Genomics Part I

Matt Simon
Dept. of Molecular Biophysics & Biochemistry
Chemical Biology Institute
January 19, 2018

What is genomics?

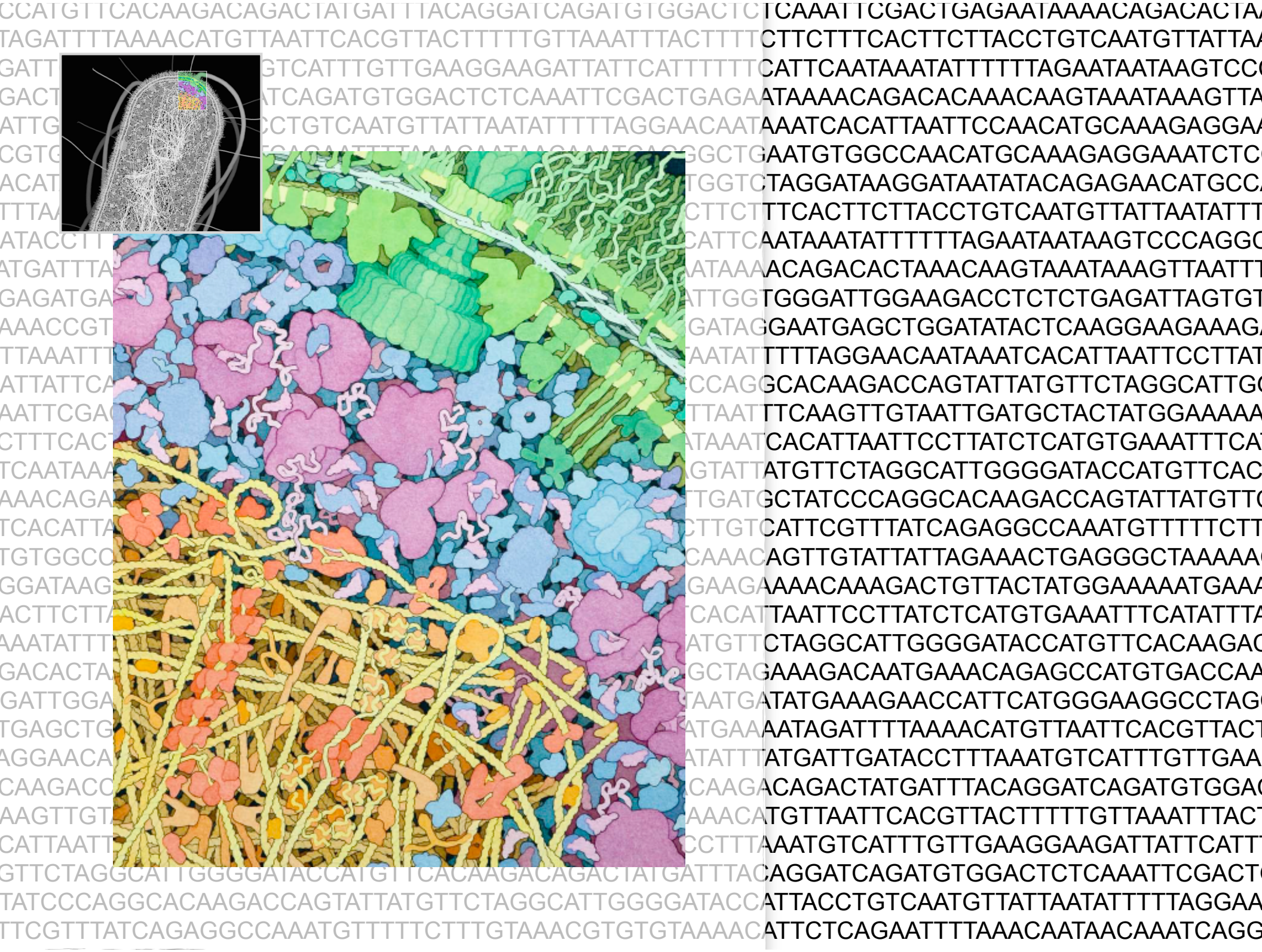
1. The **global** study of how biological **information** is encoded in genome sequence

   Genes
   Regulatory sequences
   Genetic variation

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

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

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



Overview

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

Credit: Jim Noonan for many of the slides

Importance of genomics data

Cold Spring Harbor Laboratory

bioRχiv

THE PREPRINT SERVER FOR BIOLOGY

HOME | ABOUT | SUBMIT | ALERTS / RSS | CHANNELS

Search

Advanced Search

New Results

scNMT-seq enable
methylation and t

Stephen J. Clark, Ricard
Celia Alda-Catalinas, Felix K

doi: https://doi.org/10.1101

Article

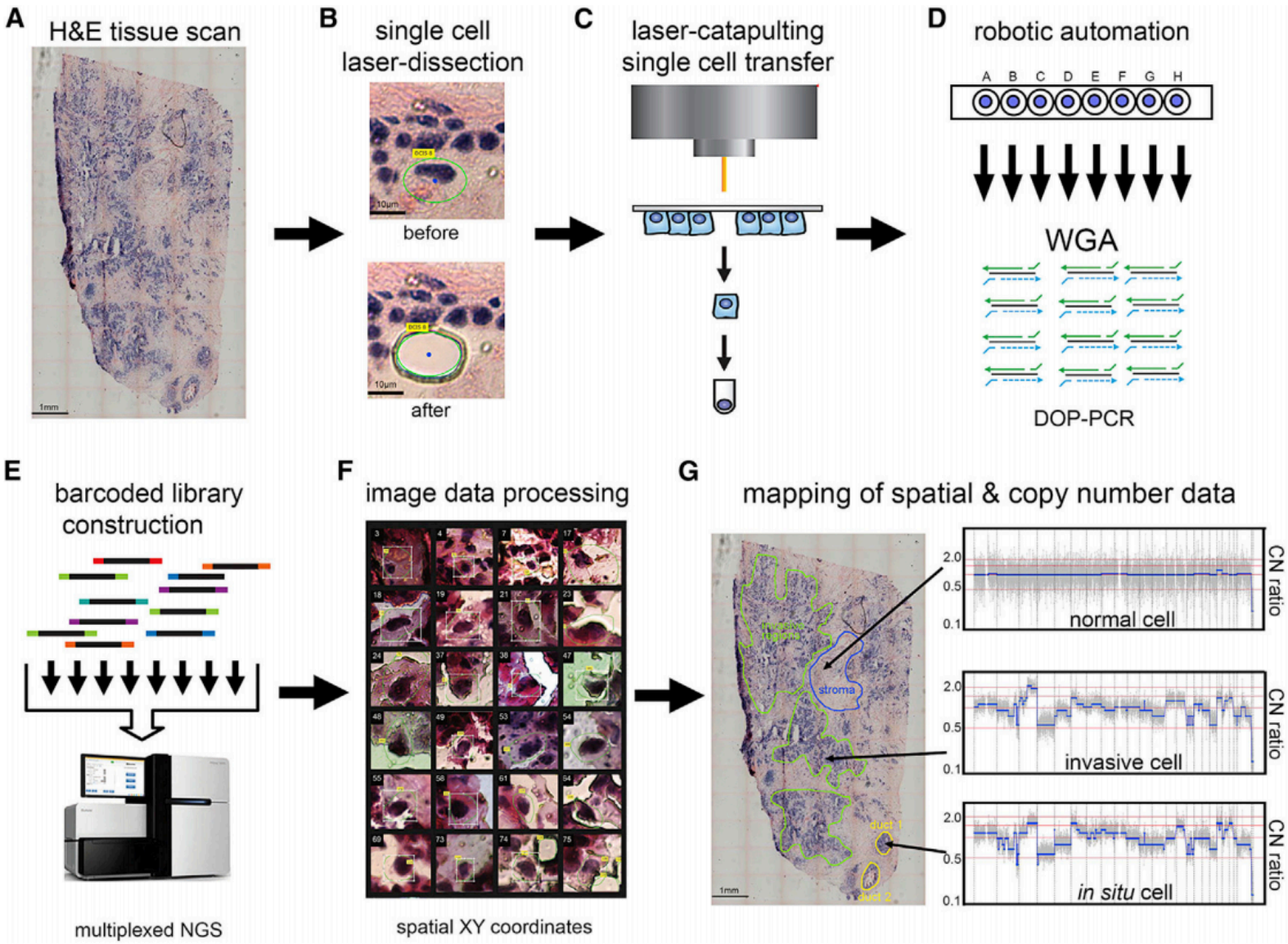
Cell

Multiclonal Invasion in Breast Tumors Identified by Topographic Single Cell Sequencing

Anna K. Casasent,[1,2] Aislyn Schalck,[1,2] Ruli Gao,[1] Emi Sei,[1] Annalyssa Long,[1] William Pangburn,[1] Tod Casasent,[3] Funda Meric-Bernstam,[4] Mary E. Edgerton,[5,*] and Nicholas E. Navin[1,2,3,6,*]

[1]Department of Genetics
[2]Graduate School of Biomedical Sciences
[3]Department of Bioinformatics and Computational Biology
[4]Department of Investigational Cancer Therapeutics
[5]Department of Pathology
The University of Texas MD Anderson Cancer Center, Houston, TX, USA
[6]Lead Contact
*Correspondence: medgerton@mdanderson.org (M.E.E.), nnavin@mdanderson.org (N.E.N.)
https://doi.org/10.1016/j.cell.2017.12.007

Importance of genomics data



Importance of genomics data

Single Cell Barcoded Library Construction

Single cell amplified DOP-PCR products that passed QC were sonicated to 250bp using the S220 acoustic sonicator (Covaris). Following sonication, TA-ligation based Illumina libraries were prepared as previously described (Gao et al., 2016). Alterations to this protocol included increased ligation time at 20°C for 30 minutes and PCR amplification cycles were adjusted according to input DNA (8 cycles for 1ug, 9 cycles for 500ng, and 10 cycles for 200ng). The insert size distributions of pooled multiplexed libraries were measured using the Bioanalyzer 2100 or Tape Station (Agilent). Multiplexed libraries were sequenced for 76 cycles using single-end or paired-end flow cells lanes on the HiSeq2000 or HiSeq4000 systems (Illumina).

Deposited Data		
Single cell copy number and exome LCM data	NCBI Sequence Read Archive	SRP116771

Data can be found in genomics databases

SRX3375957: DNA Exome Seq of Homo sapiens: Breast tissue in situ componet of synchronous ductal carcinoma in situ
1 ILLUMINA (Illumina HiSeq 4000) run: 174.2M spots, 27.7G bases, 11Gb downloads

Design: SeqCap EZ Human Exome Kit v2.0

Submitted by: MD Anderson

Study: Synchronous ductal carcinoma in situ single cell DNA sequencing
PRJNA397565 • SRP116771 • All experiments • All runs
show Abstract

Sample: DC12
SAMN07842034 • SRS2646464 • All experiments • All runs
Organism: Homo sapiens

Library:
Name: P3_DCIS_Exome
Instrument: Illumina HiSeq 4000
Strategy: WXS
Source: GENOMIC
Selection: Hybrid Selection
Layout: PAIRED

Runs: 1 run, 174.2M spots, 27.7G bases, 11Gb

Run	# of Spots	# of Bases	Size	Published
SRR6269851	174,168,383	27.7G	11Gb	2017-11-08

ID: 4712622

- 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.

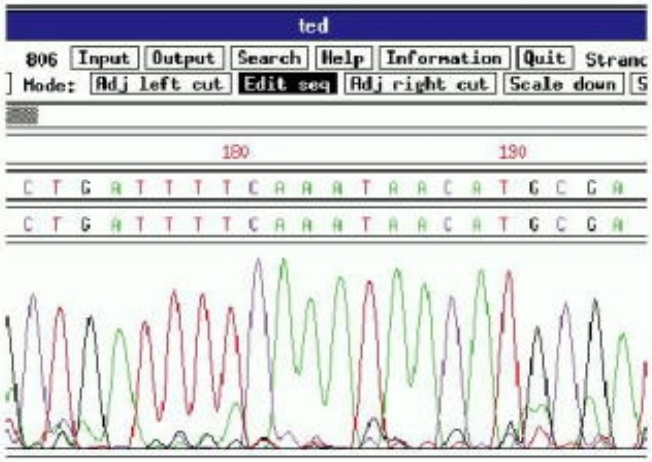
Central questions for today's lecture

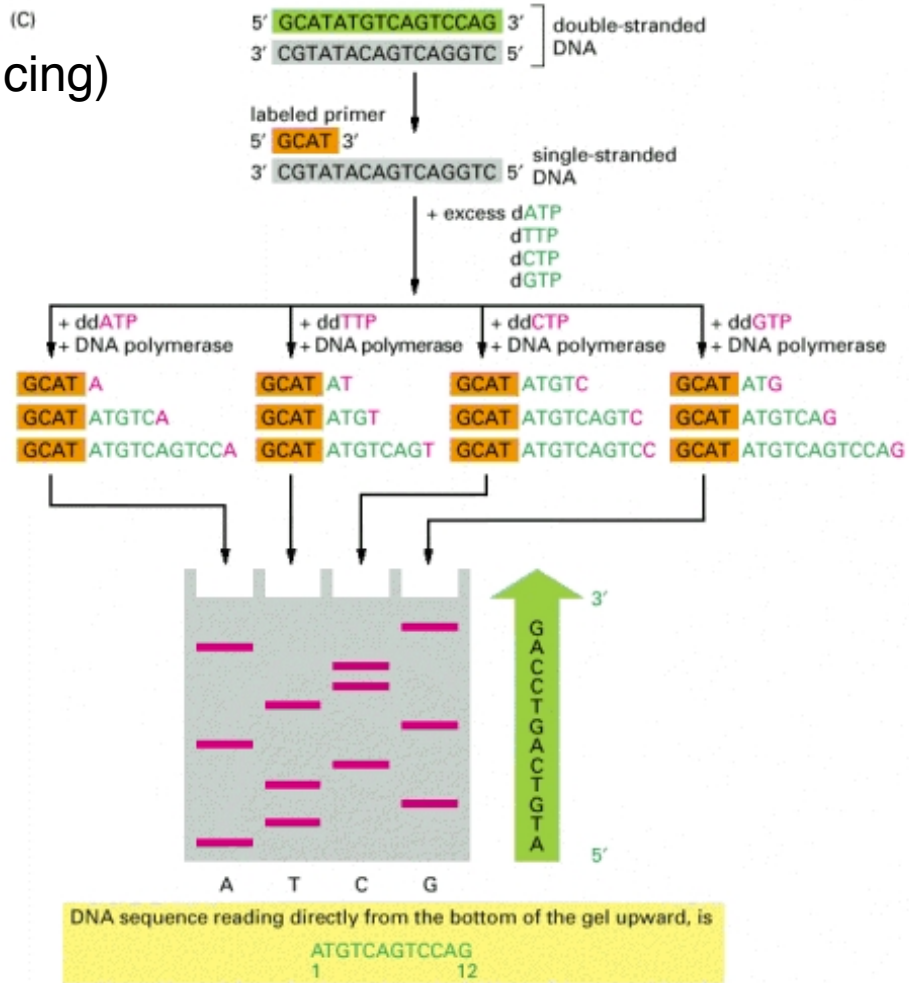
- Where do these data come from?
- How does the way we collect it influence what we know?

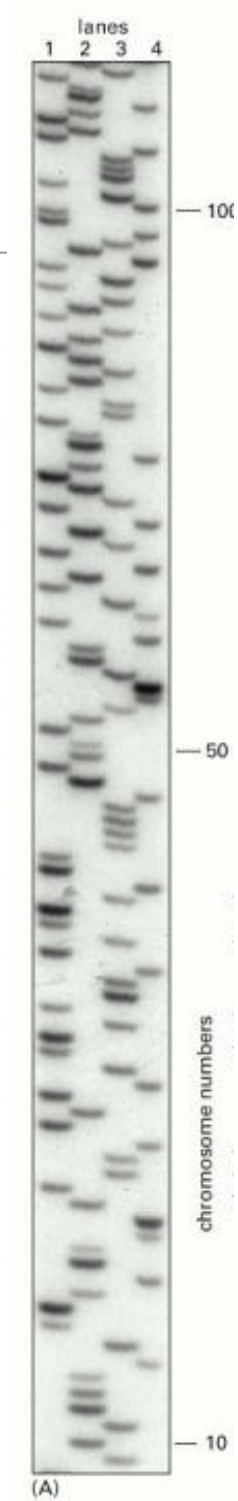
What is sequencing?

1. Yesterday (First generation sequencing)

    a. Maxam-Gilbert Sequencing
    b. Sangar Sequencing







Metrics for evaluating sequencing technology

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

What is sequencing?

1. Yesterday (First generation sequencing)

    a. Maxam-Gilbert Sequencing
    b. Sangar Sequencing

2. **Today (Second generation sequencing)**

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

3. Tomorrow (Third generation sequencing)

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

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

The steps of sequencing experiments

1. Sample preparation

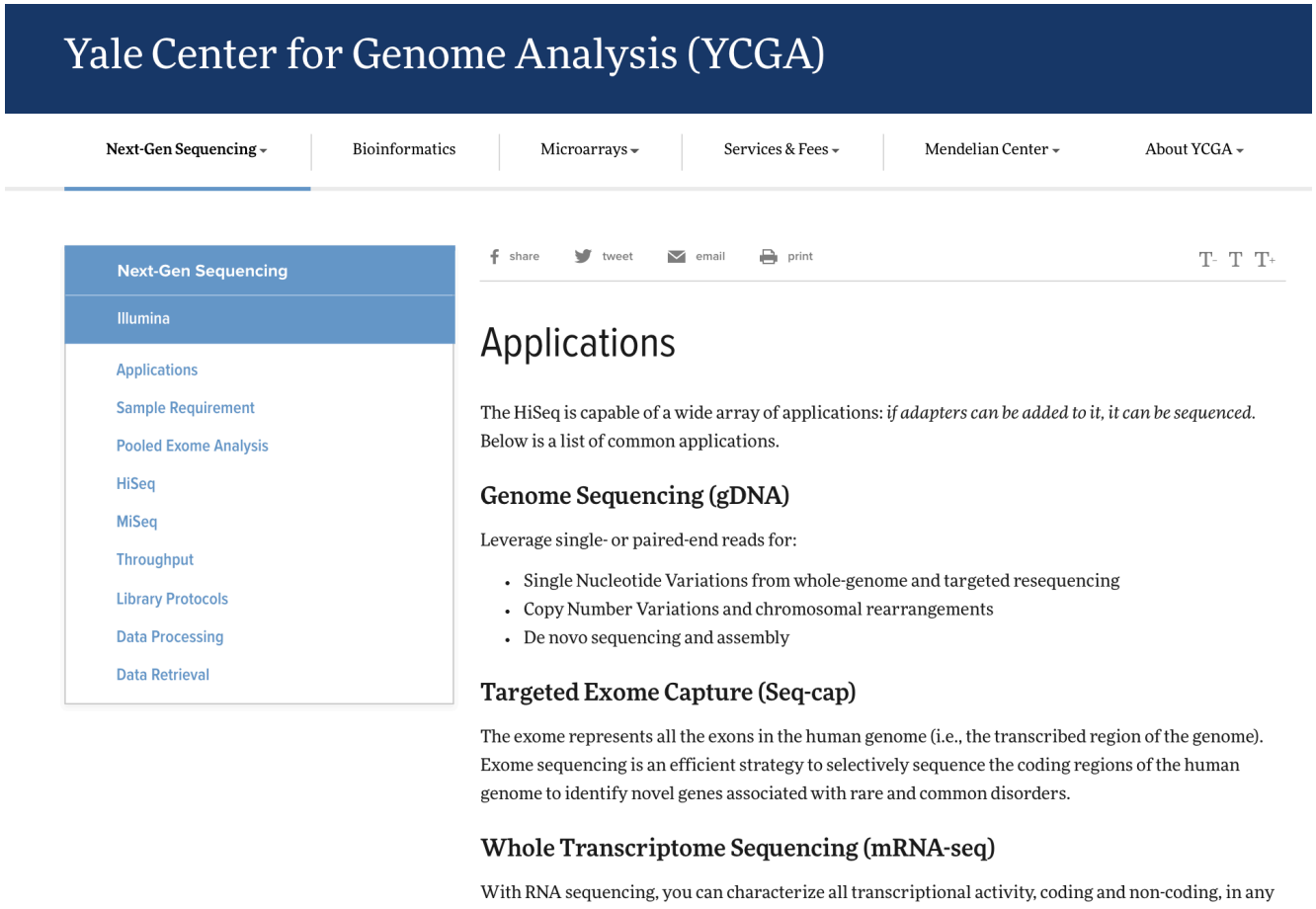
    a. Isolation
    b. Library construction

2. **Sequencing**

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

3. Data analysis

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



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

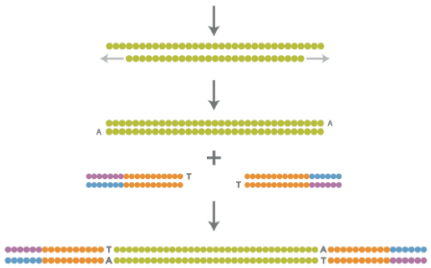
How long are the reads?

TATTGCAATATGTTAACAATCTAACAAGGAAAAAATACCCCACACAAAACAAAACACAACCCTTAGAACTGTGCTG

75 nt

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

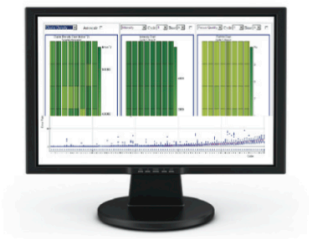
Where do these reads come from?



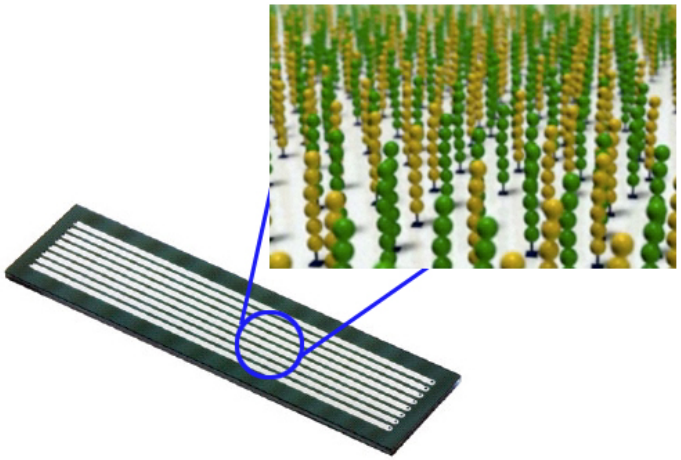
Library Preparation
~2 h [15 min hands-on (Nextera)]
< 6 h [< 3 h hands-on (TruSeq)]

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

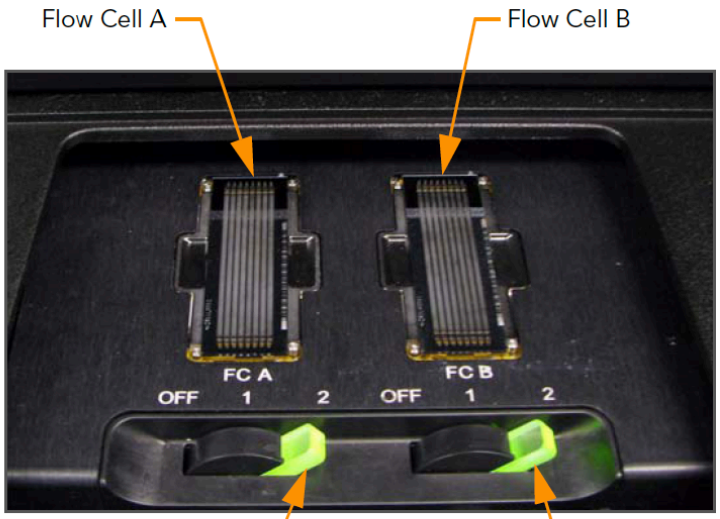
Sequencing by Synthesis
~1.5 to 11 days



CASAVA
2 days (30 min hands-on)



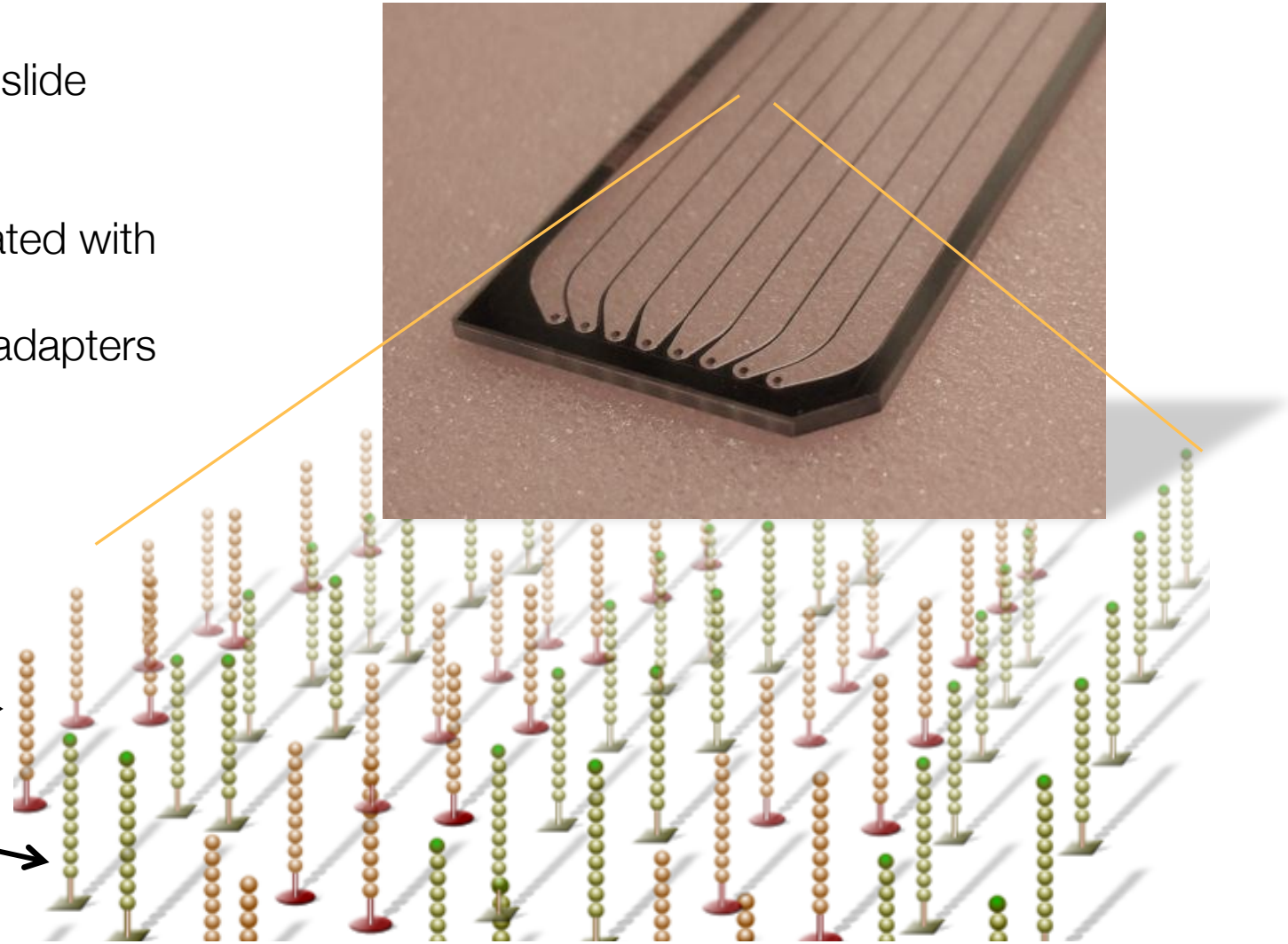
Flow cell



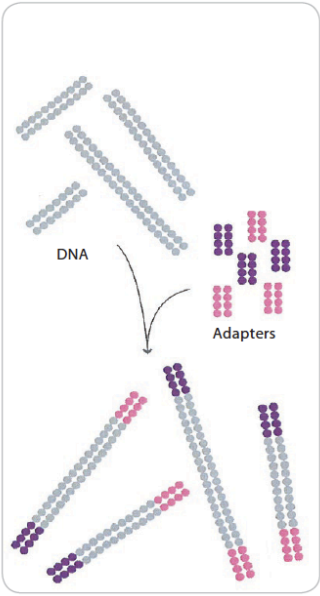
What is a flow cell?

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

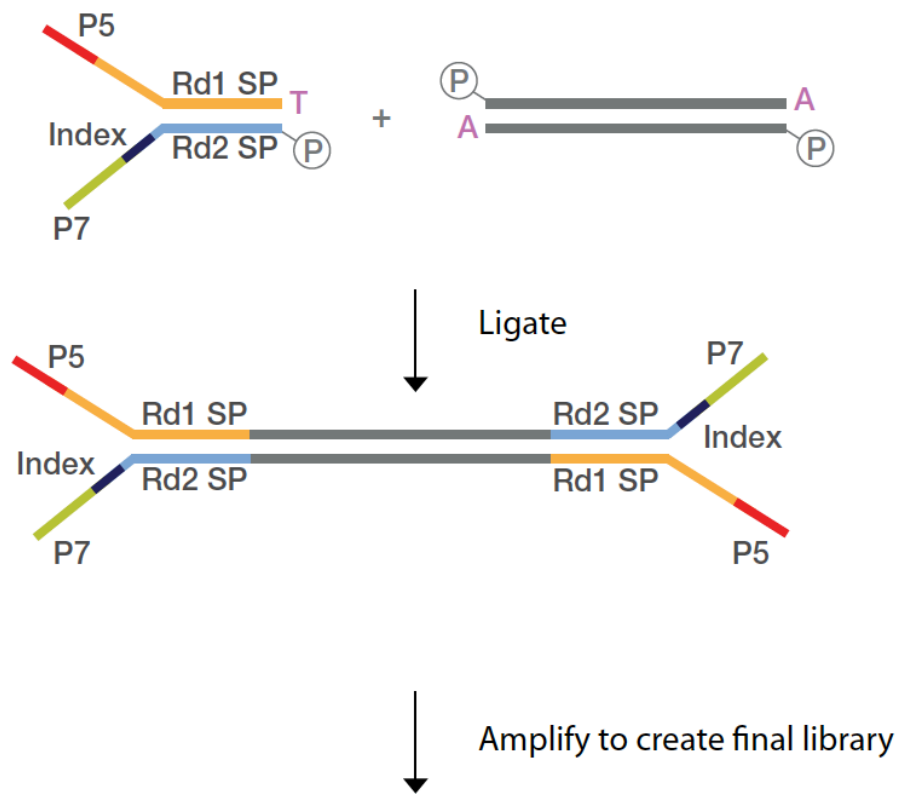


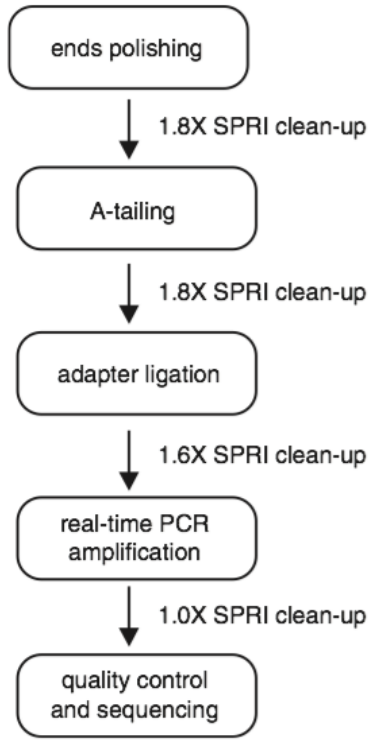
Cluster PCR
on flow cell
(8 lanes)

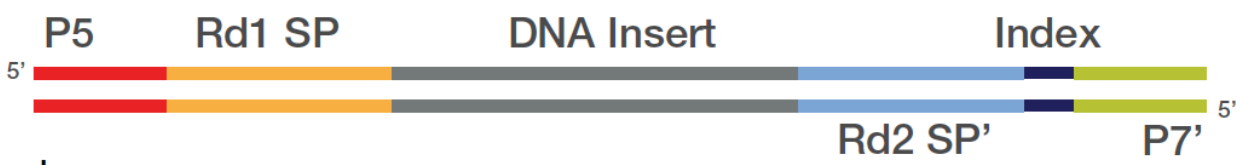


How do you make a sequencing library?

Index = unique sequence key to identify library







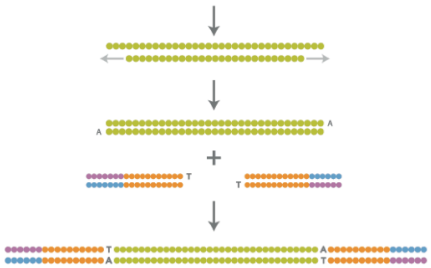
12 samples per lane

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).

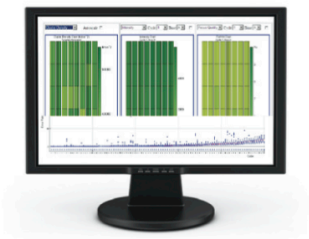
Where do these reads come from?



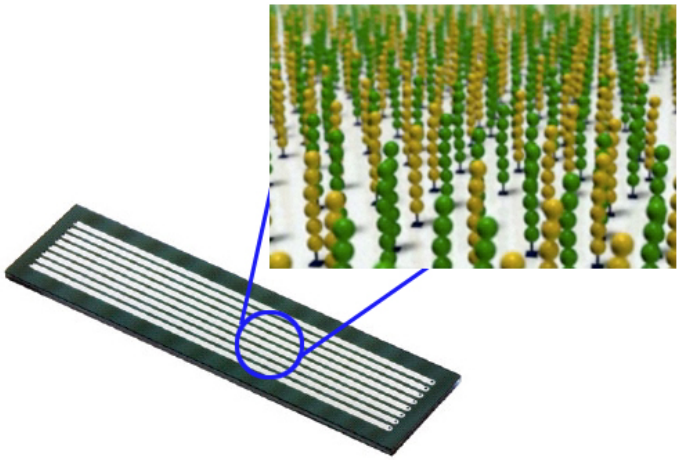
Library Preparation
~2 h [15 min hands-on (Nextera)]
< 6 h [< 3 h hands-on (TruSeq)]

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

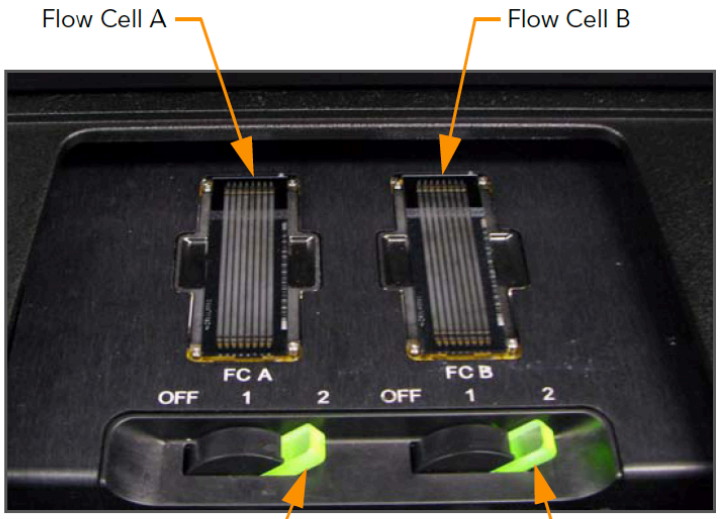
Sequencing by Synthesis
~1.5 to 11 days



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<
```

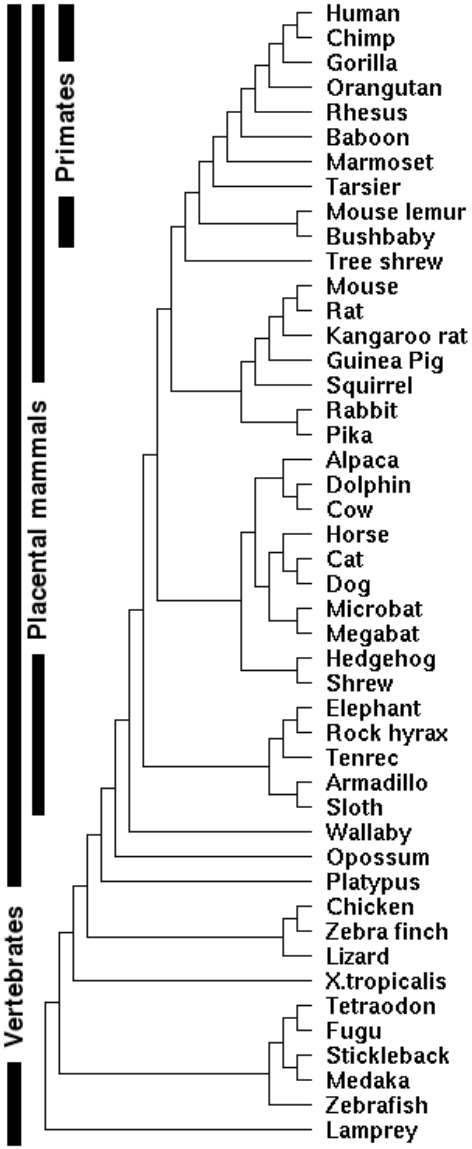
- For each single end read: Incomplete incorporation of bases.

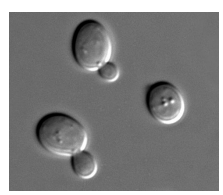
- For the size of the insert (especially for paired end analysis): Ability to get consistent clusters.

What do I do with my sequencing reads?

Source: Slate via Noonan

Many reference genomes are available



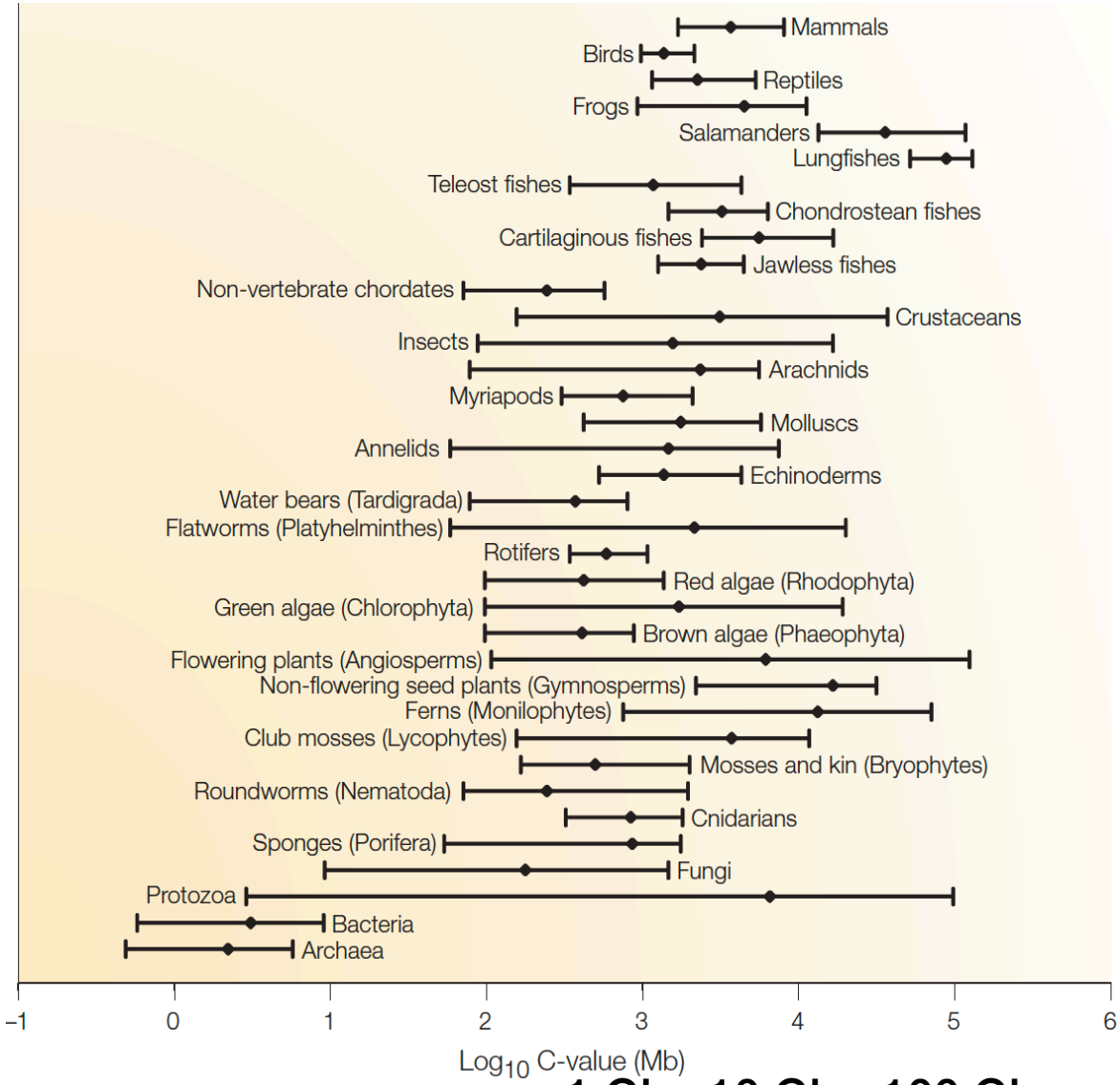


There is a wide range of genome sizes.

kb = 1000 bp
Mb = 1×10^6 bp
Gb = 1×10^9 bp
Tb = 1×10^{12} bp

Human haploid genome ~ 3 Gb

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



Sequencing of the human genome

Victory declared **2003**

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

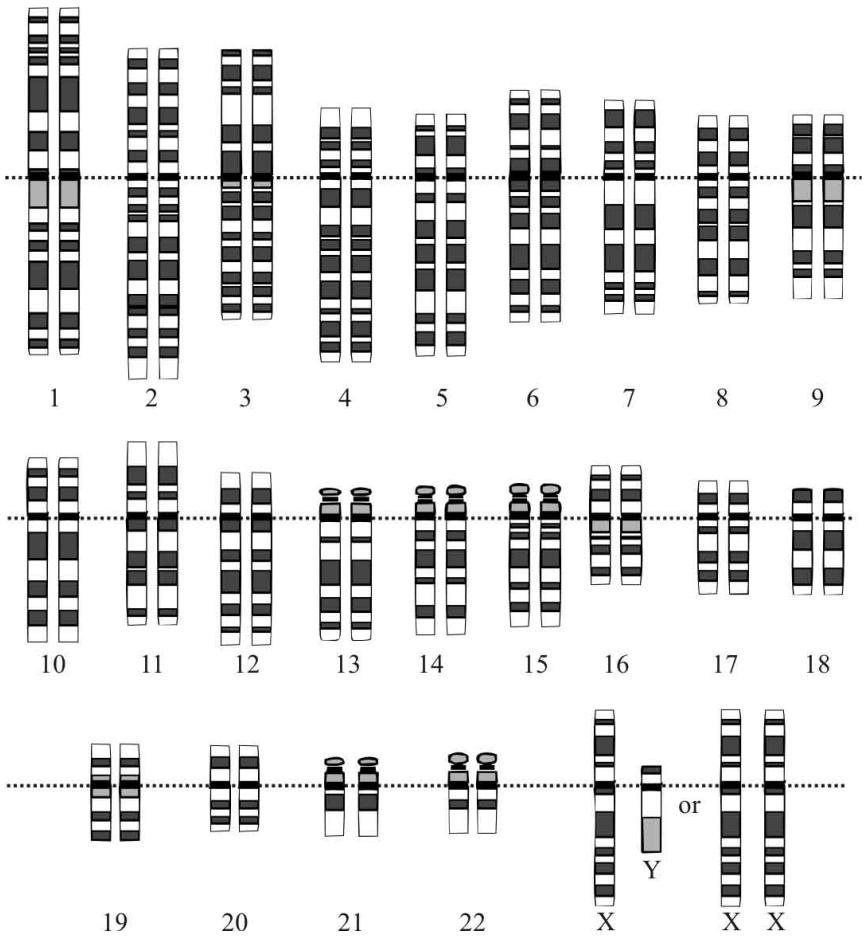
- Industrialization of Sanger sequencing, library construction, sample preparation, analysis, etc.
- $3 billion total cost
- 1 Gb/month at largest centers (2005)

Newest illumina sequencers claim 6000 Gb/run (2017)

Assembling a genome from short reads

>>10^9 sequencing reads

36 bp - 1 kb



3 Gb

How to assemble a genome

Generate reads

Find overlapping reads

Assemble reads into contigs

contig

Join contigs into scaffolds

scaffold

Scaffold_0: 12,865,123 – 12,965-110

Join scaffolds into "finished" sequence anchored on chromosomes

AGTTGTATTATTAGAAACTGAGGGCTAAAAACTGTGCACATACACAGACACACATATTATTTTAATATAGATTTTCAATAATTGGTCTAGGATAAGGATAATATACAG

Chr5: 133,876,119 – 134,876,119

Assembly quality criteria:

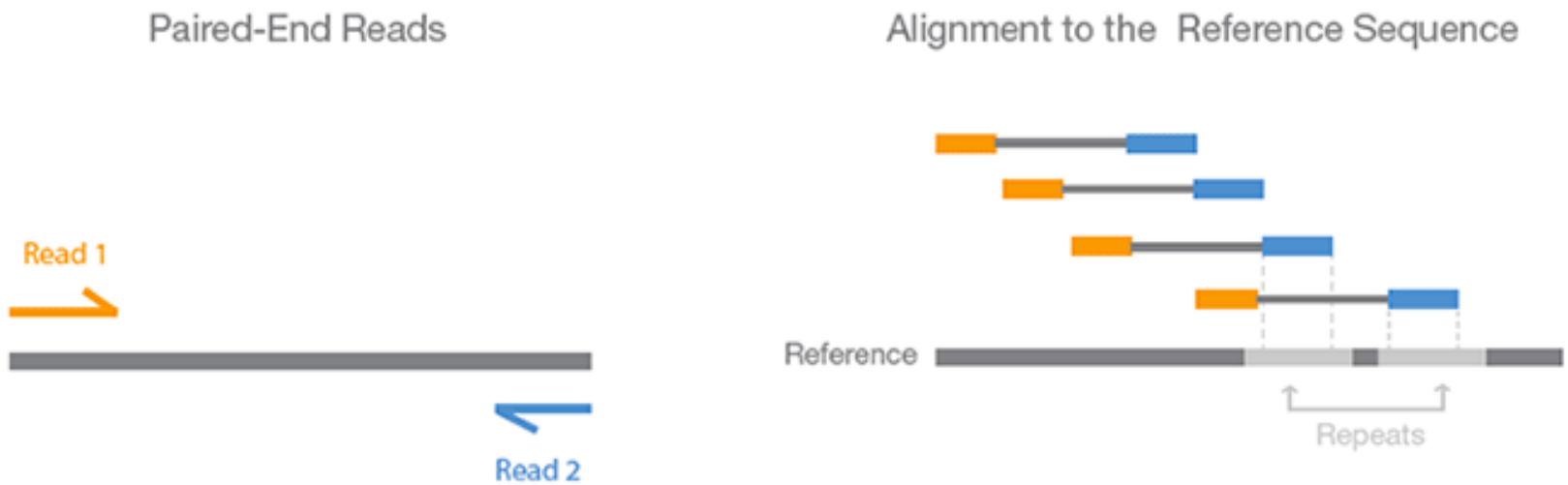
Accuracy: number of errors
(Human << 1/100,000 bp)

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

The importance of paired end reads



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

CCAAATCAAACAGTTGTATTATTAGAAACTGAGGGCTAAAAACTGTGCACATACACAGACACACATATTATTTTAATATAGATTTTCAATAATTGGTCTAGGATAAG
AAGCAAGAAGAAAACAAAGACTGTTACTATGGAAAAATGAAAATAGATTTTAAAACATGTTAATTCACGTTACTTTTTGTTAAATTTACTTTTCTTCTTTCACTTCTT
AATAAATCACATTAATTCCTTATCTCATGTGAAATTTCATATTTATGATTGATACCTTTAAATGTCATTTGTTGAAGGAAGATTATTCATTTTTTCATTCAATAAATATTT
CAGTATTATGTTCTAGGCATTGGGGATACCATGTTCACAAGACAGACTATGATTTACAGGATCAGATGTGGACTCTCAAATTCGACTGAGAATAAAACAGACACT
TAATTGATGCTAGAAAGACAATGAAACAGAGCCATGTGACCAATGAGAGAGATGAGGGTGGCAGCAGCCTGTTTTAGATAAGGTACCTGATTGGTGGGATTGG
TATGCCTTAATGATATGAAAGAACCATTCATGGGAAGGCCTAGCATTAAAAACCGTCTAGGCAGAATGAGCAGCAAGTGCAAGGGTCCTGGATAGGAATGAGCT
ATGGAAAAATGAAAATAGATTTTAAAACATGTTAATTCACGTTACTTTTTGTTAAATTTACTTTTCTTCTTTCACTTCTTACCTGTCAATGTTATTAATATTTTTAGGAA
GAAATTTCATATTTATGATTGATACCTTTAAATGTCATTTGTTGAAGGAAGATTATTCATTTTTTCATTCAATAAATATTTTTTAGAATAATAAGTCCCAGGCACAAGA
CATGTTCACAAGACAGACTATGATTTACAGGATCAGATGTGGACTCTCAAATTCGACTGAGAATAAAACAGACACTAAACAAGTAAATAAAGTTAATTTCAAGTTG
GATTTTAAAACATGTTAATTCACGTTACTTTTTGTTAAATTTACTTTTCTTCTTTCACTTCTTACCTGTCAATGTTATTAATATTTTTAGGAACAATAAATCACATTAA
ATTGATACCTTTAAATGTCATTTGTTGAAGGAAGATTATTCATTTTTTCATTCAATAAATATTTTTTAGAATAATAAGTCCCAGGCACAAGACCAGTATTATGTTCTAG
ACTATGATTTACAGGATCAGATGTGGACTCTCAAATTCGACTGAGAATAAAACAGACACAAACAAGTAAATAAAGTTAATTTCAAGTTGTAATTGATGCTATCCCA
TTGGGGATACCATTACCTGTCAATGTTATTAATATTTTTAGGAACAATAAATCACATTAATTCCAACATGCAAAGAGGAAATCTCCATATCATGCTTGTCATTCGTTT
GTGTGTAAAACATTCTCAGAATTTTAAACAATAACAAATCAGGGCTGAATGTGGCCAACATGCAAAGAGGAAATCTCCCATCTGTCCAAATCAAACAGTTGTATT
CATACACAGACACACATATTATTTTAATATAGATTTTCAATAATTGGTCTAGGATAAGGATAATATACAGAGAACATGCCAAAAGTTTAAGCAAGAAGAAAACAAAG
TAAAACATGTTAATTCACGTTACTTTTTGTTAAATTTACTTTTCTTCTTTCACTTCTTACCTGTCAATGTTATTAATATTTTTAGGAACAATAAATCACATTAATTCCT
ACCTTTAAATGTCATTTGTTGAAGGAAGATTATTCATTTTTTCATTCAATAAATATTTTTTAGAATAATAAGTCCCAGGCACAAGACCAGTATTATGTTCTAGGCAT
GATTTACAGGATCAGATGTGGACTCTCAAATTCGACTGAGAATAAAACAGACACTAAACAAGTAAATAAAGTTAATTTCAAGTTGTAATTGATGCTAGAAAGACA
AGATGAGGGTGGCAGCAGCCTGTTTTAGATAAGGTACCTGATTGGTGGGATTGGAAGACCTCTCTGAGATTAGTGTCTTCAGATATGCCTTAATGATATGAAAG
AACCGTCTAGGCAGAATGAGCAGCAAGTGCAAGGGTCCTGGATAGGAATGAGCTGGATATACTCAAGGAAGAAAGAGAAACTATGGAAAAATGAAAATAGATT
TAAATTTACTTTTCTTCTTTCACTTCTTACCTGTCAATGTTATTAATATTTTTAGGAACAATAAATCACATTAATTCCTTATCTCATGTGAAATTTCATATTTATGATTGA
TTATTCATTTTTTCATTCAATAAATATTTTTTAGAATAATAAGTCCCAGGCACAAGACCAGTATTATGTTCTAGGCATTGGGGATACCATGTTCACAAGACAGACTAT
ATTCGACTGAGAATAAAACAGACACTAAACAAGTAAATAAAGTTAATTTCAAGTTGTAATTGATGCTACTATGGAAAAATGAAAATAGATTTTAAAACATGTTAATTC
TTTCACTTCTTACCTGTCAATGTTATTAATATTTTTAGGAACAATAAATCACATTAATTCCTTATCTCATGTGAAATTTCATATTTATGATTGATACCTTTAAATGTCAT
CAATAAATATTTTTTAGAATAATAAGTCCCAGGCACAAGACCAGTATTATGTTCTAGGCATTGGGGATACCATGTTCACAAGACAGACTATGATTTACAGGATCAGA
AACAGACACAAACAAGTAAATAAAGTTAATTTCAAGTTGTAATTGATGCTATCCCAGGCACAAGACCAGTATTATGTTCTAGGCATTGGGGATACCATTACCTGTC
CACATTAATTCCAACATGCAAAGAGGAAATCTCCATATCATGCTTGTCATTCGTTTATCAGAGGCCAAATGTTTTTCTTTGTAAACGTGTGTAAAACATTCTCAGA
GTGGCCAACATGCAAAGAGGAAATCTCCCATCTGTCCAAATCAAACAGTTGTATTATTAGAAACTGAGGGCTAAAAACTGTGCACATACACAGACACACATATTA
GATAAGGATAATATACAGAGAACATGCCAAAAGTTTAAGCAAGAAGAAAACAAAGACTGTTACTATGGAAAAATGAAAATAGATTTTAAAACATGTTAATTCACGT
CTTCTTACCTGTCAATGTTATTAATATTTTTAGGAACAATAAATCACATTAATTCCTTATCTCATGTGAAATTTCATATTTATGATTGATACCTTTAAATGTCATTTGTT
ATATTTTTTAGAATAATAAGTCCCAGGCACAAGACCAGTATTATGTTCTAGGCATTGGGGATACCATGTTCACAAGACAGACTATGATTTACAGGATCAGATGTG
ACACTAAACAAGTAAATAAAGTTAATTTCAAGTTGTAATTGATGCTAGAAAGACAATGAAACAGAGCCATGTGACCAATGAGAGAGATGAGGGTGGCAGCAGCC
ATTGGAAGACCTCTCTGAGATTAGTGTCTTCAGATATGCCTTAATGATATGAAAGAACCATTCATGGGAAGGCCTAGCATTAAAAACCGTCTAGGCAGAATGAG
GAGCTGGATATACTCAAGGAAGAAAGAGAAACTATGGAAAAATGAAAATAGATTTTAAAACATGTTAATTCACGTTACTTTTTGTTAAATTTACTTTTCTTCTTTCAC
GGAACAATAAATCACATTAATTCCTTATCTCATGTGAAATTTCATATTTATGATTGATACCTTTAAATGTCATTTGTTGAAGGAAGATTATTCATTTTTTCATTCAATAA
AAGACCAGTATTATGTTCTAGGCATTGGGGATACCATGTTCACAAGACAGACTATGATTTACAGGATCAGATGTGGACTCTCAAATTCGACTGAGAATAAAACAG
AGTTGTAATTGATGCTACTATGGAAAAATGAAAATAGATTTTAAAACATGTTAATTCACGTTACTTTTTGTTAAATTTACTTTTCTTCTTTCACTTCTTACCTGTCAAT
ATTAATTCCTTATCTCATGTGAAATTTCATATTTATGATTGATACCTTTAAATGTCATTTGTTGAAGGAAGATTATTCATTTTTTCATTCAATAAATATTTTTTAGAATAA
TTCTAGGCATTGGGGATACCATGTTCACAAGACAGACTATGATTTACAGGATCAGATGTGGACTCTCAAATTCGACTGAGAATAAAACAGACACAAACAAGTAA
ATCCCAGGCACAAGACCAGTATTATGTTCTAGGCATTGGGGATACCATTACCTGTCAATGTTATTAATATTTTTAGGAACAATAAATCACATTAATTCCAACATGCA
TCGTTTATCAGAGGCCAAATGTTTTTCTTTGTAAACGTGTGTAAAACATTCTCAGAATTTTAAACAATAACAAATCAGGGCTGAATGTGGCCAACATGCAAAGAG
TGTATTATTAGAAACTGAGGGCTAAAAACTGTGCACATACACAGACACACATATTATTTTAATATAGATTTTCAATAATTGGTCTAGGATAAGGATAATATACAGAGA
CAAAGACTGTTACTATGGAAAAATGAAAATAGATTTTAAAACATGTTAATTCACGTTACTTTTTGTTAAATTTACTTTTCTTCTTTCACTTCTTACCTGTCAATGTTAT
TCCTTATCTCATGTGAAATTTCATATTTATGATTGATACCTTTAAATGTCATTTGTTGAAGGAAGATTATTCATTTTTTCATTCAATAAATATTTTTTAGAATAATAAGT

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

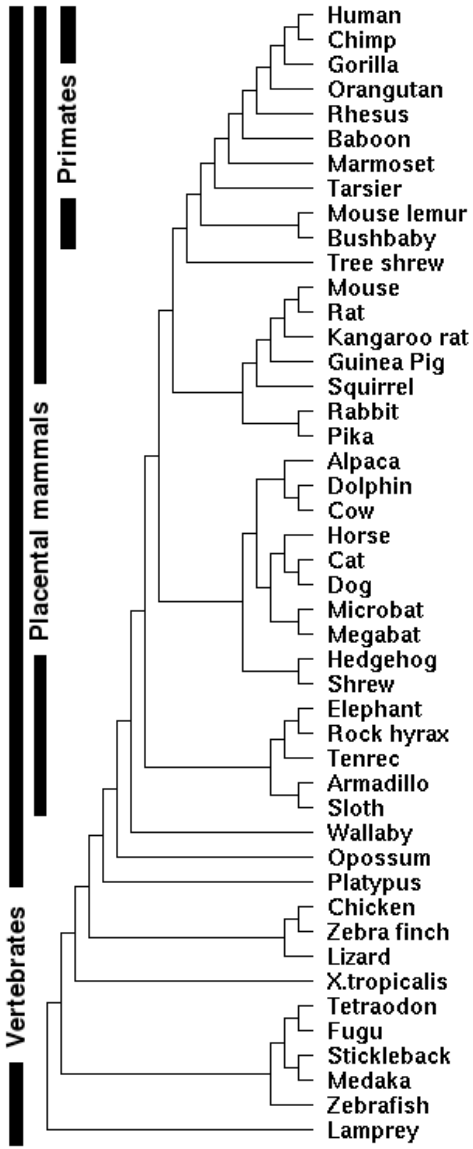
Regulatory sequences:

- 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

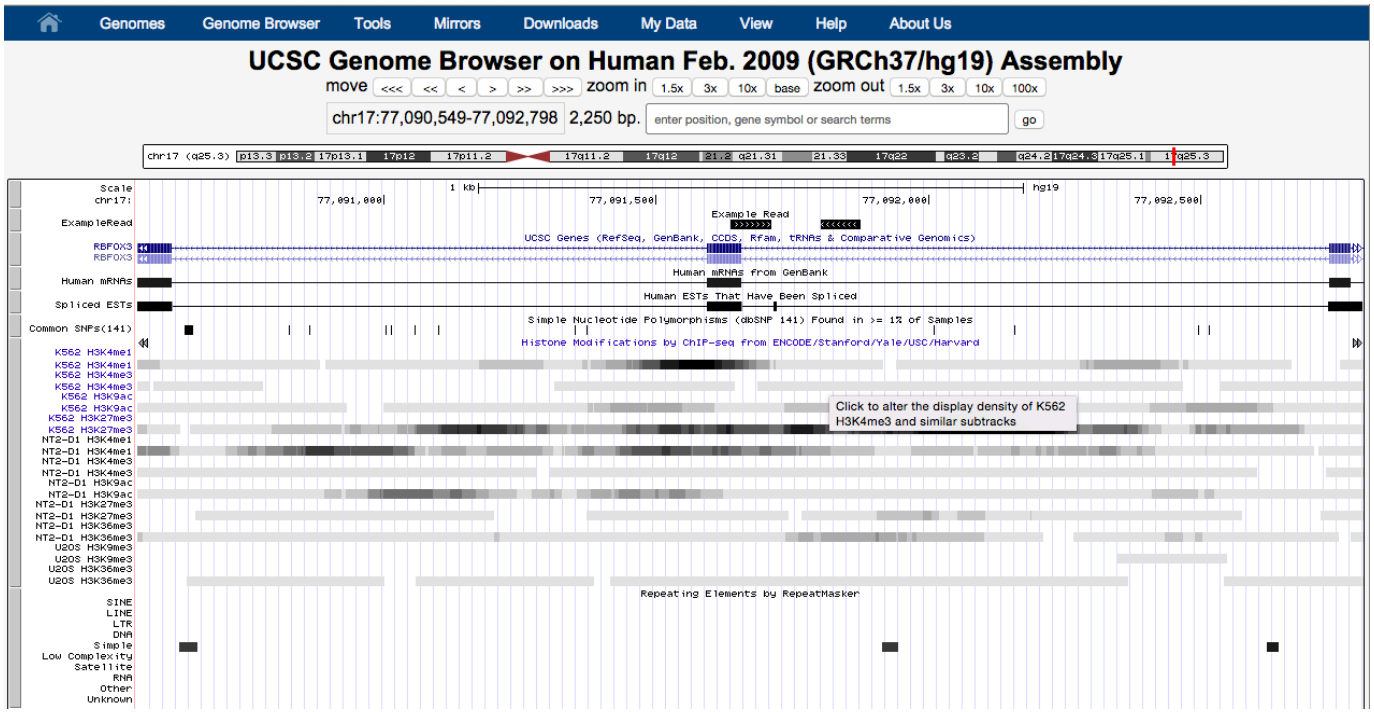
Example of a genome browser track

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



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