Genomics I

Biomedical Data Science: Mining and Modeling
CB&B 752 • MB&B 452
Matt Simon
January 15, 2020

What is genomics?

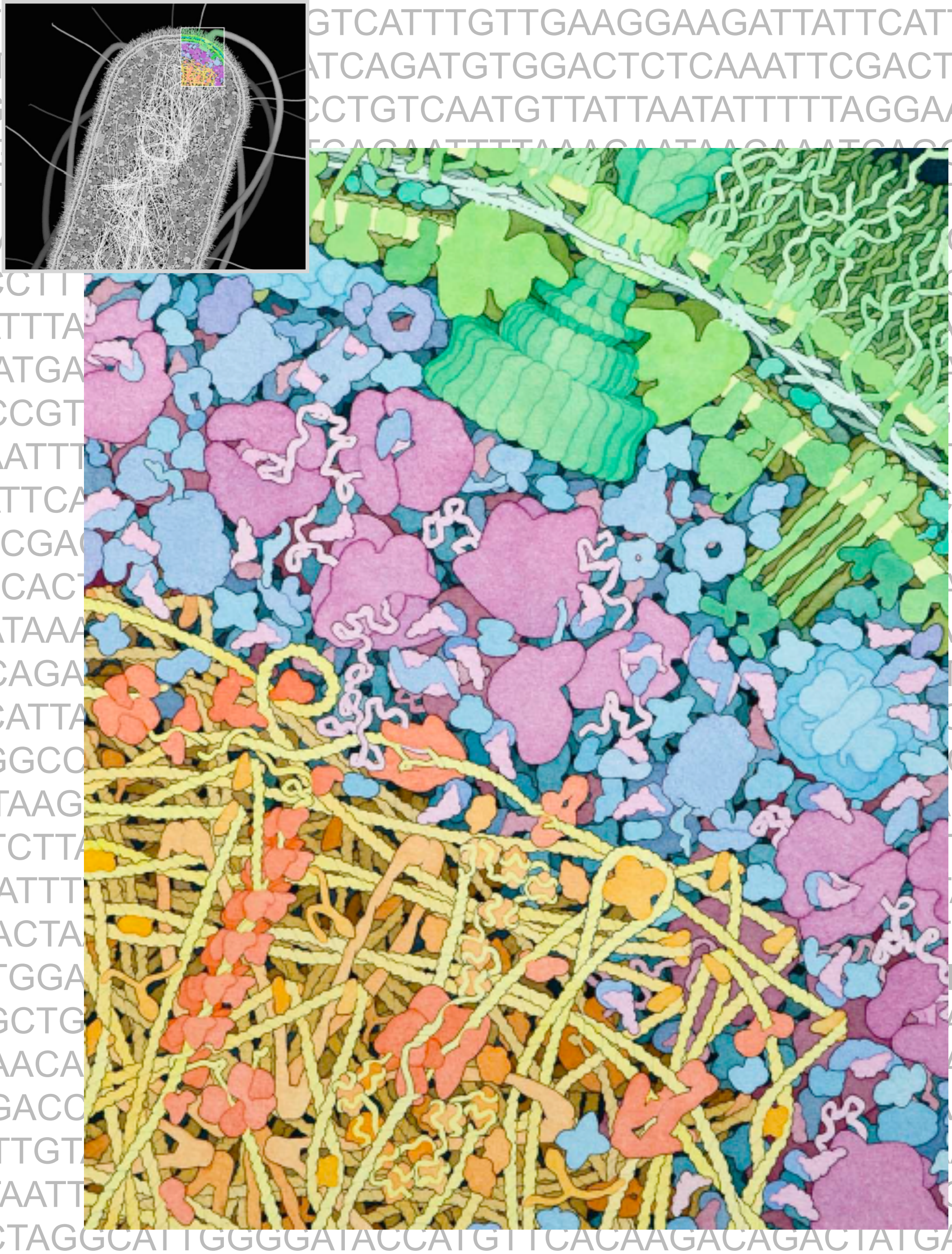
1. The **global** study of how biological **information** is encoded in genome sequence

   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.



```
CCATGTTCACAAGACAGACTATGATTTACAGGATCAGATGTGGACTCTTCAAATTCGACTGAGAATAAAACAGACACTA
TAGATTTTAAAACATGTTAATTCACGTTACTTTTTGTTAAATTTACTTTTCTTCTTTCACTTCTTACCTGTCAATGTTATTA
GATT GTCATTTGTTGAAGGAAGATTATTCATTTTTTCATTCAATAAATATTTTTTAGAATAATAAGTCC
GACT ATCAGATGTGGACTCTCAAATTCGACTGAGAATAAAACAGACACAAACAAGTAAATAAAGTTA
ATTG CCTGTCAATGTTATTAATATTTTTAGGAACAATAAATCACATTAATTCCAACATGCAAAGAGGA
CGTG GGCTGAATGTGGCCAACATGCAAAGAGGAAATCTC
ACAT TGGTCTAGGATAAGGATAATATACAGAGAACATGCC
TTTA CTTCTTTCACTTCTTACCTGTCAATGTTATTAATATTT
ATACCTT CATTCAATAAATATTTTTTAGAATAATAAGTCCCAGGC
ATGATTTA AATAAAACAGACACTAAACAAGTAAATAAAGTTAATTT
GAGATGA ATTGGTGGGATTGGAAGACCTCTCTGAGATTAGTGT
AAACCGT GATAGGAATGAGCTGGATATACTCAAGGAAGAAAG
TTAAATTT AATATTTTTAGGAACAATAAATCACATTAATTCCTTAT
ATTATTCA CCAGGCACAAGACCAGTATTATGTTCTAGGCATTGC
AATTCGAC TTAATTTCAAGTTGTAATTGATGCTACTATGGAAAAA
CTTTCACT ATAAATCACATTAATTCCTTATCTCATGTGAAATTTCA
TCAATAAA GTATTATGTTCTAGGCATTGGGGATACCATGTTCAC
AAACAGA TTGATGCTATCCCAGGCACAAGACCAGTATTATGTTC
TCACATTA CTTGTCATTCGTTTATCAGAGGCCAAATGTTTTTCTT
TGTGGCC CAAACAGTTGTATTATTAGAAACTGAGGGCTAAAAA
GGATAAG GAAGAAAACAAAGACTGTTACTATGGAAAAATGAAA
ACTTCTTA CACATTAATTCCTTATCTCATGTGAAATTTCATATTTA
AAATATTT ATGTTCTAGGCATTGGGGATACCATGTTCACAAGAC
GACACTA GCTAGAAAGACAATGAAACAGAGCCATGTGACCAA
GATTGGA TAATGATATGAAAGAACCATTCATGGAAGGCCTAG
TGAGCTG ATGAAAATAGATTTTAAAACATGTTAATTCACGTTACT
AGGAACA ATATTTATGATTGATACCTTTAAATGTCATTTGTTGAA
CAAGACC CAAGACAGACTATGATTTACAGGATCAGATGTGGAC
AAGTTGT AAACATGTTAATTCACGTTACTTTTTGTTAAATTTACT
CATTAATT CCTTTAAATGTCATTTGTTGAAGGAAGATTATTCATT
GTTCTAGGCATTGGGGATACCATGTTCACAAGACAGACTATGATTTACAGGATCAGATGTGGACTCTCAAATTCGACT
TATCCCAGGCACAAGACCAGTATTATGTTCTAGGCATTGGGGATACCATTACCTGTCAATGTTATTAATATTTTTAGGAA
TTCGTTTATCAGAGGCCAAATGTTTTTCTTTGTAAACGTGTGTAAAACATTCTCAGAATTTTAAACAATAACAAATCAGG
```

Overview

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

Credit: Jim Noonan for many of the slides

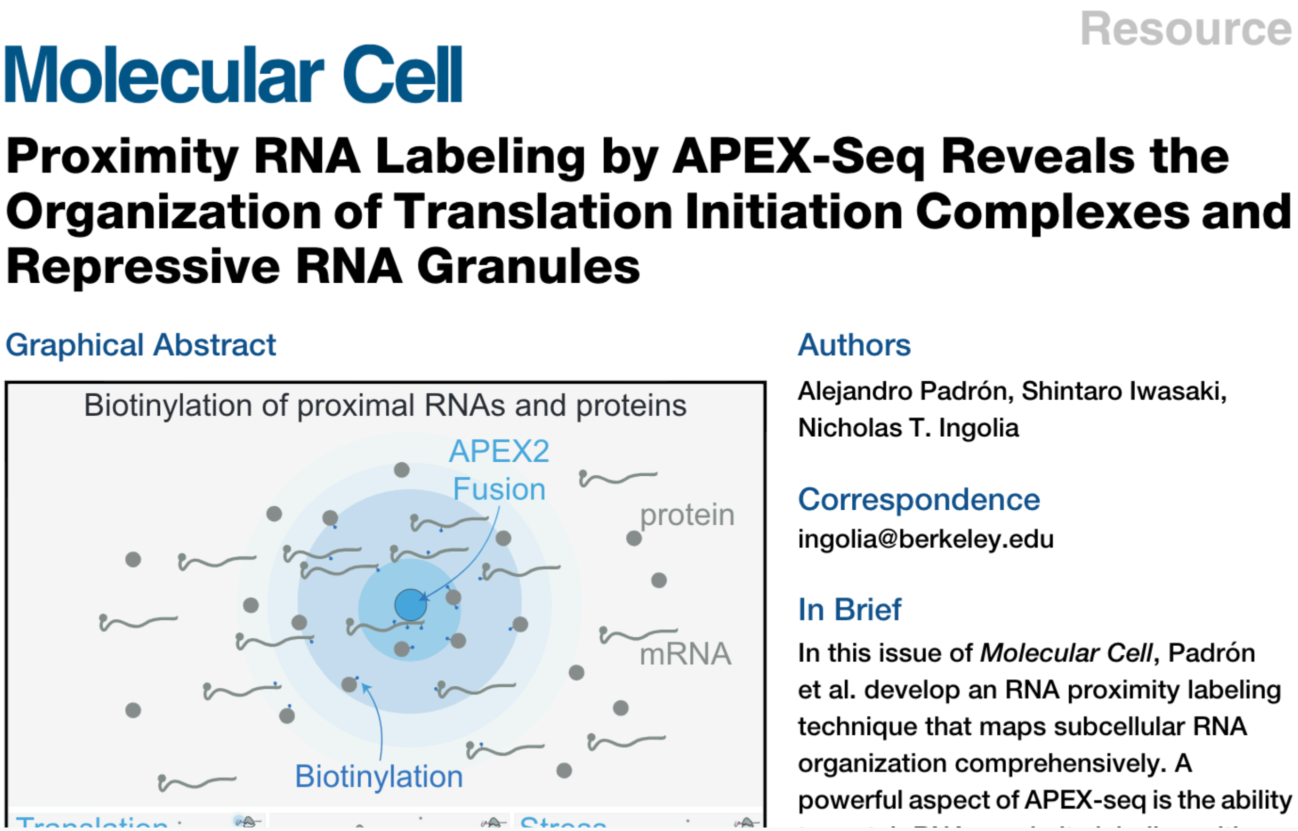
Overview

- 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



Resource

Molecular Cell

Proximity RNA Labeling by APEX-Seq Reveals the Organization of Translation Initiation Complexes and Repressive RNA Granules

Graphical Abstract

Authors

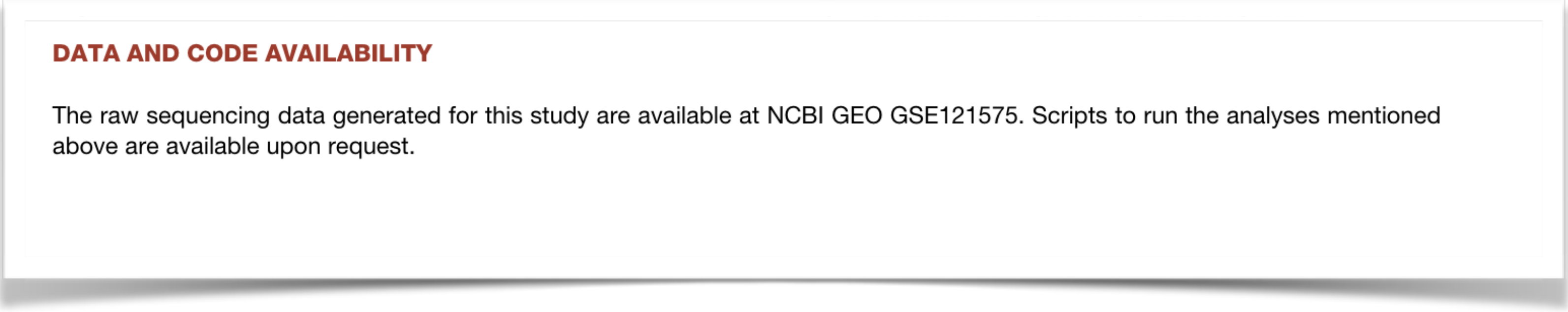
Alejandro Padrón, Shintaro Iwasaki, Nicholas T. Ingolia

Correspondence

ingolia@berkeley.edu

In Brief

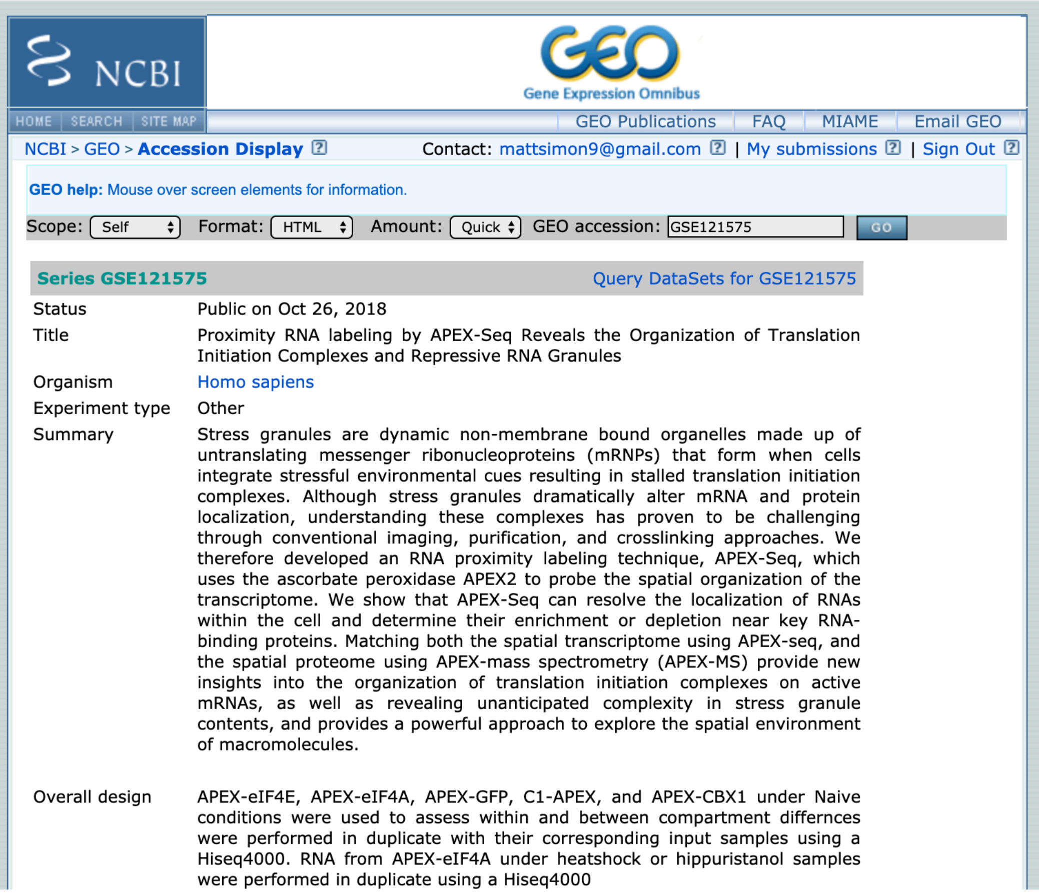
In this issue of *Molecular Cell*, Padrón et al. develop an RNA proximity labeling technique that maps subcellular RNA organization comprehensively. A powerful aspect of APEX-seq is the ability



DATA AND CODE AVAILABILITY

The raw sequencing data generated for this study are available at NCBI GEO GSE121575. Scripts to run the analyses mentioned above are available upon request.

Data can be found in genomics databases



NCBI > GEO > **Accession Display**

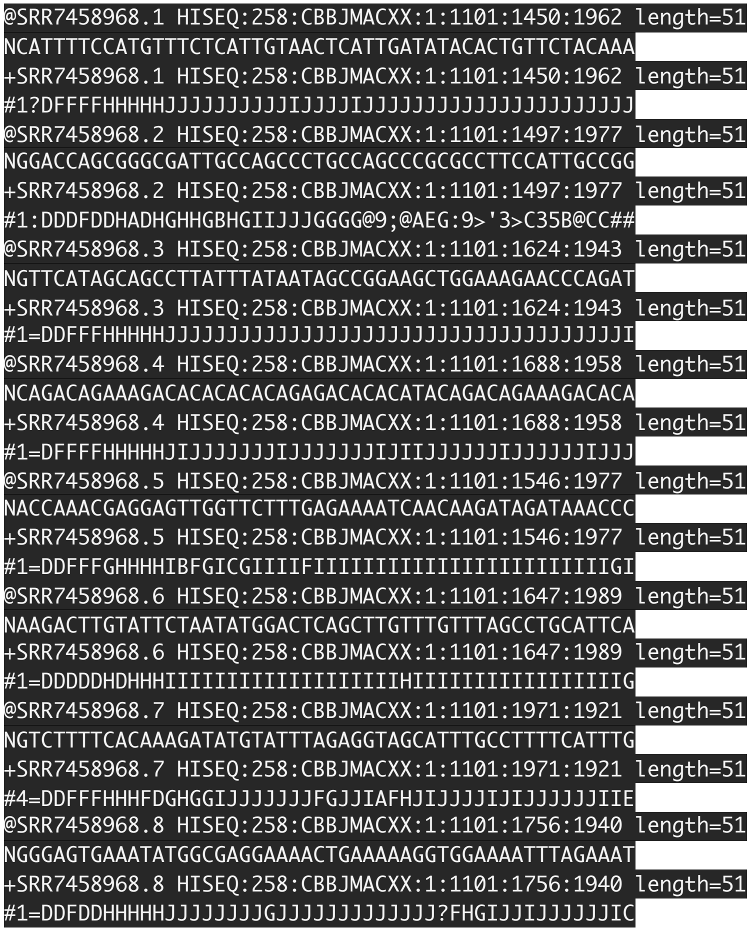
Contact: mattsimon9@gmail.com | My submissions | Sign Out

GEO help: Mouse over screen elements for information.

Scope: Self Format: HTML Amount: Quick GEO accession: GSE121575

Series GSE121575 Query DataSets for GSE121575

Status	Public on Oct 26, 2018
Title	Proximity RNA labeling by APEX-Seq Reveals the Organization of Translation Initiation Complexes and Repressive RNA Granules
Organism	Homo sapiens
Experiment type	Other
Summary	Stress granules are dynamic non-membrane bound organelles made up of untranslating messenger ribonucleoproteins (mRNPs) that form when cells integrate stressful environmental cues resulting in stalled translation initiation complexes. Although stress granules dramatically alter mRNA and protein localization, understanding these complexes has proven to be challenging through conventional imaging, purification, and crosslinking approaches. We therefore developed an RNA proximity labeling technique, APEX-Seq, which uses the ascorbate peroxidase APEX2 to probe the spatial organization of the transcriptome. We show that APEX-Seq can resolve the localization of RNAs within the cell and determine their enrichment or depletion near key RNA-binding proteins. Matching both the spatial transcriptome using APEX-seq, and the spatial proteome using APEX-mass spectrometry (APEX-MS) provide new insights into the organization of translation initiation complexes on active mRNAs, as well as revealing unanticipated complexity in stress granule contents, and provides a powerful approach to explore the spatial environment of macromolecules.
Overall design	APEX-eIF4E, APEX-eIF4A, APEX-GFP, C1-APEX, and APEX-CBX1 under Naive conditions were used to assess within and between compartment differnces were performed in duplicate with their corresponding input samples using a Hiseq4000. RNA from APEX-eIF4A under heatshock or hippuristanol samples were performed in duplicate using a Hiseq4000



```
@SRR7458968.1 HISEQ:258:CBBJMACXX:1:1101:1450:1962 length=51
NCATTTTCCATGTTTCTCATTGTAACTCATTGATATACACTGTTCTACAAA
+SRR7458968.1 HISEQ:258:CBBJMACXX:1:1101:1450:1962 length=51
#1?DFFFFHHHHHJJJJJJJJJJIJJJJIJJJJJJJJJJJJJJJJJJJJJJ
@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=DDFFFHHHHHJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJI
@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=DDFFFGHHHHIBFGICGIIIIFIIIIIIIIIIIIIIIIIIIIIIIIGI
@SRR7458968.6 HISEQ:258:CBBJMACXX:1:1101:1647:1989 length=51
NAAGACTTGTATTCTAATATGGACTCAGCTTGTTTGTTTAGCCTGCATTCA
+SRR7458968.6 HISEQ:258:CBBJMACXX:1:1101:1647:1989 length=51
#1=DDDDDHDHHHIIIIIIIIIIIIIIIIIIIHIIIIIIIIIIIIIIIIIG
@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=DDFDDHHHHHJJJJJJJJGJJJJJJJJJJJJJ?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

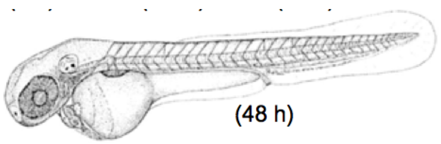
   e.g., Add known sequences to the ends.

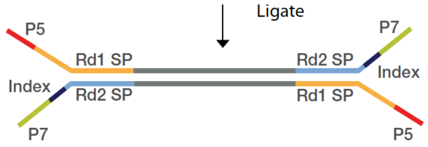
3. Sequencing

   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. First generation sequencing

    a. Maxam-Gilbert Sequencing
    b. Sangar Sequencing

2. **Second generation sequencing**

    a. Illumina Sequencing
    b. Ion Torrent

3. 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

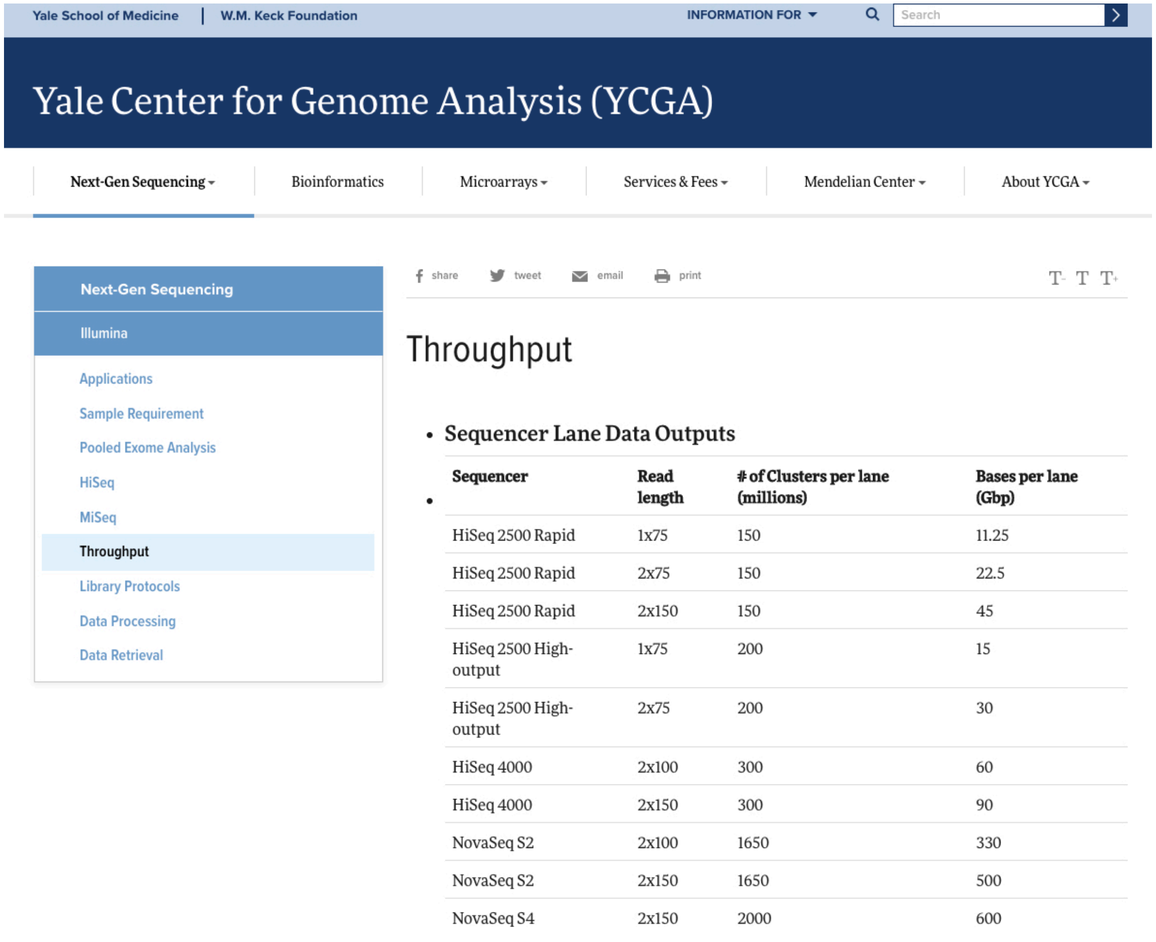
   a. Isolation
   b. Library construction

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



Throughput

- 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=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@@FFDFFHFHHHJJIJIGIIJJJJJIJJHFIJJJJJIJJJEHHJJIJJJJIIJJJJJGHHHHFBDFFFE>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
+
CCCFFFFFHHHHHIJJIIJJJIIGHHIJGGJIGIJJJJJJIJIIIJJJIIJJJJIIJGJJHCHFBDFFFDDECB
```

Generally ~ 2,000,000,000 reads/sequencing lane

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

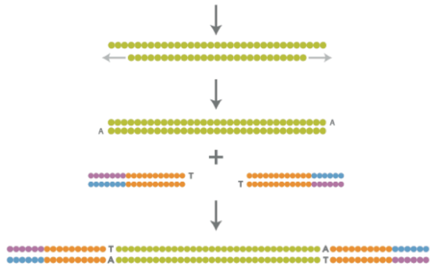
How long are the reads?

TATTGCAATATGTTAACAATCTAACAAGGAAAAAATACCCCACACAAAACAAAACACAACCCTTAGAACTGTGCTG

75 nt

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

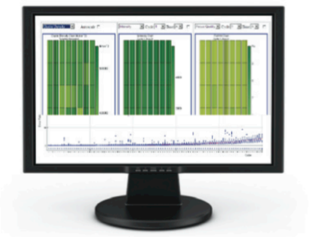
Where do these reads come from?



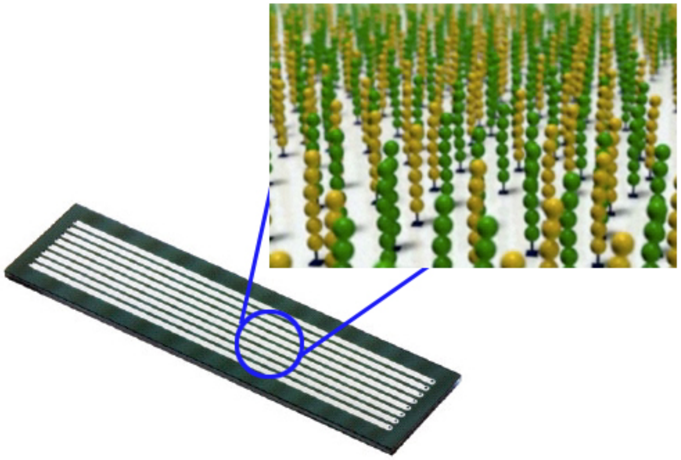
Library Preparation
~2 h [15 min hands-on (Nextera)]
< 6 h [< 3 h hands-on (TruSeq)]

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

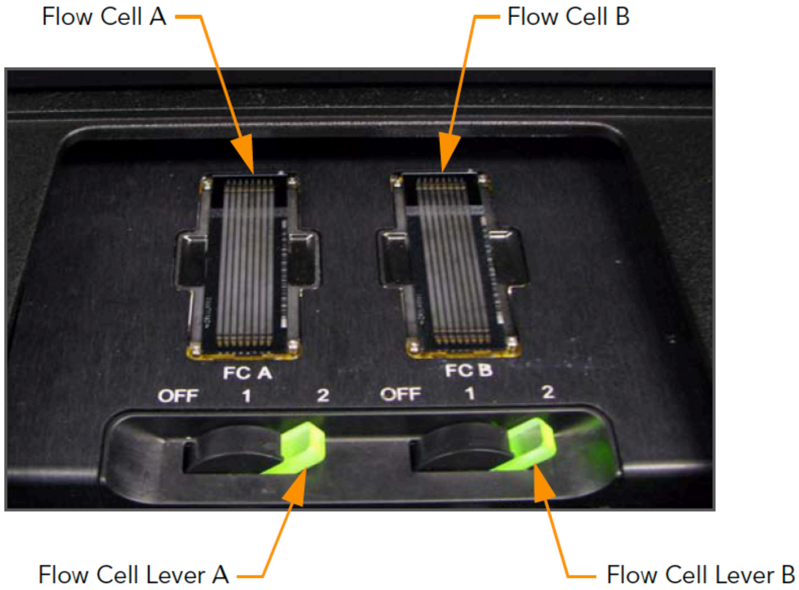
Sequencing by Synthesis
~1.5 to 11 days



CASAVA
2 days (30 min hands-on)



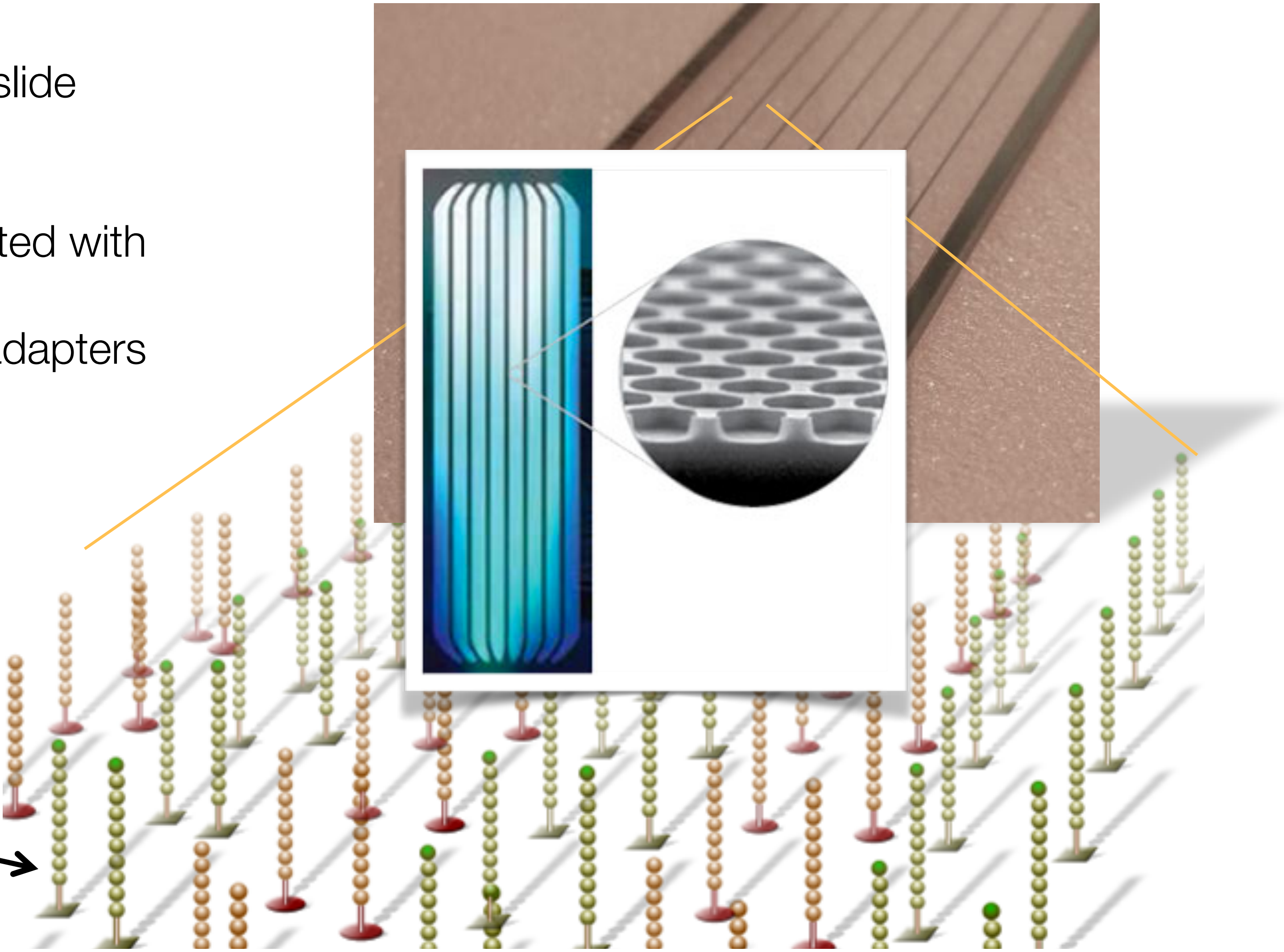
Flow cell



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

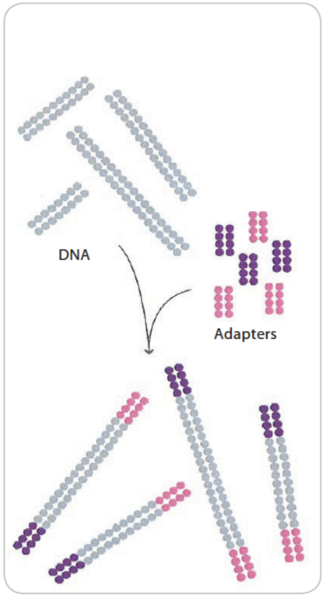


P5 oligo

P7 oligo

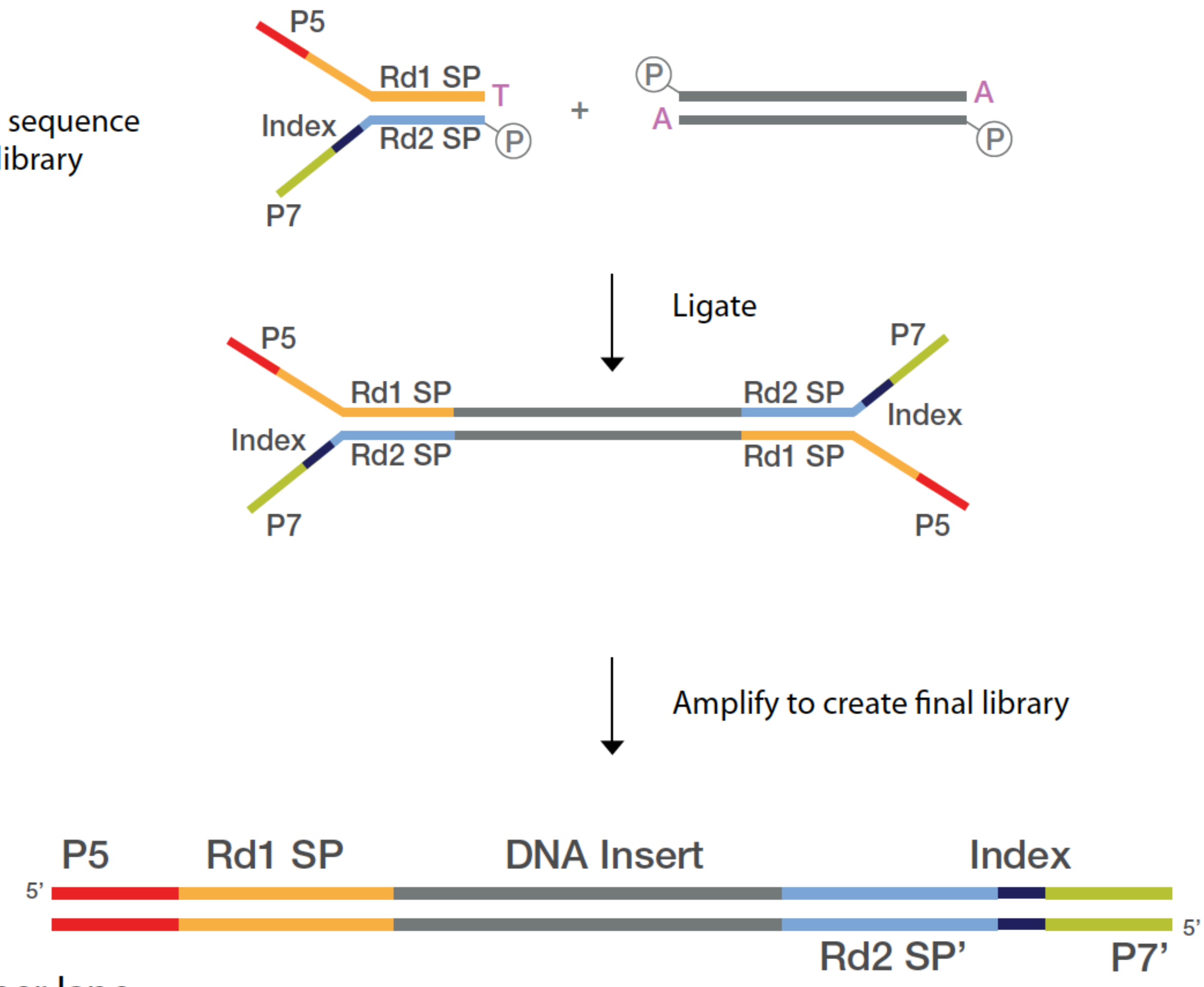
https://support.illumina.com.cn/content/illumina-marketing/amr/en/technology/next-generation-sequencing/sequencing-technology/patterned-flow-cell-technology.html

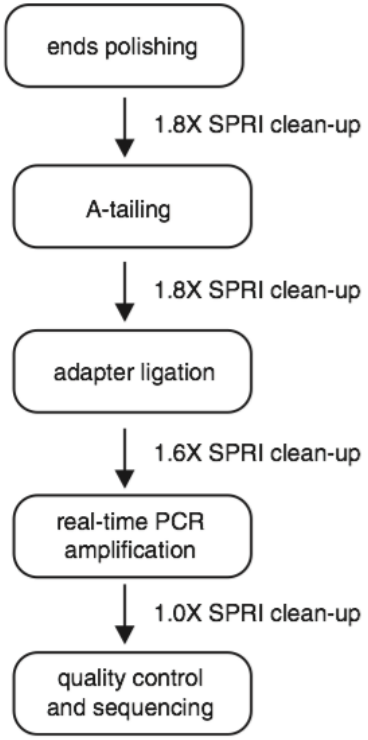
Cluster PCR
on flow cell
(8 lanes)



Optional: How do you make a sequencing library?

Index = unique sequence key to identify library





ends polishing

1.8X SPRI clean-up

A-tailing

1.8X SPRI clean-up

adapter ligation

1.6X SPRI clean-up

real-time PCR amplification

1.0X SPRI clean-up

quality control and sequencing

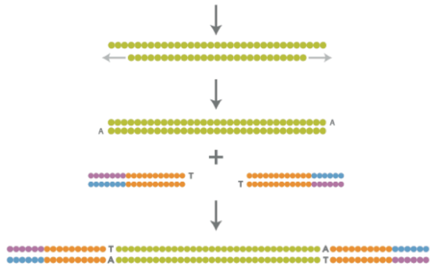
12 samples per lane

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).

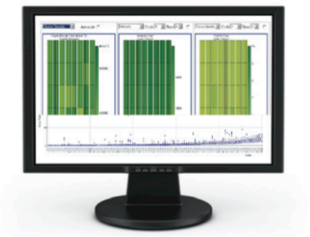
Where do these reads come from?



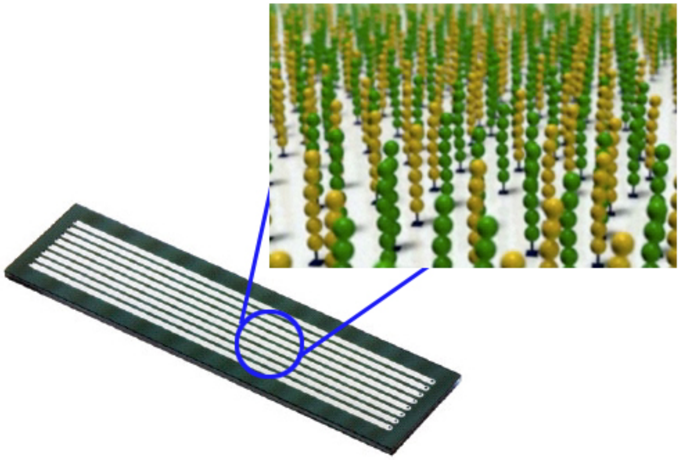
Library Preparation
~2 h [15 min hands-on (Nextera)]
< 6 h [< 3 h hands-on (TruSeq)]

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

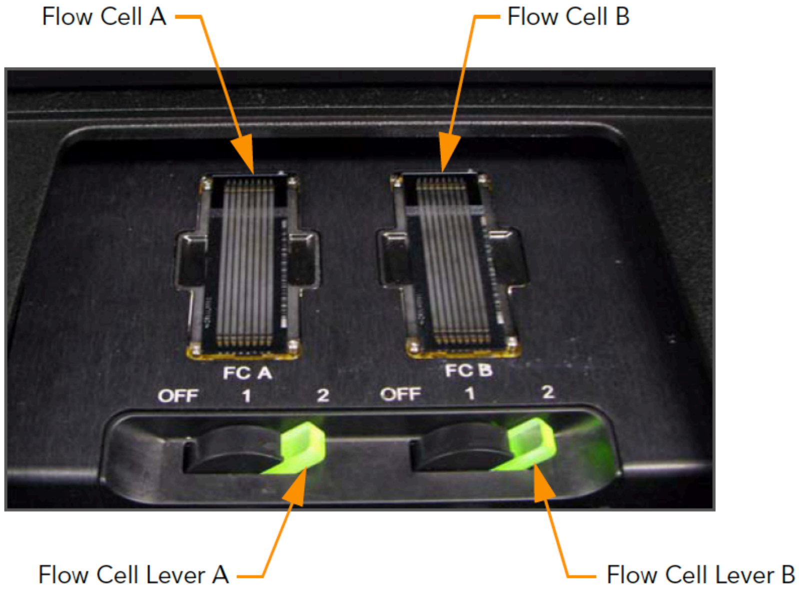
Sequencing by Synthesis
~1.5 to 11 days



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
+
#11BBDDDFFDFBFFFIIIIIIIIIIIIIFEGIIIIIFIGAGIIFIII=FEEEEEFFFDDD=@9A@BBBBB=?BB<
```

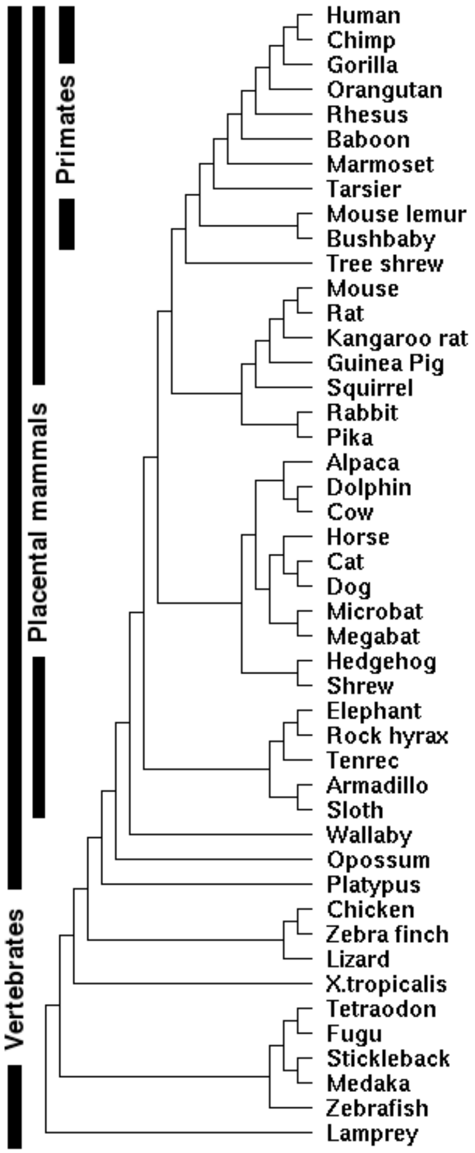
- For each single end read: Incomplete incorporation of bases.

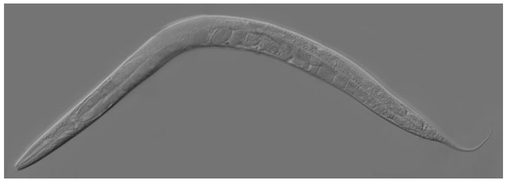
- For the size of the insert (especially for paired end analysis): Ability to get consistent clusters.

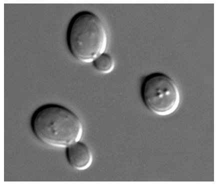
What do I do with my sequencing reads?

Source: Slate via Noonan

Many reference genomes are available





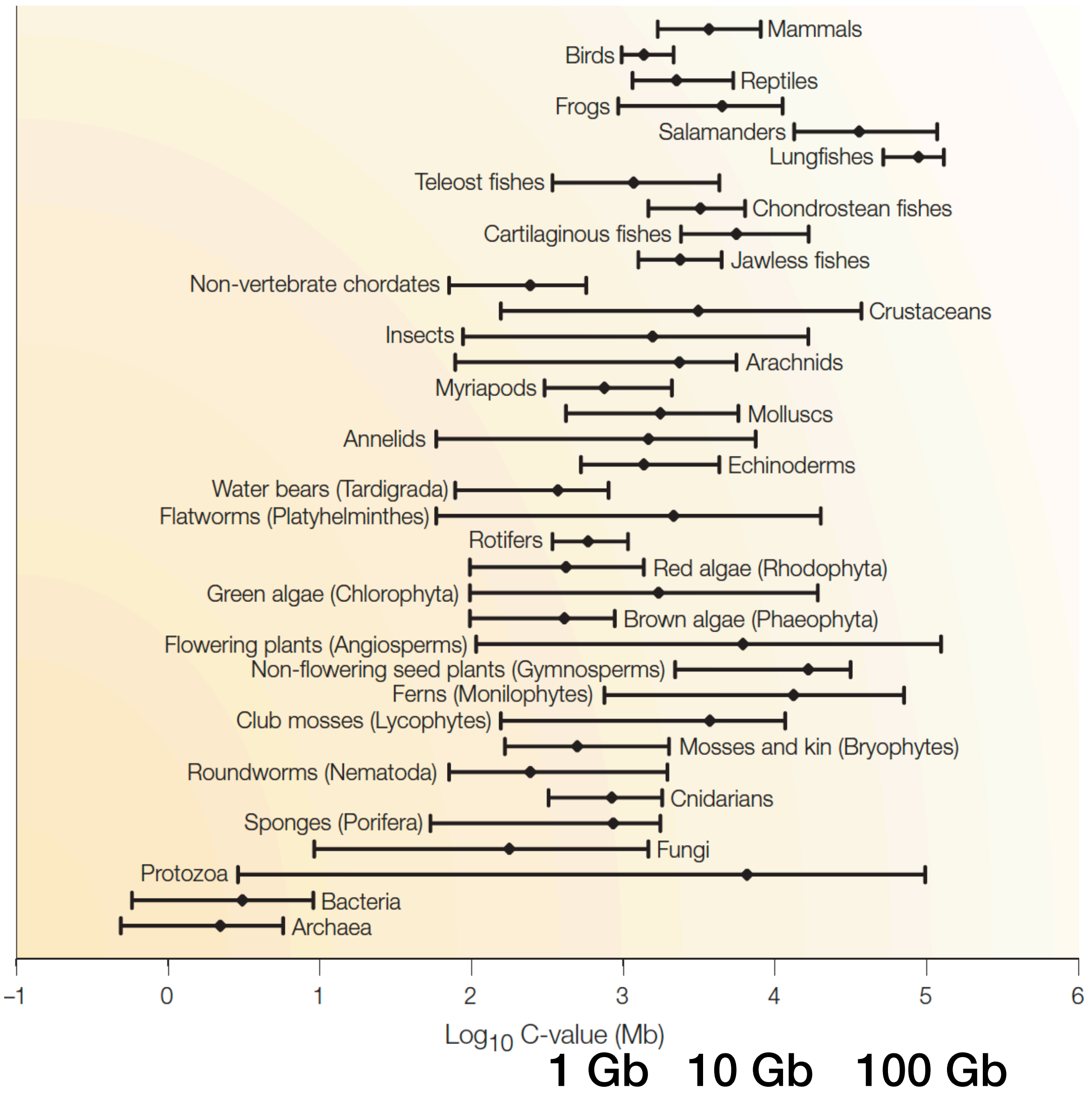


There is a wide range of genome sizes.

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

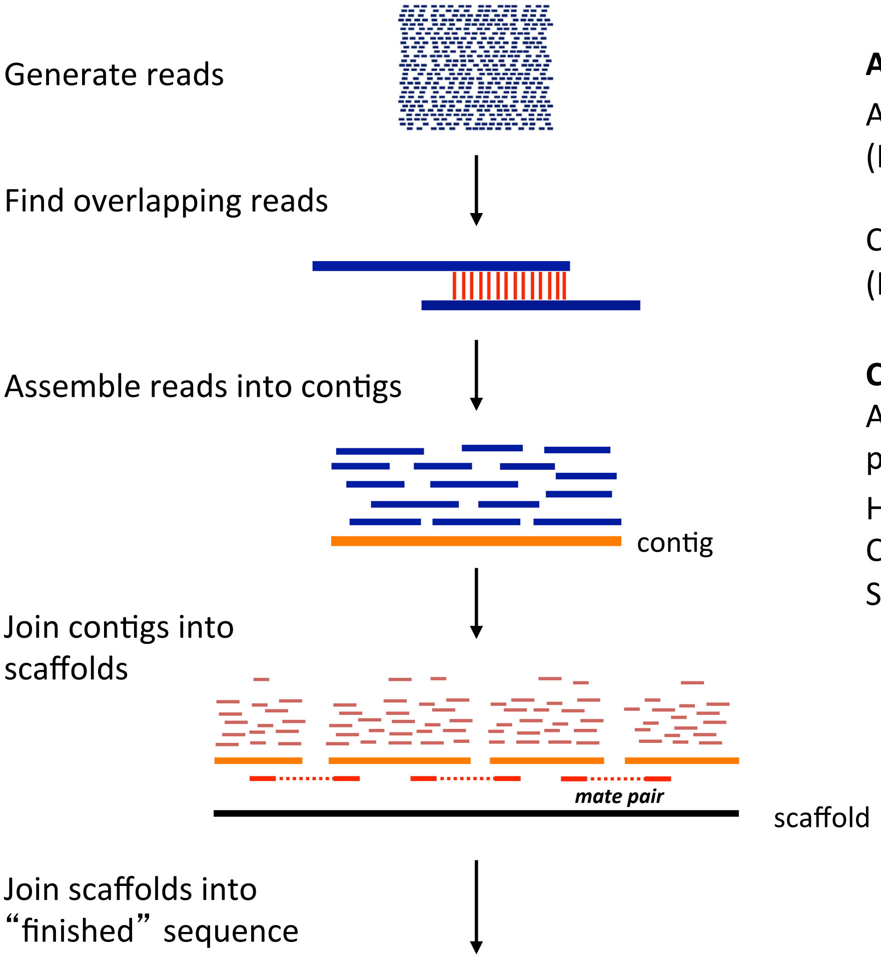
JGI
DOE JOINT GENOME INSTITUTE
US DEPARTMENT OF ENERGY
OFFICE OF SCIENCE

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



Generate reads

Find overlapping reads

Assemble reads into contigs

Join contigs into scaffolds

Join scaffolds into “finished” sequence anchored on chromosomes

AGTTGTATTATTAGAAACTGAGGGCTAAAAACTGTGCACATACACAGACACACATATTATTTTAATATAGATTTTCAATAATTGGTCTAGGATAAGGATAATATACAG

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

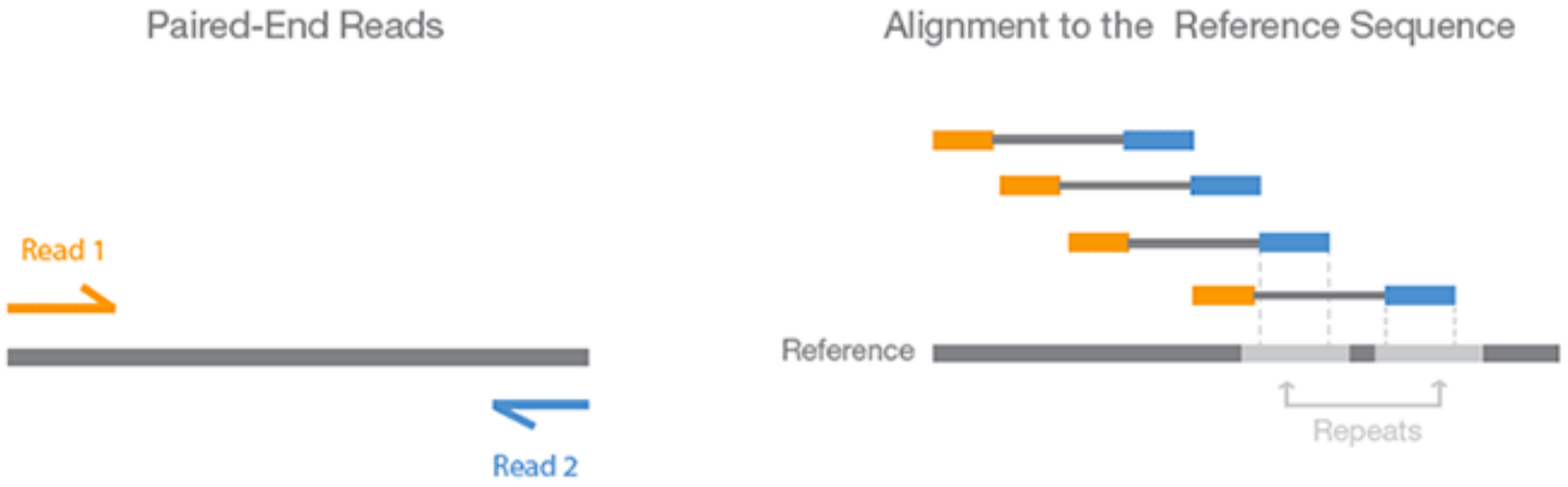
Human, Mouse, Rat:	> 20x
Chimpanzee:	~6x
Squirrel:	~2x

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.

```
CCAAATCAAACAGTTGTATTATTAGAAACTGAGGGCTAAAAACTGTGCACATACACAGACACACATATTATTTTAATATAGATTTTCAATAATTGGTCTAGGATAAGG
AAGCAAGAAGAAAACAAAGACTGTTACTATGGAAAAATGAAAATAGATTTTAAAACATGTTAATTCACGTTACTTTTTGTTAAATTTACTTTTCTTCTTTCACTTCTTA
AATAAATCACATTAATTCCTTATCTCATGTGAAATTTCATATTTATGATTGATACCTTTAAATGTCATTTGTTGAAGGAAGATTATTCATTTTTTCATTCAATAAATATTT
CAGTATTATGTTCTAGGCATTGGGGATACCATGTTCACAAGACAGACTATGATTTACAGGATCAGATGTGGACTCTCAAATTCGACTGAGAATAAAACAGACACT
TAATTGATGCTAGAAAGACAATGAAACAGAGCCATGTGACCAATGAGAGAGATGAGGGTGGCAGCAGCCTGTTTTAGATAAGGTACCTGATTGGTGGGATTGG
TATGCCTTAATGATATGAAAGAACCATTCATGGGAAGGCCTAGCATTAAAAACCGTCTAGGCAGAATGAGCAGCAAGTGCAAGGGTCCTGGATAGGAATGAGCT
ATGGAAAAATGAAAATAGATTTTAAAACATGTTAATTCACGTTACTTTTTGTTAAATTTACTTTTCTTCTTTCACTTCTTACCTGTCAATGTTATTAATATTTTTAGGAA
GAAATTTCATATTTATGATTGATACCTTTAAATGTCATTTGTTGAAGGAAGATTATTCATTTTTTCATTCAATAAATATTTTTTAGAATAATAAGTCCCAGGCACAAGA
CATGTTCACAAGACAGACTATGATTTACAGGATCAGATGTGGACTCTCAAATTCGACTGAGAATAAAACAGACACTAAACAAGTAAATAAAGTTAATTTCAAGTTG
AGATTTTAAAACATGTTAATTCACGTTACTTTTTGTTAAATTTACTTTTCTTCTTTCACTTCTTACCTGTCAATGTTATTAATATTTTTAGGAACAATAAATCACATTAAT
ATTGATACCTTTAAATGTCATTTGTTGAAGGAAGATTATTCATTTTTTCATTCAATAAATATTTTTTAGAATAATAAGTCCCAGGCACAAGACCAGTATTATGTTCTAG
ACTATGATTTACAGGATCAGATGTGGACTCTCAAATTCGACTGAGAATAAAACAGACACAAACAAGTAAATAAAGTTAATTTCAAGTTGTAATTGATGCTATCCCA
TTGGGGATACCATTACCTGTCAATGTTATTAATATTTTTAGGAACAATAAATCACATTAATTCCAACATGCAAAGAGGAAATCTCCATATCATGCTTGTCATTCGTTT
GTGTGTAAAACATTCTCAGAATTTTAAACAATAACAAATCAGGGCTGAATGTGGCCAACATGCAAAGAGGAAATCTCCCATCTGTCCAAATCAAACAGTTGTATT
CATACACAGACACACATATTATTTTAATATAGATTTTCAATAATTGGTCTAGGATAAGGATAATATACAGAGAACATGCCAAAAGTTTAAGCAAGAAGAAAACAAAGA
TTAAAACATGTTAATTCACGTTACTTTTTGTTAAATTTACTTTTCTTCTTTCACTTCTTACCTGTCAATGTTATTAATATTTTTAGGAACAATAAATCACATTAATTCCT
TACCTTTAAATGTCATTTGTTGAAGGAAGATTATTCATTTTTTCATTCAATAAATATTTTTTAGAATAATAAGTCCCAGGCACAAGACCAGTATTATGTTCTAGGCATT
TGATTTACAGGATCAGATGTGGACTCTCAAATTCGACTGAGAATAAAACAGACACTAAACAAGTAAATAAAGTTAATTTCAAGTTGTAATTGATGCTAGAAAGACA
AGATGAGGGTGGCAGCAGCCTGTTTTAGATAAGGTACCTGATTGGTGGGATTGGAAGACCTCTCTGAGATTAGTGTCTTCAGATATGCCTTAATGATATGAAAG
AACCGTCTAGGCAGAATGAGCAGCAAGTGCAAGGGTCCTGGATAGGAATGAGCTGGATATACTCAAGGAAGAAAGAGAAACTATGGAAAAATGAAAATAGATT
TAAATTTACTTTTCTTCTTTCACTTCTTACCTGTCAATGTTATTAATATTTTTAGGAACAATAAATCACATTAATTCCTTATCTCATGTGAAATTTCATATTTATGATTGA
TTATTCATTTTTTCATTCAATAAATATTTTTTAGAATAATAAGTCCCAGGCACAAGACCAGTATTATGTTCTAGGCATTGGGGATACCATGTTCACAAGACAGACTAT
ATTCGACTGAGAATAAAACAGACACTAAACAAGTAAATAAAGTTAATTTCAAGTTGTAATTGATGCTACTATGGAAAAATGAAAATAGATTTTAAAACATGTTAATTC
TTTCACTTCTTACCTGTCAATGTTATTAATATTTTTAGGAACAATAAATCACATTAATTCCTTATCTCATGTGAAATTTCATATTTATGATTGATACCTTTAAATGTCATT
CAATAAATATTTTTTAGAATAATAAGTCCCAGGCACAAGACCAGTATTATGTTCTAGGCATTGGGGATACCATGTTCACAAGACAGACTATGATTTACAGGATCAGA
AACAGACACAAACAAGTAAATAAAGTTAATTTCAAGTTGTAATTGATGCTATCCCAGGCACAAGACCAGTATTATGTTCTAGGCATTGGGGATACCATTACCTGTC
CACATTAATTCCAACATGCAAAGAGGAAATCTCCATATCATGCTTGTCATTCGTTTATCAGAGGCCAAATGTTTTTCTTTGTAAACGTGTGTAAAACATTCTCAGAA
GTGGCCAACATGCAAAGAGGAAATCTCCCATCTGTCCAAATCAAACAGTTGTATTATTAGAAACTGAGGGCTAAAAACTGTGCACATACACAGACACACATATTA
GATAAGGATAATATACAGAGAACATGCCAAAAGTTTAAGCAAGAAGAAAACAAAGACTGTTACTATGGAAAAATGAAAATAGATTTTAAAACATGTTAATTCACGT
CTTCTTACCTGTCAATGTTATTAATATTTTTAGGAACAATAAATCACATTAATTCCTTATCTCATGTGAAATTTCATATTTATGATTGATACCTTTAAATGTCATTTGTTG
AATATTTTTTAGAATAATAAGTCCCAGGCACAAGACCAGTATTATGTTCTAGGCATTGGGGATACCATGTTCACAAGACAGACTATGATTTACAGGATCAGATGTG
ACACTAAACAAGTAAATAAAGTTAATTTCAAGTTGTAATTGATGCTAGAAAGACAATGAAACAGAGCCATGTGACCAATGAGAGAGATGAGGGTGGCAGCAGCC
ATTGGAAGACCTCTCTGAGATTAGTGTCTTCAGATATGCCTTAATGATATGAAAGAACCATTCATGGGAAGGCCTAGCATTAAAAACCGTCTAGGCAGAATGAGC
GAGCTGGATATACTCAAGGAAGAAAGAGAAACTATGGAAAAATGAAAATAGATTTTAAAACATGTTAATTCACGTTACTTTTTGTTAAATTTACTTTTCTTCTTTCAC
GGAACAATAAATCACATTAATTCCTTATCTCATGTGAAATTTCATATTTATGATTGATACCTTTAAATGTCATTTGTTGAAGGAAGATTATTCATTTTTTCATTCAATAA
AAGACCAGTATTATGTTCTAGGCATTGGGGATACCATGTTCACAAGACAGACTATGATTTACAGGATCAGATGTGGACTCTCAAATTCGACTGAGAATAAAACAG
AGTTGTAATTGATGCTACTATGGAAAAATGAAAATAGATTTTAAAACATGTTAATTCACGTTACTTTTTGTTAAATTTACTTTTCTTCTTTCACTTCTTACCTGTCAAT
ATTAATTCCTTATCTCATGTGAAATTTCATATTTATGATTGATACCTTTAAATGTCATTTGTTGAAGGAAGATTATTCATTTTTTCATTCAATAAATATTTTTTAGAATAA
TTCTAGGCATTGGGGATACCATGTTCACAAGACAGACTATGATTTACAGGATCAGATGTGGACTCTCAAATTCGACTGAGAATAAAACAGACACAAACAAGTAA
ATCCCAGGCACAAGACCAGTATTATGTTCTAGGCATTGGGGATACCATTACCTGTCAATGTTATTAATATTTTTAGGAACAATAAATCACATTAATTCCAACATGCA
TCGTTTATCAGAGGCCAAATGTTTTTCTTTGTAAACGTGTGTAAAACATTCTCAGAATTTTAAACAATAACAAATCAGGGCTGAATGTGGCCAACATGCAAAGAG
TGTATTATTAGAAACTGAGGGCTAAAAACTGTGCACATACACAGACACACATATTATTTTAATATAGATTTTCAATAATTGGTCTAGGATAAGGATAATATACAGAGA
CAAAGACTGTTACTATGGAAAAATGAAAATAGATTTTAAAACATGTTAATTCACGTTACTTTTTGTTAAATTTACTTTTCTTCTTTCACTTCTTACCTGTCAATGTTAT
TTCCTTATCTCATGTGAAATTTCATATTTATGATTGATACCTTTAAATGTCATTTGTTGAAGGAAGATTATTCATTTTTTCATTCAATAAATATTTTTTAGAATAATAAGT
```

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

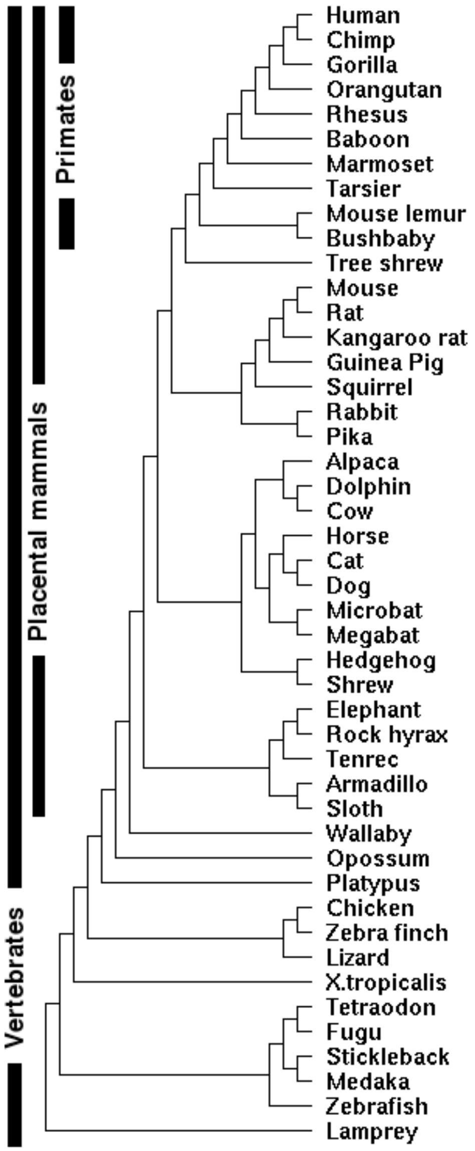
Genetic variation:
- SNPs and CNVs

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

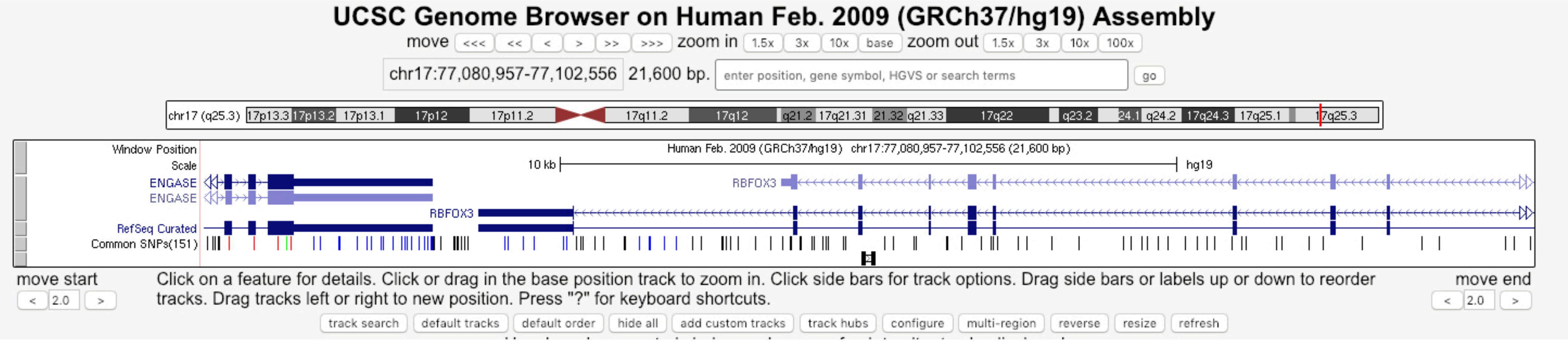
Example of a genome browser track (UCSC)

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<@@@@@@@@@?@@@?@@??????@@??@?????????????????>??????????@>???@@@?@@??????
```



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

chr17:77,080,957-77,102,556 21,600 bp.

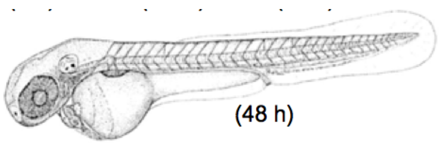
Human Feb. 2009 (GRCh37/hg19) chr17:77,080,957-77,102,556 (21,600 bp)

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.

Workflow

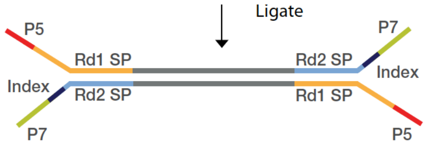
1. Isolation of sample.

   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.

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