Genomics

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

What is genomics?

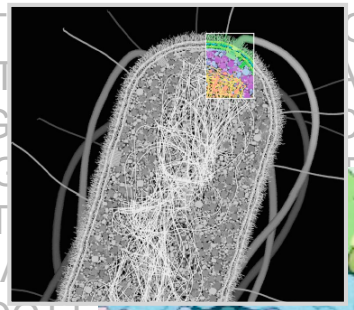
1. The **global** study of how biological **information** is encoded in genome sequence

   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.



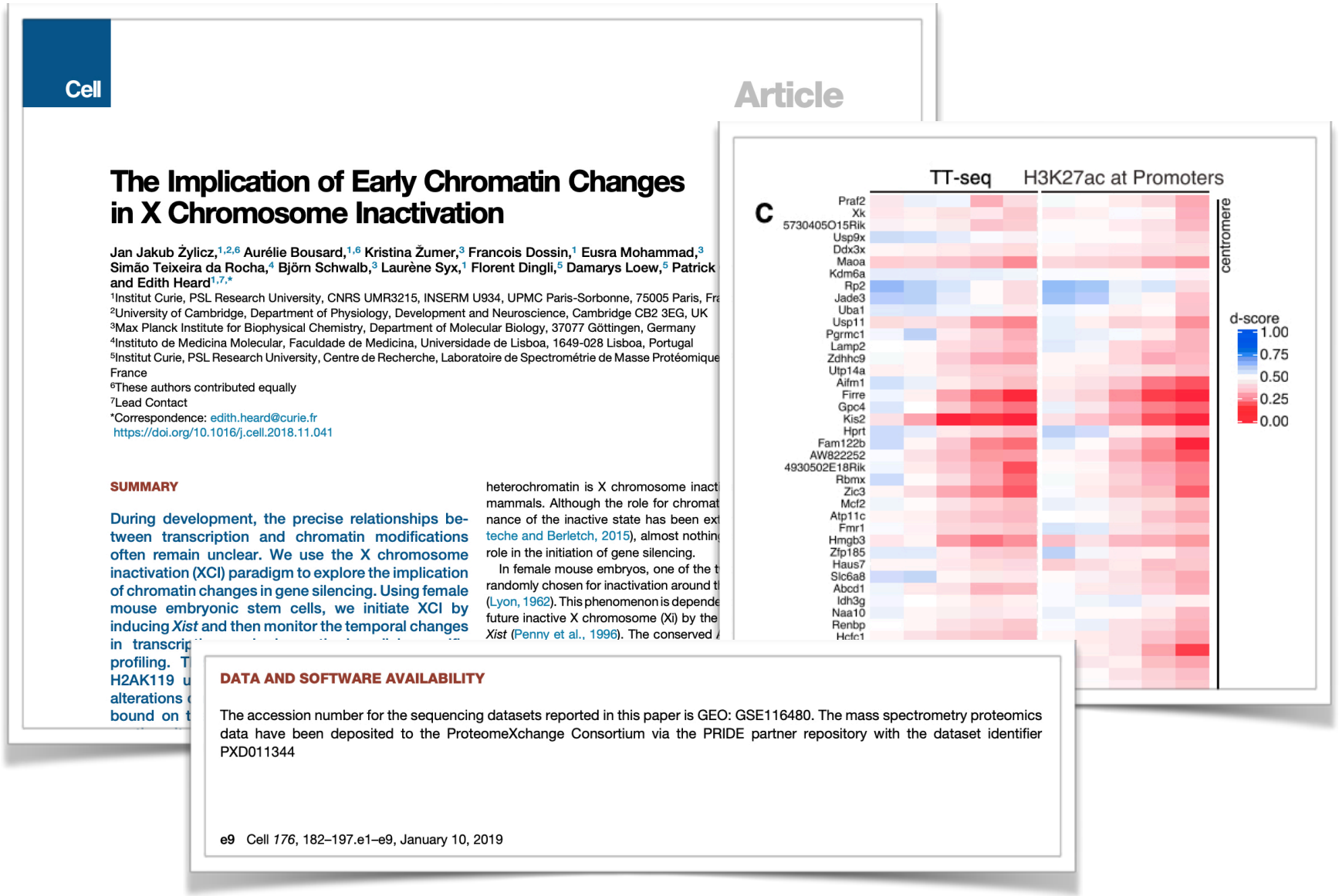
Overview

- 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.

Credit: Jim Noonan for many of the slides

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



Cell

Article

The Implication of Early Chromatin Changes in X Chromosome Inactivation

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

[1]Institut Curie, PSL Research University, CNRS UMR3215, INSERM U934, UPMC Paris-Sorbonne, 75005 Paris, Fra
[2]University of Cambridge, Department of Physiology, Development and Neuroscience, Cambridge CB2 3EG, UK
[3]Max Planck Institute for Biophysical Chemistry, Department of Molecular Biology, 37077 Göttingen, Germany
[4]Instituto de Medicina Molecular, Faculdade de Medicina, Universidade de Lisboa, 1649-028 Lisboa, Portugal
[5]Institut Curie, PSL Research University, Centre de Recherche, Laboratoire de Spectrométrie de Masse Protéomique France
[6]These authors contributed equally
[7]Lead Contact
*Correspondence: 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 transcrip
profiling. T
H2AK119 u
alterations
bound on t

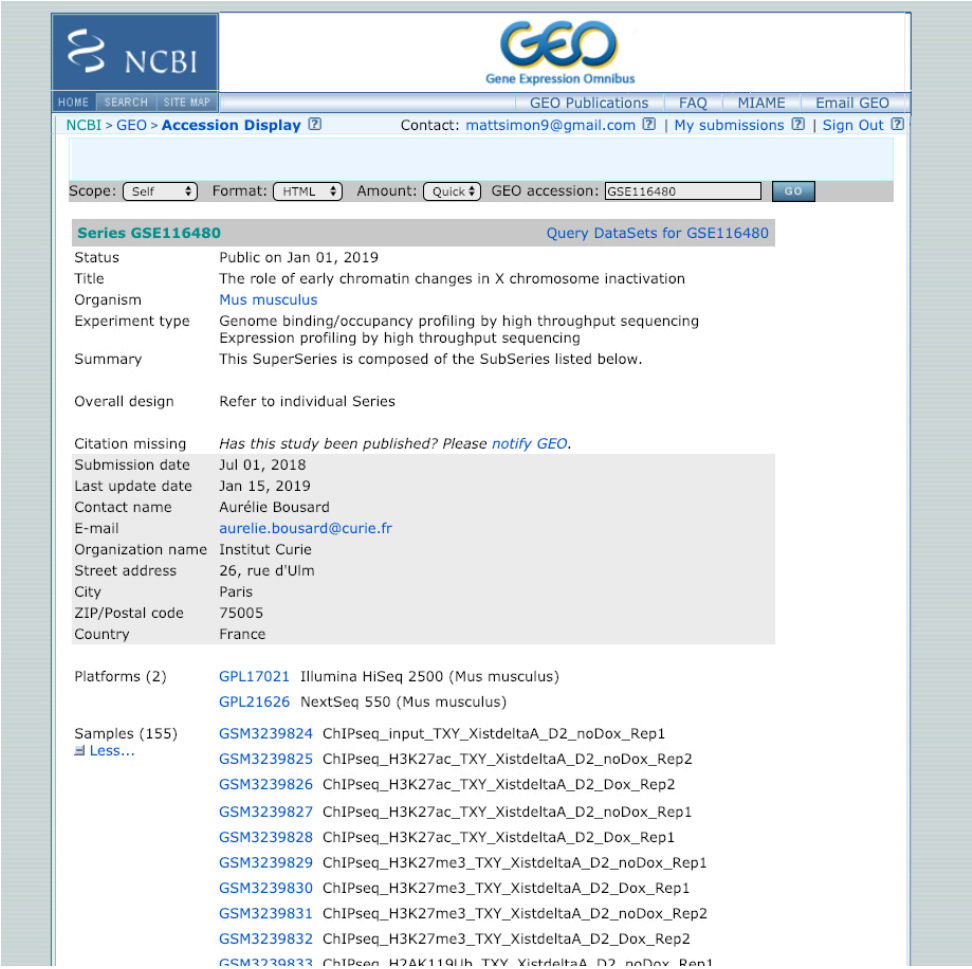
heterochromatin is X chromosome inact
mammals. Although the role for chromat
nance of the inactive state has been ex
teche and Berletch, 2015), almost nothin
role in the initiation of gene silencing.

In female mouse embryos, one of the t
randomly chosen for inactivation around t
(Lyon, 1962). This phenomenon is depende
future inactive X chromosome (Xi) by the
Xist (Penny et al., 1996). The conserved

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




```
@SRR7458968.1 HISEQ:258:CBBJMACXX:1:1101:1450:1962 length=51
NCATTTTCCATGTTTCTCATTGTAACTCATTGATATACACTGTTCTACAAA
+SRR7458968.1 HISEQ:258:CBBJMACXX:1:1101:1450:1962 length=51
#1?DFFFFHHHHHJJJJJJJJJJIJJJJIJJJJJJJJJJJJJJJJJJJJJ
@SRR7458968.2 HISEQ:258:CBBJMACXX:1:1101:1497:1977 length=51
NGGACCAGCGGGCGATTGCCAGCCCTGCCAGCCCGCGCCTTCCATTGCCGG
+SRR7458968.2 HISEQ:258:CBBJMACXX:1:1101:1497:1977 length=51
#1:DDDFDDHADHGHHGBHGIIJJJGGGG@9;@AEG:9>'3>C35B@CC##
@SRR7458968.3 HISEQ:258:CBBJMACXX:1:1101:1624:1943 length=51
NGTTCATAGCAGCCTTATTTATAATAGCCGGAAGCTGGAAAGAACCCAGAT
+SRR7458968.3 HISEQ:258:CBBJMACXX:1:1101:1624:1943 length=51
#1=DDFFFHHHHHJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJI
@SRR7458968.4 HISEQ:258:CBBJMACXX:1:1101:1688:1958 length=51
NCAGACAGAAAGACACACACACAGAGACACACATACAGACAGAAAGACACA
+SRR7458968.4 HISEQ:258:CBBJMACXX:1:1101:1688:1958 length=51
#1=DFFFFHHHHHJIJJJJJJJIJJJJJJJIJIIJJJJJJIJJJJJJIJJJ
@SRR7458968.5 HISEQ:258:CBBJMACXX:1:1101:1546:1977 length=51
NACCAAACGAGGAGTTGGTTCTTTGAGAAAATCAACAAGATAGATAAACCC
+SRR7458968.5 HISEQ:258:CBBJMACXX:1:1101:1546:1977 length=51
#1=DDFFFGHHHHIBFGICGIIIIFIIIIIIIIIIIIIIIIIIIIIIGI
@SRR7458968.6 HISEQ:258:CBBJMACXX:1:1101:1647:1989 length=51
NAAGACTTGTATTCTAATATGGACTCAGCTTGTTTGTTTAGCCTGCATTCA
+SRR7458968.6 HISEQ:258:CBBJMACXX:1:1101:1647:1989 length=51
#1=DDDDDHDHHHIIIIIIIIIIIIIIIIIIIHIIIIIIIIIIIIIIIIG
@SRR7458968.7 HISEQ:258:CBBJMACXX:1:1101:1971:1921 length=51
NGTCTTTTCACAAAGATATGTATTTAGAGGTAGCATTTGCCTTTTCATTTG
+SRR7458968.7 HISEQ:258:CBBJMACXX:1:1101:1971:1921 length=51
#4=DDFFFHHHFDGHGGIJJJJJJJFGJJIAFHJIJJJJIJIJJJJJJIIE
@SRR7458968.8 HISEQ:258:CBBJMACXX:1:1101:1756:1940 length=51
NGGGAGTGAAATATGGCGAGGAAAACTGAAAAAGGTGGAAAATTTAGAAAT
+SRR7458968.8 HISEQ:258:CBBJMACXX:1:1101:1756:1940 length=51
#1=DDFDDHHHHHJJJJJJJJGJJJJJJJJJJJJ?FHGIJJIJJJJJJIC
```

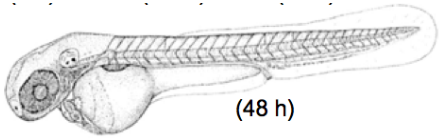
- Most journals require authors to submit their data to a database (e.g.,GEO) prior to publication.
- These databases entries contain raw data and processed data.
- These data can be use to examine the authors' claims, but also to test new hypotheses.

Central questions

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

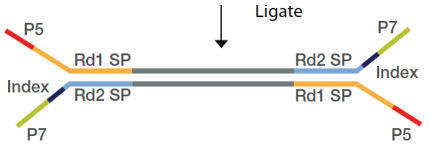
Workflow

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. Sangar 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

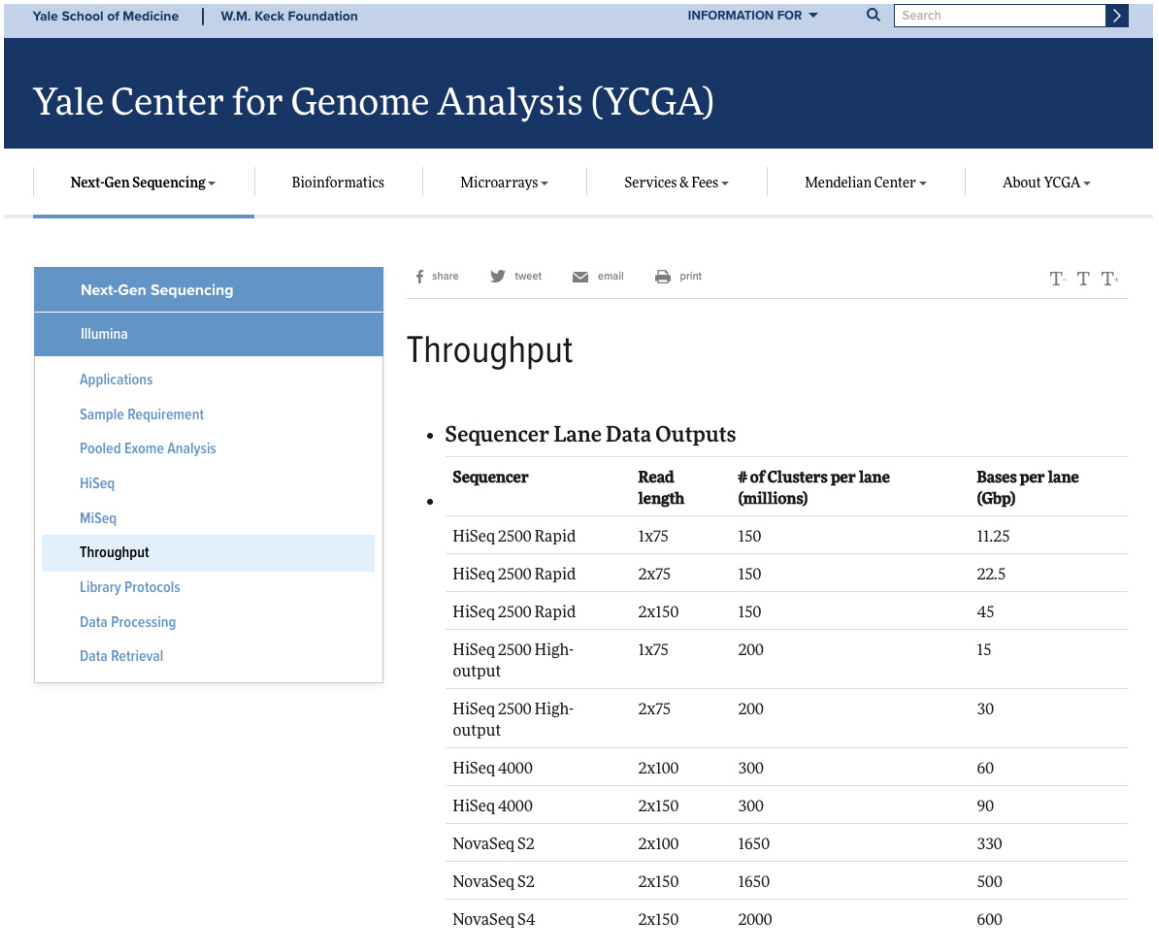
    a. Isolation
    b. Library construction

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



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=DDFFFHHHHHIJIJJJIJJJJJIJJJ?FHIDGIJ=GIHGIIIHGIJIHEHIHHGFFFFEEEDDDDDDDDDDDD
```

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

What is the output from an Illumina sequencing experiment?

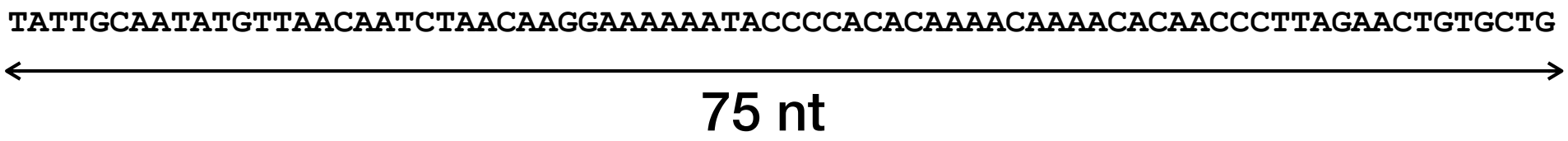
Many reads…

```
@HWI-D00306:498:HBB89ADXX:1:1101:1180:1882 1:N:0:CGATGT
NCATCACTTTCTGCACCAGCCATGACGTCAATCTTCGTCCGAACCCCAAACTCGAGATCGGAAGAGCACACGTCTG
+
#11BBDDDFFDFBFFFIIIIIIIIIIIIIIFEGIIIIIFIGAGIIFIII=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 ~ 400,000,000 reads/sequencing lane

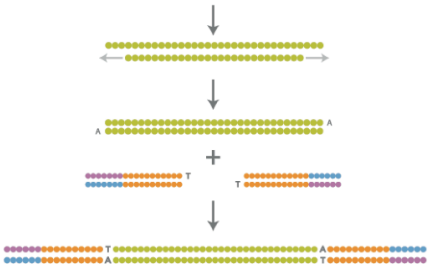
Note: This is for an Illumina HiSeq 4000 with current chemistry, but this number changes

How long are the reads?



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

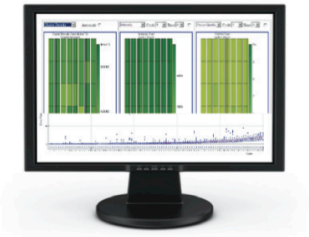
Where do these reads come from?



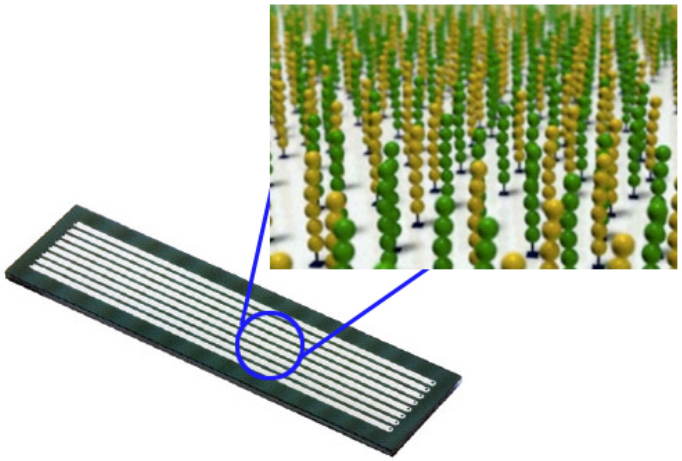
Library Preparation
~2 h [15 min hands-on (Nextera)]
< 6 h [< 3 h hands-on (TruSeq)]

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

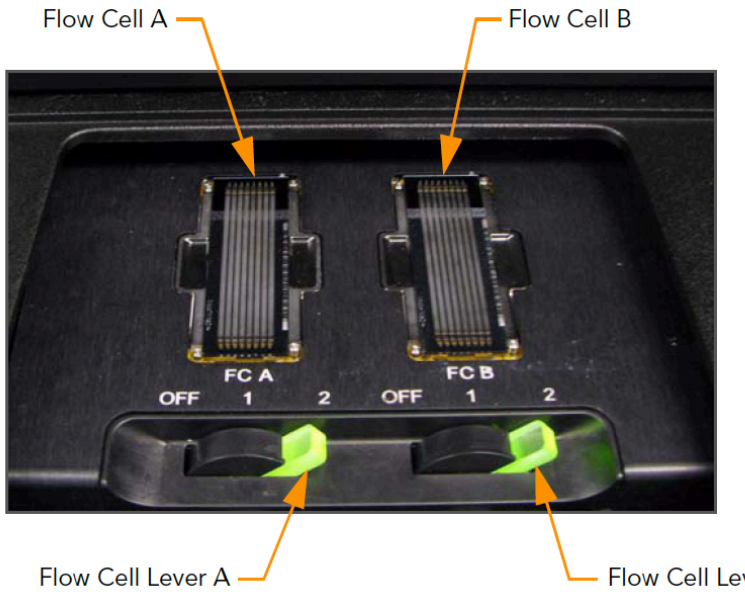
Sequencing by Synthesis
~1.5 to 11 days



CASAVA
2 days (30 min hands-on)



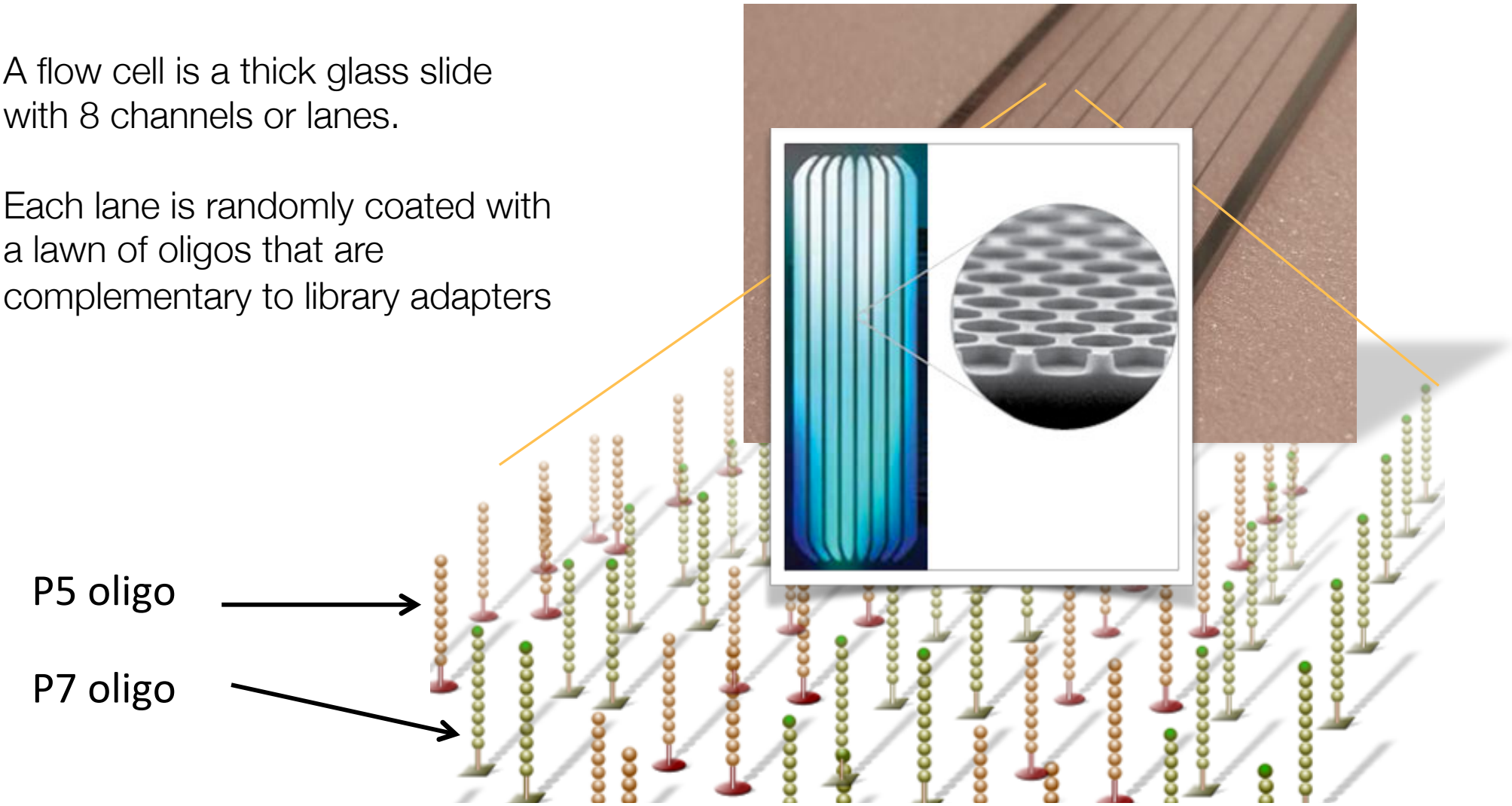
Flow cell



What is a flow cell?

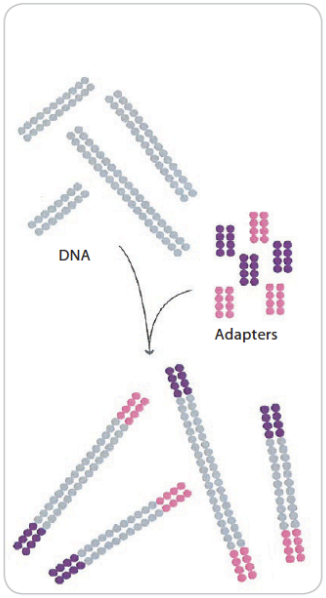
A flow cell is a thick glass slide with 8 channels or lanes.

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



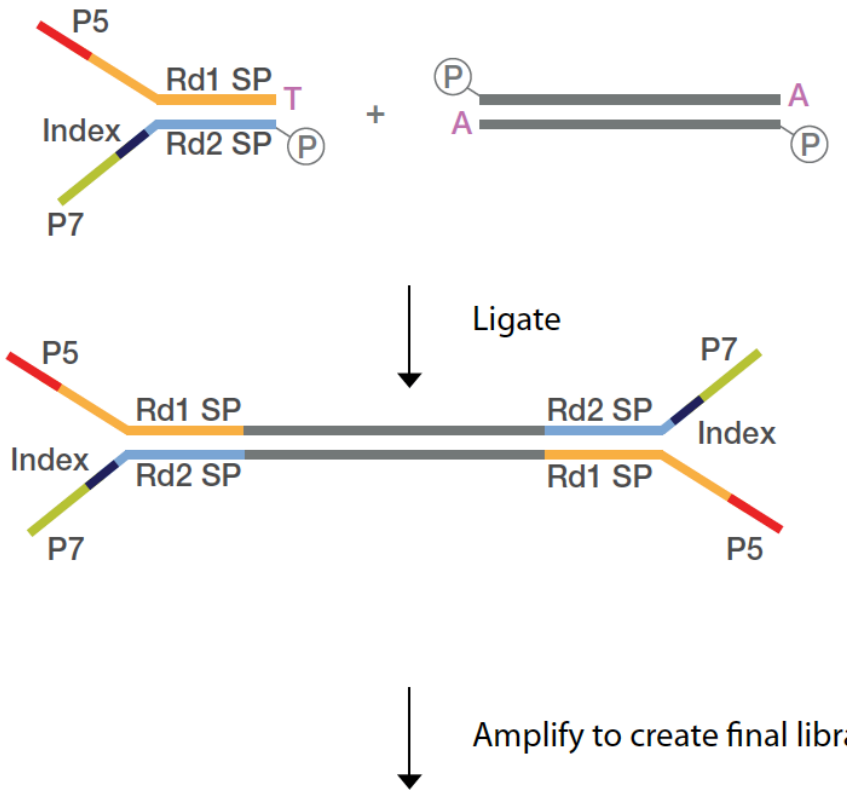
https://support.illumina.com.cn/content/illumina-marketing/amr/en/technology/next-generation-sequencing/sequencing-technology/patterned-flow-cell-technology.html

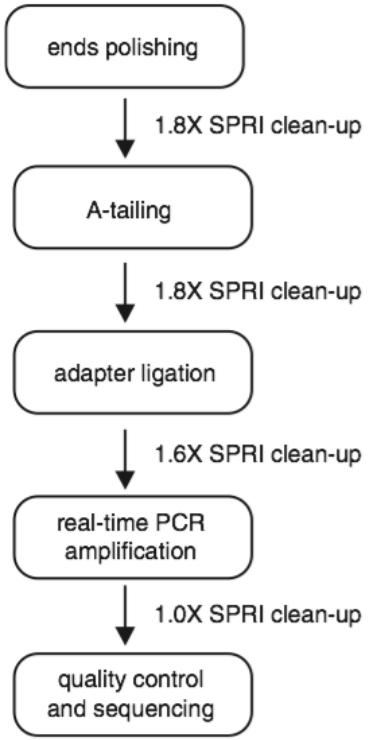
Cluster PCR
on flow cell
(8 lanes)

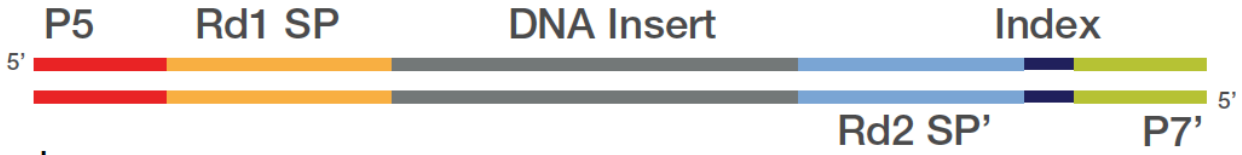


Optional: How do you make a sequencing library?

Index = unique sequence
key to identify library







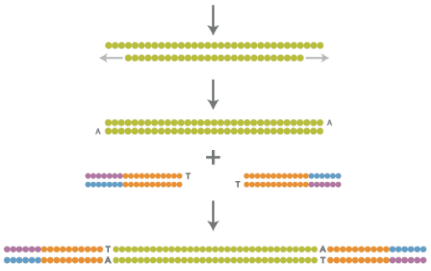
12 samples per lane

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).

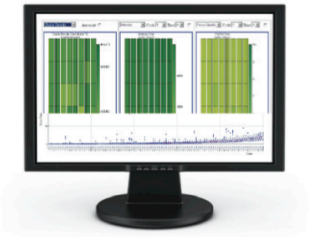
Where do these reads come from?



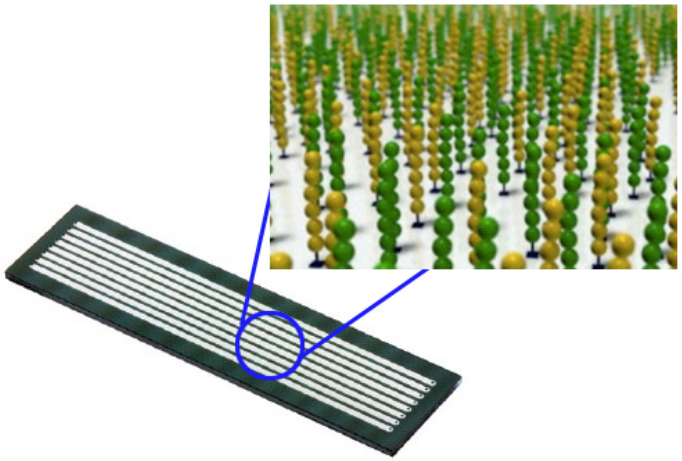
Library Preparation
~2 h [15 min hands-on (Nextera)]
< 6 h [< 3 h hands-on (TruSeq)]

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

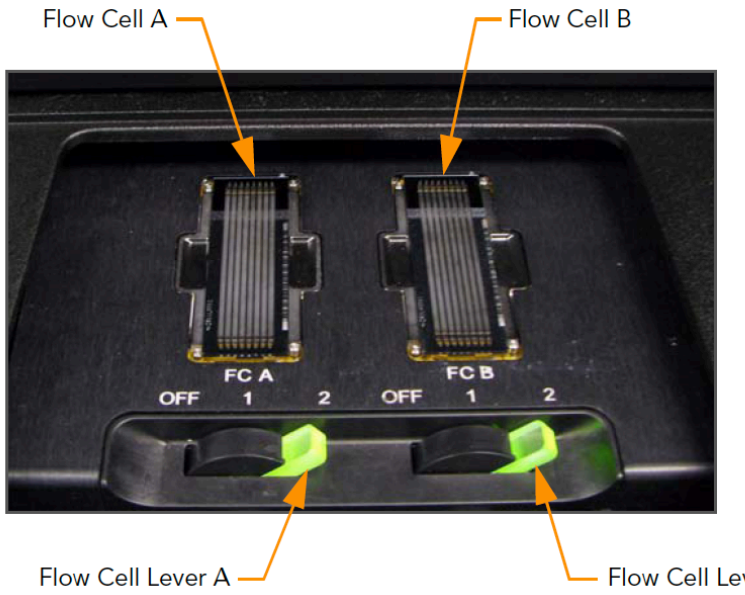
Sequencing by Synthesis
~1.5 to 11 days



CASAVA
2 days (30 min hands-on)



Flow cell



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=GIHGIIIHGIJIHEHIHHGFFFFEEEDDDDDDDDDDDD

@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. 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
NCATCACTTTCTGCACCAGCCATGACGTCAATCTTCGTCCGAACCCCAAACTCGAGATCGGAAGAGCACACGTCTG
+
#11BBDDDFFDFBFFFIIIIIIIIIIIIIIFEGIIIIIFIGAGIIFIII=FEEEEEFFFDDD=@9A@BBBBB=?BB<
```

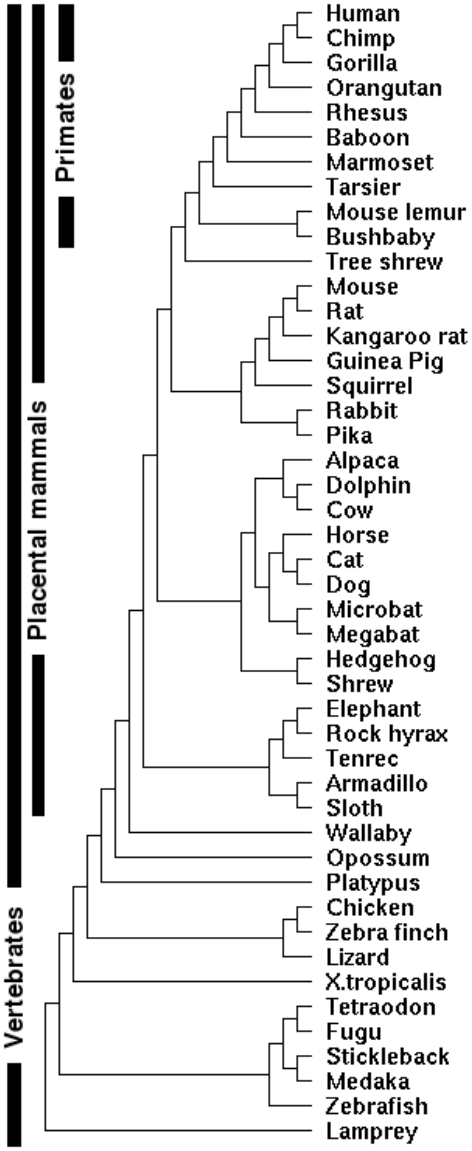
- For each single end read: Incomplete incorporation of bases.

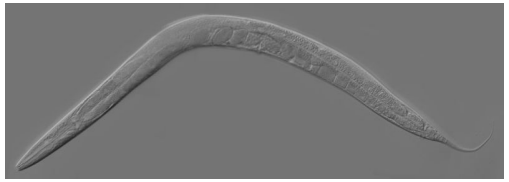
- For the size of the insert (especially for paired end analysis): Ability to get consistent clusters.

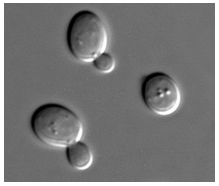
What do I do with my sequencing reads?

Source: Slate via Noonan

Many reference genomes are available





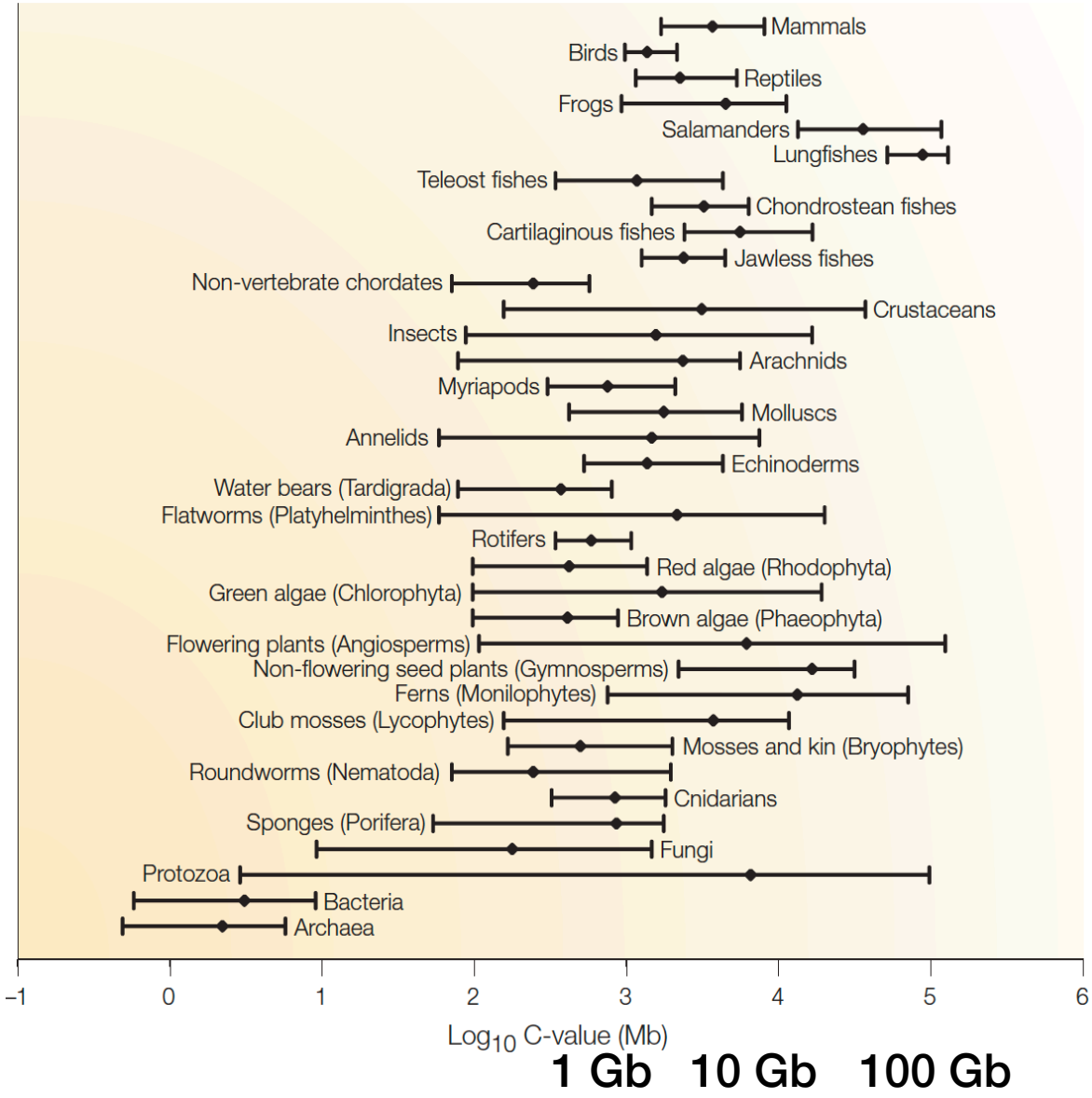


There is a wide range of genome sizes.

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**

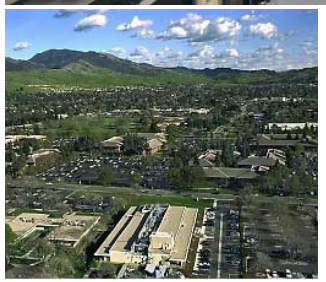
HGSC HUMAN GENOME SEQUENCING CENTER

National Human Genome Research Institute

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

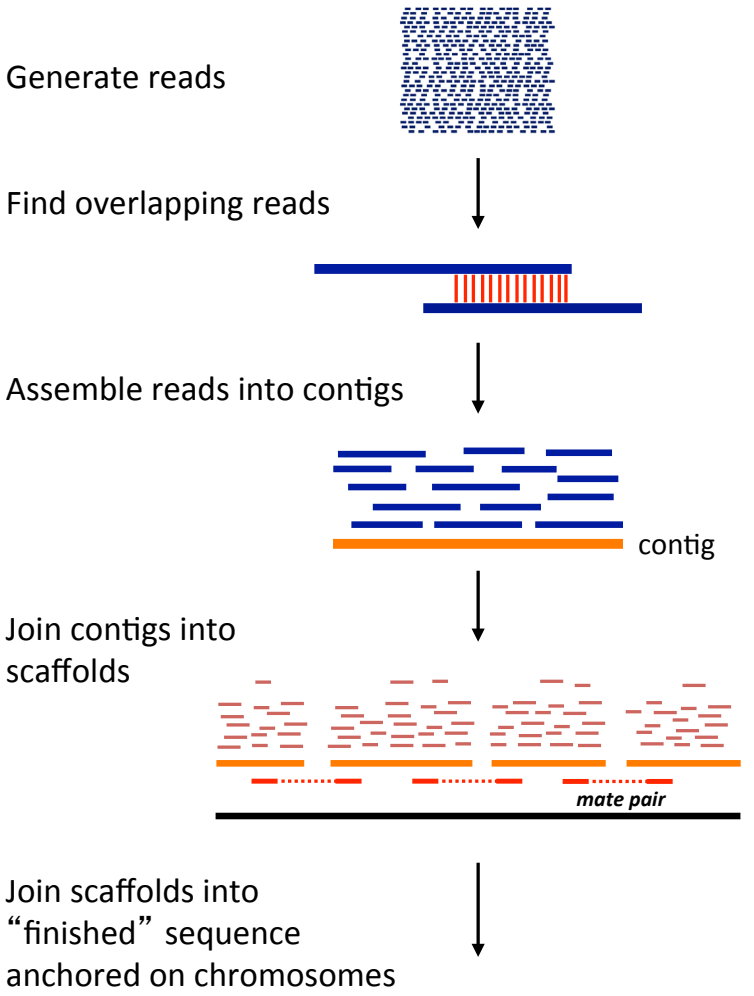
wellcome trust sanger institute

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



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

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

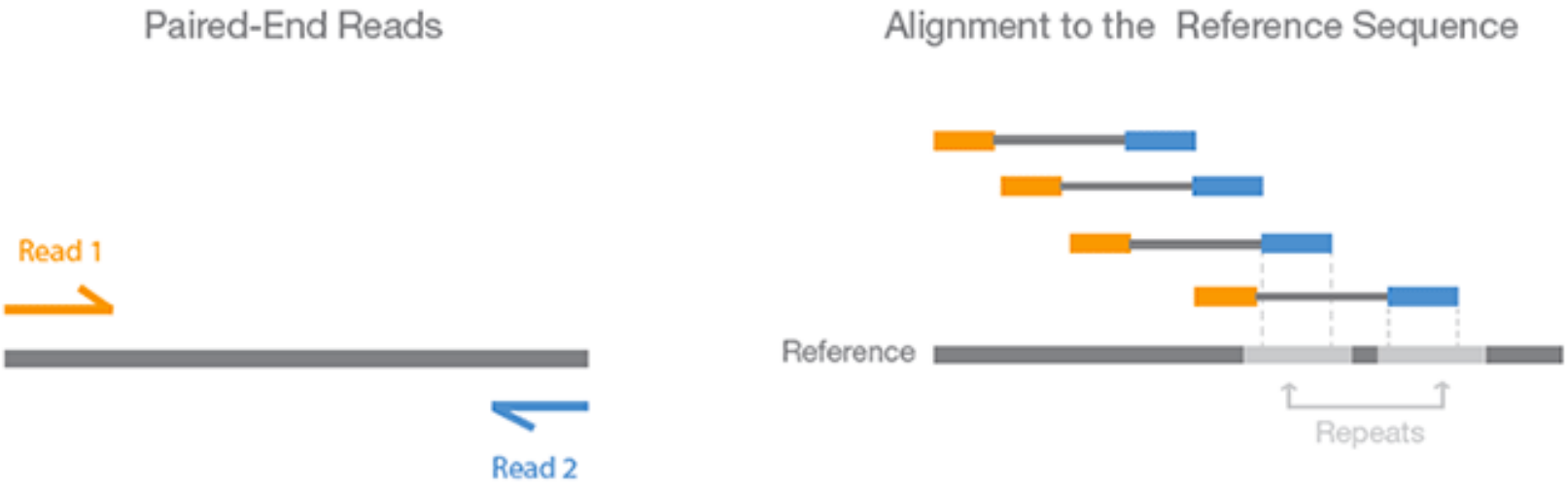
Human, Mouse, Rat: > 20x
Chimpanzee: ~6x
Squirrel: ~2x

Scaffold_0: 12,865,123 – 12,965-110

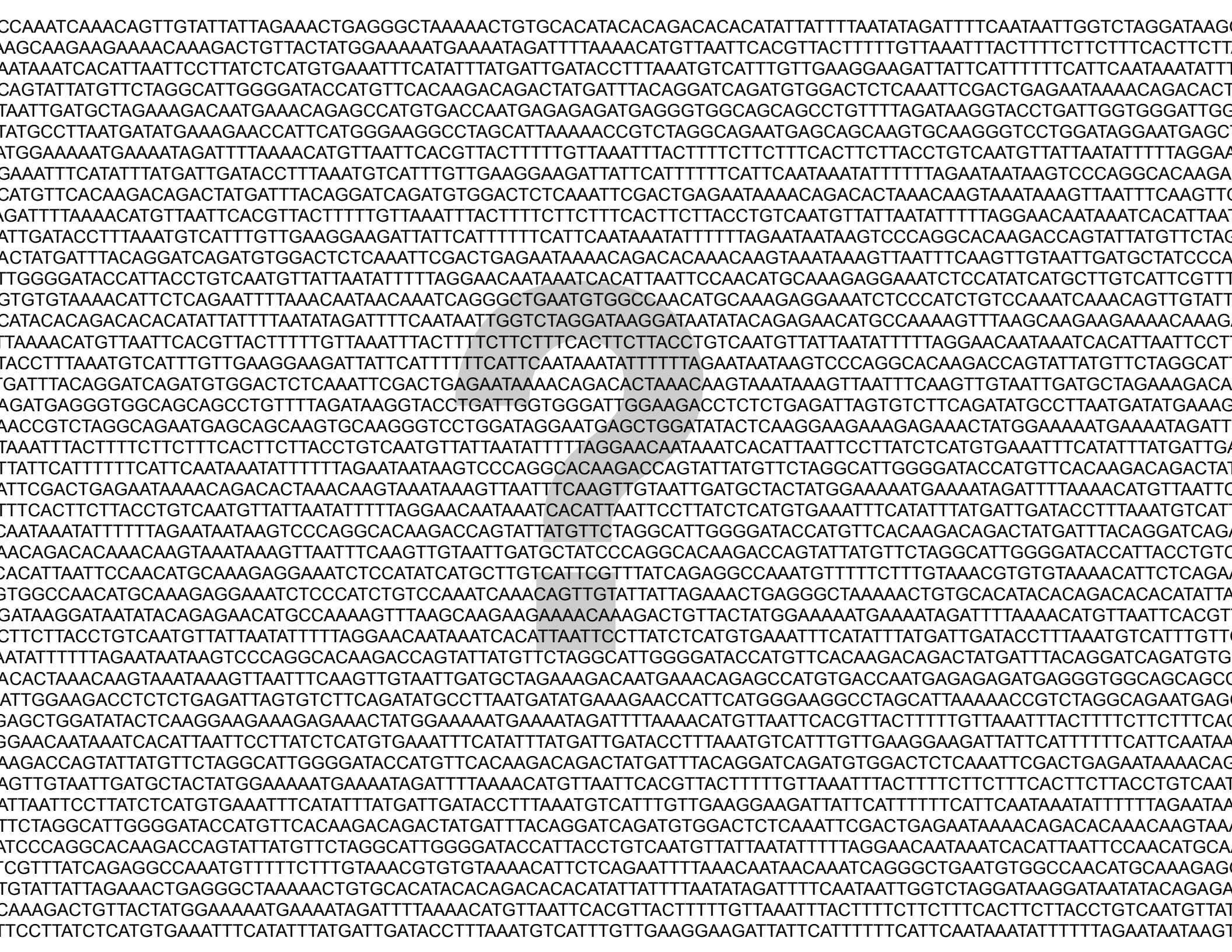
Chr5: 133,876,119 – 134,876,119

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.



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

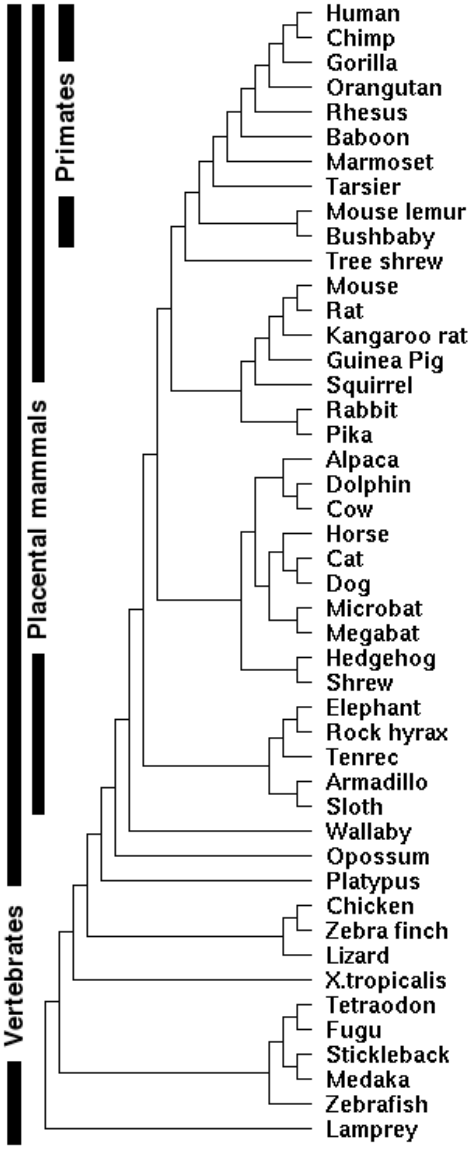
Regulatory sequences:

- Promoters
- Enhancers
- Insulators

Epigenetics:

- DNA methylation
- Chromatin

Degrees of genomic annotation vary widely



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

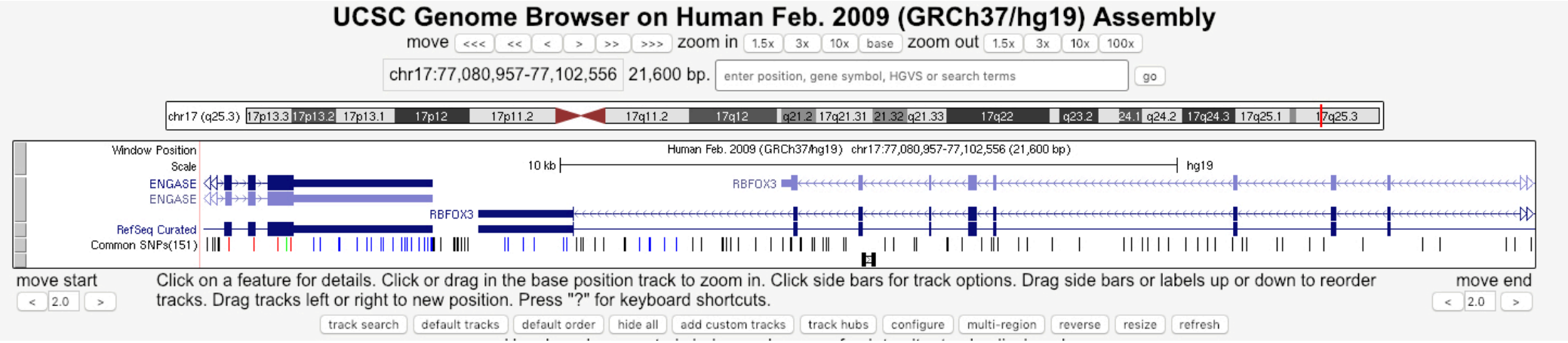
Example of a genome browser track (UCSC)

Chr5: 133,876,119 – 134,876,119

Our specific example:

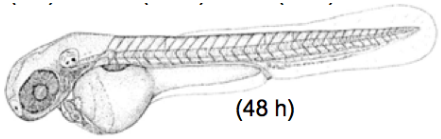
```
@HWI-ST1239:178:H0KPNADXX:2:1101:3120:1979 1:N:0:TGACCA
NCTGTAGGCTGCGTAGCCTCCCTGCAGGGTAAGTGGGAGGAGAGAGAGCAGAGGGACTTAGTGGGGCTCCCCAGGG
+
#1=DDFFFHHHHHIJIJJJIJJJJJIJJJ?FHIDGIJ=GIHGIIIHGIJIHEHIHHGFFFFEEEDDDDDDDDDDDD

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



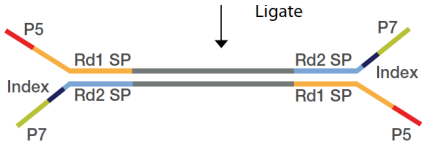
Workflow

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