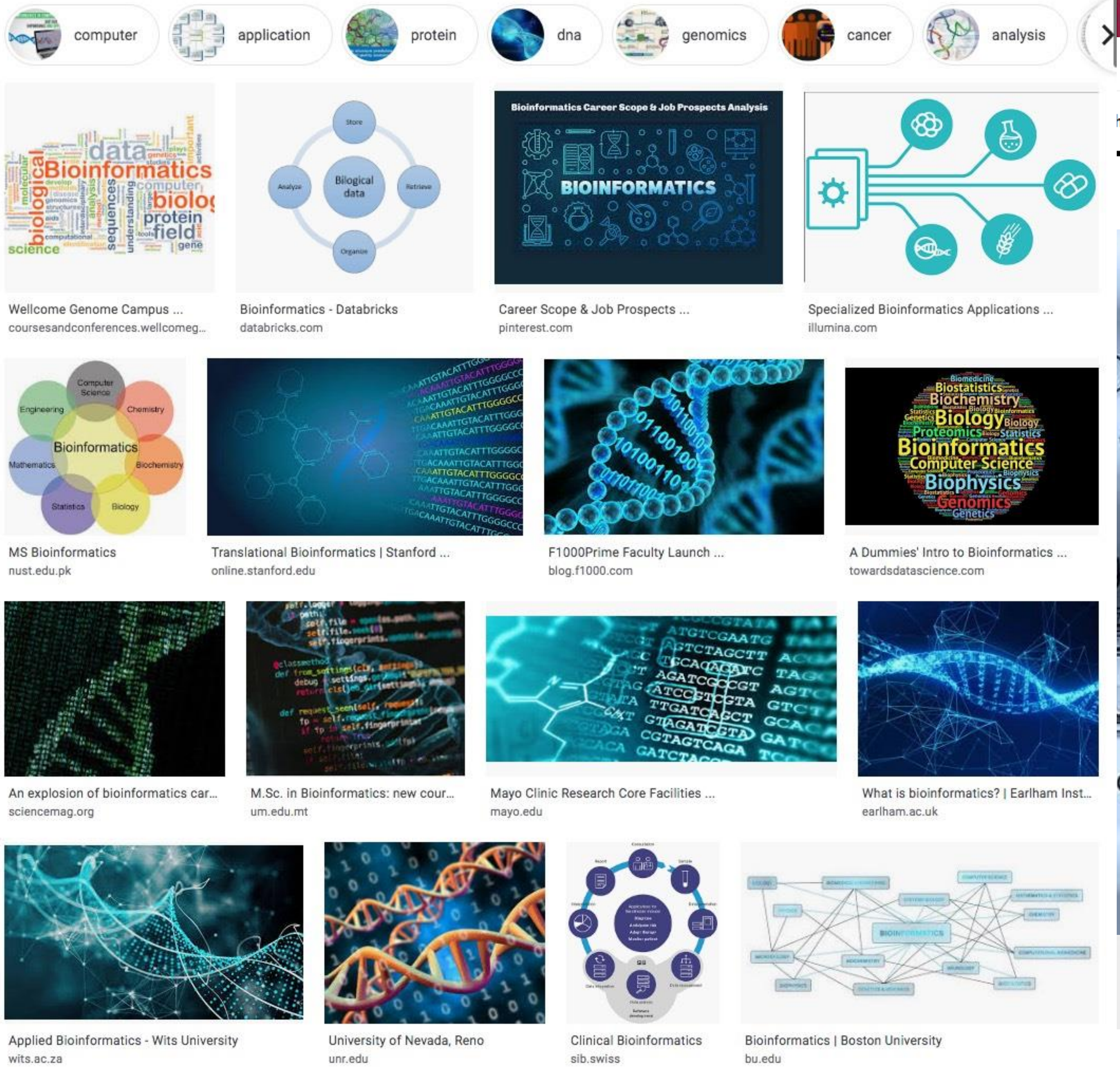
NEAR EAST UNIVERSITY
DESAM RESEARCH INSTITUTE

Third Generation Sequencing Informatics

DESAM Research Institute



Wellcome Genome Campus ...
coursesandconferences.wellcomeg...

Bioinformatics - Databricks
databricks.com

Career Scope & Job Prospects ...
pinterest.com

Specialized Bioinformatics Applications ...
illumina.com

MS Bioinformatics
nust.edu.pk

Translational Bioinformatics | Stanford ...
online.stanford.edu

F1000Prime Faculty Launch ...
blog.f1000.com

A Dummies' Intro to Bioinformatics ...
towardsdatascience.com

An explosion of bioinformatics car...
sciencemag.org

M.Sc. in Bioinformatics: new cour...
um.edu.mt

Mayo Clinic Research Core Facilities ...
mayo.edu

What is bioinformatics? | Earlham Inst...
earlham.ac.uk

Applied Bioinformatics - Wits University
wits.ac.za

University of Nevada, Reno
unr.edu

Clinical Bioinformatics
sib.swiss

Bioinformatics | Boston University
bu.edu

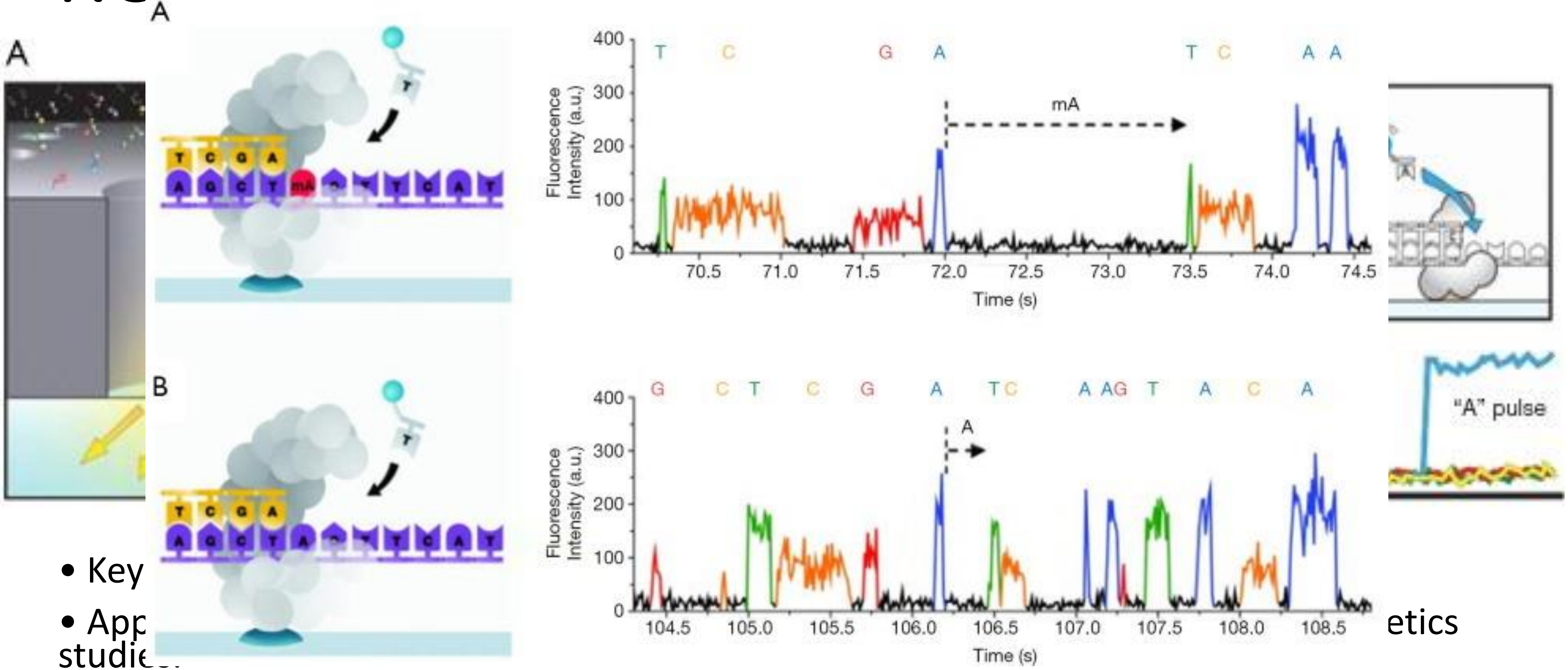
Outline

- NGS Technologies
- Limitations and Future Scope
- 3rd Gen Informatics
- Workflow of Data Analysis
- Definitions, File Types, Processes, and more

What to Expect?

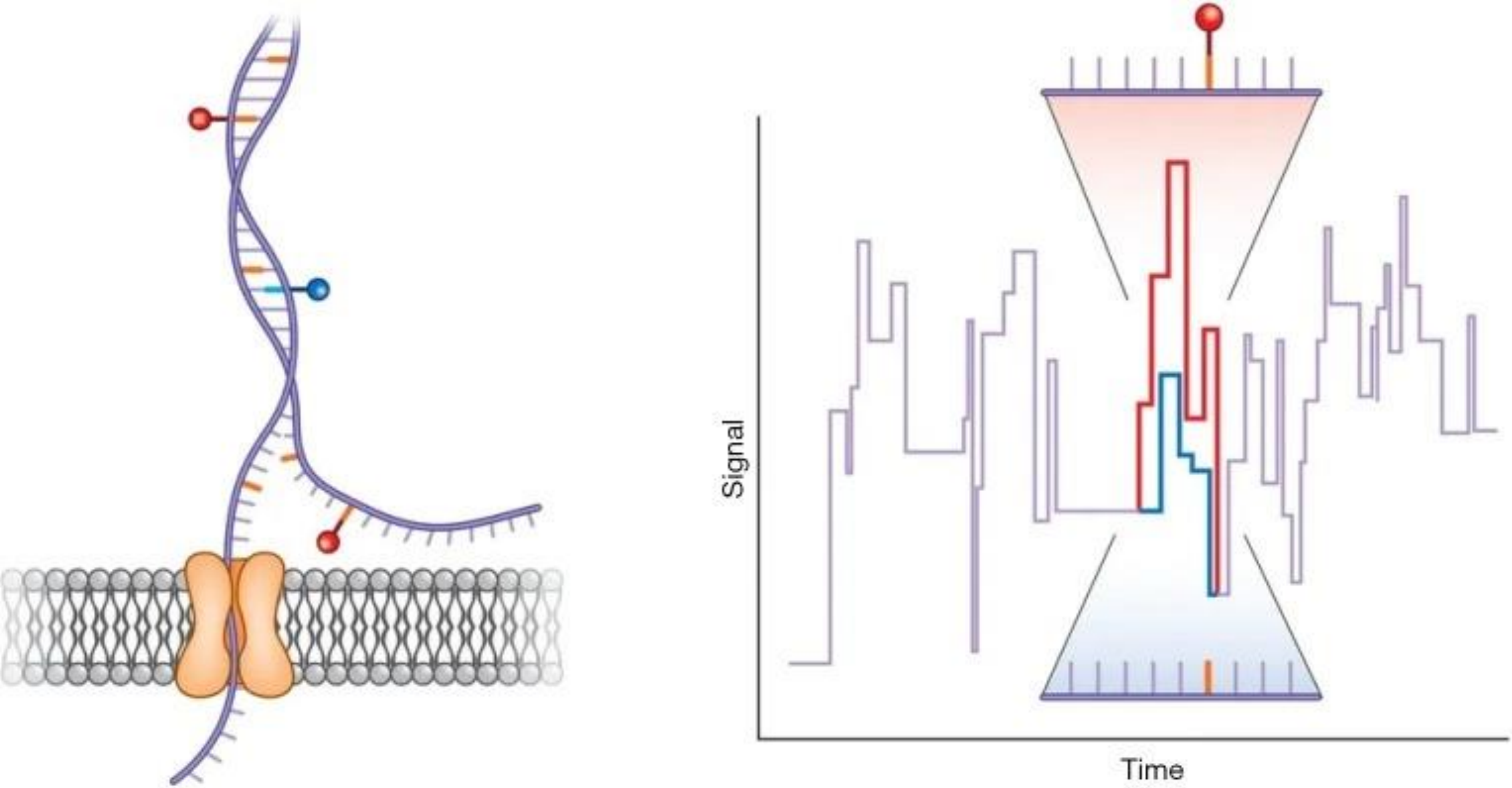
- Basic Knowledge about Sequencing Platforms
- Introduction on Bioinformatics
- How to interpret Raw Data?
- Types of Analysis and what programs we need to choose?
- Applications and Visualizations of the data

NGS Technologies:



- Key
- App etics studie

NGS Te



4. Oxf

- Prin pores.
- Key

Applic

5. 454

- Prin

•Key F

Applic

6. SOL

- Prin
- Key

Applications: Whole exome sequencing, ChIP-Seq.

NGS Technologies:

7. Helicos Sequencing (Single-Molecule Sequencing):

- Principle: Serial imaging of fluorescently labeled nucleotides during synthesis.
- Key Features: Single-molecule sequencing, no amplification bias.

Applications: Gene expression analysis, RNA-Seq.

8. Complete Genomics Sequencing:

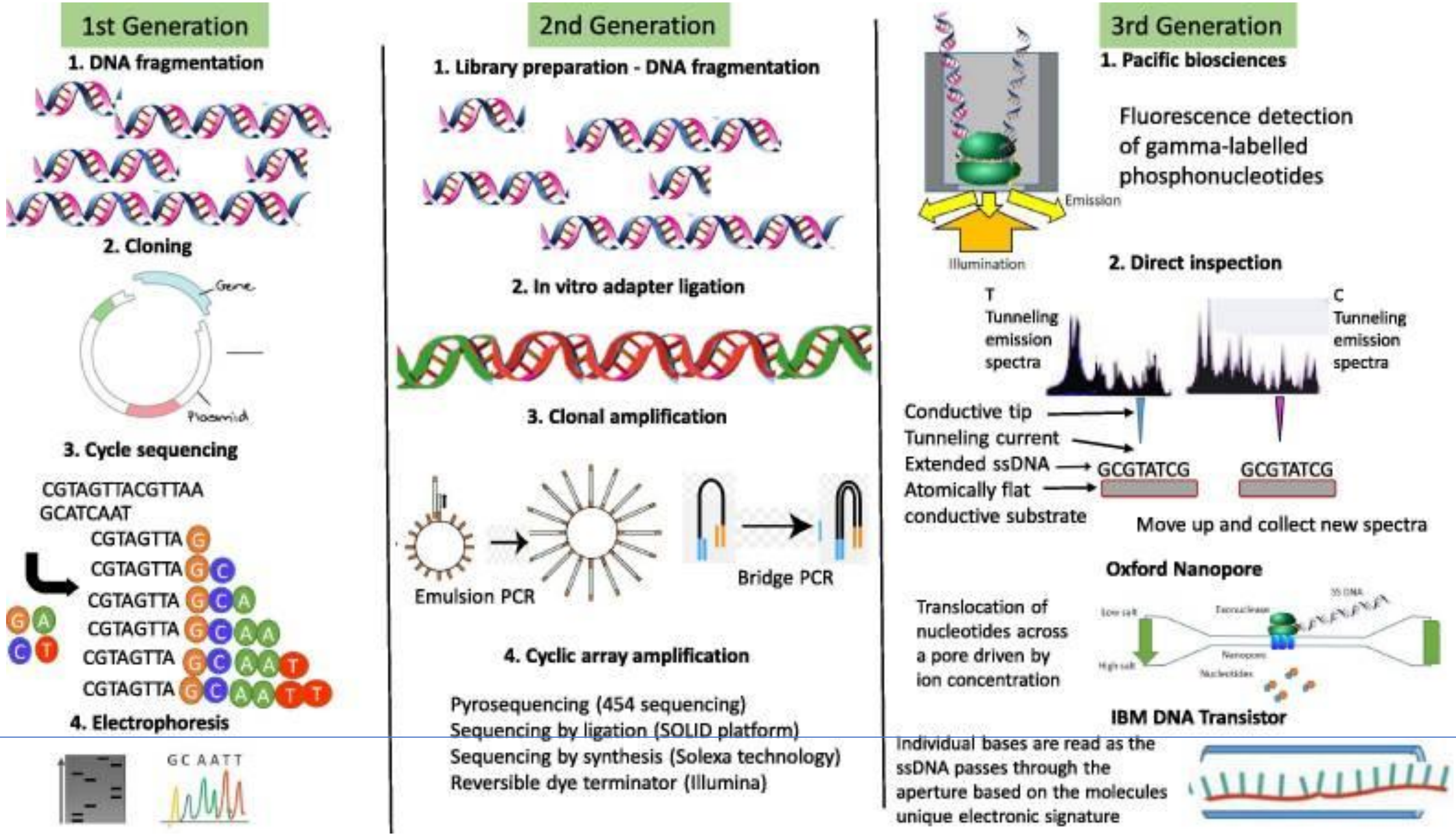
- Principle: Direct sequencing of DNA nanoarrays.
- Key Features: No need for library preparation, cost-effective for large projects.

Applications: Genome sequencing for population studies.

9. BGISEQ Sequencing:

- Principle: DNBSEQ technology using combinatorial probe-anchor synthesis.
- Key Features: High throughput, relatively low cost.

Applications: Genome sequencing, transcriptome analysis.



1st Generation

1. DNA fragmentation

2. Cloning

3. Cycle sequencing

4. Electrophoresis

2nd Generation

1. Library preparation - DNA fragmentation

2. In vitro adapter ligation

3. Clonal amplification

4. Cyclic array amplification

Pyrosequencing (454 sequencing)
Sequencing by ligation (SOLID platform)
Sequencing by synthesis (Solexa technology)
Reversible dye terminator (Illumina)

3rd Generation

1. Pacific biosciences

Fluorescence detection of gamma-labelled phosphonucleotides

2. Direct inspection

T
Tunneling emission spectra
C
Tunneling emission spectra
Conductive tip
Tunneling current
Extended ssDNA
Atomically flat conductive substrate
GCGTATCG
GCGTATCG

Move up and collect new spectra

Oxford Nanopore

Translocation of nucleotides across a pore driven by ion concentration

IBM DNA Transistor

Individual bases are read as the ssDNA passes through the aperture based on the molecules unique electronic signature

Dlamini *et al.*, 2020 - *Comput Struct Biotechnol J.*

Limitations of NGS:

1. Read Length: NGS platforms typically produce short sequence reads, which can make it challenging to assemble genomes with repetitive regions or analyze long-range structural variations.

2. Error Rates: NGS can have error rates, particularly in homopolymer regions (repeating nucleotides), which can affect variant calling accuracy and data quality.

3. Coverage Uniformity: NGS may exhibit uneven coverage across the genome, leading to gaps or regions with low sequencing depth, which can affect variant detection and downstream analysis.

4. Complex Structural Variations: Complex structural variations, such as chromosomal rearrangements and inversions, may be challenging to detect accurately using standard NGS approaches.

5. Repetitive Regions: Highly repetitive genomic regions can be difficult to sequence and assemble, leading to incomplete or inaccurate results.

6. GC Bias: NGS can exhibit GC content bias, resulting in varying sequencing depths for regions with different GC content, potentially impacting data analysis.

7. Sequence Errors: Sequencing errors, including substitutions, insertions, and deletions, can occur and require advanced error correction methods.

Limitations of NGS:

8. Library Preparation Artifacts: Artifacts introduced during library preparation, such as PCR duplicates, can lead to data redundancy and affect variant calling and quantification.

9. High Computational Demands: NGS data analysis demands significant computational resources and expertise, which can be a limitation for some research settings.

10. Data Storage and Management: Managing and storing large volumes of NGS data can be challenging and costly, particularly for long-term storage and data sharing.

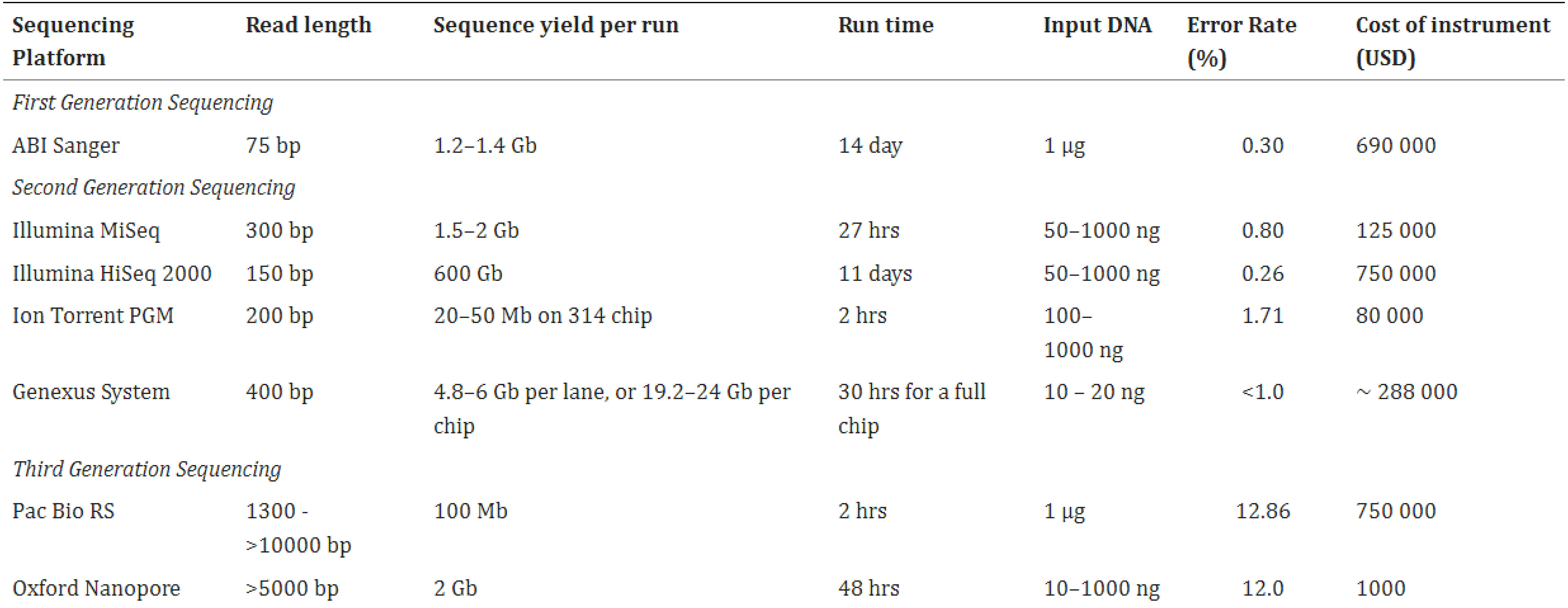
11. Sample Contamination: Contamination of samples during handling or library preparation can lead to incorrect results and necessitate rigorous quality control measures.

12. Ethical and Privacy Concerns: The generation of extensive genetic data raises ethical and privacy concerns, requiring careful data handling and consent procedures.

13. Turnaround Time: NGS workflows can have longer turnaround times compared to other diagnostic methods, impacting clinical applications where rapid results are needed.

14. Limited Detection of Epigenetic Modifications: Standard NGS techniques may not provide comprehensive information about epigenetic modifications, necessitating additional assays for detailed epigenomic studies.

15. Cost: While NGS costs have decreased over the years, it can still be expensive, particularly for large-scale projects or clinical applications.



Sequencing Platform	Read length	Sequence yield per run	Run time	Input DNA	Error Rate (%)	Cost of instrument (USD)
First Generation Sequencing						
ABI Sanger	75 bp	1.2–1.4 Gb	14 day	1 µg	0.30	690 000
Second Generation Sequencing						
Illumina MiSeq	300 bp	1.5–2 Gb	27 hrs	50–1000 ng	0.80	125 000
Illumina HiSeq 2000	150 bp	600 Gb	11 days	50–1000 ng	0.26	750 000
Ion Torrent PGM	200 bp	20–50 Mb on 314 chip	2 hrs	100–1000 ng	1.71	80 000
Genexus System	400 bp	4.8–6 Gb per lane, or 19.2–24 Gb per chip	30 hrs for a full chip	10 – 20 ng	<1.0	~ 288 000
Third Generation Sequencing						
Pac Bio RS	1300 - >10000 bp	100 Mb	2 hrs	1 µg	12.86	750 000
Oxford Nanopore	>5000 bp	2 Gb	48 hrs	10–1000 ng	12.0	1000

Dlamini *et al.*, 2020 - *Comput Struct Biotechnol J.*

Advantages of NGS:

1. High Throughput: NGS platforms can generate vast amounts of sequencing data, allowing researchers to analyze multiple samples simultaneously.

2. Speed: NGS is significantly faster than traditional Sanger sequencing, enabling quicker results for research and clinical applications.

3. Cost-Effective: The per-base cost of sequencing has decreased over the years, making NGS more accessible for various research projects and clinical diagnostics.

4. Whole Genome Sequencing: NGS allows for whole genome sequencing, providing a comprehensive view of an individual's genetic makeup and potential disease risk factors.

5. Customization: NGS can be tailored to specific research questions, enabling targeted sequencing of regions of interest or whole genome/exome sequencing as needed.

Advantages of NGS:

6. Detection of Various Variants: NGS can identify a wide range of genetic variants, including SNPs, indels, CNVs, and structural variations, enhancing its versatility.

7. Applications Across Disciplines: NGS is widely used in genomics, transcriptomics, epigenomics, metagenomics, and more, enabling diverse scientific investigations.

8. Personalized Medicine: NGS is crucial in the field of personalized medicine, where genetic information informs treatment decisions and drug selection.

9. Rare Disease Diagnosis: NGS can help diagnose rare genetic disorders by identifying disease-causing mutations in affected individuals.

10. Research Advancements: NGS has led to significant advances in genetics, genomics, and our understanding of complex diseases and traits.

Disadvantages of NGS:

1. Data Complexity: NGS generates massive volumes of data that require advanced computational and bioinformatics resources for processing and analysis.

2. Short Read Lengths: Most NGS platforms produce short reads, which can be challenging for de novo genome assembly and analyzing repetitive regions.

3. Data Storage: Storing and managing large NGS datasets can be costly and require robust infrastructure.

4. Quality Control: Ensuring data quality is critical, as sequencing errors, artifacts, and biases can impact the accuracy of results.

5. Bioinformatics Expertise: NGS data analysis requires specialized bioinformatics expertise, which may not be readily available to all researchers.

Disadvantages of NGS:

6. GC Bias: NGS can exhibit bias in sequencing GC-rich or GC-poor regions, affecting coverage uniformity.

7. Ethical and Privacy Concerns: The generation of extensive genetic data raises ethical and privacy concerns, necessitating careful data handling and consent procedures.

8. Turnaround Time: NGS workflows can have longer turnaround times compared to other diagnostic methods, which may not be suitable for urgent clinical cases.

9. Sample Contamination: Contamination of samples during handling or library preparation can lead to incorrect results and necessitate rigorous quality control measures.

10. Initial Setup Costs: While the per-base sequencing cost has decreased, the initial setup costs for acquiring NGS instruments and infrastructure can be substantial.

Future Scope of NGS technology:

1. Longer Read Lengths: Continued efforts to increase read lengths will enable more comprehensive genome assembly, better detection of structural variations, and improved characterization of complex regions.

2. Single-Molecule Sequencing: Advancements in single-molecule sequencing technologies, such as Oxford Nanopore, may provide ultra-long reads and real-time sequencing, revolutionizing genomics and metagenomics research.

3. High-Throughput Platforms: The development of high-throughput NGS platforms will enable faster and more cost-effective sequencing, making large-scale genomic projects more accessible.

Future Scope of NGS technology:

1. Longer Read Lengths: Continued efforts to increase read lengths will enable more comprehensive genome assembly, better detection of structural variations, and improved characterization of complex regions.
2. Single-Molecule Sequencing: Advancements in single-molecule sequencing technologies, such as Oxford Nanopore, may provide ultra-long reads and real-time sequencing, revolutionizing genomics and metagenomics research.
3. High-Throughput Platforms: The development of high-throughput NGS platforms will enable faster and more cost-effective sequencing, making large-scale genomic projects more accessible.

Future Scope of NGS technology:

4. Clinical Diagnostics: NGS will play an increasingly significant role in clinical diagnostics, including the identification of rare genetic disorders, monitoring cancer mutations, and guiding personalized treatment plans.

5. Epigenomics and Epitranscriptomics: NGS will continue to unravel the complexities of epigenetic modifications and RNA modifications, shedding light on gene regulation and disease mechanisms.

6. Single-Cell Sequencing: Further refinement of single-cell sequencing techniques will provide insights into cellular heterogeneity, developmental biology, and disease processes at the individual cell level.

7. Metagenomics and Microbiome Research: Advancements in metagenomic analysis will improve our understanding of microbial communities in various environments, leading to applications in health, agriculture, and environmental sciences.

Future Scope of NGS technology:

8. Structural Variation Detection: Enhanced methods for accurately detecting and characterizing structural variations will have implications for understanding genetic diseases and their underlying mechanisms.

9. Functional Genomics: Integrating NGS with functional genomics techniques will enable researchers to decipher gene function and regulatory networks with higher precision.

10. Data Analysis and Interpretation: Continued development of bioinformatics tools and AI-driven approaches will facilitate the analysis and interpretation of complex NGS datasets, making it more accessible to researchers.

11. Global Genomic Initiatives: Expanding global genomic initiatives will involve large-scale population sequencing projects to understand genetic diversity, ancestry, and disease susceptibility on a global scale.

Future Scope of NGS technology:

12. Point-of-Care Sequencing: Portable and miniaturized NGS devices for point-of-care applications, such as rapid pathogen detection, may become more widespread.

13. Ethical and Regulatory Frameworks: As NGS continues to generate vast amounts of genetic data, the development of robust ethical guidelines and regulatory frameworks for data privacy and security will be essential.

14. Environmental Genomics: NGS will be instrumental in monitoring and managing ecosystems and biodiversity, helping with conservation efforts and ecological research.

15. Therapeutic Insights: NGS will contribute to personalized medicine by identifying drug targets, optimizing treatment strategies, and predicting individual responses to therapies.

16. Emerging Sequencing Technologies: Novel sequencing technologies currently in development may bring unforeseen advancements and capabilities to genomics research.

Future Advancements and Directions:

1. Precision Medicine: Personalized treatment plans based on an individual's genomic, transcriptomic, and proteomic data will become more common, improving healthcare outcomes.

2. Functional Genomics: Advancements in functional genomics, including CRISPR-based gene editing and high-throughput screening, will enable a deeper understanding of gene function and regulation.

3. Epigenomics and Epitranscriptomics: Further exploration of epigenetic modifications and RNA modifications will provide insights into gene regulation, development, and disease.

4. Single-Cell Omics: Single-cell genomics, transcriptomics, and proteomics will reveal cellular

heterogeneity and help understand complex biological processes.

5. Multi-Omics Integration: Integrating genomics, transcriptomics, proteomics, and metabolomics data will

provide a holistic view of biological systems and disease mechanisms.

6. Artificial Intelligence (AI) and Machine Learning: Al-driven data analysis, predictive modeling, and

- drug discovery will accelerate genomics research and enable data-driven healthcare.

Future Advancements and Directions:

7. Metagenomics and Microbiome Research: Continued study of the human microbiome and environmental metagenomics will impact health, agriculture, and ecological research.

8. Long-Read Sequencing: Advancements in long-read sequencing technologies will improve genome assembly, structural variation detection, and the study of complex genomic regions.

9. Nanopore Sequencing: Further developments in nanopore sequencing will lead to ultra-long reads, real- time sequencing, and portable devices with broader applications.

10. Single-Molecule Sequencing: Refinement of single-molecule sequencing techniques will provide unprecedented insights into genomics and molecular biology.

11. Drug Discovery and Development: Genomics will play a pivotal role in identifying new drug targets, optimizing drug development, and predicting patient responses to treatments.

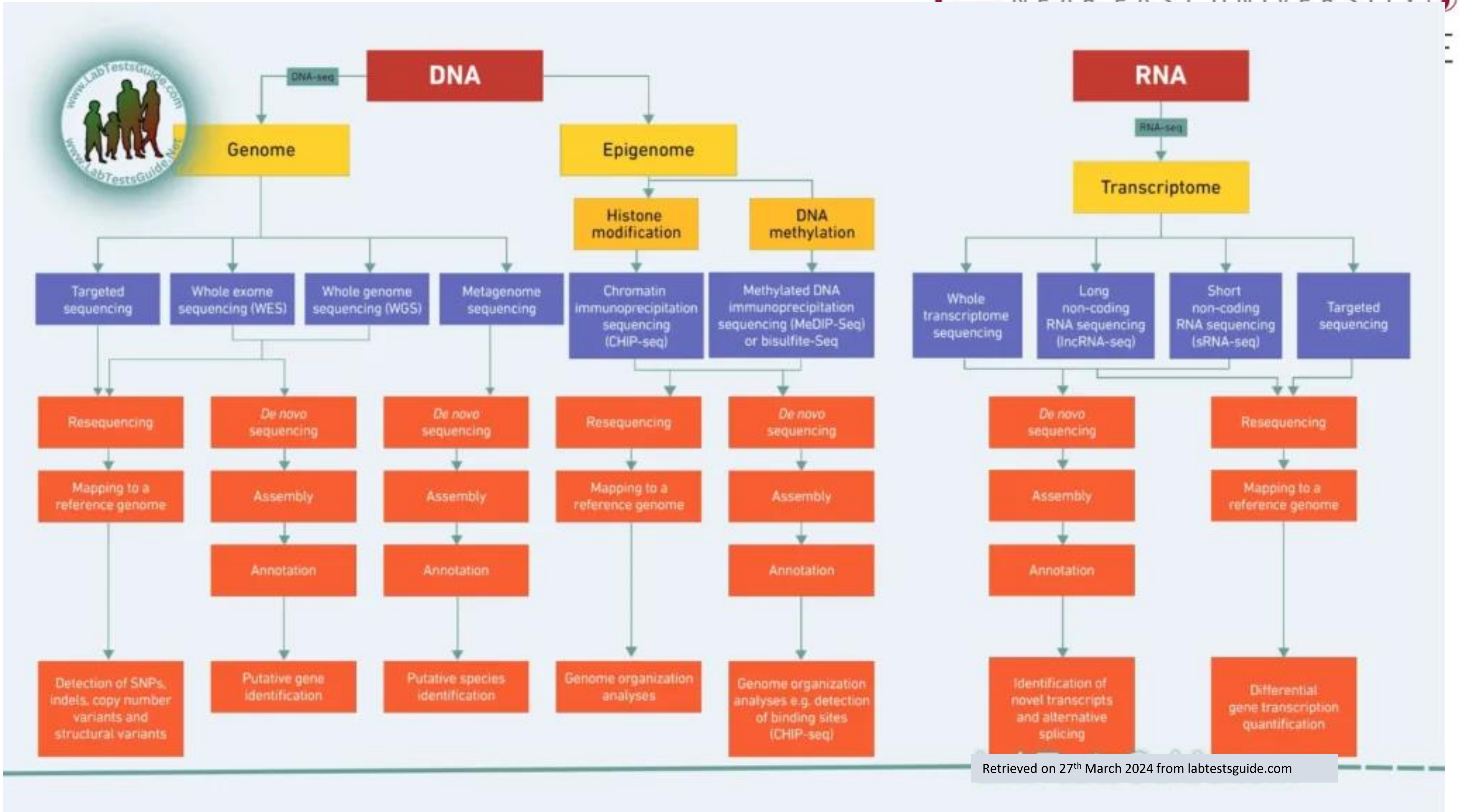
12. Cancer Genomics: Continued research into cancer genomics will lead to better diagnostics, targeted therapies, and early detection methods.

First generation → **Second generation** (next generation sequencing) → **Third generation**

First generation	Second generation (next generation sequencing)	Third generation
Sanger sequencing Maxam and Gilbert Sanger chain termination	454, Solexa, Ion Torrent, Illumina	PacBio Oxford Nanopore
Infer nucleotide identity using dNTPs, then visualize with electrophoresis	High throughput from the parallelization of sequencing reactions	Sequence native DNA in real time with single-molecule resolution
500–1,000 bp fragments	~50–500 bp fragments	Tens of kb fragments, on average

Short-read sequencing (first and second generation)

Long-read sequencing (third generation)



DNA

DNA-seq

Genome

Epigenome

Histone modification

DNA methylation

Targeted sequencing

Whole exome sequencing (WES)

Whole genome sequencing (WGS)

Metagenome sequencing

Chromatin immunoprecipitation sequencing (CHIP-seq)

Methylated DNA immunoprecipitation sequencing (MeDIP-Seq) or bisulfite-Seq

Resequencing → Mapping to a reference genome → Detection of SNPs, indels, copy number variants and structural variants

De novo sequencing → Assembly → Annotation → Putative gene identification

De novo sequencing → Assembly → Annotation → Putative species identification

Resequencing → Mapping to a reference genome → Genome organization analyses

De novo sequencing → Assembly → Annotation → Genome organization analyses e.g. detection of binding sites (CHIP-seq)

RNA

RNA-seq

Transcriptome

Whole transcriptome sequencing

Long non-coding RNA sequencing (lncRNA-seq)

Short non-coding RNA sequencing (sRNA-seq)

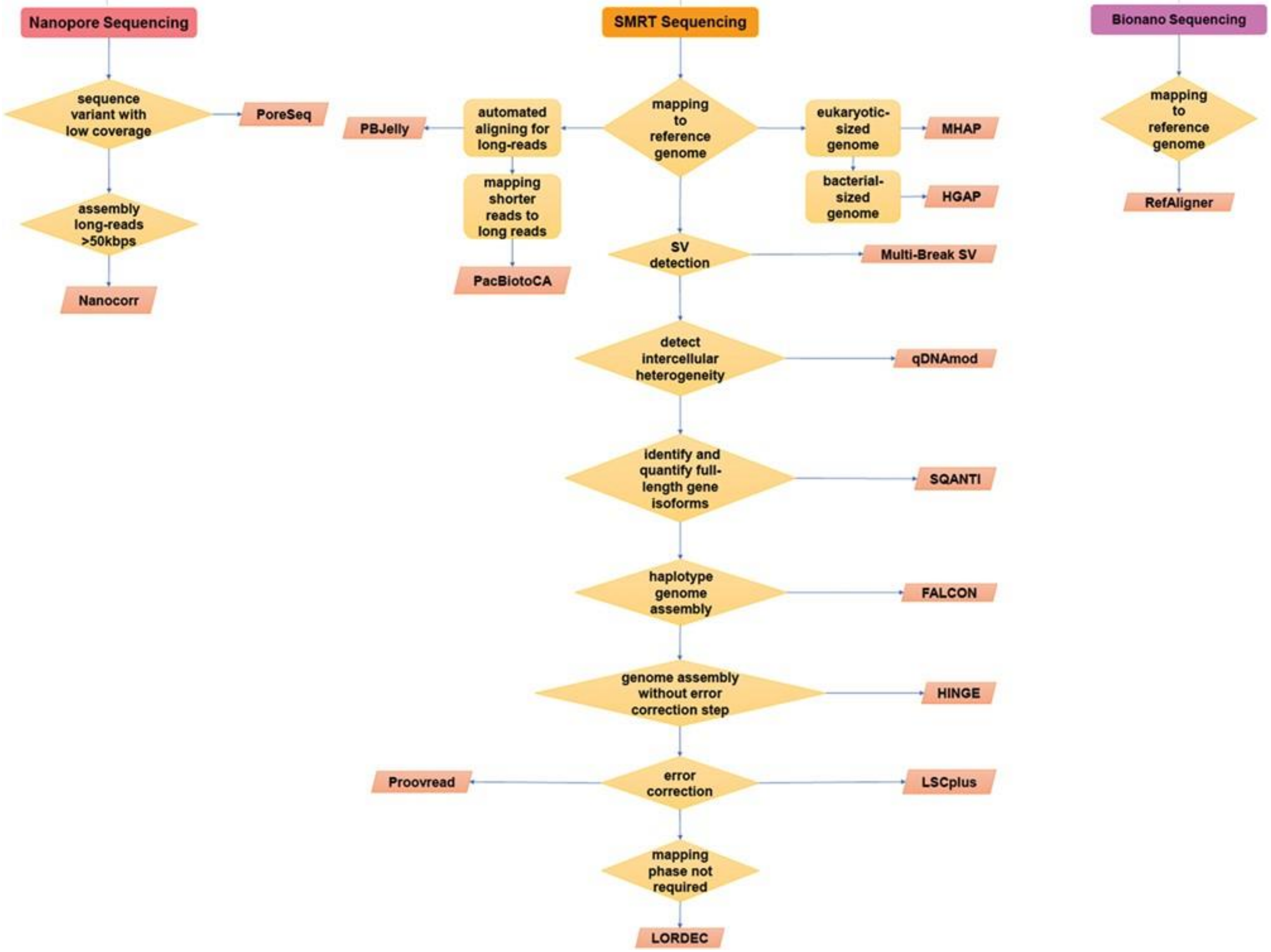
Targeted sequencing

De novo sequencing → Assembly → Annotation → Identification of novel transcripts and alternative splicing

Resequencing → Mapping to a reference genome → Differential gene transcription quantification

Retrieved on 27th March 2024 from labtestsguide.com

De Novo tools



Nanopore Sequencing

sequence variant with low coverage → PoreSeq

assembly long-reads >50kbps → Nanocorr

SMRT Sequencing

mapping to reference genome → automated aligning for long-reads → PBJelly

mapping shorter reads to long reads → PacBiotoCA

mapping to reference genome → eukaryotic-sized genome → MHAP

bacterial-sized genome → HGAP

SV detection → Multi-Break SV

detect intercellular heterogeneity → qDNAmod

identify and quantify full-length gene isoforms → SQANTI

haplotype genome assembly → FALCON

genome assembly without error correction step → HINGE

error correction → Proovread

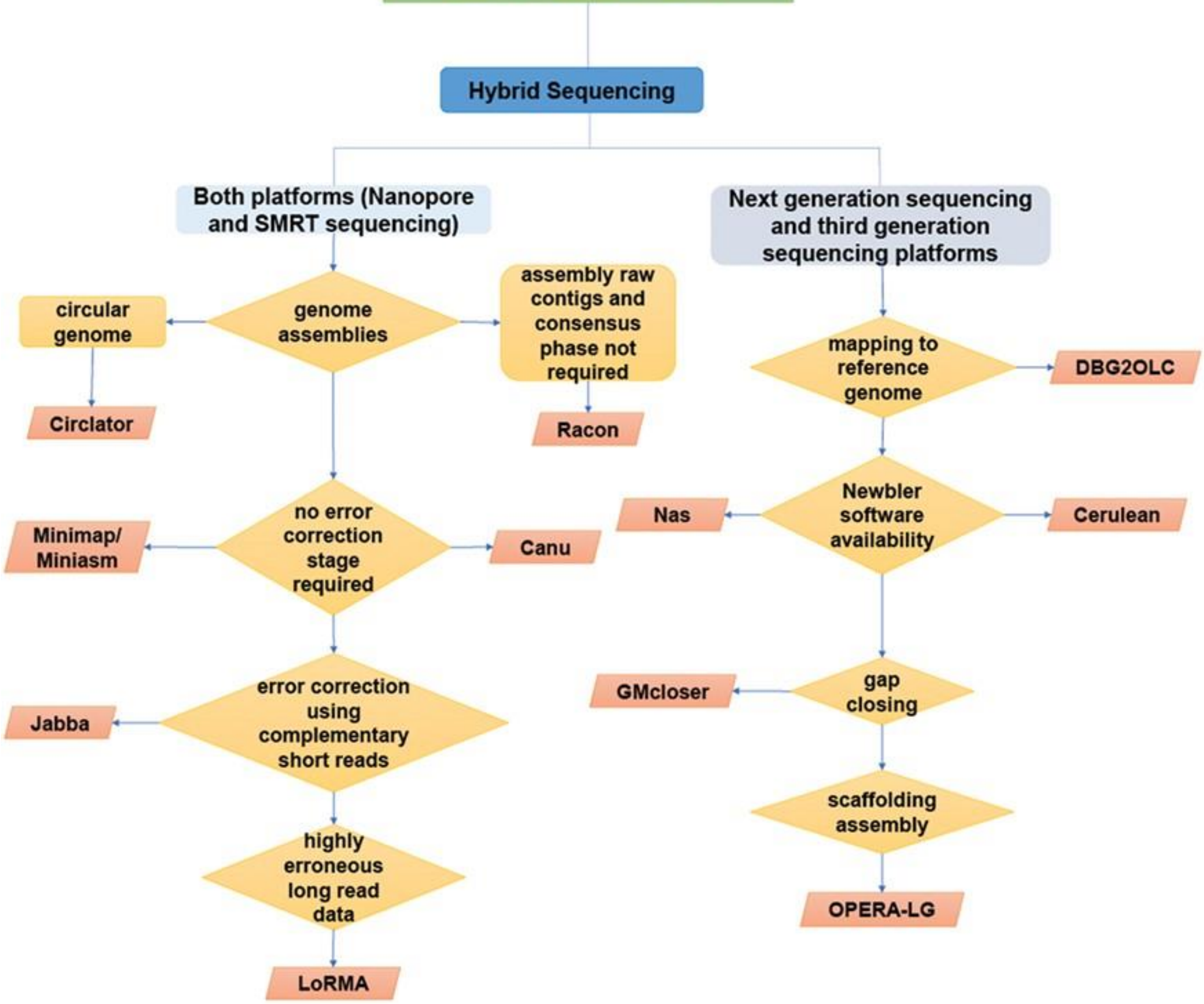
error correction → LSCplus

mapping phase not required → LORDEC

Bionano Sequencing

mapping to reference genome → RefAligner

De Novo tools



Hybrid Sequencing

Both platforms (Nanopore and SMRT sequencing)

genome assemblies

circular genome

Circlator

assembly raw contigs and consensus phase not required

Racon

no error correction stage required

Minimap/ Miniasm

Canu

error correction using complementary short reads

Jabba

highly erroneous long read data

LoRMA

Next generation sequencing and third generation sequencing platforms

mapping to reference genome

DBG2OLC

Newbler software availability

Nas

Cerulean

gap closing

GMcloser

scaffolding assembly

OPERA-LG

3rd Gen Informatics

- **Base calling:** Software process the raw data to call nucleotide bases and generate sequence reads.
- **Quality Control:** Quality control metrics are applied to assess the accuracy and reliability of the sequencing data.
- **Read Alignment:** Sequence reads are aligned or mapped to a reference genome or transcriptome to determine their genomic or transcriptomic locations.
- **Variant Calling:** Variants, such as single nucleotide polymorphism (SNPs) and insertions/deletions (indels) are identified by comparing the sequenced DNA to a reference sequence.
- **Bioinformatics Analysis:** Various bioinformatics tools and pipelines are used for downstream analysis, including gene expression quantification, pathway analysis, and functional annotation.
- **Data Interpretation:** Researchers and clinicians interpret the NGS data to gain insights into genetic variations, gene expression patterns, epigenetic modifications, and more to elucidate disease mechanisms, identify therapeutic targets, or diagnosing genetic disorders.

What is Trimming?

- Rem ng, or
  clipp ore
  than large
  choi

Palindrome Mode

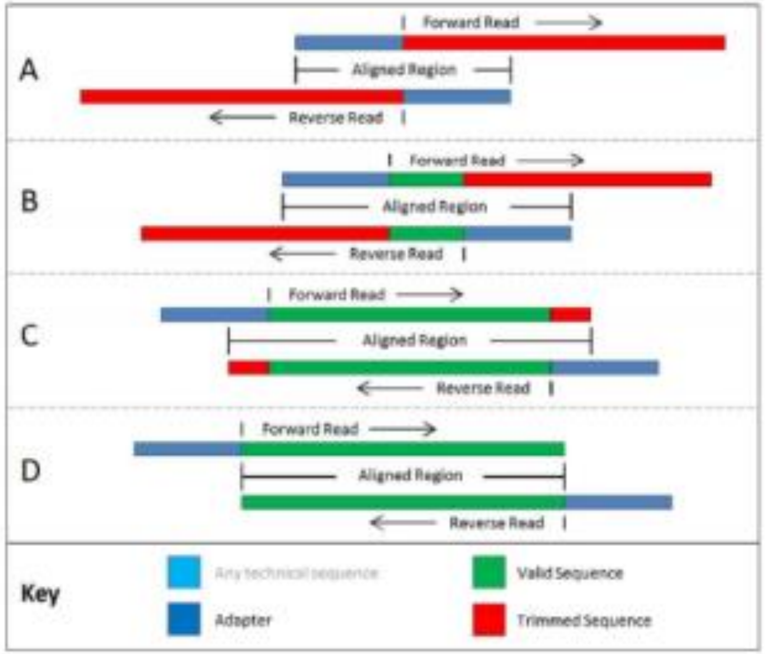


Figure modified from Bolger et al. 2014 (see link below). Caption from Bolger et al. 2014.

Fig. 2. Putative sequence alignments as tested in Palindrome Mode. The alignment process begins with the adapters completely overlapping the reads (A) testing for immediate 'read-though', then proceeds by checking for later overlap (B), including partial adapter read-though (C), finishing when the overlap indicates no read-through into the adapters (D).

https://academic.oup.com/bioinformatics/article/30/15/2114/2390096/Trimmomatic-a-flexible-trimmer-for-Illumina

sequencing cycles: 200

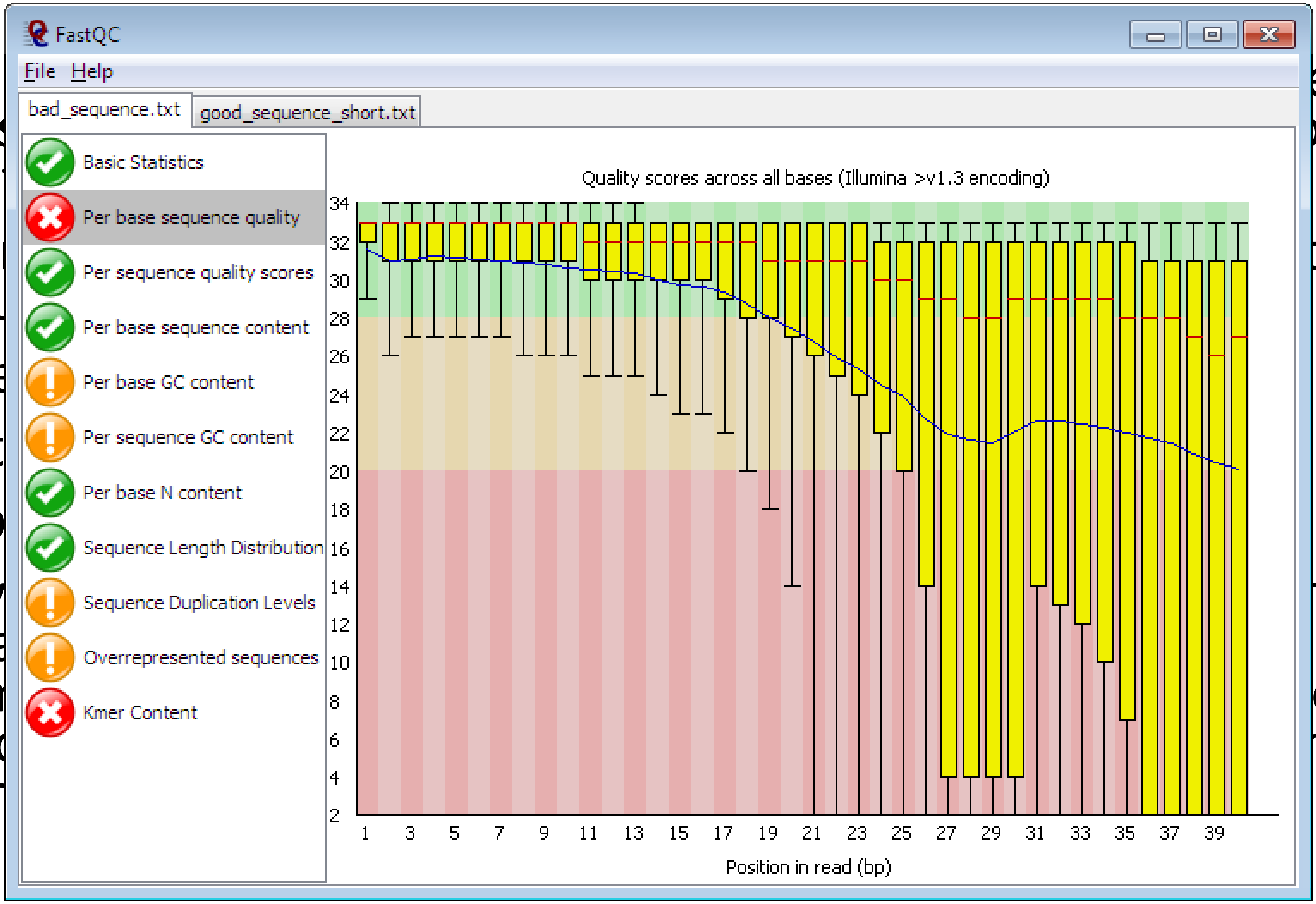
What is Quality Control?

Conducting ents of
data process l results,
are crucial s

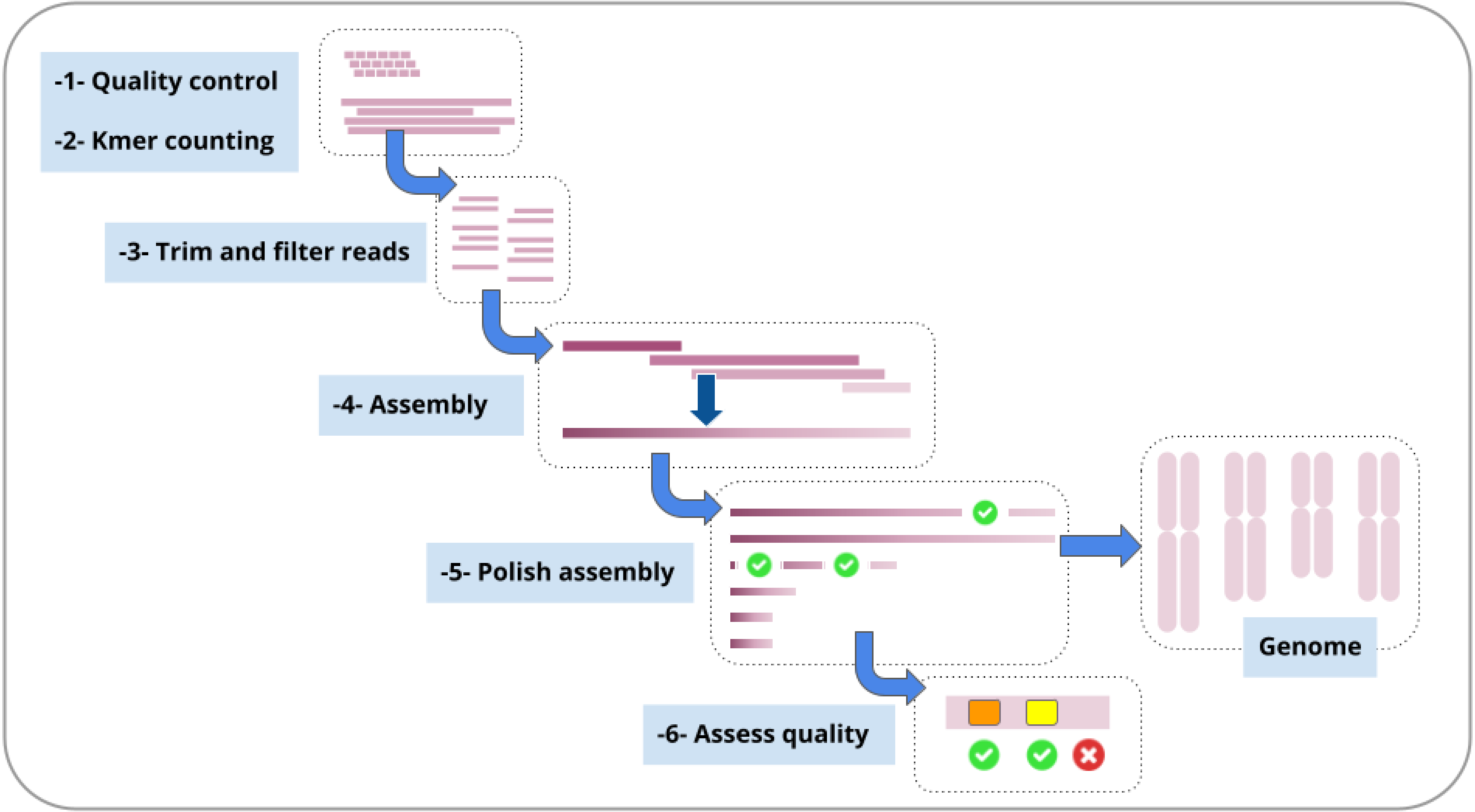
In order to i trols
should evalu

- Raw seque
- Alignment
- Reads dup

FastQC prov n raw
sequence da . It
provides a m ck
impression c hould be
aware befor



What is assembly polishing?



- bly,

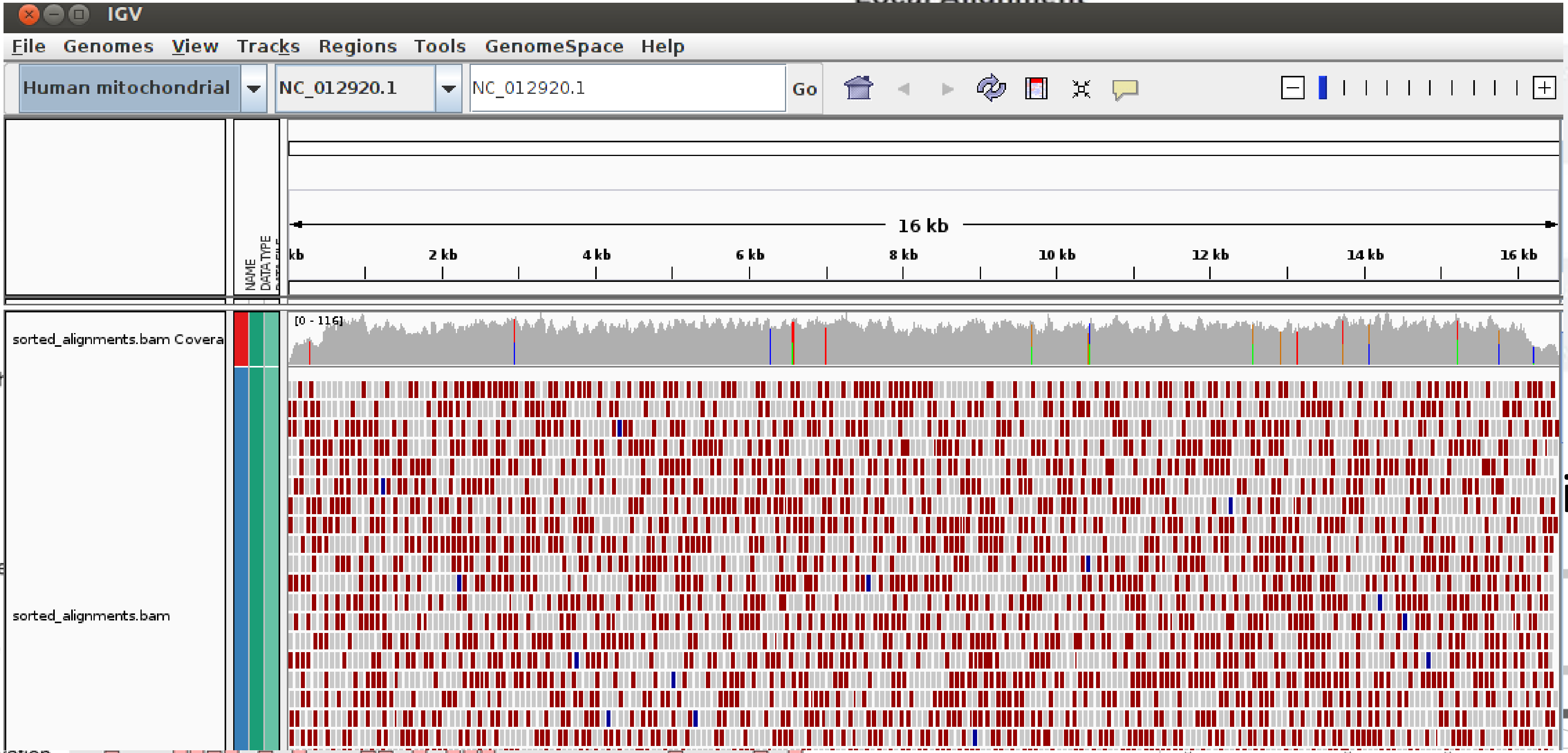
  ıbly

  racy of

What is Alignment/Mapping?

Local Alignment

•



3'

t

3' al

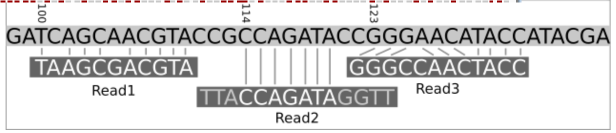
3'

ith

Conse

Sequence

Conservation



File

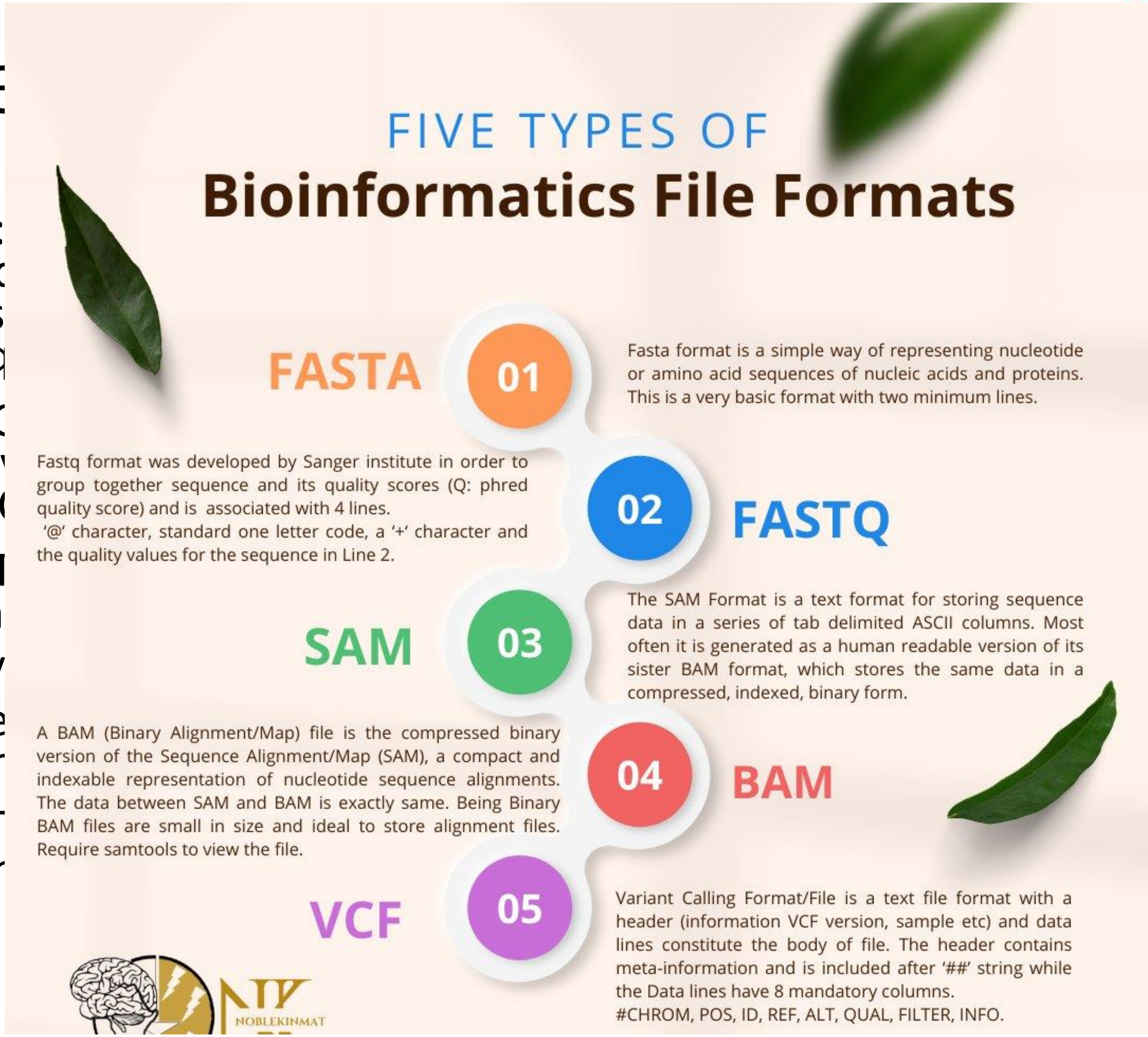
Fasta: ther nucleotide sequences or
peptic resented using single-letter
codes escription, followed by lines
of seq

GFF: C that holds information any
and ev ein sequence. Everything
from (be handled by this format.

VCF: T t used in bioinformatics for
storin

BAM/ e raw data of genome
seque esentation of the Sequence
Alignr

BED: T rmat used to store genomic
regior



FIVE TYPES OF

Bioinformatics File Formats

FASTA 01

Fasta format is a simple way of representing nucleotide or amino acid sequences of nucleic acids and proteins. This is a very basic format with two minimum lines.

02 FASTQ

Fastq format was developed by Sanger institute in order to group together sequence and its quality scores (Q: phred quality score) and is associated with 4 lines.
'@' character, standard one letter code, a '+' character and the quality values for the sequence in Line 2.

SAM 03

The SAM Format is a text format for storing sequence data in a series of tab delimited ASCII columns. Most often it is generated as a human readable version of its sister BAM format, which stores the same data in a compressed, indexed, binary form.

04 BAM

A BAM (Binary Alignment/Map) file is the compressed binary version of the Sequence Alignment/Map (SAM), a compact and indexable representation of nucleotide sequence alignments. The data between SAM and BAM is exactly same. Being Binary BAM files are small in size and ideal to store alignment files. Require samtools to view the file.

VCF 05

Variant Calling Format/File is a text file format with a header (information VCF version, sample etc) and data lines constitute the body of file. The header contains meta-information and is included after '##' string while the Data lines have 8 mandatory columns.
#CHROM, POS, ID, REF, ALT, QUAL, FILTER, INFO.

Principles of 3rd Gen Informatics

1. Data Preprocessing:
   - Quality Control: Bioinformatics tools assess the quality of sequencing data by checking for base call accuracy, sequence read length distribution, and other quality metrics.
   - Adapter Trimming: Sequencing adapters, which are added during library preparation, are removed to ensure accurate downstream analysis.
   - Read Filtering: Low-quality reads, duplicate reads, and contaminating reads are filtered out to improve data quality.
2. Sequence Alignment:
   - Reference Alignment: Sequence reads are aligned or mapped to a reference genome or transcriptome. Bioinformatics algorithms ensure that each read is correctly positioned on the reference, allowing researchers to identify genomic or transcriptomic locations.
3. Variant Calling:
   - Single Nucleotide Polymorphisms (SNPs) and Indels: Bioinformatics tools identify genetic variations, such as SNPs and indels, by comparing the sequenced DNA to the reference genome. Variant calling algorithms assess the likelihood of each variant.
   - Structural Variations: Detection of larger structural variations, such as insertions, deletions, duplications, and translocations, is performed using specialized bioinformatics tools.

Workflow of 3rd Gen Informatics

4. Gene Expression Analysis (RNA-Seq):

- Quantification: Bioinformatics tools quantify gene expression levels, isoform usage, and transcript abundances from RNA-Seq data.
- Differential Expression: Researchers identify genes that are differentially expressed between experimental conditions (e.g., diseased vs. healthy samples) to understand regulatory mechanisms and disease pathways.

5. Functional Annotation:

- Bioinformatics tools annotate sequenced genes and variants to provide functional context, including gene ontology, pathway analysis, and identification of protein domains and motifs.

6. Epigenomics Analysis:

- Bioinformatics methods are used to analyze DNA methylation, histone modifications, and chromatin accessibility data, providing insights into epigenetic regulation.

Workflow of 3rd Gen Informatics

7. Metagenomics:
    - Bioinformatics tools analyze metagenomic data to identify and characterize microorganisms in complex microbial communities, assess microbial diversity, and study their functional potential.
8. Visualization:
    - Data visualization tools and software generate plots, heatmaps, and graphs to help researchers and clinicians interpret and communicate their findings effectively.
9. Integration with Other Omics Data:
    - Bioinformatics facilitates the integration of NGS data with other omics data, such as proteomics and metabolomics, to provide a more comprehensive understanding of biological systems.

Workflow of 3rd Gen Informatics

10. Software and Pipelines:

    • Bioinformatics pipelines, often implemented in software packages (e.g., BWA, GATK, STAR), automate data analysis workflows and ensure reproducibility.

11. Custom Analysis:

    • Depending on the specific research question, bioinformaticians may develop custom scripts and

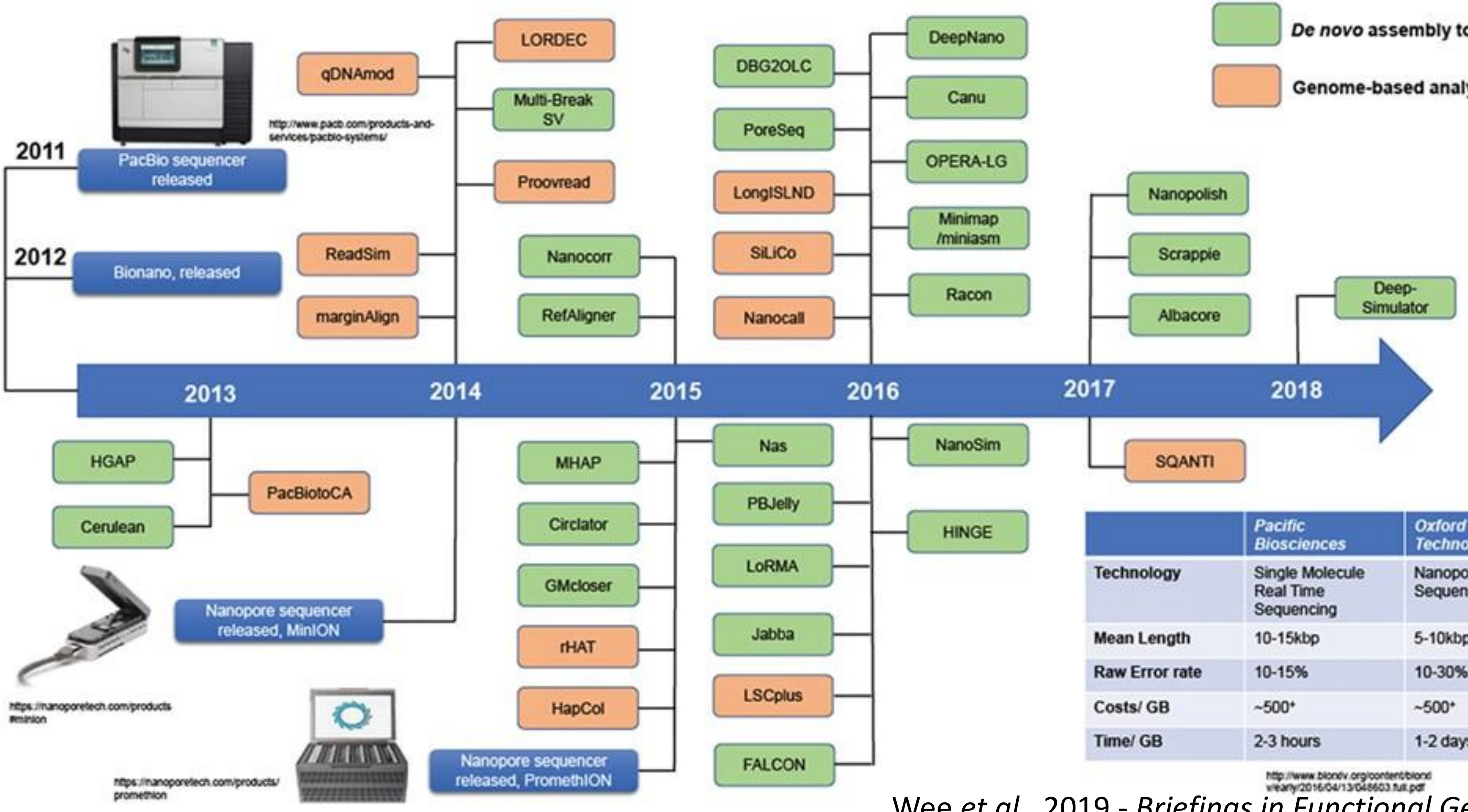
    algorithms to address unique analytical challenges.

12. Data Sharing and Databases:

    • Bioinformatics resources, databases, and data repositories (e.g., NCBI, ENCODE, GenBank) store and provide access to NGS data for the research community.

🟩 *De novo* assembly tools

🟧 Genome-based analysis tools



2011 PacBio sequencer released

http://www.pacb.com/products-and-services/pacbio-systems/

2012 Bionano, released

2013: HGAP, Cerulean, PacBiotoCA

Nanopore sequencer released, MinION

https://nanoporetech.com/products#minion

2014: qDNAmod, LORDEC, Multi-Break SV, Proovread, ReadSim, marginAlign; MHAP, Circlator, GMcloser, rHAT, HapCol

Nanopore sequencer released, PromethION

https://nanoporetech.com/products/promethion

2015: Nanocorr, RefAligner; Nas, PBJelly, LoRMA, Jabba, LSCplus, FALCON

2016: DBG2OLC, PoreSeq, LongISLND, SiLiCo, Nanocall, DeepNano, Canu, OPERA-LG, Minimap /miniasm, Racon; NanoSim, HINGE

2017: Nanopolish, Scrappie, Albacore; SQANTI

2018: Deep-Simulator

	Pacific Biosciences	*Oxford Nanopore Technologies*
Technology	Single Molecule Real Time Sequencing	Nanopore Sequencing
Mean Length	10-15kbp	5-10kbp
Raw Error rate	10-15%	10-30%
Costs/ GB	~500^{+}	~500^{+}
Time/ GB	2-3 hours	1-2 days

http://www.biorxiv.org/content/biorxiv/early/2016/04/13/048603.full.pdf

Wee *et al.*, 2019 - *Briefings in Functional Genomics*

Analysis	Methods	Platforms	Applications
Mapping and alignment	MHAP	ONT and PacBio	De novo mutations and SVs detection
	Minimap	ONT and PacBio	
	DALIGNER	ONT and PacBio	
	Canu	ONT and PacBio	
	FALCON	PacBio	
	Hinge	PacBio	
	MECAT	ONT and PacBio	
	Miniasm	ONT and PacBio	
	Spades	ONT and PacBio	
	HGAP	PacBio	
	Flye	PacBio	
	MARVEL	ONT and PacBio	
	LINKS	ONT and PacBio	
	npScarf	ONT	
	RAILLLS	ONT and PacBio	
	PBJelly	PacBio	
	Ouiver	PacBio	
	Racon	ONT and PacBio	
	BLASR	PacBio	
	BWA-MEM	ONT and PacBio	
	GraphMap	ONT and PacBio	
	LASMSA / LAST	ONT and PacBio	
	Minimap2	ONT and PacBio	
	NGMLR	ONT and PacBio	
	PBHoney	PacBio	
	SMRT-SV	PacBio	
	Sniffles	ONT and PacBio	
	HapCut2	ONT and PacBio	
	WhatsHap	ONT and PacBio	
	SIVM	PacBio	
	NextSV	PacBio	
	NanoSV	ONT	
	Picky	ONT	
	SQANTI	ONT and PacBio	RNA sequencing analysis
	TAPIS	PacBio	
	ToFU	PacBio	
	BLAT	ONT	
	Gmap	PacBio	
	BaseMods	PacBio	Methylation analysis
	Nanopolish	ONT	
	SignalAlign	ONT	
Error correction	Nanocorr	ONT	De novo
	MaSuRCA	PacBio	
	PBcR	PacBio	
	Spades	PacBio and ONT	
	FALCON-sense	PacBio	
	Pbdagcon	PacBio	

What are Genetic Variants?

1. Single Nucleotide Polymorphisms (SNPs): SNPs are single base pair changes in the DNA sequence, representing the most common type of genetic variation in the human genome. They can occur at specific positions in the genome and are associated with diverse traits and diseases.

2. Insertions and Deletions (Indels): Indels involve the insertion or deletion of one or more nucleotides in the DNA sequence, leading to length variations at specific genomic locations.

3. Copy Number Variations (CNVs): CNVs are structural variations characterized by changes in the number of copies of a DNA segment. They can include gene duplications, deletions, and complex rearrangements. 4. Tandem Repeat Variations: These variations consist of short DNA sequences repeated consecutively within a genomic region. The number of repeats can vary among individuals and impact gene expression and phenotypic traits.

5. Structural Variants (SVs): SVs are larger-scale variations involving the rearrangement, duplication, or deletion of significant DNA segments. They encompass translocations, inversions, and large insertions or deletions.

6. Point Mutations: Point mutations are changes in a single nucleotide, which can be transitions

(substitution of a purine with another purine or a pyrimidine with another pyrimidine) or transversions (substitution of a purine with a pyrimidine or vice versa).

What are Genetic Variants?

7. Frameshift Mutations: Frameshift mutations occur when nucleotides are inserted or deleted from a DNA sequence in a way that shifts the reading frame during translation, potentially resulting in non-functional proteins.

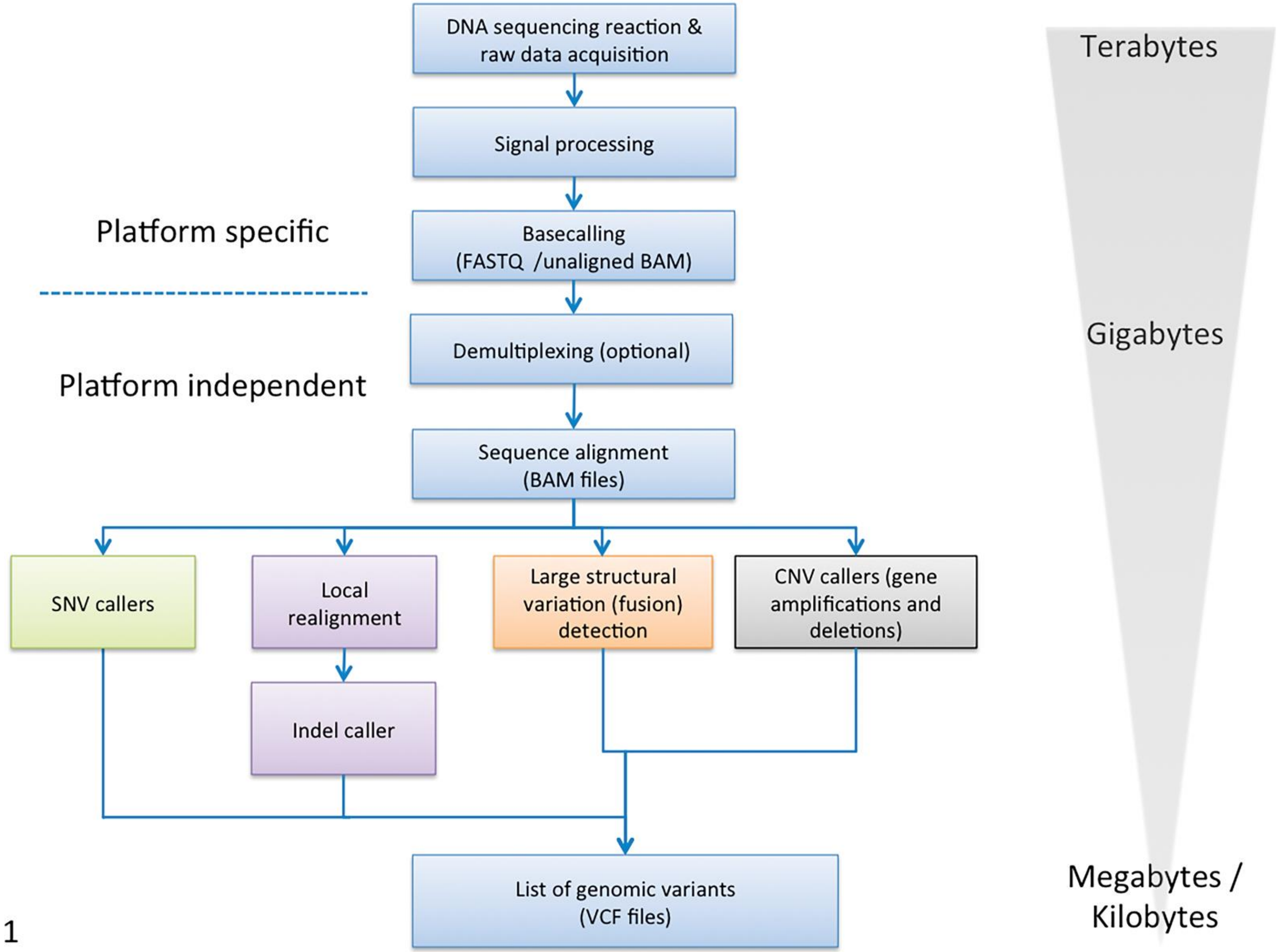
8. Missense Mutations: Missense mutations are point mutations that change a single nucleotide, leading to the substitution of one amino acid for another in the encoded protein. They can affect protein structure and function.

9. Nonsense Mutations: Nonsense mutations are point mutations that create premature stop codons in the coding sequence, resulting in truncated and often non-functional proteins.

10. Silent Mutations: Silent mutations are point mutations that do not change the amino acid sequence of the encoded protein due to the redundancy of the genetic code. They typically have no discernible phenotypic effect.

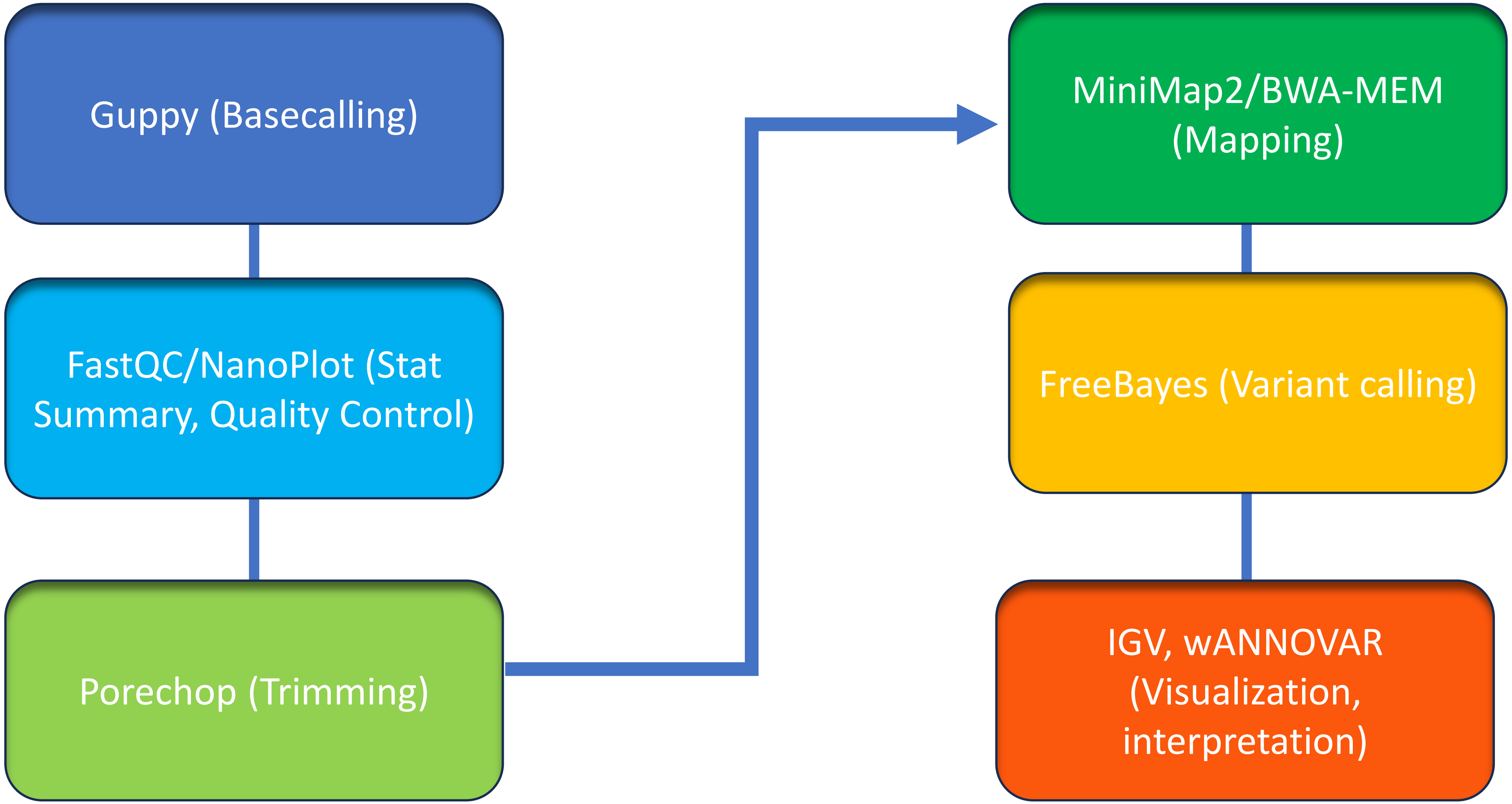
11. Splice Site Mutations: Splice site mutations affect the conserved sequences at exon-intron junctions, leading to altered mRNA splicing and potentially non-functional or dysfunctional protein products. 12. Intronic Variants: Variants located within introns (non-coding regions) of genes can influence gene expression and regulation by affecting splicing, transcription, or other regulatory elements.

NGS data analysis pipeline


File size range
DNA sequencing reaction & raw data acquisition
Signal processing
Platform specific
Basecalling (FASTQ /unaligned BAM)
Demultiplexing (optional)
Platform independent
Sequence alignment (BAM files)
SNV callers
Local realignment
Indel caller
Large structural variation (fusion) detection
CNV callers (gene amplifications and deletions)
List of genomic variants (VCF files)
Terabytes
Gigabytes
Megabytes / Kilobytes

Roy *et al.*, 2016 - *Arch Pathol Lab Med.*

Workflow for Nanopore Sequencing platform



Guppy (Basecalling)

FastQC/NanoPlot (Stat Summary, Quality Control)

Porechop (Trimming)

MiniMap2/BWA-MEM (Mapping)

FreeBayes (Variant calling)

IGV, wANNOVAR (Visualization, interpretation)

What you should have in your PCs?

- Txt editors (notepad, MSWord, notepad++, etc.)
- MSExcel for csv files
- Please register to following websites

https://usegalaxy.org

server.t-bio.info

- Please download following programs

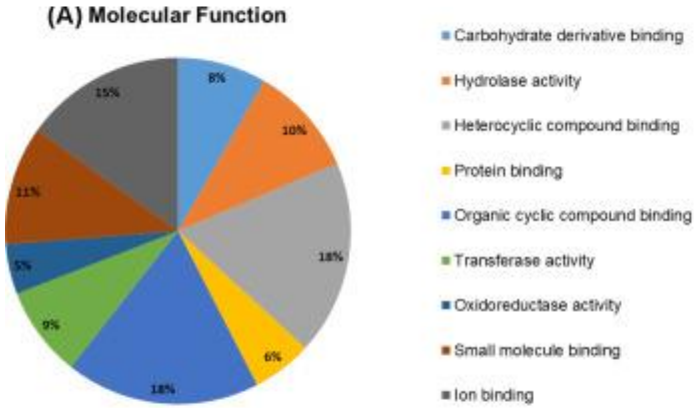
IGV (Integrative Genomics Viewer) from igv.org

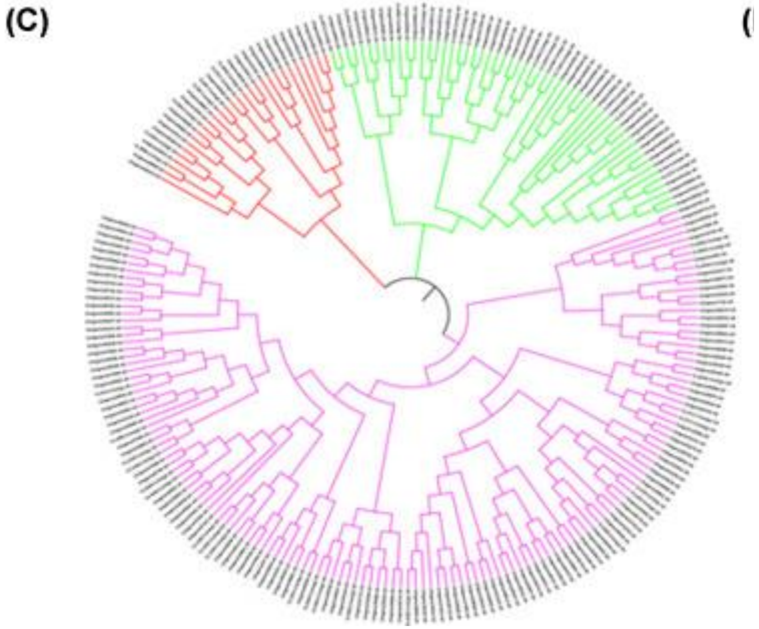
Blast2GO (Annotation tool) from blast2go.com

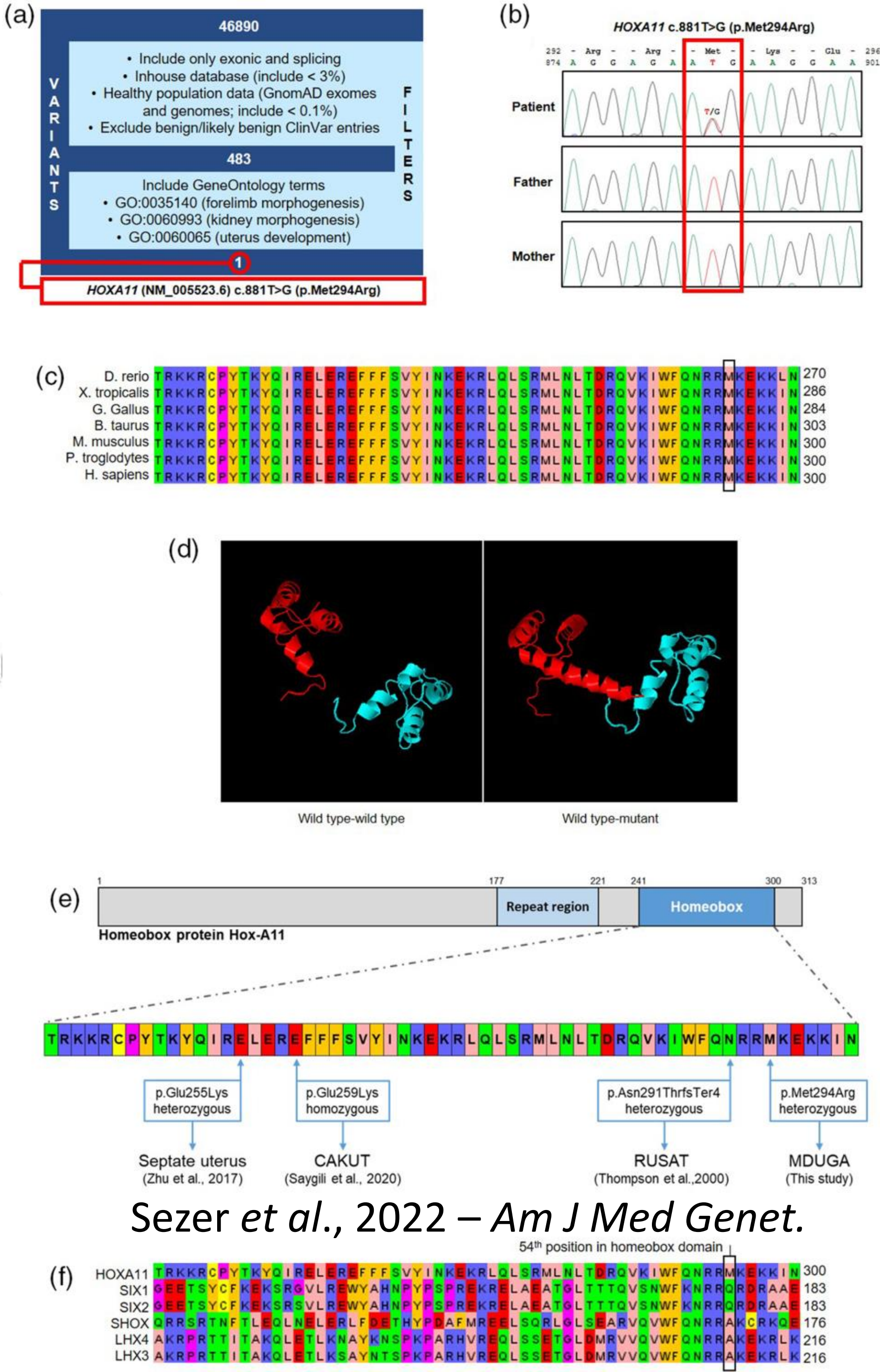
Bioedit (Seq. alignment editor) from bioedit.software.informer.com

NEAR EAST UNIVERSITY

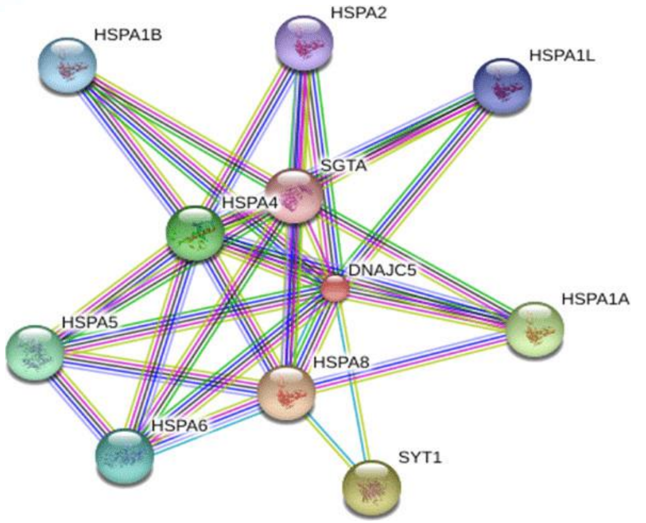
DESAM RESEARCH INSTITUTE



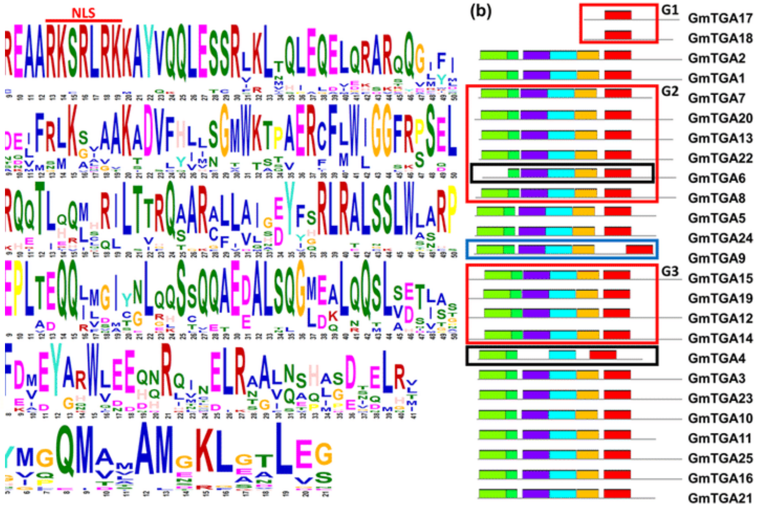



(a)
46890
VARIANTS
FILTERS
• Include only exonic and splicing
• Inhouse database (include < 3%)
• Healthy population data (GnomAD exomes and genomes; include < 0.1%)
• Exclude benign/likely benign ClinVar entries
483
Include GeneOntology terms
• GO:0035140 (forelimb morphogenesis)
• GO:0060993 (kidney morphogenesis)
• GO:0060065 (uterus development)
1
HOXA11 (NM_005523.6) c.881T>G (p.Met294Arg)
(b)
HOXA11 c.881T>G (p.Met294Arg)
Arg Arg Met Lys Glu
Patient
T/G
Father
Mother
(c)
D. rerio 270
X. tropicalis 286
G. Gallus 284
B. taurus 303
M. musculus 300
P. troglodytes 300
H. sapiens 300
TRKKRCPYTKYQIRELEREFFFSVYINKEKRLQLSRMLNLTDRQVKIWFQNRRMKEKKIN
(d)
Wild type-wild type
Wild type-mutant
(e)
1
177
221
241
300
313
Repeat region
Homeobox
Homeobox protein Hox-A11
TRKKRCPYTKYQIRELEREFFFSVYINKEKRLQLSRMLNLTDRQVKIWFQNRRMKEKKIN
p.Glu255Lys heterozygous
p.Glu259Lys homozygous
p.Asn291ThrfsTer4 heterozygous
p.Met294Arg heterozygous
Septate uterus (Zhu et al., 2017)
CAKUT (Saygili et al., 2020)
RUSAT (Thompson et al.,2000)
MDUGA (This study)
(f)
54th position in homeobox domain
HOXA11 300
SIX1 183
SIX2 183
SHOX 176
LHX4 216
LHX3 216

Sezer *et al.*, 2022 – *Am J Med Genet.*



Barbera *et al.*, 2017 - *Proteomics*



llah *et al.*, 2019 – *Sci. Rep.*

NEAR EAST UNIVERSITY
DESAM RESEARCH INSTITUTE

Thanx For Your Attention...