

# Genomics Part I

---

Matt Simon  
Dept. of Molecular Biophysics & Biochemistry  
Chemical Biology Institute  
January 20, 2017

# 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**

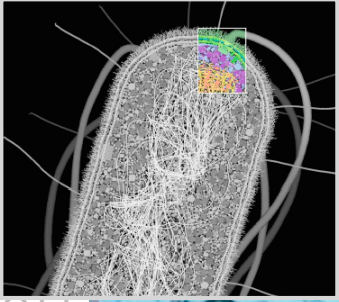
Gene expression and regulation

Cellular identity, differentiation and development

Phenotypic variation among individuals and species

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

CCATGTTCAACAAGACAGAC TATGATTACAGGATCAGATGGGACTCTCAAATTCGACTGAGAATAAAACAGACACTA  
TAGATTTTAAAACATGTTAATTCACGTTACTTTTTGTTAAATTTACTTTTTCTTCTTTCACTTCTTACCTGTCAATGTTATTAA  
GATTGTCATTTGTTGAAGGAAGATTATTCATTTTTTTCATTCAATAAATATTTTTTAGAATAATAAGTCC  
GACTATCAGATGTGGACTCTCAAATTCGACTGAGAATAAAACAGACACAAACAAGTAAATAAAGTTA  
ATTGCTGTCAATGTTATTAATATTTTTAGGAACAATAAATCACATTAATTCCAACATGCAAAGAGGAA  
CGTCACAAATTTTAAAACAATAAACAATTCAGGGCTGAATGTGGCCAACATGCAAAGAGGAAATCTC  
ACATTTGGTCTAGGATAAGGATAATATACAGAGAACATGCC  
TTATACTTCTTTCACTTCTTACCTGTCAATGTTATTAATATTT  
ATACCTTCATTCAATAAATATTTTTTAGAATAATAAGTCCCAGGC  
ATGATTTAATAAAACAGACACTAAACAAGTAAATAAAGTTAATTT  
GAGATGAATTGGTGGGATTGGAAGACCTCTCTGAGATTAGTGT  
AAACCGTATAGAGGAAATGAGCTGGATATACTCAAGGAAGAAAG  
TTAAATTTAATTTTTAGGAACAATAAATCACATTAATTCCTTAT  
ATTATTCAACAGGCACAAGACCAGTATTATGTTCTAGGCATTGC  
AATTCGACTTAATTTCAAGTTGTAATTGATGCTACTATGGAAAA  
CTTTCACATAAATCACATTAATTCCTTATCTCATGTGAAATTTCA  
TCAATAAAGTATTATGTTCTAGGCATTGGGGATACCATGTTTAC  
AAACAGATTGATGCTATCCCAGGCACAAGACCAGTATTATGTT  
TCACATTCTTGTCATTCGTTTATCAGAGGCCAAATGTTTTTCTT  
TGTGGCOCAAACAGTTGTATTATTAGAACTGAGGGCTAAAAA  
GGATAAGGAAGAAAACAAGACTGTTACTATGGAAAATGAA  
ACTTCTTACATTAATTCCTTATCTCATGTGAAATTTCATATTTA  
AAATATTTATGTTCTAGGCATTGGGGATACCATGTTTACAAGAC  
GACACTAGCTAGAAAGACAATGAAACAGAGCCATGTGACCA  
GATTGGATAATGATATGAAAGAACCATTTCATGGGAAGGCCTAG  
TGAGCTGATGAAAATAGATTTTTAAAACATGTTAATTCACGTTACT  
AGGAACAATATTTATGATTGATACCTTTAAATGTCATTTGTTGAA  
CAAGACOCAGACAGACTATGATTTACAGGATCAGATGTGGAC  
AAGTTGTAACATGTTAATTCACGTTACTTTTTGTTAAATTTACT  
CATTAAATCCTTTAAATGTCATTTGTTGAAGGAAGATTATTCATT  
GTTCTAGGCATGGGGATACCATGTTTACAGGATCAGATGTGGACTCTCAAATTCGACT  
TATCCCAGGCACAAGACCAGTATTATGTTCTAGGCATTGGGGATACCATTACCTGTCAATGTTATTAATATTTTTAGGAA  
TTCGTTTATCAGAGGCCAAATGTTTTCTTTGTTAAACGTGTGTAAACATTCTCAGAATTTTAAACAATAACAATCAGG



# Overview

---

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



# Importance of genomics data

## Genomic hallmarks of localized, non-indolent prostate cancer

Micha  
Julie I  
Natali  
Musar  
Zhang  
+ et

Affilia

Nature  
Receiv  
2017

PI

Abstr

Abstrac

ACCEPTED MANUSCRIPT

## An efficient targeted nuclease strategy for high-resolution mapping of DNA



< Previous Article

Volume 93, Issue 2, p362–378, 18 January 2017

Pete

Howa

DOI: h

Publis

Cite a

[ -

Article

Switch to Standard View

## lncRNA Functional Networks in Oligodendrocytes Reveal Stage-Specific Myelination Control by an *IncOL1/Suz12* Complex in the CNS

Danyang He, Jincheng Wang, Yulan Lu, Yaqi Deng, Chuntao Zhao, Lingli Xu, Yinhuai Chen, Yueh-Chiang Hu, Wenhao Zhou, Q. Richard Lu<sup>6</sup>✉

<sup>6</sup> Lead Contact

DOI: <http://dx.doi.org/10.1016/j.neuron.2016.11.044> |

(CUT&RUN), a chromatin profiling strategy in which

controlled cleavage by micrococcal nuclease relea

complexes into the supernatant for paired-end DNA sequencing. Unlike

Chromatin Immunoprecipitation (ChIP), which fragments and solubilizes total

### ▼ Accession Numbers

The accession number for the RNA-seq, ChIP-seq data, and lncRNA annotation reported in this paper is GEO: GSE82211.

# Data can be found in genomics databases

The screenshot shows the NCBI GEO Accession Display page for GSE82211. The page includes the NCBI logo, the GEO logo (Gene Expression Omnibus), and navigation links for HOME, SEARCH, SITE MAP, GEO Publications, FAQ, MIAME, and Email GEO. The main content area displays the accession number GSE82211 and provides a search interface with options for Scope (Self), Format (HTML), and Amount (Quick). The series information is as follows:

|                        |  |
|------------------------|--|
| <b>Series GSE82211</b> | <a href="#">Query DataSets for GSE82211</a>  |
| Status                 | Public on Dec 28, 2016   |
| Title                  | The lncRNA genomic landscape of oligodendrocytes reveals myelination control by a lncOL1/Suz12 complex in the CNS  |
| Organisms              | <a href="#">Mus musculus</a> ; <a href="#">Rattus norvegicus</a>   |
| Experiment type        | Genome binding/occupancy profiling by high throughput sequencing<br>Non-coding RNA profiling by high throughput sequencing<br>Expression profiling by high throughput sequencing |
| Summary                | This SuperSeries is composed of the SubSeries listed below.  |
| Overall design         | Refer to individual Series   |
| Citation missing       | <i>Has this study been published? Please <a href="#">notify GEO</a>.</i>   |
| Submission date        | Jun 03, 2016   |
| Last update date       | Jan 19, 2017   |
| Contact name           | Richard Lu   |
| Organization name      | Cincinnati Children's Hospital Medical Center  |
| Department             | CBDI   |
| Lab                    | Lu Lab,T6.525  |
| Street address         | 3333 Burnet Ave  |
| City                   | Cincinnati   |
| State/province         | OH   |

- Most journals require authors to submit their data to a database (e.g., GEO) prior to publication.
- These databases entries contain raw data and processed data.
- These data can be use to examine the authors' claims, but also to test new hypotheses.

# Central questions for today's lecture

---

- Where do these data come from?
- How does the way we collect it influence what we know?



# 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. Sanger 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

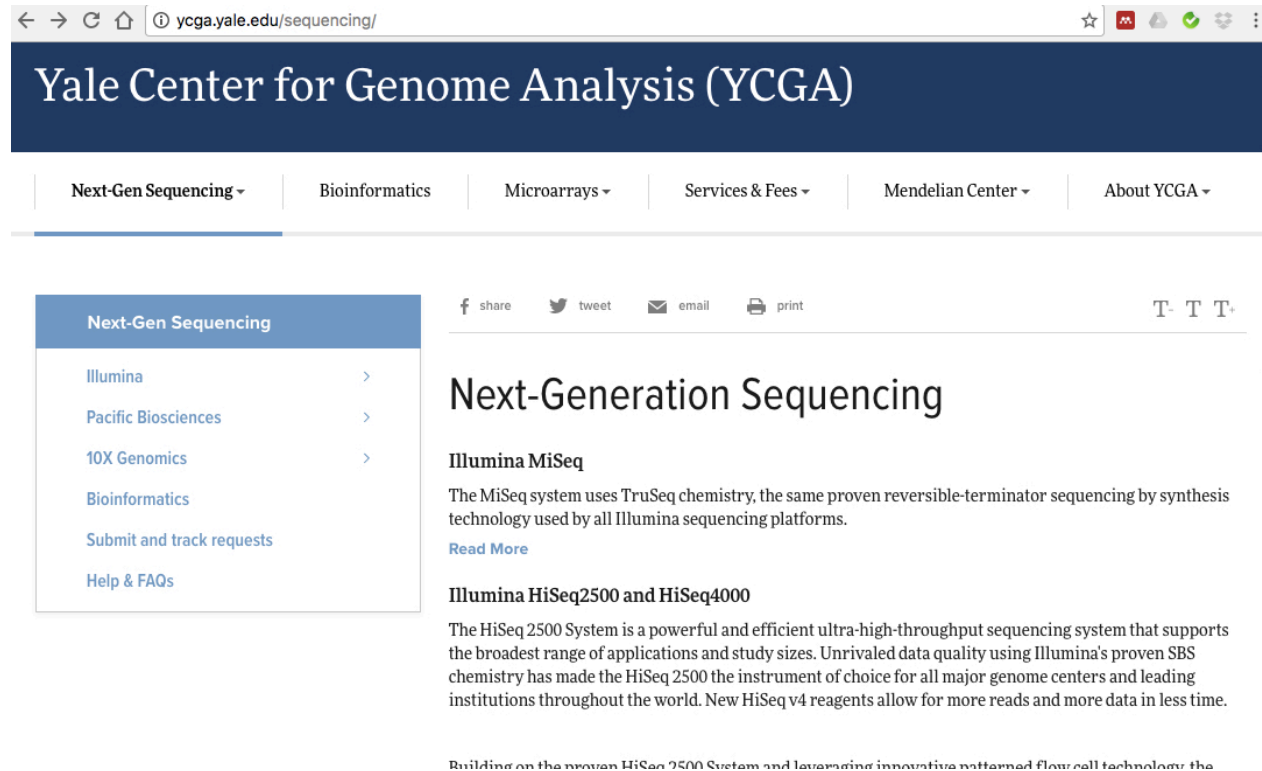
- a. Isolation
- b. Library construction

## 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



The screenshot shows the Yale Center for Genome Analysis (YCGA) website. The browser address bar displays "ycga.yale.edu/sequencing/". The main header is "Yale Center for Genome Analysis (YCGA)". The navigation menu includes "Next-Gen Sequencing", "Bioinformatics", "Microarrays", "Services & Fees", "Mendelian Center", and "About YCGA". The "Next-Gen Sequencing" dropdown menu is open, showing options: "Illumina", "Pacific Biosciences", "10X Genomics", "Bioinformatics", "Submit and track requests", and "Help & FAQs". The main content area is titled "Next-Generation Sequencing" and features social media sharing icons (Facebook, Twitter, Email, Print) and text: "Illumina MiSeq", "The MiSeq system uses TruSeq chemistry, the same proven reversible-terminator sequencing by synthesis technology used by all Illumina sequencing platforms.", "Read More", "Illumina HiSeq2500 and HiSeq4000", "The HiSeq 2500 System is a powerful and efficient ultra-high-throughput sequencing system that supports the broadest range of applications and study sizes. Unrivaled data quality using Illumina's proven SBS chemistry has made the HiSeq 2500 the instrument of choice for all major genome centers and leading institutions throughout the world. New HiSeq v4 reagents allow for more reads and more data in less time.", and "Building on the proven HiSeq 2500 System and leveraging innovative patterned flow cell technology, the".

# 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=DDEFFHHHHHIJIJJJIJJJJJIJJJ?FHIDGIJ=GIHGIIIHGIJIHEHIHHGFFFFEEEDDDDDDDDDDDDD
```


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



# How long are the reads?

---

TATTGCAATATGTTAACAATCTAACAAGGAAAAAATACCCACACAAAACAAAACACAACCCTTAGAACTGTGCTG

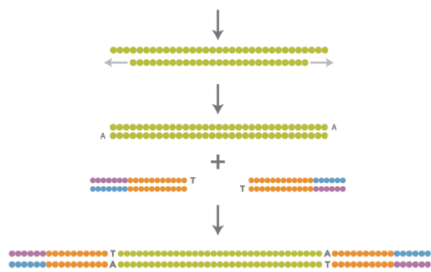


75 nt

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?



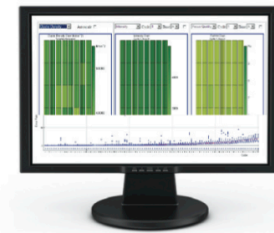
Library Preparation  
~2 h [15 min hands-on (Nextera)]  
< 6 h [< 3 h hands-on (TruSeq)]



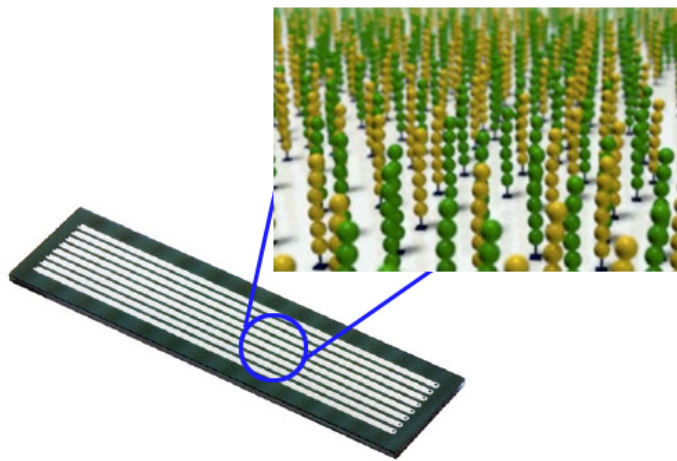
Cluster Generation  
~5 h (<10 min hands-on)



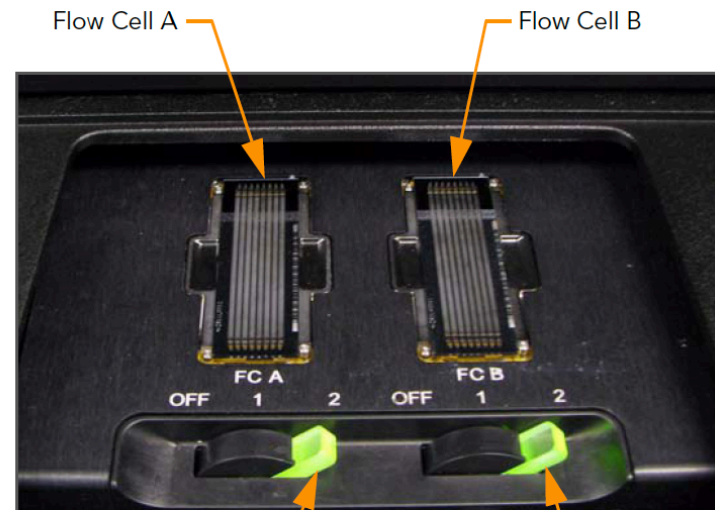
Sequencing by Synthesis  
~1.5 to 11 days



CASAVA  
2 days (30 min hands-on)



Flow cell



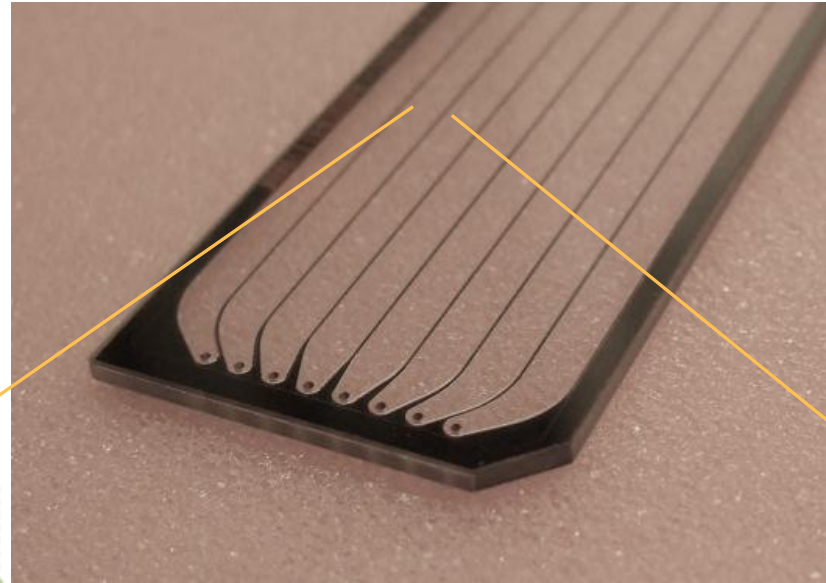
Flow Cell Lever A

Flow Cell Lever B

# What is a flow cell?

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

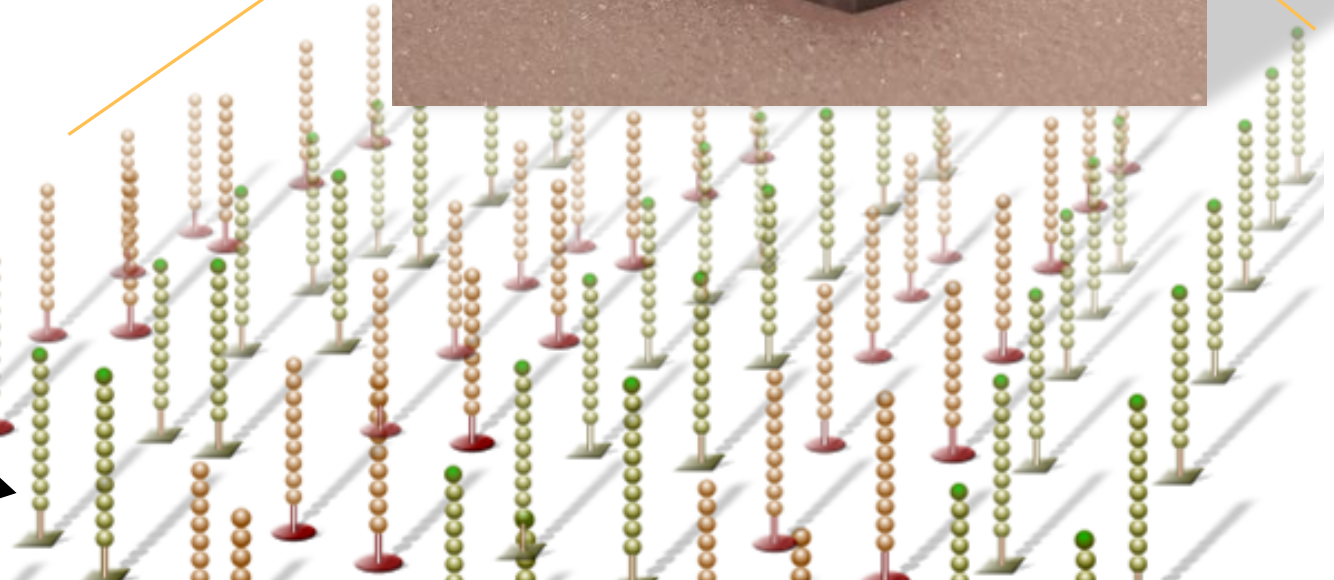
Each lane is randomly coated with a lawn of oligos that are complementary to library adapters



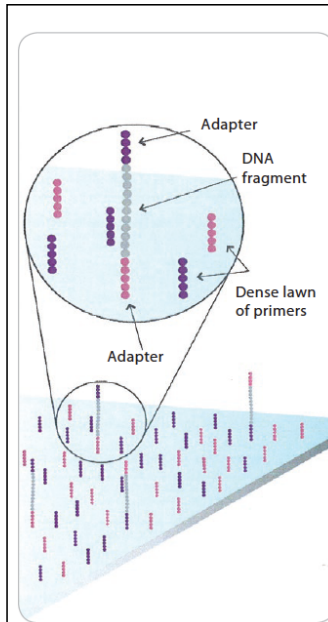
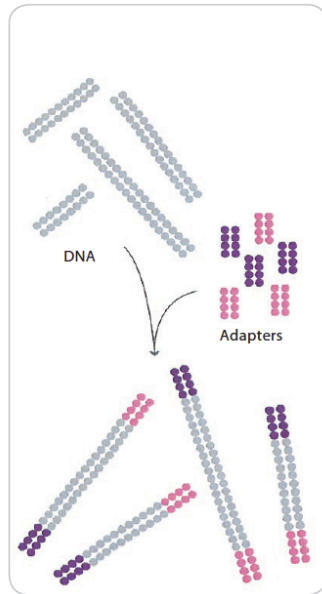
P5 oligo



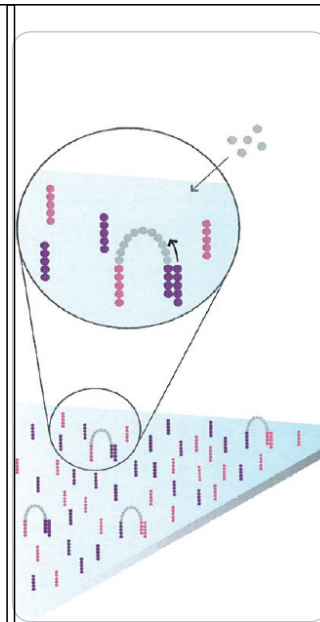
P7 oligo



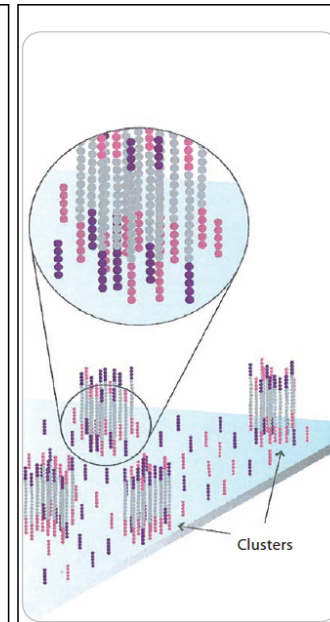
Cluster PCR  
on flow cell  
(8 lanes)



Attach to flow cell

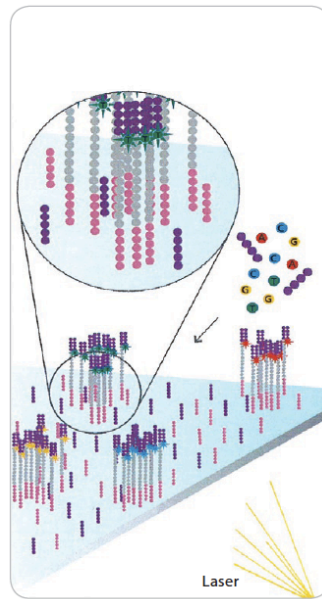


'bridge PCR'



Cluster generation

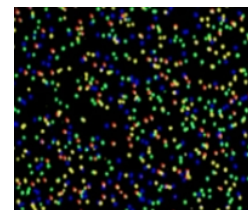
Sequencing  
by synthesis  
with reversible  
dye terminators



Add base

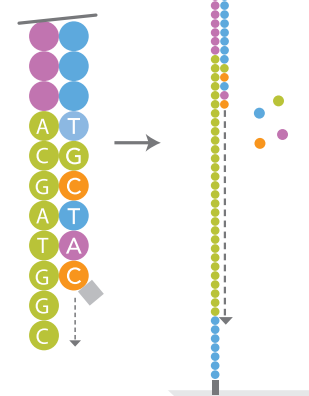


Scan flow cell



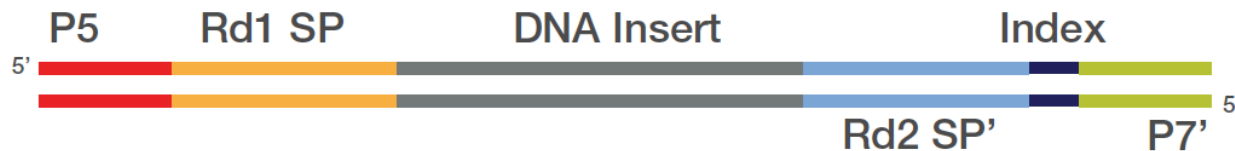
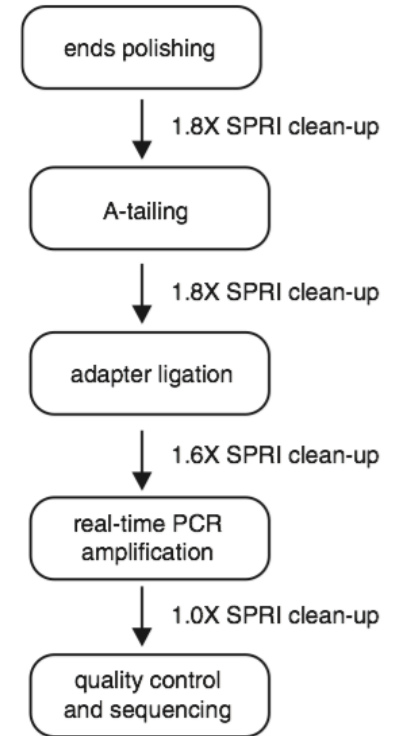
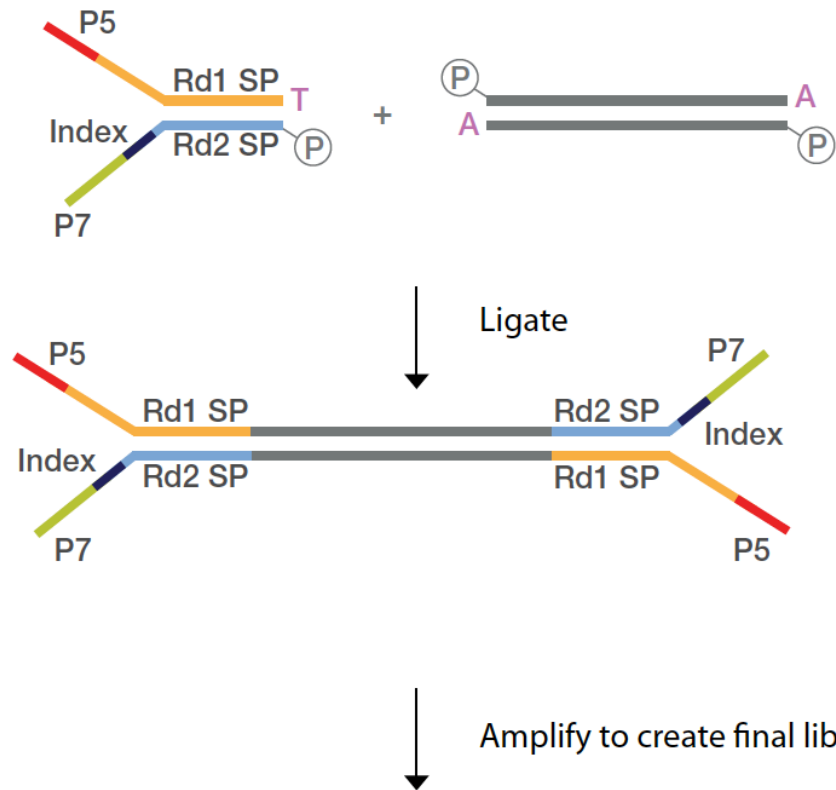
1 cycle

Reverse  
termination  
Add next base



# How do you make a sequencing library?

Index = unique sequence key to identify library



12 samples per lane

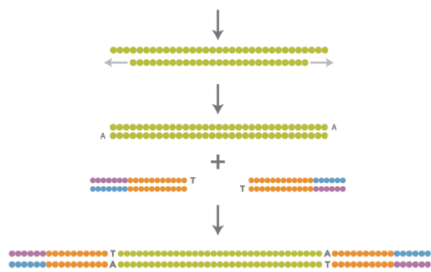
## Potential sources of bias:

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).



# Where do these reads come from?



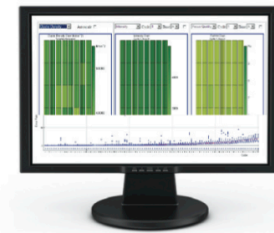
Library Preparation  
~2 h [15 min hands-on (Nextera)]  
< 6 h [< 3 h hands-on (TruSeq)]



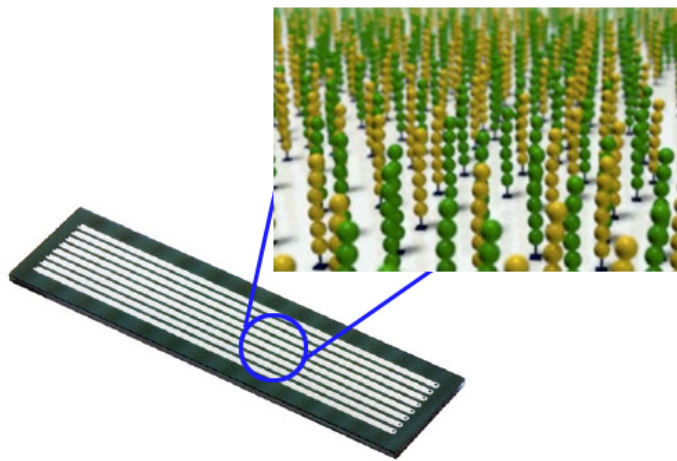
Cluster Generation  
~5 h (<10 min hands-on)



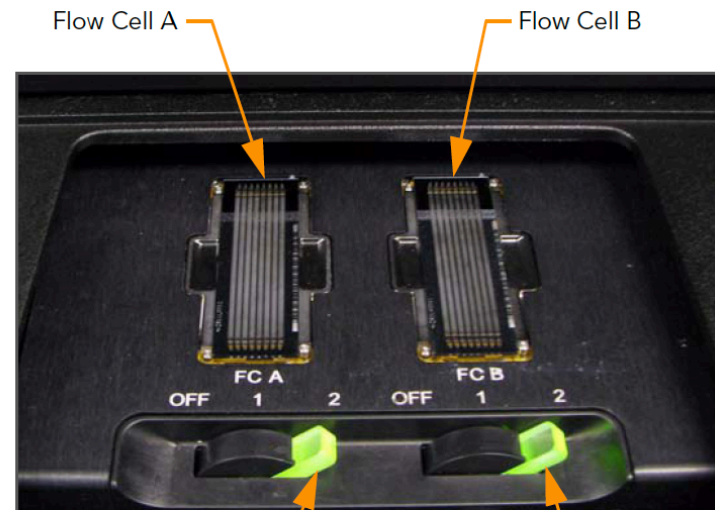
Sequencing by Synthesis  
~1.5 to 11 days



CASAVA  
2 days (30 min hands-on)



Flow cell



Flow Cell Lever A

Flow Cell Lever B



# 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=DDEFFHHHHHIJIJJJIJJJJJIJJJ?FHIDGIJ=GIHGIIIHGIJIHEHIHHGFFFFEEDDDDDDDDDDDDD

@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
NCATCACTTTCTGCACCAGCCATGACGTCAATCTTCGTCCGAACCCCAAACCTCGAGATCGGAAGAGCACACGTCTG
+
#11BBDDDFDFBFFFIIIIIIIIIIIIIFEGIIIIIFIGAGIIFIII=FEFFFFFFFDDD=@9A@BBBBB=?BB<
```

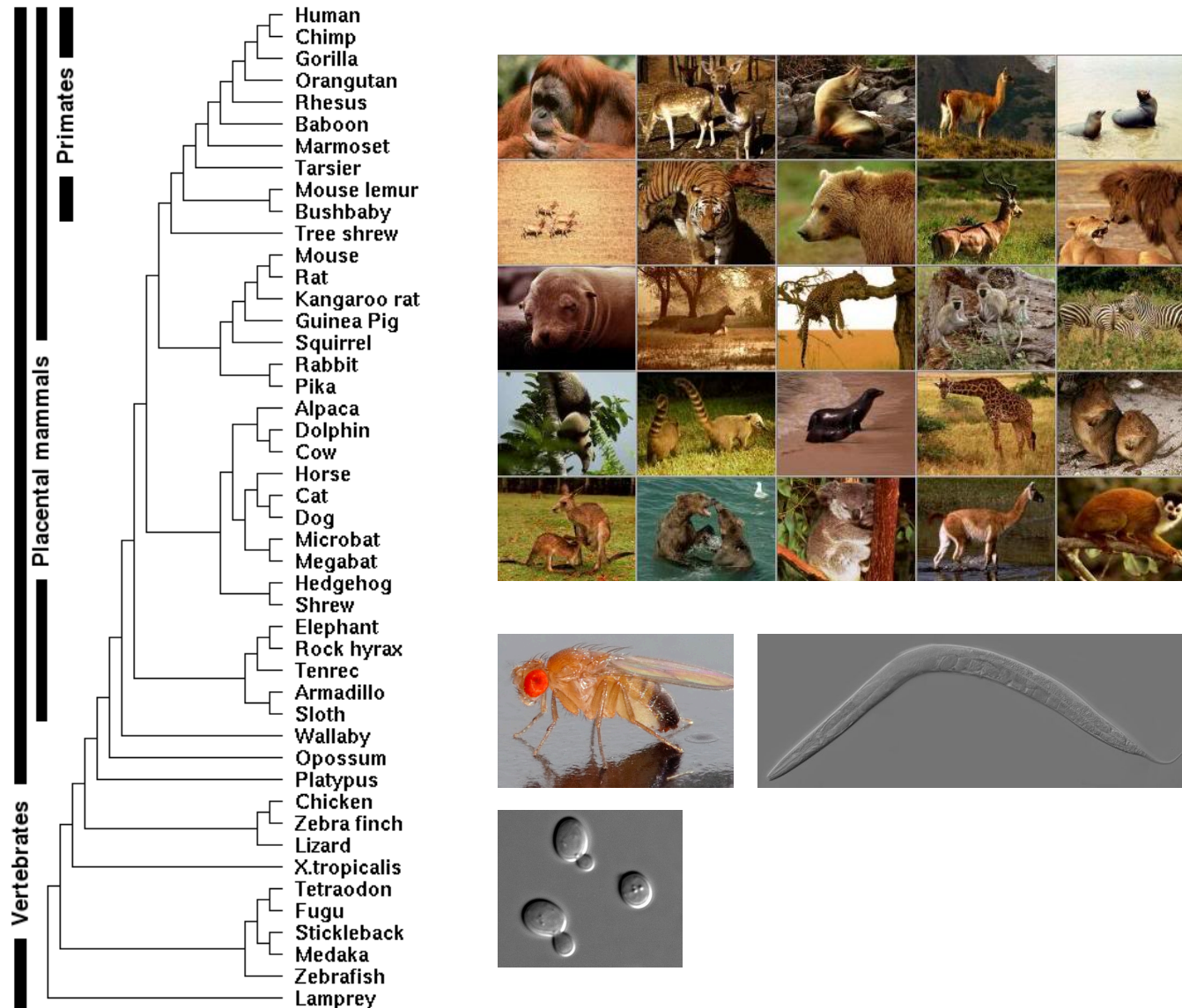
- For each single end read: Incomplete incorporation of bases.
- For the size of the insert (especially for paired end analysis): Ability to get consistent clusters.

# What do I do with my sequencing reads?

---



# Many reference genomes are available



# There is a wide range of genome sizes.

kb = 1000 bp

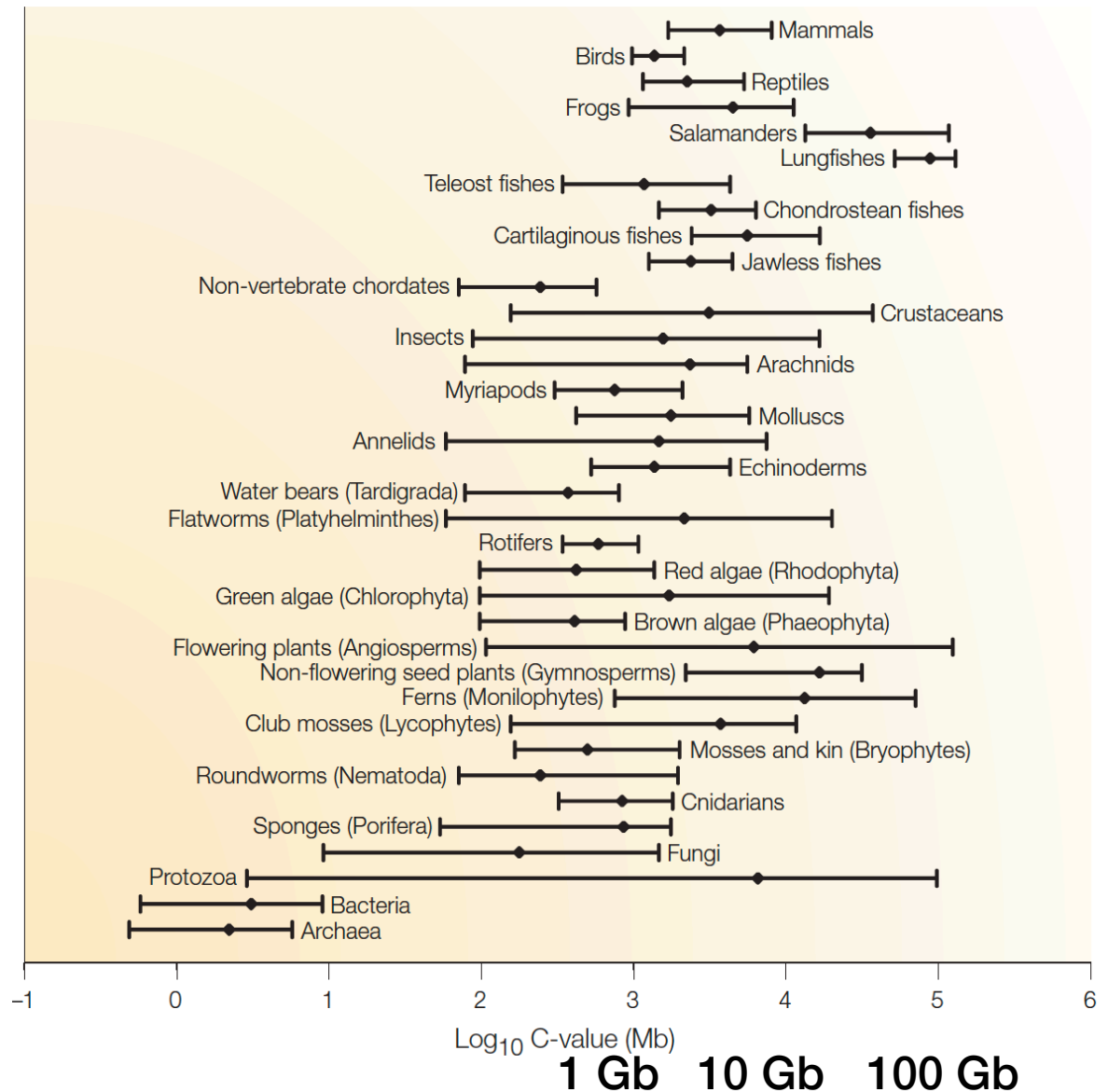
Mb =  $1 \times 10^6$  bp

Gb =  $1 \times 10^9$  bp

Tb =  $1 \times 10^{12}$  bp

Human haploid  
genome ~ 3 Gb

75 nt x  $3 \times 10^8$  reads/lane is  
about the right scale, but the  
amount of **coverage** necessary  
depends on application.





# Sequencing of the human genome

---

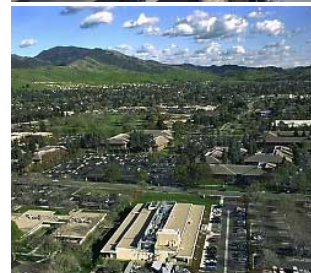
Victory declared **2003**



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



National Human  
Genome Research  
Institute



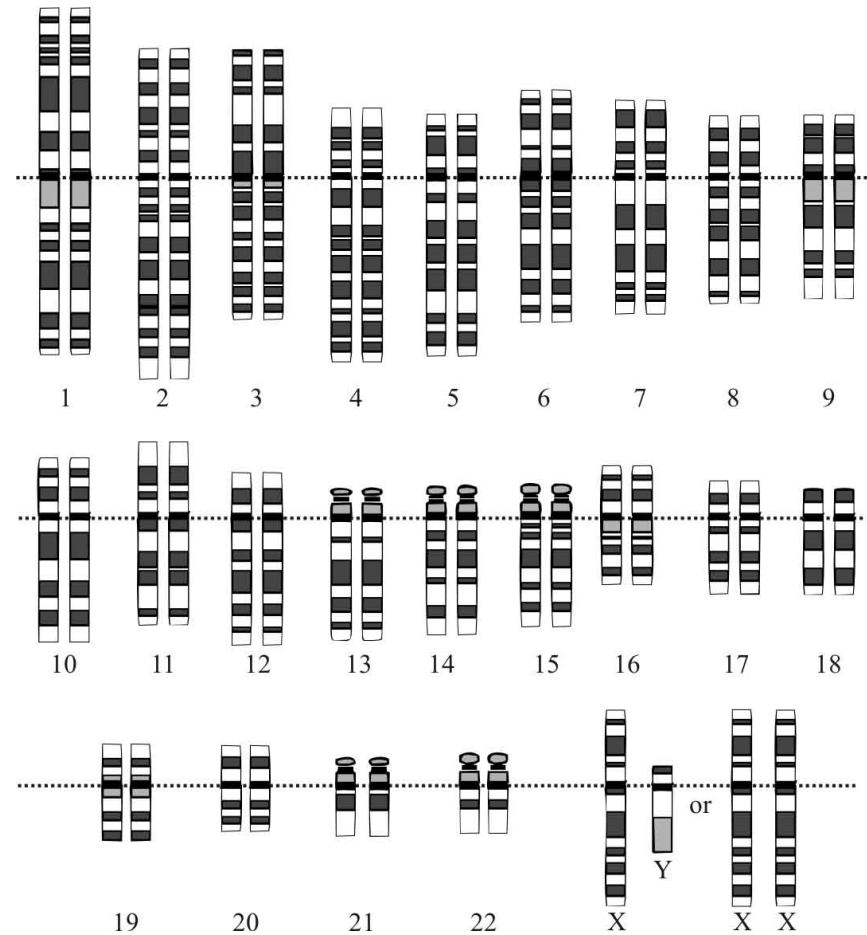
Newest illumina sequencers claim 6000 Gb/run (2017)

# Assembling a genome from short reads



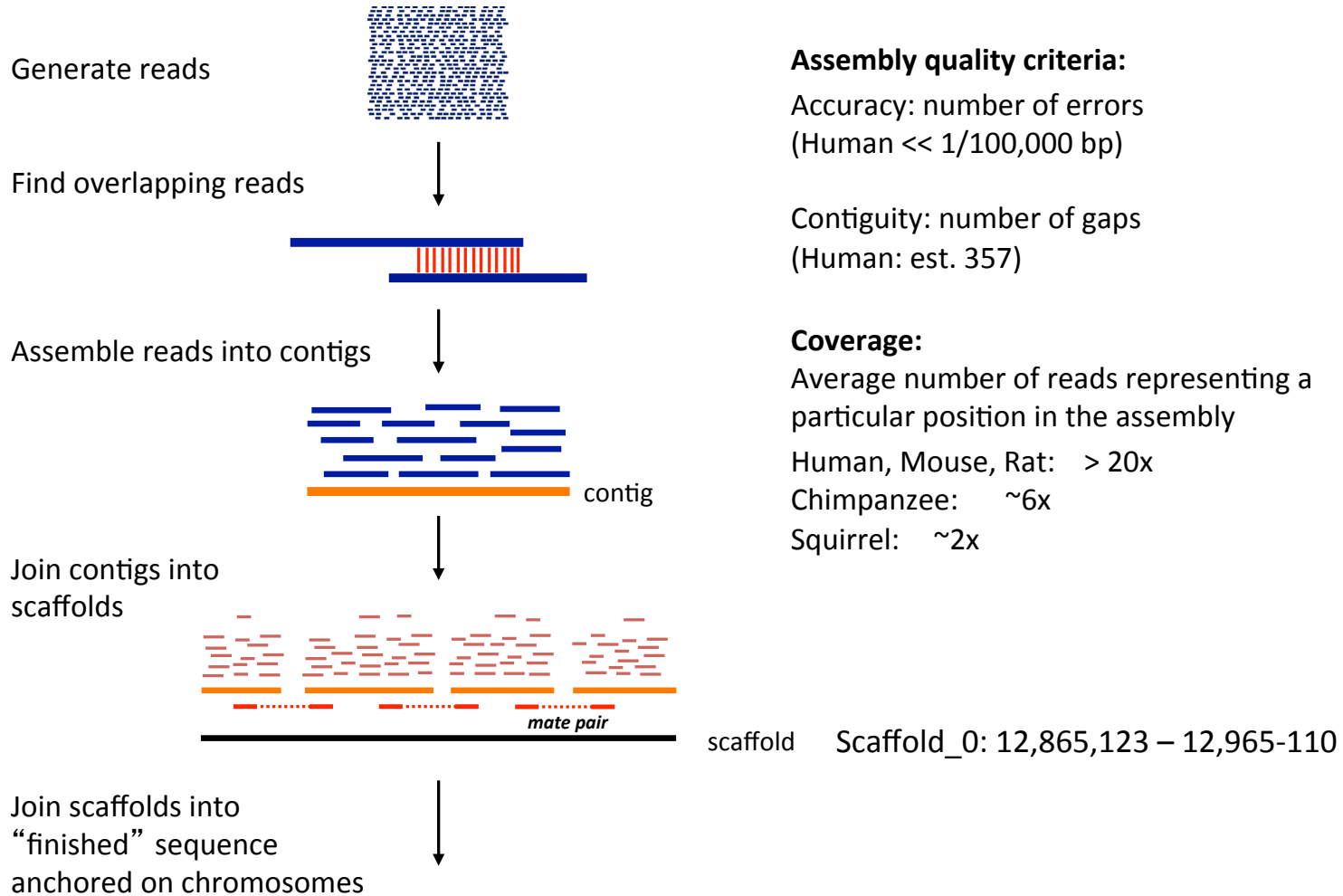
$\gg 10^9$  sequencing reads

36 bp - 1 kb



3 Gb

# How to assemble a genome

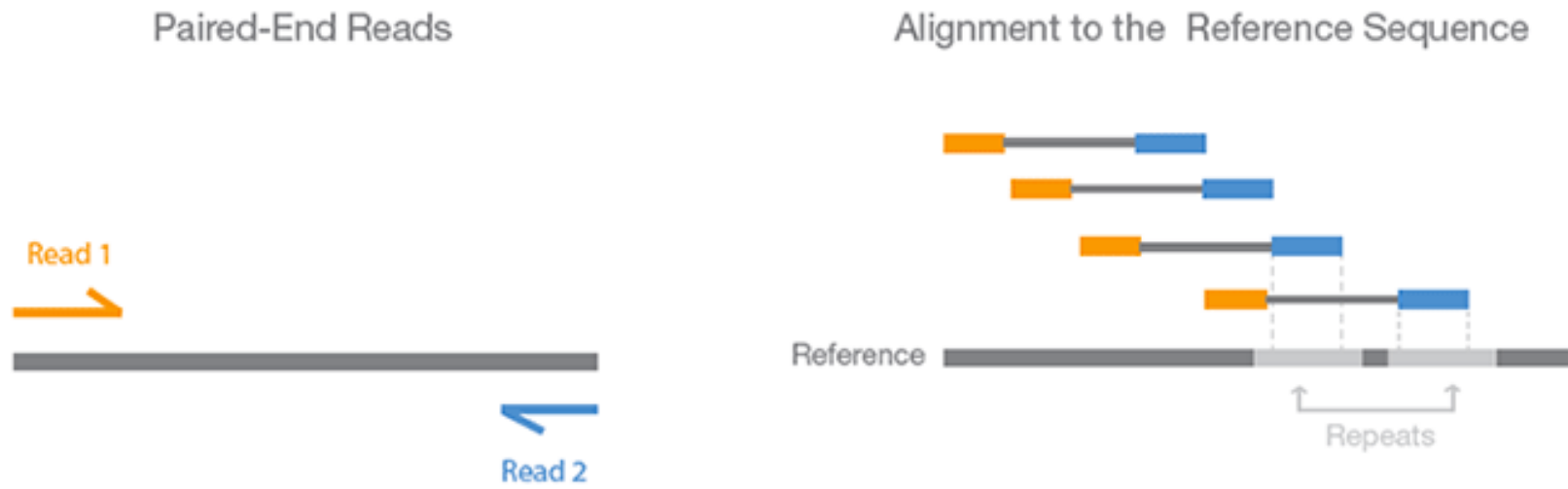


AGTTGTATTATTAGAACTGAGGGCTAAAACTGTGCACATACACAGACACACATATTATTTTAATATAGATTTTCAATAATTGGTCTAGGATAAGGATAATATACAG

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.

CCAAATCAAACAGTTGTATTATTAGAAACTGAGGGCTAAAACTGTGCACATACACAGACACACATATTATTTAATATAGATTTTCAATAATTGGTCTAGGATAAG  
AGCAAGAAGAAAACAAAGACTGTTACTATGGAAAAATGAAAATAGATTTTAAAACATGTTAATTCACGTTACTTTTTGTTAAATTTACTTTTCTTCTTTCACTTCTT  
AATAAATCACATTAATTCCTTATCTCATGTGAAATTTTATGATTGATACCTTTAAATGTCATTTGTTGAAGGAAGATTATTCATTTTTTTCATTCAATAAATATTT  
CAGTATTATGTTCTAGGCATTGGGGATACCATGTTTACAAGACAGACTATGATTTACAGGATCAGATGTGGACTCTCAAATTCGACTGAGAATAAAAACAGACACT  
TAATTGATGCTAGAAAGACAATGAAACAGAGCCATGTGACCAATGAGAGAGATGAGGGTGGCAGCAGCCTGTTTTAGATAAGGTACCTGATTGGTGGGATTGG  
TATGCCTTAATGATATGAAAGAACCATTGATGGGAAGGCCTAGCATTAAAACCGTCTAGGCAGAATGAGCAGCAAGTGCAAGGGTCTGGATAGGAATGAGC  
ATGGAAAAATGAAAATAGATTTTAAAACATGTTAATTCACGTTACTTTTTGTTAAATTTACTTTTCTTCTTTCACTTCTTACCTGTCAATGTTATTAATATTTT  
GAAATTTTCAATTTATGATTGATACCTTTAAATGTCATTTGTTGAAGGAAGATTATTCATTTTTTTCATTCAATAAATATTTTTTAGAATAATAAGTCCCAGGCACAAGA  
CATGTTTACAAGACAGACTATGATTTACAGGATCAGATGTGGACTCTCAAATTCGACTGAGAATAAAAACAGACACTAAACAAGTAAATAAAGTTAATTTCAAGTT  
GATTTTAAAACATGTTAATTCACGTTACTTTTTGTTAAATTTACTTTTCTTCTTTCACTTCTTACCTGTCAATGTTATTAATATTTTTTAGGAACAATAAATCACATTA  
ATTGATACCTTTAAATGTCATTTGTTGAAGGAAGATTATTCATTTTTTTCATTCAATAAATATTTTTTAGAATAATAAGTCCCAGGCACAAGACCAGTATTATGTTCTAG  
ACTATGATTTACAGGATCAGATGTGGACTCTCAAATTCGACTGAGAATAAAAACAGACACAAACAAGTAAATAAAGTTAATTTCAAGTTGTAATTGATGCTATCCCA  
TTGGGGATACCATTACCTGTCAATGTTATTAATATTTTTTAGGAACAATAAATCACATTAATTTCCAACATGCAAAGAGGAAATCTCCATATCATGCTTGTCAATTCGTT  
GTGTGTAACAAATTCTCAGAATTTTAAACAATAACAAATCAGGGCTGAATGTGGCCAACATGCAAAGAGGAAATCTCCATCTGTCCAAATCAAACAGTTGTATT  
CATAACAGACACACATATTATTTAATATAGATTTTCAATAATTGGTCTAGGATAAGGATAATATACAGAGAACATGCCAAAAGTTAAGCAAGAAGAAAACAAAG  
TAAAACATGTTAATTCACGTTACTTTTTGTTAAATTTACTTTTCTTCTTTCACTTCTTACCTGTCAATGTTATTAATATTTTTTAGGAACAATAAATCACATTAATTCCT  
TACCTTTAAATGTCATTTGTTGAAGGAAGATTATTCATTTTTTTCATTCAATAAATATTTTTTAGAATAATAAGTCCCAGGCACAAGACCAGTATTATGTTCTAGGCAT  
GATTTACAGGATCAGATGTGGACTCTCAAATTCGACTGAGAATAAAAACAGACACTAAACAAGTAAATAAAGTTAATTTCAAGTTGTAATTGATGCTAGAAAGACA  
AGATGAGGGTGGCAGCAGCCTGTTTTAGATAAGGTACCTGATTGGTGGGATTGGAAGACCTCTCTGAGATTAGTGTCTTCAGATATGCCTTAATGATATGAAAG  
AACCGTCTAGGCAGAATGAGCAGCAAGTGCAAGGGTCTGGATAGGAATGAGCTGGATATACTCAAGGAAGAAAGAGAAACTATGGAAAAATGAAAATAGATT  
TAAATTTACTTTTCTTCTTTCACTTCTTACCTGTCAATGTTATTAATATTTTTTAGGAACAATAAATCACATTAATTCCTTATCTCATGTGAAATTTTCAATTTATGATTGA  
TATTTCATTTTTTTCATTCAATAAATATTTTTTAGAATAATAAGTCCCAGGCACAAGACCAGTATTATGTTCTAGGCATTGGGGATACCATGTTTACAAGACAGACTAT  
ATTTCGACTGAGAATAAAAACAGACACTAAACAAGTAAATAAAGTTAATTTCAAGTTGTAATTGATGCTACTATGGAAAAATGAAAATAGATTTTAAAACATGTTAATTC  
TTTCACTTCTTACCTGTCAATGTTATTAATATTTTTTAGGAACAATAAATCACATTAATTCCTTATCTCATGTGAAATTTTCAATTTATGATTGATACCTTTAAATGTCAT  
CAATAAATATTTTTTAGAATAATAAGTCCCAGGCACAAGACCAGTATTATGTTCTAGGCATTGGGGATACCATGTTTACAAGACAGACTATGATTTACAGGATCAG  
AACAGACACAAACAAGTAAATAAAGTTAATTTCAAGTTGTAATTGATGCTATCCAGGCACAAGACCAGTATTATGTTCTAGGCATTGGGGATACCATACCTGT  
CACATTAATTTCCAACATGCAAAGAGGAAATCTCCATATCATGCTTGTCAATTCGTTTATCAGAGGCCAAATGTTTTTCTTGTAAACGTGTGTAAACATTCTCAGA  
GTGGCCAACATGCAAAGAGGAAATCTCCATCTGTCCAAATCAAACAGTTGTATTATTAGAAACTGAGGGCTAAAACTGTGCACATACACAGACACACATATTA  
GATAAGGATAATATACAGAGAACATGCCAAAAGTTTAAAGCAAGAAGAAAACAAAGACTGTTACTATGGAAAAATGAAAATAGATTTTAAAACATGTTAATTCACGT  
CTTCTTACCTGTCAATGTTATTAATATTTTTTAGGAACAATAAATCACATTAATTCCTTATCTCATGTGAAATTTTCAATTTATGATTGATACCTTTAAATGTCATTTGTT  
ATATTTTTTAGAATAATAAGTCCCAGGCACAAGACCAGTATTATGTTCTAGGCATTGGGGATACCATGTTTACAAGACAGACTATGATTTACAGGATCAGATGTG  
ACACTAAACAAGTAAATAAAGTTAATTTCAAGTTGTAATTGATGCTAGAAAGACAATGAAACAGAGCCATGTGACCAATGAGAGAGATGAGGGTGGCAGCAGCC  
ATTGGAAGACCTCTCTGAGATTAGTGTCTTCAGATATGCCTTAATGATATGAAAGAACCATTGATGGGAAGGCCTAGCATTAAAACCGTCTAGGCAGAATGAG  
GAGCTGGATATACTCAAGGAAGAAAGAGAAACTATGGAAAAATGAAAATAGATTTTAAAACATGTTAATTCACGTTACTTTTTGTTAAATTTACTTTTCTTCTTTCA  
GGAACAATAAATCACATTAATTCCTTATCTCATGTGAAATTTTCAATTTATGATTGATACCTTTAAATGTCATTTGTTGAAGGAAGATTATTCATTTTTTTCATTCAATA  
AAGACCAGTATTATGTTCTAGGCATTGGGGATACCATGTTTACAAGACAGACTATGATTTACAGGATCAGATGTGGACTCTCAAATTCGACTGAGAATAAAAACAG  
AGTTGTAATTGATGCTACTATGGAAAAATGAAAATAGATTTTAAAACATGTTAATTCACGTTACTTTTTGTTAAATTTACTTTTCTTCTTTCACTTCTTACCTGTCAAT  
ATTAATTCCTTATCTCATGTGAAATTTTCAATTTATGATTGATACCTTTAAATGTCATTTGTTGAAGGAAGATTATTCATTTTTTTCATTCAATAAATATTTTTTAGAATA  
TTCTAGGCATTGGGGATACCATGTTTACAAGACAGACTATGATTTACAGGATCAGATGTGGACTCTCAAATTCGACTGAGAATAAAAACAGACACAAACAAGTAA  
ATCCAGGCACAAGACCAGTATTATGTTCTAGGCATTGGGGATACCATACCTGTCAATGTTATTAATATTTTTTAGGAACAATAAATCACATTAATTTCCAACATGCA  
TCGTTTATCAGAGGCCAAATGTTTTTCTTGTAAACGTGTGTAAACATTCTCAGAATTTTAAACAATAACAAATCAGGGCTGAATGTGGCCAACATGCAAAGAG  
GTGATTATTAGAAACTGAGGGCTAAAACTGTGCACATACACAGACACACATATTATTTAATATAGATTTTCAATAATTGGTCTAGGATAAGGATAATATACAGAGA  
CAAAGACTGTTACTATGGAAAAATGAAAATAGATTTTAAAACATGTTAATTCACGTTACTTTTTGTTAAATTTACTTTTCTTCTTTCACTTCTTACCTGTCAATGTT  
TTCTTATCTCATGTGAAATTTTCAATTTATGATTGATACCTTTAAATGTCATTTGTTGAAGGAAGATTATTCATTTTTTTCATTCAATAAATATTTTTTAGAATAATAAGT

# What types of annotation do we have/want?

---

**~3 billion bp**

```
ACAATAAATCACATTAATTCTTATCTCATGTGAAATTCATATTTATGATTG
ATACCTTTAAATGTCATTTGTTGAAGGAAGATTATTCATTTTTTCATTCAAT
AAATATTTTTAGAAATAAAGTCCCAGGCACAAGACCAGTATTATGTTCT
AGGCATTGGGGATACCATGTTCAAGACAGACTATGATTACAGGATC
AGATGTGGACTCTCAAATTCGACTGAGAATAAACAGACACTAAACAAG
TAAATAAAGTTAATTTCAAGTTGTAATTGATGCTAGAAAAGACAATGAACA
GAGCCATGTGACCAATGAGAGAGATGAGGGTGGCAGCAGCCTGTTTTA
GATAAGGTACCTGATTGGTGGGATTGGAAGACCTCTCTGAGATTAGTGT
CTTCAGATATGCCTTAATGATATGAAAGAACCATTTCATGGGAAGGCCTAG
CATTAAAAACCGTCTAGGCAGAATGAGCAGCAAGTGCAAGGGTCCTGG
ATAGGAATGAGCTGGATATACTCAAGGAAGAAAGAGAAACTATGGAAAA
ATGAAAAATAGATTTTAAACATGTTAATTCACGTTACTTTTTGTTAAATTTA
CTTTTCTTCTTTCACCTCTTACCTGTCAATGTTAATATTTTTAGGAACA
ATAAATCACATTAATTCCTTATCTCATGTGAAATTCATATTTATGATTGATA
CCTTTAAATGTCATTTGTTGAAGGAAGATTATTCATTTTTTCATTCAATAAA
TATTTTTAGAAATAAAGTCCCAGGCACAAGACCAGTATTATGTTCTAGG
CATTGGGGATACCATGTTCAAGACAGACTATGATTACAGGATCAGAT
GTGGACTCTCAAATTCGACTGAGAATAAACAGACACTAAACAAGTAAAT
AAAGTTAATTTCAAGTTGTAATTGATGCTACTATGGAAAAATGAAAAATAGA
TTTTAAACATGTTAATTCACGTTACTTTTTGTTAAATTTACTTTTCTCTTT
CACTTCTTACCTGTCAATGTTAATATTTTTAGGAACAATAAATCACATT
AATTCCTTATCTCATGTGAAATTCATATTTATGATTGATACCTTTAAATGT
CATTTTGTGAAGGAAGATTATTCATTTTTTCATTCAATAAATATTTTTAGA
ATAATAAGTCCCAGGCACAAGACCAGTATTATGTTCTAGGCATTGGGGAT
ACCATGTTCAAGACAGACTATGATTTACAGGATCAGATGTGGACTCTC
AAATTCGACTGAGAATAAACAGACACAACAAGTAAATAAAGTTAATTT
CAAGTTGTAATTGATGCTATCCAGGCACAAGACCA....
```

## **Genes:**

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

## **Genetic variation:**

- SNPs and CNVs

## **Sequence conservation**

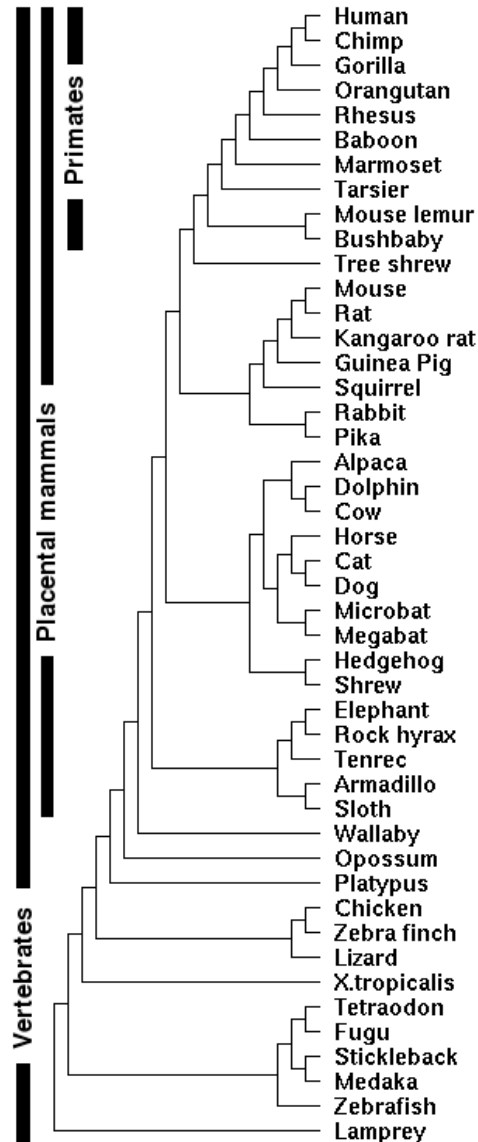
## **Regulatory sequences:**

- Promoters
- Enhancers
- Insulators

## **Epigenetics:**

- DNA methylation
- Chromatin

# Degrees of genomic annotation vary widely



## 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](http://genome.ucsc.edu)):

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

## **ENSEMBL** ([ensembl.org](http://ensembl.org)):

Visualization, data recovery, simple analysis

## **Integrative Genomics Viewer**

([broadinstitute.org/software/igv/](http://broadinstitute.org/software/igv/)):

Local genome viewer (visualize local and remote data)

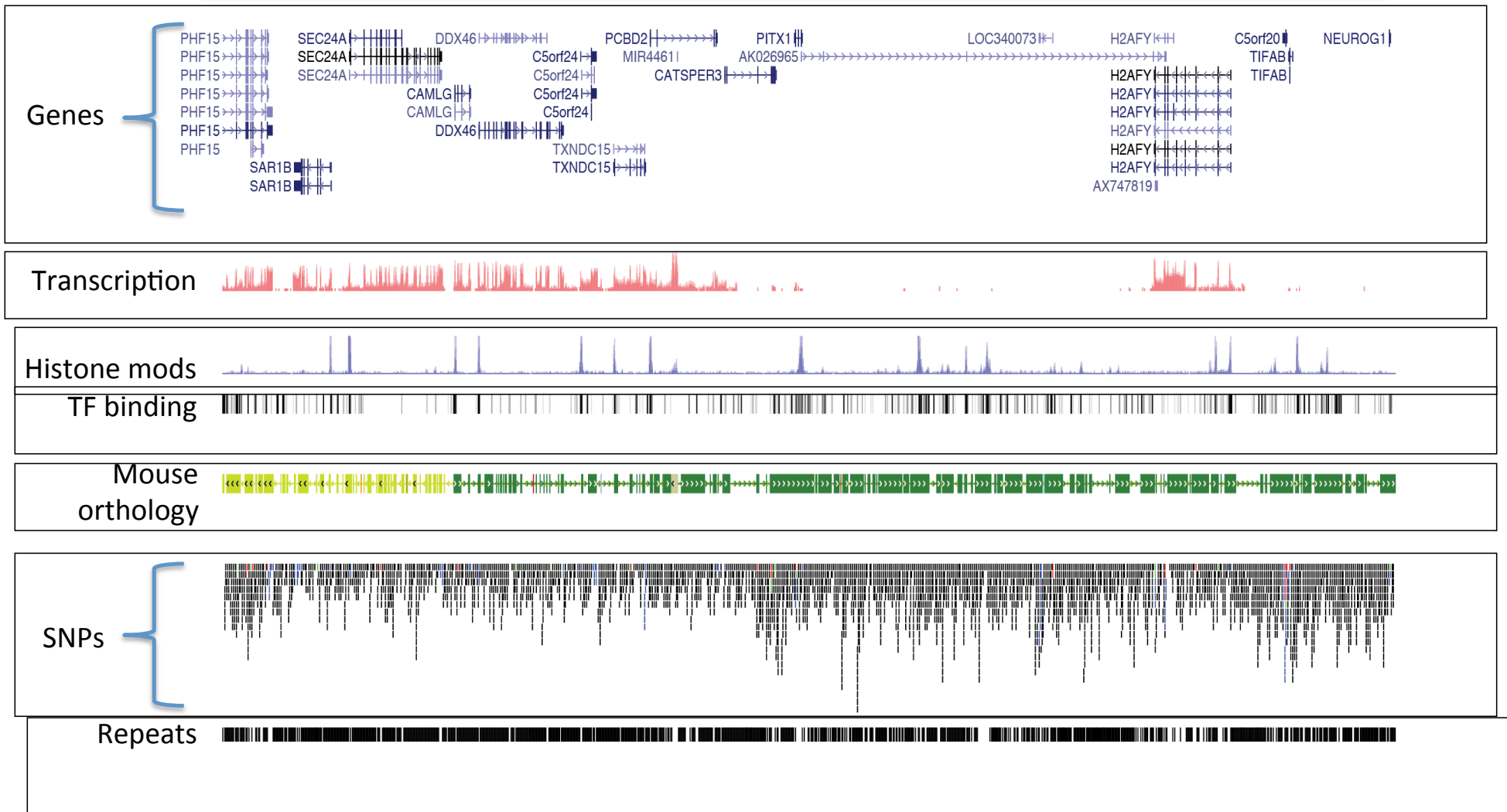
## **Galaxy** ([main.g2.bx.psu.edu](http://main.g2.bx.psu.edu)):

Complex data analysis and workflows



# Example of a genome browser track

Chr5: 133,876,119 – 134,876,119





# 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.

# 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**

Extra slides (in case there is time)

---

# 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
  - Whole-genome and targeted resequencing
- Counting applications

## 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



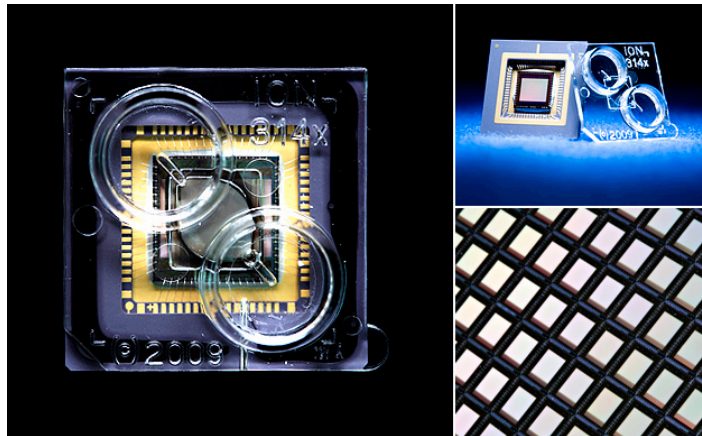
- Run-times
  - 50 cycle – 4 hours
  - 300 cycle – 27 hours
- Two sequencing options
  - 50 cycles
  - 300 cycles (2x150 bp)
- One lane
  - 6-7 million clusters
  - Up to 8 billion bases (300 cycles)

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

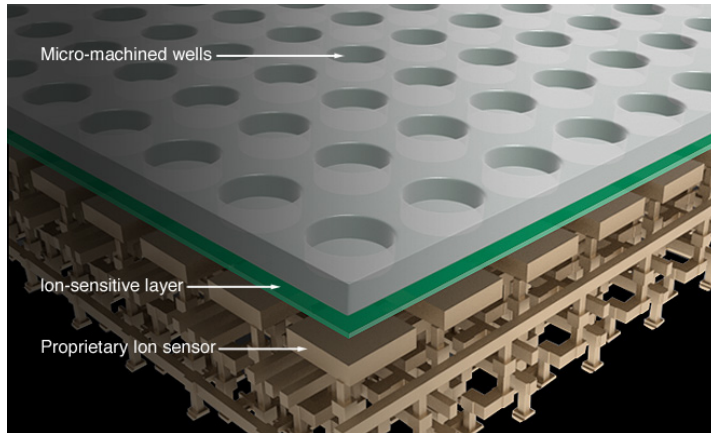
# Ion Torrent and Ion Proton



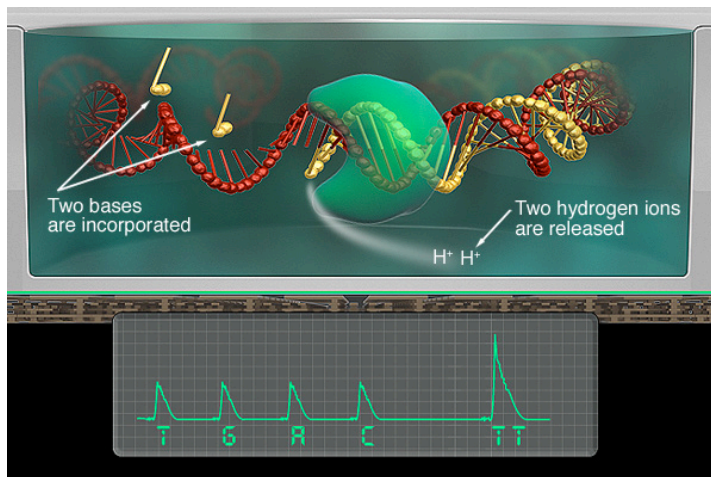
Sequencing on  
semiconductor chip



# Ion Torrent sequencing chemistry

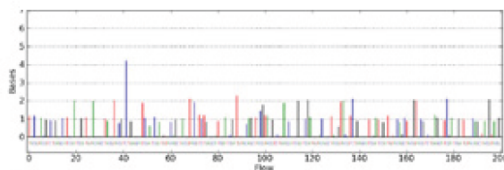


When a nucleotide is incorporated into a strand of DNA, a proton is released as a byproduct.



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

Single-Well Ionogram for [2406, 1991]



# 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

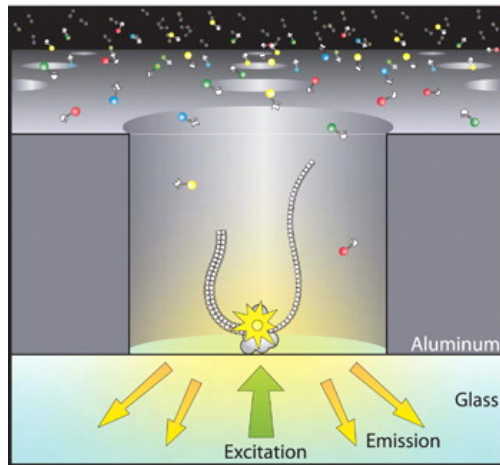
## Pacific Biosciences

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



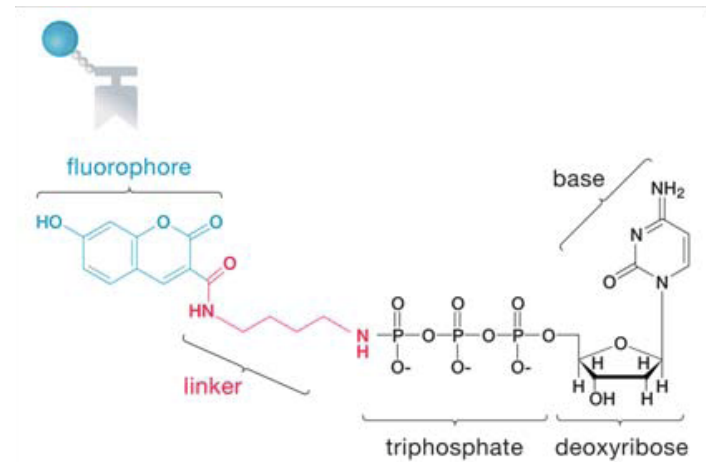
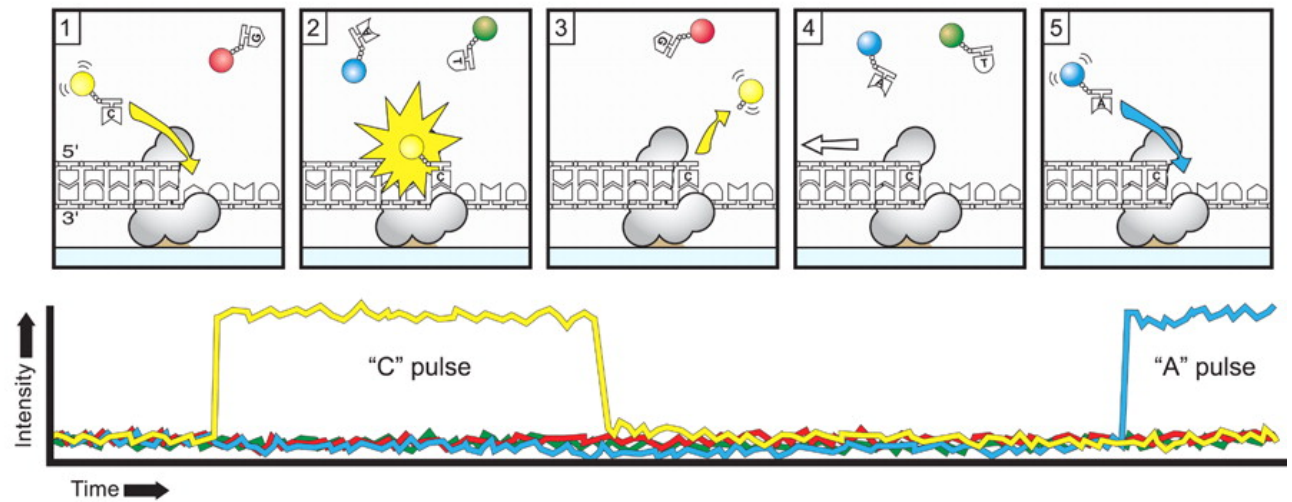
# Sequencing in real time: Pacific Biosciences

## A SMRT cells



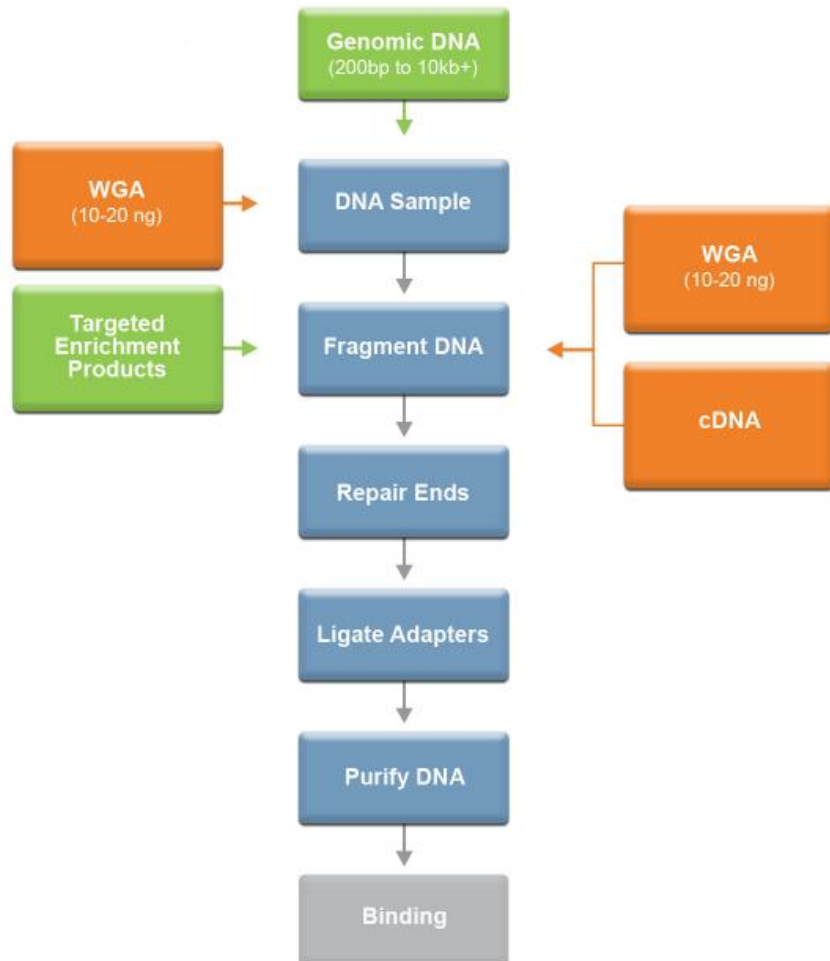
Zero Mode Waveguides

## B

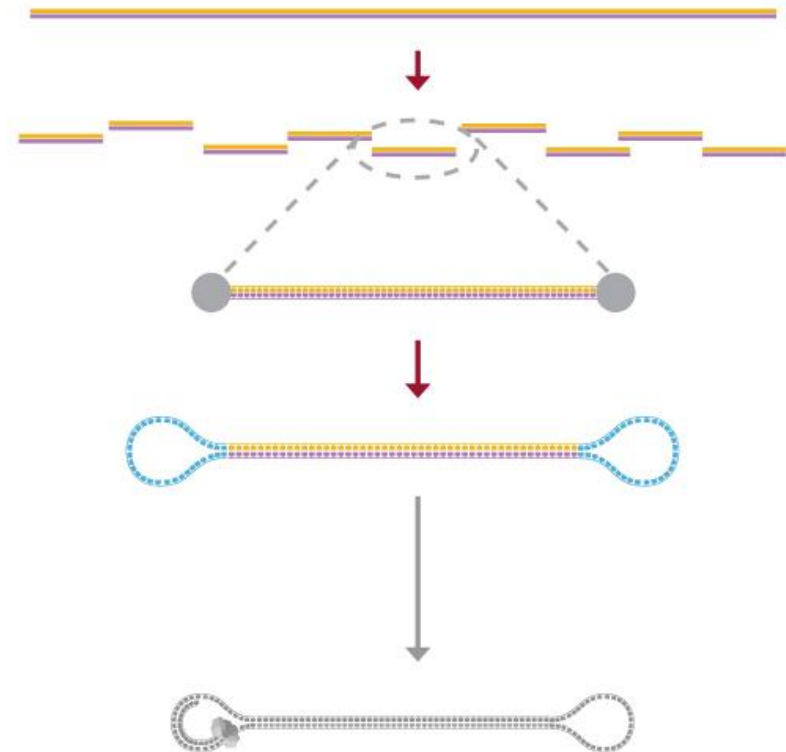


# PacBio sequencing strategy

## Sample Preparation



## Building of SMRTbell





# Applications


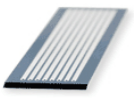
- Targeted sequencing
  - SNP and structure variants detection
  - Repetitive regions
  - Full length transcript profiling
- De novo assembly and genome finishing
  - Bacterial genomes
  - Fungal genomes
  - Gap-captured sequencing
  - Targeted captured sequencing
- Base modifications detection
  - Methylation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

\*\*Projects at YCGA



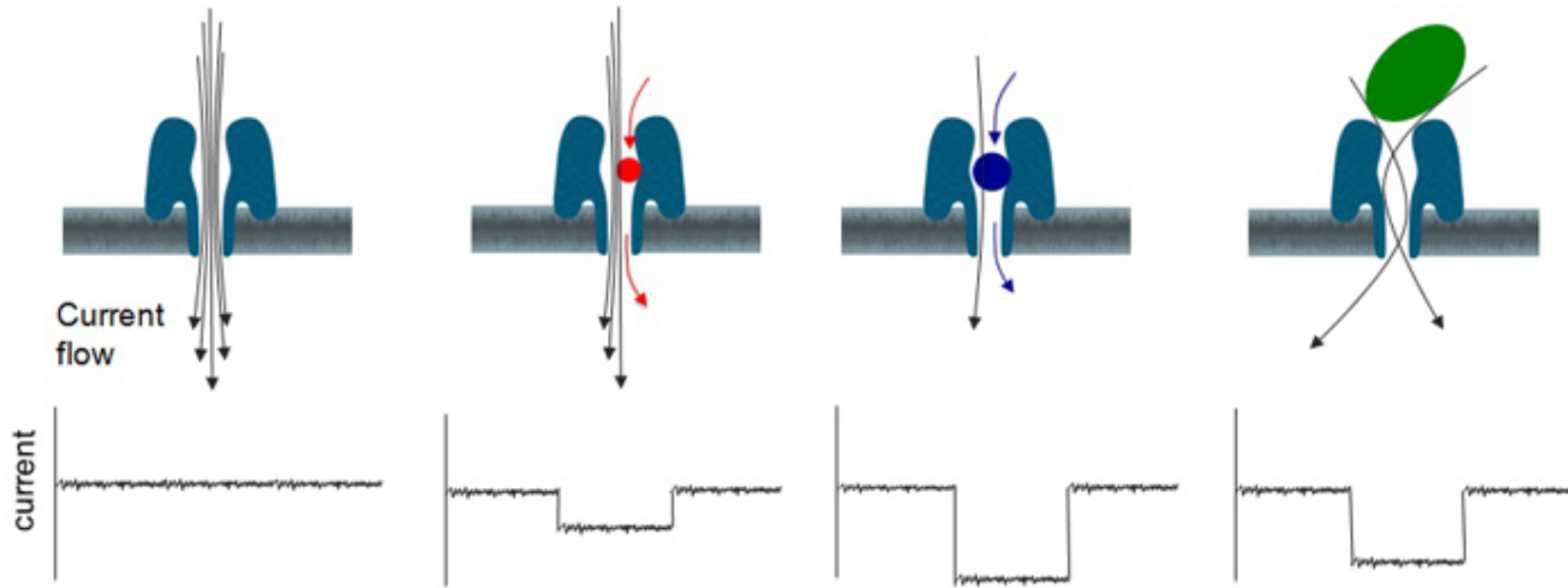
YCGA PacBio RS

# PacBio vs Illumina

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



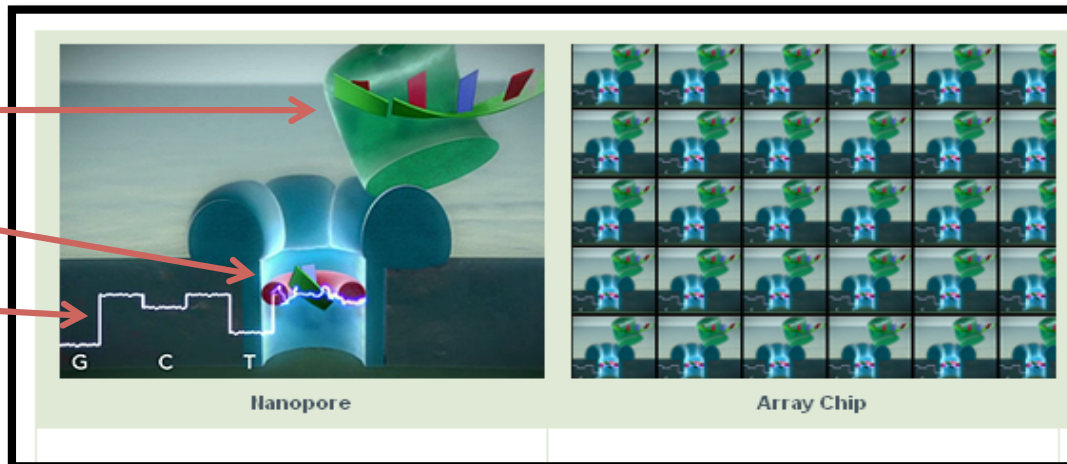
# Oxford Nanopore



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