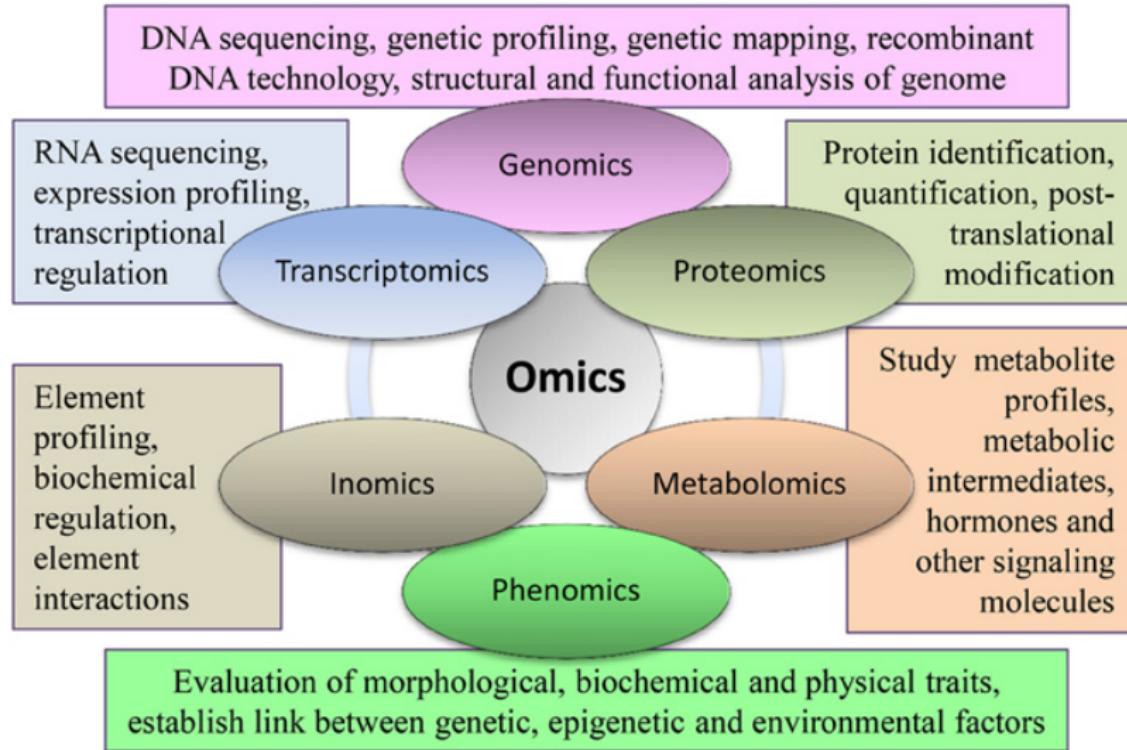


Genomic technologies

Mikhail Dozmorov

Spring 2018

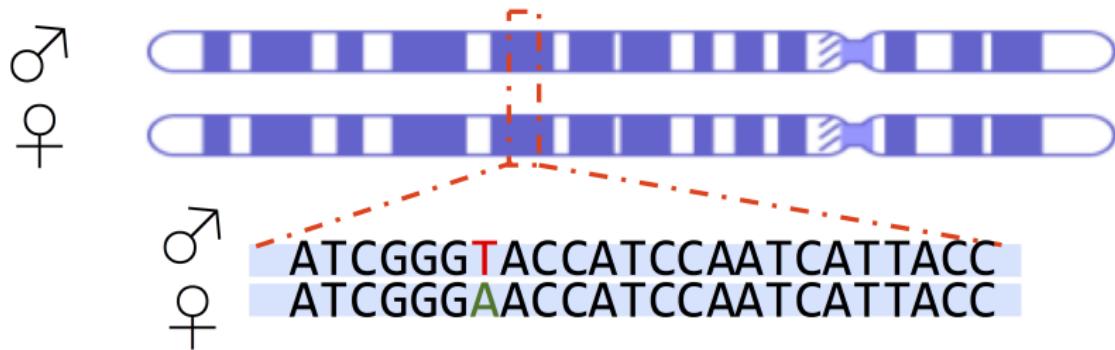
Age of OMICS



Genome in a nutshell

Genome arithmetics

- Haploid (one copy) human genome has 23 chromosomes, autosomes (chromosome 1-22) and one sex chromosome (X, Y)
- Human genome is *diploid* - comprised of a paternal and a maternal “haplotype”. Together, they form our “genotype” of 46 chromosomes

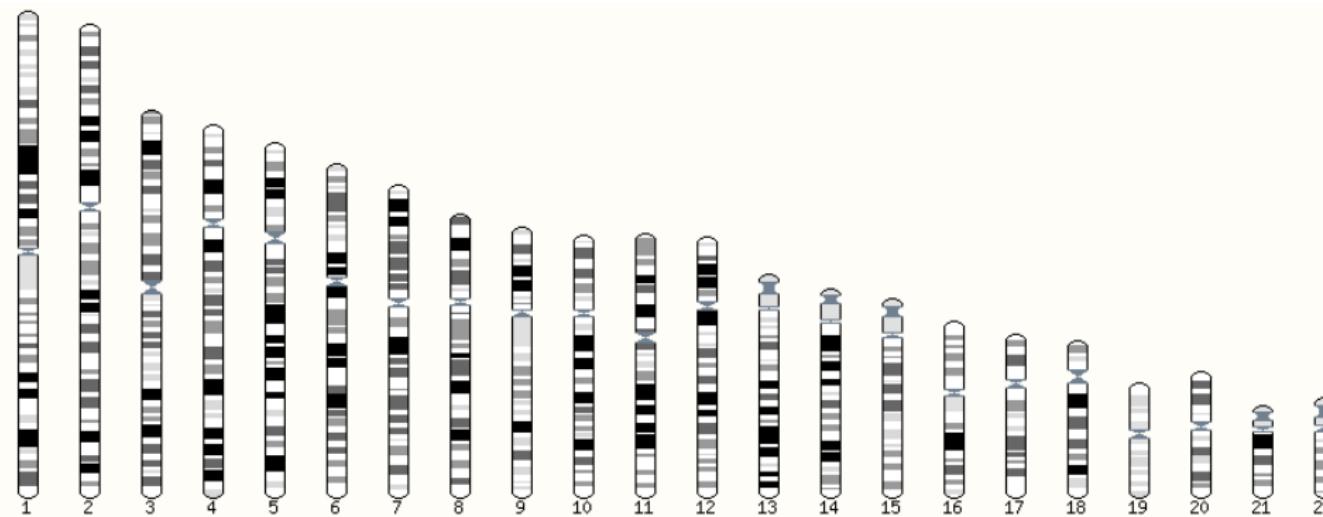


Genome arithmetics

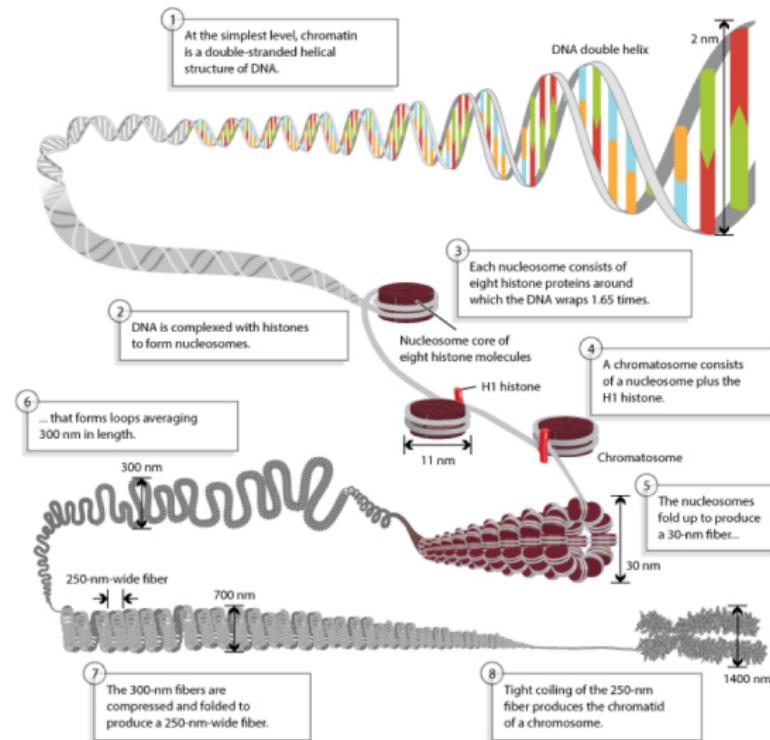
- One genome per cell, located in the nucleus - most of the time (Red blood cells lack chromosomes)
- Mitochondria (cell powerhouses) have their own genomes - many mitochondrial genomes (Liver cells have 1000-2000 mito)
- A typical human is comprised of roughly 40 trillion human cells (excluding trillions of bacterial cells in our gut)
- If stretched out, each haploid genome would be roughly 2 meters - each cell has 4 meters of DNA (1 m = 3.28 ft)
- $40 \text{ trillion} * 4 \text{ meters} = 160 \text{ trillion meters.}$
- $160 \text{ trillion meters} / 1609.34 = 99,750,623,441 \text{ miles}$
- $99,750,623,441 / 92,960,000 = 1,073.05 \text{ trips to the sun.}$

Genome arithmetics

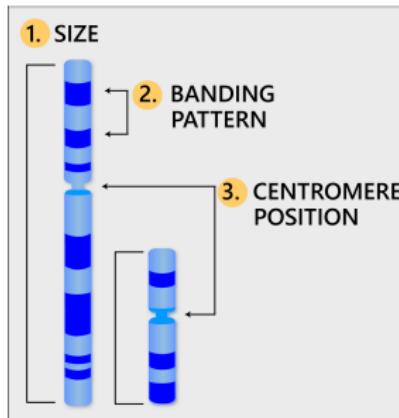
- ~3,235 billion base pairs (haploid)
- ~20,000 protein coding genes
- ~200,000 coding transcripts (isoforms of a gene that each encode a distinct protein product)



The human genome from a micro to macro scale

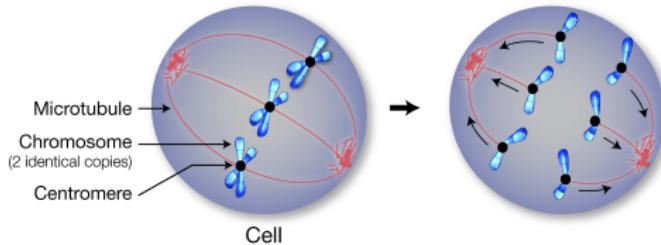


The basic structure of a chromosome



- **Size.** This is the easiest way to tell chromosomes apart.
- **Banding pattern.** The size and location of Giemsa bands make each chromosome unique.
- **Centromere position.** Centromeres appear as a constriction. They have a role in the separation of chromosomes into daughter cells during cell division (mitosis and meiosis).

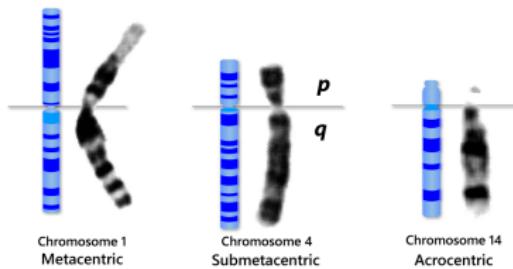
The role of the centromere



- Centromeres are required for chromosome separation during cell division.
- The centromeres are attachment points for microtubules, which are protein fibers that pull duplicate chromosomes toward opposite ends of the cell before it divides.
- This separation ensures that each daughter cell will have a full set of chromosomes.
- Each chromosome has only one centromere.

<http://learn.genetics.utah.edu/content/basics/readchromosomes/>

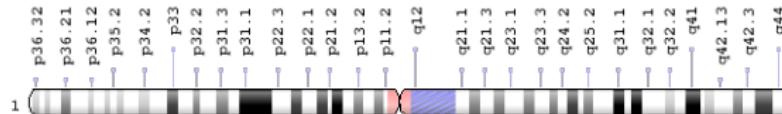
Centromere positions



The position of the centromere relative to the ends helps scientists tell chromosomes apart. Centromere position can be described as:

- **Metacentric** - the centromere lies near the center of the chromosome.
- **Submetacentric** - the centromere that is off-center, so that one chromosome arm is longer than the other. The short arm is designated "p" (for petite), and the long arm is designated "q" (because it follows the letter "p").
- **Acrocentric** - the centromere is very near one end.

Chromosome Giemsa banding (G-banding)



- Heterochromatic regions, which tend to be rich with adenine and thymine (AT-rich) DNA and relatively gene-poor, stain more darkly with Giemsa and result in G-banding
- Less condensed (“open”) chromatin, which tends to be (GC-rich) and more transcriptionally active, incorporates less Giemsa stain, resulting in light bands in G-banding.
- Cytogenetic bands are labeled p1, p2, p3, q1, q2, q3, etc., counting from the centromere out toward the telomeres. At higher resolutions, sub-bands can be seen within the bands.
- For example, the locus for the CFTR (cystic fibrosis) gene is 7q31.2, which indicates it is on chromosome 7, q arm, band 3, sub-band 1, and sub-sub-band 2. (Say 7,q,3,1 dot 2)

Gene content

- “There appear to be about $30,000 \pm 40,000$ protein-coding genes in the human genome – only about twice as many as in worm or fly. However, the genes are more complex, with more alternative splicing generating a larger number of protein products.”
- Over time this has evolved to an estimate of approximately 20,000 protein coding genes, which reflects roughly the number of genes in fly and worm

Table 21 Characteristics of human genes

	Median	Mean
Internal exon	122 bp	145 bp
Exon number	7	8.8
Introns	1,023 bp	3,365 bp
3' UTR	400 bp	770 bp
5' UTR	240 bp	300 bp
Coding sequence (CDS)	1,100 bp 367 aa	1,340 bp 447 aa
Genomic extent	14 kb	27 kb

<http://www.nature.com/nature/journal/v409/n6822/full/409860a0.html>

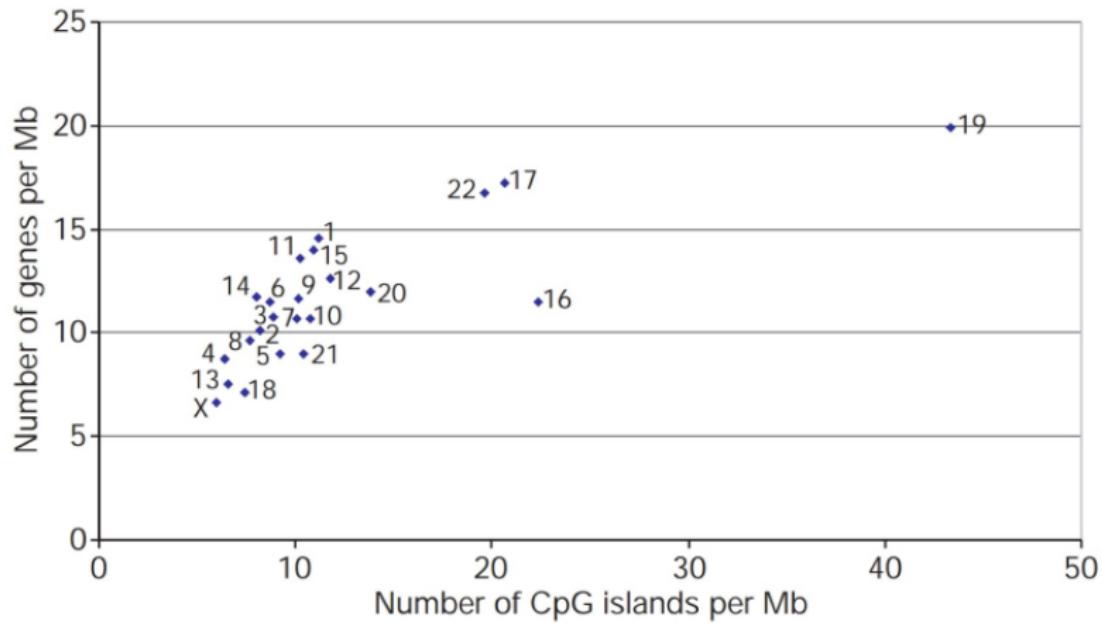
Genes are unevenly distributed across chromosomes

- Highly expressed genes positively correlated with:
 - Very short indels
 - High gene density
 - High GC content
 - High density of Short interspersed nuclear elements (SINE) repeats
 - Low density of Long interspersed nuclear elements (LINE) repeats
 - Both housekeeping and tissue-specific expression
- The opposite is true for lowly expressed genes

Versteeg, Rogier, Barbera D. C. van Schaik, Marinus F. van Batenburg, Marco Roos, Ramin Monajemi, Huib Caron, Harmen J. Bussemaker, and Antoine H. C. van Kampen. "The Human Transcriptome Map Reveals Extremes in Gene Density, Intron Length, GC Content, and Repeat Pattern for Domains of Highly and Weakly Expressed Genes." *Genome Research* 13, no. 9 (September 2003): 1998–2004. <https://doi.org/10.1101/gr.1649303>.

Genes are unevenly distributed across chromosomes

Chromosome 19 is the most gene dense chromosome in the human genome



GENCODE – Annotation Gene Features

- ~21,000 protein coding genes
- PolyA+
 - Almost completely spliced before nuclear export – co-transcriptional splicing “first transcribed – first spliced”
 - Most have at least 2 dominate splice forms
 - Show allele specific expression – potential imprinting
- PolyA-
 - Many are lncRNAs
 - Also shows allele specific expression

GENCODE – Annotation Gene Features

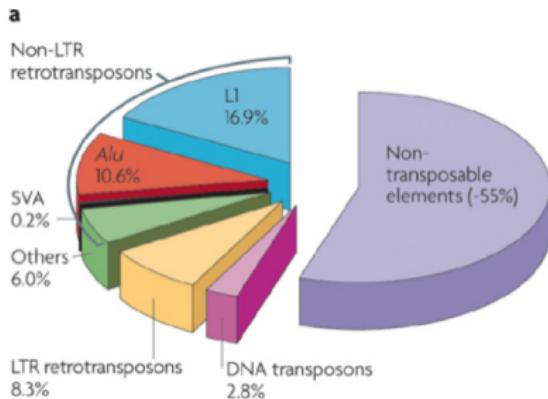
- Most (62%) of the genome is transcribed – the genome is pervasively transcribed
 - <5% can be identified as exons
- ~12,000 pseudogenes – results of duplications
 - 876 are transcribed – can have regulatory function by serving as decoys
 - Infrequently spliced
- ~10,000 lncRNA = noncoding RNAs >200bp
 - 92% are not translated
 - Many show tissue-specific expression – more so than protein coding genes
 - 33% are primate specific but few are human specific – most new genes are in this category
 - Poorly spliced – most are two exon transcripts

GENCODE – Annotation Gene Features

- ~9000 small RNAs - many of the lncRNA transcripts are processed into stable small RNAs
 - tRNA, miRNA, siRNA, snRNA, snoRNA
- ~82,000 – 128,000 transcription start sites - depending on detection method
 - ~44% are near annotated transcripts
- ~5,000 RNA edits occur post transcription
 - Mostly A to G(I) conversions (APOBEC pathway)
 - 94% are in transcribed repeat elements
 - Remaining are mostly in introns, 3'UTRs
 - Very few (123) in protein coding sequences

Half of the human genome is low complexity

- Retrotransposons - fossil records of evolution
 - McClintock's "jumping genes" in maize
 - Retrotransposons use a "copy/paste" mechanism - transcribed to RNA and then reverse transcribed into DNA and insert
 - DNA transposons use a "cut/paste" mechanism - excise themselves and insert to another place



Repeats

- Repetitive DNA not driven by retrotransposition (e.g., ATATATATATATATAT...)
- CpG islands - clusters of CG dinucleotides (The “p” represents the phosphate bond between the nucleotides on the same strand. Needed to distinguish between hydrogen bond between C and G on complementary DNA strands)

Table 10 Number of CpG islands by GC content

GC content of island	Number of islands	Percentage of islands	Nucleotides in islands
Total	28,890	100	19,818,547
>80%	22	0.08	5,916
70–80%	5,884	20	3,111,965
60–70%	18,779	65	13,110,924
50–60%	4,205	15	3,589,742

<https://www.nature.com/nature/journal/v409/n6822/full/409860a0.html>

Genome variability

A typical genome differs from the reference genome at 4.1 to 5.0 million sites - Single Nucleotide Polymorphisms (SNPs)

- Over 99.9% are SNPs or short indels
- Only 1-4% are rare (frequency <0.5% in the population)
- Contains 2,100 – 2,500 structural variants, which affect more bases (~20 million bases)
- ~1,000 large deletions
- ~1,094 Alu, L1, SINE (short interspersed nuclear element), VNTR (variable number tandem repeat) insertions
- ~160 CNVs
- ~10 inversions
- ~ 4 NUMTs (nuclear mitochondrial DNA variations)

Genome variability

- 149-182 protein truncating variants
- ~2,000 variants associated with complex traits
- 24-30 variants associated with rare disease
- On average 74 *de novo* SNVs per individual

Conrad DF, Keebler JE, DePristo MA, Lindsay SJ, Zhang Y, Casals F, Idaghdour Y, Hartl CL, Torroja C, Garimella KV, Zilversmit M, Cartwright R, Rouleau GA, Daly M, Stone EA, Hurles ME, Awadalla P; 1000 Genomes Project. Variation in genome-wide mutation rates within and between human families. Nat Genet. 2011 Jun 12;43(7):712-4. doi: 10.1038/ng.862.
<https://www.nature.com/articles/ng.862>

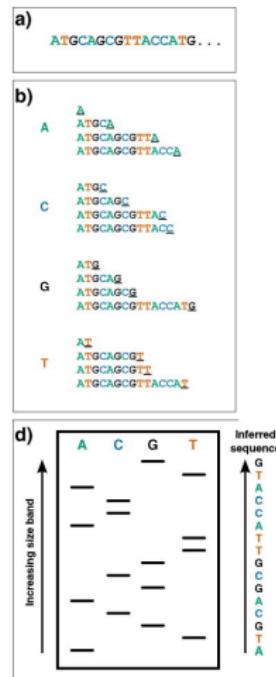
Genome sequencing

Why sequence a reference genome?

- Determine the “complete” sequence of a human haploid genome.
- Identify the sequence and location of every protein coding gene.
- Use as a “map” with which to track the location and frequency of genetic variation in the human genome.
- Unravel the genetic architecture of inherited and somatic human diseases.
- Understand genome and species evolution

DNA sequencing: Maxam-Gilbert, Sanger

- 1) Sequencing by synthesis (not degradation)
- 2) Radioactive primers hybridize to DNA
- 3) Polymerase + dNTPs (normal dNTPs) + ddNTP (dideoxynucleotides terminators) at low concentration
- 4) 1 lane per base, visually interpret ladder



https://en.wikipedia.org/wiki/Maxam%20%93Gilbert_sequencing

<https://www.youtube.com/watch?v=bEFLBf5WEtc>

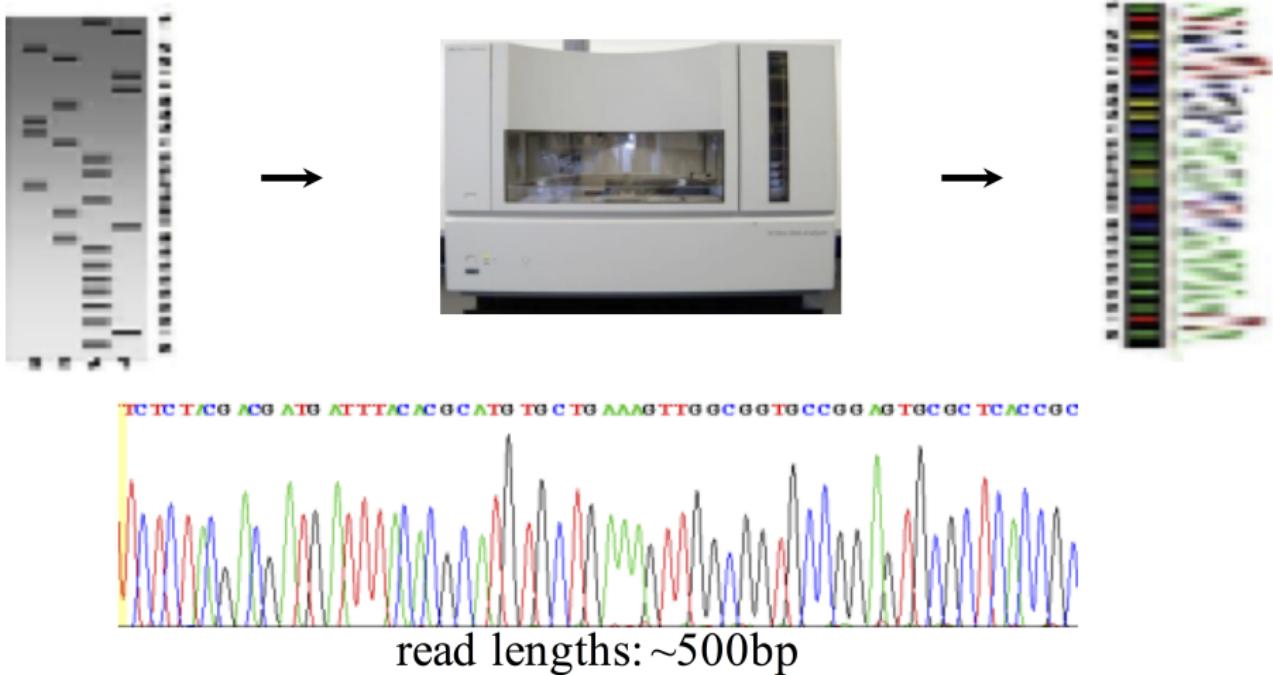
Shotgun genome sequencing milestones

- 1977: Bacteriophage $\Phi X147$ (5kb)
- 1995: *H. Influenza* (1Mb);
- 1996: Yeast (12mb);
- 2000: *Drosophila* (165Mb);
- 2002: Human (3Gb)

https://en.wikipedia.org/wiki/Phi_X_174

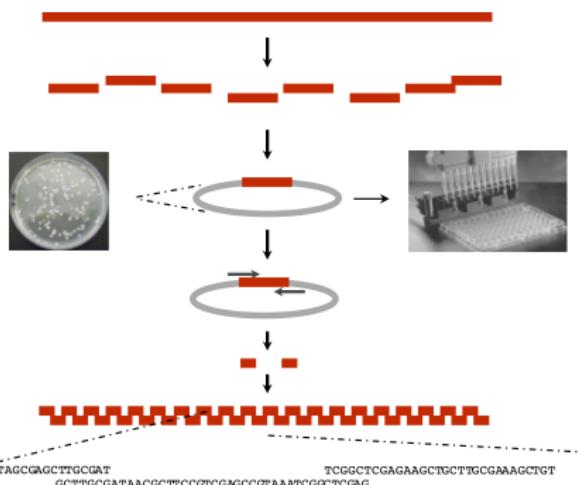
Sequencing on a scale

How to sequence a human genome: Lee Hood automation



Shotgun genome sequencing (Sanger, 1979)

- ① Fragment the genome (or large Bacterial artificial chromosome (BAC) clones)
- ② Clone 2-10kb fragments into plasmids; pick lots of colonies; purify DNA from each
- ③ Use a primer to plasmid to sequence into genomic DNA
- ④ Assemble the genome from overlapping “reads”



ATCGCC6TACTAGCGAGCTTCGAT
GCTTGCATTAACGCTCCGTGAGCCGAAATCGCTCGAG

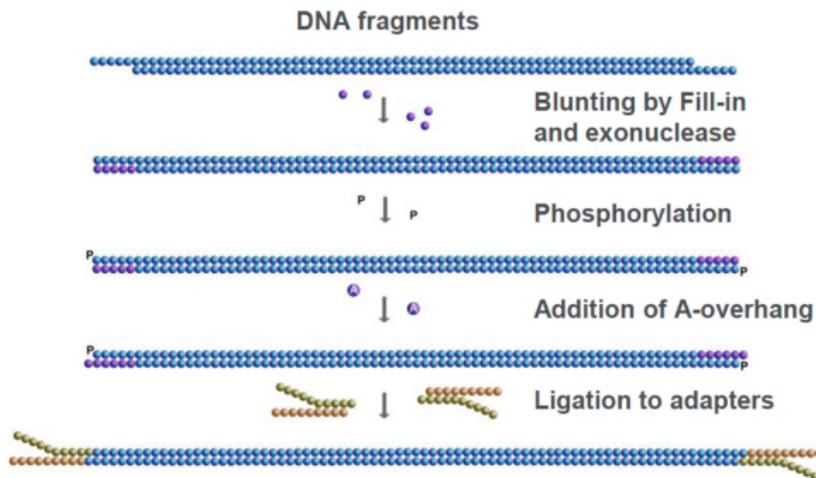
Sequencing in a nutshell

- Cut the long DNA into smaller segments (several hundreds to several thousand bases).
- Sequence each segment: start from one end and sequence along the chain, base by base.
- The process stops after a while because the noise level is too high.
- Results from sequencing are many sequence pieces. The lengths vary, usually a few thousands from Sanger, and several hundreds from NGS.
- The sequence pieces are called “reads” for NGS data.

Massively Parallel DNA sequencing instruments

- All MPS platforms require a library obtained either by amplification or ligation with custom linkers (adapters)
- Each library fragment is amplified on a solid surface (either bead or flat *Si*-derived surface) with covalently attached adapters that hybridize the library adapters
- Direct step-by-step detection of the nucleotide base incorporated by each amplified library fragment set
- Hundreds of thousands to hundreds of millions of reactions detected per instrument run = “massively parallel sequencing”
- A “digital” read type that enables direct quantitative comparisons
- Shorter read lengths than capillary sequencers

Library Construction for MPS



- Shear high molecular weight DNA with sonication
- Enzymatic treatments to blunt ends
- Ligate synthetic DNA adapters (each with a DNA barcode), PCR amplify
- Quantitate library
- Proceed to WGS, or do exome or specific gene hybrid capture

PCR-related Problems in MPS

- PCR is an effective vehicle for amplifying DNA, however...
- In MPS library construction, PCR can introduce preferential amplification (“jackpotting”) of certain fragments
- Duplicate reads with exact start/stop alignments
- Need to “de-duplicate” after alignment and keep only one pair
- Low input DNA amounts favor jackpotting due to lack of complexity in the fragment population

PCR-related Problems in MPS

- PCR also introduces false positive artifacts due to substitution errors by the polymerase
- If substitution occurs in early PCR cycles, error appears as a true variant
- If substitution occurs in later cycles, error typically is drowned out by correctly copied fragments in the cluster
- Cluster formation is a type of PCR (“bridge amplification”) • Introduces bias in amplifying high and low G+C fragments
- Reduced coverage at these loci is a result

Hybrid Capture

- Hybrid capture - fragments from a whole genome library are selected by combining with probes that correspond to most (not all) human exons or gene targets.
- The probe DNAs are biotinylated, making selection from solution with streptavidin magnetic beads an effective means of purification.
- An “exome” by definition, is the exons of all genes annotated in the reference genome.
- Custom capture reagents can be synthesized to target specific loci that may be of clinical interest.

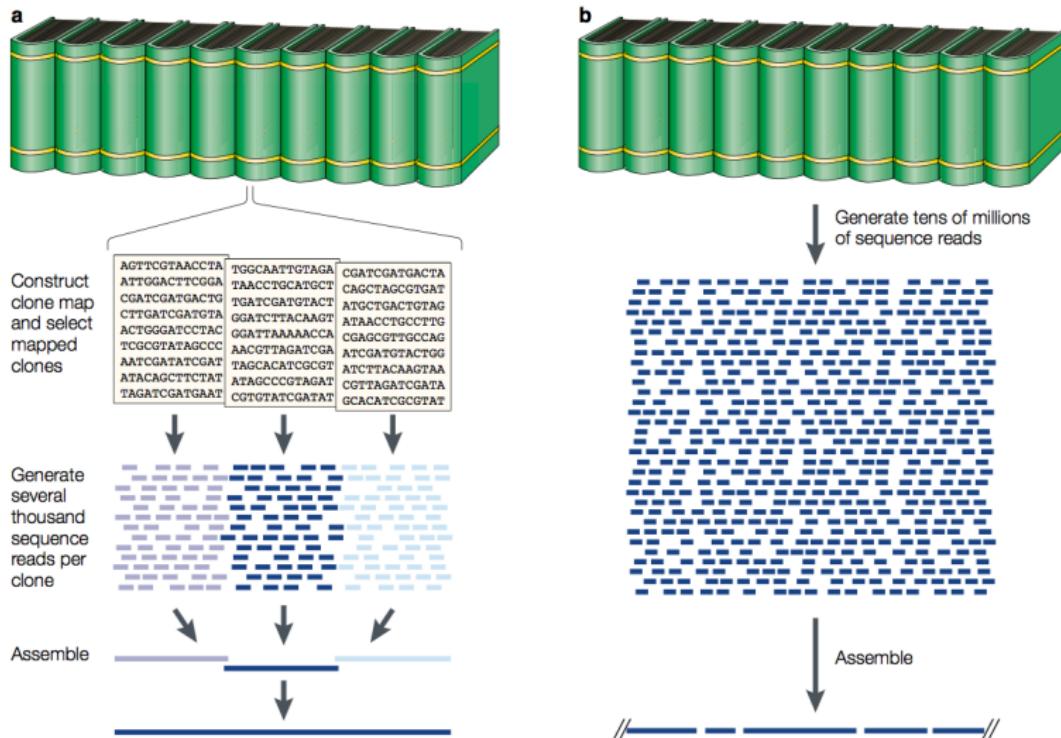
The Human Genome project

Early days of human genome sequencing



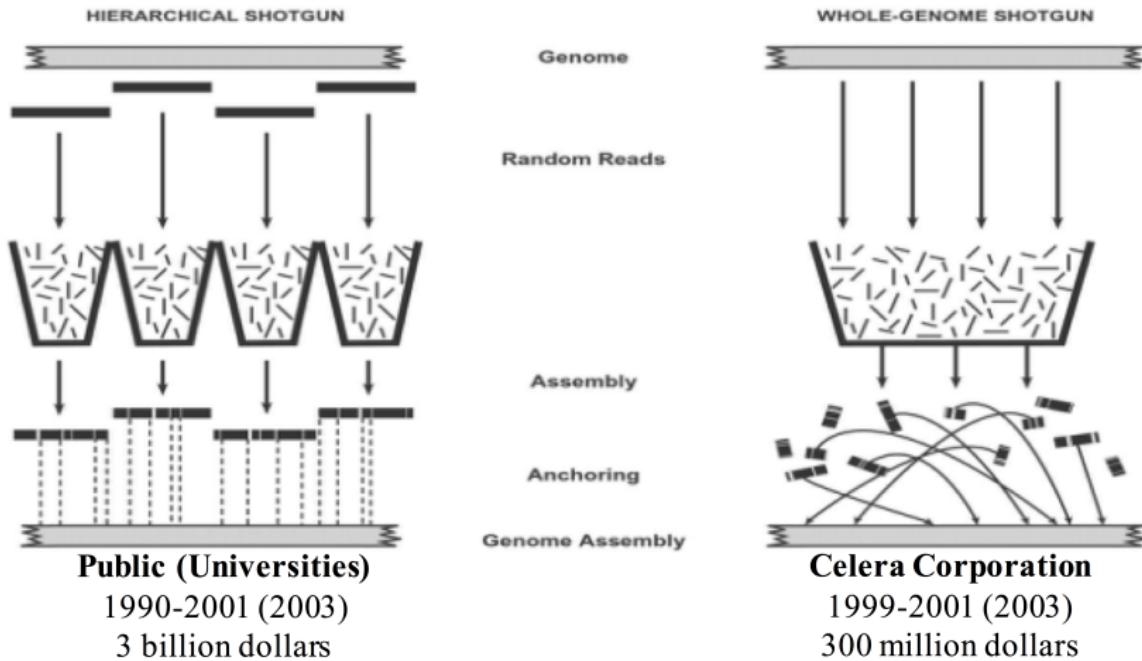
Green, Eric D., James D. Watson, and Francis S. Collins. "Human Genome Project: Twenty-Five Years of Big Biology." *Nature* 526, no. 7571 (October 1, 2015): 29–31. doi:10.1038/526029a.

Two shotgun-sequencing strategies



https://www.nature.com/nrg/journal/v2/n8/full/nrg0801_573a.html

The competing human genome projects



A first map of the human genome

articles

Initial sequencing and analysis of the human genome

International Human Genome Sequencing Consortium*

*A partial list of authors appears on the opposite page. Affiliations are listed at the end of the paper.

The human genome holds an extraordinary trove of information about human development, physiology, medicine and evolution. Here we report the results of an international collaboration to produce and make freely available a draft sequence of the human genome. We also present an initial analysis of the data, describing some of the insights that can be gleaned from the sequence.

<http://www.nature.com/nature/journal/v409/n6822/full/409860a0.html>

Human genome is sequenced!

'All the News
That's Fit to Print'

The New York Times

VOL. CXLIX . . No. 51,432

Copyright © 2000 The New York Times

NEW YORK, TUESDAY, JUNE 27, 2000

\$1 beyond the greater New York metropolitan area.

75 CENTS

Late Edition
New York: Today, afternoon thunderstorms, high 88. Tonight, showers end, low 67. Tomorrow, partly cloudy with showers late, high 81. Yesterday, high 88, low 74. Weather map, Page D8.

Genetic Code of Human Life Is Cracked by Scientists

JUSTICES REAFFIRM MIRANDA RULE, 7-2; A PART OF 'CULTURE'

By LINDA GREENHOUSE

WASHINGTON, June 26 — The Supreme Court reaffirmed the Miranda decision today by a 7-to-2 vote that erased a shadow over one of the most famous "rulings of modern times." The majority held that the Miranda warnings "have become part of our national culture."

The court said in an opinion by Chief Justice William H. Rehnquist that because the 1966 Miranda decision "announced a constitutional rule," a statute by which Congress had sought to overrule the decision was not required by the Constitution.

Miranda had appeared to be in jeopardy, because changes that had gone out recently rediscarded law, by which Congress had tried to overrule Miranda 32 years ago, and because of the court's perceived hostility to the original decision.

The chief justice said, though, that the 1968 law, which replaced the Miranda warnings with a case-by-case test of whether a confession was voluntary, could be upheld only if the Supreme Court decided to overturn Miranda. But with Miranda having

Justices Antonin Scalia and Clarence Thomas cast the dissenting votes.

The decision overturned a ruling last year by the federal appeals court in Richmond, Va., which held that Congress was entitled to the last word because Miranda's presumption that a confession was not voluntary unless preceded by the warnings was not required by the Constitution.

The decision today — only 14 pages long — is Chief Justice Rehnquist's typically spare style — brought an abrupt end to one of the older episodes in the court's recent history, an intense and strangely delayed re-fighting of a previous generation's battle over the rights of criminal suspects. In *Arizona v. Arizona*, a case from the Warren Court, and Chief Justice Rehnquist, despite his record as an early and tenacious critic of the decision, evidently did not want its reputation to be an imprint of his own tenure.

There was considerable drama in the courtroom today as the chief justice announced that he would rule

The Book of Life

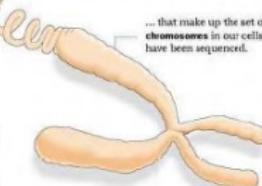
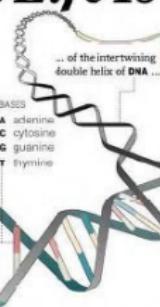
The three billion base pairs ...

BASE PAIRS
Runs between the strands of the double helix



BASES

A adenine
C cytosine
G guanine
T thymine



By ordering the base units, scientists hope to locate the genes and determine their functions.

The New York Times

Science Times

A SPECIAL ISSUE

■ Putting the genome to work.

■ Some information has already paid research dividends.

■ Two research methods, two results.

■ From Mendel to helix to genome.

■ More articles, charts and photos of the genome effort.

Section F

Francis S. Collins, head of the Human Genome Project, left, with J. Craig Venter, head of Celera Genomics, after the announcement yesterday that they had finished the first survey of the human genome.



Paul Haring/The New York Times

A SHARED SUCCESS

2 Rivals' Announcement Marks New Medical Era, Risks and All

By NICHOLAS WADE

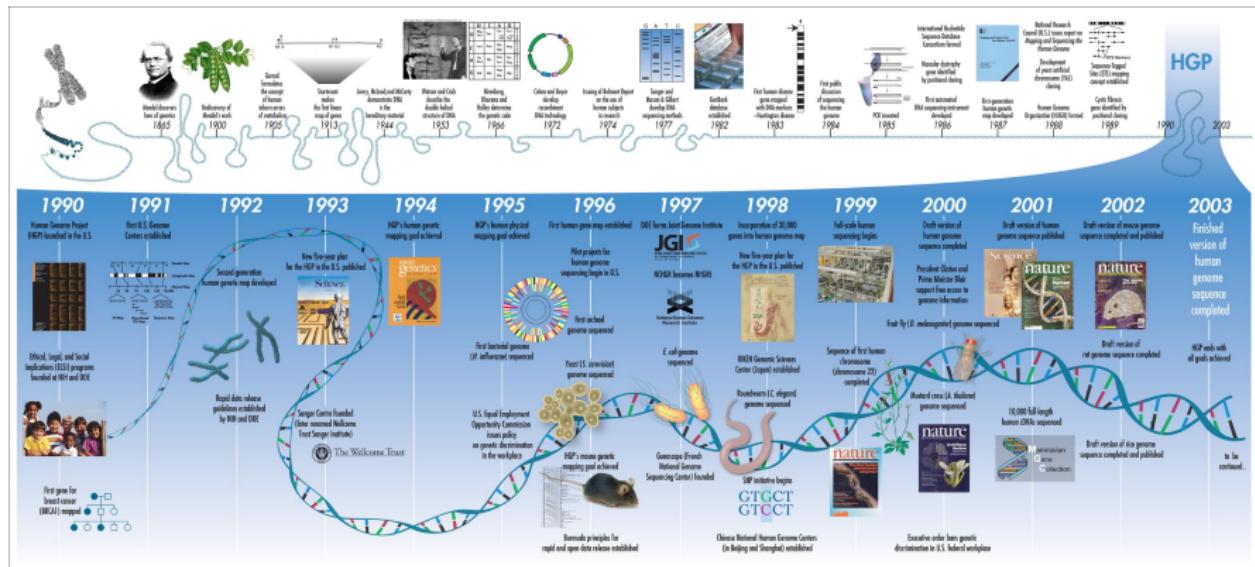
WASHINGTON, June 26 — In an achievement that represents a pinnacle of human self-knowledge, two rival groups of scientists said today that they had deciphered the hereditary script, the set of instructions that defines the human organism.

"Today we are learning the language in which God created us," President Clinton said at a White House ceremony attended by members of the two teams, Dr. James D. Watson, codiscoverer of the structure of DNA, and, via satellite, Prime Minister Tony Blair of Britain. [Excerpts, Page D8.]

The teams' leaders, Dr. J. Craig Venter, president of Celera Genomics, and Dr. Francis S. Collins, director of the National Human Genome Research Institute, praised each other's contributions and signaled a spirit of cooperation from now on, even though the two efforts will remain formally independent.

The human genome, the ancient script that has now been deciphered, consists of two sets of 23 strand DNA

The Human Genome roadmap



<https://www.davidstreams.com/mis-apuntes/human-genome-project/>

Evolution of post-Sanger sequencing technologies

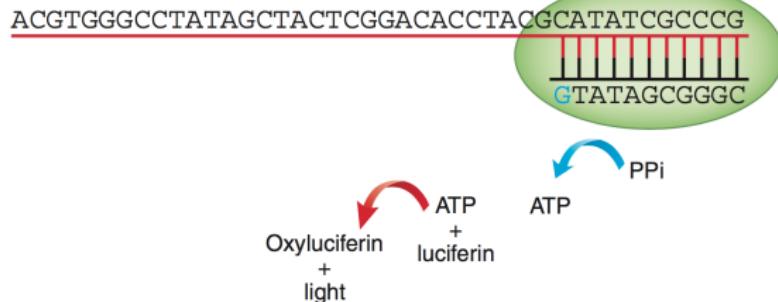
Evolution of sequencing technologies

- “Massively parallel” sequencing
- “High-throughput” sequencing
- “Ultra high-throughput” sequencing
- “Next generation” sequencing (NGS)
- “Second generation” sequencing

Evolution of sequencing technologies

- 2005: 454 (Roche)
- 2006: Solexa (Illumina)
- 2007: ABI/SOLiD (Life Technologies)
- 2010: Complete Genomics
- 2011: Pacific Biosciences
- 2010: Ion Torrent (Life Technologies)
- 2015: Oxford Nanopore Technologies

454 pyrosequencing

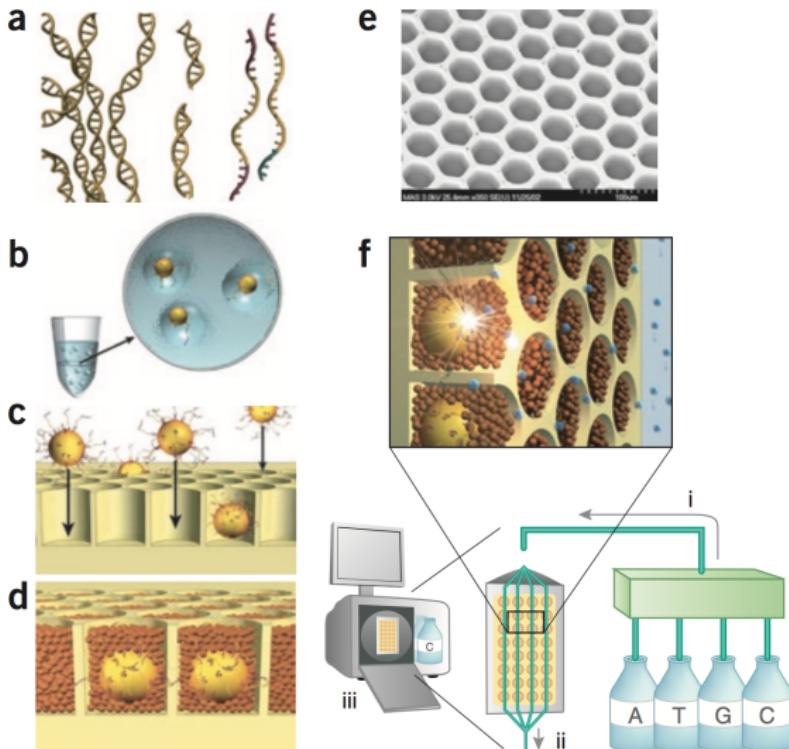


- ① Hybridize sequencing primer
- ② Add DNA polymerase, ATP sulfurylase, luciferase, apyrase & substrates (adenosine 5' phosphosulfate (APS) and luciferin)
- ③ Nucleotide incorporation catalyzes chain reaction that results in light
- ④ Add bases sequentially: add A, take a picture - did it flash? :: wash :: add T - did it flash? :: wash :: add G - did it flash? :: wash :: add C - did it flash? :: wash. Repeat ~500 times

<https://www.nature.com/nbt/journal/v26/n10/full/nbt1485.html>

454 pyrosequencing

- 1) Fragment DNA
- 2) Bind to beads, emulsion PCR amplification
- 3) Remove emulsion, place beads in wells
- 4) Solid phase pyrophosphate sequencing reaction
- 5) Scanning electron micrograph



<https://www.nature.com/nbt/journal/v26/n10/full/nbt1485.html>

454 sequencing: summary

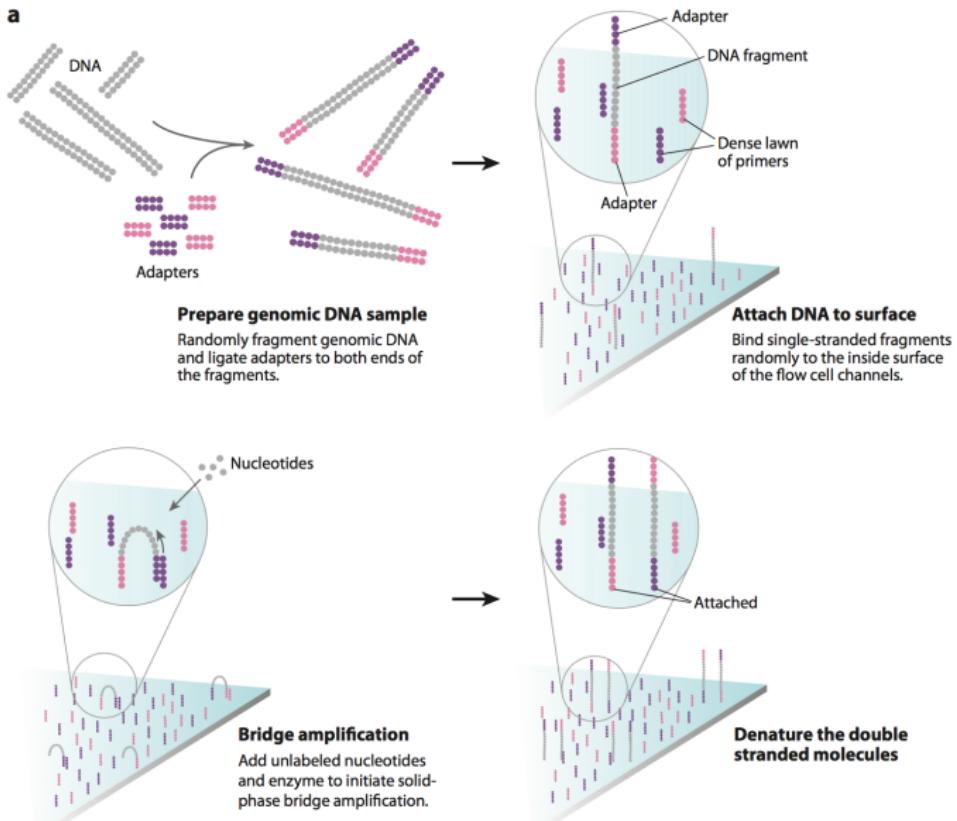
- First post-Sanger technology (2005)
- Used to sequence many microorganisms & Jim Watson's genome (for \$2M in 2007)
- Longer reads than Illumina, but much lower yield (~500bp)
- Rapidly outpaced by other technologies - now essentially obsolete

Solexa (Illumina) sequencing (2006)

- PCR amplify DNA fragments
- Immobilize fragments on a solid surface, amplify
- Reversible terminator sequencing with 4 color dye-labelled nucleotides

Video of Illumina sequencing, <http://www.youtube.com/watch?v=77r5p8IBwJk> (1.5m)

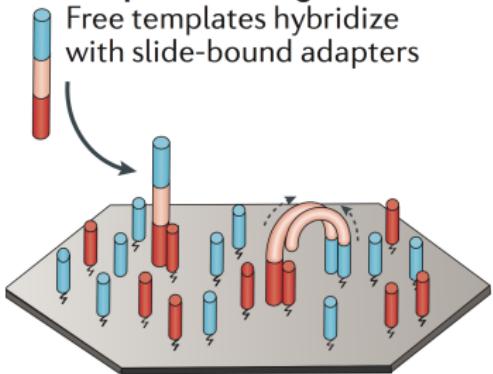
Solexa (Illumina) sequencing (2006)

a

Cluster amplification by “bridge” PCR

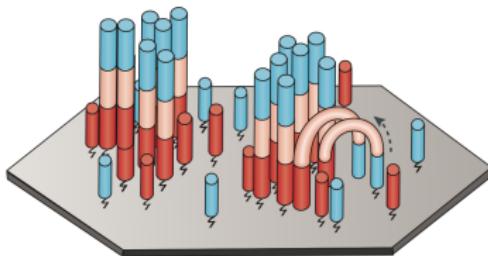
Template binding

Free templates hybridize with slide-bound adapters



Bridge amplification

Distal ends of hybridized templates interact with nearby primers where amplification can take place

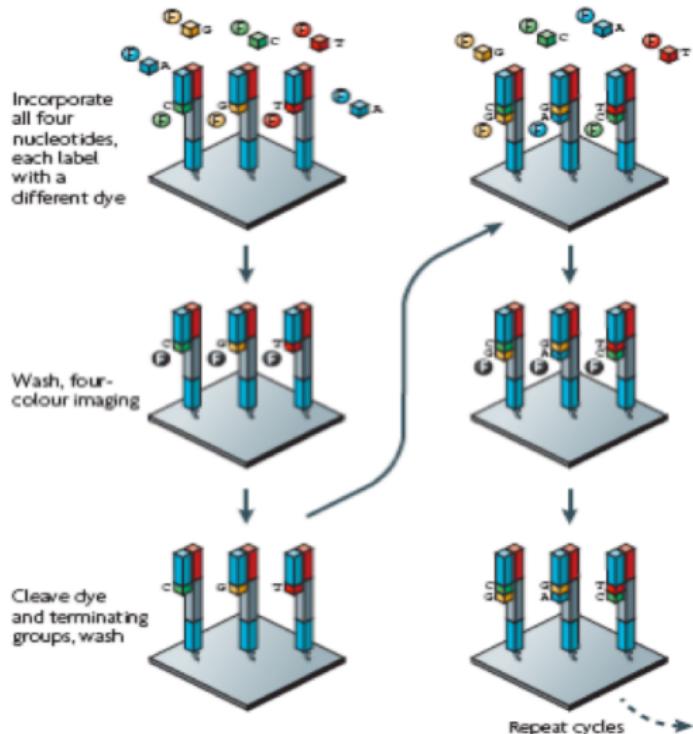


Cluster generation

After several rounds of amplification, 100–200 million clonal clusters are formed

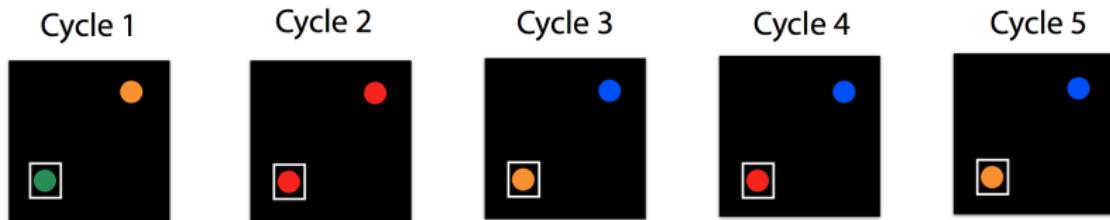
<https://binf.snipcademy.com/lessons/ngs-techniques/bridge-pcr>

Clonal amplification



Base calling

- 6 cycles with base-calling



"Base caller" software looks at this cluster across all images and "calls" the complementary nucleotides: **TACAC**, corresponding to the template sequence



TACAC is a "sequence read," or "read." Actual reads are usually 100 or more nucleotides long.

Illumina sequencers



- **Illumina HiSeq:** ~3 billion paired 100bp reads, ~600Gb, \$10K, 8 days (or “rapid run” ~90Gb in 1-2 days)
- **Illumina X Ten:** ~6 billion paired 150bp reads, 1.8Tb, <3 days, ~1000 / genome(\$\$), (or “rapid run” ~90Gb in 1-2 days)
- **Illumina NextSeq:** One human genome in <30 hours

<http://www.businesswire.com/news/home/20150112006333/en/Illumina-Expands-World%2E2%80%99s-Comprehensive-Next-Generation-Sequencing-Portfolio>

Solexa (Illumina) sequencing: summary

Advantages:

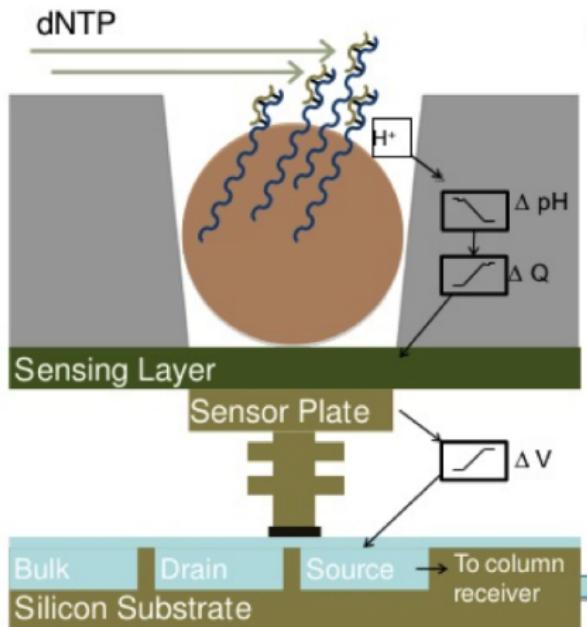
- Best throughput, accuracy and read length for any 2nd gen. sequencer
- Fast & robust library preparation

Disadvantages:

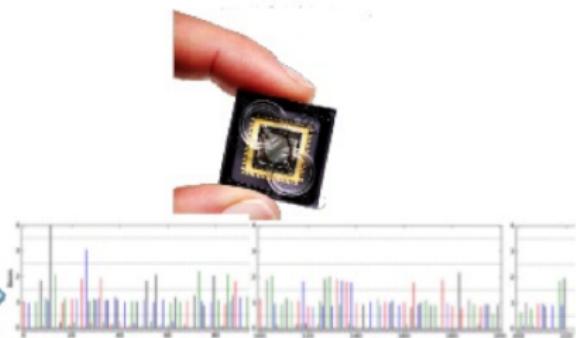
- Inherent limits to read length (practically, 150bp)
- Some runs are error prone

Video of Illumina sequencing <https://www.youtube.com/watch?v=womKfikWIxM> (5m)

ION Torrent-pH Sensing of Base Incorporation



- DNA → Ions → Sequence
 - Nucleotides flow sequentially over Ion semiconductor chip
 - One sensor per well per sequencing reaction
 - Direct detection of natural DNA extension
 - Millions of sequencing reactions per chip
 - Fast cycle time, real time detection



Platforms: Ion Torrent



PGM



Proton

- Three sequencing chips available:
 - 314 = up to 100 Mb
 - 316 = up to 1 Gb
 - 318 = up to 2 Gb
- 2-7 hour/run
- up to 400 bp read length
- 400kreads up to 5 Mreads

- Two human exomes (Proton 1 chip) or one genome (@20X-Proton 2 chip) per run
- Ion One Touch or Ion Chef preparatory modules
 - 2-4 hour/run
 - ~200 bp average read length
 - Proton 1 produces 60-80 Mreads \geq 50 bp

- Low substitution error rate, in/dels problematic, no paired end reads
- Inexpensive and fast turn-around for data production
- Improved computational workflows for analysis

Pacific Biosciences

HOW IT WORKS

DNA is copied by an enzyme in PacBio's machine

The DNA letters used to make the copy have been tagged to emit tiny flashes of colored light.

A camera can catch these tiny flashes thanks to a 50-nanometer hole that screens out other light.



- Long reads
 - Structural variant discovery
 - *De novo* genome assembly

<https://www.forbes.com/forbes/2009/1005/revolutionaries-science-genomics-gene-machine.html>

Pacific Biosciences: summary

Key Points:

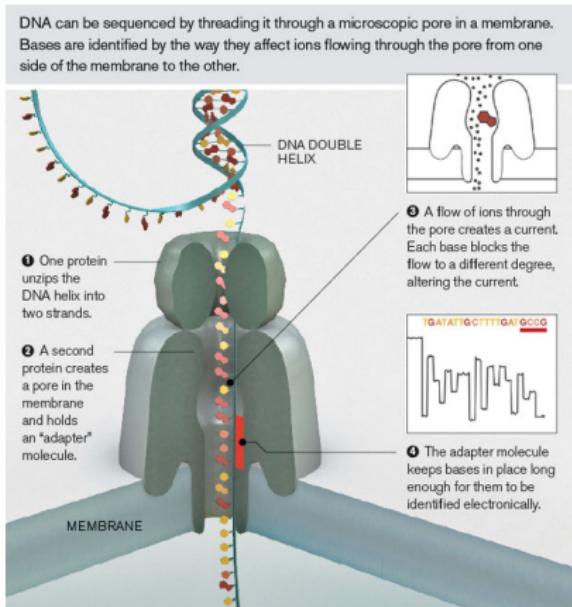
- 1 DNA molecule and 1 polymerase in each well (zero-mode waveguide)
- 4 colors flash in real time as polymerase acts
- Methylated cytosine has distinct pattern
- No *theoretical* limit to DNA fragment length

Caveats:

- Higher error rate (1-2%), but they are random
- Lower throughput, roughly 5 gigabases per run

Nanopore sequencing

- Nearly 30-years old technology



<http://www2.technologyreview.com/news/427677/nanopore-sequencing/>

Nanopore sequencing

- Nanopore sequencing with ONT is accurate and relatively reliable
- Current yield per run (“R9.4” chemistry): ~5 Gbp, 97% identity (i.e., 3% error rate)



<https://www.technologyreview.com/s/600887/with-patent-suit-illumina-looks-to-tame-emerging-british-rival-oxford-nanopore/>

Video of Ion Torrent chemistry, <http://www.youtube.com/watch?v=yVf2295JqUg> (2.5m)

Nanopore sequencing

- Key advantage - portability



Faria et al. *Genome Medicine* (2016) 8:97
DOI 10.1186/s13073-016-0356-2

Genome Medicine

COMMENT

Open Access



Mobile real-time surveillance of Zika virus in Brazil

Nuno Rodrigues Faria¹, Ester C. Sabino², Marcio R. T. Nunes^{3,4}, Luiz Carlos Junior Alcantara⁵, Nicholas J. Loman⁶ and Oliver G. Pybus¹

Video of Nanopore DNA sequencing technology <https://www.youtube.com/watch?v=CE4dW64x3Ts> (4.5m)

<https://phys.org/news/2016-08-nasa-dna-sequencing-space-success.html>

Nanopore for human genome sequencing

Nanopore sequencing and assembly of a human genome with ultra-long reads

Miten Jain^{1,13}, Sergey Koren^{2,13}, Karen H Miga^{1,13}, Josh Quick^{3,13}, Arthur C Rand^{1,13}, Thomas A Sasani^{4,5,13}, John R Tyson^{6,13}, Andrew D Beggs⁷, Alexander T Dilthey², Ian T Fiddes¹, Sunir Malla⁸, Hannah Marriott⁸, Tom Nieto⁷, Justin O'Grady⁹, Hugh E Olsen¹, Brent S Pedersen^{4,5}, Arang Rhie², Hollian Richardson⁹, Aaron R Quinlan^{4,5,10}, Terrance P Snutch⁶, Louise Tee⁷, Benedict Paten¹, Adam M Phillippy², Jared T Simpson^{11,12}, Nicholas J Loman³ & Matthew Loose⁸

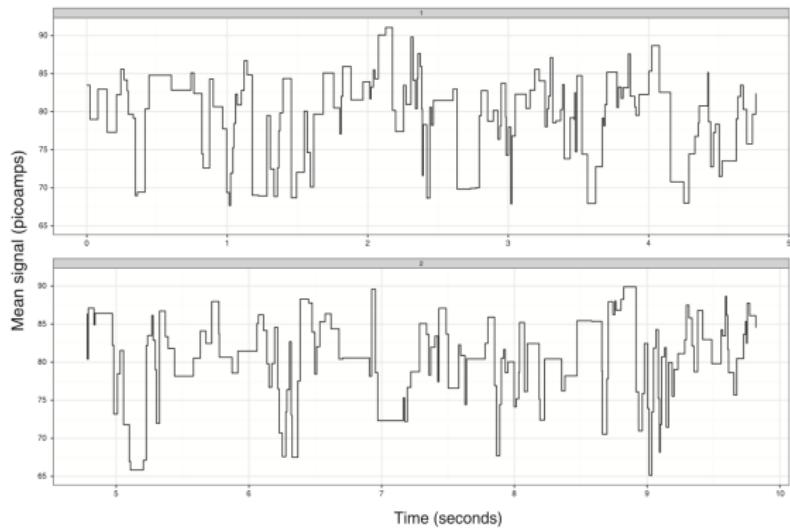
- Closes 12 gaps
- Phased the entire major histocompatibility complex (MHC) region, one of the most gene-dense and highly variable regions of the genome

Jain, Miten, Sergey Koren, Karen H Miga, Josh Quick, Arthur C Rand, Thomas A Sasani, John R Tyson, et al. "Nanopore Sequencing and Assembly of a Human Genome with Ultra-Long Reads." *Nature Biotechnology*, January 29, 2018.
<https://doi.org/10.1038/nbt.4060>.

<https://www.genengnews.com/gen-exclusives/first-nanopore-sequencing-of-human-genome/77901044>

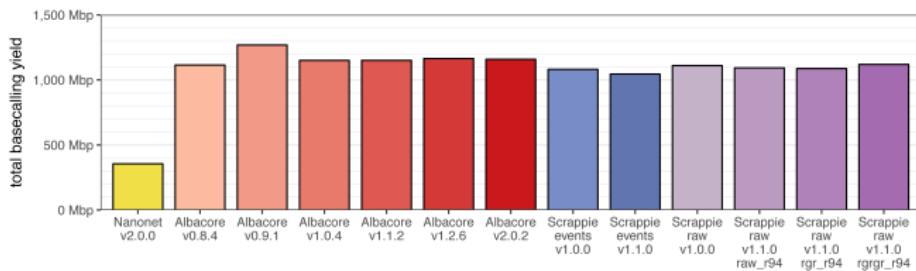
Nanopore technology

- Nanopore sequencing yields raw signals reflecting modulation of the ionic current at each pore by a DNA molecule.
- The resulting time-series of nanopore translocation, 'events', are base-called by proprietary software running as a cloud service.



Nanopore base callers

- Proper base calling is a paramount, as it defines whether the technology is good or bad.
- Nanonet, Albacore, Scrappie
- Most modern basecallers use neural networks.



<https://github.com/rrwick/Basecalling-comparison>

Nanopore analysis

- The resulting files for each sequenced read are stored in 'FAST5' format, an application of the HDF5 format.
- `poretools` - a toolkit for analyzing nanopore sequence data.

BIOINFORMATICS APPLICATIONS NOTE

Vol. 30 no. 23 2014, pages 3399–3401
doi:10.1093/bioinformatics/btu555

Sequence analysis

Advance Access publication August 20, 2014

Poretools: a toolkit for analyzing nanopore sequence data

Nicholas J. Loman^{1,*} and Aaron R. Quinlan^{2,*}

¹Institute of Microbiology and Infection, University of Birmingham, Birmingham B15 2TT, UK and ²Department of Public Health Sciences, University of Virginia, Charlottesville 22932, VA, USA

<https://github.com/arq5x/poretools>

<https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btu555>

PacBio vs. Oxford Nanopore sequencing

	PacBio ¹		Oxford Nanopore ²	
Instrument Specifications	RS II (P6-C4)	Sequel	MinION	PromethION
Average read length	10 – 15 kb	10 – 15 kb	Variable (up to 900 kb) ^{3,4}	*
Error rate	10 – 15 %	10 – 15 %	5 – 15 % ^{4,5}	*
Output	500 Mb – 1 Gb	5 Gb – 10 Gb	~5 Gb ⁴	*
# of reads	~50k	~500k	Variable (up to 1M) ^{6,7}	*
Instrument price/Access fee ^a	\$700k	\$350k	\$1000 ⁸	\$135k bundle ⁹
Run price	~\$400	~\$850	\$500-\$900 ⁷	*

<https://blog.genohub.com/2017/06/16/pacbio-vs-oxford-nanopore-sequencing/>

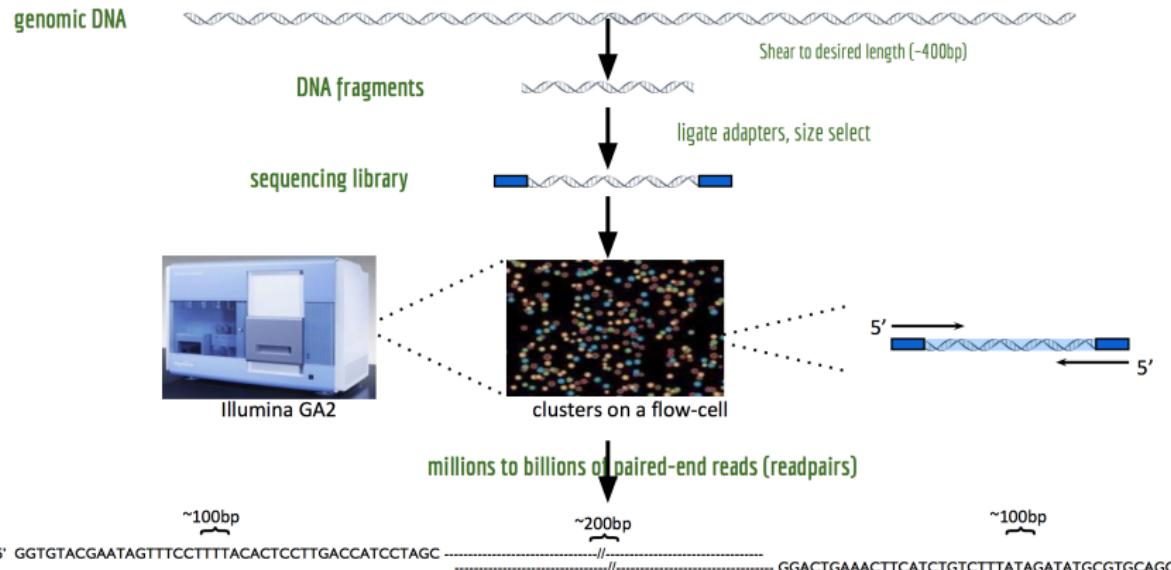
Single-end vs. paired-end sequencing

Single-end vs. paired-end sequencing

- Single-end sequencing: sequence one end of the DNA segment.
- Paired-end sequencing: sequence both ends of a DNA segments.
 - Result reads are “paired”, separated by certain length (the length of the DNA segments, usually a few hundred bps).
 - Paired-end data can be used as single-end, but contain extra information which is useful in some cases, e.g., detecting structural variations in the genome.
 - Modeling technique is more complicated.

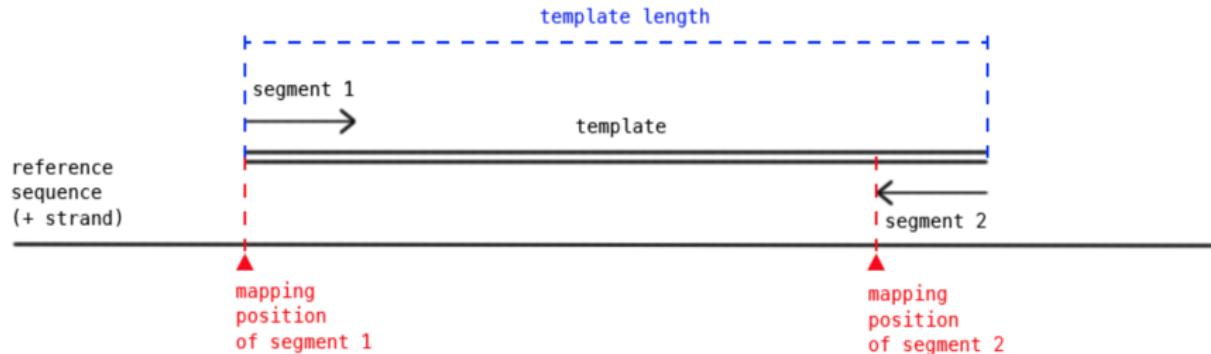
Paired-end sequencing - a workaround to sequence longer fragments

- Read one end of the molecule, flip, and read the other end
- Generate pair of reads separated by up to 500bp with inward orientation



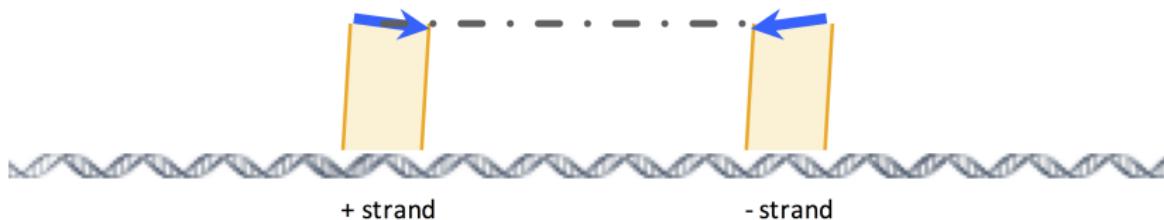
Templates and segments

- Template – DNA/RNA molecule which was subjected to sequencing – “Insert size” - template length
 - “Segment” – part of the template which was “read” by a sequencing machine (represented by a “sequencing read”)



Advantages of paired-end sequencing

- Alignment of the read pair to the reference genome gives coordinates describing where in the human genome the read pair came from



Sequencing applications

Applications

- NGS has a wide range of applications.
 - DNA-seq: sequence genomic DNA.
 - RNA-seq: sequence RNA products.
 - ChIP-seq: detect protein-DNA interaction sites.
 - Bisulfite sequencing (BS-seq): measure DNA methylation strengths.
 - A lot of others.
- Basically replaced microarrays with better data: greater dynamic range and higher signal-to-noise ratios.

DNA-seq (Whole-Genome sequencing)

- Sequence the untreated genomic DNA.
 - Obtain DNA from cells, cut into small pieces then sequence the segments.
- Goals:
 - Compare with the reference genome and look for genetic variants:
 - Single nucleotide polymorphisms (SNPs)
 - Insertions/deletions (indels),
 - Copy number variations (CNVs)
 - Other structural variations (gene fusion, etc.).
 - *De novo* assembly of a new genome.

Variations of DNA-seq

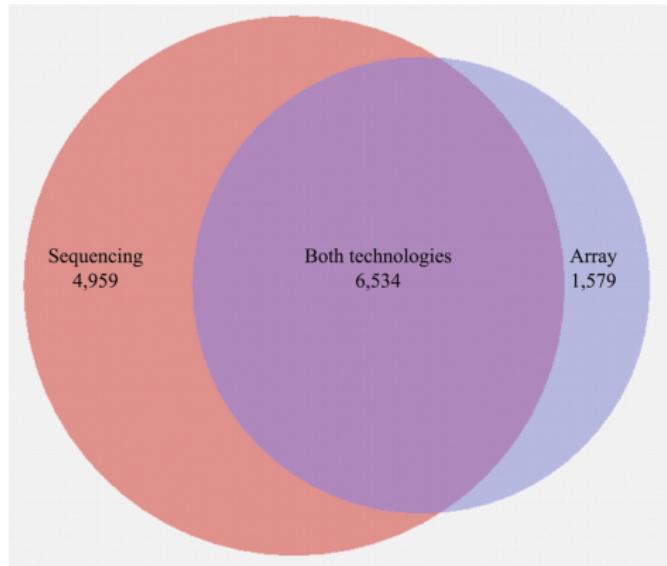
- Targeted sequencing, e.g., exome sequencing.
 - Sequence the genomic DNA at targeted genomic regions.
 - Cheaper than whole genome DNA-seq, so that money can be spent to get bigger sample size (more individuals).
 - The targeted genomic regions need to be “captured” first using technologies like microarrays.
- Metagenomic sequencing.
 - Sequence the DNA of a mixture of species, mostly microbes, in order to understand the microbial environments.
 - The goal is to determine number of species, their genome and proportions in the population.
 - *De novo* assembly is required. But the number and proportions of species are unknown, so it poses challenge to assembly.

RNA-seq

- Sequence the “transcriptome”: the set of RNA molecules.
- Goals:
 - Catalogue RNA products.
 - Determine transcriptional structures: alternative splicing, gene fusion, etc.
 - Quantify gene expression: the sequencing version of gene expression microarray.

Sequencing vs. microarray

- Very good agreement
- More information



<https://www.ncbi.nlm.nih.gov/pubmed/18550803>

ChIP-seq

- Chromatin-Immunoprecipitation (ChIP) followed by sequencing (seq): sequencing version of ChIP-chip.
- Used to detect locations of certain “events” on the genome:
 - Transcription factor binding.
 - DNA methylations and histone modifications.
- A type of “captured” sequencing. ChIP step is to capture genomic regions of interest.

What matters is what you feed into the sequencing machine

*Seq

I am maintaining an up-to-date annotated bibliography of ***Seq assays** (functional genomics assays based on high-throughput sequencing) on this page. The bibliography is also available in BibTeX. I also maintain a page with a list of reviews and survey papers about ***Seq**.

RNA structure

dsRNA-Seq: Qi Zheng et al., "Genome-Wide Double-Stranded RNA Sequencing Reveals the Functional Significance of Base-Paired RNAs in *Arabidopsis*," *PLoS Genet* 6, no. 9 (September 30, 2010): e1001141, doi:10.1371/journal.pgen.1001141.

FRAG-Seq: Jason G. Underwood et al., "FragSeq: Transcriptome-wide RNA Structure Probing Using High-throughput Sequencing," *Nature Methods* 7, no. 12 (December 2010): 995–1001, doi:10.1038/nmeth.1529.

SHAPE-Seq: (a) Julius B. Lucks et al., "Multiplexed RNA Structure Characterization with Selective 2'-hydroxyl Acylation Analyzed by Primer Extension Sequencing (SHAPE-Seq)," *Proceedings of the National Academy of Sciences* 108, no. 27 (July 5, 2011): 11063–11068, doi:10.1073/pnas.1106501108.

(b) Sharon Aviran et al., "Modeling and Automation of Sequencing-based Characterization of RNA Structure," *Proceedings of the National Academy of Sciences* (June 3, 2011), doi:10.1073/pnas.1106541108.

PARTE-Seq: Yue Wan et al., "Genome-wide Measurement of RNA Folding Energies," *Molecular Cell* 48, no. 2 (October 26, 2012): 169–181, doi:10.1016/j.molcel.2012.08.008.

PARS-Seq: Michael Kertesz et al., "Genome-wide Measurement of RNA Secondary Structure in Yeast," *Nature* 467, no. 7311 (September 2, 2010): 103–107, doi:10.1038/nature09322.

Structure-Seq: Yiliang Ding et al., "In Vivo Genome-wide Profiling of RNA Secondary Structure Reveals Novel Regulatory Features," *Nature* advance online publication (November 24, 2013), doi:10.1038/nature12756.

DMS-Seq: Silvi Rouskin et al., "Genome-wide Probing of RNA Structure Reveals Active Unfolding of mRNA Structures In Vivo," *Nature* advance online publication (December 15, 2013), doi:10.1038/nature12894.

Chromatin structure, accessibility and nucleosome positioning

Nucleo-Seq: Anton Valouev et al., "Determinants of Nucleosome Organization in Primary Human Cells," *Nature* 474, no. 7352 (June 23, 2011): 516–520, doi:10.1038/nature10002.

DNaseI-Seq: Gregory E. Crawford et al., "Genome-wide Mapping of DNase Hypersensitive Sites Using Massively Parallel Signature Sequencing (MPSS)," *Genome Research* 16, no. 1 (January 1, 2006): 123–131, doi:10.1101/gr.407410.

DNASEI-Seq: Jay R. Hesselberth et al., "Global Mapping of protein-DNA Interactions In Vivo by Digital Genomic Footprinting," *Nature Methods* 6, no. 4 (April 2009): 283–289, doi:10.1038/nmeth.1313.

Sono-Seq: Raymond K. Auerbach et al., "Mapping Accessible Chromatin Regions Using Sono-Seq," *Proceedings of the National Academy of Sciences* 106, no. 35 (September 1, 2009): 14926–14931, doi:10.1073/pnas.0905443106.

Hi-C-Seq: Erez Lieberman-Aiden et al., "Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome," *Science* 326, no. 5950 (October 9, 2009): 289–293, doi:10.1126/science.1181369.

ChiA-PET-Seq: Melissa J. Fullwood et al., "An Oestrogen-receptor- α -bound Human Chromatin Interactome," *Nature* 462, no. 7269 (November 5, 2009): 58–64, doi:10.1038/nature08497.

FAIRE-Seq: Hironori Waki et al., "Global Mapping of Cell Type-Specific Open Chromatin by FAIRE-seq Reveals the Regulatory Role of the NF1 Family in Adipocyte Differentiation," *PLoS Genet* 7, no. 10 (October 20, 2011): e1002311,

NOme-Seq: Theresa K. Kelly et al., "Genome-wide Mapping of Nucleosome Positioning and DNA Methylation Within Individual DNA Molecules," *Genome Research* 22, no. 12 (December 1, 2012): 2497–2506, doi:10.1101/gr.143008.112.

ATAC-Seq: Jason D. Buenrostro et al., "Transposition of Native Chromatin for Fast and Sensitive Epigenomic Profiling of Open Chromatin, DNA-binding Proteins and Nucleosome Position," *Nature Methods* advance online publication (October 6, 2013), doi:10.1038/nmeth.2688.

Protein-DNA binding

ChIP-Seq: David S. Johnson et al., "Protein-DNA Interactions," *Science* 302, doi:10.1126/science.1141713.

ChIP-Seq: Tarjei S. Mikkelsen et al., "State in Pluripotent and Lineage-differentiating Stem Cells," *Nature* 448, no. 7153 (August 2, 2007): 553–560, doi:10.1038/nature05933.

HITS-Flip-Seq: Razvan Nutiu et al., "Landscapes on a High-throughput Scale," *Biochemistry* 29, no. 7 (July 2000): 1663–1670.

Chip-exo-Seq: Ho Sung Rheu et al., "Genome-wide Protein-DNA Interaction Resolution," *Cell* 147, no. 6 (December 12, 2011): e1002610, doi:10.1016/j.cell.2011.11.013.

PB-Seq: Michael J. Guertin et al., "Probing Transcription Factor Binding Interactions," *Cell* 149, no. 4 (March 29, 2012): e1002610, doi:10.1016/j.cell.2012.03.010.

AHT-ChIP-Seq: Sarah Aldridge et al., "Automated Robotic Protocol for High-Throughput ChIP-Seq," *Genome Biology* 14, no. R124, doi:10.1186/gb-2013-14-R124.

Protein-protein interaction

PDZ-Seq: Andreas Ernst et al., "Proteins at the PDZ-Seq: Protein-Protein Interactions Analyzed by High-Throughput Sequencing," *Molecular BioSystems* 10, doi:10.1039/c0mb00061b.

Small molecule-protein interaction

PD-Seq: Daniel Arango et al., "Molecular Targets of the Flavonoid Revealed by the Combinatorial Approach," *Proceedings of the National Academy of Sciences* 110, no. 46 (June 11, 2013): E2153–E2162, doi:10.1073/pnas.1308277110.

Small molecule-DNA interaction

Chem-Seq: Lars Anders et al., "Chemical-Seq: Chemical Sequencing of Small Molecules," *Nature Biotechnology* 31, no. 10 (October 2013): 901–906, doi:10.1038/nbt.2776.

Evolution of sequencing technologies

Technology	Brief description
ChIP-seq	Locate protein-DNA interaction or histone modification sites.
CLIP-seq	Map protein-RNA binding sites
RNA-seq	Quantify expression
SAGE-seq	Quantify expression
RIP-seq	capture TF-bound transcripts
GRO-seq	evaluate promoter-proximal pausing
BS-seq	Profile DNA methylation patterns
MeDIP-seq	Profile DNA methylation patterns
TAB-seq	Profile DNA hydroxyl-methylation patterns
MIRA-seq	Profile DNA methylation patterns
ChiRP-seq	Map lncRNA occupancy
DNase-seq	Identify regulatory regions
FAIRE-seq	Identify regulatory regions
FRT-seq	Quantify expression
Repli-seq	Assess DNA replication timing
MNase-seq	Identify nucleosome position
Hi-C	Infer 3D genome organization
ChIA-PET	Detect long distance chromosome interactions
4C-seq	Detect long distance chromosome interaction
Sono-seq	Map open-chromatin sites
NET-seq	determine <i>in vivo</i> position of all active RNAP complexes.
NA-seq	Map Nuclease-Accessible Sites

Developments in next generation sequencing: instruments, read lengths, throughput.

