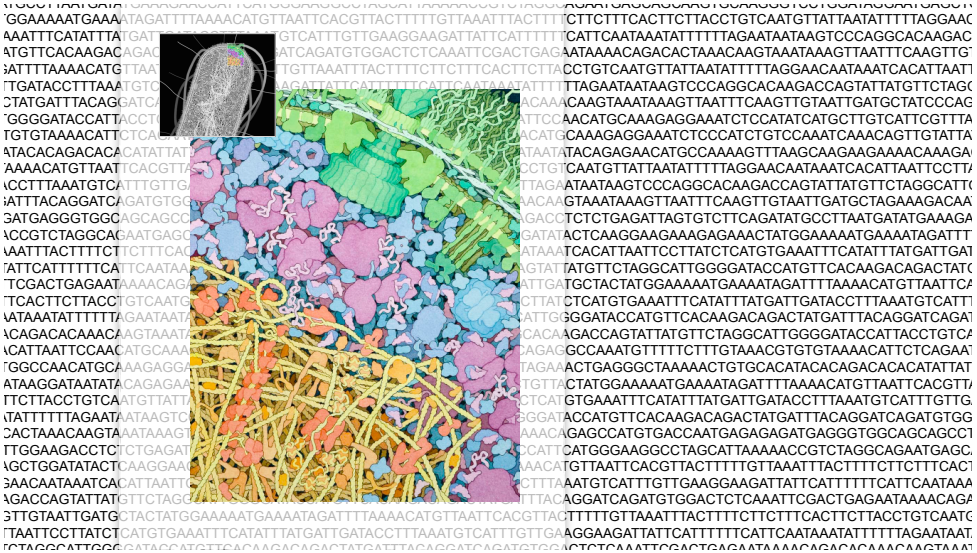
Genomics I

Biomedical Data Science: Mining and Modeling
CB&B 752 · MB&B 452
Matt Simon
Jan 31, 2022



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.

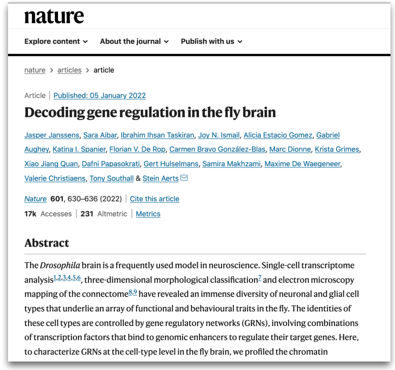
Genomics II: (Wednesday's lecture): Focus on applications of sequencing technology.

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


nature

Explore content ⌄ About the journal ⌄ Publish with us ⌄

nature > articles > article

Article | Published: 05 January 2022

Decoding gene regulation in the fly brain

Jasper Janssens, Sara Aibar, Ibrahim Ihsan Taskiran, Joy N. Ismail, Alicia Estacio Gomez, Gabriel Aughey, Katina I. Spanier, Florian V. De Rop, Carmen Bravo González-Blas, Marc Dionne, Krista Grimes, Xiao Jiang Quan, Dafni Papasokrati, Gert Hulselmans, Samira Makhzami, Maxime De Waegeneer, Valerie Christiaens, Tony Southall & Stein Aerts

Nature **601**, 630–636 (2022) | Cite this article

17k Accesses | **231** Altmetric | Metrics

Abstract

The *Drosophila* brain is a frequently used model in neuroscience. Single-cell transcriptome analysis[1,2,3,4,5,6], three-dimensional morphological classification[7] and electron microscopy mapping of the connectome[8,9] have revealed an immense diversity of neuronal and glial cell types that underlie an array of functional and behavioural traits in the fly. The identities of these cell types are controlled by gene regulatory networks (GRNs), involving combinations of transcription factors that bind to genomic enhancers to regulate their target genes. Here, to characterize GRNs at the cell-type level in the fly brain, we profiled the chromatin

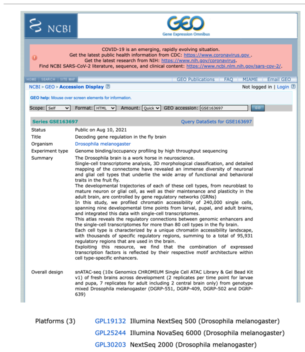
Methods: scRNA-seq, scATAC-seq, DamID, CUT&Tag, ChIP-seq

Data availability

The data generated for this study have been deposited in NCBI's Gene Expression Omnibus and are accessible through GEO Series accession numbers GSE163697 and GSE181494 (DGRP lines). We also provide a dedicated website to browse the results of the analyses and processed data (https://flybrain.aertslab.org), which provides link-outs to the SCope session (http://scope.aertslab.org/#/Fly_Brain/), UCSC hub (http://genome.ucsc.edu/cgi-bin/hgTracks?db=dm6&hubUrl=http://ucsctracks.aertslab.org/papers/FlyBrain/hub.txt), the eGRNs in NDEx, the DeepExplainer plots of enhancers and other information. The following online databases were used: FlyBase (https://flybase.org/), FlyMine (https://www.flymine.org/flymine), icis-Target (https://gbiomed.kuleuven.be/apps/lcb/i-cisTarget/), FlyLight (https://flweb.janelia.org/cgi-bin/flew.cgi), CIS-BP (http://cisbp.ccbr.utoronto.ca/), ENCODE (https://www.encodeproject.org/, ENCFF704WGH). The following publicly accessible datasets were also used: GSE107451 (scRNA-seq adult brain), GSE157202 (scRNA-seq larval brain), GSE101581 (scATAC-seq embryo). The neural network is from Özel et al[5].

https://www.nature.com/articles/s41586-021-04262-z.pdf

Raw data can be found in genomics databases


NCBI GEO Gene Expression Omnibus

COVID-19 is an emerging, rapidly evolving situation.
Get the latest public health information from CDC: https://www.coronavirus.gov .
Get the latest research from NIH: https://www.nih.gov/coronavirus.
Find NCBI SARS-CoV-2 literature, sequence, and clinical content: https://www.ncbi.nlm.nih.gov/sars-cov-2/.

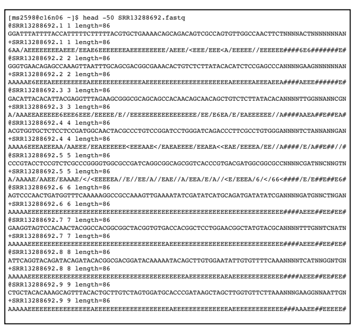
HOME SEARCH SITE MAP | GEO Publications | FAQ | MIAME | Email GEO

NCBI > GEO > Accession Display | Not logged in | Login

GEO help: Mouse over screen elements for information.

Scope: Self Format: HTML Amount: Quick GEO accession: GSE163697 GO

Series GSE163697	Query DataSets for GSE163697
Status	Public on Aug 10, 2021
Title	Decoding gene regulation in the fly brain
Organism	Drosophila melanogaster
Experiment type	Genome binding/occupancy profiling by high throughput sequencing
Summary	The Drosophila brain is a work horse in neuroscience. Single-cell transcriptome analysis, 3D morphological classification, and detailed mapping of the connectome have revealed an immense diversity of neuronal and glial cell types that underlie the wide array of functional and behavioral traits in the fruit fly. The developmental trajectories of each of these cell types, from neuroblast to mature neuron or glial cell, as well as their maintenance and plasticity in the adult brain, are controlled by gene regulatory networks (GRNs) In this study, we profiled chromatin accessibility of 240,000 single cells, spanning nine developmental time points from larval, pupal, and adult brains, and integrated this data with single-cell transcriptomes. This atlas reveals the regulatory connections between genomic enhancers and the single-cell transcriptomes for more than 80 cell types in the fly brain. Each cell type is characterized by a unique chromatin accessibility landscape, with thousands of specific regulatory regions, summing to a total of 95,931 regulatory regions that are used in the brain. Exploiting this resource, we find that the combination of expressed transcription factors is reflected by their respective motif architecture within cell type-specific enhancers.
Overall design	snATAC-seq (10x Genomics CHROMIUM Single Cell ATAC Library & Gel Bead Kit v1) of fresh brains across development (2 replicates per time point for larvae and pupa, 7 replicates for adult including 2 central brain only) from genotype mixed Drosophila melanogaster (DGRP-551, DGRP-409, DGRP-502 and DGRP-639)
Platforms (3)	GPL19132 Illumina NextSeq 500 (Drosophila melanogaster)
	GPL25244 Illumina NovaSeq 6000 (Drosophila melanogaster)
	GPL30203 NextSeq 2000 (Drosophila melanogaster)


```
[ms2598@c16n06 ~]$ head -50 SRR13288692.fastq
@SRR13288692.1 1 length=86
GGATTTATTTTACCATTTTTCTTTTTACGTGCTGAAAACAGCAGACAGTCGCCAGTGTTGGCCAACTTCTNNNNACTNNNNNNNAN
+SRR13288692.1 1 length=86
6AA/AEEEEEEEEAEEE/EEAE6EEEEEEEAEEEEEEEEE/AEEE/<EEE/EEE<A/EEEEE//EEEEEE####6E6#######E#
@SRR13288692.2 2 length=86
GGGTGAACAGAGCCAAAGTTAATTTGCAGCGACGGCGAAACACTGTCTCTTATACACATCTCCGAGCCCANNNNGAAGNNNNNNAN
+SRR13288692.2 2 length=86
AAAAAE6EEEAEEEEEEEEEAEEEEEEEEEEEEEAEEEEEEEEEEEEEEEEEEEEEEAEEEEEAEEEAEEA###AEEE######E#
@SRR13288692.3 3 length=86
GACATTACACATTACGAGGTTTAGAAGCGGGCGCAGCAGCCACAACAGCAACAGCTGTCTCTTATACACANNNNTTGGNNANNCGN
+SRR13288692.3 3 length=86
A/AAAEEAEEEEE6EEE6EEE/EEEEE/E//EEEEEEEEEEEEEEEE/EE/E6EA/E/EAEEEEEE//A###AAEA##E##EA#
@SRR13288692.4 4 length=86
ACGTGGTGCTCTCCTCCGATGGCAACTACGCCCTGTCCGGATCCTGGGATCAGACCCTTCGCCTGTGGGANNNNTCTANNANNGAN
+SRR13288692.4 4 length=86
AAAA6EEEAEEEAA/AAEEE/EEAEEEEEE<EEEAAE</EAEAEEEE/EEAEA<<EAE/EEEEA/EE//A####/E/A##E##//#
@SRR13288692.5 5 length=86
CCCGTACCTCCGTCTCGCCCGGGGTGGCGCCGATCAGGCGGCAGCGGTCACCCGTGACGATGGCGGCGCCNNNNCGATNNCNNGTN
+SRR13288692.5 5 length=86
A/AAAAE/AAEE/EAAAE/</<EEEEEA//E//EE/A//EAE//A/EEA/E/A//<E/EEEA/6/</66<####/E/E##E##E6#
@SRR13288692.6 6 length=86
AGTCCCAACTGATGGTTTCAAAAAGGCCGCCAAAGTTGAAAATATCGATATCATGCAGATGATATATCGANNNNGATGNNCTNGAN
+SRR13288692.6 6 length=86
AAAAAEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEE###AEEE##EE#EE#
@SRR13288692.7 7 length=86
GAAGGTAGTCCACAACTACGGCCACGGCGGCTACGGTGTGACCACGGCTCCTGGAACGGCTATGTACGCANNNNTTTGNNTCNATN
+SRR13288692.7 7 length=86
AAAAAEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEAEEEEEEEEEEEEEEEEEEEEEEEEEEEEEE###AEEE##EE#EE#
@SRR13288692.8 8 length=86
ATTCAGGTACAGATACAGATACACGGCGACGGATACAAAAATACAGCTTGTGGAATATTGTGTTTTCAAANNNNTCATNNGGNTGN
+SRR13288692.8 8 length=86
AAAAAEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEAEEEEEEEEEAEAEEEEEEEEEEEE###AEEE##EE#EE#
@SRR13288692.9 9 length=86
CTGCTACACAAAGCAGTTTACACTGCTTGTCTAGTGGATGCACCCGATAAGCTAGCTTGGTGTTCTTAAANNNGAAGGNNAATTGN
+SRR13288692.9 9 length=86
AAAAAEEEEEEEEEEEEEEEEEAEEEEEEEEEEEEEEEEEEEEEAEEEEEEEEEEEEEEEAEEEEEEEEEEEE###AAAEE##EEEEE#
```

- 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.

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

Central questions

Where do these data come from?

How does the way we collect it influence what we know?

Which best describes your wet lab experience?

(A) I have never conducted research that requires molecular biology.

(B) I've done a lot of molecular biology (cloning, etc.) but only worked with Sanger sequencing.

(C) I've generated at least one deep sequencing data set.

(D) I've done a lot of deep sequencing.

Which best describes your experience with analysis of sequencing data?

(A) I have no relevant experience with DNA sequencing data.

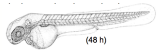
(B) I've read/thought about DNA sequencing data but never worked with it.

(C) I've worked with some DNA sequencing data.

(D) I've worked with a lot of DNA sequencing data.

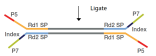
Workflow

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

Quality
- Accuracy per base

What is sequencing?

One-at-a-time methods

a. Maxam-Gilbert Sequencing
b. Sangar Sequencing

Short read deep sequencing

a. Illumina Sequencing
b. Ion Torrent

Long read deep sequencing

a. **Nanopore** based
b. **Pacific Bioscience** Sequencing

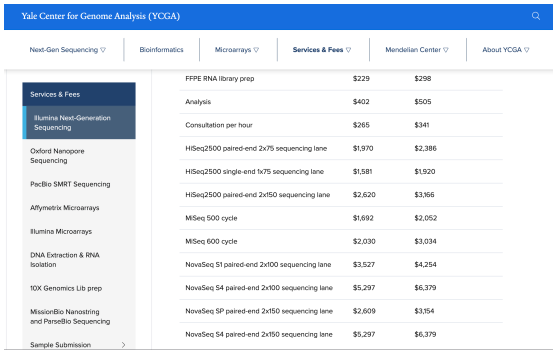
Sequencing technology	Platform	Data type	Read length (kb)		Read accuracy (%)	Throughput per flow cell (Gb)		Estimated cost per Gb (US$)	Maximum throughput per year (Gb)[a]
			N50	Maximum		Mean	Maximum		
Pacific Biosciences (PacBio)	RS II[b]	CLR	5–15	>60	87–92	0.75–1.5	2	333–933[c]	4,380
	Sequel	CLR	25–50	>100		5–10	20	98–195[d]	17,520
	Sequel II	CLR	30–60	>200		50–100	160	13–26[e]	93,440
		HiFi	10–20	>20	>99	15–30	35	43–86[e]	10,220
Oxford Nanopore Technologies (ONT)	MinION/GridION	Long	10–60	>1,000	87–98	2–20	30	50–500[f]	21,900 (MinION) 109,500 (GridION)
		Ultra-long	100–200	>1,500		0.5–2	2.5	500–2,000[f]	913 (MinION) 4,563 (GridION)
	PromethION	Long	10–60	>1,000		50–100	180	21–42[f]	3,153,600
Illumina	NextSeq 550	Single-end	0.075–0.15	0.15	>99.9	16–30	>30	50–63[g]	>47,782
		Paired-end	0.075–0.15 (×2)	0.15 (×2)		32–120	>120	40–60[g]	>70,080
	NovaSeq 6000	Single-end	0.05–0.25	0.25		65–3,000	>3,000	10–35[h]	>1,194,545
		Paired-end	0.05–0.25 (×2)	0.25 (×2)					

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

Logsdon (**2020**) *Nat Rev Genetics*

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

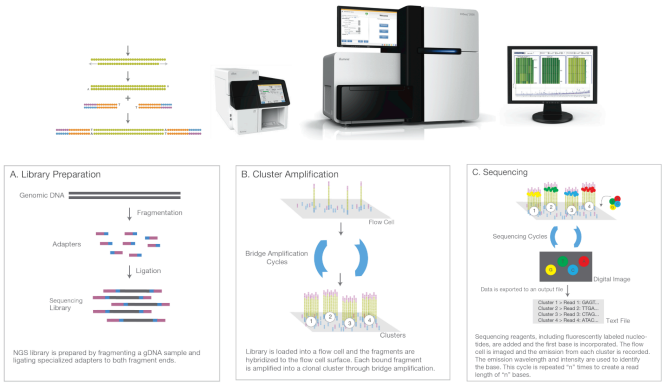


Retrieved Jan 30, 2022:
https://medicine.yale.edu/keck/ycga/services/illuminaprices/

What is the most raw form of data recorded in an Illumina sequencing experiment?

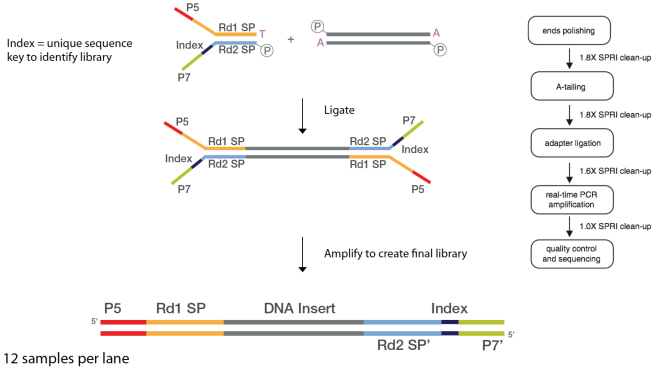
(A) A chromatogram.
(B) A string of letters.
(C) A series of images.
(D) A readout of genomic locations.

Where do these reads come from?



https://www.illumina.com/content/dam/illumina-marketing/documents/products/illumina_sequencing_introduction.pdf

Optional: Library preparation using ligation



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).

Optional: Library preparation using tagmentation

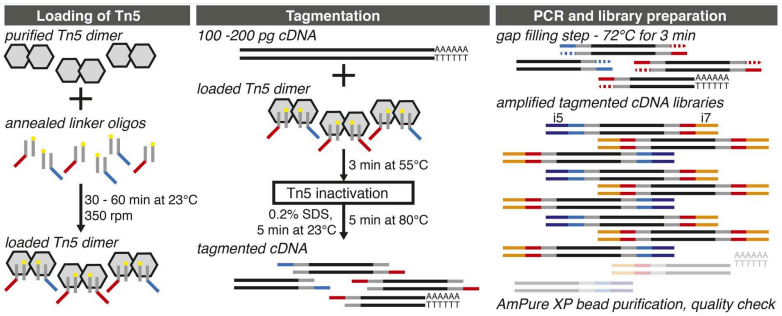
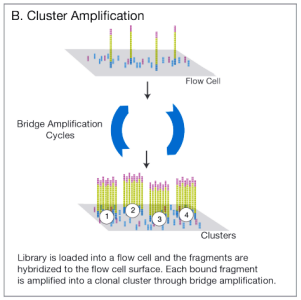
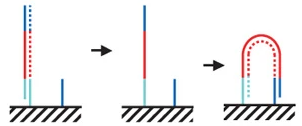


Figure from: Hennig BP, Velten L, Racke I, Tu CS, Thoms M, Rybin V, Besir H, Remans K, Steinmetz LM. G3 (Bethesda). 2018 Jan 4;8(1):79-89. PMID: 29118030.

Cluster amplification.

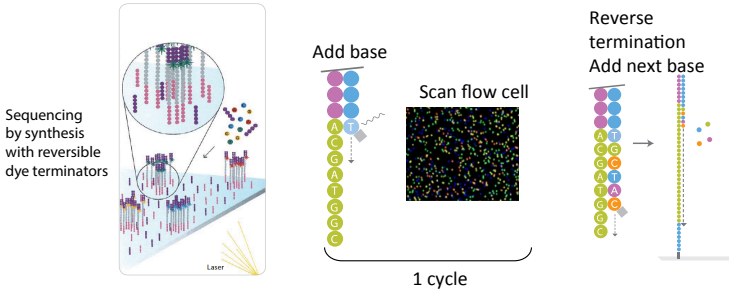




- Separate each individual molecule (randomly).
- Give each molecule an address (spatial location).
- Pack as many on as possible but avoid overlaps.

https://www.illumina.com/content/dam/illumina-marketing/documents/products/illumina_sequencing_introduction.pdf

Sequencing by synthesis



https://www.illumina.com/content/dam/illumina-marketing/documents/products/illumina_sequencing_introduction.pdf

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

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
+
#11BBDDDFFDFBFFFIIIIIIIIIIIIIIIFEGIIIIIFIGAGIIFIII=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 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=GIHGIIIHGIJIHEHIHHGFFFFEEEDDDDDDDDDDD

@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 is the output from an Illumina sequencing experiment?

Many reads…

```
@HWI-D00306:498:HBB89ADXX:1:1101:1180:1882 1:N:0:CGATGT
NCATCACTTTCTGCACCAGCCATGACGTCAATCTTCGTCCGAACCCCAAACTCGAGATCGGAAGAGCACACGTCTG
+
#11BBDDDFFDFBFFFIIIIIIIIIIIIIIIFEGIIIIIFIGAGIIFIII=FEEEEEFFFDDD=@9A@BBBBB=?BB<

@HWI-D00306:498:HBB89ADXX:1:1101:1167:1902 1:N:0:CGATGT
TATTGCAATATGTTAACAATCTAACAAGGAAAAAATACCCCACACAAAACAAAACACAACCCTTAGAACTGTGCTG
+
B@@FFDFFHFHHHJJIJIGIIJJJJJJIJJHFIJJJJJIJJJEHHJJIJJJJIIJJJJJJGHHHHFBDFFFE>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
+
CCCFFFFFHHHHHHHIJJIIJJJIIGHHIJGGJIGIJJJJJJJIJIIIJJJIIJJJJIIJGJJHCHFBDFFFDDECB
```

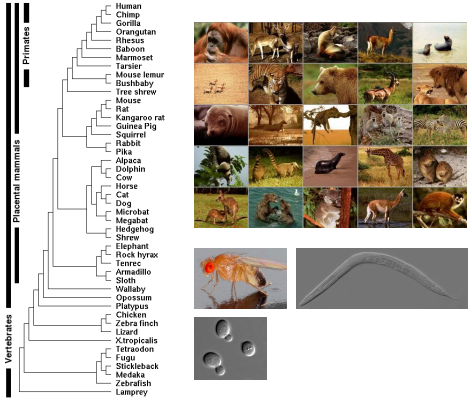
Generally ~ 2,000,000,000 reads/sequencing lane

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

What do I do with my sequencing reads?

Source: Slate via Noonan

Many reference genomes are available



A 75 nt sequencing read matches to a reference genome perfectly, except for one mismatch. What might have caused this?

(A) An error introduced during library preparation.

(B) An error in a base call during sequencing.

(C) A single nucleotide difference between the genome of the biological sample and the reference genome.

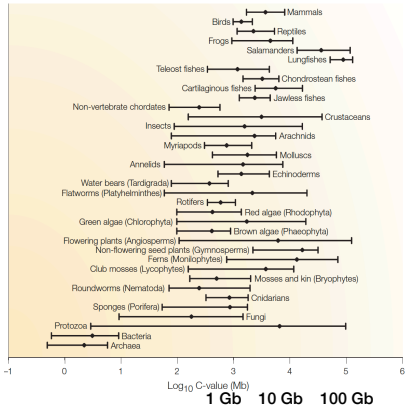
(D) Any of the above.

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

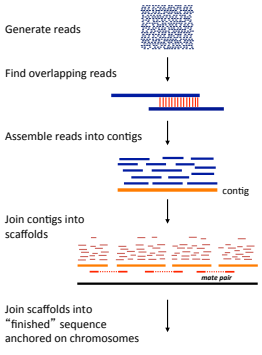
Victory declared **2003**

HGSC
National Human Genome Research Institute
JGI DOE JOINT GENOME INSTITUTE
welcome 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 2 billion reads 2x150 bp. $5000 -> <$100/genome.

How to assemble a genome



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

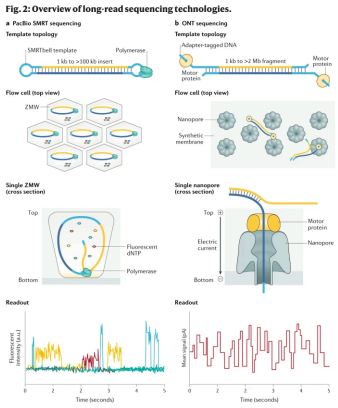
Contiguity: number of gaps (Human: est. 357)

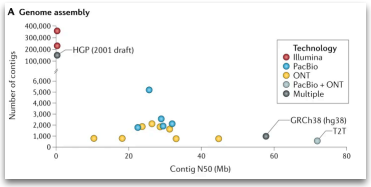
Coverage:
Average number of reads representing a particular position in the assembly
Human, Mouse, Rat: > 20x
Chimpanzee: ~6x
Squirrel: ~2x

scaffold Scaffold_0: 12,865,123 – 12,965-110

Chr5: 133,876,119 – 134,876,119

The importance of long read sequencing





Logsdon (**2020**) *Nat Rev Genetics*

What types of annotation do we have/want?

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

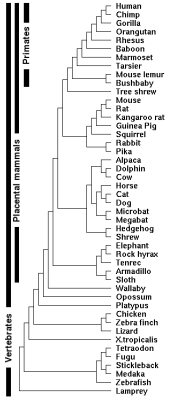
Genetic variation:
- SNPs and CNVs

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

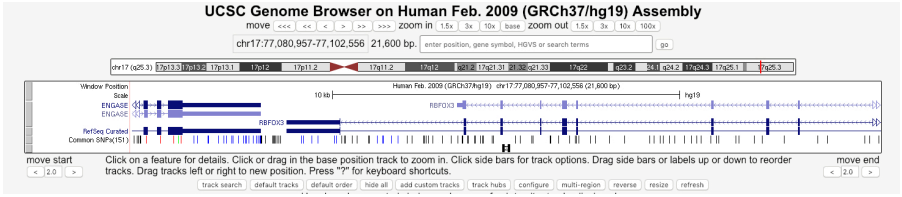
Example of a genome browser track (UCSC)

Chr5: 133,876,119 – 134,876,119

Our specific example:

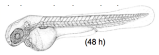
```
@HWI-ST1239:178:H0KPNADXX:2:1101:3120:1979 1:N:0:TGACCA
NCTGTAGGCTGCGTAGCCTCCCTGCAGGGTAAGTGGGAGGAGAGAGAGCAGAGGGACTTAGTGGGGCTCCCCAGGG
+
#1=DDFFFHHHHHIJIJJJIJJJJJIJJJ?FHIDGIJ=GIHGIIIHGIJIHEHIHHGFFFFEEEDDDDDDDDDDDD

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



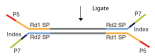
Workflow

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.

Applications of sequencing technology next week.

Conclusions

- Sequencing technology is central to our understanding of biology.
- The decrease in cost and increase in throughput make sequencing data increasingly ubiquitous.

✴ Sophisticated use of data from genomics requires an integrated understanding of the biological experiment, sample preparation and down stream computational analyses of the data.