

Genes:

- Coding, noncoding, miRNA, etc.
- Isoforms
- Expression

Genetic variation:

- SNPs and CNVs

Sequence conservation

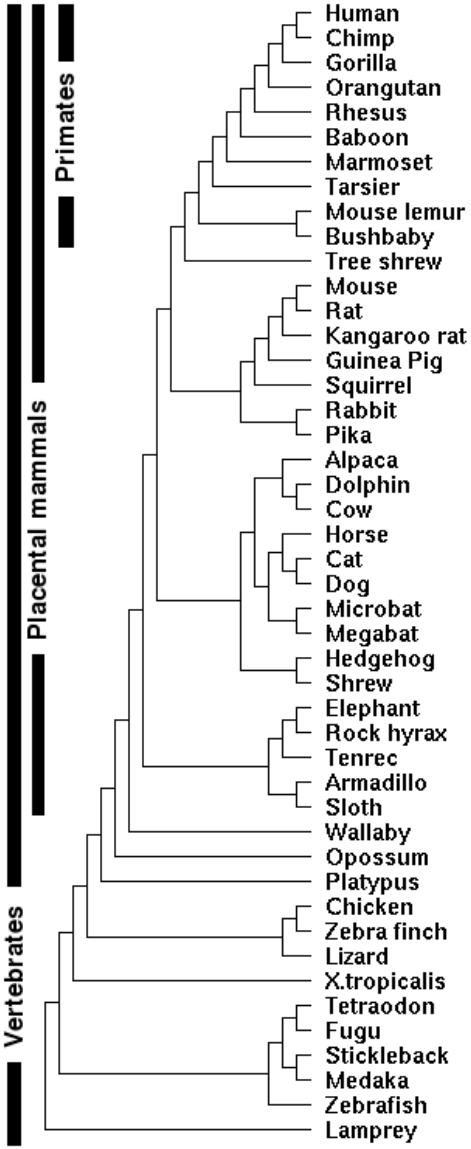
Regulatory sequences:

- Promoters
- Enhancers
- Insulators

Epigenetics:

- DNA methylation
- Chromatin

Degrees of genomic annotation vary widely



Human, Mouse (Fly, Worm, Yeast):

- Chromosome assemblies
- Dense gene and regulatory maps, variation, etc.

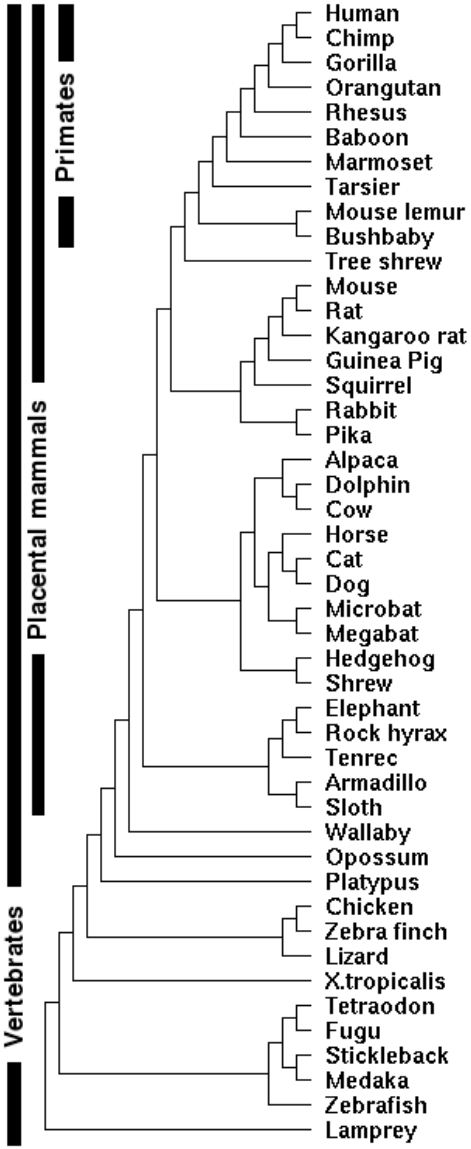
Other models (Dog, Chicken, Zebrafish):

- Chromosome assemblies
- Partial gene maps; variation; little regulatory data

Low coverage vertebrate genomes:

- Scaffold assemblies
- Few annotated genes
- Used for comparative purposes

Degrees of genomic annotation vary widely



ENCODE and modENCODE

Human, Mouse (Fly, Worm, Yeast):

- Chromosome assemblies
- Dense gene and regulatory maps, variation, etc.

Other models (Dog, Chicken, Zebrafish):

- Chromosome assemblies
- Partial gene maps; variation; little regulatory data

Low coverage vertebrate genomes:

- Scaffold assemblies
- Few annotated genes
- Used for comparative purposes

Where do you look for existing annotations?

UCSC Genome Browser (genome.ucsc.edu):

Visualization, data recovery, simple analysis
(also http://genome-preview.ucsc.edu/)

ENSEMBL (ensembl.org):

Visualization, data recovery, simple analysis

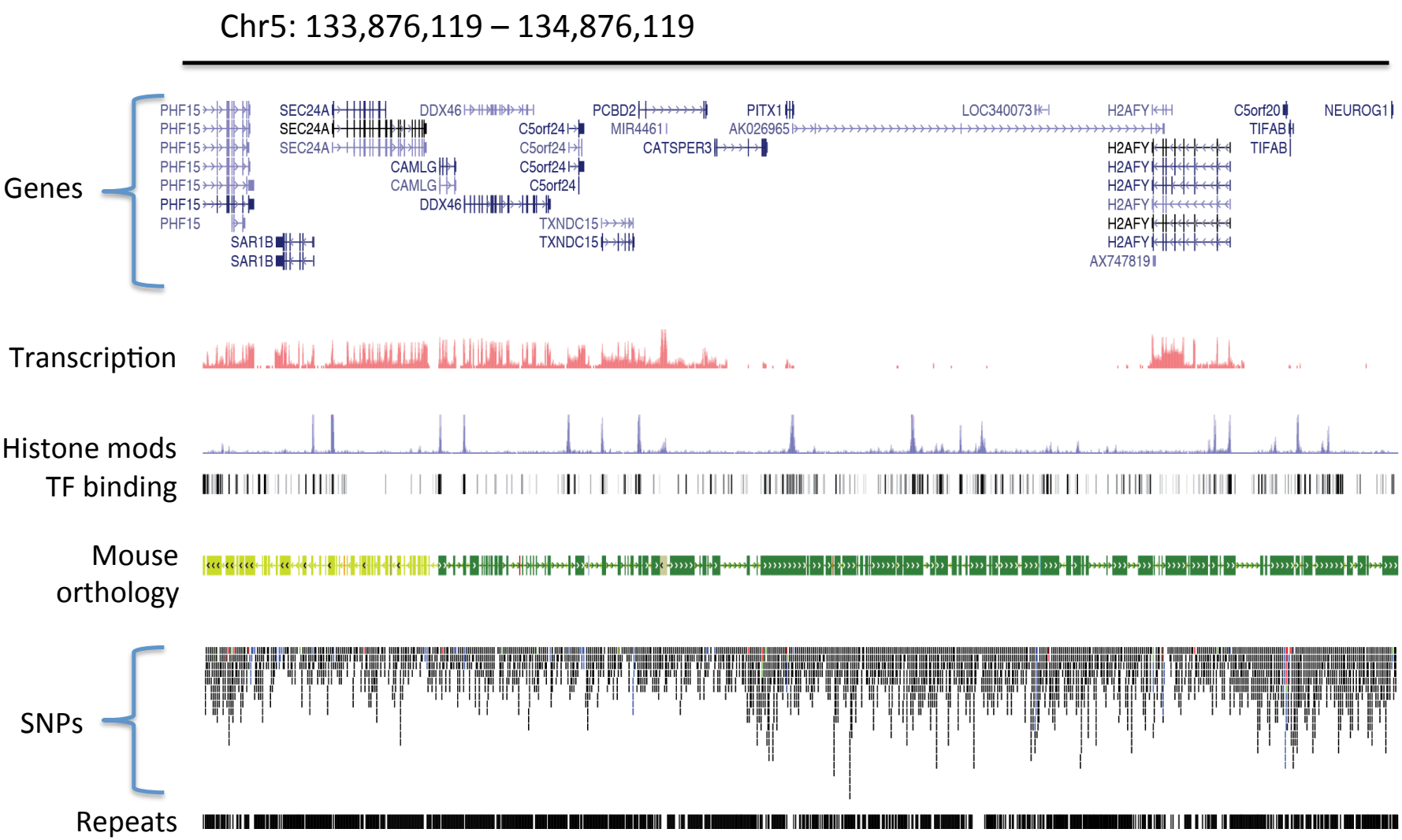
Integrative Genomics Viewer
(broadinstitute.orgsoftware/igv/):

Local genome viewer (visualize local and remote data)

Galaxy (main.g2.bx.psu.edu):

Complex data analysis and workflows

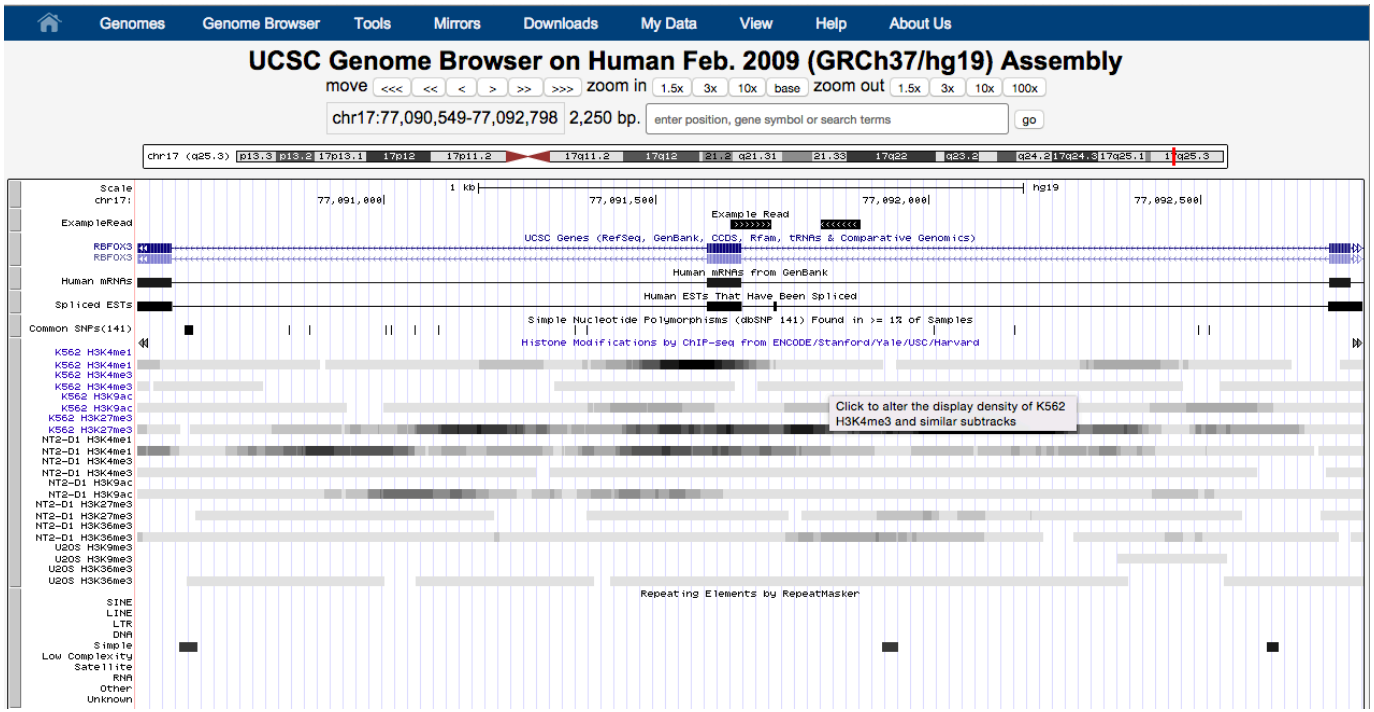
Example of a genome browser track



Our specific example:

```
@HWI-ST1239:178:H0KPNADXX:2:1101:3120:1979 1:N:0:TGACCA
NCTGTAGGCTGCGTAGCCTCCCTGCAGGGTAAGTGGGAGGAGAGAGAGCAGAGGGACTTAGTGGGGCTCCCCAGGG
+
#1=DDFFFHHHHHIJIJJJIJJJJJIJJJ?FHIDGIJ=GIHGIIIHGIJIHEHIHHGFFFFEEEDDDDDDDDDDDD

@HWI-ST1239:178:H0KPNADXX:2:1101:3120:1979 2:N:0:TGACCA
NNACCTAGCCATCTGCAGTCCTCGGTCCTGTGTTAGACCAGAACTAGGTGCCCAGGCCAGGTACCACCTAATCCTT
+
##4<@@@@@@@@@@?@@@?@@??????@@??@??????????????????>??????????@>???@@@?@@??????
```



How else can sequence contribute to our understanding of the regulation of our genomes?

1. Examine transcription: RNA-seq
2. Probe genomic binding sites of proteins (e.g., TFs): ChIP-seq
3. Probe histone modifications: ChIP-seq
4. Probe DNA-methylation: methyl-Seq
5. Examine genomic variation.
6. Probe genomic binding sites of RNAs (e.g., TFs): CHART-seq
7. Examine the conformation of the genome through DNA-DNA interactions: 4C/5C/Hi-C/&c.
8. Probe RNA-protein interactions. (e.g., CLIP)

Applications of sequencing technology next week.

Second-generation sequencing

"Democratizing" sequencing production

- Massive parallelization
- Reduction in per-base cost
- Eliminate need for huge infrastructure
- Millions of reads - >1Gb sequence per run

Novel sequencing applications

- RNA-seq
- ChIP-seq
- Methyl-seq

} Counting applications

- Whole-genome and targeted resequencing

Challenges

- Read length
- Quality
- Data analysis and storage

HiSeq 2500

1 Instrument – 2 Run Modes

High Output Mode

600 Gb in ~10.5 days
Current v3 flow cell
Current v3 reagents
cBot required

Rapid Run Mode

120Gb in ~1 day
New 2-lane flow cell
New reagents
No cBot required

User configurable

6 human genomes in 10.5 days

Highest Output

1 human genome in a day

Fastest turnaround

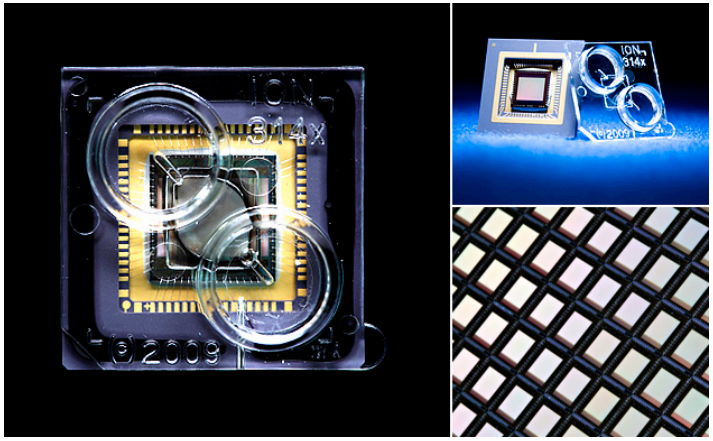
MiSeq

- Run-times
  - 50 cycle – 4 hours
  - 300 cycle – 27 hours
- Two sequencing options
  - 50 cycles
  - 300 cycles (2x150 bp)
- One lane
  - 6-7 million clusters
  - Up to 8 billion bases (300 cycles)

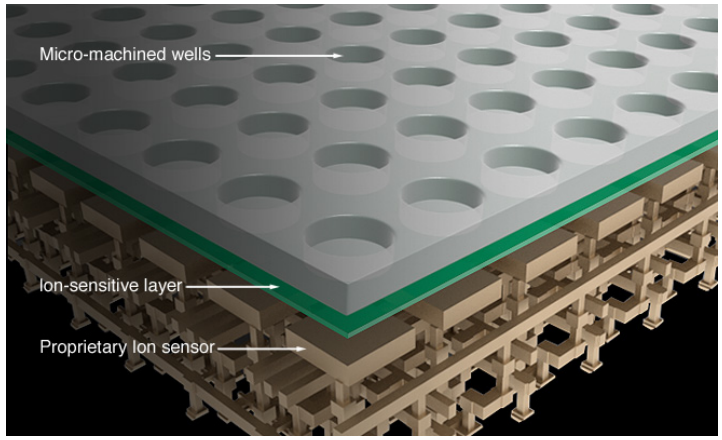
Ideal for: R&D, CLIA, small genomes and projects where longer reads are important

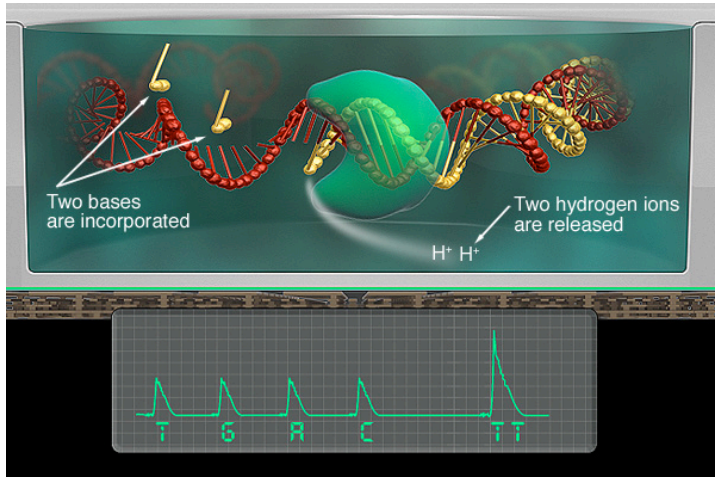
Ion Torrent and Ion Proton

Sequencing on semiconductor chip

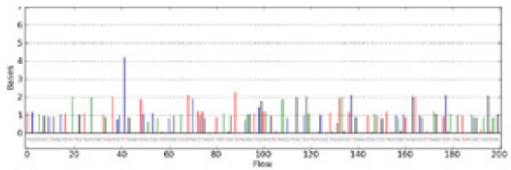


Ion Torrent sequencing chemistry





Single-Well Ionogram for (2406, 1991)



When a nucleotide is incorporated into a strand of DNA, a proton is released as a byproduct.

The H^+ ion carries a charge which the PGM's ion sensor can detect as a base.

Advantages and limitations

Advantages

- Low equipment cost
- Rapid run times: 3 to 4 hours
- Simple Chemistry

Limitations

- Homopolymers detection
- Error rates
- Slow on introducing newer chips: Overpromise
- PGM and Proton: two separate systems
- Library prep: Emulsion PCR

Toward third-generation sequencing

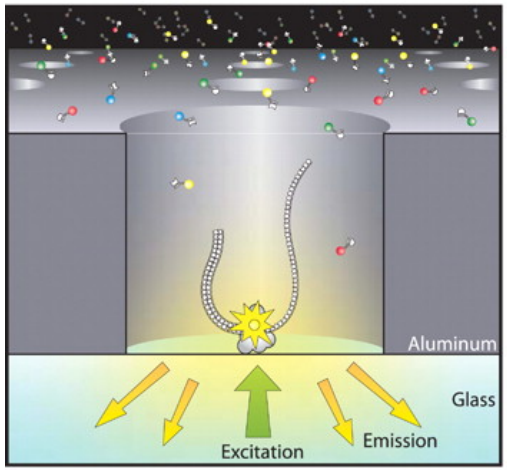
High-throughput single molecule sequencing in real time at low cost

Pacific Biosciences

- Sequence in real time with fluorescent NTPs
- Rate limited by processivity of polymerase
- Very long reads possible (6 kb)
- Not well parallelized (few reads)

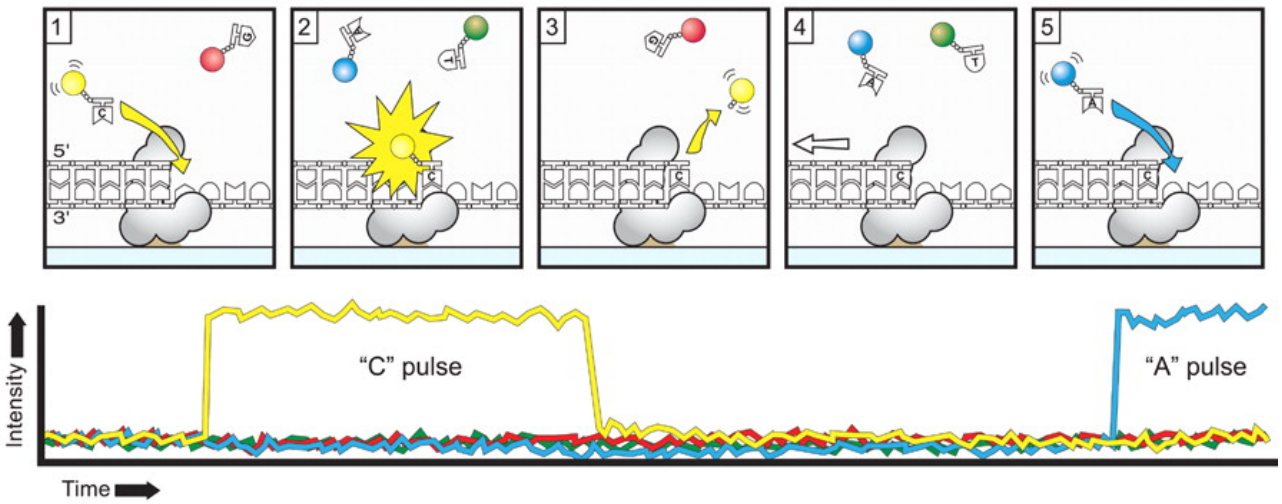
Sequencing in real time: Pacific Biosciences

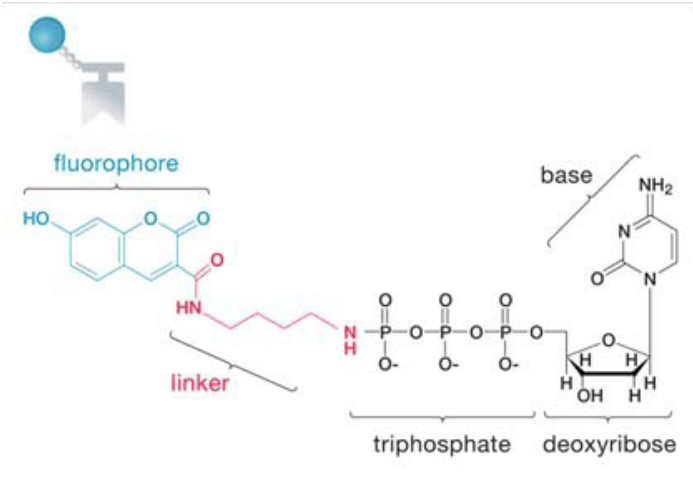
A SMRT cells



Zero Mode Waveguides

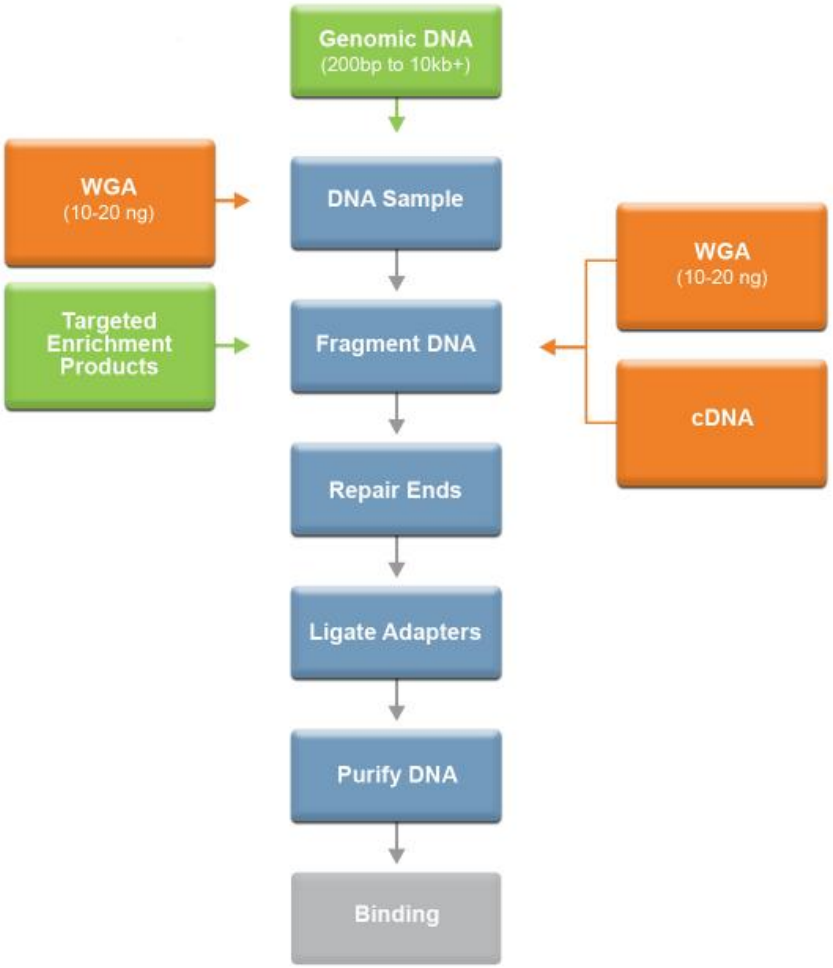
B



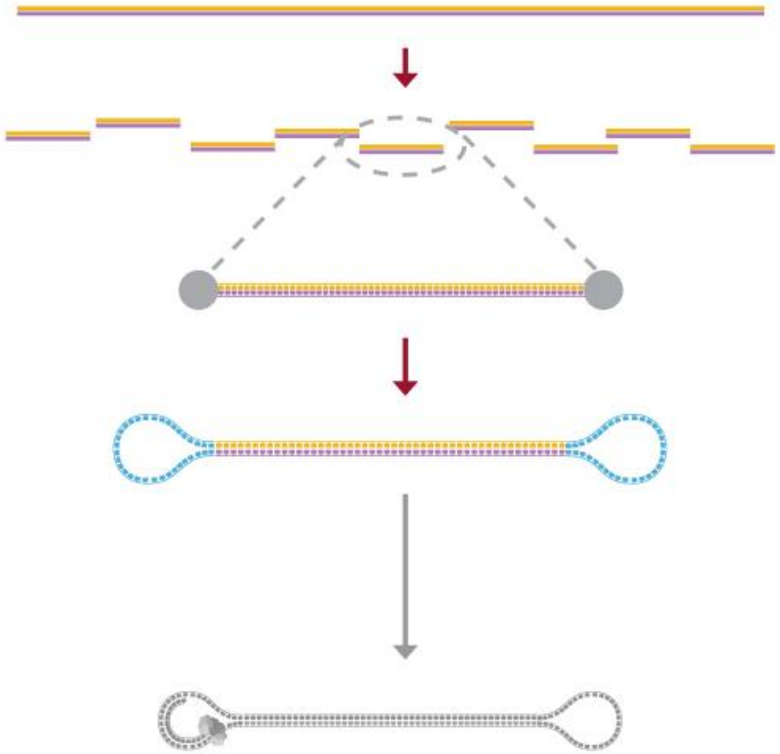


PacBio sequencing strategy

Sample Preparation



Building of SMRTbell



Applications

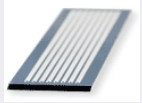
- Targeted sequencing
  - SNP and structure variants detection
  - Repetitive regions
  - Full length transcript profiling
- De novo assembly and genome finishing
  - Bacterial genomes
  - Fungal genomes
  - Gap-captured sequencing
  - Targeted captured sequencing
- Base modifications detection
  - Methylation
  - DNA damage

**Projects at YCGA

YCGA PacBio *RS*

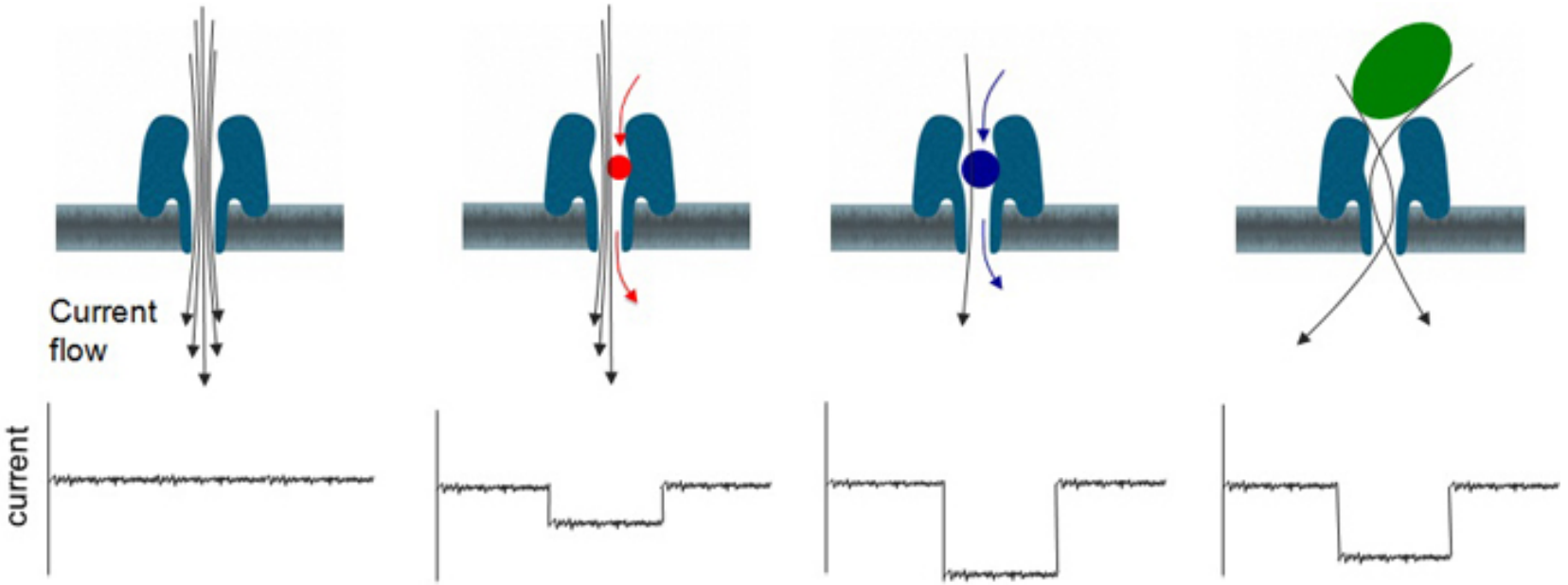
Shrikant Mane

PacBio vs Illumina

	PacBio *RS (Third generation)*	Illumina HiSeq (Second generation)
Sequencing Chemistry	Sequencing by synthesis (SBS) Single Molecule Real Time (SMRT)	Sequencing by synthesis (SBS)
Sequencing substrate	 Smart Cell made up of 150,000 ZMWs	 Flow cell has made of 8 separate lanes
Data output per day	1 to 2 billion/ day. $1.5/ Mb	60 billion/day at a cost of $.06 per Mb
Read Length	Average up to 5 Kb	50bp to 150bp
Error rates	Raw: 10-15 %. With 30x coverage: Q50 (< 0.01)	0.5 to 1 %
Sample Library	SMRT Bell template (Single-strand circular DNA) 250 bp to 10 Kb insert	dsDNA with adaptors (175 bp to 1 Kb)

Shrikant Mane

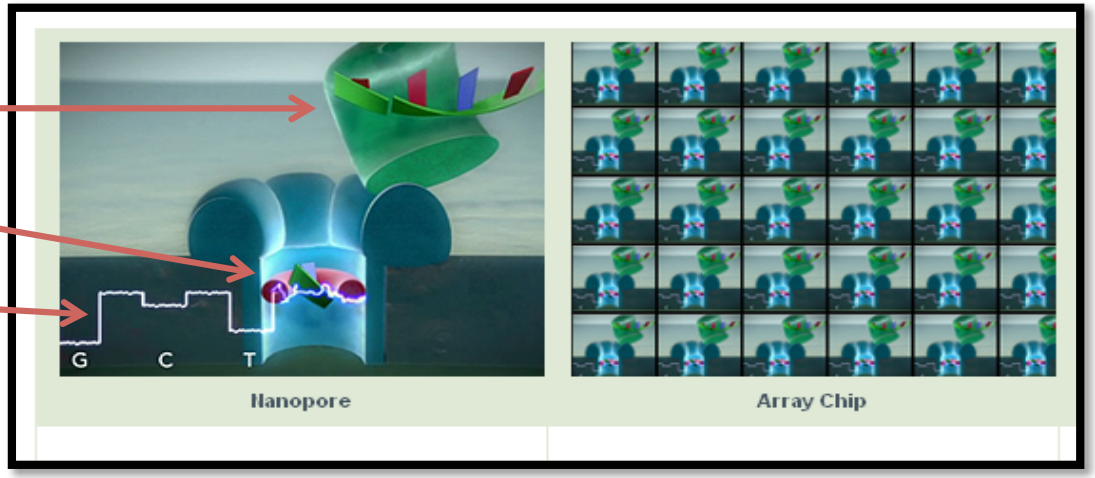
Oxford Nanopore



Exonuclease

Cyclodextrin

Lipid bilayer



Advantages and limitations

- Nanopores offer a label-free, electrical, single-molecule DNA sequencing method
- No costly fluorescent labeling reagents
- No need for expensive optical hardware and sophisticated instrumentation to detect DNA bases
- Runs as long as needed
- High error rates ~5%
- Not available yet

Conclusions

- High-throughput sequencing has become democratized - moved out of industrial-scale genome centers
- Sequence is no longer limiting - next generation of sequencers will make sequencing very inexpensive
- Earlier methods for counting / resequencing applications are largely obsolete
- Scale of data production outstripping our ability to store and analyze it
- Next: Applications of the technology