Hôpital général juif
Jewish General Hospital

McGill | Faculty of Medicine

Hôpital général juif
Jewish General Hospital
Institut Lady Davis | Lady Davis Institute

Tips & Considerations to incorporate genomics into your skin research projects

SkIN Canada webinar – June 7 2023

Philippe Lefrançois MD, PhD, FRCPC, DABD

Assistant Professor Dept of Medicine- McGill

Attending Staff – JGH Dermatology
Principal Investigator - LDI

philippe.lefrancois2@mcgill.ca

philippe.lefrancois.med@ssss.gouv.qc.ca

DISCLOSURE OF RELATIONSHIPS WITH INDUSTRY

Philippe Lefrançois, MD, PhD, FRCPC

DISCLOSURES

I do not have any relevant relationships with industry.

Plan

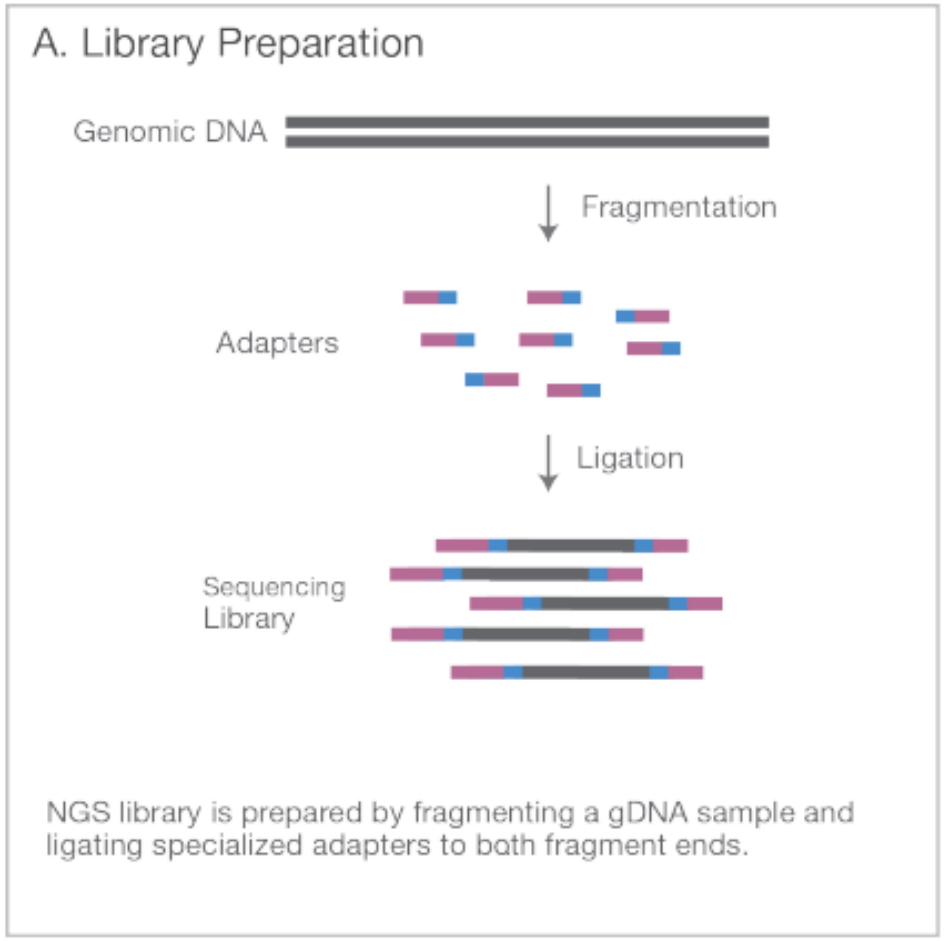
- High-throughput sequencing technologies
- Sequencing needed
- Computational analyses – Primary vs. Secondary vs. Tertiary
- Free resources

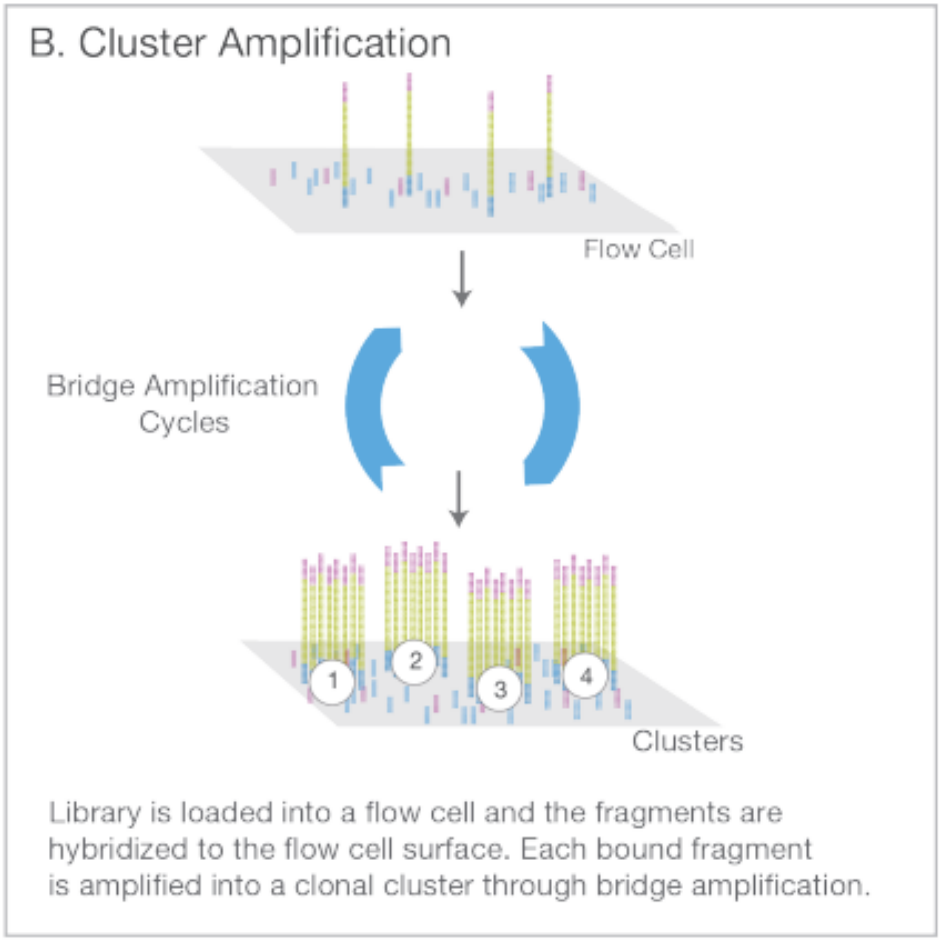
- DISCLAIMER: Goal is to present an overview

Plan

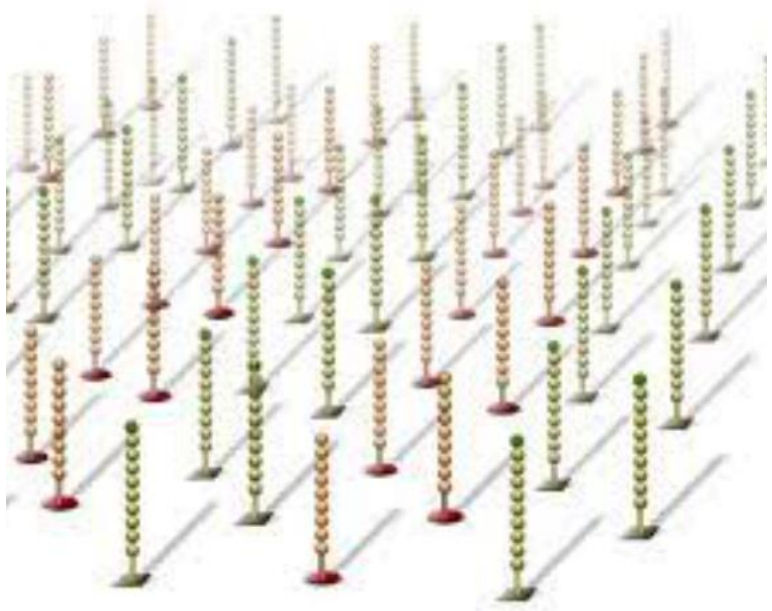
- **High-throughput sequencing technologies**
- Sequencing needed
- Computational analyses – Primary vs. Secondary vs. Tertiary
- Free resources

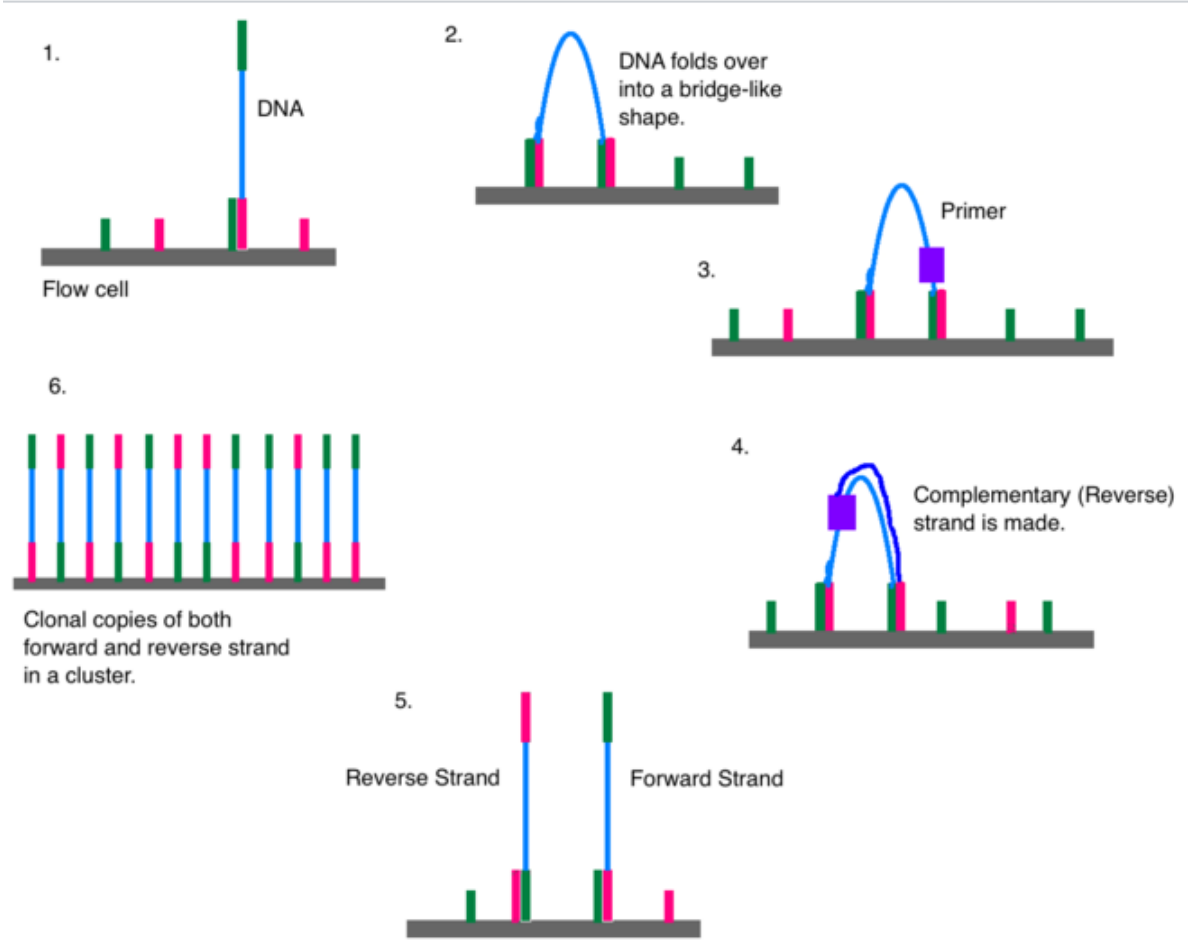
Illumina – Sequencing by synthesis (2nd gen)



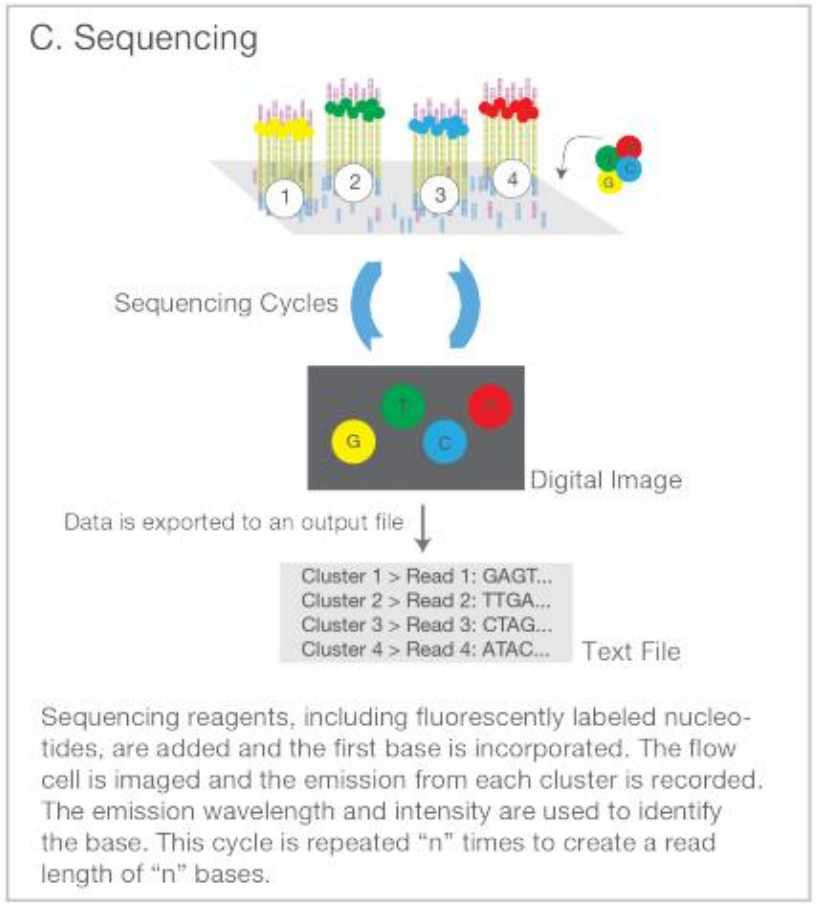


Illumina – Bridge amplification on flowcell

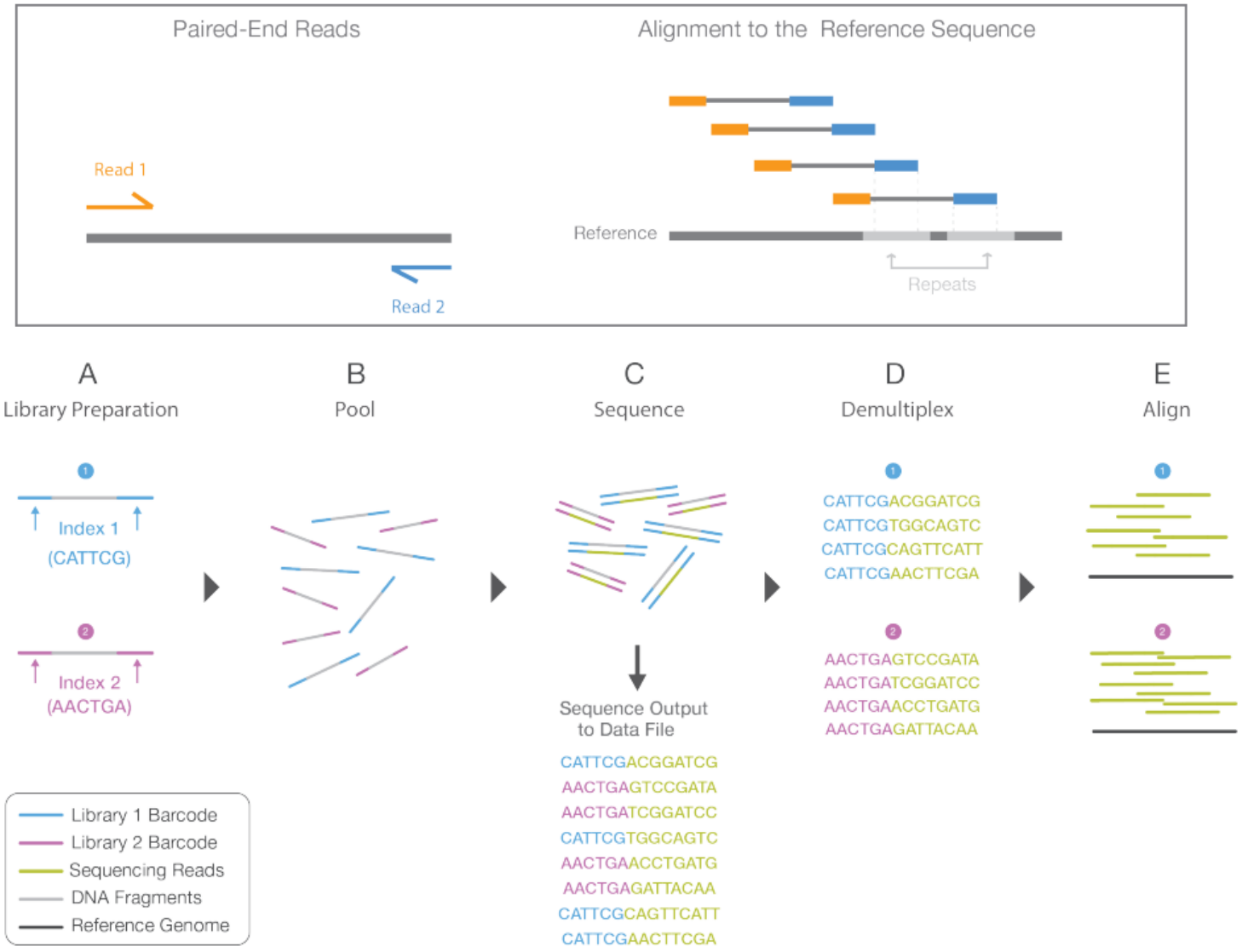




Illumina – Sequencing by synthesis

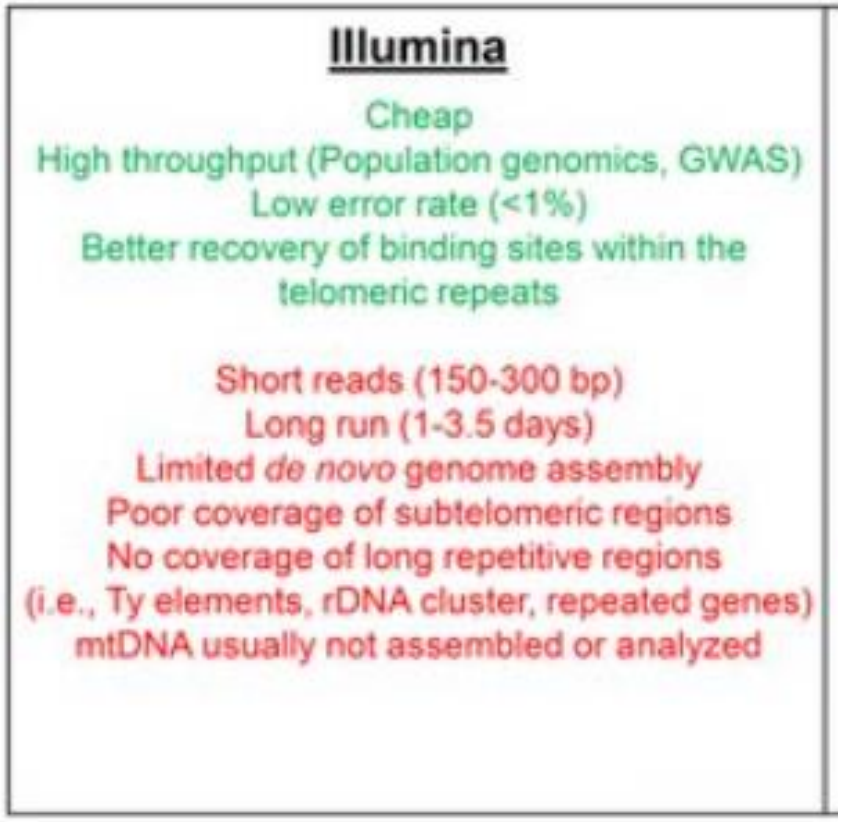


Illumina – Paired end reads + Multiplexing

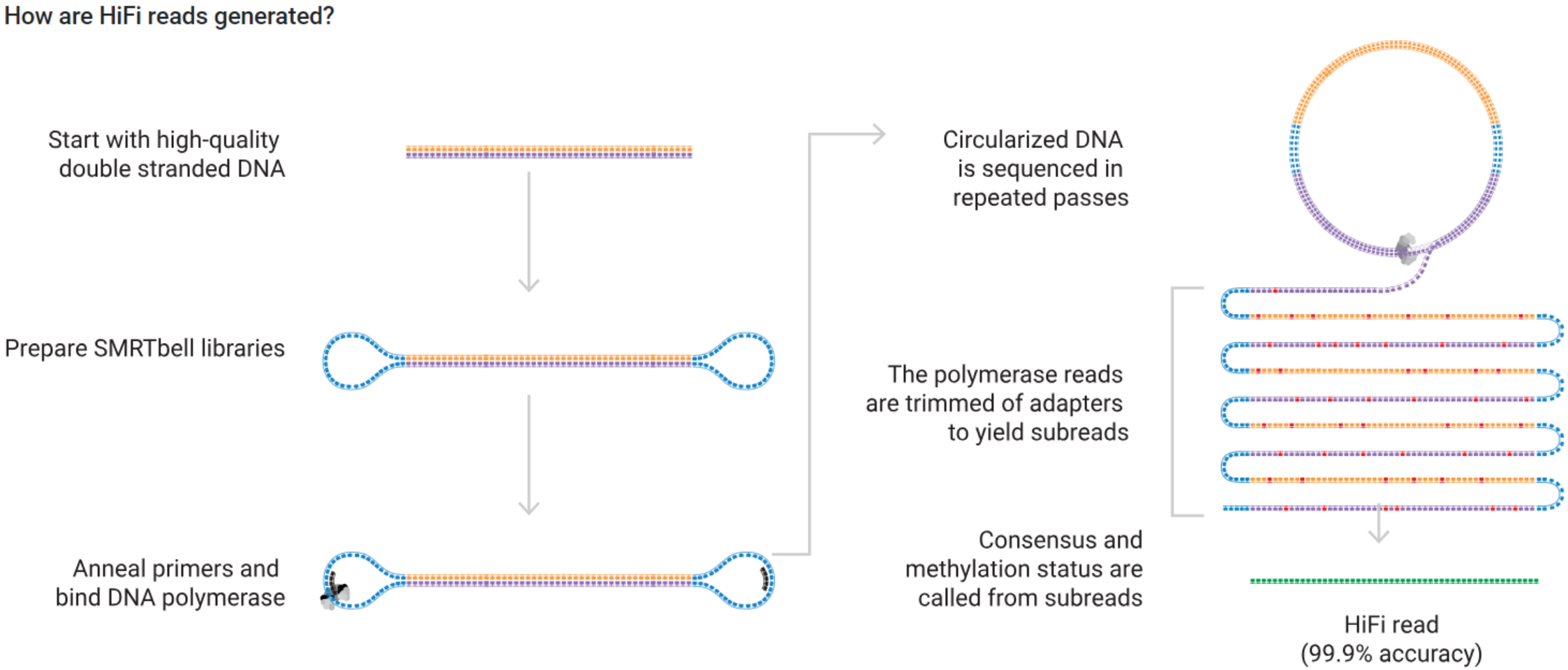


Illumina – Pros & Cons

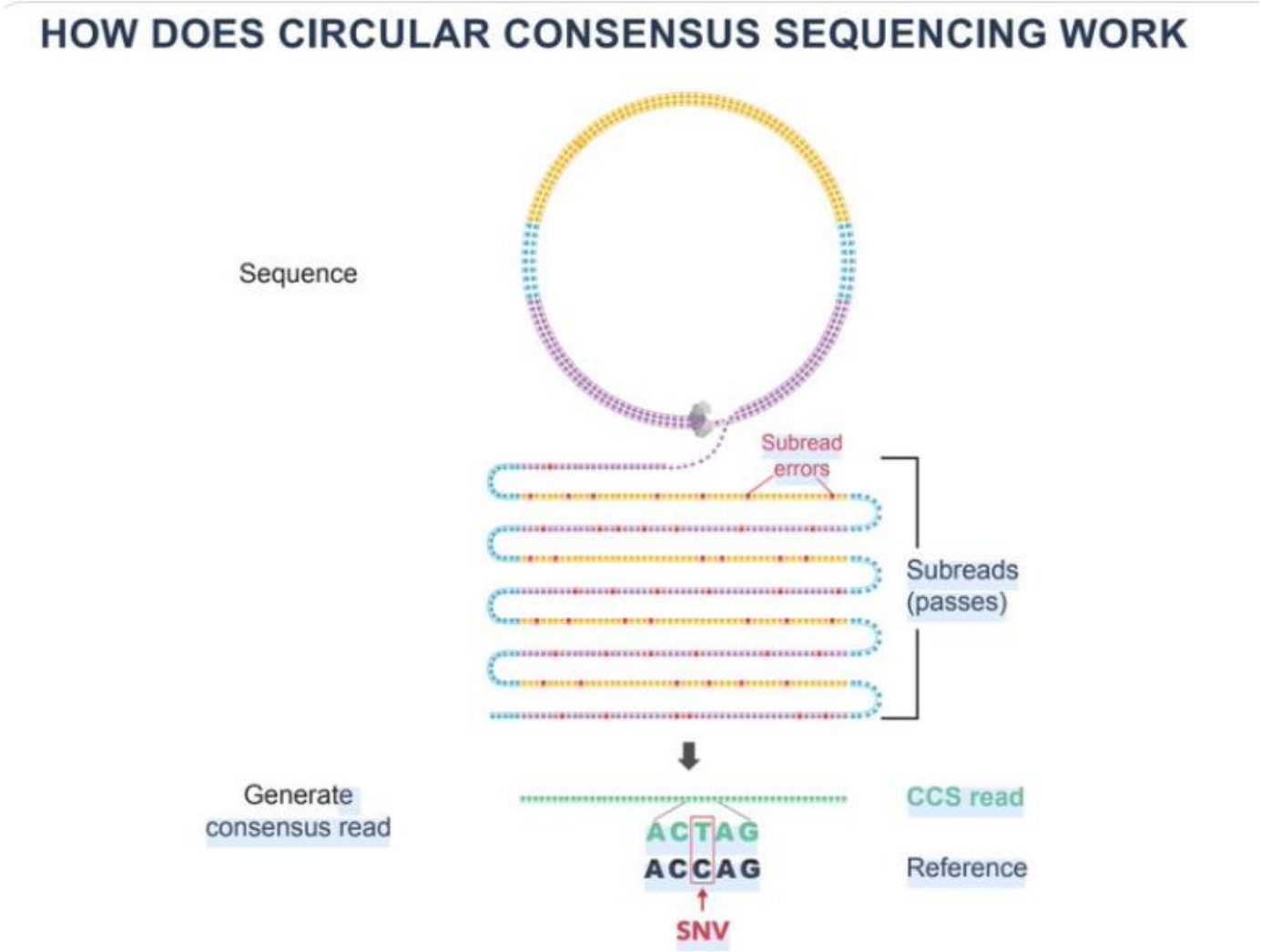
- Reads between 50-150 bp, SE or PE
- Up to 2.5 billion reads per flowcell lanes (NovaSeq), up to 8 lanes
  - 20 B reads
- Quantitative (# reads proportional with #DNA/RNA molecules)
- Issues with GC-rich content (underrepresentation)
- Anything de novo difficult



PacBio – SMRT single-molecule real-time (3rd gen)

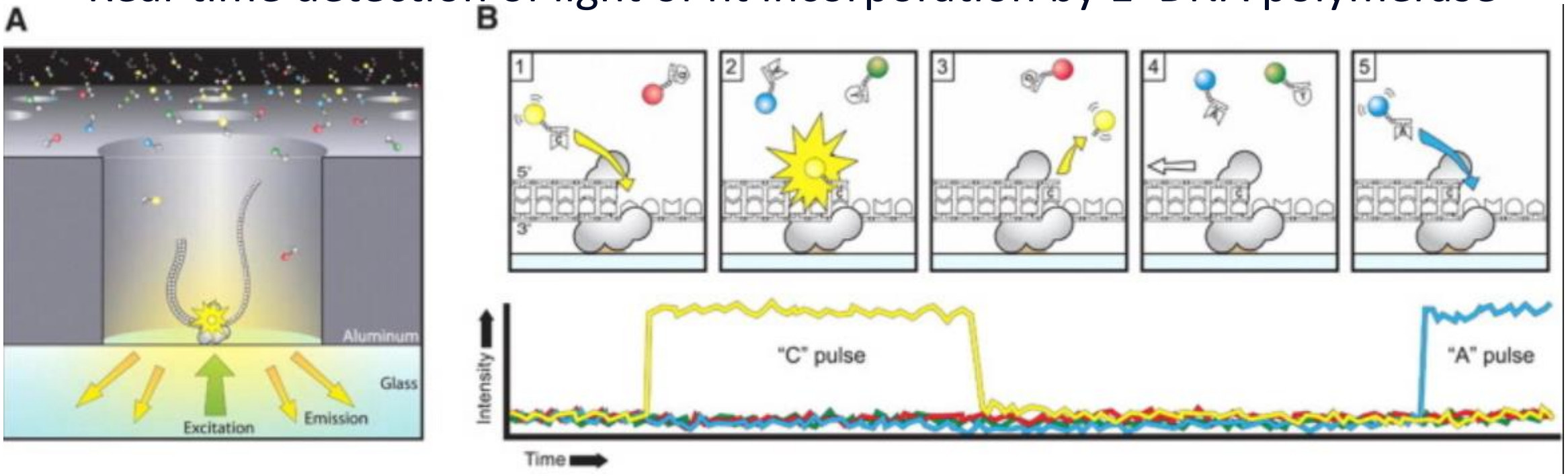


PacBio – Circular consensus reads



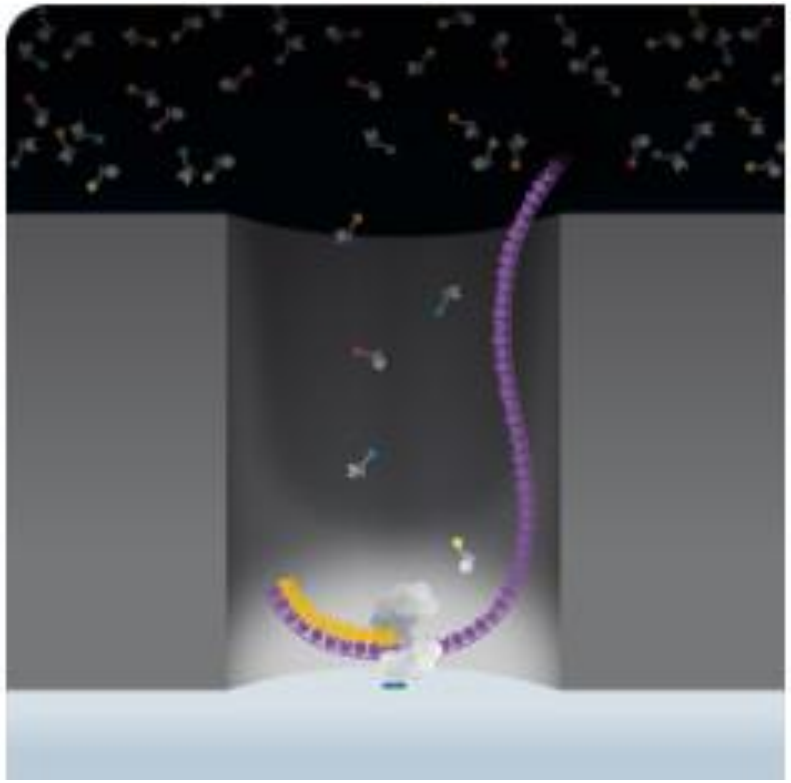
PacBio – 1 molecule/well, DNA poly at bottom

- Real-time detection of light of nt incorporation by 1 DNA polymerase



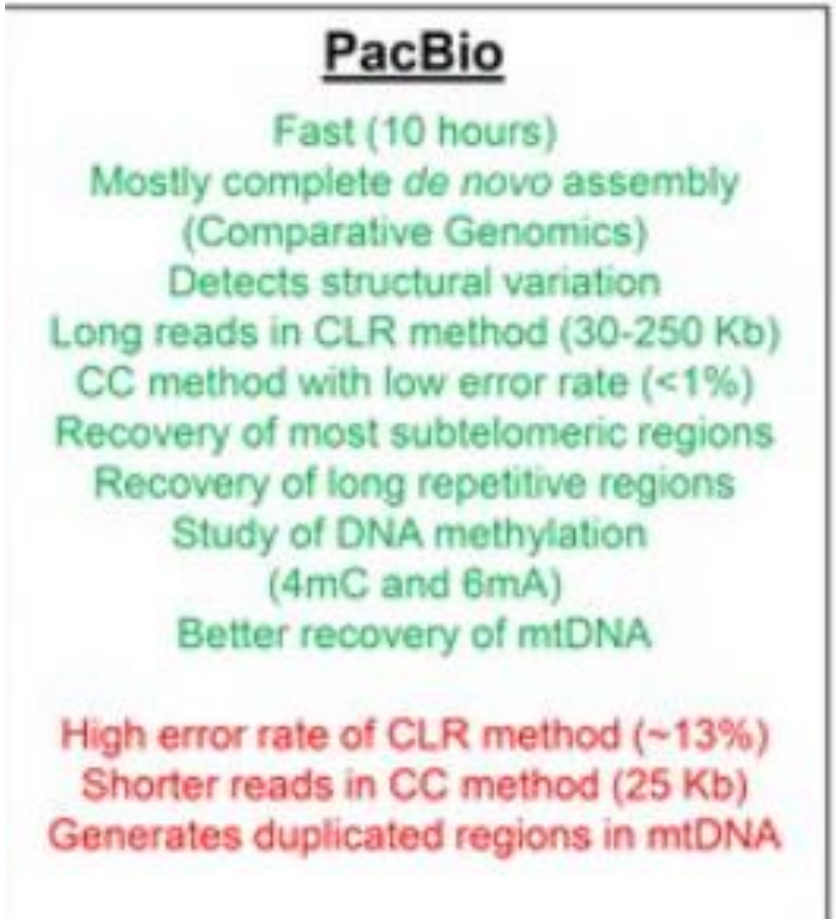
PacBio – Zero-mode waveguide

- A nanophotonic device for confining light to a small observation volume. This can be, for example, a small hole in a conductive layer whose diameter is too small to permit the propagation of light in the wavelength range used for detection.
- Light illumination of only bottom 30 nm, ZMW smaller than wavelength of light



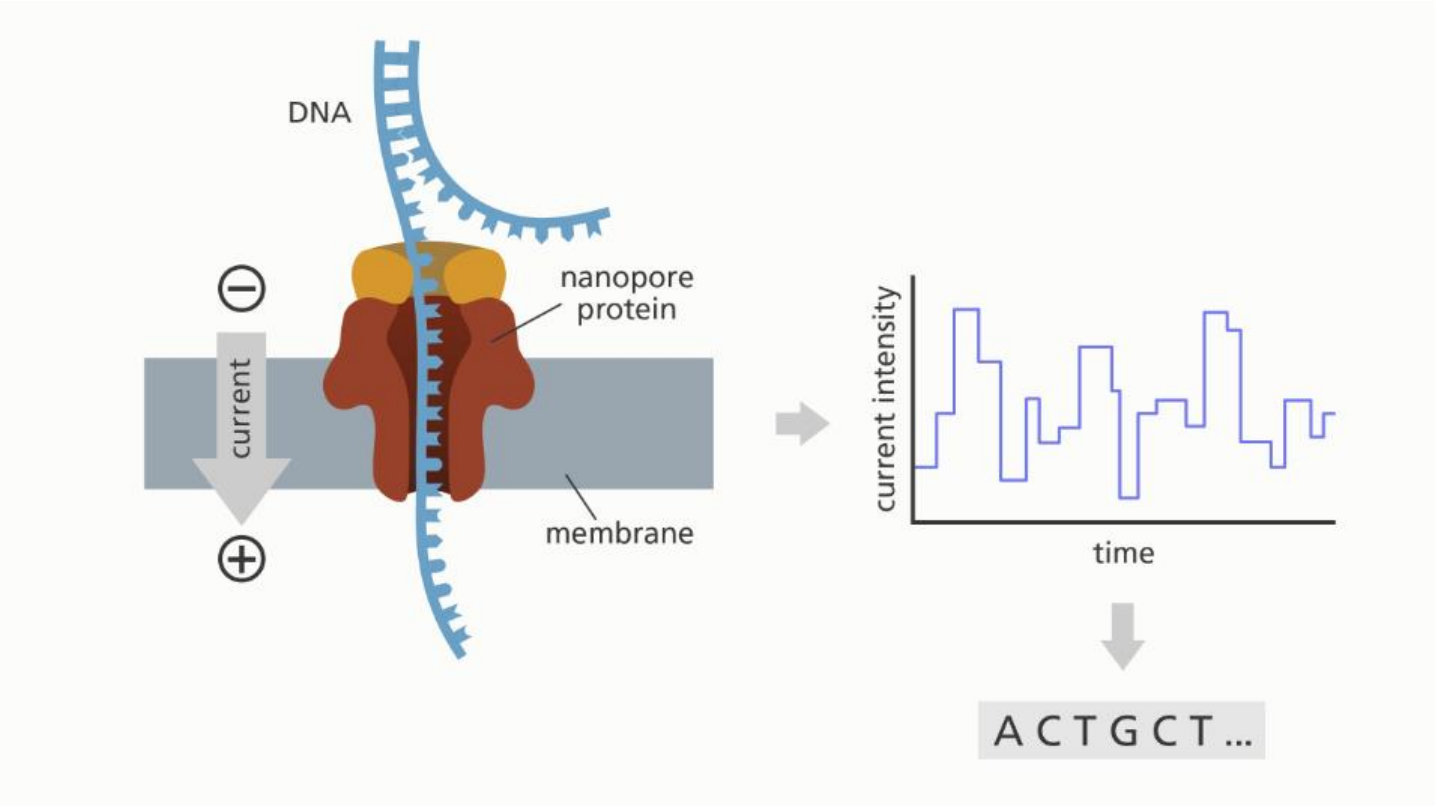
PacBio – Pros & Cons

- Reads up to 50-250 kb
- About 0.5-4 M reads per flowcell
- Real-time analysis and detection
- Single-molecule, no PCR artifact before flowcell loading
- Can look directly at DNA methylation
- Circular consensus mode 99.9%+ accuracy



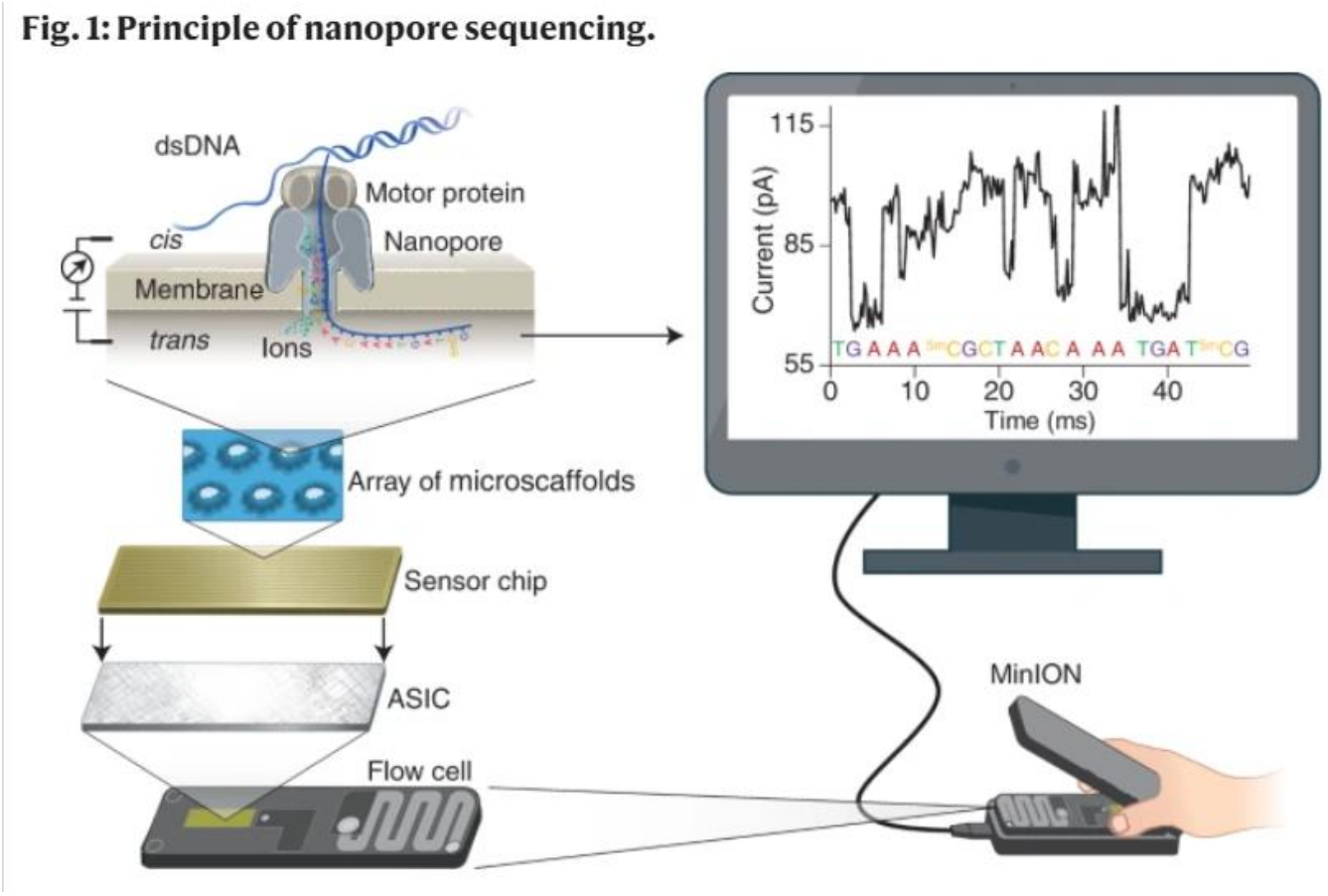
Oxford Nanopore – Nanopore sequencing (3rd gen)

- Pore forming protein alpha-hemolysin
- Inner diameter 1 nm – single strand of DNA/RNA molecule
- Voltage to attract negatively charge DNA/RNA through nanopore -> current variation based on A/C/G/T -> sequence
- Motor protein on top
- No polymerase



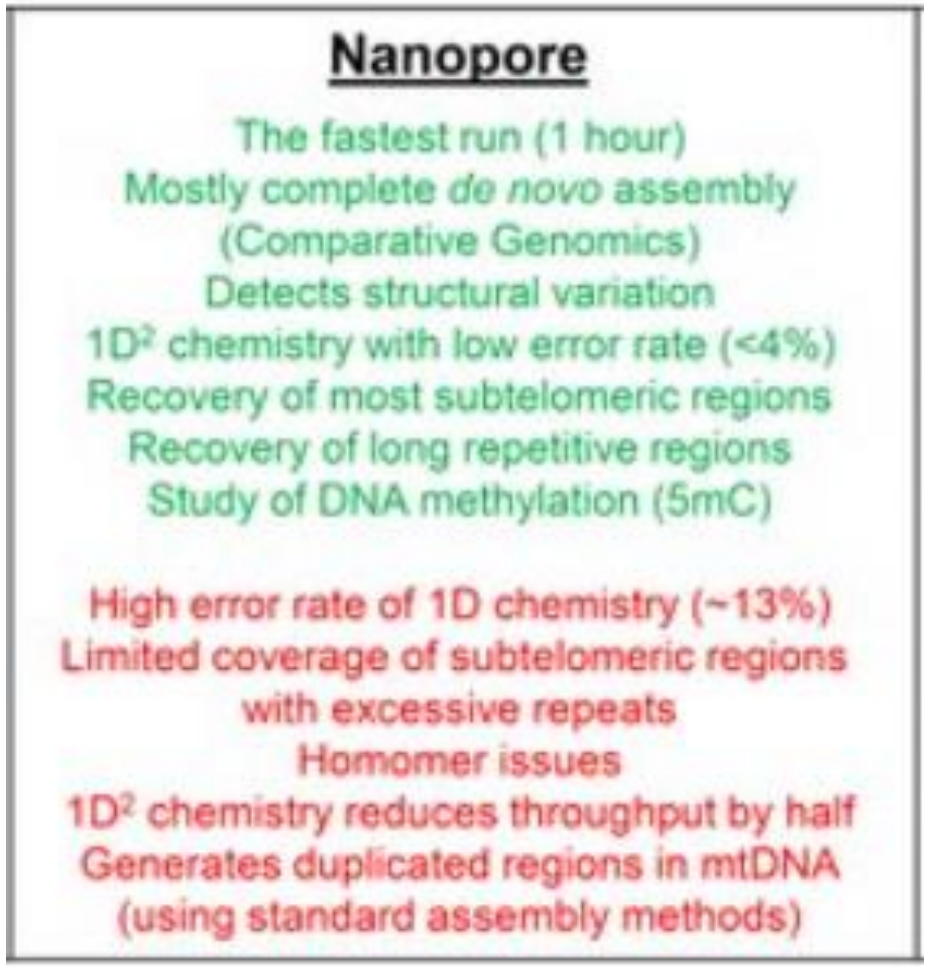
Oxford Nanopore – Scalability

- Size of cell phone to microwave
- Runs on a laptop!
- Flowcell with microscaffolds



Oxford Nanopore – Pros & Cons

- Reads up to 4 Mb
- About 0.2-1 M reads per flowcell
- Real-time analysis and detection
- Single-molecule, no PCR artifact before flowcell loading
- Can look directly at DNA methylation
- Issues initially with error rate -> new chemistries towards Illumina error rate

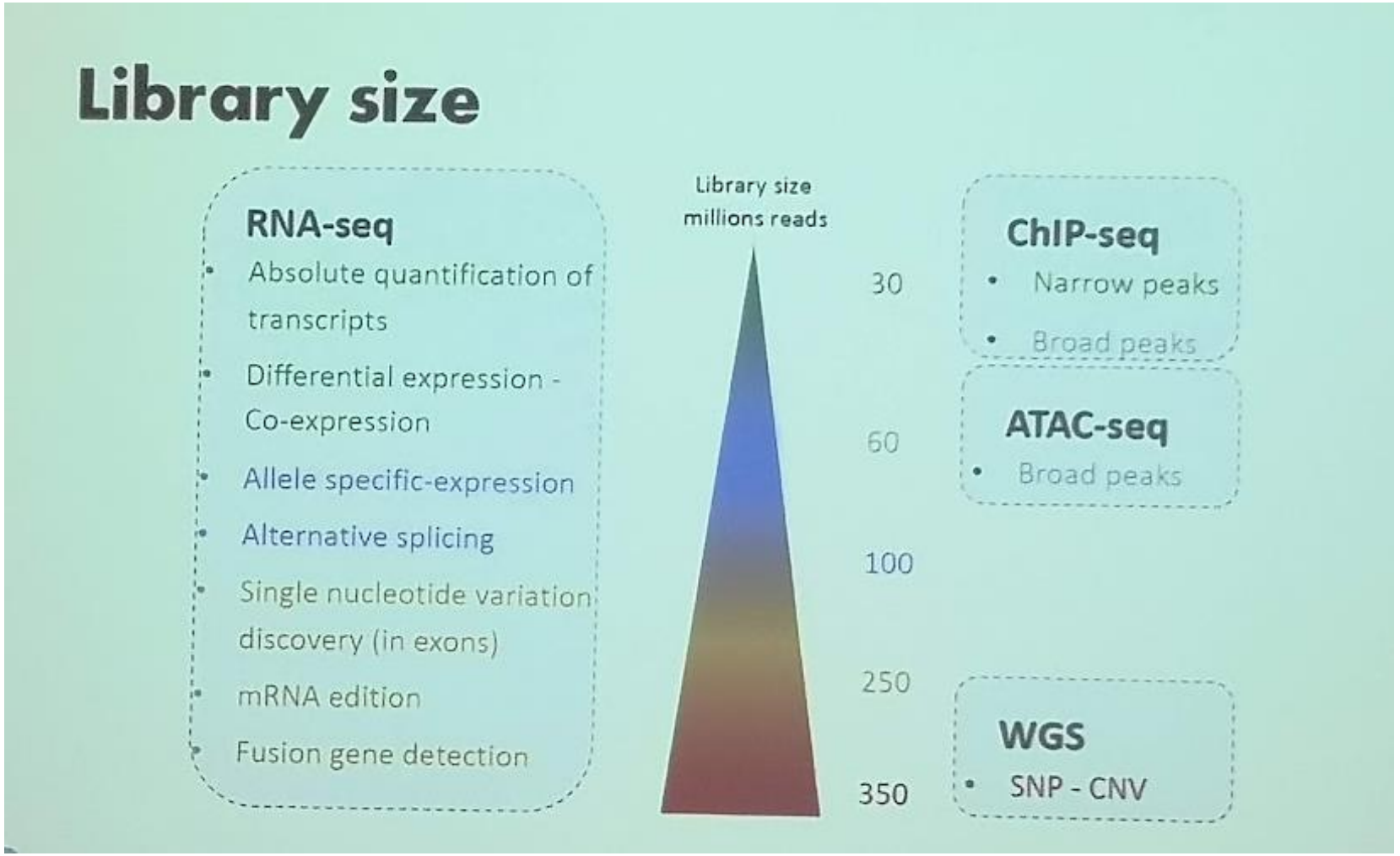


Plan

- High-throughput sequencing technologies
- **Sequencing needed**
- Computational analyses – Primary vs. Secondary vs. Tertiary
- Free resources

Sequencing needed - Illumina

- Slide from IRCM bioinfo

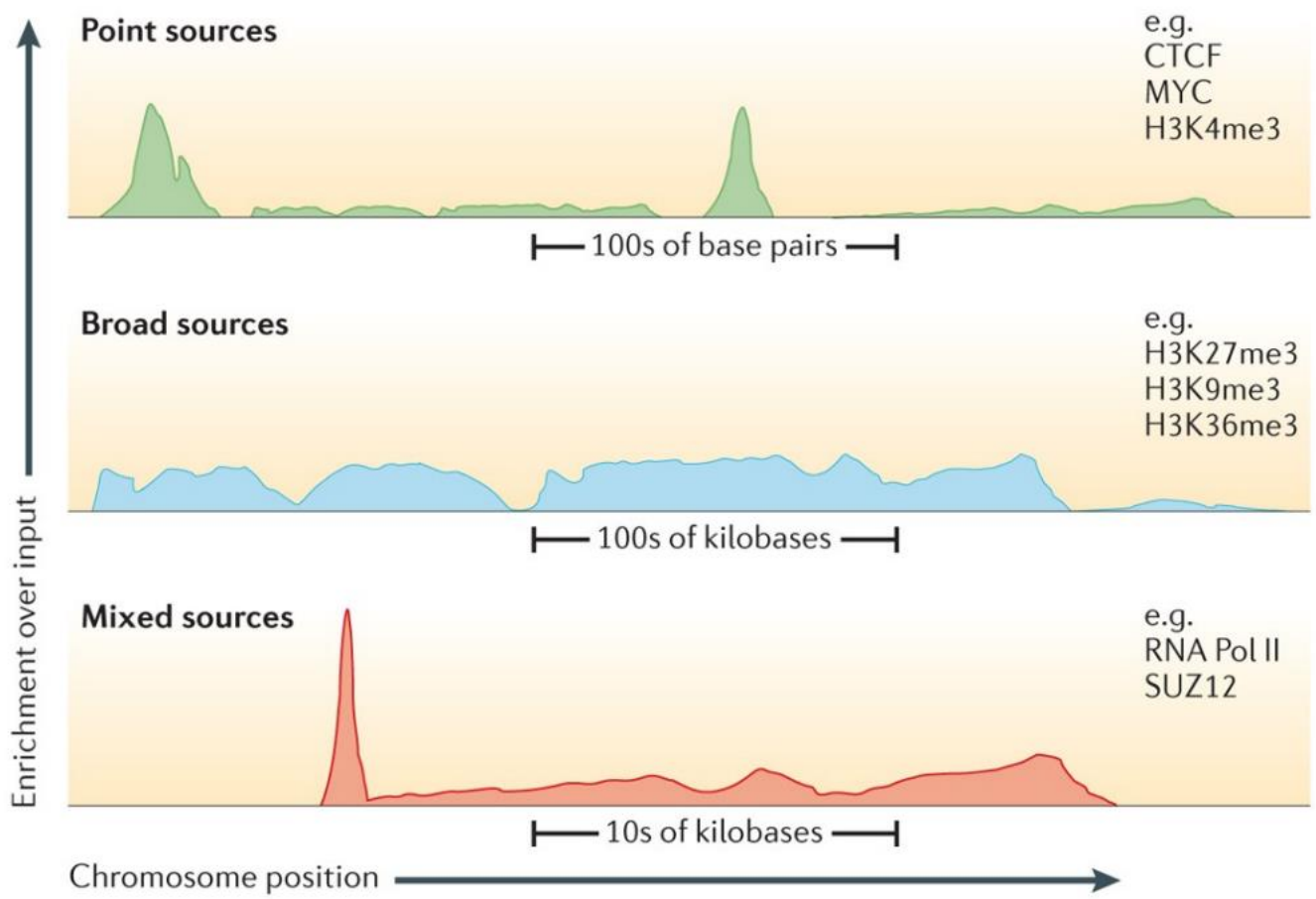


Sequencing needed – Design is key

- RNA Seq
  - 20-30 M reads sufficient for differential gene expression, gene counting
  - 60-80 M reads probably OK for allele-specific expression, alternative transcript
  - 100-200 M reads for RNA editing, SNV calls on RNA, fusion
- Exome-Seq
  - 50-100X coverage for most applications
- Whole Genome
  - 250-350 M reads for decent coverage (10-30X)

Sequencing needed – Design is key

- ChIP-Seq
  - Depends on frequency and size of peaks
  - 30-40 M reads for sharp peaks
  - 60-100 M reads for broad peaks
- ATAC-Seq
  - Similar to ChIP-Seq



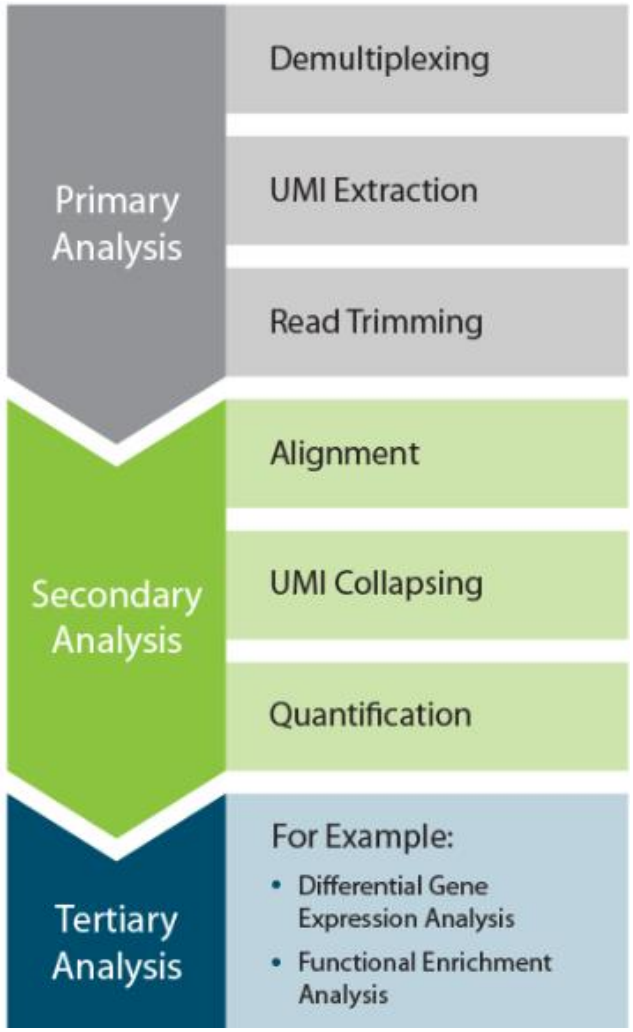
Sequencing needed – Other considerations

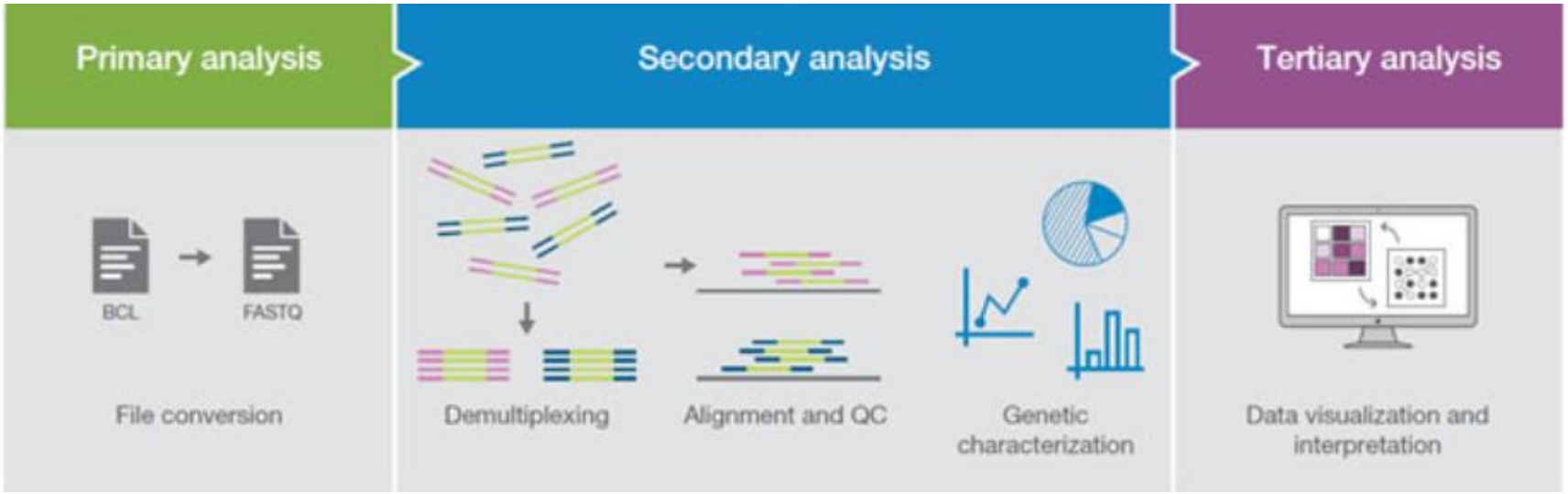
- Targeted (re)sequencing – size of panel + depth of coverage
- Sample heterogeneity
- Mutational burden
- Frequency of particular events (fusion transcripts, alternative splicing)
- Experimental design – paired vs. not
  - Lesional vs. normal
- Hard to sequence areas (subtelomeric, large CNV regions)

Plan

- High-throughput sequencing technologies
- Sequencing needed
- **Computational analyses – Primary vs. Secondary vs. Tertiary**
- Free resources

Computational bio nomenclature





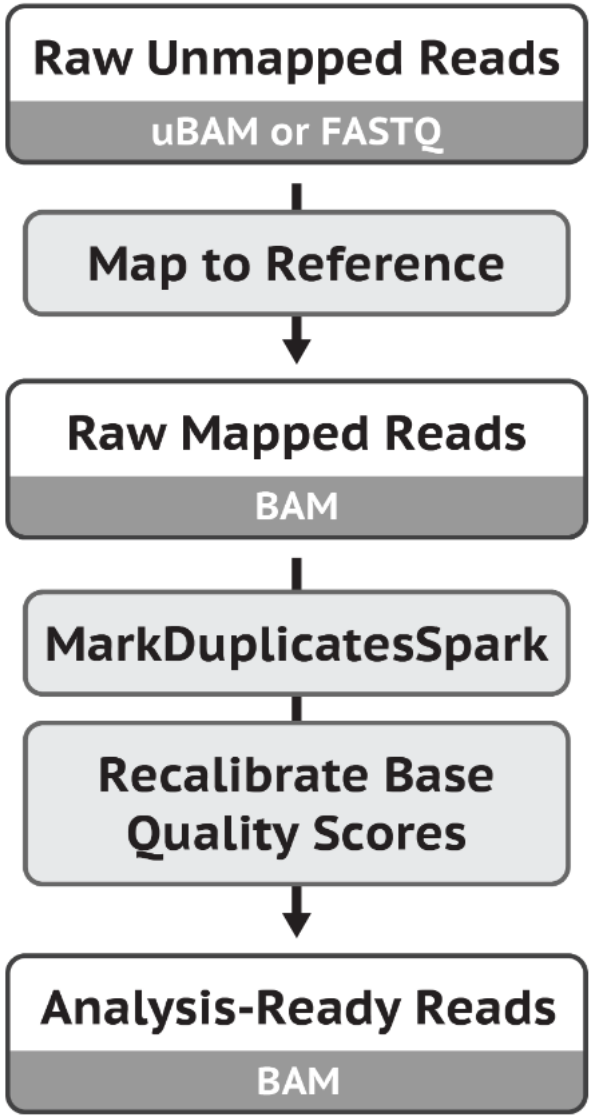
Primary – From image to fastq

- On sequencer

- In practice, people will call secondary -> primary; and tertiary -> secondary…. Because sequencer does true primary

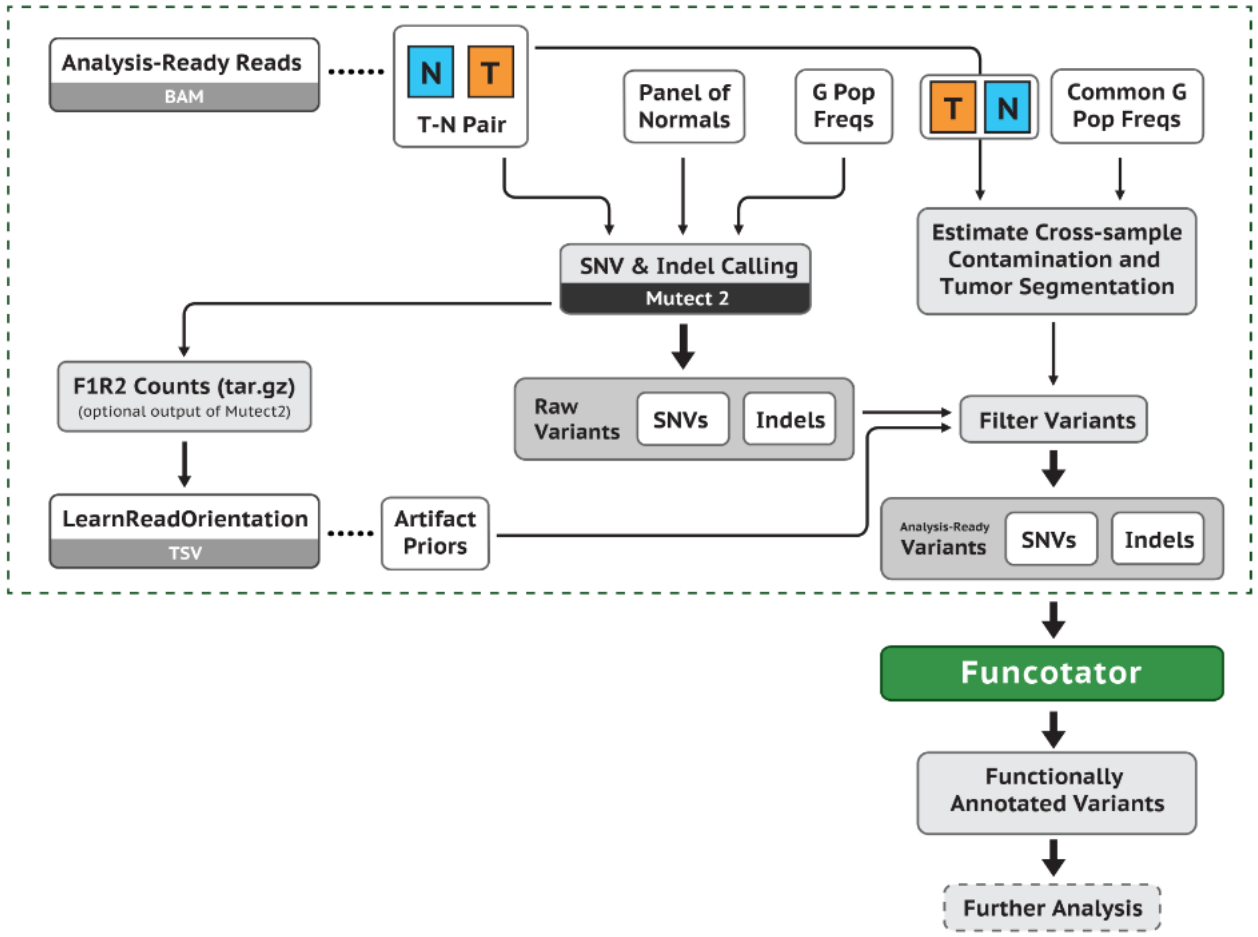
Secondary – From Fastq to counts/variants

- GATK (Genomic Analysis ToolKit)
- Pre-processing



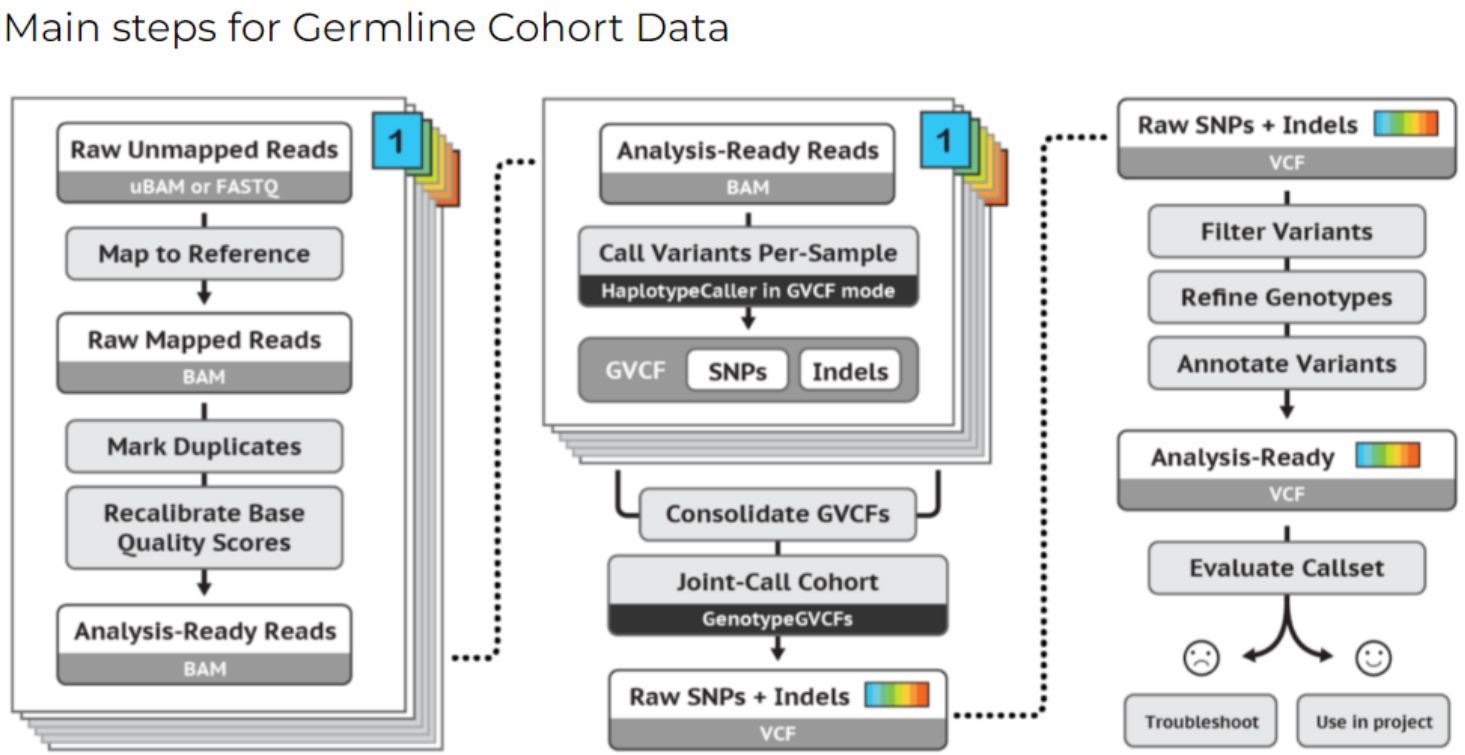
Secondary – From Fastq to counts/variants

- GATK (Genomic Analysis ToolKit)
- Somatic short variants



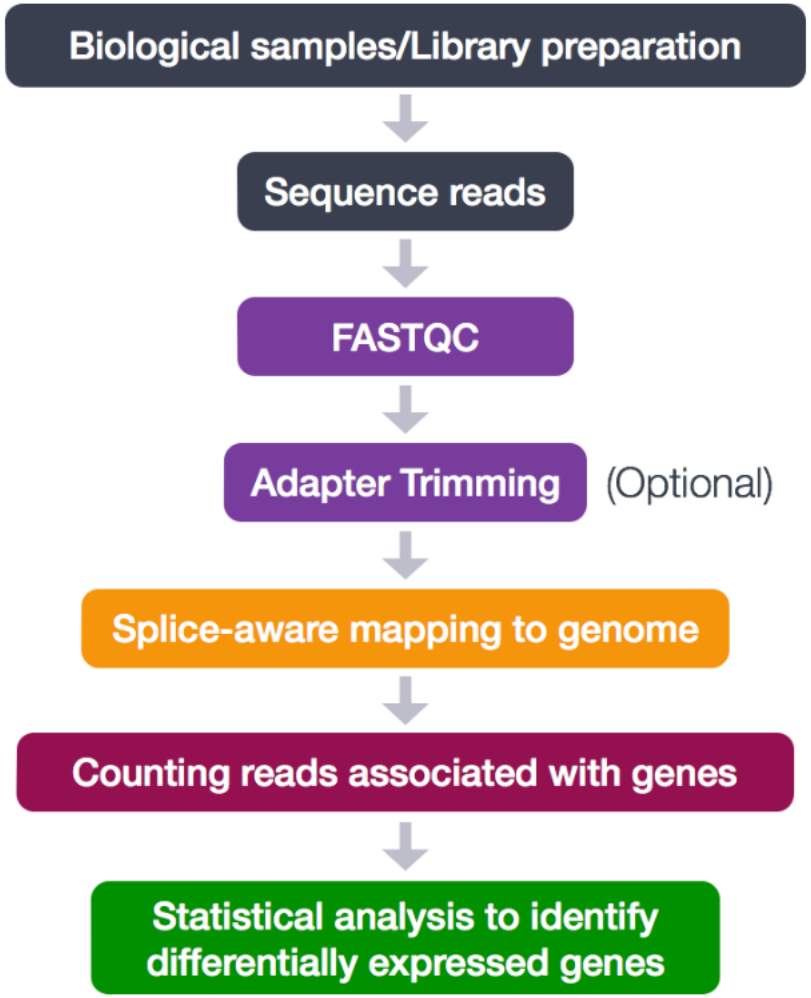
Secondary – From Fastq to counts/variants

- GATK (Genomic Analysis ToolKit)
- Germline short variants



Secondary – From Fastq to counts/variants

- RNA-Seq workflow (basic)

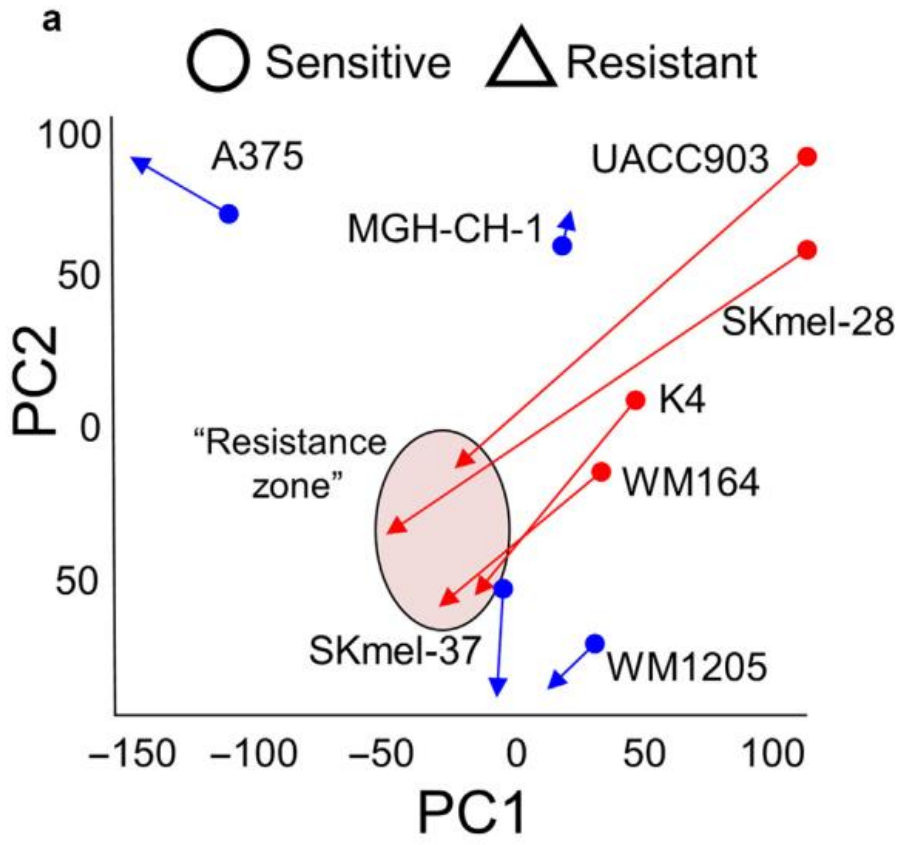


Tertiary - Examples

- RNA – Differentially-expressed genes, pathway analysis
- DNA – Phenotype-variant correlations, waterfall plots
- ChIP-Seq – Motif finding, co-regulations
- ATAC-Seq – Footprinting

Tertiary – Key concepts

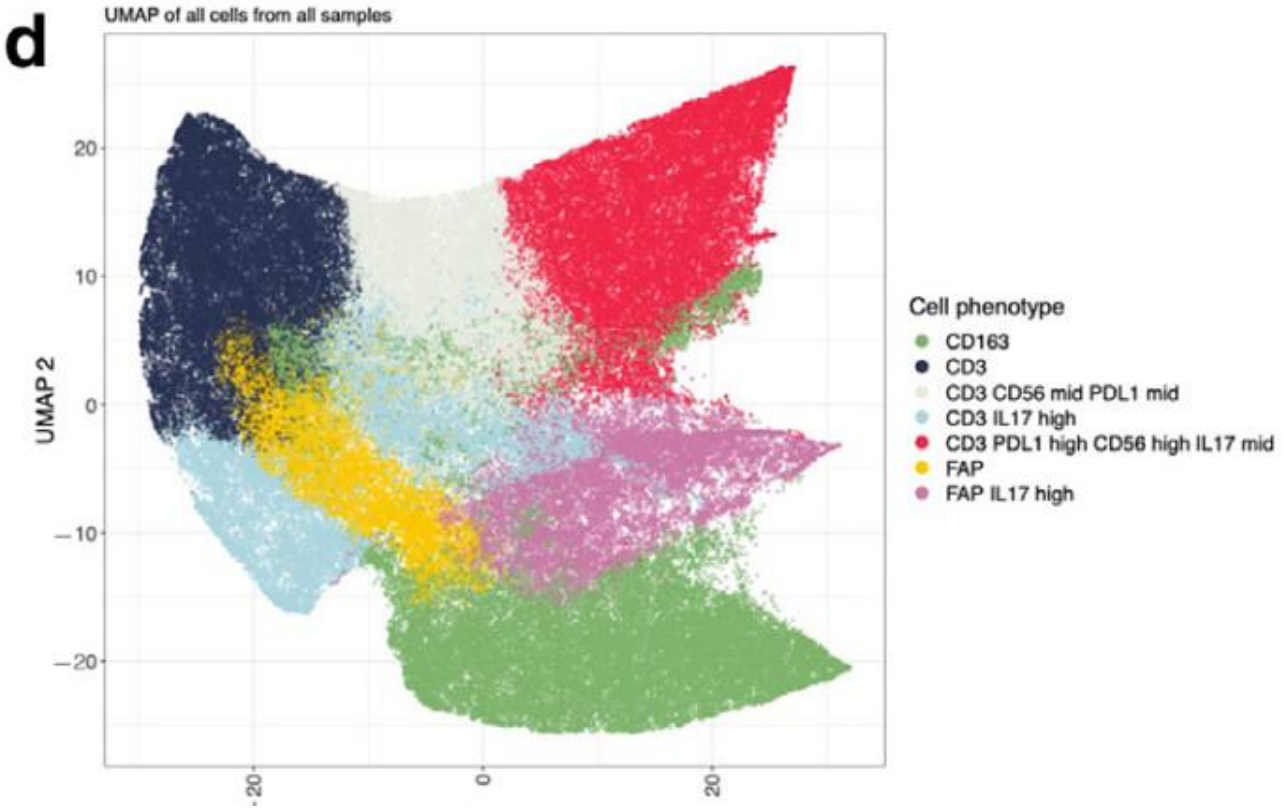
- Dimension reduction
- Multiple hypothesis testing correction
  - Bonferroni (based on alpha; adjusted p-value)
  - Benjamini-Hochberg (based on FDR; Q-value)
- Data visualization

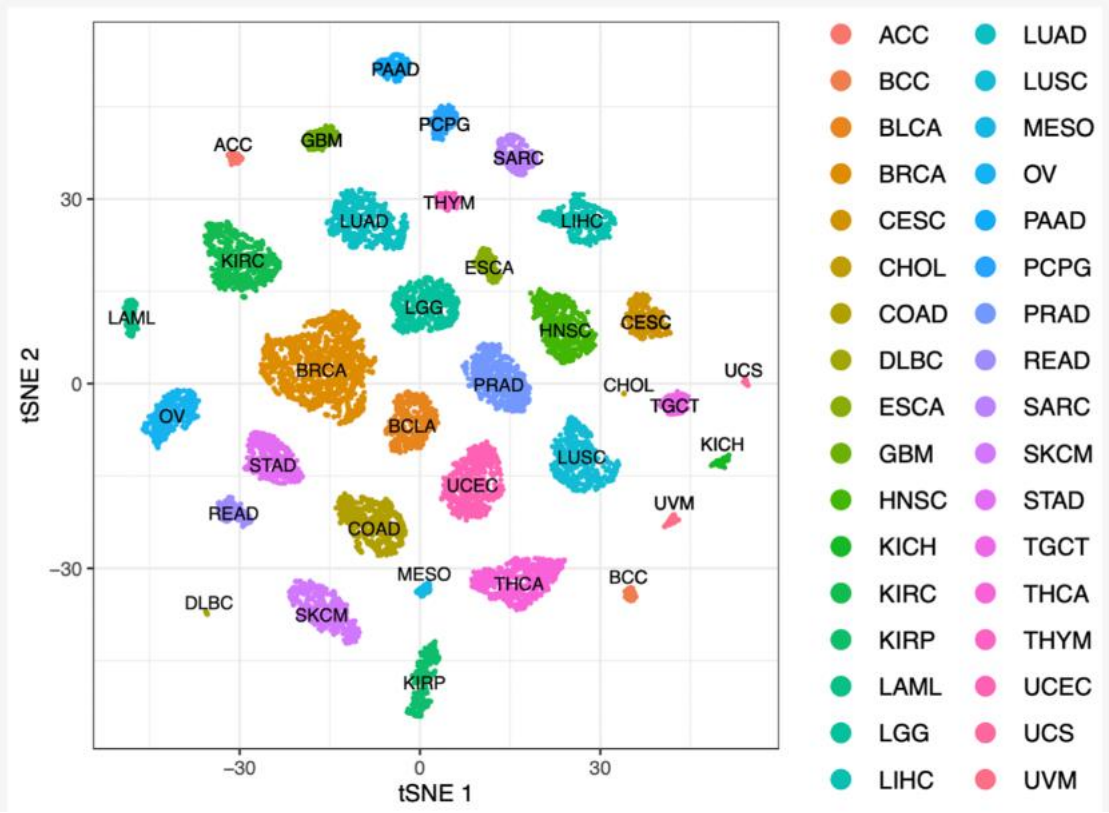


Tertiary – Very high dimension representations

- t-SNE (t-stochastic neighbor embedding)
- UMAP (uniform manifold approximation and projection)

- Can show individual cell (single cell) or each sample (1000s)





Plan

- High-throughput sequencing technologies
- Sequencing needed
- Computational analyses – Primary vs. Secondary vs. Tertiary
- **Free resources**

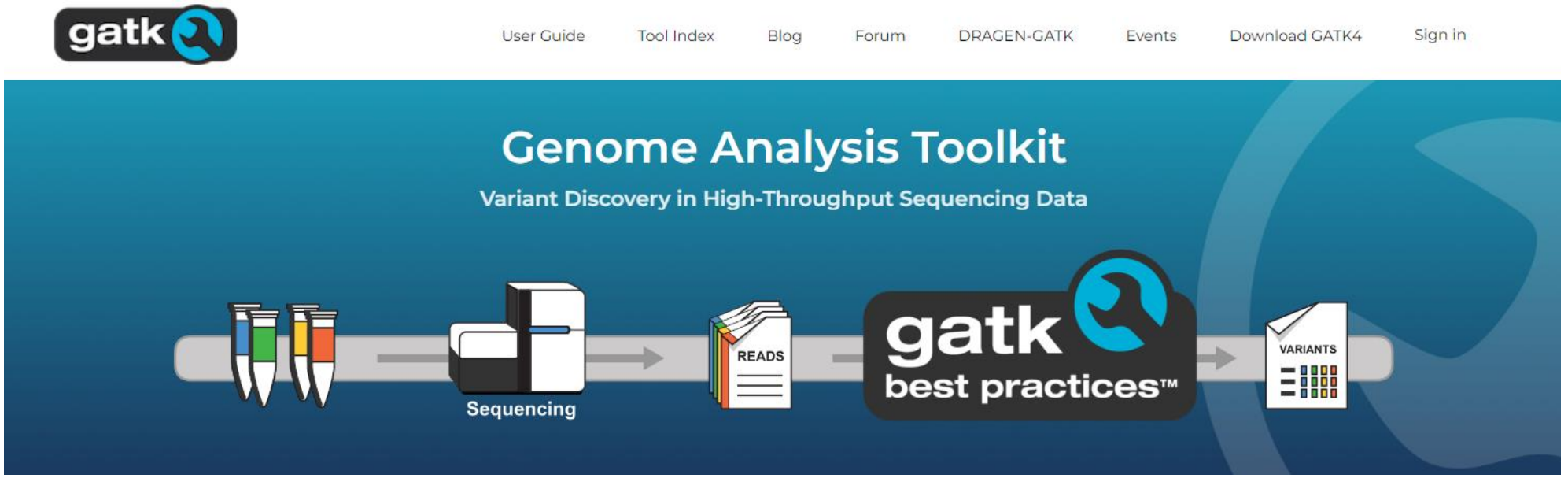
The “Alliance” (DRAC) – formerly ComputeCanada

Digital Research Alliance of Canada | Alliance de recherche numérique du Canada

- High-performance computing clusters, pan-Canadian
- alliancecan.ca
- Free basic account for PI with university affiliation
- Can apply for more (Research Allocation Competitions)
- Excellent wiki: https://docs.alliancecan.ca/wiki/Main_Page
- UNIX-based, Bash programming

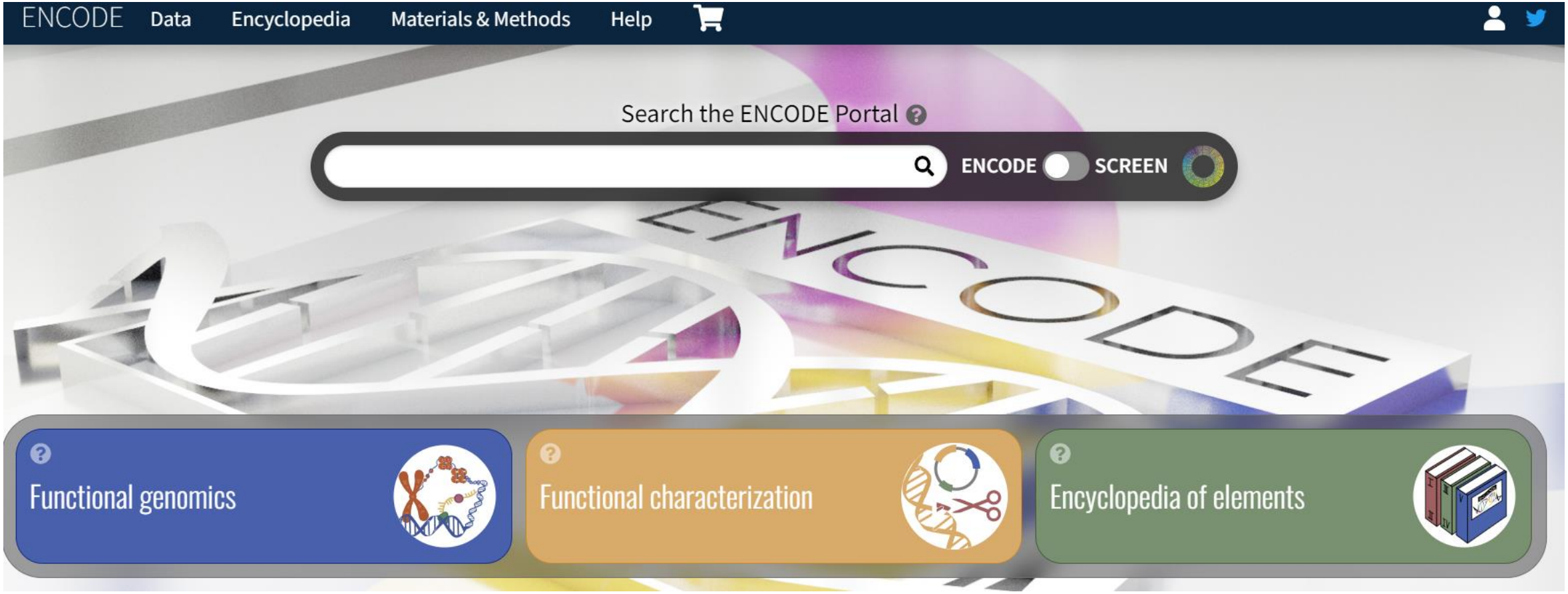
GATK (Genomic Analysis ToolKit)

- Standardized workflows for variant discovery



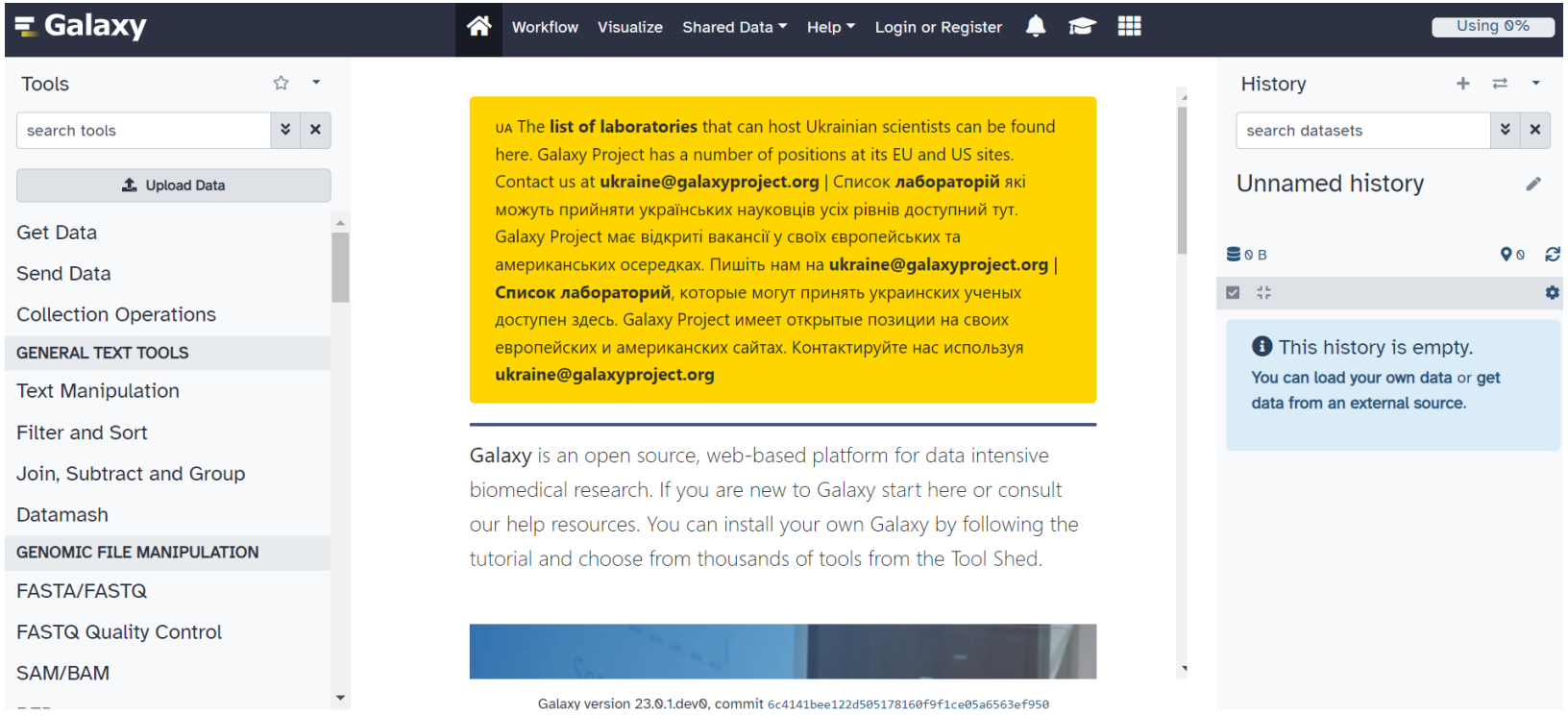
ENCODE

- For gene regulation, but highly integrated with standardized analyses



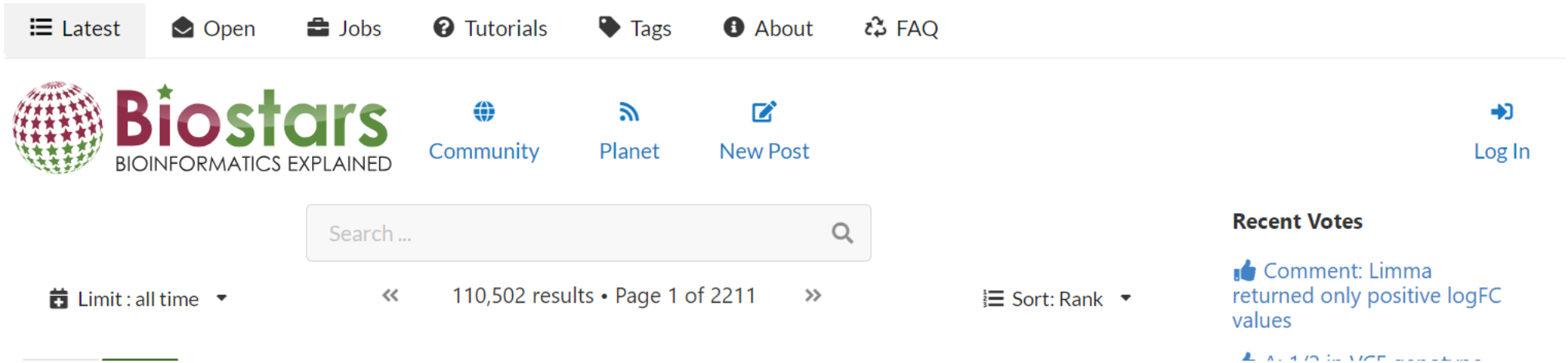
UseGalaxy

- Free, convivial, good for beginners without programming knowledge
- GUI makes it easy to start standard analyses
- Cloud-based -> https://usegalaxy.org/



Biostars

- Forum with Q&As and tutorials for computational biology



R & Bioconductor

- For tertiary analyses mostly, also some secondary analyses
- Statistical and programming language, object-oriented so easier to learn

CRAN
Mirrors
What's new?
Search
CRAN Team

About R
R Homepage

The Comprehensive R Archive Network

Download and Install R

Precompiled binary distributions of the base system and contributed packages, **Windows and Mac** users most likely want one of these versions of R:

Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

Search:

Home | Install | Help | Developers | About

About *Bioconductor*

The mission of the *Bioconductor* project is to develop, support, and disseminate free open source software that facilitates rigorous and reproducible analysis of data from current and emerging biological

Bioc2023 Conference»

This is a hybrid in-person (Boston, MA, USA) and virtual conference from August 2-4, 2023.

Register for Bioc2023

Abstract Submission is now closed.

Important Notice!»

On March 8th, the Bioconductor Core Team will rename the default branch on `git.bioconductor.org` to `devel`.

This changes affects maintainers of packages.

Challenges

- Many programming languages (Bash, UNIX/Linux, R, Python, Perl, java)
- Dealing with terabytes of data -> summary files can be several Gbs
- Vignettes, papers, repositories not always clear, rarely updated
- Advanced statistics (non-parametric, MCMC, HMM, multivariate, resampling, sliding windows)
- Power issue (n is small, p is extremely large)
- Rapidly evolving
- Storage, transfer and data security

Questions?

Interested by skin cancer research?

Please email me: philippe.lefrancois2@mcgill.ca