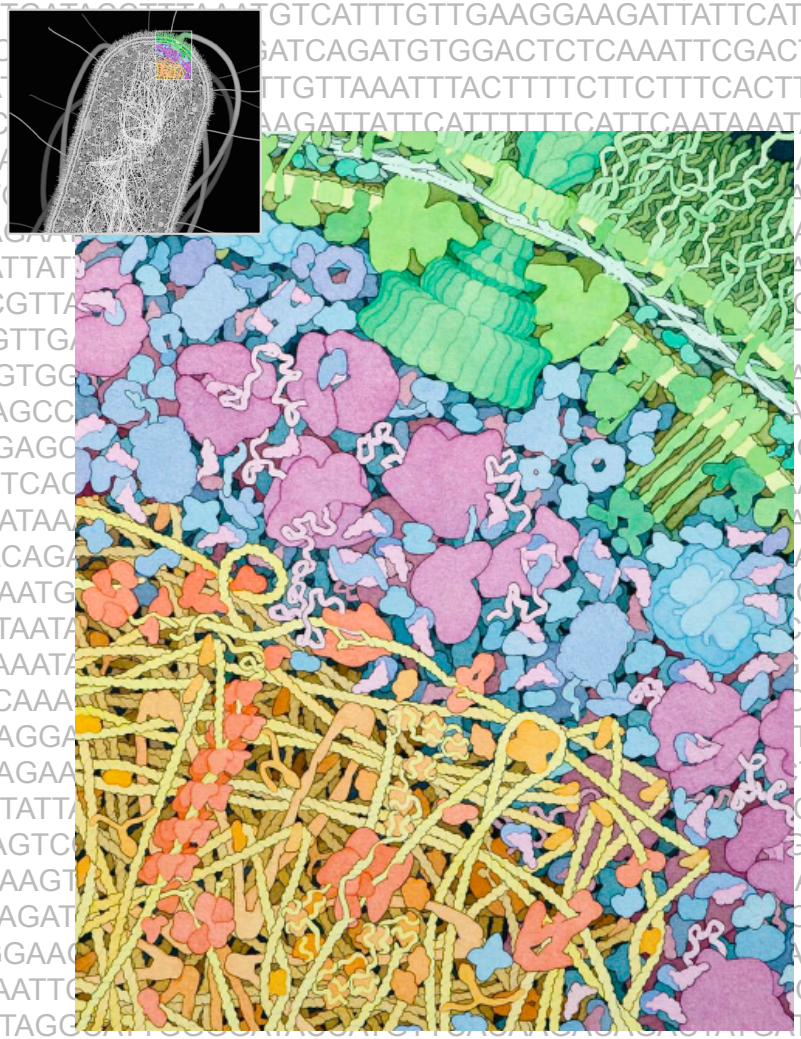
Genomics I

Biomedical Data Science: Mining and Modeling
CB&B 752 • MB&B 452
Matt Simon
Jan 27, 2025



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

Molecular Cell Supports

Meet the author

View all

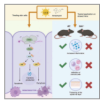
Most read (last 30 days)

Article • Open Access

The ribotoxic stress response drives acute inflammation, cell death, and epidermal thickening in UV-irradiated skin *in vivo*

Vind et al.

Published online: November 25, 2024

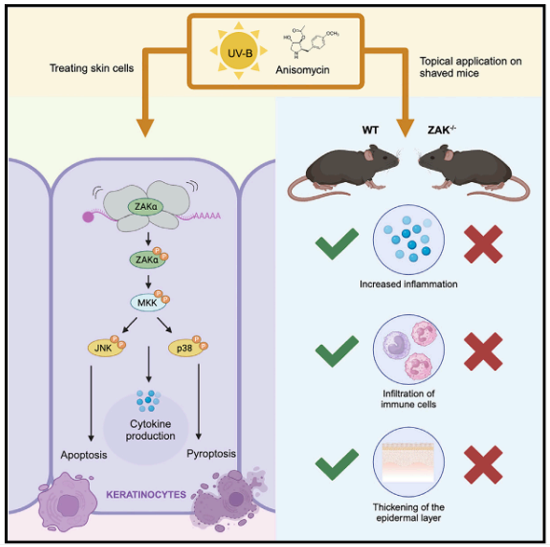


Molecular Cell

Article

The ribotoxic stress response drives acute inflammation, cell death, and epidermal thickening in UV-irradiated skin *in vivo*

Graphical abstract



Authors

Anna Constance Vind, Zhenzhen Wu, Muhammad Jasrie Firdaus, ..., Mads Gyrd-Hansen, Franklin L. Zhong, Simon Bekker-Jensen

Correspondence

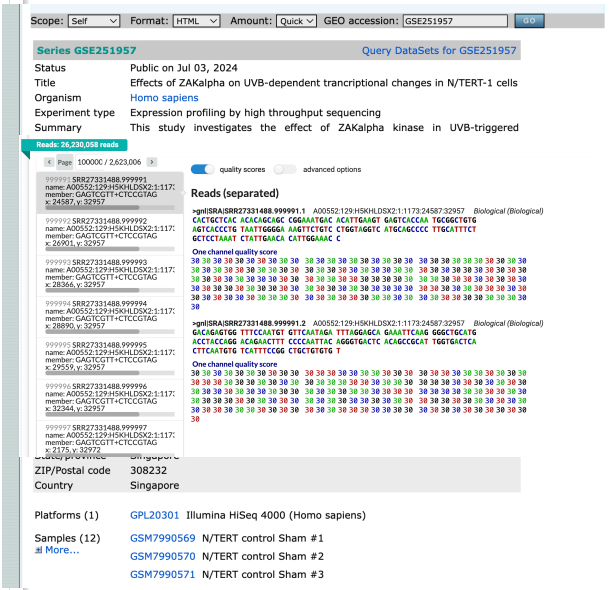
vind@sund.ku.dk (A.C.V.), franklin.zhong@ntu.edu.sg (F.L.Z.), sbj@sund.ku.dk (S.B.-J.)

In brief

The acute skin reaction to sunburn, encompassing keratinocyte cell death, inflammation, and epidermal thickening, has traditionally been ascribed to DNA damage responses. Using a mouse model deficient for the *Zak* gene, Vind et al. demonstrate that these reactions depend on sensing of cytoplasmic RNA damage rather than nuclear DNA damage.

Data and code availability

- RNA sequencing raw data files have been deposited in NCBI's Gene Expression Omnibus (GEO) archive, with the accession code GEO: GSE251957.



https://www.cell.com/molecular-cell/fulltext/S1097-2765(24)00884-0

Raw data can be found in genomics databases

Experiment attributes:

GEO Accession: GSM4802676

Links:

NCBI link: NCBI Entrez (gds)

Runs: 1 run, 43.5M spots, 3.3G bases, 1.1Gb

Run	# of Spots	# of Bases	Size	Published
SRR15101033	43,549,074	3.3G	1.1Gb	2022-11-10

```
[ms2598@c16n06 ~]$ head -50 SRR13288692.fastq
@SRR13288692.1 1 length=86
GGATTTATTTTACCATTTTTCTTTTTACGTGCTGAAAACAGCAGACAGTCGCCAGTGTTGGCCAACTTCTNNNNACTNNNNNNNAN
+SRR13288692.1 1 length=86
6AA/AEEEEEEEAEEE/EEAE6EEEEEEEAEEEEEEEEE/AEEE/<EEE/EEE<A/EEEEE//EEEEEE####6E6#######E#
@SRR13288692.2 2 length=86
GGGTGAACAGAGCCAAAGTTAATTTGCAGCGACGGCGAAACACTGTCTCTTATACACATCTCCGAGCCCANNNNGAAGNNNNNNAN
+SRR13288692.2 2 length=86
AAAAAE6EEEAEEEEEEEEEAEEEEEEEEEEEEEAEEEEEEEEEEEEEEEEEEEEEAEEEEEAEEEAEEA####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
AAAAAEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEAEEEEEEEEEEEEEEEEEEEEEEEEEE####AEEE##EE#EE#
@SRR13288692.8 8 length=86
ATTCAGGTACAGATACAGATACACGGCGACGGATACAAAAATACAGCTTGTGGAATATTGTGTTTTCAAANNNNTCATNNGGNTGN
+SRR13288692.8 8 length=86
AAAAAEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEAEEEEEEEEAEAEEEEEEEEEEEE####AEEE##EE#EE#
@SRR13288692.9 9 length=86
CTGCTACACAAAGCAGTTTACACTGCTTGTCTAGTGGATGCACCCGATAAGCTAGCTTGGTGTTCTTAAANNNGAAGGNNAATTGN
+SRR13288692.9 9 length=86
AAAAAEEEEEEEEEEEEEEEAEEEEEEEEEEEEEEEEEEEAEEEEEEEEEEEEEAEEEEEEEEEEEE###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.

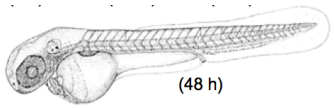
(B) I've read/thought about DNA sequencing data but never worked with it.

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

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

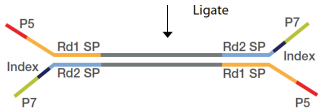
Workflow

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

Spatial -omics

a. **DNA probe** based
b. **Antibody probe** based

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*
Bressan (**2023**) *Science*

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

Sequencing

Service	Yale Rate	Non-Yale Rate
MiSeq 500 Cycle	$1,905	$2,485
NextSeq Usage	$1,039	$1,358
NovaSeq X Plus 25B 2x150	$2,936	$3,825
NovaSeq SP 2x150	$2,617	$3,410

…per 1.25 billion reads!

Retrieved Jan 27, 2025:
https://medicine.yale.edu/genetics/research/ycga/service-fees/

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

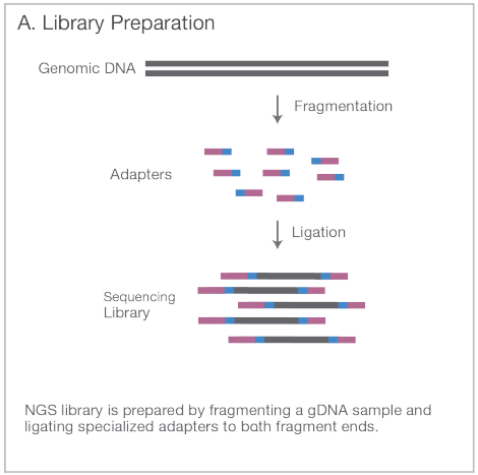
(A) A chromatogram.

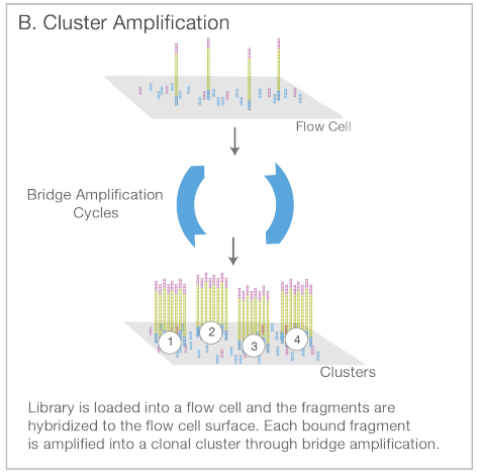
(B) A string of letters.

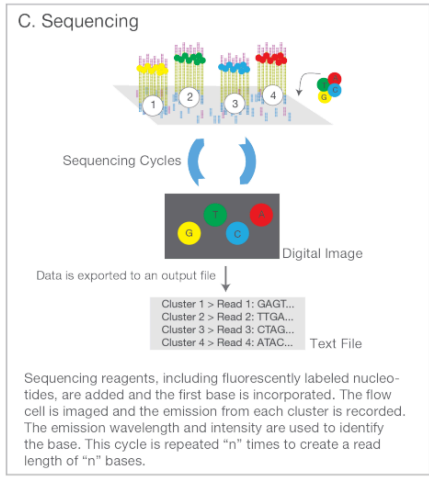
(C) A series of images.

(D) A readout of genomic locations.

Where do these reads come from?



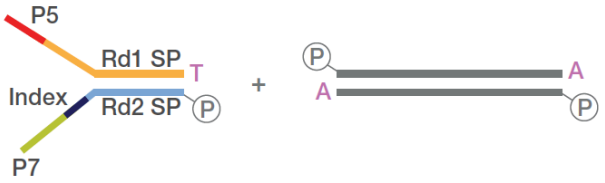


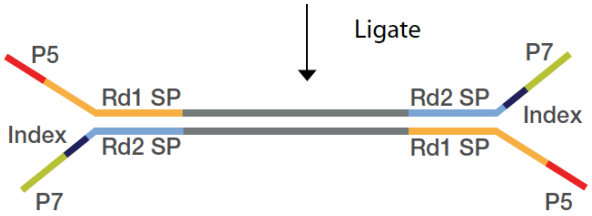


https://www.illumina.com/content/dam/illumina-marketing/documents/products/illumina_sequencing_introduction.pdf

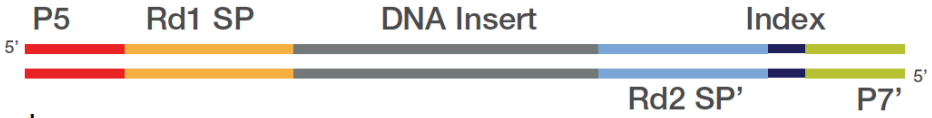
Optional: Library preparation using ligation

Index = unique sequence key to identify library

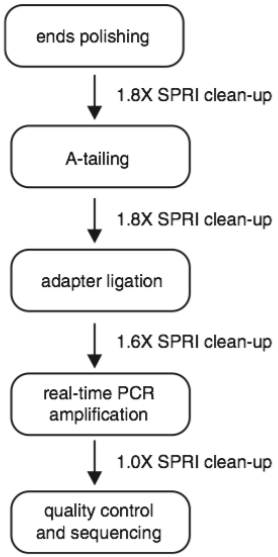




Amplify to create final library



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).

Optional: Library preparation using tagmentation

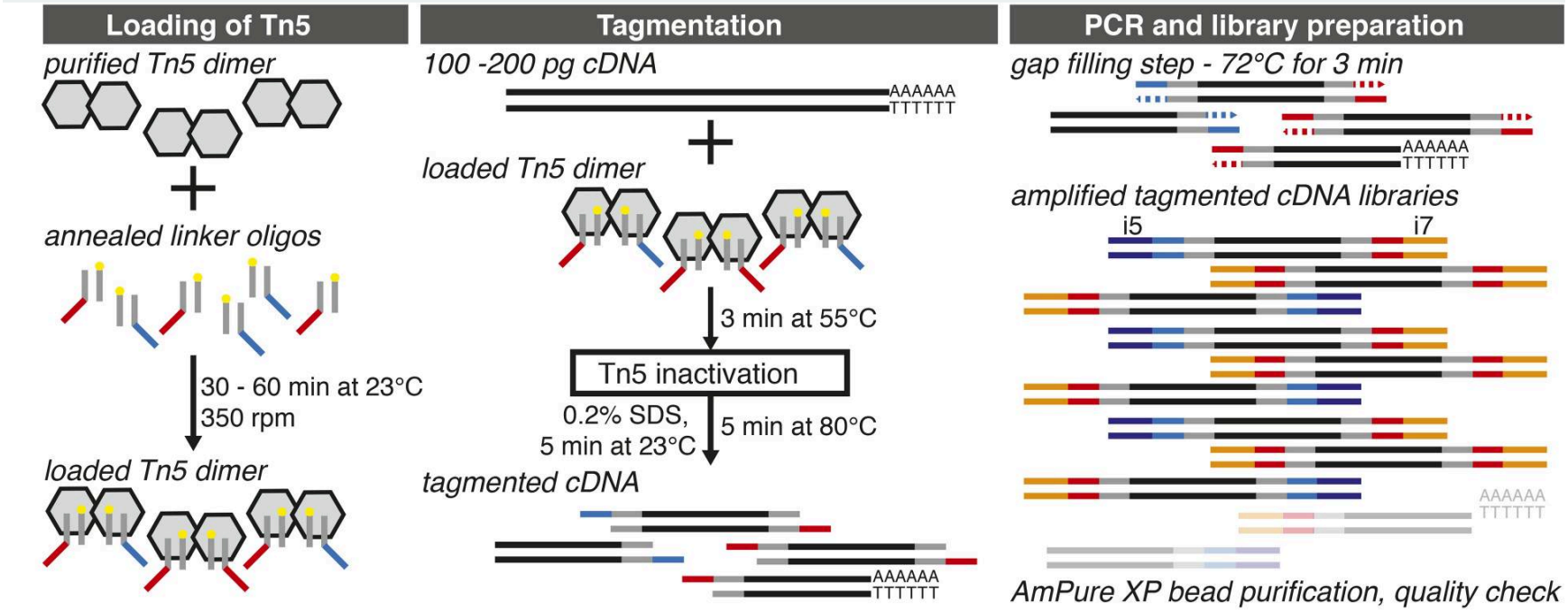
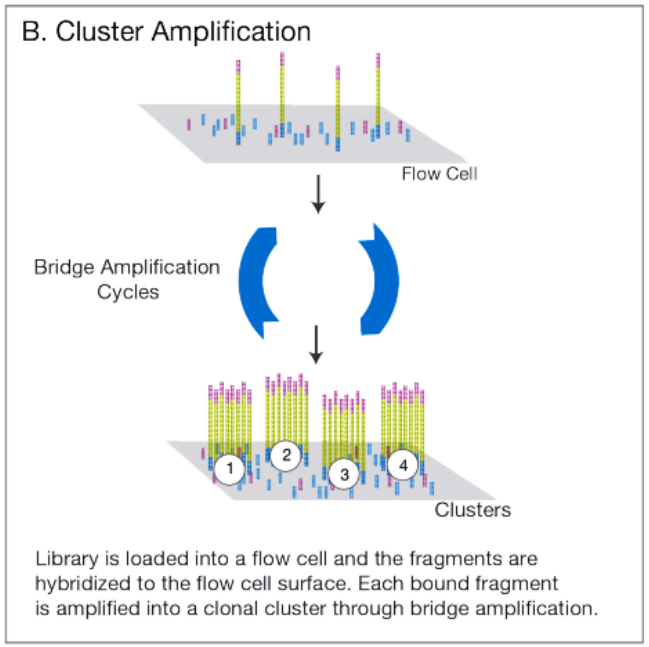
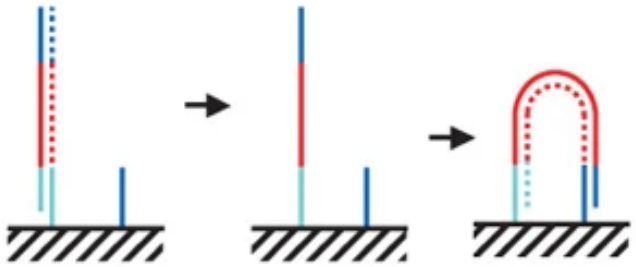
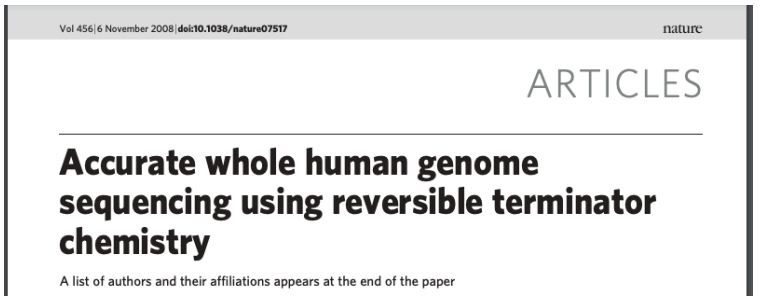


Figure from: Hennig BP, Velten L, Racke I, Tu CS, Thoms M, Rybin V, Besir H, Remans K, Steinmetz LM. G3 (Bethesda). 2018 Jan 4;8(1):79-89. PMID: 29118030.

Cluster amplification.





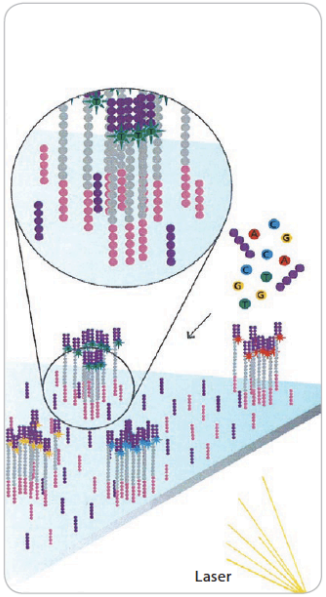


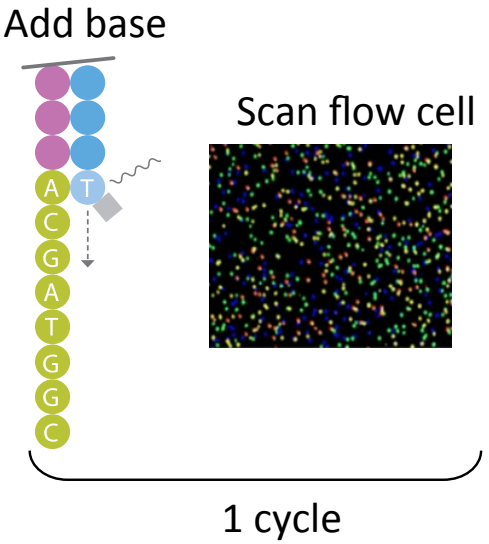
- Separate each individual molecule (randomly).
- Give each molecule an address (spatial location).
- Pack as many on as possible but avoid overlaps.

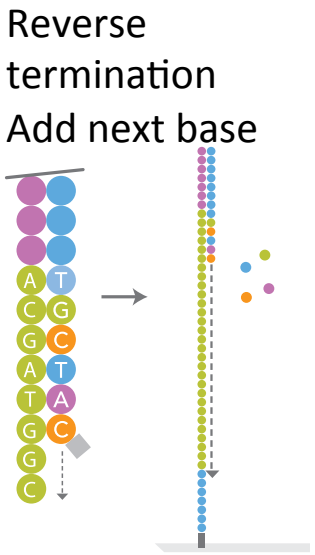
https://www.illumina.com/content/dam/illumina-marketing/documents/products/illumina_sequencing_introduction.pdf

Sequencing by synthesis

Sequencing by synthesis with reversible dye terminators

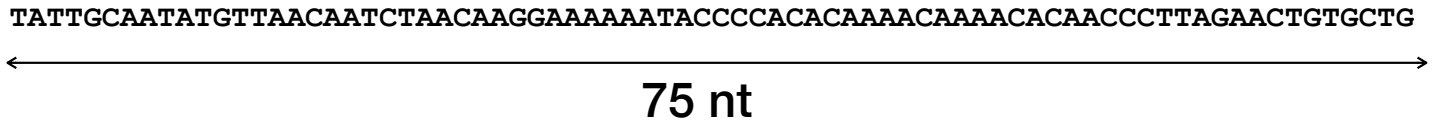






https://www.illumina.com/content/dam/illumina-marketing/documents/products/illumina_sequencing_introduction.pdf

How long are the reads?



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
+
#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 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
+
#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
```

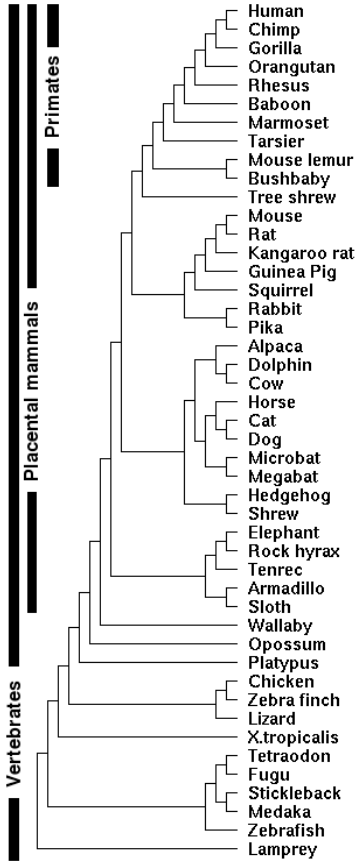
Generally ~ 2,000,000,000 reads/sequencing lane

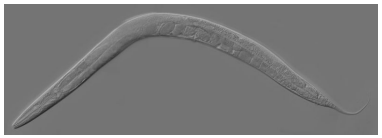
Note: This is for an Illumina NovaSeq with current chemistry, but this number changes

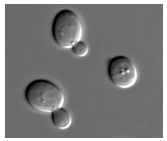
What do I do with my sequencing reads?

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.

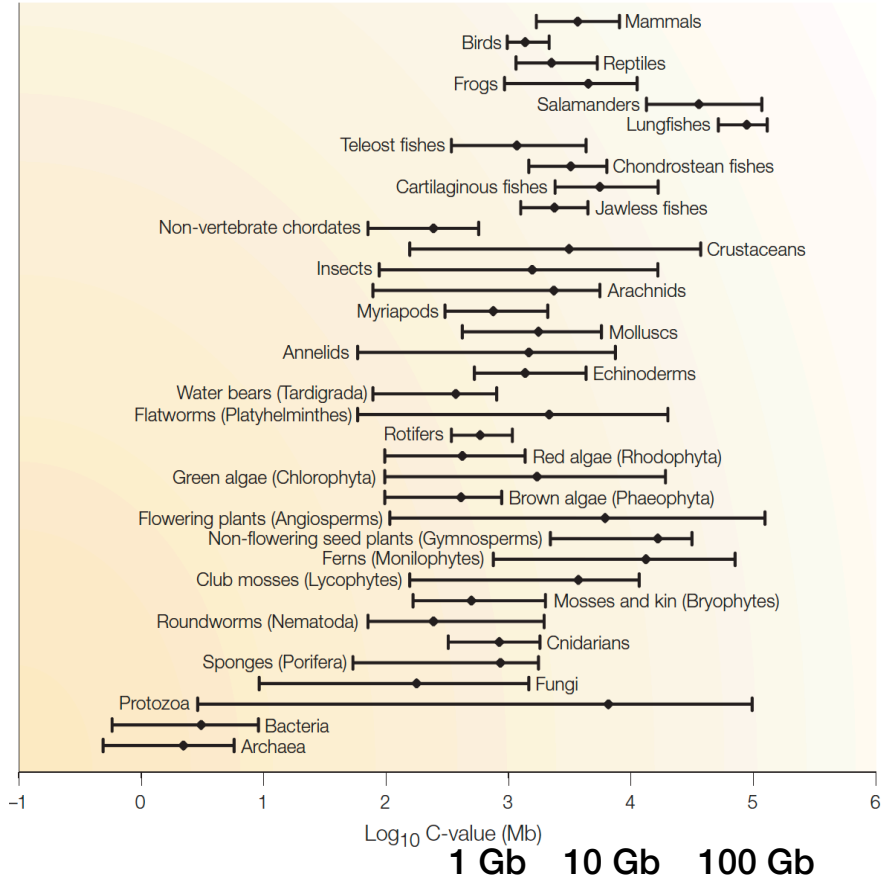
(D) Any of the above.

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**

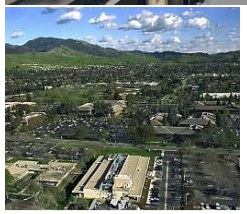
HGSC HUMAN GENOME SEQUENCING CENTER

National Human Genome Research Institute

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

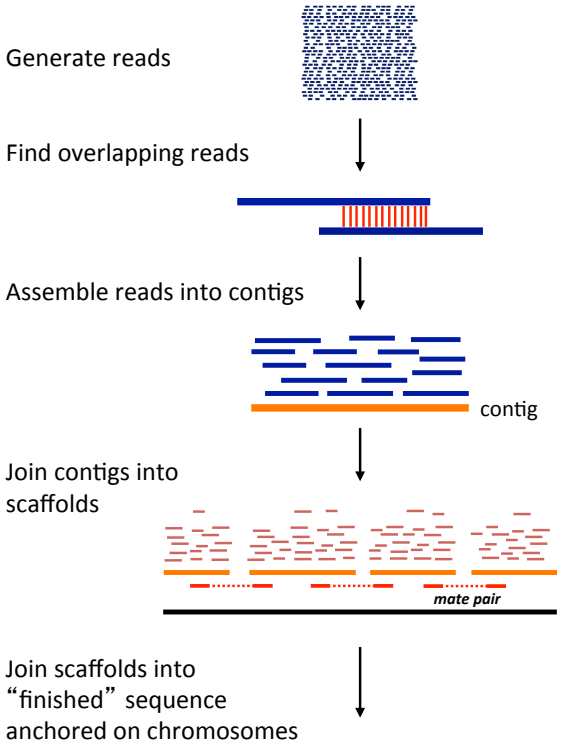
wellcome trust sanger institute

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



Novaseq 1 billion reads 2x150 bp. $3000 -> <$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)

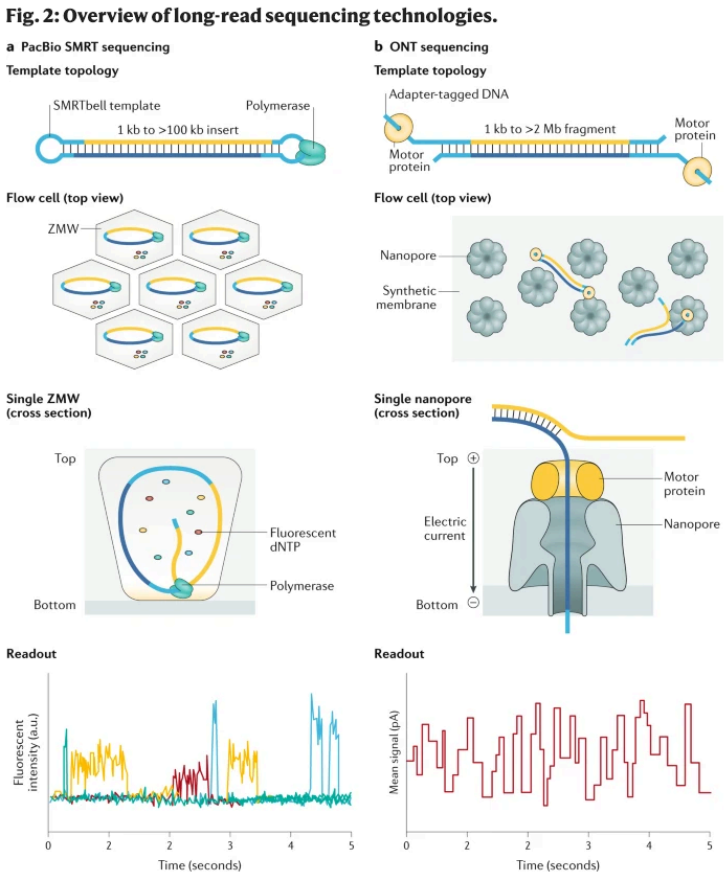
Coverage:
Average number of reads representing a particular position in the assembly

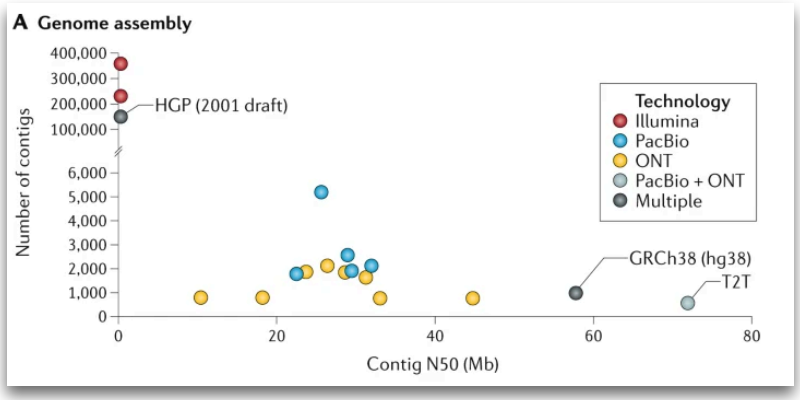
Human, Mouse, Rat: > 20x
Chimpanzee: ~6x
Squirrel: ~2x

Scaffold_0: 12,865,123 – 12,965-110

Chr5: 133,876,119 – 134,876,119

The importance of long read sequencing





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

The importance of long read sequencing

SPECIAL SECTION COMPLETING THE HUMAN GENOME

RESEARCH ARTICLE

HUMAN GENOMICS

The complete sequence of a human genome

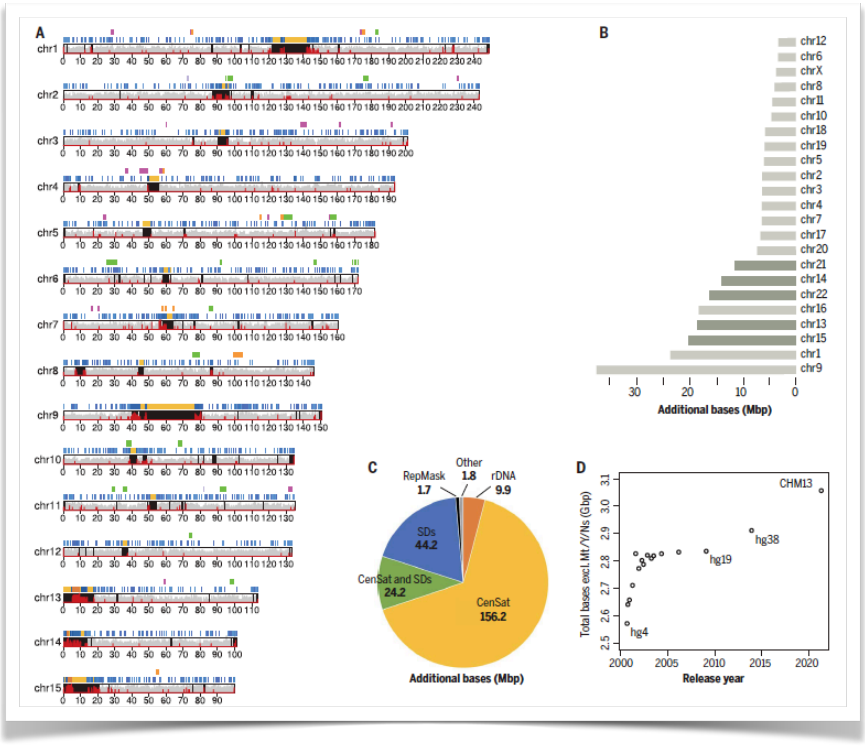
Sergey Nurk[1†], Sergey Koren[1†], Arang Rhie[1†], Mikko Rautiainen[1†], Andrey V. Bzikadze[2], Alla Mikheenko[3], Mitchell R. Vollger[4], Nicolas Altemose[5], Lev Uralsky[6,7], Ariel Gershman[8], Sergey Aganezov[9†], Savannah J. Hoyt[10], Mark Diekhans[11], Glennis A. Logsdon[4], Michael Alonge[9], Stylianos E. Antonarakis[12], Matthew Borchers[13], Gerard G. Bouffard[14], Shelise Y. Brooks[14], Gina V. Caldas[15], Nae-Chyun Chen[9], Haoyu Cheng[16,17], Chen-Shan Chin[18], William Chow[19], Leonardo G. de Lima[13], Philip C. Dishuck[4], Richard Durbin[19,20], Tatiana Dvorkina[3], Ian T. Fiddes[21], Giulio Formenti[22,23], Robert S. Fulton[24], Arkarachai Fungtammasan[25], Erik Garrison[21,26], Patrick G. S. Grady[10], Tina A. Graves-Lindsay[24], Ira M. Hall[27], Nancy F. Hansen[28], Gabrielle A. Hartley[29], Marina Haukness[30], Kerstin Howe[31], Michael W. Hunkapiller[32], Chirag Jain[1,33], Miten Jain[34], Erich D. Jarvis[22,23], Peter Kerpedjiev[35], Melanie Kirsche[9], Mikhail Kolmogorov[36], Jonas Korlach[29], Milinn Kremitzki[24], Heng Li[30,37], Valerie V. Maduro[38], Tobias Marschall[39,40], Ann M. McCartney[3], Jennifer McDaniel[28], Danny E. Miller[4,16], James C. Mullikin[14,28], Eugene W. Myers[41,42], Nathan D. Olson[28], Benedict Paten[21], Paul Peluso[29], Pavel A. Pevzner[32], David Porubsky[4], Tamara Potapova[13], Evgeny I. Rogaev[6,7,38,39], Jeffrey A. Rosenfeld[40], Steven L. Salzberg[3,41], Valerie A. Schneider[42], Fritz J. Sedlazeck[43], Kishwar Shafin[19], Colin J. Shew[44], Alaina Shumate[21], Ying Sims[19], Arian F. A. Smit[45], Daniela C. Soto[4,46], Ivan Sović[25,46], Jessica M. Storer[45], Aaron Streets[5,47], Beth A. Sullivan[48], Françoise Thibaud-Nissen[49], James Torrance[19], Justin Wagner[28], Brian P. Walenz[1], Aaron Wenger[29], Jonathan M. D. Wood[19], Chunlin Xiao[28], Stephanie M. Yan[43], Alice C. Young[24], Samantha Zarate[9], Urvashi Surti[50], Rajiv C. McCoy[40], Megan Y. Dennis[44], Ivan A. Alexandrov[2,51], Jennifer L. Gerton[13,52], Rachel J. O'Neill[53], Winston Timp[8,41], Justin M. Zook[28], Michael C. Schatz[9,43], Evan E. Eichler[4,54]*, Karen H. Miga[21,54]*, Adam M. Phillippy[1]*

Since its initial release in 2000, the human reference genome has covered only the euchromatic fraction of the genome, leaving important heterochromatic regions unfinished. Addressing the remaining 8% of the genome, the Telomere-to-Telomere (T2T) Consortium presents a complete 3.055 billion–base pair sequence of a human genome, T2T-CHM13, that includes gapless assemblies for all chromosomes except Y, corrects errors in the prior references, and introduces nearly 200 million base pairs of sequence containing 1956 gene predictions, 99 of which are predicted to be protein coding. The completed regions include all centromeric satellite arrays, recent segmental duplications, and the short arms of all five acrocentric chromosomes, unlocking these complex regions of the genome to variational and functional studies.

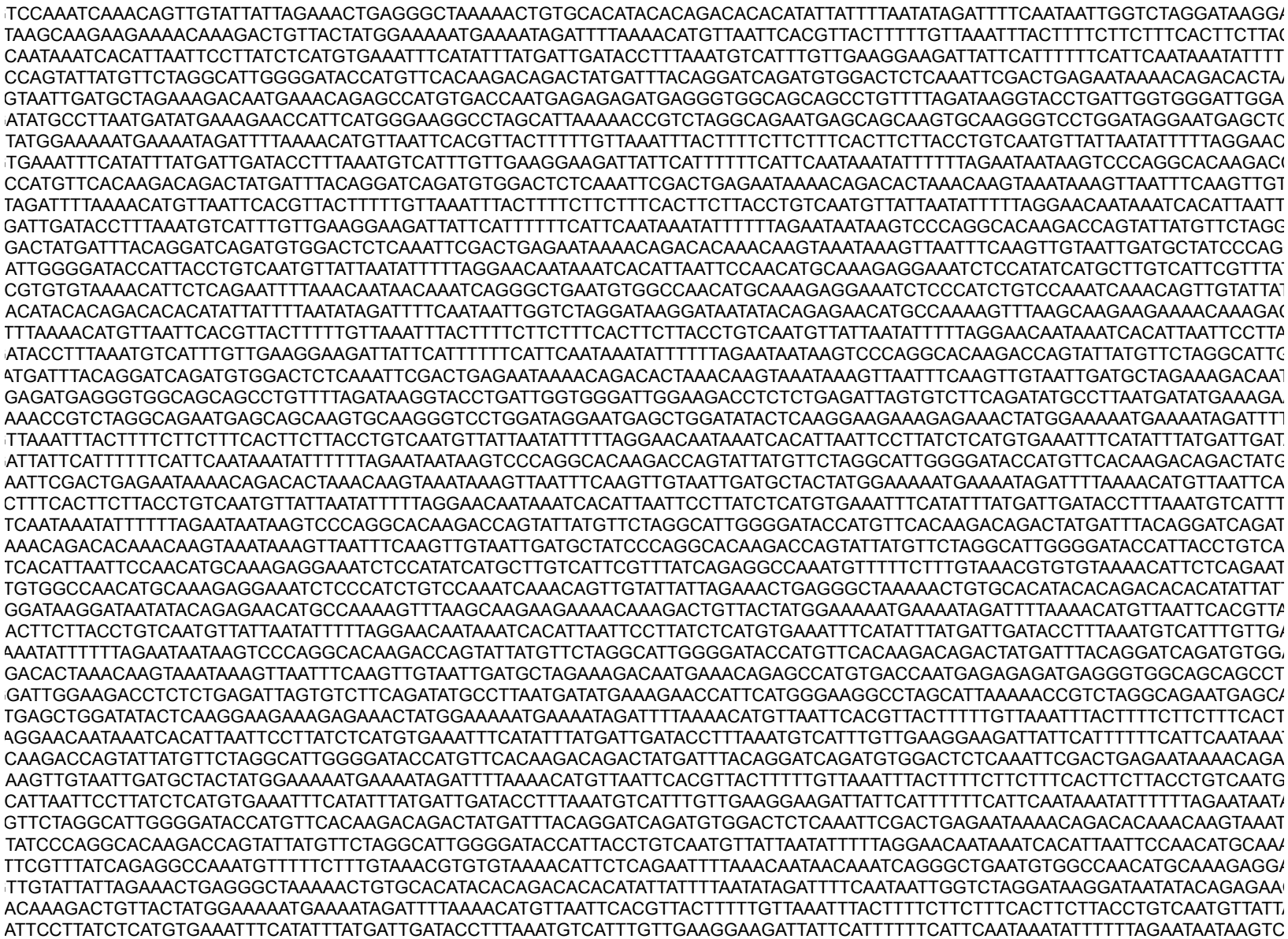
The current human reference genome was released by the Genome Reference Consortium (GRC) in 2013 and most recently patched in 2019 (GRCh38.p13) (*1*). This reference traces its origin to the publicly funded Human Genome Project (*2*) and has been continually improved over the past two decades. Unlike the competing Celera effort (*3*) and most modern sequencing projects based on "shotgun" sequence assembly (*4*), the GRC assembly was constructed from sequenced bacterial artificial chromosomes (BACs) that were ordered and oriented along the human genome by means of radiation hybrid, genetic linkage, and fingerprint maps. However, limitations of BAC cloning led to an underrepresentation of repetitive sequences, and the opportunistic assembly of BACs derived from multiple individuals resulted in a mosaic of haplotypes. As a result, several GRC assembly gaps are unsolvable because of incompatible structural polymorphisms on their flanks, and many other repetitive and polymorphic regions were left unfinished or incorrectly assembled (*5*).

The GRCh38 reference assembly contains 151 mega–base pairs (Mbp) of unknown sequence distributed throughout the genome, including pericentromeric and subtelomeric regions, recent segmental duplications, ampliconic gene arrays, and ribosomal DNA (rDNA) arrays, all of which are necessary for fundamental cellular processes (Fig. 1A). Some of the largest reference gaps include human satellite (HSat) repeat arrays and the short arms of all five acrocentric chromosomes, which are represented in GRCh38 as multimegabase stretches of unknown bases (Fig. 1, B and C). In addition to these apparent gaps, other regions of GRCh38 are artificial or are otherwise incorrect. For example, the centromeric alpha satellite arrays are represented as computationally generated models of alpha satellite monomers to serve as decoys for resequencing analyses (*6*), and sequence assigned to the short arm of chromosome 21 appears falsely duplicated and poorly assembled (*7*). When compared with other human genomes, GRCh38 also shows a genome-wide deletion bias that is indicative of incomplete assembly (*8*). Despite finishing efforts from both the Human Genome Project (*9*) and GRC (*1*) that improved the quality of the reference, there was limited

[1]Genome Informatics Section, Computational and Statistical Genomics Branch, National Human Genome Research Institute, National Institutes of Health, Bethesda, MD, USA. [2]Graduate Program in Bioinformatics and Systems Biology, University of California, San Diego, La Jolla, CA, USA. [3]Center for Algorithmic Biotechnology, Institute of Translational Biomedicine, Saint Petersburg State University, Saint Petersburg, Russia. [4]Department of Genome Sciences, University of Washington School of Medicine, Seattle, WA, USA. [5]Department of Bioengineering, University of California, Berkeley, Berkeley, CA, USA. [6]Kazus University of Science and Technology, Sochi, Russia. [7]Vavilov Institute of General Genetics, Moscow, Russia. [8]Department of Molecular Biology and Genetics, Johns Hopkins University, Baltimore, MD, USA. [9]Department of Computer Science, Johns Hopkins University, Baltimore, MD, USA. [10]Institute for Systems Genomics and Department of Molecular and Cell Biology, University of Connecticut, Storrs, CT, USA. [11]UC Santa Cruz Genomics Institute, University of California, Santa Cruz, Santa Cruz, CA, USA. [12]University of Geneva Medical School, Geneva, Switzerland. [13]Stowers Institute for Medical Research, Kansas City, MO, USA. [14]NIH Intramural Sequencing Center, National Human Genome Research Institute, National Institutes of Health, Bethesda, MD, USA. [15]Department of Molecular and Cell Biology, University of California, Berkeley, Berkeley, CA, USA. [16]Department of Data Sciences, Dana-Farber Cancer Institute, Boston, MA, USA. [17]Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA. [18]DNAnexus, Mountain View, CA, USA. [19]Wellcome Sanger Institute, Cambridge, UK. [20]Department of Genetics, University of Cambridge, Cambridge, UK. [21]Biomedical Engineering, University of California, Santa Cruz, Santa Cruz, CA, USA. [22]Laboratory of Neurogenetics of Language and The Vertebrate Genome Lab, The Rockefeller University, New York, NY, USA. [23]Howard Hughes Medical Institute, The Rockefeller University, New York, NY, USA. [24]Department of Genetics, Washington University School of Medicine, St. Louis, MO, USA. [25]University of Tennessee Health Science Center, Memphis, TN, USA. [26]McDonnell Genome Institute, Washington University in St. Louis, St. Louis, MO, USA. [27]Department of Genetics, Yale University School of Medicine, New Haven, CT, USA. [28]Comparative Genomics Analysis Unit, Cancer Genetics and Comparative Genomics Branch, National Human Genome Research Institute, National Institutes of Health, Bethesda, MD, USA. [29]Pacific Biosciences, Menlo Park, CA, USA. [30]Department of Computational and Data Sciences, Indian Institute of Science, Bangalore KA, India. [31]Reannzor Genomics LLC, Oakland, CA, USA. [32]Department of Computer Science and Engineering, University of California, San Diego, La Jolla, CA, USA. [33]Undiagnosed Diseases Program, National Human Genome Research Institute, National Institutes of Health, Bethesda, MD, USA. [34]Heinrich Heine University Düsseldorf, Medical Faculty, Institute for Medical Biometry and Bioinformatics, Düsseldorf, Germany. [35]Biosystems and Biomaterials Division, National Institute of Standards and Technology, Gaithersburg, MD, USA. [36]Department of Pediatrics, Division of Genetic Medicine, University of Washington and Seattle Children's Hospital, Seattle, WA, USA. [37]Max Planck Institute of Molecular Cell Biology and Genetics, Dresden, Germany. [38]Department of Psychiatry, University of Massachusetts Medical School, Worcester, MA, USA. [39]Faculty of Biology, Lomonosov Moscow State University, Moscow, Russia. [40]Cancer Institute of New Jersey, New Brunswick, NJ, USA. [41]Department of Biomedical Engineering, Johns Hopkins University, Baltimore, MD, USA. [42]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD, USA. [43]Human Genome Sequencing Center, Baylor College of Medicine, Houston, TX, USA. [44]Genome Center, MIND Institute, Department of Biochemistry and Molecular Medicine, University of California, Davis, CA, USA. [45]Institute for Systems Biology, Seattle, WA, USA. [46]Digital BioLogic d.o.o., Ivanić-Grad, Croatia. [47]Chan Zuckerberg Biohub, San Francisco, CA, USA. [48]Department of Molecular Genetics and Microbiology, Duke University School of Medicine, Durham, NC, USA. [49]Department of Biology, Johns Hopkins University, Baltimore, MD, USA. [50]Department of Pathology, University of Pittsburgh, Pittsburgh, PA, USA. [51]Research Center of Biotechnology of the Russian Academy of Sciences, Moscow, Russia. [52]Department of Biochemistry and Molecular Biology, University of Kansas Medical School, Kansas City, MO, USA. [53]Howard Hughes Medical Institute, University of Washington, Seattle, WA, USA. [54]Department of Biomolecular Engineering, University of California, Santa Cruz, Santa Cruz, CA, USA
*Corresponding author. Email: eee@gs.washington.edu (E.E.E.); khmiga@ucsc.edu (K.H.M.); adam.phillippy@nih.gov (A.M.P.)
†These authors contributed equally to this work. ‡Present address: Oxford Nanopore Technologies Inc., Lexington, MA, USA.



https://www.science.org/doi/10.1126/science.abj6987



What types of annotation do we have/want?

~3 billion bp

Genes:

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

Genetic variation:

- SNPs and CNVs

Sequence conservation

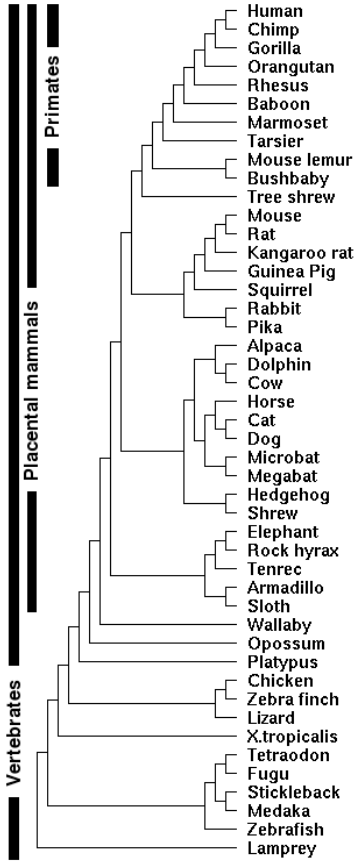
Regulatory sequences:

- 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

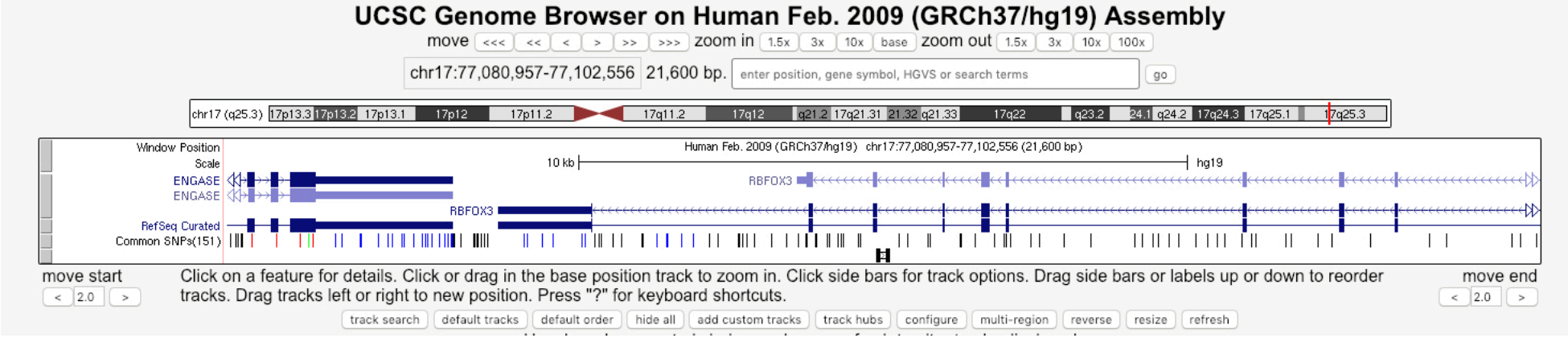
Example of a genome browser track (UCSC)

Chr5: 133,876,119 – 134,876,119

Our specific example:

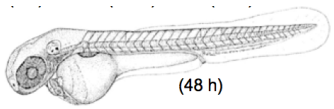
```
@HWI-ST1239:178:H0KPNADXX:2:1101:3120:1979 1:N:0:TGACCA
NCTGTAGGCTGCGTAGCCTCCCTGCAGGGTAAGTGGGAGGAGAGAGAGCAGAGGGACTTAGTGGGGCTCCCCAGGG
+
#1=DDFFFHHHHHIJIJJJIJJJJJIJJJ?FHIDGIJ=GIHGIIIHGIJIHEHIHHGFFFFEEEDDDDDDDDDDDD

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



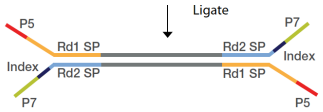
Workflow

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 class.

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.