# Next Generation Sequencing (NGS) and its platforms

Phuc-Loi Luu, PhD
Luu.p.loi@googlemail.com
loi.lp@pacificinformatics.com.vn

23.04.2023

# Agenda 23 April 2023

| Time | Content | Presenter |
|---|---|---|
| 19:00-19:15 | From microarray to sequencing technology | Loi |
| 19:15-19:30 | Sanger sequencing | Phuoc (Sisc group) |
| 19:30-20:00 | Illumina platform | Huong (Biomedic JSC) |
| 20:00-20:30 | MGI platform | Van (Research Instruments) |
| 20:30-21:00 | Ion Torrent platform | Phuoc (Sisc group) |
| 21:00-21:30 | Bioinformatics analysis of NGS | Loi |
| 21:30-22:00 | Discussion | All |

# Module II: DNA-seq

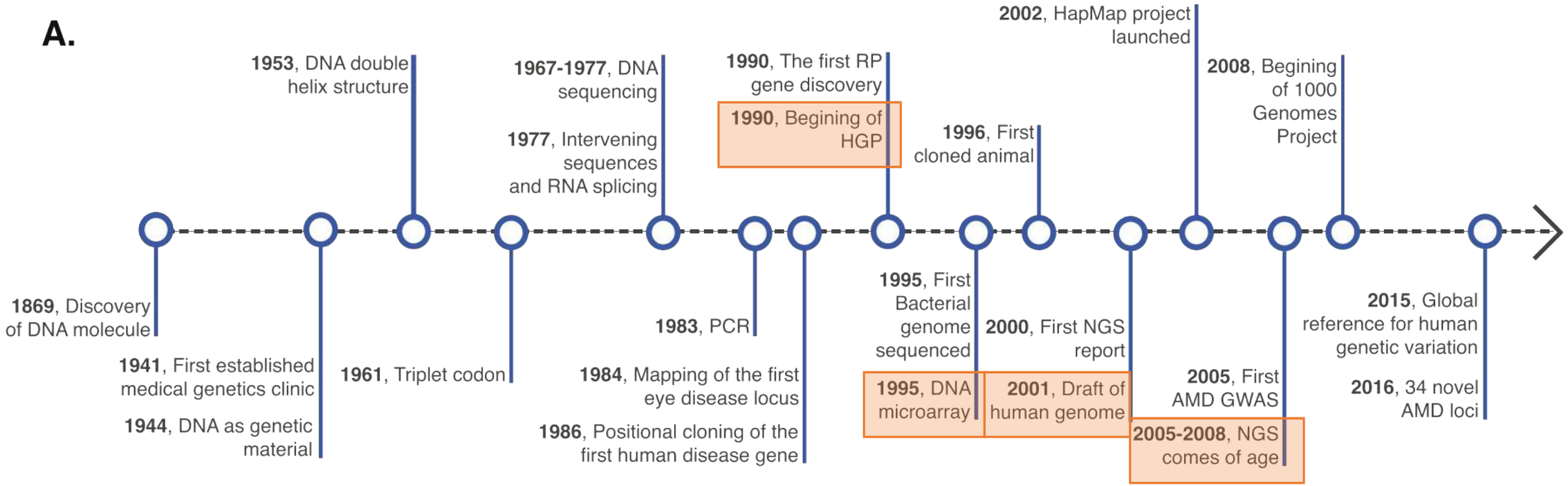
| | | | |
|---|---|---|---|
| 12 | 1) Human genome structure, function and clinical considerations https://do<br>2) Human reference genome and gene annotation https://docs.google.com/ | 04/09/2023 | Loi |
| 13 | Introduction to GWAS https://drive.google.com/file/d/1dPWxG3N_AB31mT | 04/16/2023 | Loi and Hoang |
| 14 | Introduction to NGS Technology | 04/23/2023 | Loi and more |
| 14' | 30/04 Holiday Break | | |
| 15 | DNA-seq: Raw Data Processing and Alignment/Mapping | 05/07/2023 | Duy |
| 16 | DNA-seq: Variant Calling and Variant Filtering | 05/14/2023 | Minh and Duy |
| 17 | DNA-seq: Variant Annotation and Variant Annotation databases | 05/21/2023 | Minh |
| 18 | Cancer Gene Panel DNA testing | 05/28/2023 | Thinh |
| 19 | Calling Copy Number Variant (CNV) and Clinical application | 06/04/2023 | Giang, Nhu |
| 20 | Calling Structural Variant (SV) and Clinical application | 06/11/2023 | Du, Nhu |
| 21 | Introduction to DNA denovo assembly | 06/18/2023 | Tan |
| 22 | Introduction to Microbiomics/Metagenomics | 06/25/2023 | Tien |
| 23 | Exam | 07/02/2023 | Loi |
| | | | |
| | Link youtube | | |
| | https://www.youtube.com/@vpivnpathoinformatics8930 | | |

Short read

Long read:

1) Pacbio
2) Oxford NanoPore

# Timeline of human genetics and genomic technologies




http://dx.doi.org/10.1016/j.preteyeres.2016.06.001

Age-related macular degeneration (AMD)
Human Genome Project (HGP)

From genotyping with DNA Microarray to SNP discovery with High Throughput Sequencing

# DNA Microarray (dChip)

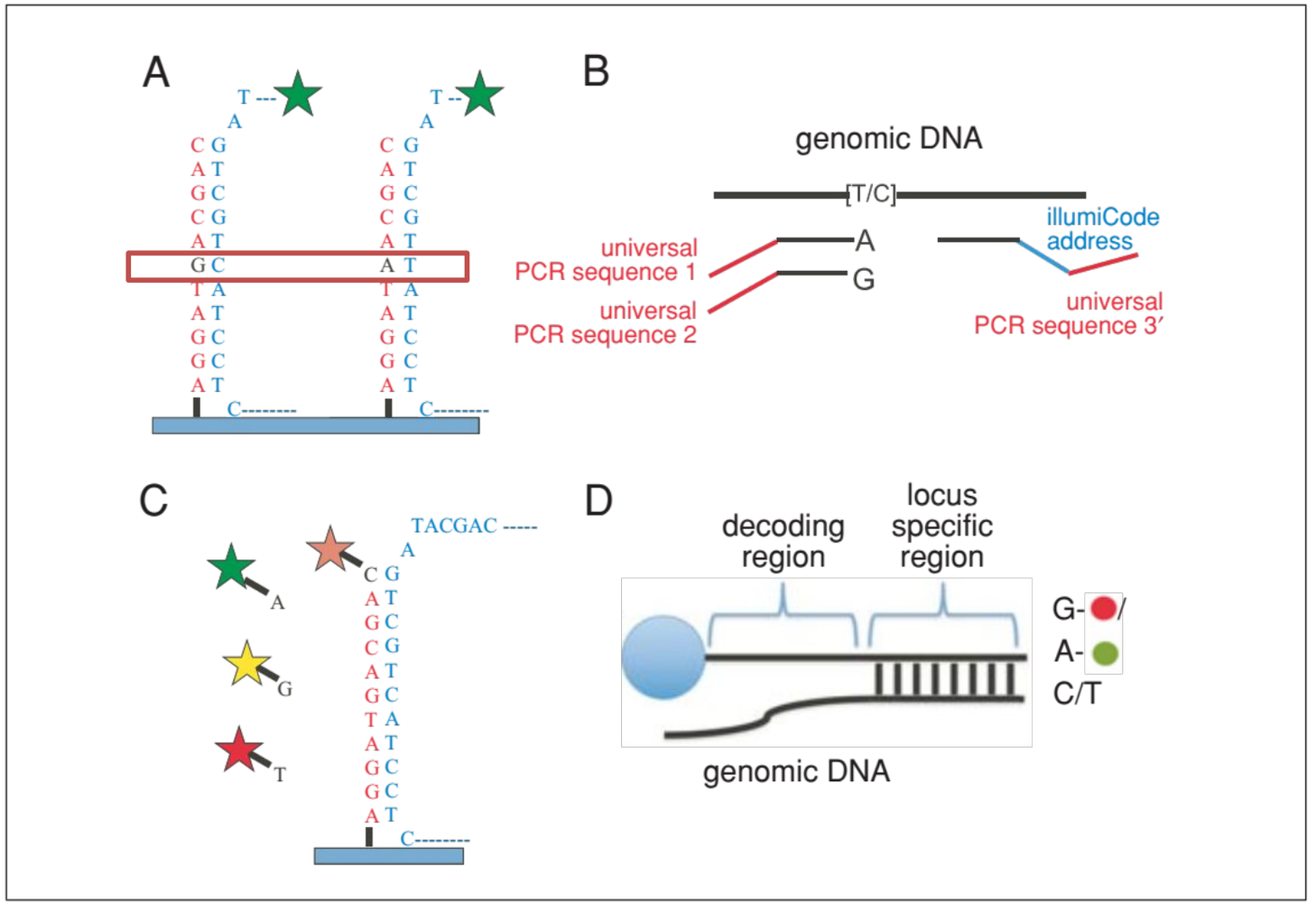


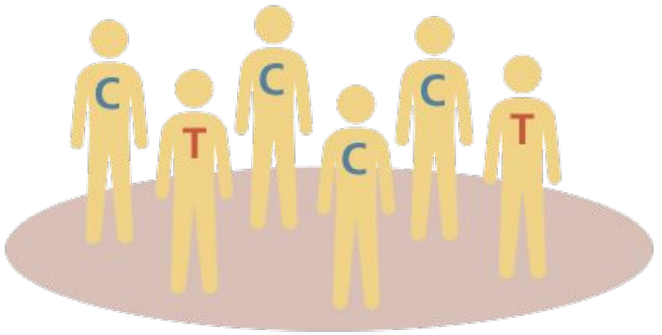

**Figure 22.1.4 SNP detection strategies for arrays.**
**(A) Allele discrimination by hybridization.** Oligos that are complementary to each allele are placed on the array and labeled genomic DNA is hybridized to the array. The variant position is placed in the center of the oligo (typically 25 bp on Affymetrix arrays), as this position has the greatest effect on hybridization. Typically, multiple array positions are used for each allele to improve signal-to-noise.
**(B) Illumina's "Golden Gate Assay."** Two allele-specific oligos are each tailed with a different universal primer (1 and 2) and hybridized in solution to genomic DNA. A third oligo that is complementary to the same locus is tailed with a "barcode" sequence and a third universal primer (3). Polymerase is used to extend the allele-specific primers across the genomic sequence, and the extended products are ligated to the third oligo. PCR is performed using primers complementary to universal sequences 1, 2, and 3. The PCR primers complementary to the universal sequences 1 and 2 are labeled with a unique fluorophore. The barcode sequence on the third oligo allows the PCR product to be uniquely detected on an array containing oligos complementary to the barcode sequence. The use of multiple barcodes (one for each locus of interest) allows the assay to be multiplexed to sample many loci.
**(C) Arrayed primer extension (APEX)**. In this assay, the array contains DNA oriented with the 5_x0003_ end attached to the array and the 3_x0003_ end stopping one nucleotide short of the SNP. Genomic DNA is fragmented and hybridized to the array, and the oligo on the array is extended in a single nucleotide dye terminator sequencing reaction.
**(D) Illumina's Infinium assay**. This assay is similar to the APEX assay except that the oligo to be extended is on a bead and the single nucleotide that is added is labeled with a nucleotide-specific hapten as opposed to a fluorophore. The haptens are then detected by staining with fluorescently labeled proteins that bind each hapten.

DOI: 10.1002/0471142727.mb2201s101

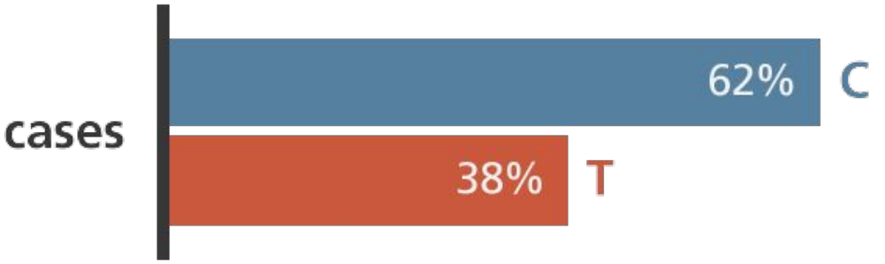
# Experiment design

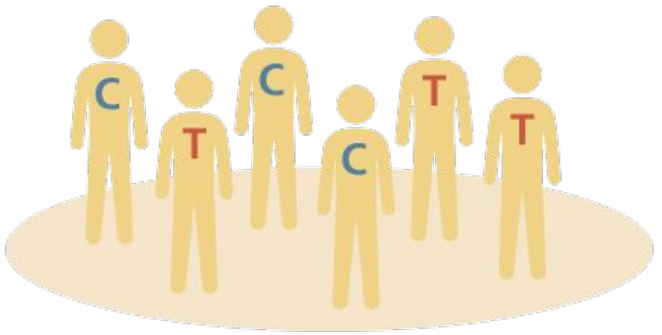
| | Microarrays | Whole genome sequencing |
|---|---|---|
| **Variant types** | SNPs, CNVs | ~All |
| **Loci probed** | 500k - 5m | ~All (3b) |
| **% variation captured** | 60%-85% of MAF>1% ($r^2$>0.8) | ~90%+ of genome |
| **Cost per sample (approx, 2014)** | $100-$500 | $1000-5000 |
| **Data file size** | <100Mb | 100Gb |
| **Data storage cost** (Amazon cloud 2014, 1000 samples) | $50 / year | $50,000 / year |

# Example of heart disease




**cases** (n=1,000)
people with heart disease




**controls** (n=1,000)
people without heart disease


https://www.yourgenome.org/stories/genome-wide-association-studies

# GWAS ON MIGRAINE (GORMLEY ET AL. 2016) (1/3)

- 60,000 cases and 315,000 controls from 22 studies
- 38 loci with convincing association
- Highlights vascular system

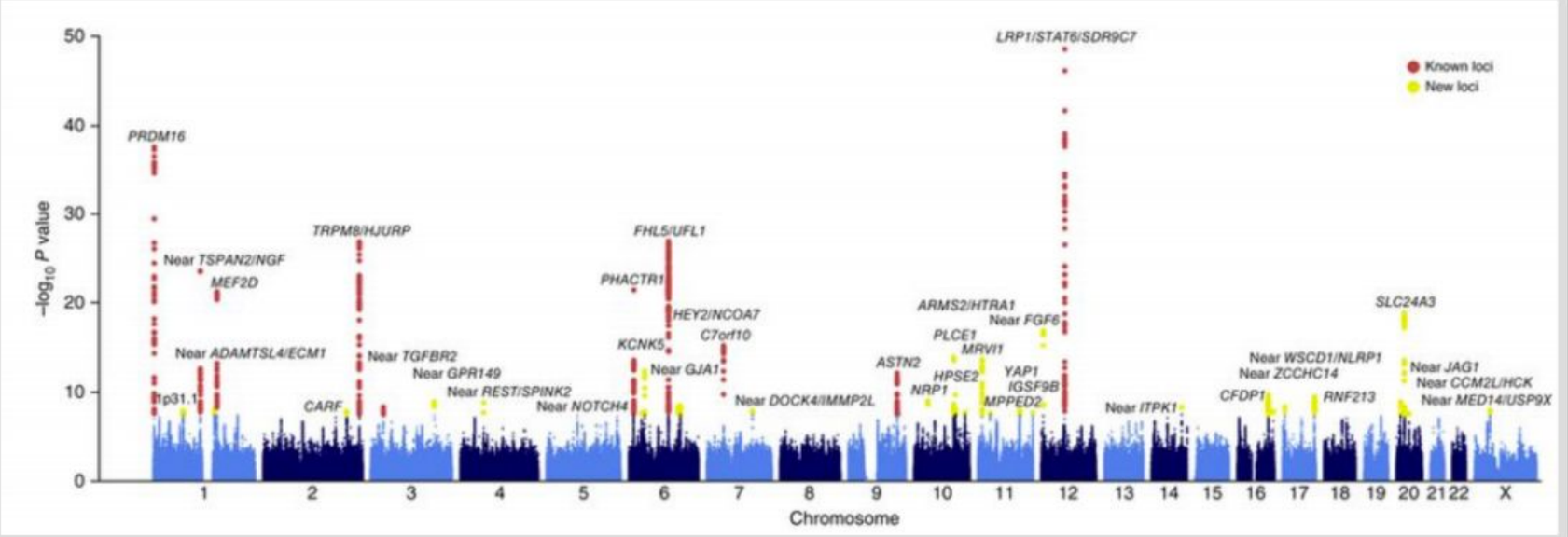
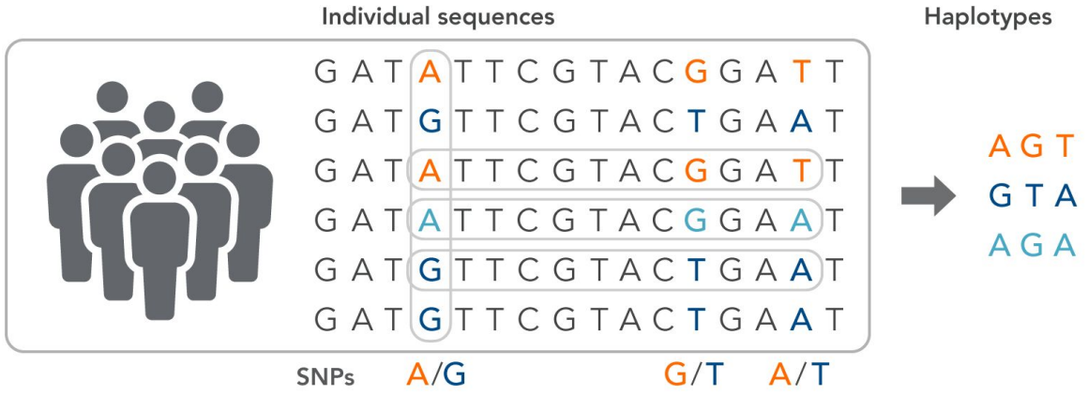



Fig. 1 of Gormley et al.
Manhattan plot of results.

# Whole genome sequencing (WGS), Whole Exome Sequencing (WES) and Target Sequencing with NGS

One individual


Multi individual

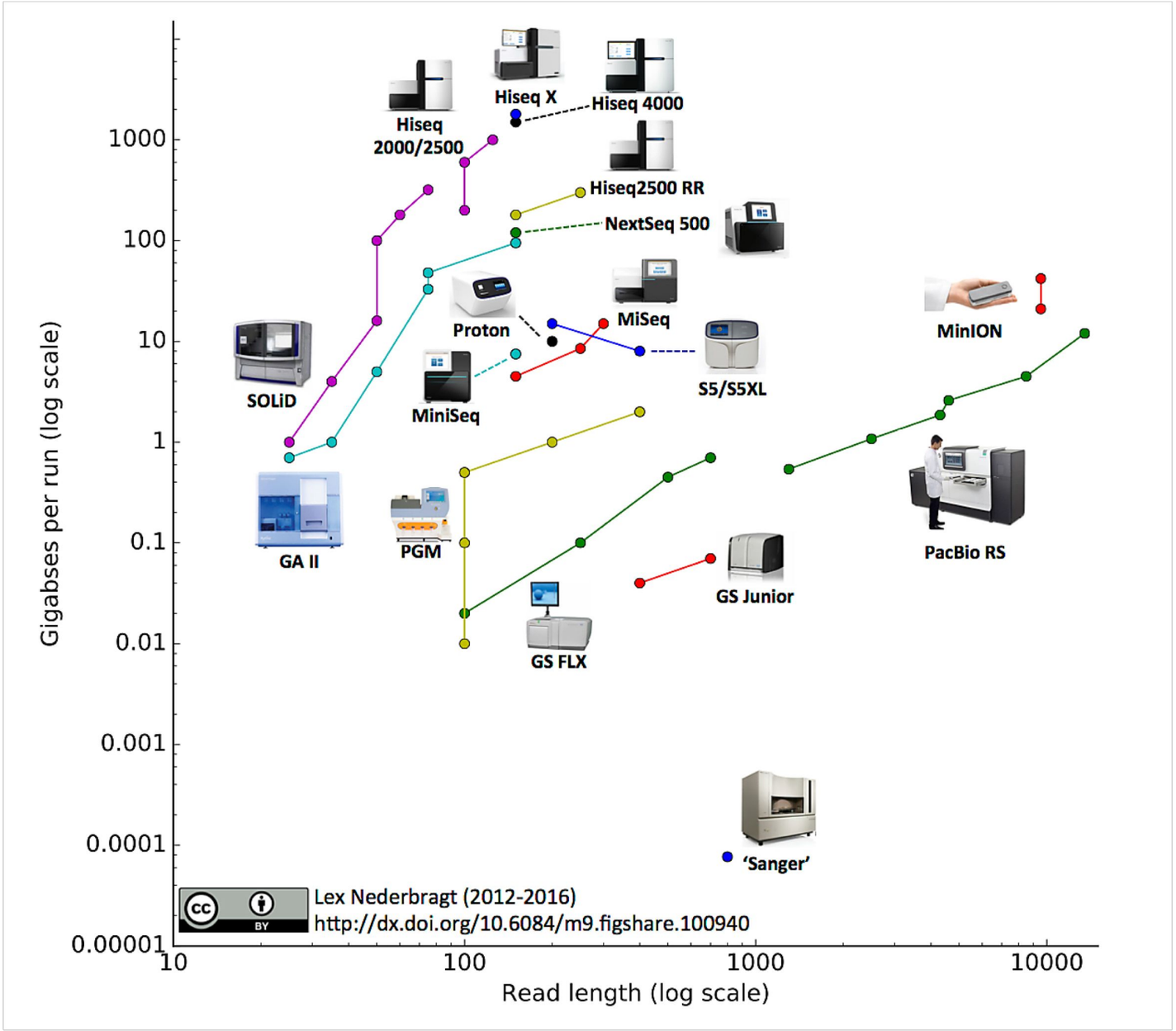



https://online.stat.psu.edu/stat555/node/106/

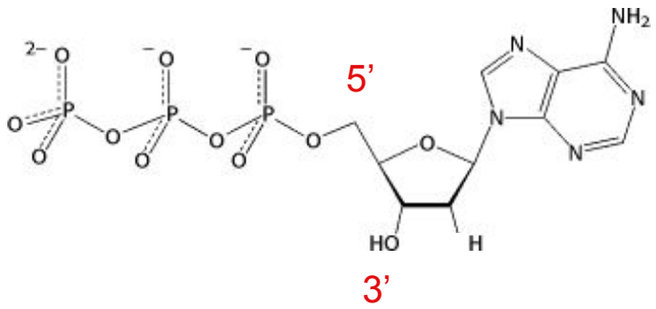
# Four Fundamentally Different Approaches to DNA Sequencing

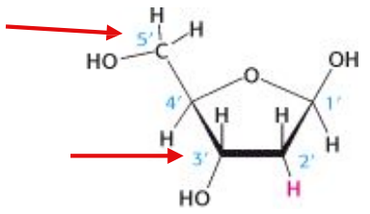
- Chemical degradation of DNA
  - Maxam-Gilbert
  - Obsolete
- Sequencing by synthesis ("SBS")
  - Uses DNA polymerase in a primer extension reaction
  - Most common approach
  - Fred Sanger developed it ("Sanger sequencing")
  - Illumina, Pacific Biosciences (being bought by Illumina), Ion Torrent, 454
- Ligation-based
  - Sequencing using short probes that hybridize to the template
  - SOLiD, BGI-Seq (Complete Genomics)
- Nanopore
  - Inferring sequence by change in electrical current as ssDNA is pulled though a nanopore
  - Oxford Nanopore, NABsys, Genia

# Throughput and Read Length in Sequencing



# 5’ and 3’




Base plus sugar

“nucleoside”

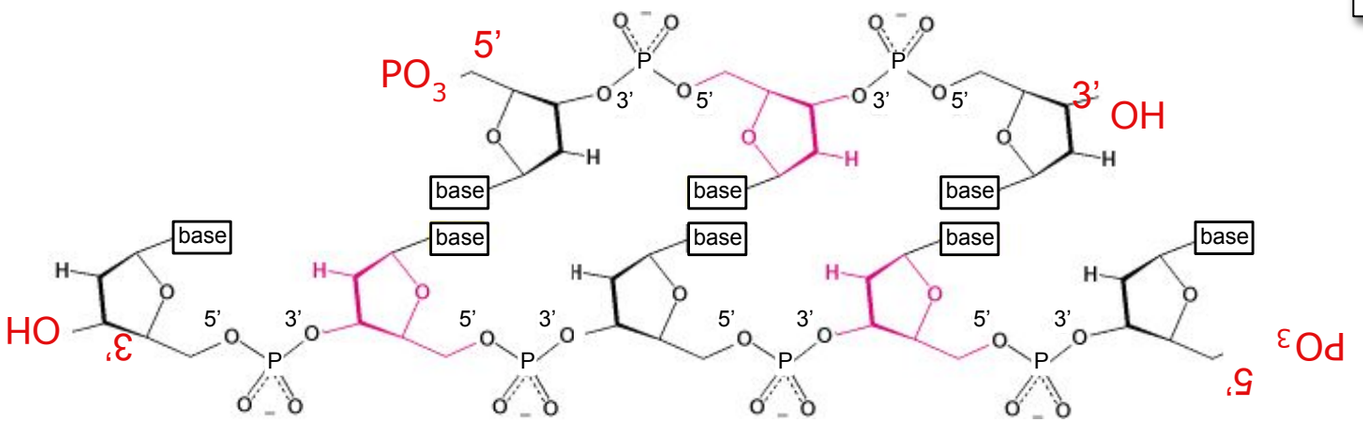
| | |
|---|---|
| Adenine | Adenosine |
| Guanine | Guanosine |
| Cytosine | Cytidine |
| Thymine | Thymidine |

in DNA: “deoxyadenosine”

plus triphosphate

“deoxynucleotide”

“2’-deoxyadenosine 5’-triphosphate” = dATP




Antiparallel

If I throw in DNA polymerase and free nucleotide, which end gets extended?

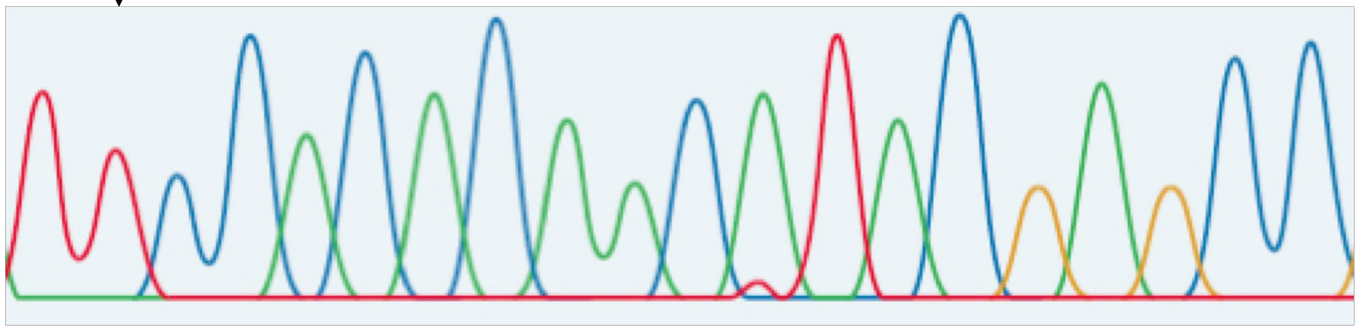
Adapted From Berg et al: Biochemistry 5th ed. Freeman+Co, 2002

# Fluorescent Sanger Sequencing Trace

**Lane signal**

(Real fluorescent signals from a lane/capillary are much uglier than this).
Various algorithms to boost signal/noise, correct for dye-effects, mobility differences, etc., generates the 'final' trace (for each capillary of the run)

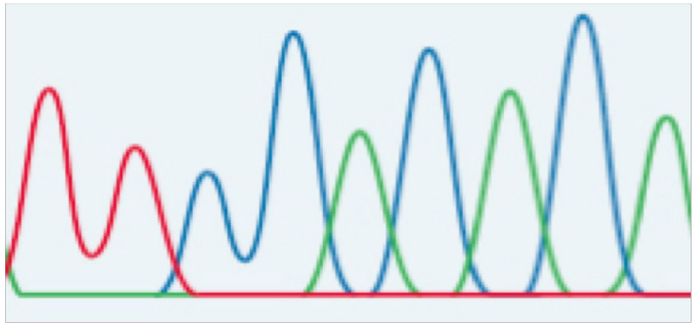
**Trace**



Recombinant DNA: Genes and Genomes. 3rd Edition (Dec06). WH Freeman Press.

# Sanger Base Calling

Low quality over here…



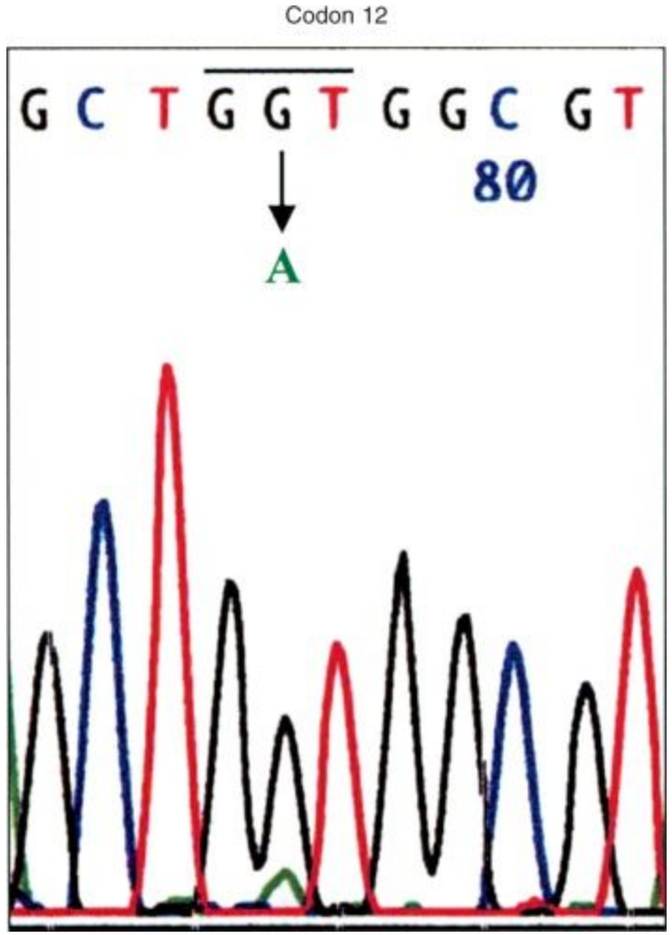
…low quality over there

Base Caller
(Phred)

```
... 44  45  46  47  48  49  50  51  52  53  54  55 ... 718 719 720 ...
...N    A   G   C   G   T   T   C   C   G   C   G ...  A   N   N ...
...0    3  20  25  40  88  95  99  99  99  99  99 ... 10   0   0 ...
```

Quality score = -10 * log(probability of error) or
P=10^-Q/10  For Q20, probability of error = 1/100
For Q99, probability of error ~$10^{-10}$

# Sanger Base Calling

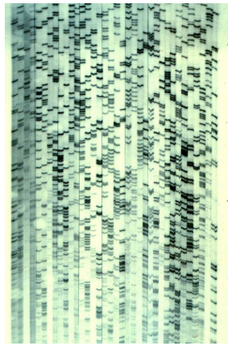


# Phred: *The* base-calling program for ABI sequencing

- Algorithm based on ideas about what might go wrong in a sequencing reaction and in electrophoresis
- Tested the algorithm on a huge dataset of "gold standard" sequences
  (finished human and *C. elegans* sequences generated by highly-redundant sequencing)
- Compared the results of phred with the ABI Basecaller
- Phred was considerably more accurate (40-50% fewer errors), particularly for indels and particularly for the higher quality sequences

(Ewing et al., 1998, *Genome Research* **8**: 175-185; Ewing and Green 1998, *Genome Research* **8**: 186-194)

# Progress of Sanger Sequencing Technology



Radioactive polyacrylamide slab gel

Low throughput, labor intensive

AB slab gel sequencers (370, 373, 377)

Fluorescent sequencing 1990-1999

6 runs/day

96 reads/run

500 bp/read

288,000 bp/day

AB capillary sequencers (3700, 3730)

1998-now

24 runs/day

96 reads/run

550 – 1,000 bp/read

1-2 million bp/day

~1,000-fold increase in throughput since 1985 accomplished by incremental improvements of the same underlying technology

2nd Generation Sequencing Technologies have >1e6x more throughput than 3730

# Whole Genome Sequencing

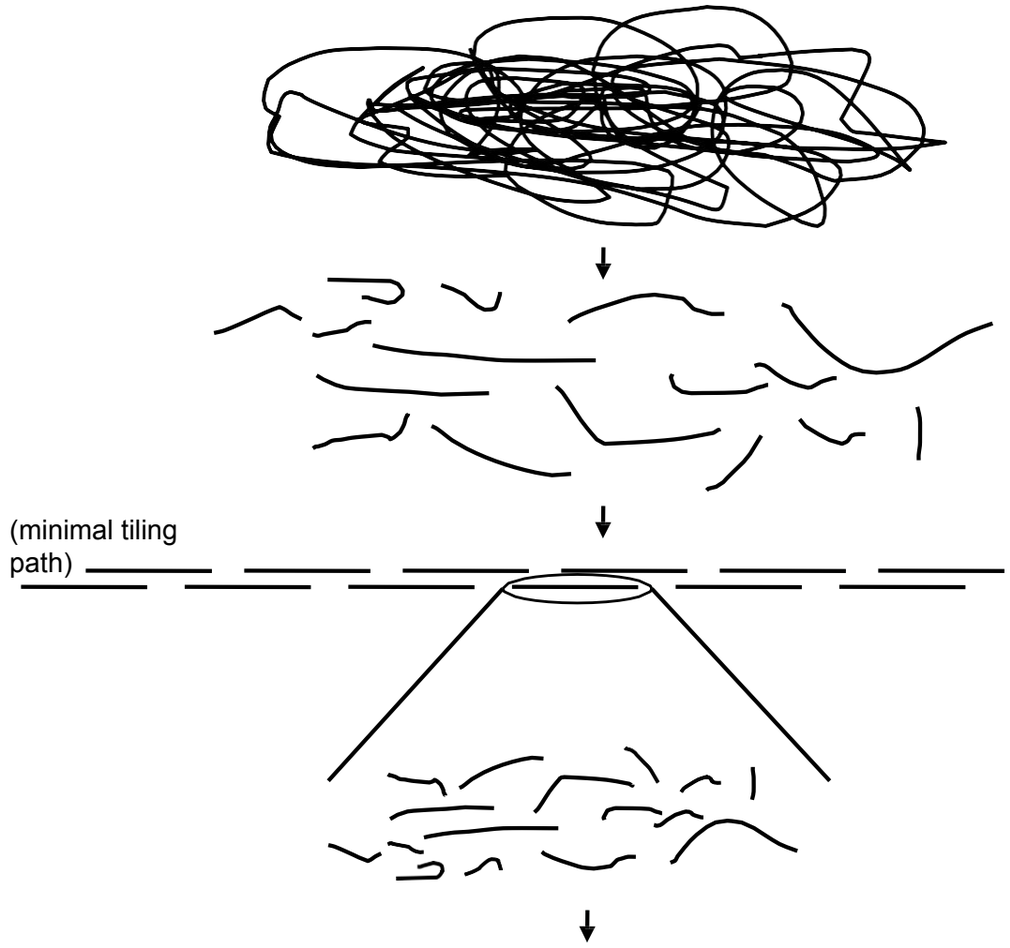
- Two main challenges:
  - Getting sufficient “coverage” of the genome
    - A function of read length, number of reads, complexity of library, and size of genome
  - Assembling the sequence reads into a complete genome
    - A function of coverage, and repeat size (relative to read lengths) and repeat frequency

# Overcoming repeats

- Most problematic when:
  - Repeats are longer than read lengths
  - Repeats are present in many copies
- Recognize based on coverage
- Resolve with longer range continuity information:
  - Paired-end reads
  - Multiple insert size libraries
    - Plasmids
    - Fosmids
    - BAC ends
    - Other tricks (which I'll come to later)

# Whole Genome Sequencing Approaches

## Hierarchical Shotgun Approach




Genomic DNA

BAC library

Organized, Mapped Large Clone Contigs

Shotgun Clones

GAAATATGTTCTTGTAATTTAAGCTGACACTCCTAATTTAGCTCTTGTCCTCTACTGAGTCTACCTAATTATATGTATGGATTGACTTGG

AGCTCTTGTCCTCTACTGAGTCTACCTAATTATATGTATGGATTGACTTGGTGTTTTCTCTTTTTCTTAAATAGTAATGCAGAAAGCCTGGAGA

GAGAG

Reads

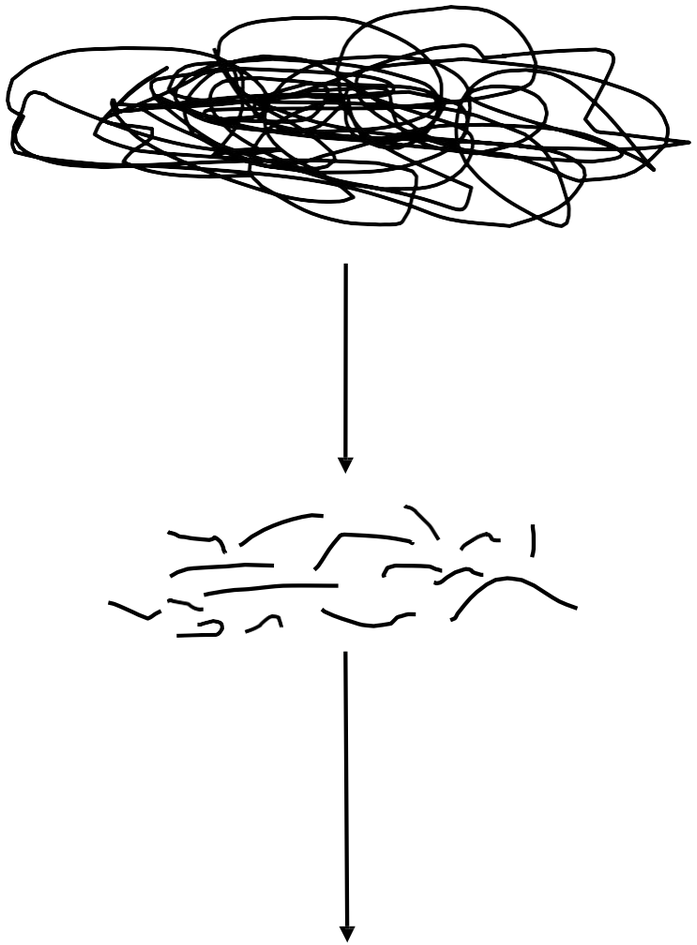
GAAATATGTTCTTGTAATTTAAGCTGACACTCCTAATTTAGCTCTTGTCCTCTACTGAGTCTACCTAATTATATGTATGGATTGACTTGGTGTTTTCTCTTTTTCTTAAATAGTAATGCAGAAAGCCTGGAGA

GAGAG

Assembly

# Whole Genome Sequencing Approaches

Shotgun Approach



Genomic DNA

Shotgun Clones

GAAATATGTTCTTGTAATTTAAGCTGACACTCCTAATTTAGCTCTTGTCCTCTACTGAGTCTACCTAATTATATGTATGGATTGACTTGG

AGCTCTTGTCCTCTACTGAGTCTACCTAATTATATGTATGGATTGACTTGGTGTTTTCTCTTTTTCTTAAATAGTAATGCAGAAAGCCTGGAGA
GAGAG

Reads

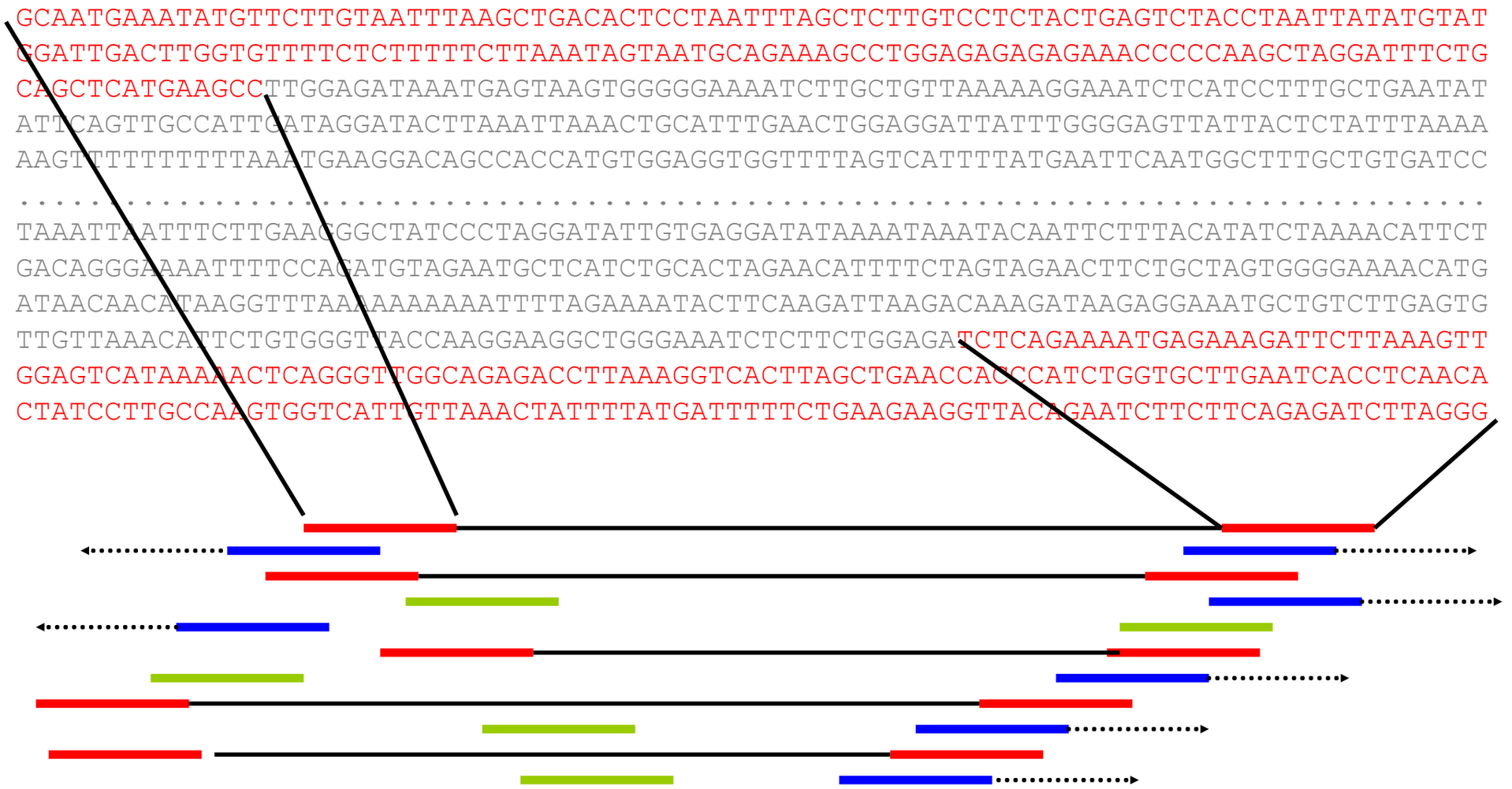
GAAATATGTTCTTGTAATTTAAGCTGACACTCCTAATTTAGCTCTTGTCCTCTACTGAGTCTACCTAATTATATGTATGGATTGACTTGGTGTTTTCTCTTTTTCTTAAATAGTAATGCAGAAAGCCTGGAG
AGAGAG

Assembl
y

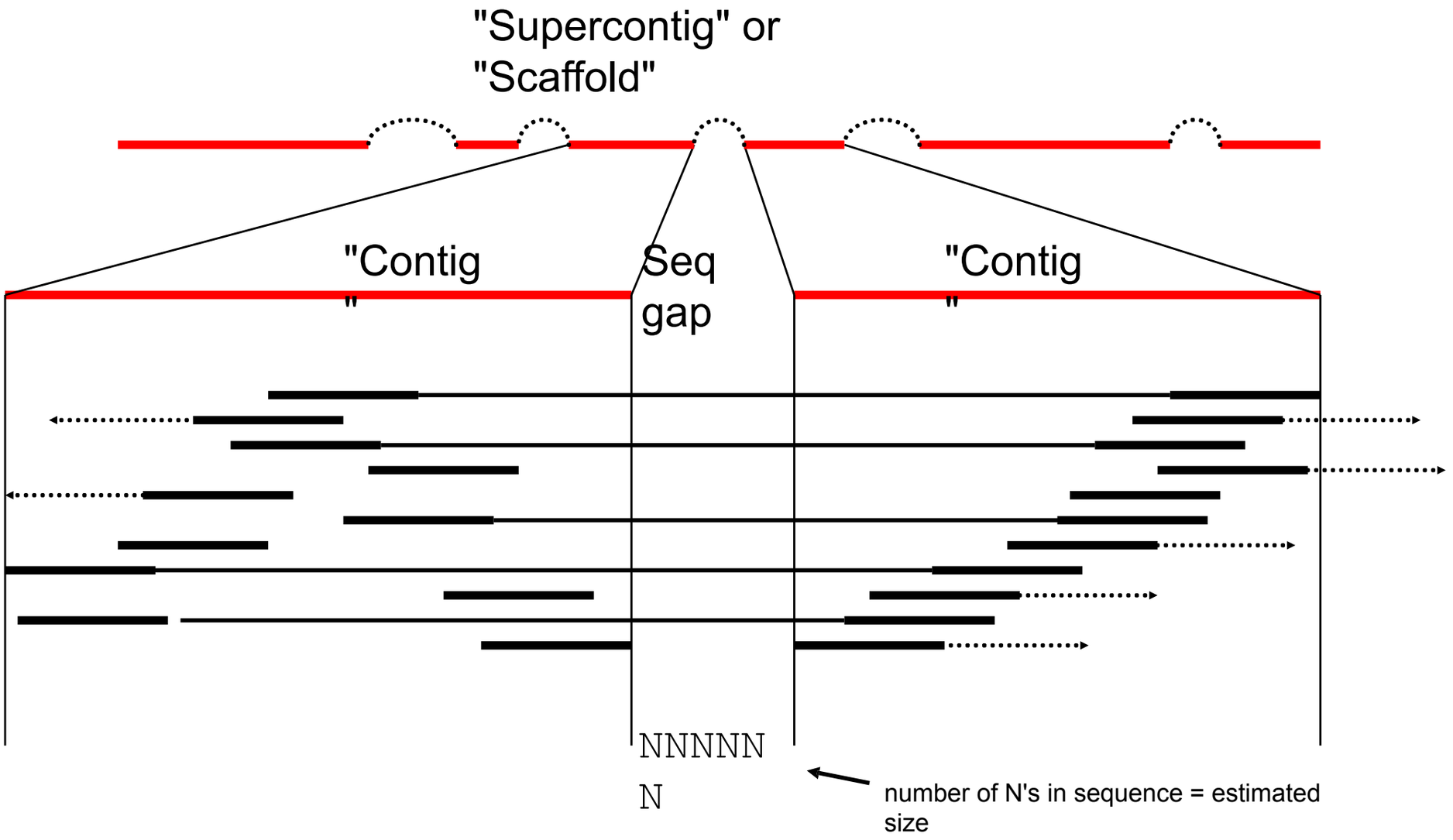
# Single End Read

GCAATGAAATATGTTCTTGTAATTTAAGCTGACACTCCTAATTTAGCTCTTGTCCTCTACTGAGTCTACCTAATTATATGTATG
GATTGACTTGGTGTTTTCTCTTTTTCTTAAATAGTAATGCAGAAAGCCTGGAGAGAGAGAAACCCCCAAGCTAGGATTTCTGCA
GCTCATGAAGCCTTGGAGATAAATGAGTAAGTGGGGGAAAATCTTGCTGTTAAAAAGGAAATCTCATCCTTTGCTGAATATATT
CAGTTGCCATTGATAGGATACTTAAATTAAACTGCATTTGAACTGGAGGATTATTTGGGGAGTTATTACTCTATTTAAAAAAGT
TTTTTTTTTAAATGAAGGACAGCCACCATGTGGAGGTGGTTTTAGTCATTTTATGAATTCAATGGCTTTGCTGTGATCCTAAAT
TAATTTCTTGAAGGGCTATCCCTAGGATATTGTGAGGATATAAAATAAATACAATTCTTTACATATCTAAAACATTCTGACAGG
GAAAATTTTCCAGATGTAGAATGCTCATCTGCACTAGAACATTTTCTAGTAGAAC**TTCTGCTAGTGGGGAAAACATGATAACAA**
**CATAAGGTTTAAAAAAAAAATTTTAGAAAATACTTCAAGATTAAGACAAAGATAAG**AGGAAATGCTGTCTTGAGTGTTGTTAAA
CATTCTGTGGGTTACCAAGGAAGGCTGGGAAATCTCTTCTGGAGATCTCAGAAAATGAGAAAGATTCTTAAAGTTGGAGTCATA
AAAACTCAGGGTTGGCAGAGACCTTAAAGGTCACTTAGCTGAACCACCCATCTGGTGCTTGAATCACCTCAACACTATCCTTGC
CAAGTGGTCATTGTTAAACTATTTTATGATTTTTCTGAAGAAGGTTACAGAATCTTCTTCAGAGATCTTAGGGAAAAAAAAAAA
AGATTGTCGTGAGAGTTGAAAATCCTGCCATTGTAACCAGTTGATCTACGGTTTCTGATTCTGTCATGCAACATATTTATTTTC
CAGTTTCTTGTCATCTACAAATTCGATATGCCTGCCTTCTGTGTGTCATCCATATTTCTGAGAAAAATATGAAGGCCAGGAATA
GAGCCCTGTGACATGACATAGAAACTACCCTCCAGGTTCATGTCTTCATGAATCACCATCTTTTGTATTGTTCACTCAATTACT
AAGCCACCCAGTTACACTGTGACTCAGCTCATATTTCTCCATTTGGATCTTAAGAATGCCAATCGTAGCTGCGGATCTTAAATT
TATAGTAAATCTATTACAGTAAATTAAGCTAGCACAATCTGATTTATTTATTCTTAGTGAATATAAGCTGGCTTCTAGTCGTCA
CTACTTTCTTTTTAAAGTGCTTGGAGACCATTCCTTTAATAATCCATTAGAATATCTTTCCAAATCACTGTGTTCTGTAGTTTG
GGAAGTCTGCCTTCTTCCCCTTTTTGAAAATTTATGCTACATTTATCATCTCATCTTCTAGCACCTCTCCATTCTTTGTGATTC
CTCAACTATCCACAGAGAGCAATTCCATGGCCTGCCTACAAGGTCTTTCGGTTTCCTGGGATTTGCCCATCCAGTCCAGTAATT
CATTTAGAATGGATCAATTATTTGCTATCTTACATCTTTTTACCCATTTTAGAGTTTAATTTCTTCTCCCTTTTTCAGTCTGAC
AGTCATTCTCCTTGATAGAGAAGCCAGGAACAAAATAGGAGGGAGAGAGTTTTGCTTTTTCTTTATTATCTACTGCTTTTAACA
ATAAACCTTCCTTGTTTTGATGTTATTATGTTGTTTGTCTTTTTTTTTTTACTTATTTGCCTTTGTGACATGGGGACGGTGATAG
GGCCTTAAATATAATTTTAAAATAGGGAATAAATGGTTGTCTTTAGTATTTTATTTTGTTTTATTATTATTATTATTATTGTTA
TTTTTGCAAGCTTCAGCTAATTTGGAATTGTAGCTCTCCTGACATTATTCTTATAAGCTCATTCCACTCTCTTATAGACCATCA
TTACATGCCCTCTTTCCATCTTTTAAAATATGTCCTTTAAAAATCTGACCTGGGAGAAATCTCTGTGAAGCCGTGTTGGTTACT
TAAGTGCCACCCCTCTTTTCTTCCTGAGAGGATCATTTGTGATTGCAGTTACAGTTGA

# Paired End Sequencing Reads

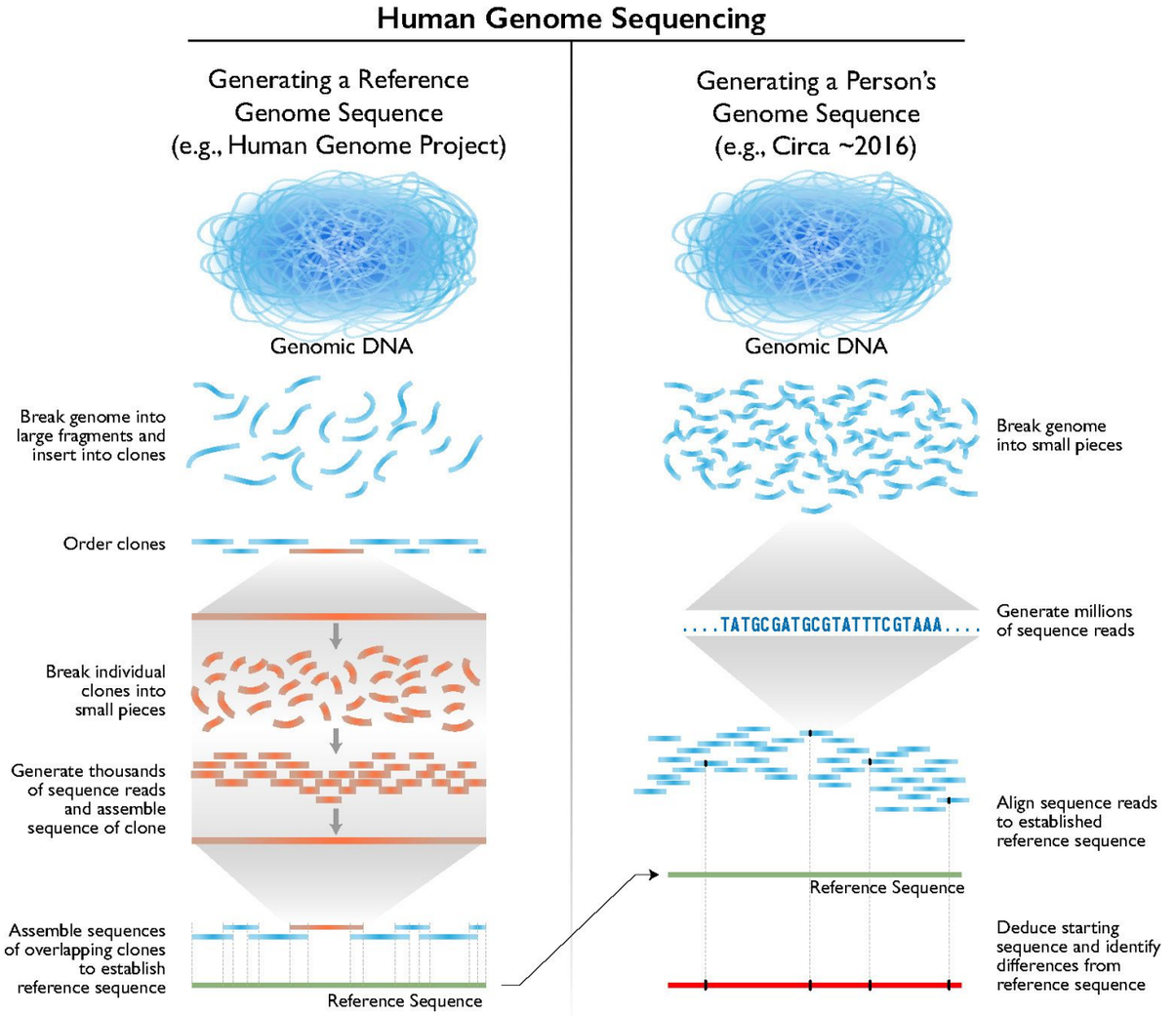


# Assembly: Contigs and Supercontigs



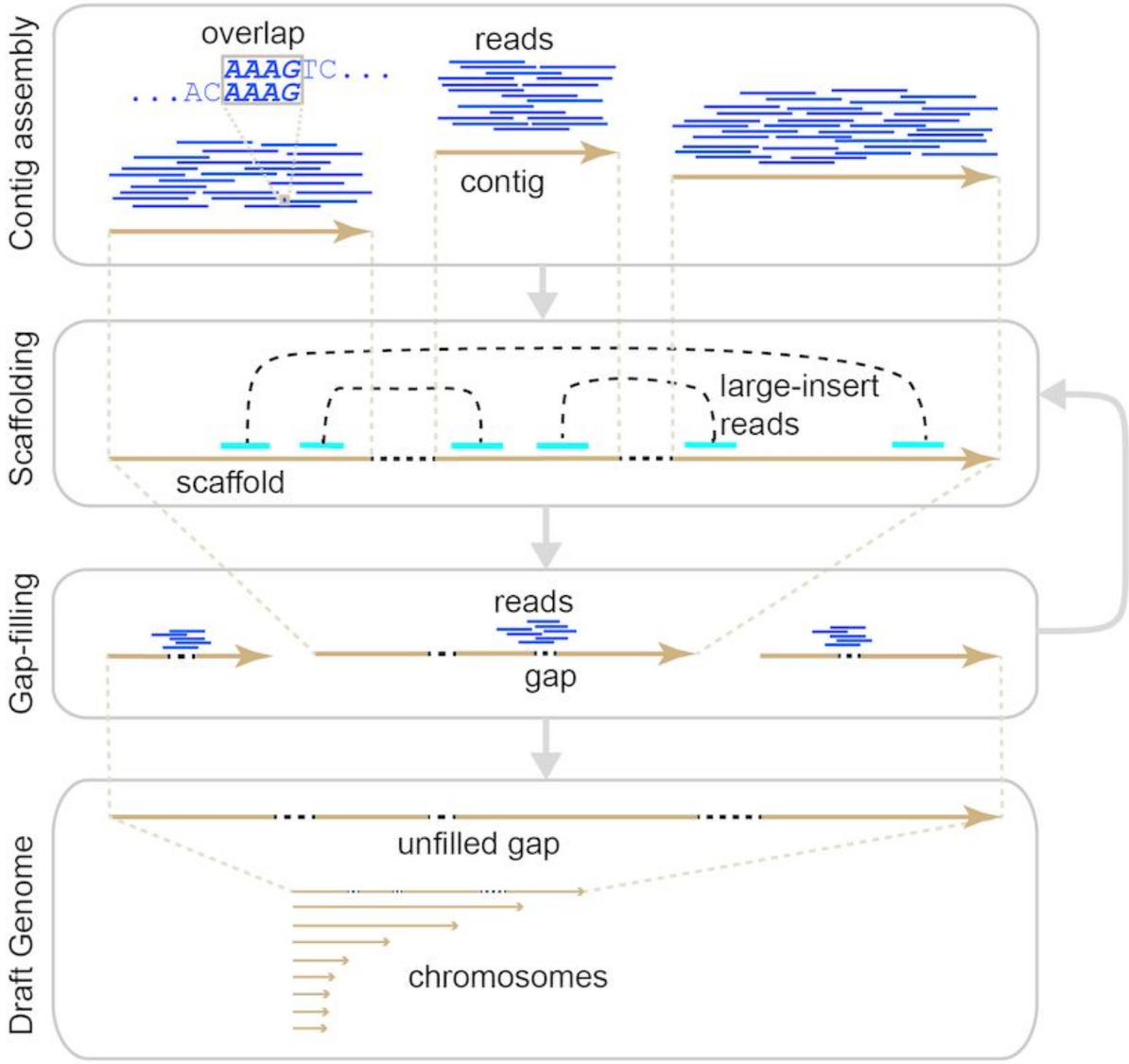
# Human Genome Project: shotgun method

Once significant human genome sequencing began for the HGP, a 'draft' human genome sequence (as described above) was produced over a 15-month period (from April 1999 to June 2000). The estimated cost for generating that initial 'draft' human genome sequence is **~$300** million worldwide



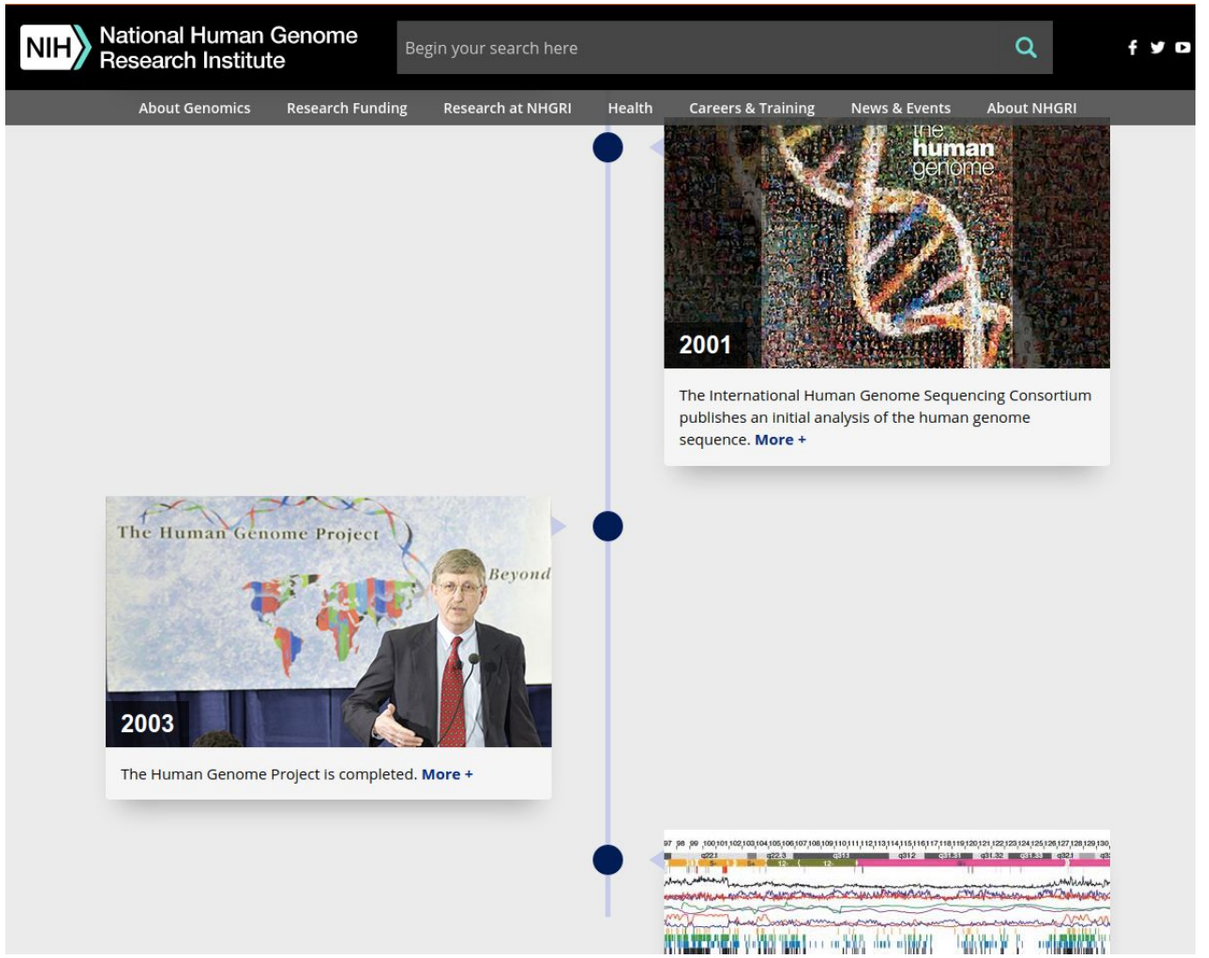

https://www.genome.gov/about-genomics/fact-sheets/Sequencing-Human-Genome-cost

# Assemblers



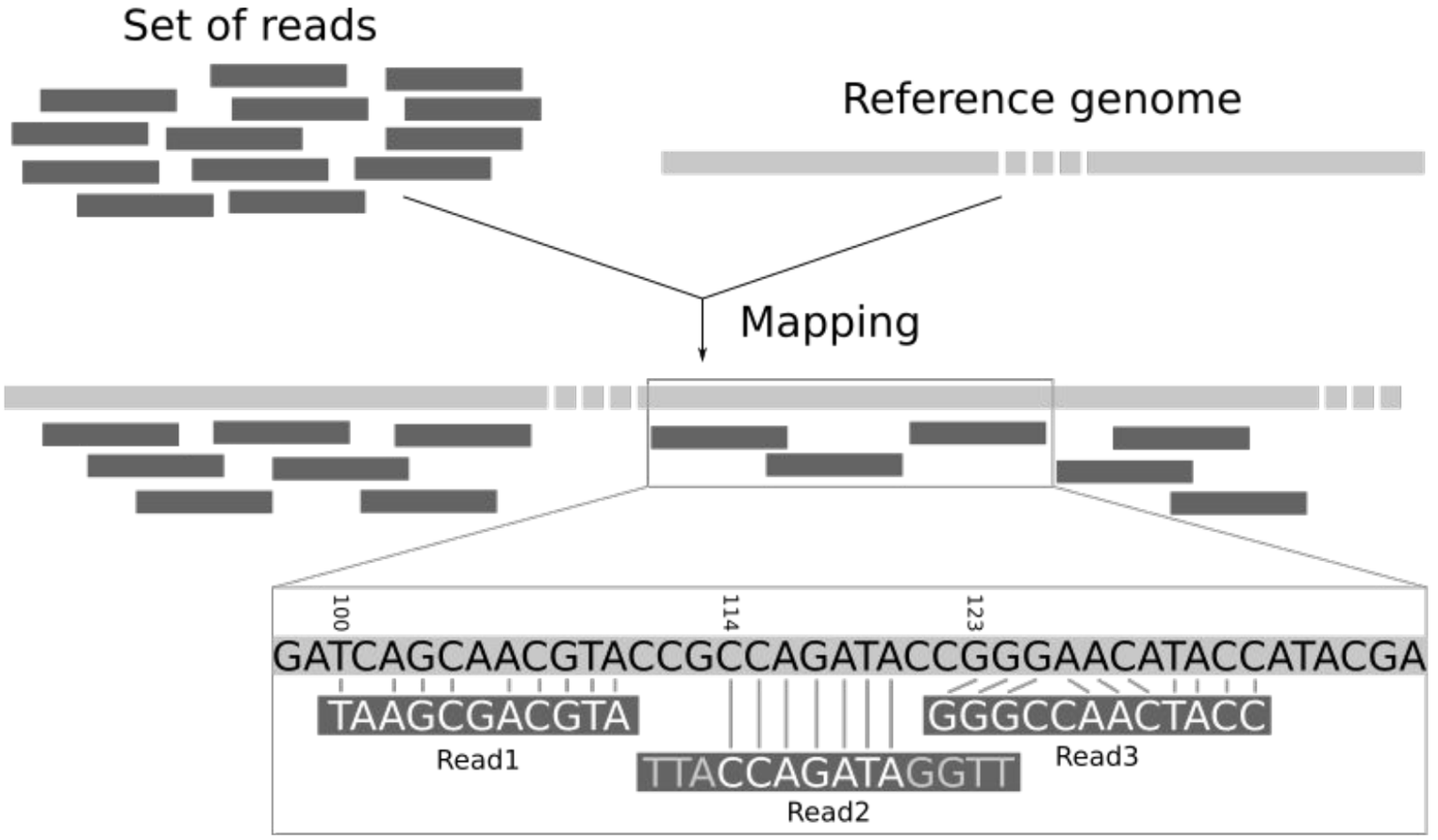

10.1093/BIB/BBW096

# Human Genome Project: History




https://www.genome.gov/human-genome-project/Timeline-of-Events

# Mappers (Aligners)



# High Throughput Sequencing

"*The cost of DNA sequencing has plunged orders of magnitude in the last 25 years. Back in 1990, sequencing 1 million nucleotides cost the equivalent of 15 tons of gold (adjusted to 1990 price). At that time, this amount of material was equivalent to the output of all United States gold mines combined over two weeks. Fastforwarding to the present, sequencing 1 million nucleotides is equivalent to the value of ~30 g of aluminum. This is approximately the amount of material needed to wrap five breakfast sandwiches at a New York City food car.*"

Erlich Y. (2015). A vision for ubiquitous sequencing. *Genome Res.* **25(10)**:1411-6.

# The Players

- Commercially available systems:
  - 454, Helicos – both commercially dead
  - Illumina – most prevalent technology
  - SOLiD (Life Technologies) - dead
  - Ion Torrent (Life Technologies)
  - Complete Genomics – acquired by BGI, now BGI-Seq
  - Pacific Biosciences – acquired by Illumina
  - Oxford Nanopore
- Next generation approaches
  - Illumina Nanopore (probably abandoned)
  - NABsys, Genia, Noblegen – all be dead
  - Roswell Biotech – 4$^{th}$ generation? CMOS-based.

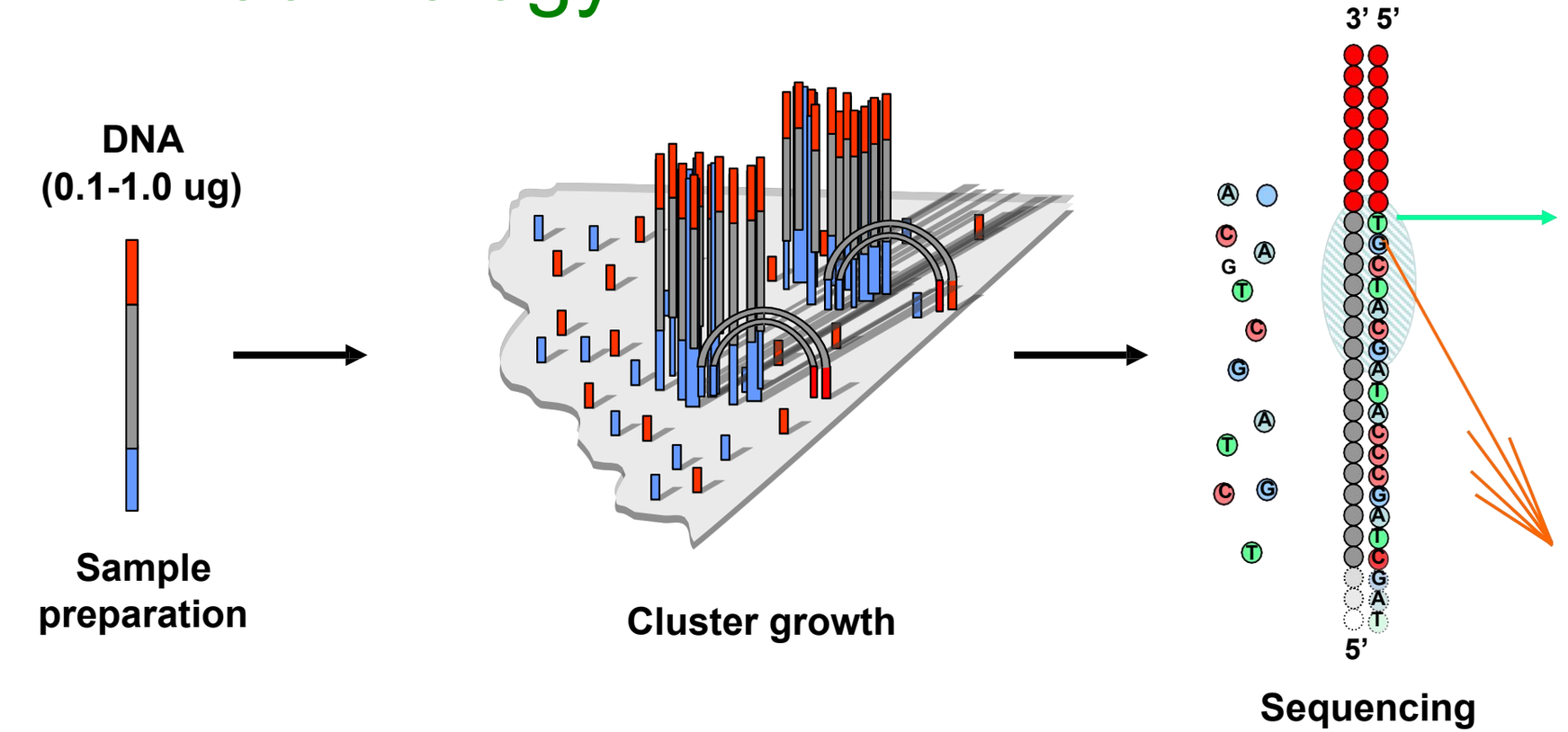
# Sequencing Platforms

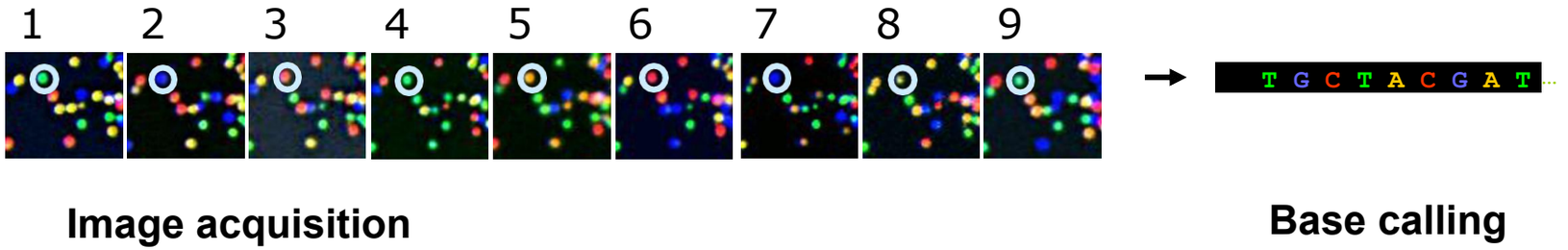
| Platform | Reads x run: (M) | Read length: | Run time: (d) | Yield: (Gb) | Rate: (Gb/d) | per-Gb: ($) | hg-30x: ($) | Machine: ($) |
|---|---|---|---|---|---|---|---|---|
| **iSeq 100 1fcell** | 4 | 250* | 077-1.28 | 1.2-2 | 1.56 | 521 | $62,500 | 19.9K |
| **MiniSeq 1fcell** | 25 | 150* | 1 | 7.5 | 7.5 | 233 | $28,000 | 49.5K |
| **MiSeq 1fcell** | 25 | 300* | 2 | 15 | 7.5 | 66 | $8,000 | 99K |
| **NextSeq 550 1fcell** | 400 | 150* | 1.2 | 120 | 100 | 50 | $5,000 | 250K |
| **HiSeq 2500 RR 2fcells** | 600 | 100* | 1.125 | 120 | 106.6 | 51.2 | $6,144 | 740K |
| **HiSeq 2500 V3 2fcells** | 3000 | 100* | 11 | 600 | 55 | 39.1 | $4,692 | 690K |
| **HiSeq 2500 V4 2fcells** | 4000 | 125* | 6 | 1000 | 166 | 31.7 | $3,804 | 690K |
| **HiSeq 4000 2fcells** | 5000 | 150* | 3.5 | 1500 | 400 | 20.5 | $2,460 | 900K |
| **HiSeq X 2fcells** | 6000 | 150* | 3 | 1800 | 600 | 7.08 | $850 | 1M |
| **NovaSeq S1 2fcells** | 3300 | 150* | 1.66 | 1000 | 600 | 18.75 | $1,800 | 999K |
| **NovaSeq S2 2fcells** | 6600 | 150* | 1.66 | 2000 | 1200 | 17.5 | $1,564 | 999K |
| **NovaSeq S4 2fcells** | 20000 | 150* | 1.83 | 6000 | 3600 | 10.67 | $700 | 999K |
| **Illumina PacBio RSII** | 0.88 | 20K** | 4.3 | 12 | 2.8 | 200 | $24,000 | 695K |
| **Illumina PacBio Sequel 16cells v6.0 2018** | 6.4 | 45K** | 6.6 | 160-320 | 24-48 | 80 | $9,600 | 350K |
| **Illumina PacBio Q1 2019** | -- | 45K** | -- | 192 | -- | 6.6 | $1,000 | 350K |
| **SmidgION 1fcell** | -- | 500-2,000,000 | TBC | TBC | TBC | TBC | -- | -- |
| **Flongle 1fcell** | -- | 500-2,000,000 | 1 | 0.1/1.8-3.3 | -- | 90-30 | $2,700 - $8,100 | -- |
| **MinION Mk 1B 1fcell** | -- | 500-2,000,000 | 3 | 17/30-50 | -- | 50-12.5 | $1,125 - $2,700 | -- |
| **GridION X5 5fcells** | -- | 500-2,000,000 | 3 | 85/150-250 | -- | 47.5/15.70-7 | $675 - $1,575 | -- |
| **PromethION 48fcells** | -- | 500-2,000,000 | 2.6 | 3000/7000-15000 | -- | 14/7-3.5 | $315 - $1,400 | -- |

https://docs.google.com/spreadsheets/d/1GMMfhyLK0-q8XkIo3YxlWaZA5vVMuhU1kg41g4xLkXc/edit?hl=en_GB&hl=en_GB#gid=515231169

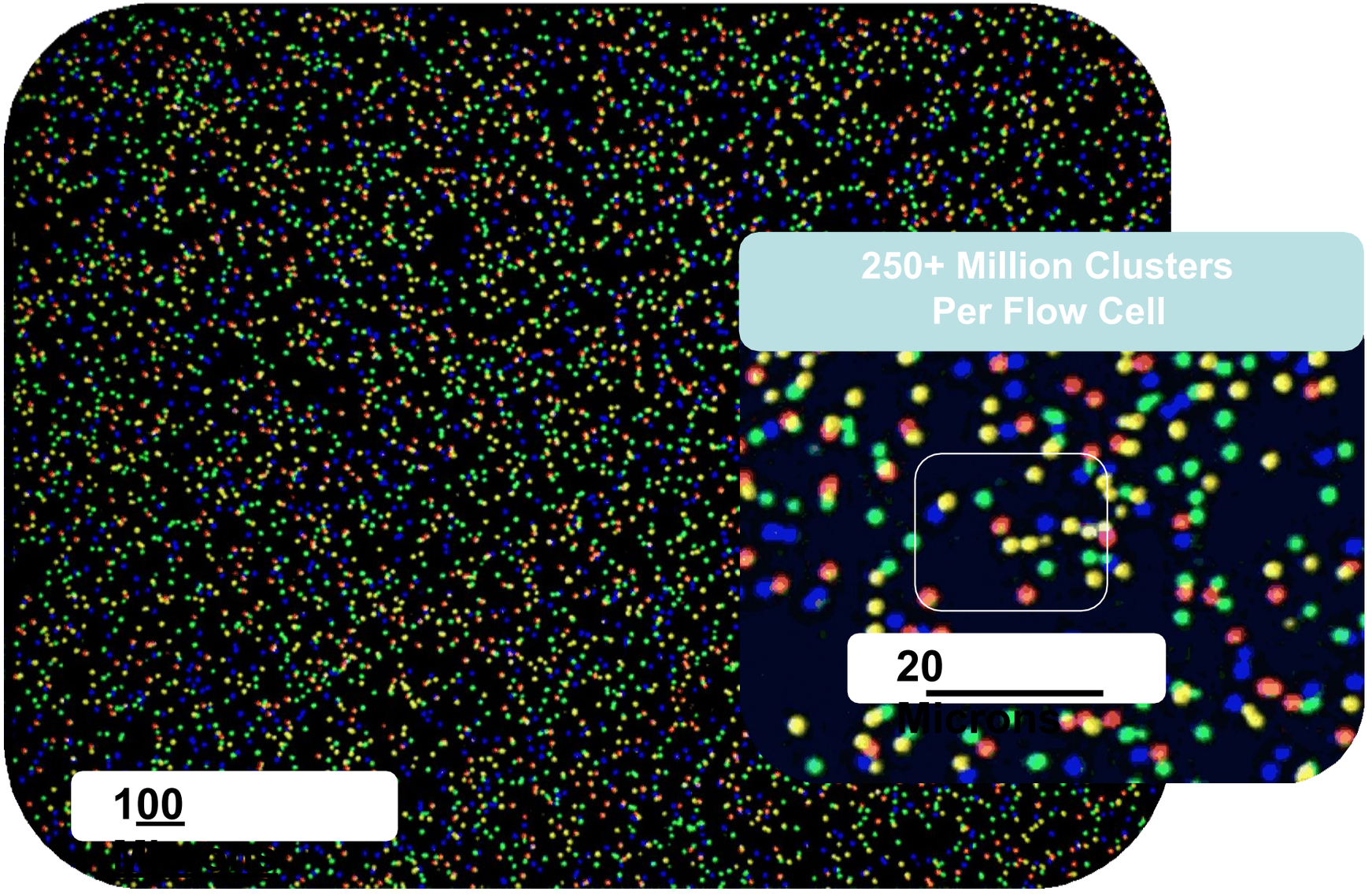
# Illumina Sequencing Technology

*Robust Reversible Terminator Chemistry Foundation*



# Illumina Sequence Visualization



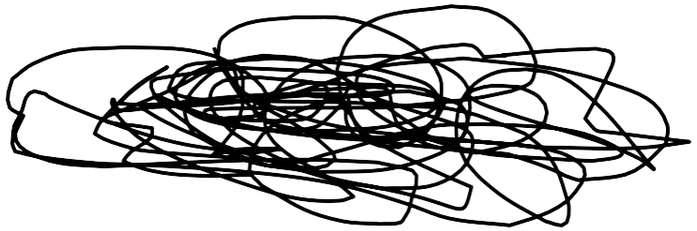
# Image Processing, Base Calling

- Image processing algorithms find signals in each panel, align signals from different panels, etc.
  - Machines ship with server or small cluster that does image analysis while run is happening
- Sequence data after base calling much reduced in size (tens of gigabytes) => more manageable but still large amounts that add up over time
- Unsustainable to keep image data; people discard the images, and just keep the sequences (fastq format).

# Recent Illumina Innovations

- Patterned flow cells (HiSeq 3000/HiSeq 4000 systems)
  - Allows denser cluster spacing
  - Avoids cluster overlap
  - Image analysis easier
- Two-Channel SBS (NextSeq)
  - Two, rather than 4 colors
  - Leads to faster sequencing times
- Synthetic Long Reads
  - We may discuss later, but not widely used
- MiniSeq
  - 2nd generation MiSeq – smaller, cheaper, more reads
- NovaSeq – announced 2017
  - Up to 10 billion reads in 2 days, per flowcell (2018)
  - Towards $100 genome
- iSeq
  - released last year. Benchtop ($20K) platform for 6M reads

# How do we make an Illumina Genomic DNA library?



Double-stranded genomic DNA

Fragment (Covaris)

Polish, add dA overhang
Add adaptors, size select

Sequence

# How Much Sequence?

- HiSeq 2500 can give ~250 million reads/lane of paired end 100bp reads
- This is 50Gb of sequence
- This is ~4,000x coverage yeast (12Mb).
- This is an obvious waste of resources (it's also ~500x *C. elegans*, and ~500x *D. melanogaster*)
- How can we sequence on a HiSeq and not waste all these resources when sequencing smaller genomes?
- HiSeq 3000/HiSeq 4000
  - Patterned flow cells (not random clusters)
  - Almost twice as much data, half the time

# What are the data?

- Illumina produces data in fastq format.

```
@SEQ_ID
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!"*((((***+))%%%++)(%%%%).1***-+*"))**55CCF>>>>>>CCCCCCC65
```

’@’ followed by a sequence Identifier

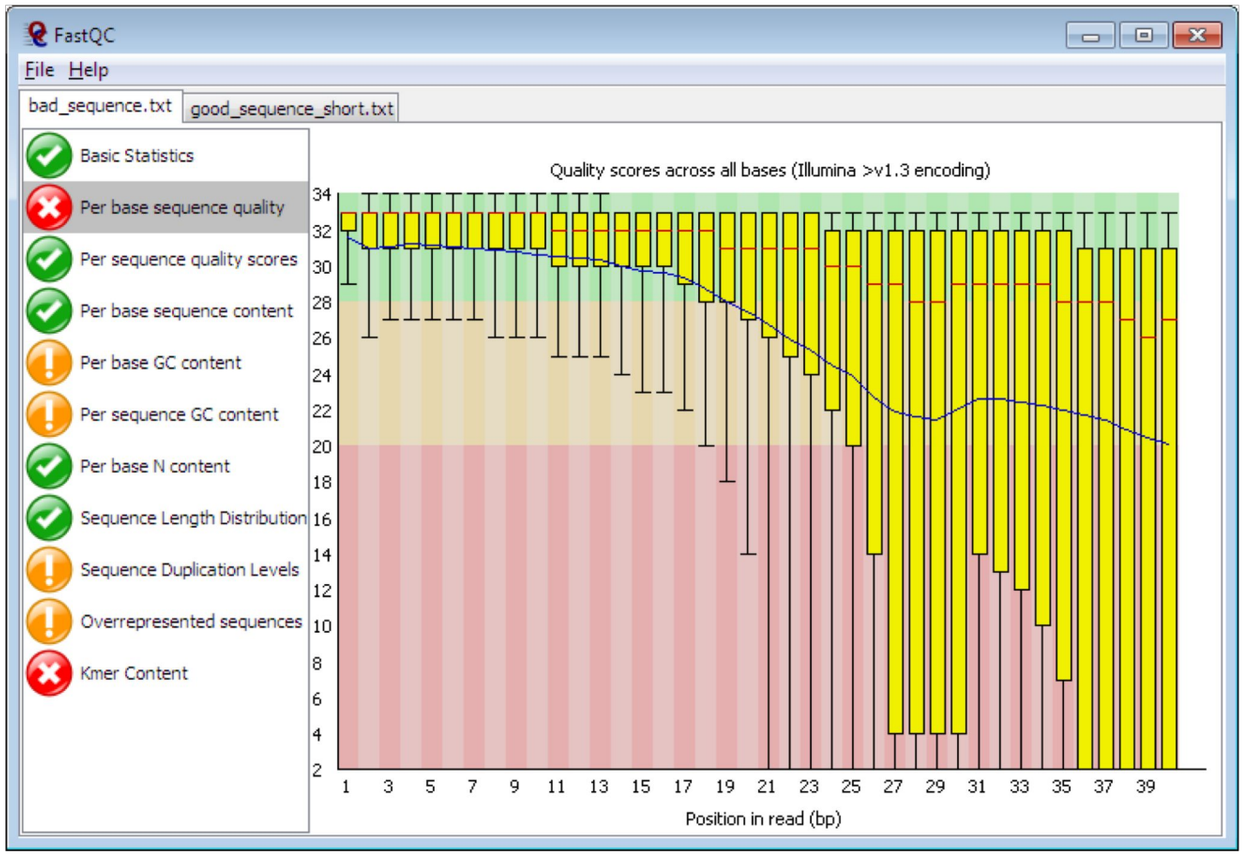
The sequence

’+’, optionally followed by a sequence Identifier

The quality scores

# FastQC

http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc/



# *De novo* Assembly of Short Reads

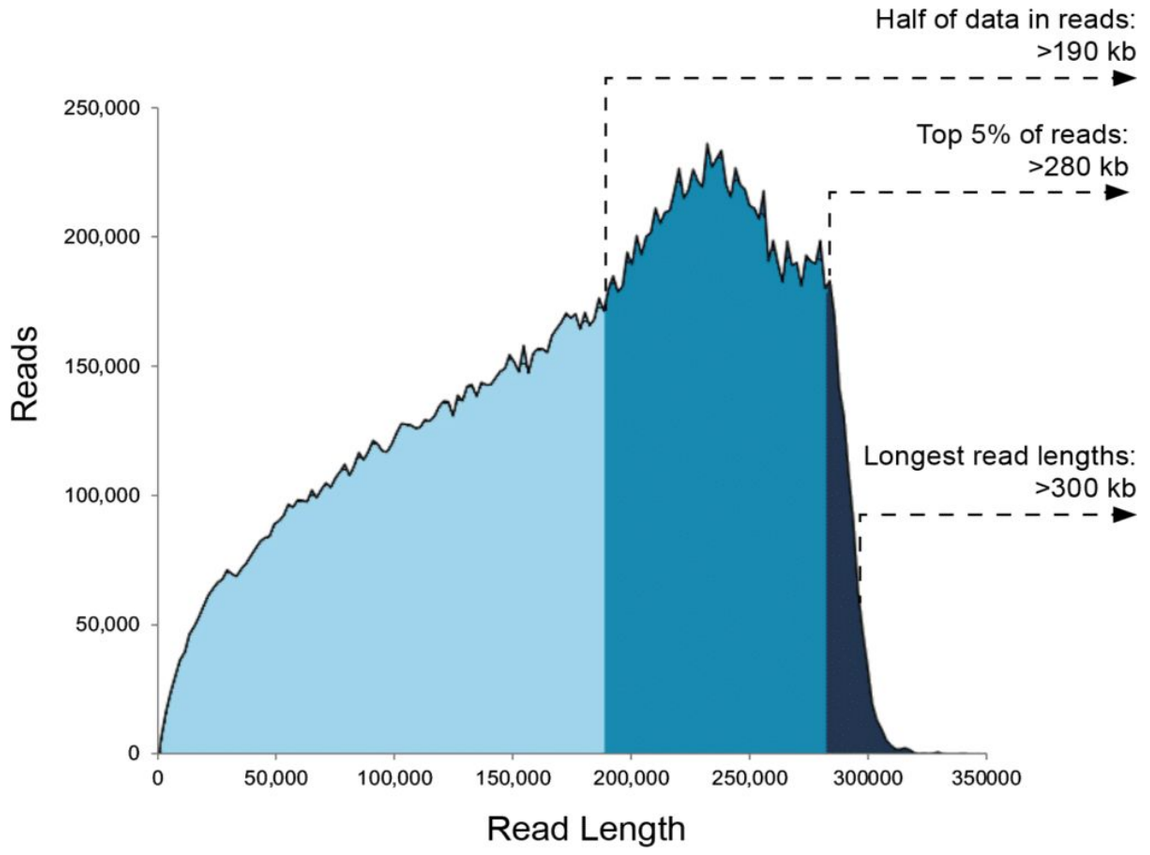
- Several methods available
- Short reads require long overlaps
  - e.g., 33 bp reads must overlap by 20 bp
  - end-trimming helps, to remove low quality bases.
- Most *de novo* short read assemblers use a k-mer hashing based approach and de Bruijn graphs.
- The central challenge of genome assembly is resolving repeat regions.

# *De novo* Assembly Strategies

- Many, many different algorithms and open source (as well as closed source) software for short read sequence assembly.
- Choice of tool depends on exactly what you are trying to assemble:
  - Genome size
  - Genome complexity
  - Level of polymorphism
  - Genome vs. transcriptome
  - Sequence coverage you have (more is generally better)
  - Paired-end vs. single end (you should really have paired-end data)
- E.g.
  - Velvet (Zerbino and Birney, 2008)
    - Uses DeBruijn graph algorithm plus error correction
  - SGA (Simpson and Durbin, 2010)
    - Use String Graph – lower memory requirements, but takes longer
  - SOAPdenovo2 (Li et al, 2012)
    - Also uses DeBruijn graphs with error correction

# Pacific Biosciences

- Single Molecule Real Time (SMRT) DNA Sequencing
- Light is detected when fluorescent nucleotides are incorporated into a growing DNA strand
- Half of all data in reads >45kb; can get >200kb



# Pacific Biosciences

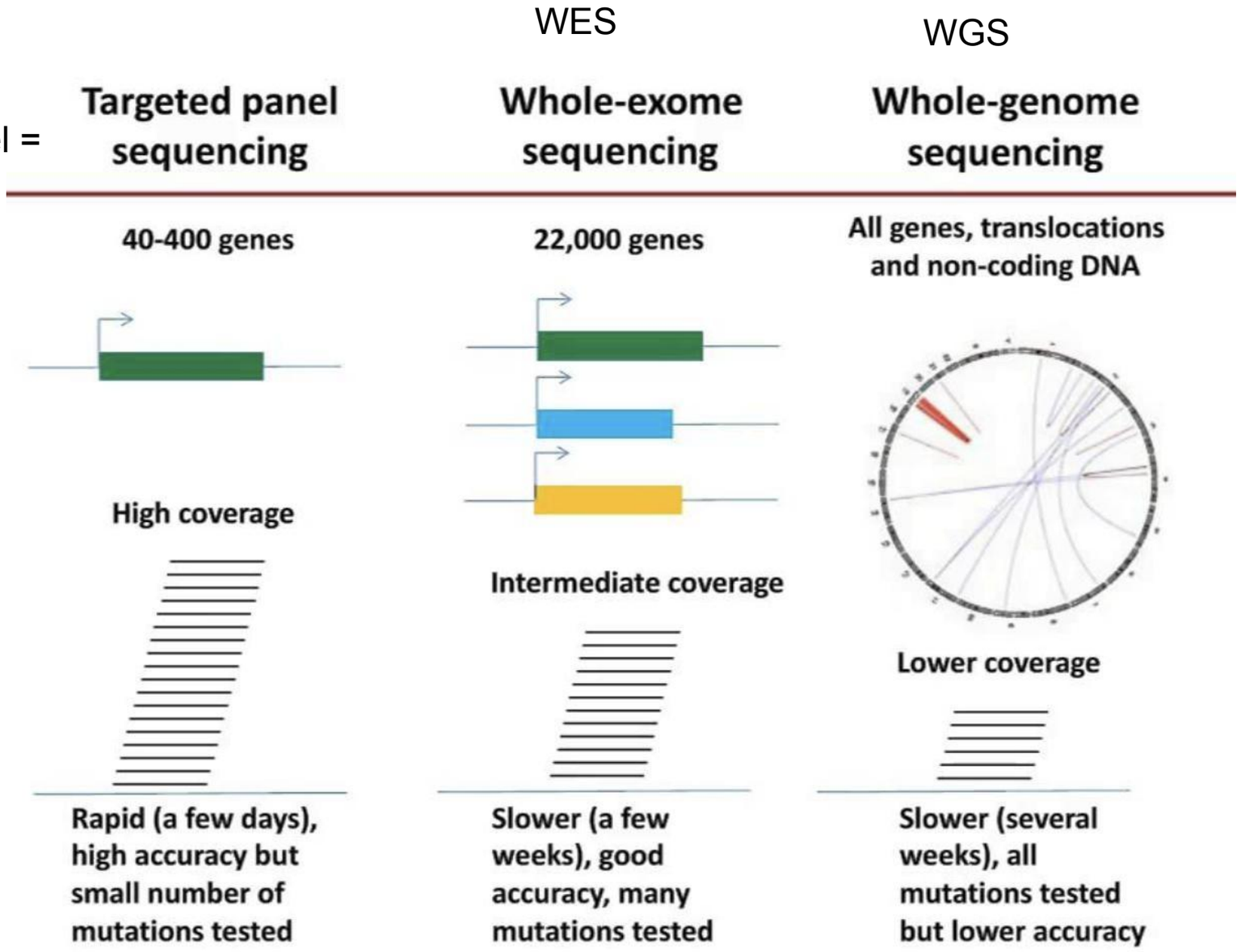
- Single Molecule Real Time (SMRT) DNA Sequencing
- Light is detected when fluorescent nucleotides are incorporated into a growing DNA strand
- Half of all data in reads >45kb; can get >200kb
- Accuracy now ~99% (Q20); with 40x coverage, consensus approached 99.999% accuracy (Q50)
- Observation of DNA modifications
- Throughput per run is low (~6 million reads), but run time is short (~6 hours)

# Oxford Nanopore

- MinION, GridION, PromethION products
- DNA “sequenced” as it is dragged through a nanopore, based on change in conductance
- Can also detect modified bases
- High error rate (5-15%), but improving
- But very long reads –>2 million bp read reported
- Tons of papers/data/tools released
- Nanopore is a game changer for genome assembly
- Has also been reported recently to directly sequence RNA – great for seeing isoforms

# Clinical DNA sequencing

Gene panel =

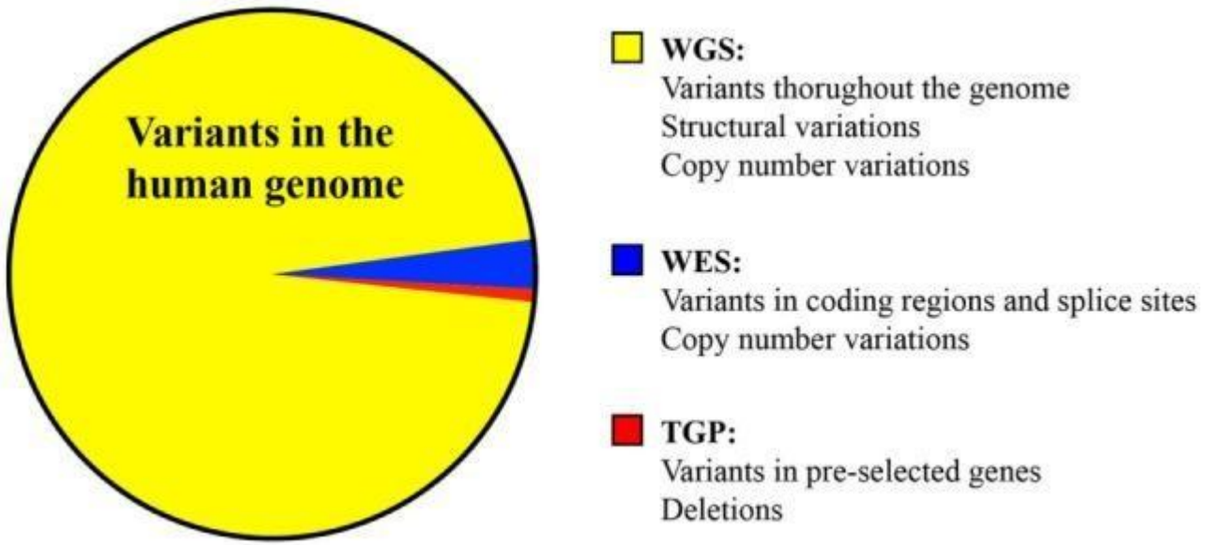



https://nebula.org/blog/invitae-review/

# Comparison

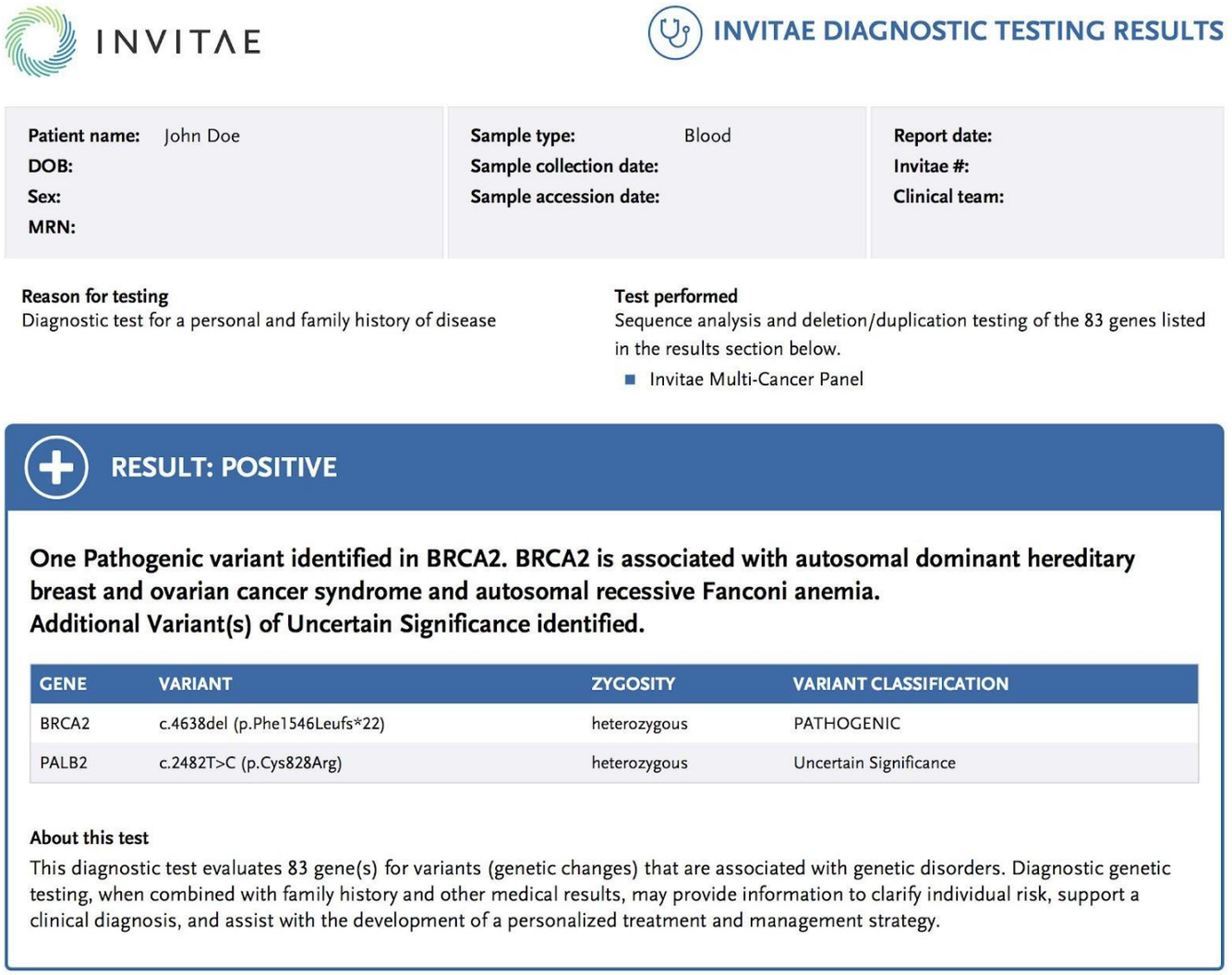
Review > Front Immunol. 2017 Jul 24;8:847. doi: 10.3389/fimmu.2017.00847.
eCollection 2017.

## Uses of Next-Generation Sequencing Technologies for the Diagnosis of Primary Immunodeficiencies




doi: 10.3389/fimmu.2017.00847.

# Gene panel outcome example



INVITAE

INVITAE DIAGNOSTIC TESTING RESULTS

| | | |
|---|---|---|
| **Patient name:** John Doe | **Sample type:** Blood | **Report date:** |
| **DOB:** | **Sample collection date:** | **Invitae #:** |
| **Sex:** | **Sample accession date:** | **Clinical team:** |
| **MRN:** | | |

**Reason for testing**
Diagnostic test for a personal and family history of disease

**Test performed**
Sequence analysis and deletion/duplication testing of the 83 genes listed in the results section below.

- Invitae Multi-Cancer Panel

**RESULT: POSITIVE**

**One Pathogenic variant identified in BRCA2. BRCA2 is associated with autosomal dominant hereditary breast and ovarian cancer syndrome and autosomal recessive Fanconi anemia.**
**Additional Variant(s) of Uncertain Significance identified.**

| GENE | VARIANT | ZYGOSITY | VARIANT CLASSIFICATION |
|---|---|---|---|
| BRCA2 | c.4638del (p.Phe1546Leufs*22) | heterozygous | PATHOGENIC |
| PALB2 | c.2482T>C (p.Cys828Arg) | heterozygous | Uncertain Significance |

**About this test**
This diagnostic test evaluates 83 gene(s) for variants (genetic changes) that are associated with genetic disorders. Diagnostic genetic testing, when combined with family history and other medical results, may provide information to clarify individual risk, support a clinical diagnosis, and assist with the development of a personalized treatment and management strategy.

https://nebula.org/blog/invitae-review/

# Gene panel

Find the right test for you

| | Invitae Cancer Screen<br>Looks at 61 genes to assess your risk of developing an inherited form of cancer<br>$250 | Invitae Cardio Screen<br>Looks at 77 genes to assess your risk of developing an inherited form of cardiovascular (heart) disease<br>$250 | Invitae Genetic Health Screen<br>Looks at 147 genes to assess your risk of developing an inherited form of cancer or cardiovascular (heart) disease and more<br>$350 |
|---|---|---|---|
| Breast cancer | ● | | ● |
| Colorectal cancer | ● | | ● |
| Cutaneous melanoma | ● | | ● |
| Gastric cancer | ● | | ● |
| Ovarian cancer | ● | | ● |
| Pancreatic cancer | ● | | ● |
| Prostate cancer | ● | | ● |
| Renal cell cancer | ● | | ● |
| Thyroid cancer | ● | | ● |
| Uterine cancer | ● | | ● |
| Aortopathies | | ● | ● |
| Arrhythmias | | ● | ● |
| Cardiomyopathies | | ● | ● |
| Genetic forms of high blood pressure and high cholesterol | | ● | ● |
| Thrombophilia | | ● | ● |
| Additional conditions | | | |
| Alpha-1 antitrypsin deficiency | | | ● |
| Hereditary hemochromatosis | | | ● |
| Malignant hyperthermia susceptibility | | | ● |
| OTC deficiency | | | ● |
| Wilson disease | | | ● |
| | $250 | $250 | $350 |

We accept credit card and HSA/FSA payments but do not accept insurance payments for proactive testing, as this test is rarely covered by insurance.

For details on the genes tested, view our full gene list.

https://nebula.org/blog/invitae-review/

# Whole Exome Sequencing of Patients with Steroid-Resistant Nephrotic Syndrome

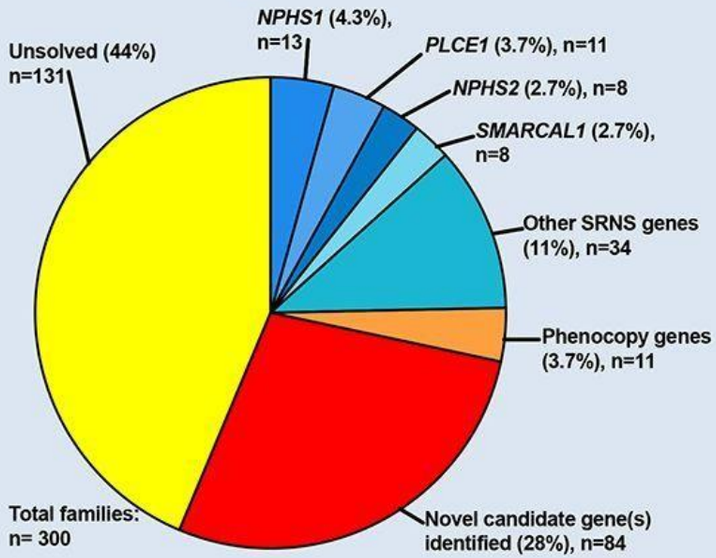
## METHODS

**335** individuals from **300** families with steroid resistant nephrotic syndrome

Whole exome sequencing performed

Variant calling and filtering

## OUTCOMES: 25% of families had a causative mutation detected in candidate genes




**CONCLUSION** A potentially causative genetic mutation can be identified in many patients with steroid-resistant nephrotic syndrome.

Jillian K. Warejko, Weizhen Tan, et al. Whole Exome Sequencing of Patients with Steroid-Resistant Nephrotic Syndrome. CJASN doi: 10.2215/CJN.04120417.

CJASN
Clinical Journal of American Society of Nephrology

ANNALS OF ONCOLOGY
ESMO GOOD SCIENCE BETTER MEDICINE BEST PRACTICE

ORIGINAL ARTICLES BREAST TUMORS | VOLUME 30, ISSUE 7, P1071-1079, JULY 01, 2019

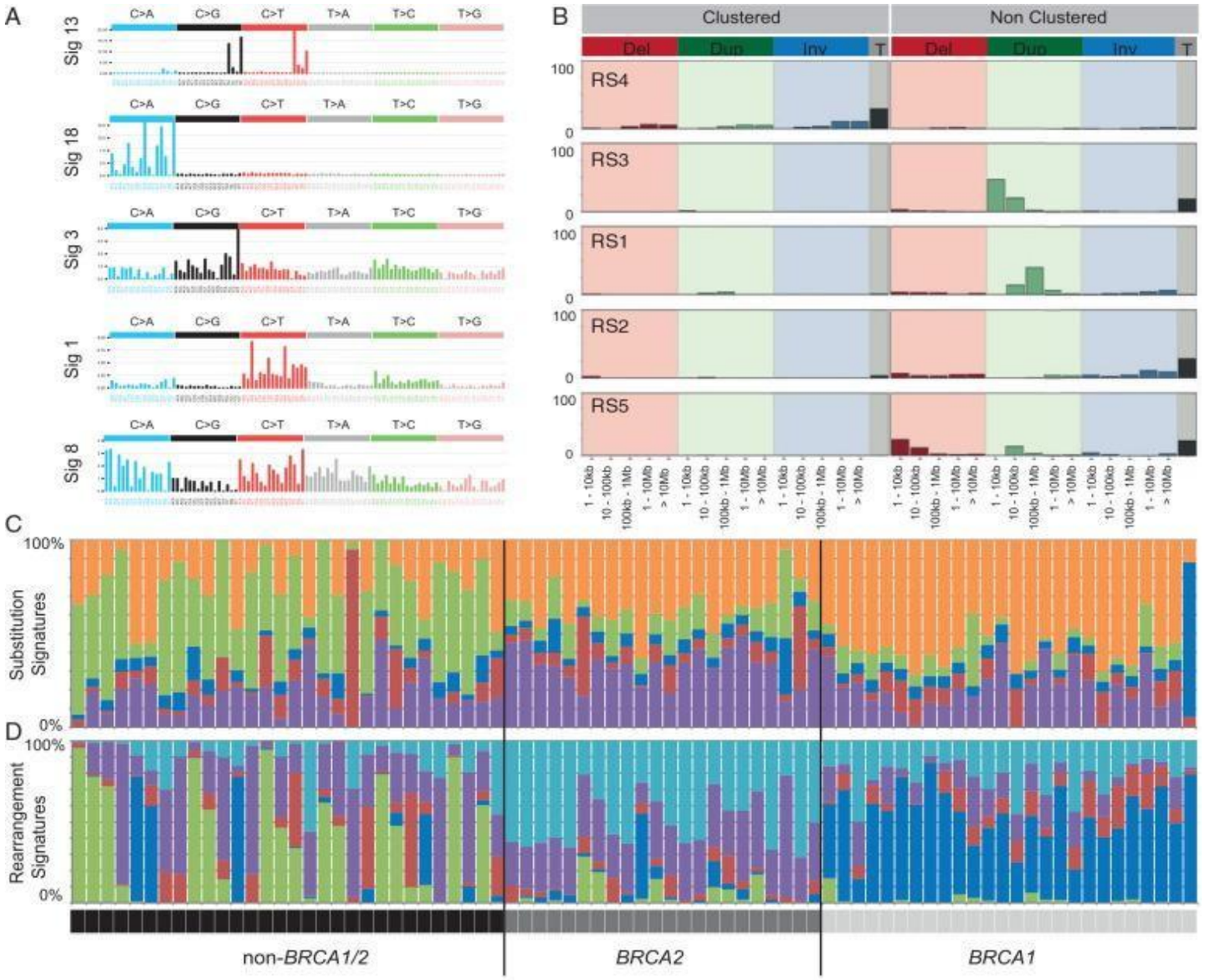
# Whole-genome sequencing reveals clinically relevant insights into the aetiology of familial breast cancers

K. Nones † • J. Johnson † • F. Newell • ... G. Chenevix-Trench • N. Waddell † • P

Show all authors • Show footnotes

Open Access • DOI: https://doi.org/10.1093/annonc/mdz132


DOI:https://doi.org/10.1093/annonc/mdz132




10.1093/annonc/mdz132

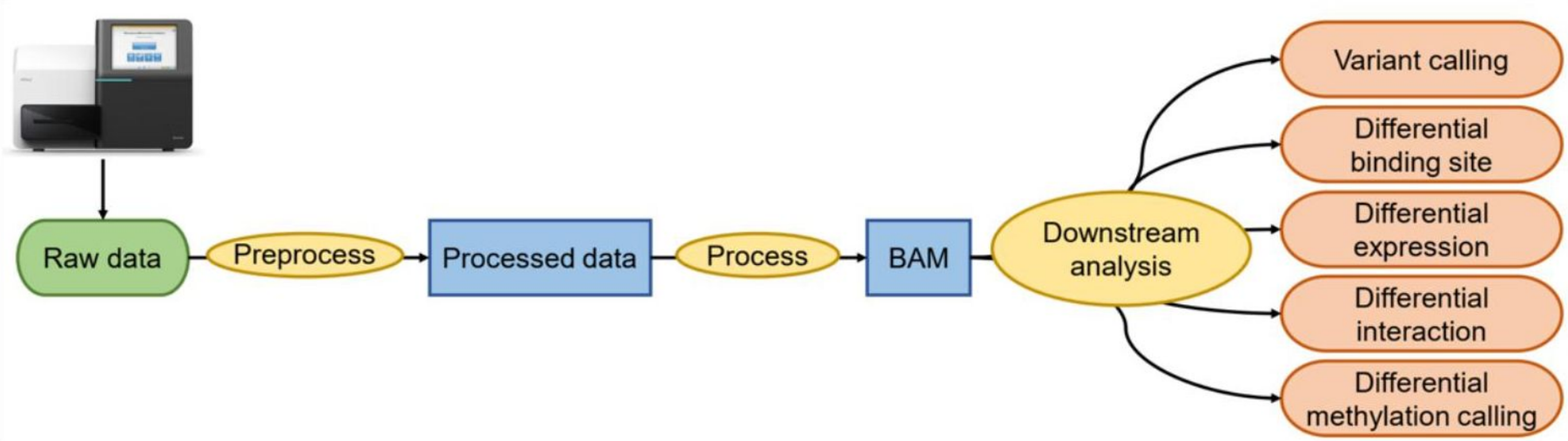
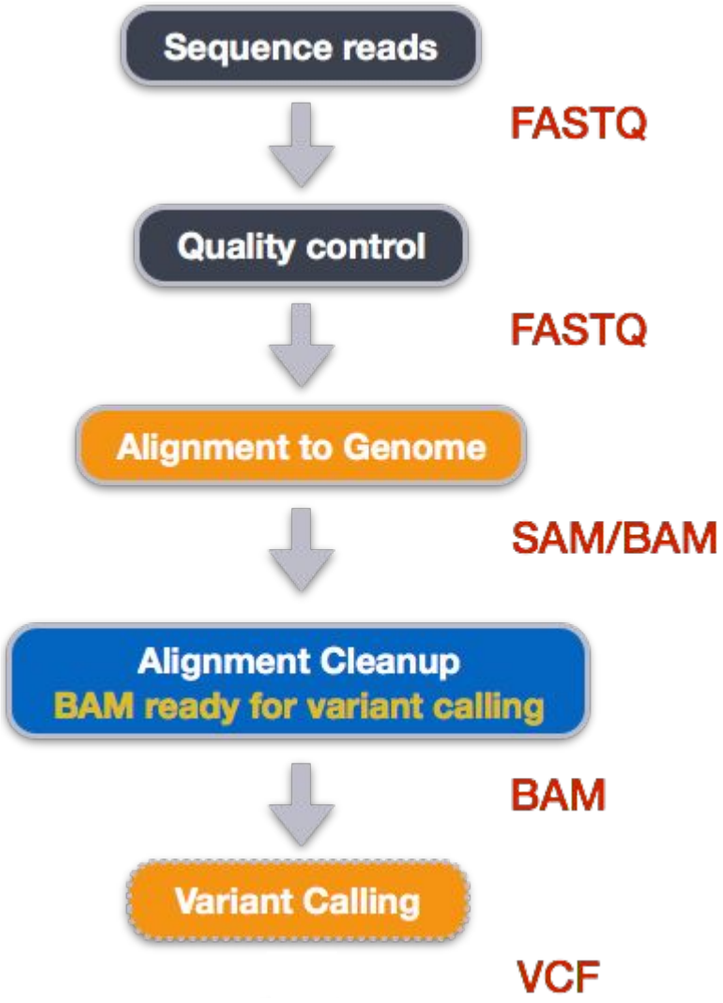
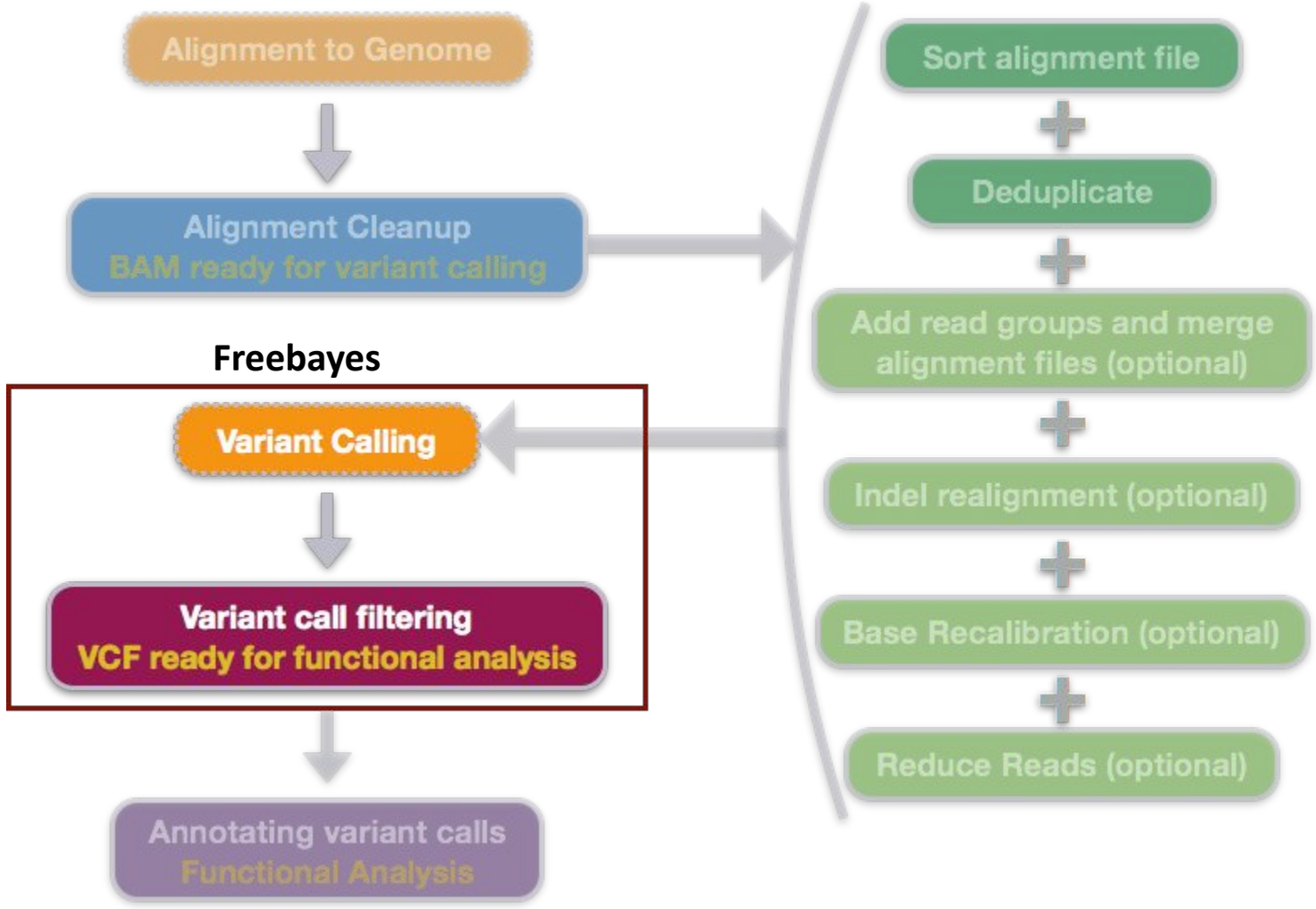



**Figure 1.** Sequential steps of a bioinformatic pipeline for genomic/epigenomic sequencing data

# Variant calling workflow of DNA-seq




Simple version




Call variants with Freebayes

# File format in Variant calling

1) • Sequencing
2) • Alignment
3) • Variant Calling
4) • Annotation
5) • Filtering

Fastq file: output from sequencing, **step 1**

```
@SN638:981:HK7HWBCXX:2:1101:14784:2782 1:N:0:TTAGGC
CATCATCGAGGACAGCGCCGGTGACCTGGCGGCCCGCATCGGTGCCCCCC
+
GGGGGIIIIIIIIIIIIIIIIIIIIIIIGIIIIIIIIIIIIIIGIIII
```

SAM/BAM file: output from alignment, **step 2**

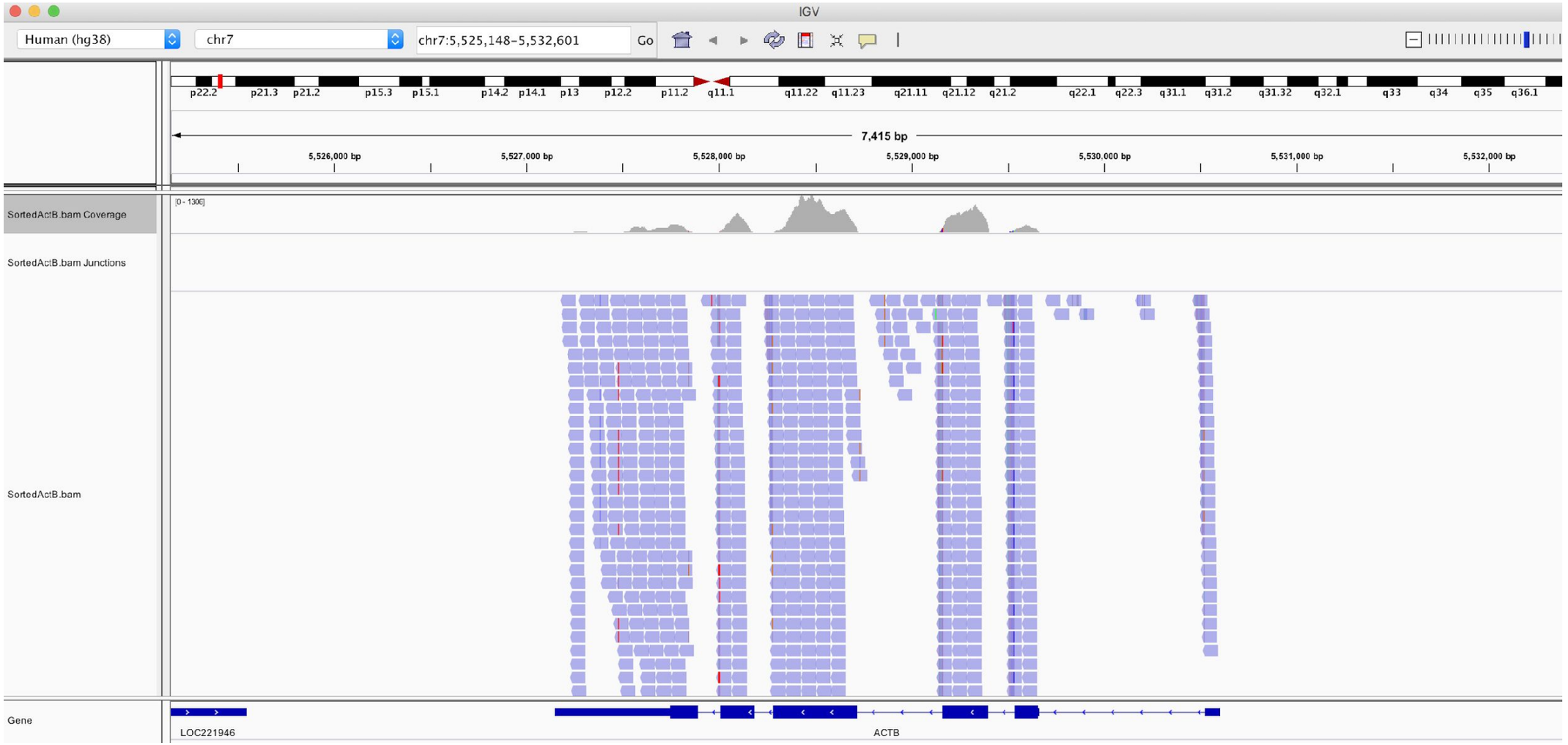
```
K00188:264:HG3WJBBXX:1:1116:14692:35180#0       121     ctg2    128     58      101M    =       128     0
AAGTCTCGACCAAGTGGTTCAGATGGTGACACAGATGTTAGCCCCATCCACCATTCAGTTGCCGTTTTGATAGCTGGAAATCCTGTAAACACAATGCTGAG
FJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJFFFAA    NM:i:10
```

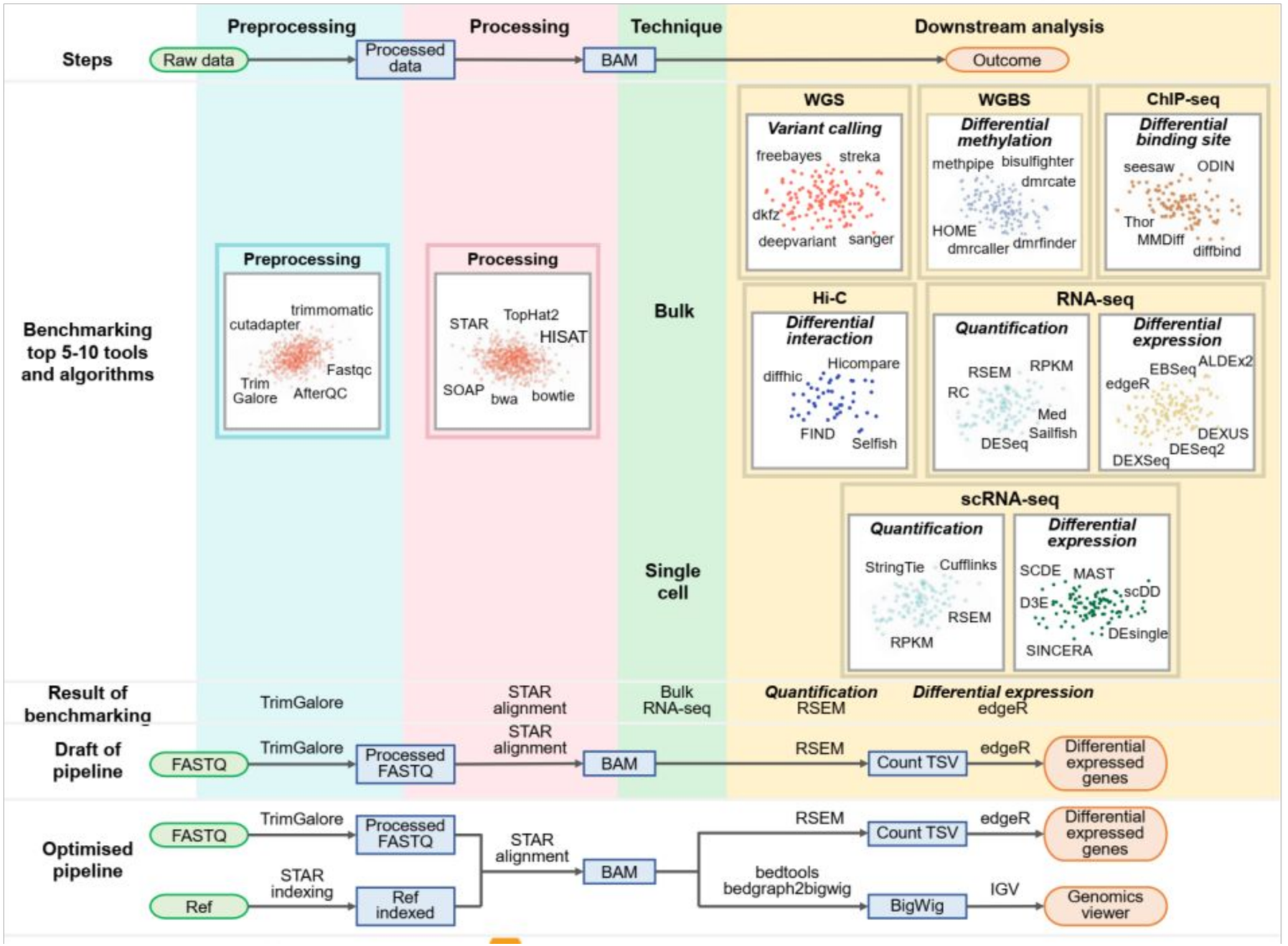
VCF file: output from variant calling, **step 3**

```
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
#CHROM POS     ID        REF     ALT     QUAL    FILTER  INFO                              FORMAT  sample07 ...
20      14370   rs6054257       G       A       29      PASS    NS=3;DP=14;AF=0.5;DB;H2           GT:GQ:DP:HQ     0|0:48:1:51,51
20      17330   .               T       A       3       q10     NS=3;DP=11;AF=0.017               GT:GQ:DP:HQ     0|0:49:3:58,50
20      111069  rs6040355       A       G,T     67      PASS    NS=2;DP=10;AF=0.333,0.667;AA=T;DB GT:GQ:DP:HQ     1|2:21:6:23,27
20      123027  .               T       .       47      PASS    NS=3;DP=13;AA=T             GT:GQ:DP:HQ     0|0:54:7:56,60
20      123457  microsat1       GTC     G,GTCT  50      PASS    NS=3;DP=9;AA=G              GT:GQ:DP        0/1:35:4
```

https://doi.org/10.1093/bfgp/elv054

# BAM in IGV


IGV
Human (hg38)
chr7
chr7:5,525,148-5,532,601
Go
p22.2 p21.3 p21.2 p15.3 p15.1 p14.2 p14.1 p13 p12.2 p11.2 q11.1 q11.22 q11.23 q21.11 q21.12 q21.2 q22.1 q22.3 q31.1 q31.2 q31.32 q32.1 q33 q34 q35 q36.1
7,415 bp
5,526,000 bp
5,527,000 bp
5,528,000 bp
5,529,000 bp
5,530,000 bp
5,531,000 bp
5,532,000 bp
SortedActB.bam Coverage
[0 - 1306]
SortedActB.bam Junctions
SortedActB.bam
Gene
LOC221946
ACTB


Steps
Preprocessing
Processing
Technique
Downstream analysis
Raw data
Processed data
BAM
Outcome
Benchmarking top 5-10 tools and algorithms
Preprocessing
trimmomatic
cutadapter
Fastqc
Trim Galore
AfterQC
Processing
STAR
TopHat2
HISAT
SOAP
bwa
bowtie
Bulk
Single cell
WGS
Variant calling
freebayes
streka
dkfz
deepvariant
sanger
WGBS
Differential methylation
methpipe
bisulfighter
dmrcate
HOME
dmrcaller
dmrfinder
ChIP-seq
Differential binding site
seesaw
ODIN
Thor
MMDiff
diffbind
Hi-C
Differential interaction
diffhic
Hicompare
FIND
Selfish
RNA-seq
Quantification
RSEM
RPKM
RC
Med
Sailfish
DESeq
Differential expression
EBSeq
ALDEx2
edgeR
DEXUS
DESeq2
DEXSeq
scRNA-seq
Quantification
StringTie
Cufflinks
RSEM
RPKM
Differential expression
SCDE
MAST
scDD
D3E
DEsingle
SINCERA
Result of benchmarking
TrimGalore
STAR alignment
Bulk RNA-seq
Quantification RSEM
Differential expression edgeR
Draft of pipeline
FASTQ
TrimGalore
Processed FASTQ
STAR alignment
BAM
RSEM
Count TSV
edgeR
Differential expressed genes
Optimised pipeline
FASTQ
TrimGalore
Processed FASTQ
Ref
STAR indexing
Ref indexed
STAR alignment
BAM
RSEM
Count TSV
edgeR
Differential expressed genes
bedtools bedgraph2bigwig
BigWig
IGV
Genomics viewer

# Sanger sequencing?

1. Khi nào sử dụng Sanger sequencing (không phải microarray, không phải NGS)?
2. Chiều dài có thể giải?
3. Thời gian giải?
4. Độ chính xác?
5. Lâm sàng có chấp nhận kết quả của Sanger không và tại sao?
6. Sanger sequencers có IVD không?
7. Tiến Lê: Anh có thể cho em hỏi có trường hợp các peak có chiều cao khác nhau hoặc chồng lên nhau? mình xử lý các trường hợp này ntn. Và gtt 16s có phải phân lập ko hay có thể sử dụng mẫu môi trường Như NGS? e cảm ơn!

# Illumina sequencers?

1. Chuẩn bị thư viện Library prep đòi hỏi được đào tạo mức độ nào?
2. Thời gian chuẩn bị thư viện Library prep bao lâu? Nó có phụ thuộc vào hệ máy không ví dụ iSeq100, MiSeq, NextSeq550 và NovaSeq?
3. THUY DUONG THI NGOC: Cho mình hỏi là đối với giải trình tự NGS trên Illunina thì khi đầu tư mua máy Illumina thì cần thêm thiết bị nào khác để cho các bước chuẩn bị mẫu không ah?
4. Base quality có khác nhau ở cách kênh màu 1-Dye/2-Dye/4-Dye?
5. Nguyên tắc của Illumina khác gì so với Sanger mà giải được rất rất nhiều trình tự so với Sanger?
6. Làm sao giải quyết vấn đề Optical Duplication?
7. Việc mua máy Illumina phải kèm theo mua những hóa chất gì và những hóa chất đó dùng cho những bước nào?
8. Lâm sàng có chấp nhận kết quả của Illumina không và tại sao?
9. Basespace có thể dùng để lưu trữ dữ liệu: được miễn phí bao nhiêu TB? Và sẽ tính tiền từ bao nhiêu TB? Và tiền lưu trữ sẽ tính như thế nào?
10. Và basespace phân tích bioinformatics bao gồm những pipeline nào? Miễn phí bao nhiêu, tính tiền bao nhiêu? Cách tính tiền như thế nào?
11. Có software nào thay cho Basespace trong lâm sàn?

# MGI sequencers?

1. Chuẩn bị thư viện Library prep đòi hỏi được đào tạo mức độ nào?
2. Kit convert có đắc không so với việc mua trực tiếp library kit từ MGI?
3. Thời gian chuẩn bị thư viện MGI khoảng bao lâu?
4. Nguyên tắc giải trình tự của MGI là gì và có khác gì so với Sanger?
5. Trung bình base quality của MGI là bao nhiêu?
6. Việc mua máy MGI phải kèm theo mua những hóa chất gì và những hóa chất đó dùng cho những bước nào?
7. MGI có máy cho lâm sàng không và sự chấp nhận kết quả xét nghiệm của MGI?
8. MGI có cloud software như Basespace không? Nếu không thì có gì tương tự?
9. Có software nào dành riêng cho lâm sàn?
10. Nguyễn Huy Thịnh: cho mình hỏi là đối với SBS của MGI thì khi gắn nucleotide đầu tiên (vd như nucleotide A), có nguy cơ nào A này sẽ bám vào các vị trí T khác bất kì trên sợi DNA ngoài vị trí mong muốn trên primer không ạ?
11. Tien Vuong Quang: Hệ máy BGI cho file đầu ra theo định dạng nào thế ạ? Và em muốn hỏi thêm phần mềm phân tích thường được sử dụng là gì?
12. Lan Hương: Trong ứng dụng pháp y, các bộ marker nào được sử dụng trong MGI? Có phù hợp với các chuẩn CODIS của FBI hay không? MGI đã được công nhận như "chuẩn àng Sanger" chưa? Việc phân tích kết quả từ MGI có thể thực hiện trên phần mềm GeneMapper hay GeneMarker hay không? Các loại mẫu có thể thực hiện trên máy MGI? Nếu gặp mẫu hỗn hợp thì MGI có thể giải quyết được không?

# Ion Torrent sequencers?

1. Chuẩn bị thư viện Library prep đòi hỏi được đào tạo mức độ nào?
2. Kit bắt buộc theo máy?
3. Thời gian chuẩn bị thư viện Ion Torrent khoảng bao lâu?
4. Nguyên tắc giải trình tự của Ion Torrent là gì và có khác gì so với Sanger?
5. Trung bình base quality của Ion Torrent là bao nhiêu?
6. Việc mua máy Ion Torrent phải kèm theo mua những hóa chất gì và những hóa chất đó dùng cho những bước nào?
7. Việc mua máy Ion Torrent phải kèm theo mua máy tính phân tích bioinformatics luôn không? Nếu không mua thì có thể tự làm phân tích bioinformatics được không?
8. Ion Torrent có máy cho lâm sàng không và sự chấp nhận kết quả xét nghiệm của Ion Torrent?
9. Ion GeneStudio có cloud software như Basespace không? Nếu không thì có gì tương tự?
10. Có software nào dành riêng cho lâm sàn?
11. Phương pháp H+ chính xác như thế nào so với MGI và Illumina? Chứng minh ntn?
12. duynhat le: cho e hỏi a phước là đối với hệ thống Ion Torrent mình có thể làm library riêng cho các tác nhân rồi pool mẫu vào chạy sequence nhiều tác nhân trong 1 chip dc không ạ
13. Nguyễn Thị Lan Hương: (hỏi Phước) Cải tiến của thế hệ S5 so với Ion Torrent đời trước là gì? (đời cũ toàn làm tay, quá cực). Nếu mua máy thì việc phân tích kết quả hãng có training cho bên mua không?
14. MIFOLAB (TS. Trương Huỳnh Anh Vũ): Cần phải xác nhận giá trị sử dụng phương pháp xác định target cụ thệ nào đó ạ...

# Mix

1. Trần Gia Huy: Đề tài của em cần giải trình tự Whole Genome Sequencing ở thực vật (đã có draft genome trên ncbi) Xin hỏi sử dung công nghệ nào của NGS là tối ưu nhất? Xin cảm ơn.
2. Tô Thị Thùy Ninh: E muốn hỏi về việc Ứng dụng của NGS trong sàng lọc NIPT ạ, trong quá trình sàng lọc bất thường NST 13,18,21 thì theo e được biết Down(+) trong NIPT độ tin cậy sẽ cao hơn so với trisomy13,18(+) thì sự khác biệt ở đây là do đâu ạ?
3. Nhung trả lời: Hi bạn, với hệ thống ILMN thì riêng với NIPT thì server phân tích cùng pipeline phân tích đều được nghiên cứu trên lâm sàng và có IVD. Độ tin cậy của các NST ko có sự quá khác biệt đâu ạ! Ngoài ra với giải pháp mới của ILMN là VeriSeq NIPT v2 thì là phát hiện toàn bộ bất thường số lượng trên tất cả các NST và các vi mất đoạn > 7Mb ạ
4. Nguyễn Phương: Novaseq và máy MGI này là 2 máy giống chức năng nhau hay riêng biệt ạ
5. Tiến Vương Quang: Thưa thầy Lợi, với kinh nghiệm phân tích tin sinh của thầy thì thầy có thể chia sẻ những suy nghĩ, đánh giá / so sánh của thầy về chất lượng đầu ra của các hệ máy giải trình tự từ các hãng khác nhau không ạ?
6. Vy Nguyễn: Em muốn biết thêm về Parse Biosciences được không ạ? Coverage? Depth? Số lượng cells load được mỗi đợt chạy? Cost per cell? Nguyên lý hoạt động? Đặc trưng so với các dòng máy khác của 10X, Smart-seq2? (https://www.youtube.com/watch?v=pVyZvX8N5ww&list=PLXtgXP89Tyn_N5HAs-SWgbMKr_iOIhYrw)

# Mix

1. Oanh Pham: Dạ cho em hỏi trong trường hợp nào thì kết quả NGS cần được confirm lại bằng Sanger ạ?
2. Quynh Nhu Nguyen: Em Xin hỏi lý thuyết cơ bản; Từ kết quả của các máy, minh sẽ có trình tự 1 sợi đơn DNA. Vậy tại sao trong file FASTA lại có read1 và read2. 2 cái read này từ đâu mà có?
3. Xin hỏi anh Lợi là có thể dùng NGS data từ mẫu của 2 genotype khác nhau của cùng 1 loài để làm assembly hay không?