

# Genomics

---

Matt Simon  
Dept. of Molecular Biophysics & Biochemistry  
Chemical Biology Institute  
January 18, 2019

# 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**

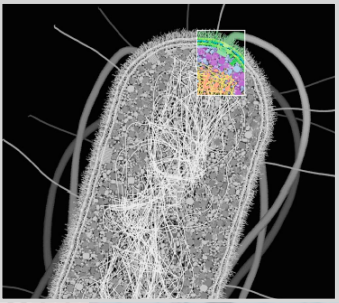
Gene expression and regulation

Cellular identity, differentiation and development

Phenotypic variation among individuals and species

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

CCATGTTCAACAAGACAGAC TATGAT TACAGGATCAGATG GGGACTCTCAAATTCGACTGAGAATAAAACAGACACTA  
TAGATTTTAAAACATGTTAATTCACGTTACTTTTTGTTAAATTTACTTTTTCTTCTTTCACTTCTTACCTGTCAATGTTATTAA  
GATT GTCATTTGTTGAAGGAAGATTATTCATTTTTT CATTCAATAAATATTTTTTAGAATAATAAGTCC  
GACT ATCAGATGTGGACTCTCAAATTCGACTGAGAATAAAACAGACACAAACAAGTAAATAAAGTTA  
ATTG CCTGTCAATGTTATTAATATTTTTTAGGAACAATAAATCACATTAATTCCAACATGCAAAGAGGAA  
CGTC ECAGCAATTTTAAAACAATAAACAATTCAGGGCTGAATGTGGCCAACATGCAAAGAGGAAATCTC  
ACAT TGGTCTAGGATAAGGATAATATACAGAGAACATGCC  
TTA CTTCTTTCACTTCTTACCTGTCAATGTTATTAATATTT  
ATACCTT CATTCAATAAATATTTTTTAGAATAATAAGTCCCAGGC  
ATGATTTA AATAAAACAGACACTAAACAAGTAAATAAAGTTAATTT  
GAGATGA ATTGGTGGGATTGGAAGACCTCTCTGAGATTAGTGT  
AAACCGT GATAGGAATGAGCTGGATATACTCAAGGAAGAAAG  
TTAAATTT AATATTTTTAGGAACAATAAATCACATTAATTCCTTAT  
ATTATTCA CCAGGCACAAGACCAGTATTATGTTCTAGGCATTGC  
AATTCGA TTAATTTCAAGTTGTAATTGATGCTACTATGGAAAAA  
CTTTCAC AATAATCACATTAATTCCTTATCTCATGTGAAATTTCA  
TCAATAAA GTATTATGTTCTAGGCATTGGGGATACCATGTTTAC  
AAACAGA TTGATGCTATCCCAGGCACAAGACCAGTATTATGTT  
TCACATTA CTTGTCAATTCGTTTATCAGAGGCCAAATGTTTTTCTT  
TGTGGCO CAAACAGTTGTATTATTAGAACTGAGGGCTAAAAA  
GGATAAG GAAGAAAACAAGACTGTTACTATGGAAAATGAA  
ACTTCTT CACATTAATTCCTTATCTCATGTGAAATTTCATATTTA  
AAATATTT ATGTTCTAGGCATTGGGGATACCATGTTTACAAGAC  
GACACTA GCTAGAAAGACAATGAAACAGAGCCATGTGACCAA  
GATTGGA TAATGATATGAAAGAACCATTTCATGGGAAGGCCTAG  
TGAGCTG ATGAAAATAGATTTTTAAAACATGTTAATTCACGTTACT  
AGGAACA ATATTTATGATTGATACCTTTAAATGTCATTTGTTGAA  
CAAGACO CAAGACAGACTATGATTTACAGGATCAGATGTGGAC  
AAGTTGT AACATGTTAATTCACGTTACTTTTTGTTAAATTTAC  
CATTAAAT CCTTTAAATGTCATTTGTTGAAGGAAGATTATTCATT  
GTTCTAGGCAT TGGGGATACCATGTTCAACAAGACAGAC TATGATTTACAGGATCAGATGTGGACTCTCAAATTCGACT  
TATCCCAGGCACAAGACCAGTATTATGTTCTAGGCATTGGGGATACCATTACCTGTCAATGTTATTAATATTTTTTAGGAA  
TTCGTTTATCAGAGGCCAAATGTTTTCTTTGTAAACGTGTGTAAACATTCTCAGAATTTTAAACAATAACAATCAGG



# Overview

---

- Sequencing data: from wet lab to fastq.
- Applications to studying genomes and much much more.
- \*Sophisticated use of data from genomics requires an integrated understanding of the biological experiment, sample preparation and down stream computational analyses of the data.

# Importance of genomics data: these data are central to most biomedical and biological sciences

Cell

## The Implication of Early Chromatin Changes in X Chromosome Inactivation

Jan Jakub Żylicz,<sup>1,2,6</sup> Aurélie Bousard,<sup>1,6</sup> Kristina Žumer,<sup>3</sup> Francois Dossin,<sup>1</sup> Eusra Mohammad,<sup>3</sup> Simão Teixeira da Rocha,<sup>4</sup> Björn Schwalb,<sup>3</sup> Laurène Syx,<sup>1</sup> Florent Dingli,<sup>5</sup> Damarys Loew,<sup>5</sup> Patrick and Edith Heard<sup>1,7,\*</sup>

<sup>1</sup>Institut Curie, PSL Research University, CNRS UMR3215, INSERM U934, UPMC Paris-Sorbonne, 75005 Paris, France

<sup>2</sup>University of Cambridge, Department of Physiology, Development and Neuroscience, Cambridge CB2 3EG, UK

<sup>3</sup>Max Planck Institute for Biophysical Chemistry, Department of Molecular Biology, 37077 Göttingen, Germany

<sup>4</sup>Instituto de Medicina Molecular, Faculdade de Medicina, Universidade de Lisboa, 1649-028 Lisboa, Portugal

<sup>5</sup>Institut Curie, PSL Research University, Centre de Recherche, Laboratoire de Spectrométrie de Masse Protéomique France

<sup>6</sup>These authors contributed equally

<sup>7</sup>Lead Contact

\*Correspondence: [edith.heard@curie.fr](mailto:edith.heard@curie.fr)

<https://doi.org/10.1016/j.cell.2018.11.041>

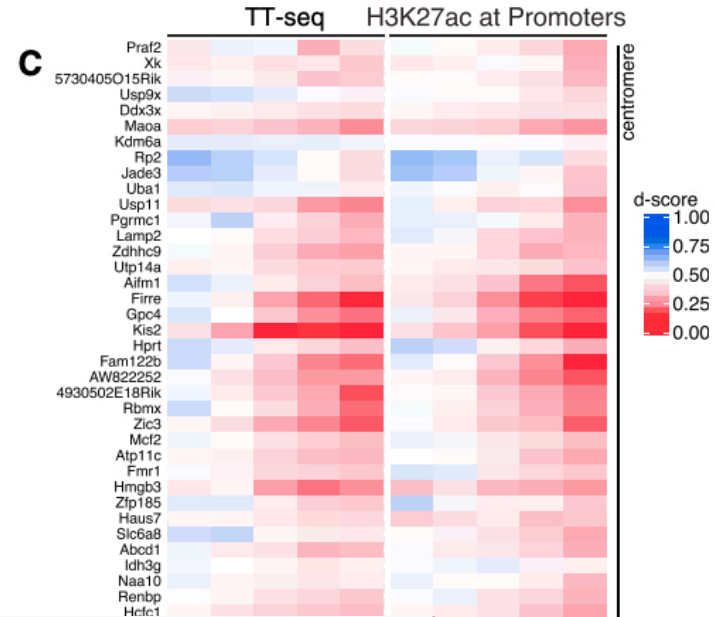
### SUMMARY

During development, the precise relationships between transcription and chromatin modifications often remain unclear. We use the X chromosome inactivation (XCI) paradigm to explore the implication of chromatin changes in gene silencing. Using female mouse embryonic stem cells, we initiate XCI by inducing *Xist* and then monitor the temporal changes in transcription. We use nascent transcription profiling. The H2AK119 ubiquitination and H3K27ac alterations are found to be bound on the

heterochromatin is X chromosome inactivation in mammals. Although the role for chromatin modifications in the maintenance of the inactive state has been extensively studied (Penny and Berletch, 2015), almost nothing is known about their role in the initiation of gene silencing.

In female mouse embryos, one of the two X chromosomes is randomly chosen for inactivation around the time of gastrulation (Lyon, 1962). This phenomenon is dependent on the expression of the future inactive X chromosome (*Xi*) by the *Xist* (Penny et al., 1996). The conserved

Article



### DATA AND SOFTWARE AVAILABILITY

The accession number for the sequencing datasets reported in this paper is GEO: GSE116480. The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE partner repository with the dataset identifier PXD011344

# Data can be found in genomics databases

The screenshot shows the NCBI GEO Accession Display page for GSE116480. The page includes the NCBI logo, the GEO logo (Gene Expression Omnibus), and navigation links like HOME, SEARCH, and SITE MAP. The main content area displays the following information:

- Series GSE116480** (Query DataSets for GSE116480)
- Status:** Public on Jan 01, 2019
- Title:** The role of early chromatin changes in X chromosome inactivation
- Organism:** *Mus musculus*
- Experiment type:** Genome binding/occupancy profiling by high throughput sequencing; Expression profiling by high throughput sequencing
- Summary:** This SuperSeries is composed of the SubSeries listed below.
- Overall design:** Refer to individual Series
- Citation missing:** Has this study been published? Please [notify GEO](#).
- Submission date:** Jul 01, 2018
- Last update date:** Jan 15, 2019
- Contact name:** Aurélie Bousard
- E-mail:** [aurelie.bousard@curie.fr](mailto:aurelie.bousard@curie.fr)
- Organization name:** Institut Curie
- Street address:** 26, rue d'Ulm
- City:** Paris
- ZIP/Postal code:** 75005
- Country:** France

Platforms (2):

- [GPL17021](#) Illumina HiSeq 2500 (*Mus musculus*)
- [GPL21626](#) NextSeq 550 (*Mus musculus*)

Samples (155) [Less...](#)

- [GSM3239824](#) ChIPseq\_input\_TXY\_XistdeltaA\_D2\_noDox\_Rep1
- [GSM3239825](#) ChIPseq\_H3K27ac\_TXY\_XistdeltaA\_D2\_noDox\_Rep2
- [GSM3239826](#) ChIPseq\_H3K27ac\_TXY\_XistdeltaA\_D2\_Dox\_Rep2
- [GSM3239827](#) ChIPseq\_H3K27ac\_TXY\_XistdeltaA\_D2\_noDox\_Rep1
- [GSM3239828](#) ChIPseq\_H3K27ac\_TXY\_XistdeltaA\_D2\_Dox\_Rep1
- [GSM3239829](#) ChIPseq\_H3K27me3\_TXY\_XistdeltaA\_D2\_noDox\_Rep1
- [GSM3239830](#) ChIPseq\_H3K27me3\_TXY\_XistdeltaA\_D2\_Dox\_Rep1
- [GSM3239831](#) ChIPseq\_H3K27me3\_TXY\_XistdeltaA\_D2\_noDox\_Rep2
- [GSM3239832](#) ChIPseq\_H3K27me3\_TXY\_XistdeltaA\_D2\_Dox\_Rep2
- [GSM3239833](#) ChIPseq\_H3K27me3\_TXY\_XistdeltaA\_D2\_noDox\_Rep1

```
@SRR7458968.1 HISEQ:258:CBBJMACXX:1:1101:1450:1962 length=51
NCATTTTCCATGTTTCTCATTGTAACCTATTGATATACACTGTTCTACAAA
+SRR7458968.1 HISEQ:258:CBBJMACXX:1:1101:1450:1962 length=51
#1?DFFFFHHHHJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ
@SRR7458968.2 HISEQ:258:CBBJMACXX:1:1101:1497:1977 length=51
NGGACCAGCGGGCGATTGCCAGCCCTGCCAGCCCGCCCTCCATTGCCGG
+SRR7458968.2 HISEQ:258:CBBJMACXX:1:1101:1497:1977 length=51
#1:DDDFDDHADHGHGHBHGIIJJGGGG@9;@AEG:9>'3>C35B@CC##
@SRR7458968.3 HISEQ:258:CBBJMACXX:1:1101:1624:1943 length=51
NGTTCATAGCAGCCTTATTATAATAGCCGGAAGCTGGAAGAACCAGAT
+SRR7458968.3 HISEQ:258:CBBJMACXX:1:1101:1624:1943 length=51
#1=DDFFHHHHJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ
@SRR7458968.4 HISEQ:258:CBBJMACXX:1:1101:1688:1958 length=51
NCAGACAGAAAGACACACACAGAGACACATACAGACAGAAAGACACA
+SRR7458968.4 HISEQ:258:CBBJMACXX:1:1101:1688:1958 length=51
#1=DDFFHHHHJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ
@SRR7458968.5 HISEQ:258:CBBJMACXX:1:1101:1546:1977 length=51
NACCAACGAGGAGTTGTTCTTTGAGAAAATCAACAAGATAGATAAACCC
+SRR7458968.5 HISEQ:258:CBBJMACXX:1:1101:1546:1977 length=51
#1=DDFFGHHHHIBFGICGIIIFIIIIIIIIIIIIIIIIIIIIIIIIIIIGI
@SRR7458968.6 HISEQ:258:CBBJMACXX:1:1101:1647:1989 length=51
NAAGACTTGTATTCTAATATGGACTCAGCTTGTGTTTGTAGCCTGCATTCA
+SRR7458968.6 HISEQ:258:CBBJMACXX:1:1101:1647:1989 length=51
#1=DDDDHDHHHIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIG
@SRR7458968.7 HISEQ:258:CBBJMACXX:1:1101:1971:1921 length=51
NGTCTTTTCAAAAGATATGTTTAGAGGTAGCATTGCTTTTCATTG
+SRR7458968.7 HISEQ:258:CBBJMACXX:1:1101:1971:1921 length=51
#4=DDFFHHHHFDGHHGIIJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ
@SRR7458968.8 HISEQ:258:CBBJMACXX:1:1101:1756:1940 length=51
NGGGAGTGAAATATGGCGAGGAAACTGAAAAGGTGAAAATTTAGAAT
+SRR7458968.8 HISEQ:258:CBBJMACXX:1:1101:1756:1940 length=51
#1=DDFDDHHHHJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ
```

- 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 used to examine the authors' claims, but also to test new hypotheses.

# Central questions

---

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

# Workflow

---

## 1. Isolation of sample.

*e.g.*, Isolate DNA and shear.

## 2. Library preparation

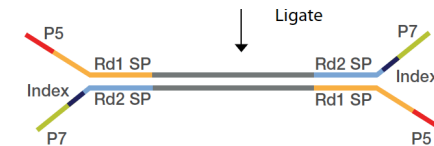
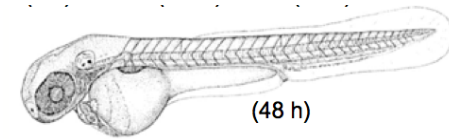
*e.g.*, Add known sequences to the ends.

## 3. Sequencing

*e.g.*, Illumina Novaseq

## 4. Analysis

*e.g.*, Map to genome and interpret.





# 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

## 3. Tomorrow (Third generation sequencing)

- a. Nanopore based
- b. Pacific Bioscience Sequencing

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 header includes 'Yale School of Medicine | W.M. Keck Foundation' and 'INFORMATION FOR' with a search bar. The main navigation bar lists 'Next-Gen Sequencing', 'Bioinformatics', 'Microarrays', 'Services & Fees', 'Mendelian Center', and 'About YCGA'. The 'Next-Gen Sequencing' menu is open, showing options like 'Illumina', 'Applications', 'Sample Requirement', 'Pooled Exome Analysis', 'HiSeq', 'MiSeq', 'Throughput' (highlighted), 'Library Protocols', 'Data Processing', and 'Data Retrieval'. The 'Throughput' section is titled 'Throughput' and contains a table of 'Sequencer Lane Data Outputs'.

Sequencer	Read length	# of Clusters per lane (millions)	Bases per lane (Gbp)
HiSeq 2500 Rapid	1x75	150	11.25
HiSeq 2500 Rapid	2x75	150	22.5
HiSeq 2500 Rapid	2x150	150	45
HiSeq 2500 High-output	1x75	200	15
HiSeq 2500 High-output	2x75	200	30
HiSeq 4000	2x100	300	60
HiSeq 4000	2x150	300	90
NovaSeq S2	2x100	1650	330
NovaSeq S2	2x150	1650	500
NovaSeq S4	2x150	2000	600

<http://ycga.yale.edu/sequencing/illumina/>

# 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=GIHGIIIHGIJIHEHIHHGFFFFEEDDDDDDDDDDDDD
```

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
NCATCACTTTCTGCACCAGCCATGACGTCAATCTTCGTCCGAACCCCAAACCTCGAGATCGGAAGAGCACACGTCTG
+
#11BBDDDFDFBFFFIIIIIIIIIIIFEGIIIIIFIGAGIIFIII=FEFFFFFFDDDD=@9A@BBBBB=?BB<

@HWI-D00306:498:HBB89ADXX:1:1101:1167:1902 1:N:0:CGATGT
TATTGCAATATGTTAACAATCTAACAAGGAAAAAATACCCACACAAAACAAAACACAACCCTTAGAACTGTGCTG
+
B@@FFDFFHFHHHJJIJIGIIJJJJJJJJJHFIJJJJJJJJJEHHJJJJJJJJJGHHHHFBDFFFE>CEEC
@HWI-D00306:498:HBB89ADXX:1:1101:1190:1928 1:N:0:CGATGT
ACCAAGCCACAATAAGTTAGTGTTCATAGTACATGCTGAGTTATTTGATCCCGTATCTATACTGCTACTGTC
+
@<@DDDDD8CDDDGE?2<AFFBCCEEHEIEGHIIEGEIDD@CDGFFFEFIDGCFCDABFG>FBFGFGIEIFFDDDD
@HWI-D00306:498:HBB89ADXX:1:1101:1157:1931 1:N:0:CGATGT
CTGAGATTCTTTGCCATAGTCCTTAACCACTACGCAACTGCAACCAACCACCTTCCGTGGTTTGCCTCTCGATCG
+
CCCFHHHHHHHHIJJIIJJJIIGHHIJGGJIGIJJJJJJJJJIIIIJJJJIIJJJJIIJGJJHCHFBDFFFDDECB
```


Generally ~ 400,000,000 reads/sequencing lane

Note: This is for an Illumina HiSeq 4000 with current chemistry, but this number changes

# How long are the reads?

---

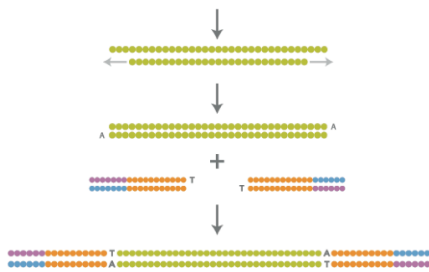
TATTGCAATATGTTAACAATCTAACAAGGAAAAAATACCCACACAAAACAAAACACAACCCTTAGAACTGTGCTG



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?



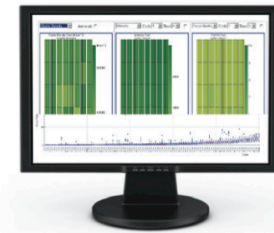
Library Preparation  
~2 h [15 min hands-on (Nextera)]  
< 6 h [< 3 h hands-on (TruSeq)]



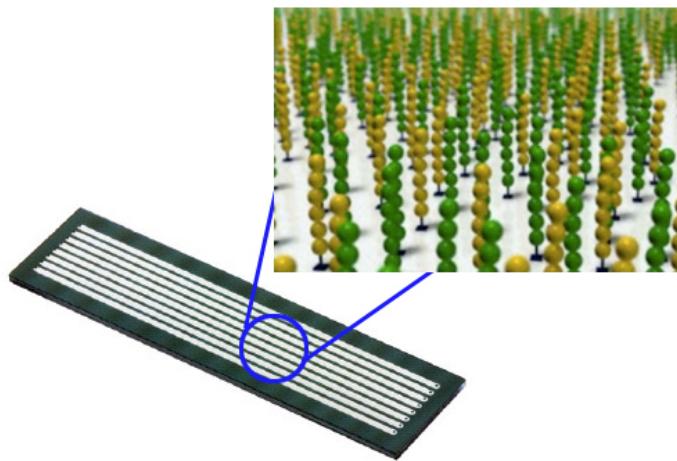
Cluster Generation  
~5 h (<10 min hands-on)



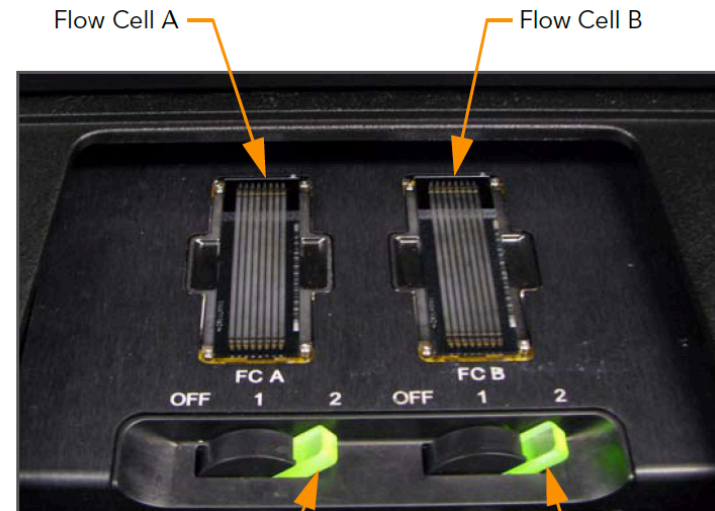
Sequencing by Synthesis  
~1.5 to 11 days



CASAVA  
2 days (30 min hands-on)



Flow cell



Flow Cell Lever A

Flow Cell Lever B

# What is a flow cell?

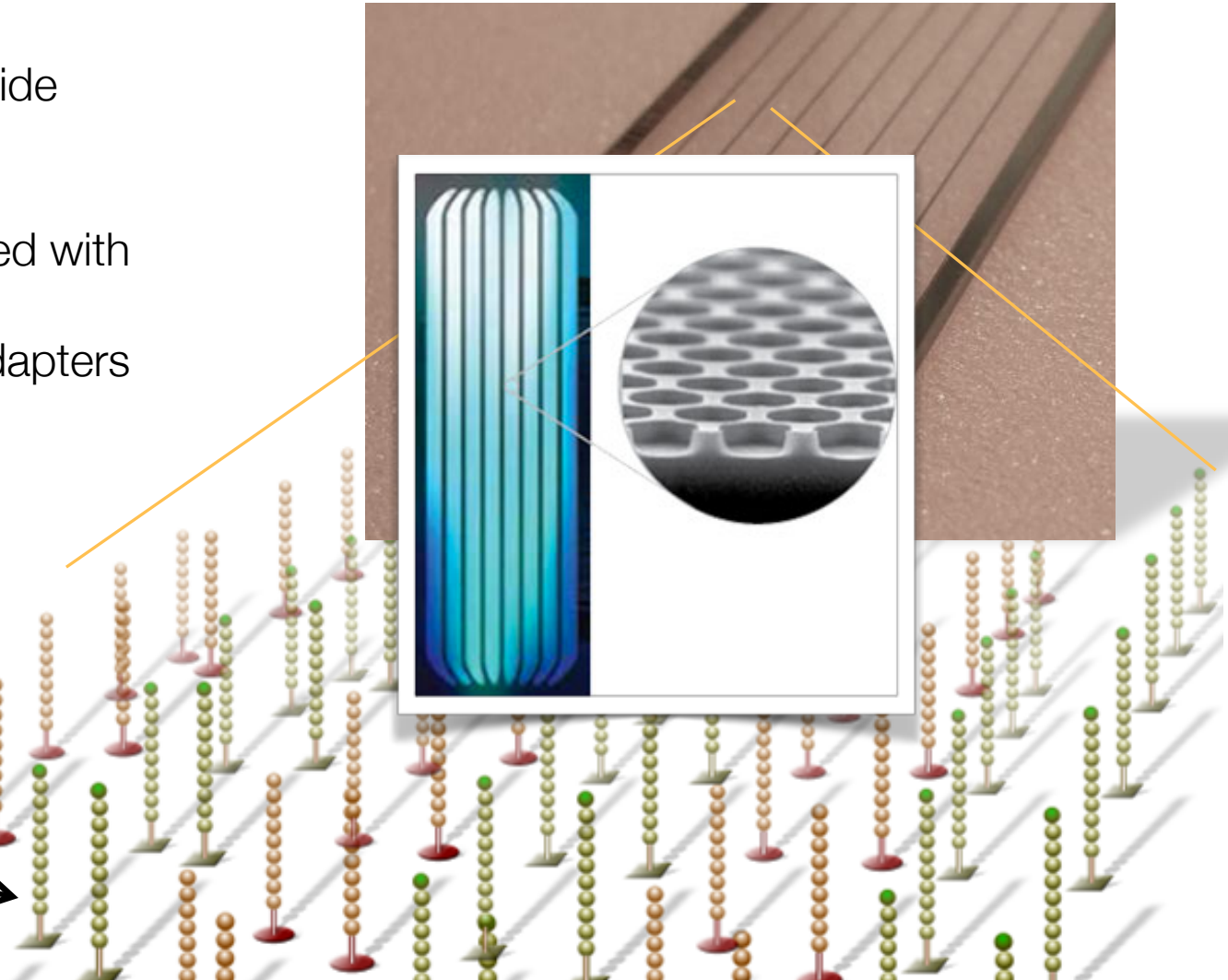
A flow cell is a thick glass slide with 8 channels or lanes.

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

P5 oligo

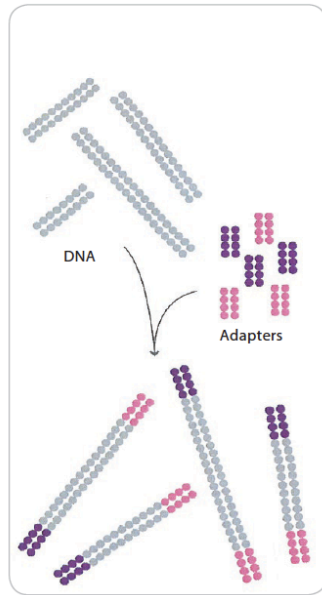


P7 oligo



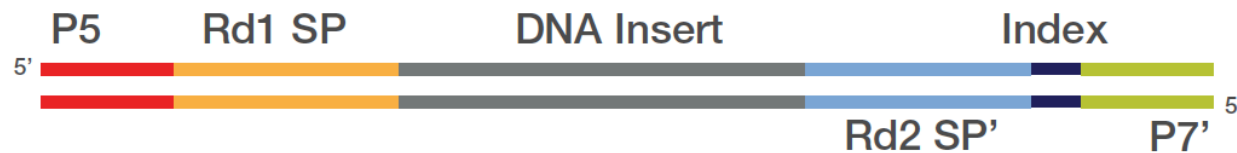
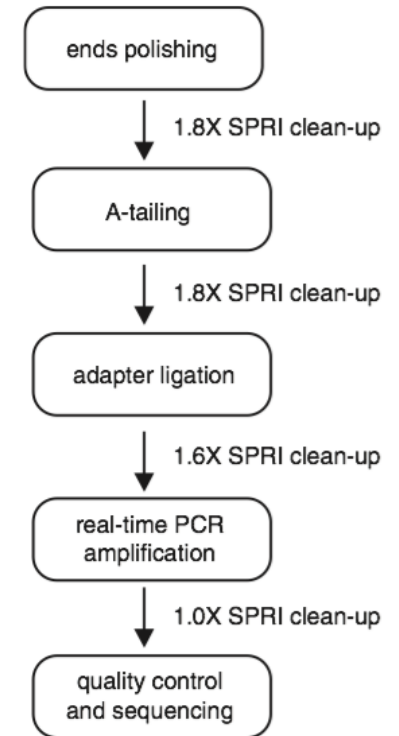
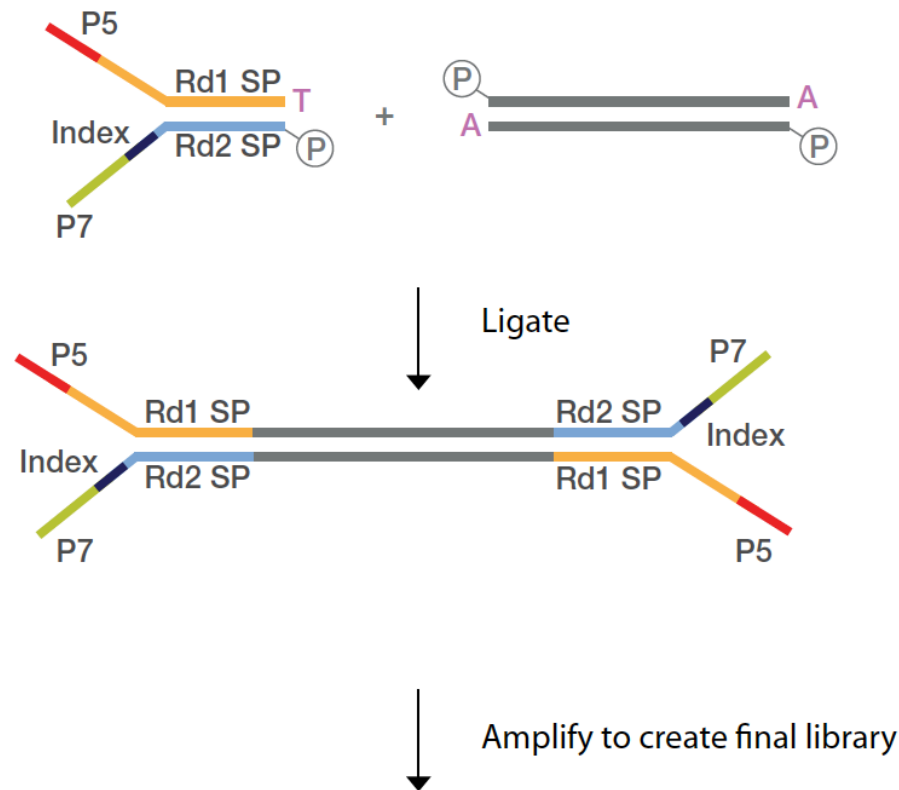


Cluster PCR  
on flow cell  
(8 lanes)



# Optional: How do you make a sequencing library?

Index = unique sequence key to identify library



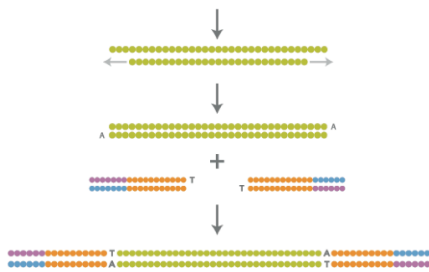
12 samples per lane

## 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?



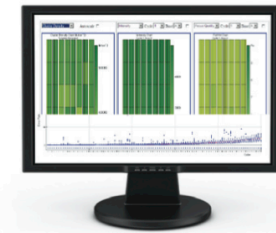
Library Preparation  
~2 h [15 min hands-on (Nextera)]  
< 6 h [< 3 h hands-on (TruSeq)]



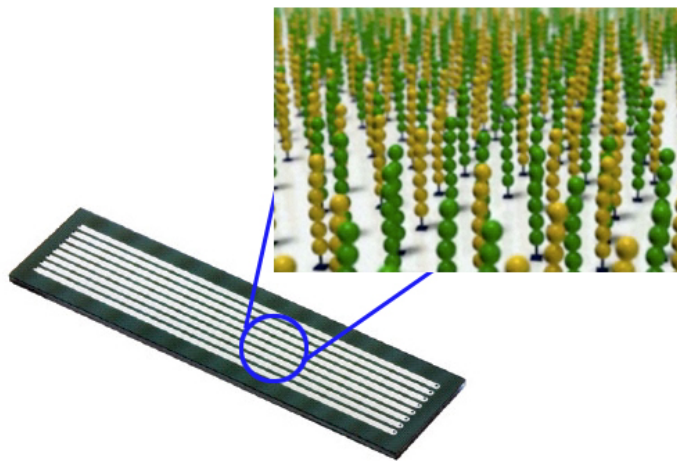
Cluster Generation  
~5 h (<10 min hands-on)



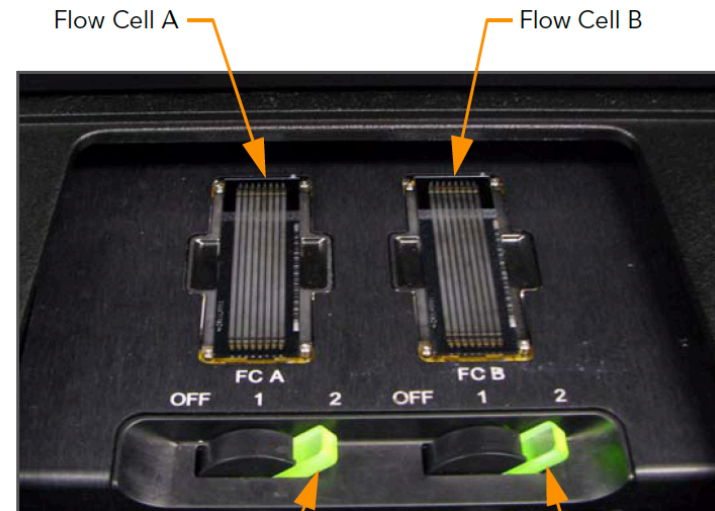
Sequencing by Synthesis  
~1.5 to 11 days



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
NNACCTAGCCATCTGCAGTCCTCGGTCCTGTGTTAGACCAGAAGTAGGTGCCCAGGCCAGGTACCACCTAATCCTT
+
##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. Sequence
3. Quality score identifier “+”
4. 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<
```

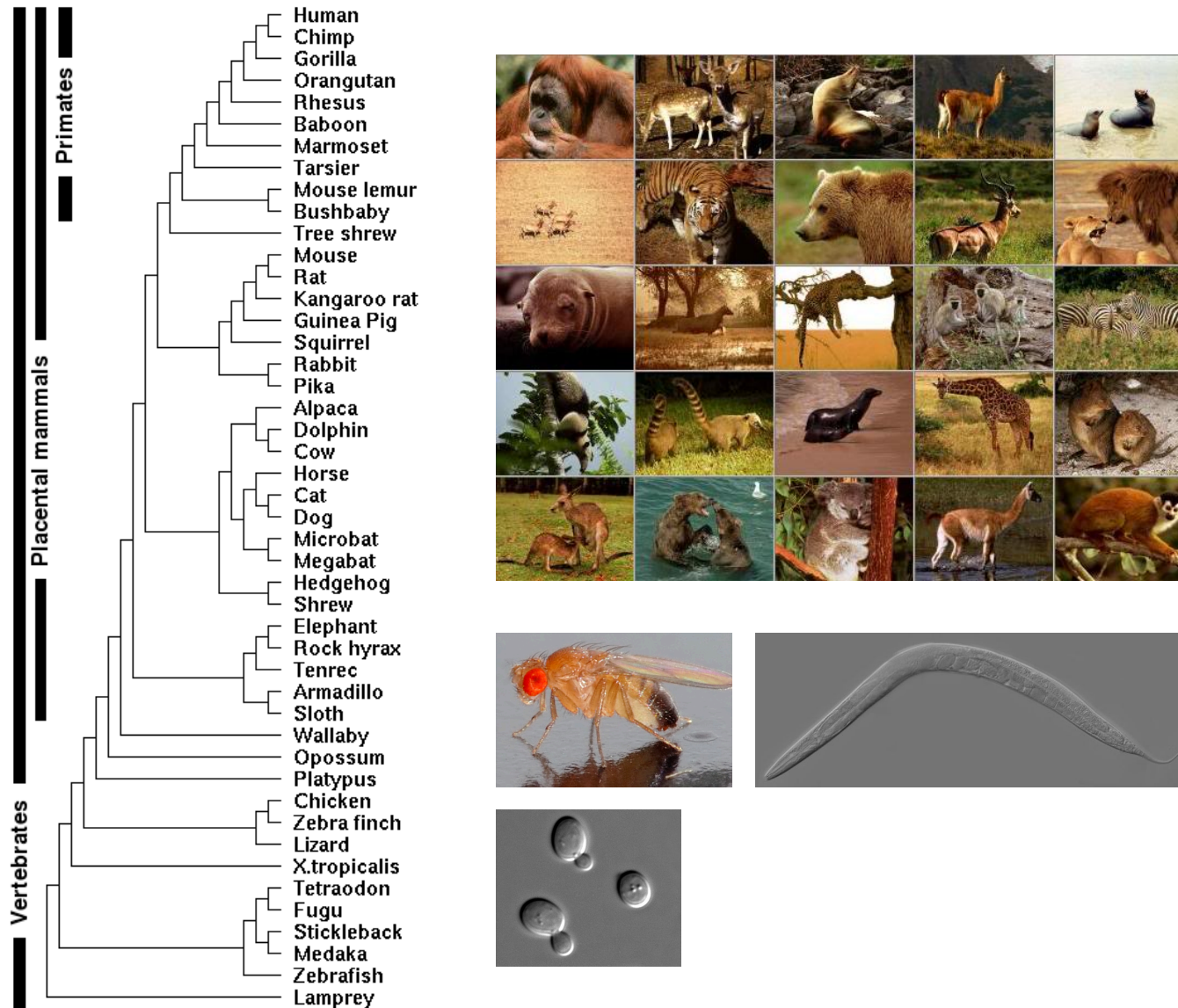
- For each single end read: Incomplete incorporation of bases.
- For the size of the insert (especially for paired end analysis): Ability to get consistent clusters.

# What do I do with my sequencing reads?

---



# Many reference genomes are available

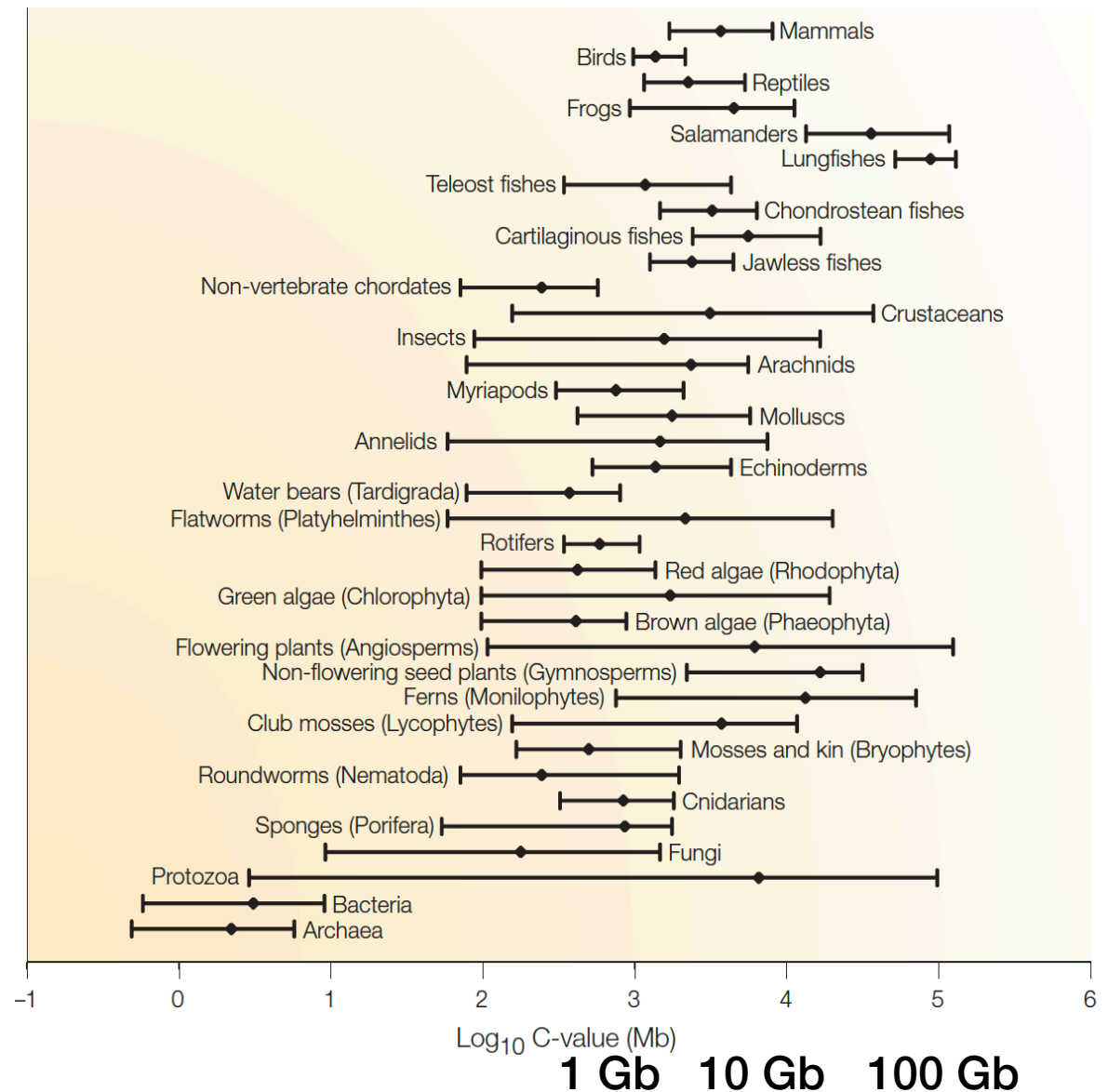


# There is a wide range of genome sizes.

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

---

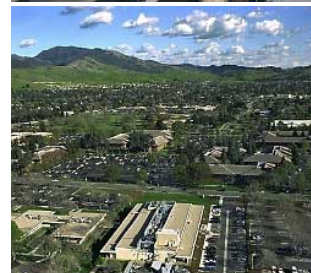
Victory declared **2003**



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

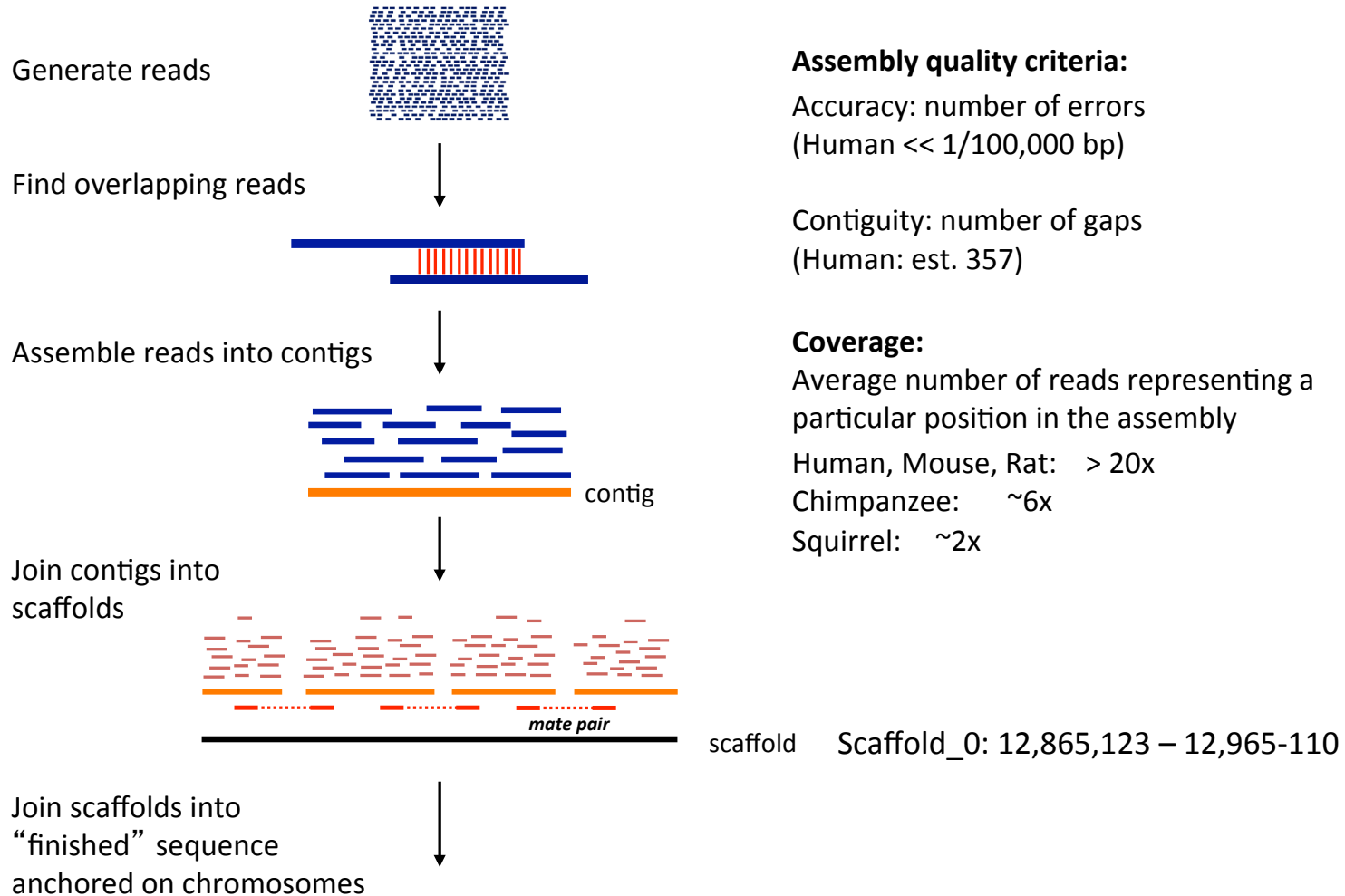


National Human  
Genome Research  
Institute



Novaseq 20 billion reads 2x150 bp. \$1000 -> \$100/genome.

# How to assemble a genome

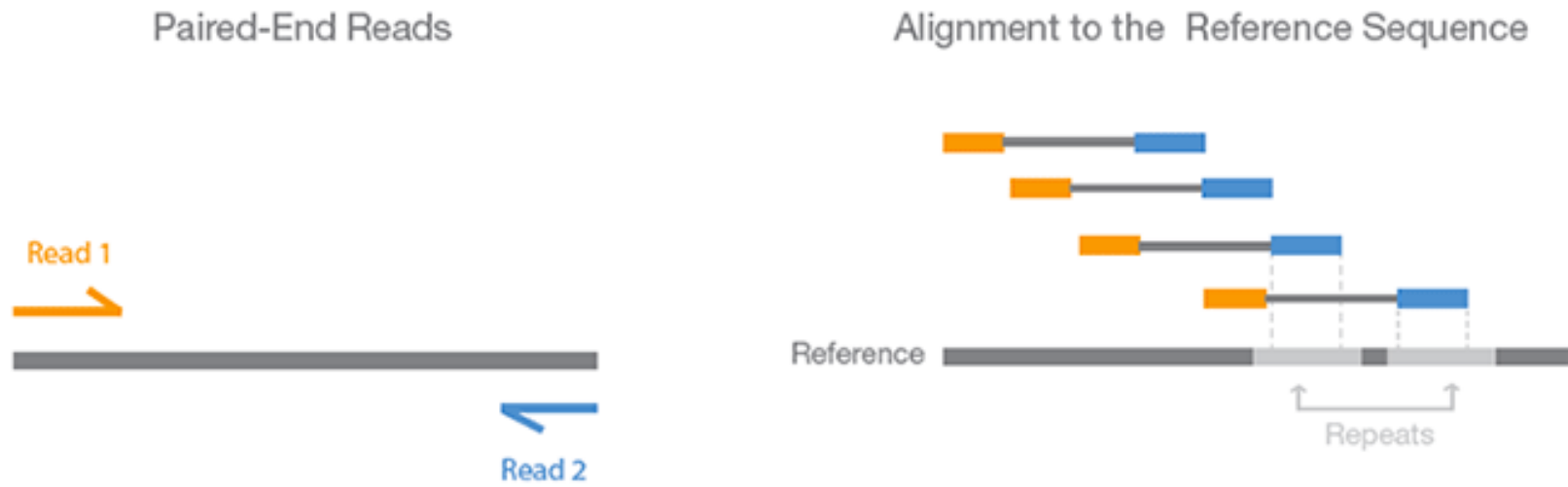


AGTTGTATTATTAGAACTGAGGGCTAAAACTGTGCACATACACAGACACACATATTATTTTAATATAGATTTTCAATAATTGGTCTAGGATAAGGATAATATACAG

There are various

# The importance of paired end reads

---



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

CCAAATCAAACAGTTGTATTATTAGAAACTGAGGGCTAAAACTGTGCACATACACAGACACACATATTATTTAATATAGATTTTCAATAATTGGTCTAGGATAAG  
AAGCAAGAAGAAAACAAAGACTGTTACTATGGAAAAATGAAAATAGATTTTAAAACATGTTAATTCACGTTACTTTTTGTTAAATTTACTTTTCTTCTTTCACTTCTT  
AATAAATCACATTAATTCCTTATCTCATGTGAAATTTATGATTGATACCTTTAAATGTCATTTGTTGAAGGAAGATTATTCATTTTTTTCATTCAATAAATATTT  
CAGTATTATGTTCTAGGCATTGGGGATACCATGTTCCACAAGACAGACTATGATTTACAGGATCAGATGTGGACTCTCAAATTCGACTGAGAATAAAAACAGACACT  
TAATTGATGCTAGAAAGACAATGAAACAGAGCCATGTGACCAATGAGAGAGATGAGGGTGGCAGCAGCCTGTTTTAGATAAGGTACCTGATTGGTGGGATTGG  
TATGCCTTAATGATATGAAAGAACCATTGATGGGAAGGCCTAGCATTAAAAACCGTCTAGGCAGAATGAGCAGCAAGTGCAAGGGTCCCTGGATAGGAATGAGC  
ATGGAAAAATGAAAATAGATTTTAAAACATGTTAATTCACGTTACTTTTTGTTAAATTTACTTTTCTTCTTTCACTTCTTACCTGTCAATGTTATTAATATTTT  
GAAATTTTCATATTTATGATTGATACCTTTAAATGTCATTTGTTGAAGGAAGATTATTCATTTTTTTCATTCAATAAATATTTTTTAGAATAATAAGTCCCAGGCACAAGA  
CATGTTCCACAAGACAGACTATGATTTACAGGATCAGATGTGGACTCTCAAATTCGACTGAGAATAAAAACAGACACTAAACAAGTAAATAAAGTTAATTTCAAGTT  
GATTTTAAAACATGTTAATTCACGTTACTTTTTGTTAAATTTACTTTTCTTCTTTCACTTCTTACCTGTCAATGTTATTAATATTTTTTAGGAACAATAAATCACATTA  
ATTGATACCTTTAAATGTCATTTGTTGAAGGAAGATTATTCATTTTTTTCATTCAATAAATATTTTTTAGAATAATAAGTCCCAGGCACAAGACCAGTATTATGTTCTAG  
ACTATGATTTACAGGATCAGATGTGGACTCTCAAATTCGACTGAGAATAAAAACAGACACAAACAAGTAAATAAAGTTAATTTCAAGTTGTAATTGATGCTATCCCA  
TTGGGGATACCATTACCTGTCAATGTTATTAATATTTTTTAGGAACAATAAATCACATTAATTCCAACATGCAAAGAGGAAATCTCCATATCATGCTTGTCAATTCGTT  
GTGTGTAAAACATTCTCAGAATTTTAAACAATAACAAATCAGGGCTGAATGTGGCCAACATGCAAAGAGGAAATCTCCATCTGTCCAAATCAAACAGTTGTATT  
CATAACAGACACACATATTTTAAATAGATTTTCAATAATTGGTCTAGGATAAGGATAATATACAGAGAACATGCCAAAAGTTAAGCAAGAAGAAAACAAAG  
TAAAACATGTTAATTCACGTTACTTTTTGTTAAATTTACTTTTCTTCTTTCACTTCTTACCTGTCAATGTTATTAATATTTTTTAGGAACAATAAATCACATTAATTCCT  
TACCTTTAAATGTCATTTGTTGAAGGAAGATTATTCATTTTTTTCATTCAATAAATATTTTTTAGAATAATAAGTCCCAGGCACAAGACCAGTATTATGTTCTAGGCAT  
TGATTTACAGGATCAGATGTGGACTCTCAAATTCGACTGAGAATAAAAACAGACACTAAACAAGTAAATAAAGTTAATTTCAAGTTGTAATTGATGCTAGAAAGACA  
AGATGAGGGTGGCAGCAGCCTGTTTTAGATAAGGTACCTGATTGGTGGGATTGGAAGACCTCTCTGAGATTAGTGTCTTCAGATATGCCTTAATGATATGAAAG  
AACCGTCTAGGCAGAATGAGCAGCAAGTGCAAGGGTCCCTGGATAGGAATGAGCTGGATATACTCAAGGAAGAAAGAGAAACTATGGAAAAATGAAAATAGATT  
TAAATTTACTTTTCTTCTTTCACTTCTTACCTGTCAATGTTATTAATATTTTTTAGGAACAATAAATCACATTAATTCCTTATCTCATGTGAAATTTTCATATTTATGATTGA  
TTATTCATTTTTTTCATTCAATAAATATTTTTTAGAATAATAAGTCCCAGGCACAAGACCAGTATTATGTTCTAGGCATTGGGGATACCATGTTCCACAAGACAGACTAT  
ATTCGACTGAGAATAAAAACAGACACTAAACAAGTAAATAAAGTTAATTTCAAGTTGTAATTGATGCTACTATGGAAAAATGAAAATAGATTTTAAAACATGTTAATTC  
TTTTCACTTCTTACCTGTCAATGTTATTAATATTTTTTAGGAACAATAAATCACATTAATTCCTTATCTCATGTGAAATTTTCATATTTATGATTGATACCTTTAAATGTCAT  
CAATAAATATTTTTTAGAATAATAAGTCCCAGGCACAAGACCAGTATTATGTTCTAGGCATTGGGGATACCATGTTCCACAAGACAGACTATGATTTACAGGATCAGA  
AACAGACACAAACAAGTAAATAAAGTTAATTTCAAGTTGTAATTGATGCTATCCAGGCACAAGACCAGTATTATGTTCTAGGCATTGGGGATACCATACCTGT  
CACATTAATTCCAACATGCAAAGAGGAAATCTCCATATCATGCTTGTCAATTCGTTTATCAGAGGCCAAATGTTTTTCTTTGTAACGTGTGTAAAACATTCTCAGA  
GTGGCCAACATGCAAAGAGGAAATCTCCATCTGTCCAAATCAAACAGTTGTATTATTAGAAACTGAGGGCTAAAACTGTGCACATACACAGACACACATATTA  
GATAAGGATAATATACAGAGAACATGCCAAAAGTTTAAAGCAAGAAGAAAACAAAGACTGTTACTATGGAAAAATGAAAATAGATTTTAAAACATGTTAATTCACGT  
CTTCTTACCTGTCAATGTTATTAATATTTTTTAGGAACAATAAATCACATTAATTCCTTATCTCATGTGAAATTTTCATATTTATGATTGATACCTTTAAATGTCATTTGTT  
AATATTTTTTAGAATAATAAGTCCCAGGCACAAGACCAGTATTATGTTCTAGGCATTGGGGATACCATGTTCCACAAGACAGACTATGATTTACAGGATCAGATGT  
ACACTAAACAAGTAAATAAAGTTAATTTCAAGTTGTAATTGATGCTAGAAAGACAATGAAACAGAGCCATGTGACCAATGAGAGAGATGAGGGTGGCAGCAGCC  
ATTGGAAGACCTCTCTGAGATTAGTGTCTTCAGATATGCCTTAATGATATGAAAGAACCATTGATGGGAAGGCCTAGCATTAAAAACCGTCTAGGCAGAATGAG  
GAGCTGGATATACTCAAGGAAGAAAGAGAAACTATGGAAAAATGAAAATAGATTTTAAAACATGTTAATTCACGTTACTTTTTGTTAAATTTACTTTTCTTCTTTCA  
GGAAACAATAAATCACATTAATTCCTTATCTCATGTGAAATTTTCATATTTATGATTGATACCTTTAAATGTCATTTGTTGAAGGAAGATTATTCATTTTTTTCATTCAATA  
AAGACCAGTATTATGTTCTAGGCATTGGGGATACCATGTTCCACAAGACAGACTATGATTTACAGGATCAGATGTGGACTCTCAAATTCGACTGAGAATAAAAACAG  
AGTTGTAATTGATGCTACTATGGAAAAATGAAAATAGATTTTAAAACATGTTAATTCACGTTACTTTTTGTTAAATTTACTTTTCTTCTTTCACTTCTTACCTGTCAAT  
ATTAATTCCTTATCTCATGTGAAATTTTCATATTTATGATTGATACCTTTAAATGTCATTTGTTGAAGGAAGATTATTCATTTTTTTCATTCAATAAATATTTTTTAGAATA  
TTCTAGGCATTGGGGATACCATGTTCCACAAGACAGACTATGATTTACAGGATCAGATGTGGACTCTCAAATTCGACTGAGAATAAAAACAGACACAAACAAGTAA  
ATCCAGGCACAAGACCAGTATTATGTTCTAGGCATTGGGGATACCATACCTGTCAATGTTATTAATATTTTTTAGGAACAATAAATCACATTAATTCCAACATGCA  
TCGTTTATCAGAGGCCAAATGTTTTTCTTTGTAACGTGTGTAAAACATTCTCAGAATTTTAAACAATAACAAATCAGGGCTGAATGTGGCCAACATGCAAAGAG  
GTGATTATTAGAAACTGAGGGCTAAAACTGTGCACATACACAGACACACATATTTTAAATAGATTTTCAATAATTGGTCTAGGATAAGGATAATATACAGAGA  
CAAAGACTGTTACTATGGAAAAATGAAAATAGATTTTAAAACATGTTAATTCACGTTACTTTTTGTTAAATTTACTTTTCTTCTTTCACTTCTTACCTGTCAATGTT  
TTCTTATCTCATGTGAAATTTTCATATTTATGATTGATACCTTTAAATGTCATTTGTTGAAGGAAGATTATTCATTTTTTTCATTCAATAAATATTTTTTAGAATAATAAGT

# What types of annotation do we have/want?

---

**~3 billion bp**

```
ACAATAAATCACATTAATTCCTTATCTCATGTGAAATTCATATTTATGATTG
ATACCTTTAAATGTCATTTGTTGAAGGAAGATTATTCATTTTTTCATTCAAT
AAATATTTTTAGAAATAATAAGTCCCAGGCACAAGACCAGTATTATGTTCT
AGGCATTGGGGATACCATGTTCAAGACAGACTATGATTACAGGATC
AGATGTGGACTCTCAAATTCGACTGAGAATAAACAGACACTAAACAAG
TAAATAAAGTAAATTCGAAGTTGTAATTGATGCTAGAAAGACAATGAAACA
GAGCCATGTGACCAATGAGAGAGATGAGGGTGGCAGCAGCCTGTTTTA
GATAAGGTACCTGATTGGTGGGATTGGAAGACCTCTCTGAGATTAGTGT
CTTCAGATATGCCTTAATGATATGAAAGAACCATTTCATGGGAAGGCCTAG
CATTA AAAACCGTCTAGGCAGAATGAGCAGCAAGTGCAAGGGTCCTGG
ATAGGAATGAGCTGGATATACTCAAGGAAGAAAGAGAAACTATGGAAAA
ATGAAAAATAGATTTTAAACATGTTAATTCACGTTACTTTTTGTTAAATTTA
CTTTCTCTTTCACTTCTTACCTGTCAATGTTATTAATTTTTAGGAACA
ATAAATCACATTAATTCCTTATCTCATGTGAAATTCATATTTATGATTGATA
CCTTTAAATGTCATTTGTTGAAGGAAGATTATTCATTTTTTCATTCAATAAA
TATTTTTTAGAATAATAAGTCCCAGGCACAAGACCAGTATTATGTTCTAGG
CATTGGGGATACCATGTTCAAGACAGACTATGATTACAGGATCAGAT
GTGGACTCTCAAATTCGACTGAGAATAAACAGACACTAAACAAGTAAAT
AAAGTAAATTCAGTTGTAATGATGCTACTATGGAAAAATGAAAAATAGA
TTTTAAACATGTTAATTCACGTTACTTTTTGTTAAATTTACTTTCTCTTT
CACTTCTACCTGTCAATGTTAATATTTTTAGGAACAATAAATCACATT
AATTCCTTATCTCATGTGAAATTCATATTTATGATTGATACCTTTAAATGT
CATTTGTTGAAGGAAGATTATTCATTTTTTCATTCAATAAATATTTTTAGA
ATAATAAGTCCCAGGCACAAGACCAGTATTATGTTCTAGGCATTGGGGAT
ACCATGTTCAAGACAGACTATGATTTACAGGATCAGATGTGGACTCTC
AAATTCGACTGAGAATAAACAGACACAACAAGTAAATAAAGTAAATTT
CAAGTTGTAATGATGCTATCCCAGGCACAAGACCA....
```

## **Genes:**

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

## **Genetic variation:**

- SNPs and CNVs

## **Sequence conservation**

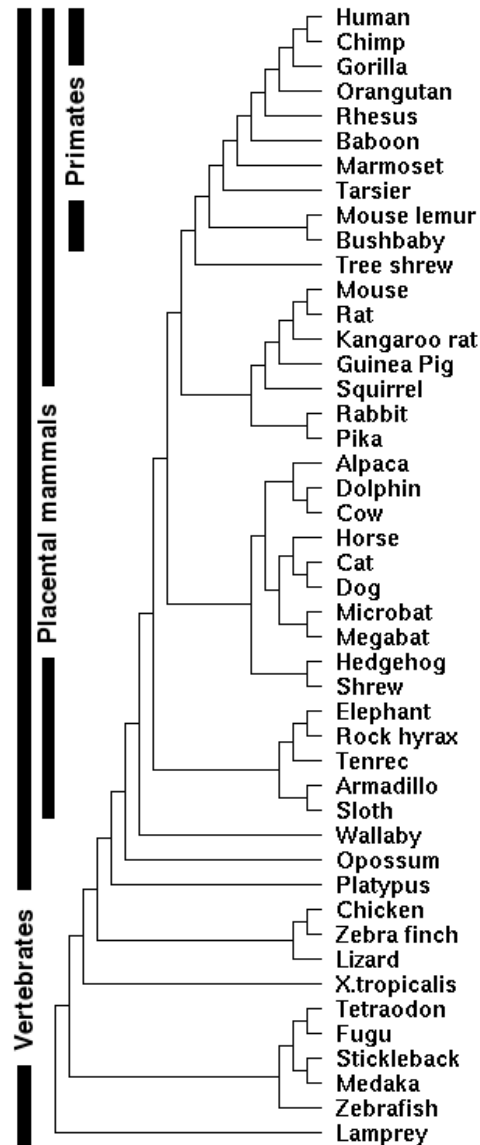
## **Regulatory sequences:**

- 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 (UCSC)

---

Chr5: 133,876,119 – 134,876,119

---

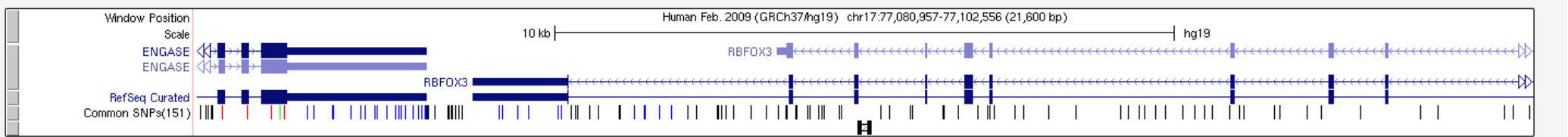
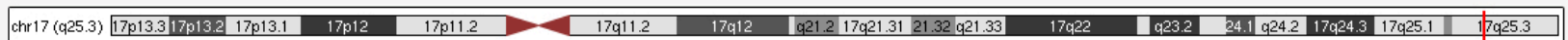


# Our specific example:

```
@HWI-ST1239:178:H0KPNADXX:2:1101:3120:1979 1:N:0:TGACCA
NCTGTAGGCTGCGTAGCCTCCCTGCAGGGTAAGTGGGAGGAGAGAGAGACAGAGGGACTTAGTGGGGCTCCCCAGGG
+
#1=DDFFFHHHHHIJIJJJIJJJJJIJJJ?FHIDGIJ=GIHGI IIHGIJJIHEHIHHGFFFFEEEDDDDDDDDDDDDD
@HWI-ST1239:178:H0KPNADXX:2:1101:3120:1979 2:N:0:TGACCA
NNACCTAGCCATCTGCAGTCCCTCGGTCCTGTGTTAGACCAGAACTAGGTGCCCAGGCCAGGTACCACCTAATCCTT
+
##4<@@@@@@@@@?@@@?@@?????@??@????????????????????>????????????@>????@@@@?@???????
```

## UCSC Genome Browser on Human Feb. 2009 (GRCh37/hg19) Assembly

move <<< << < > >> >>> zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x 100x



move start < 2.0 > Click on a feature for details. Click or drag in the base position track to zoom in. Click side bars for track options. Drag side bars or labels up or down to reorder tracks. Drag tracks left or right to new position. Press "?" for keyboard shortcuts. move end < 2.0 >

# Workflow

---

## 1. Isolation of sample.

*e.g.*, Isolate DNA and shear.

## 2. Library preparation

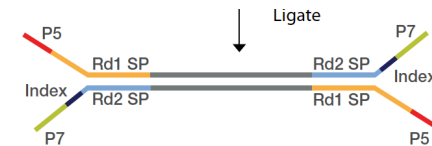
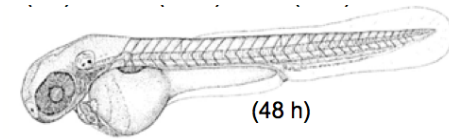
*e.g.*, Add known sequences to the ends.

## 3. Sequencing

*e.g.*, Illumina Novaseq

## 4. Analysis

*e.g.*, Map to genome and interpret.



# Using sequencing to annotate the genome

---

1. Where are the cis-acting regulatory elements in DNA?
  - A. DNase I hyper-sensitivity mapping (DNase-Seq).
  - B. FAIRE to map regulatory elements.
2. Where do transcription factors bind?
  - C. ChIP-seq of transcription factors (or in high res, ChIP-exo)
  - D. Nucleosome mapping (MNase-Seq).
3. Where are different histone modifications found?
  - E. ChIP-Seq of histone modifications.
  - F. ChIP-Seq of chromatin writers, readers and erasers.
4. Where is RNA polymerase transcribing?
  - G. ChIP-Seq of polymerase.
  - H. GRO-Seq, NET-Seq and TT-Seq to measure RNA in the polymerase active site..
5. How is the genome organized in 3D?
  - I. 4C/5C/Hi-C to measure chromatin conformation.

# 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