

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.

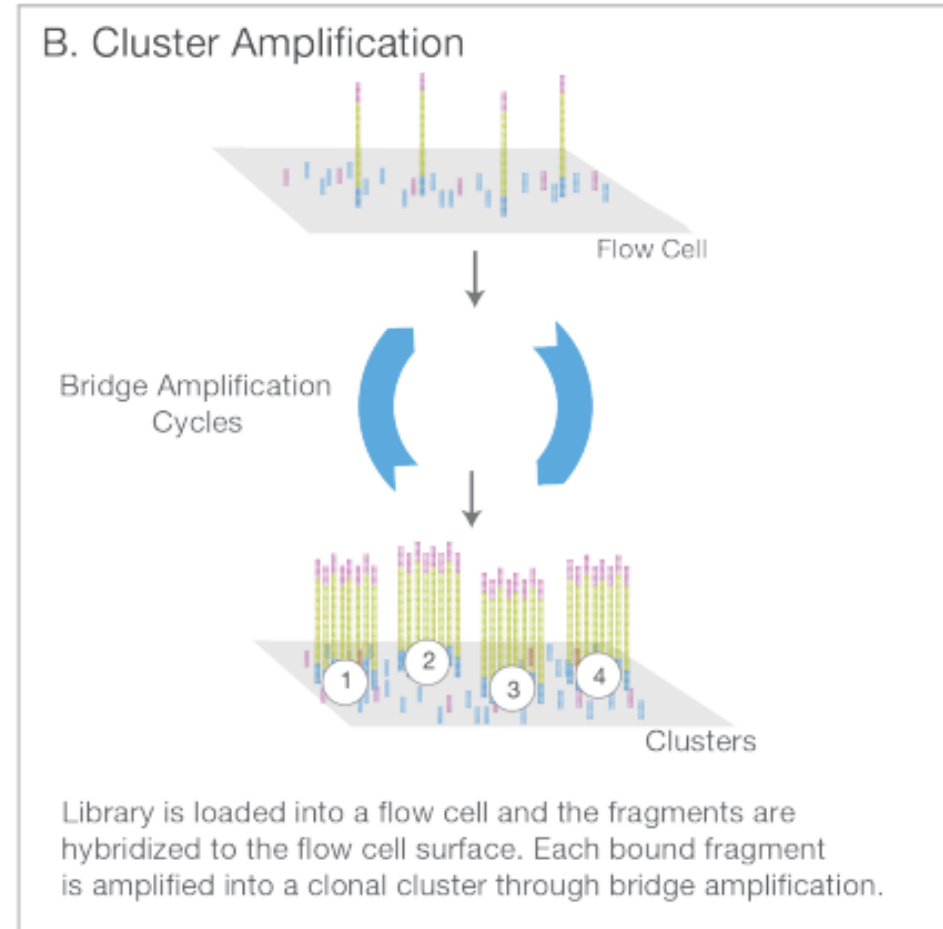
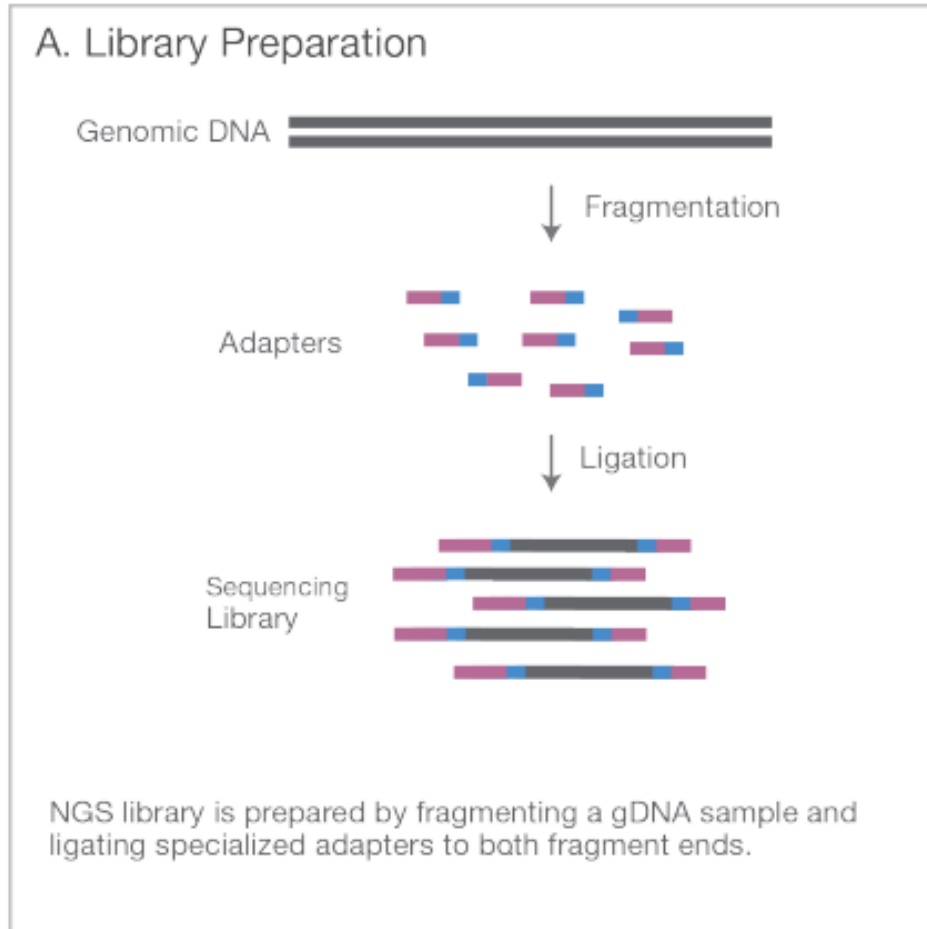
Plan

- High-throughput sequencing technologies
 - Sequencing needed
 - Computational analyses – Primary vs. Secondary vs. Tertiary
 - Free resources
-
- **DISCLAIMER:** Goal is to present an overview

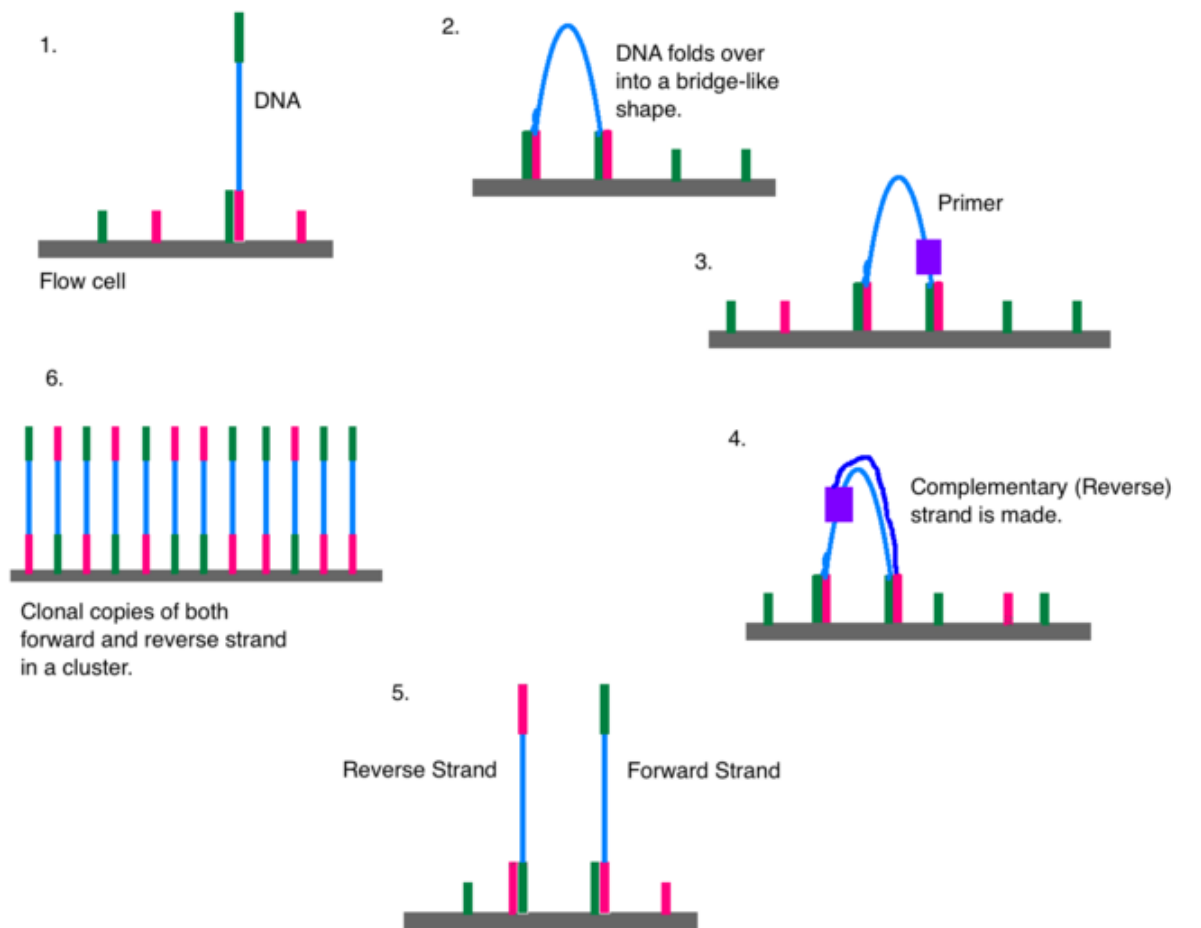
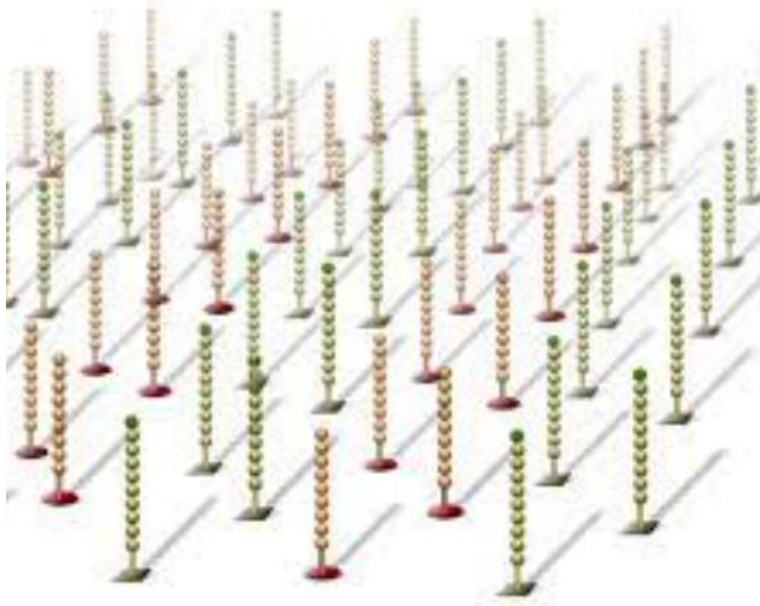
Plan

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

Illumina – Sequencing by synthesis (2nd gen)

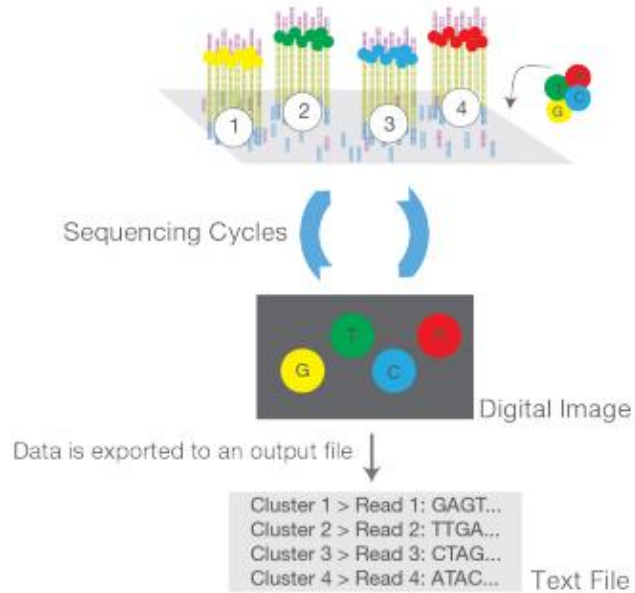


Illumina – Bridge amplification on flowcell



Illumina – Sequencing by synthesis

C. Sequencing



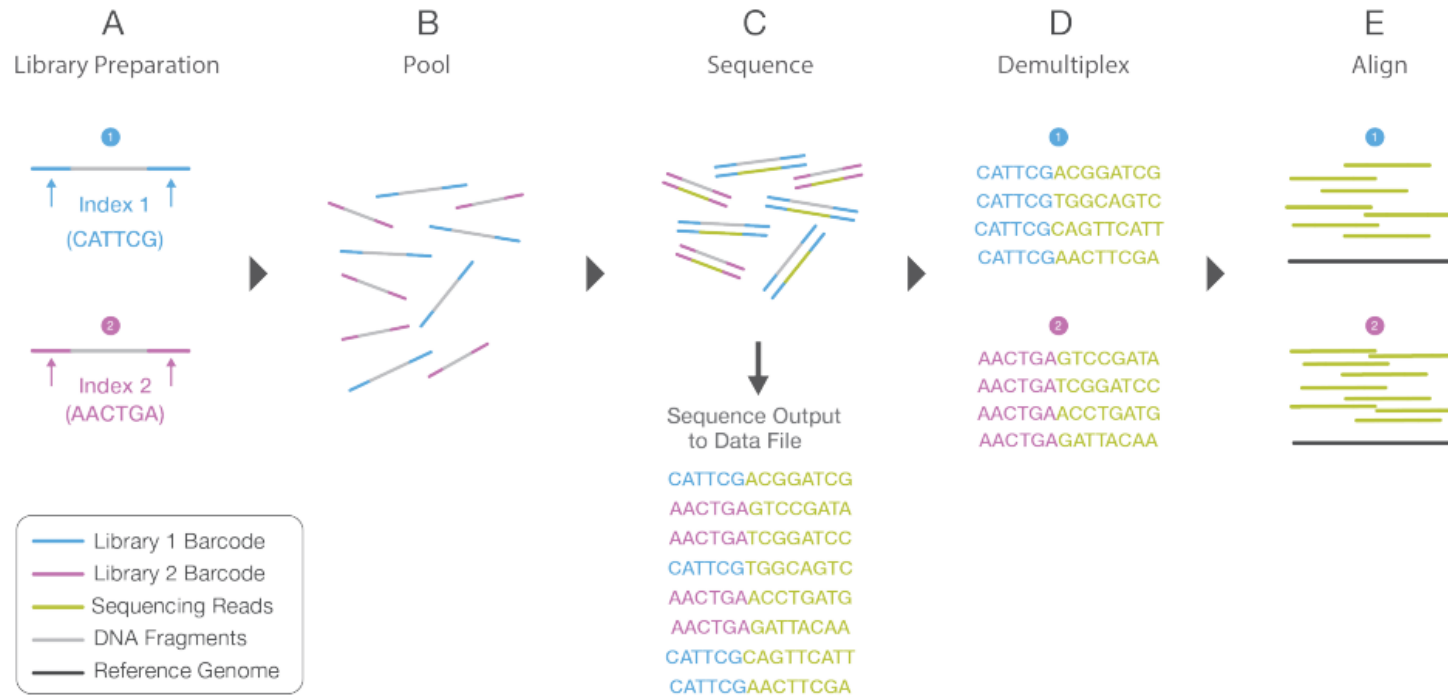
Sequencing reagents, including fluorescently labeled nucleotides, are added and the first base is incorporated. The flow cell is imaged and the emission from each cluster is recorded. The emission wavelength and intensity are used to identify the base. This cycle is repeated "n" times to create a read length of "n" bases.

D. Alignment and Data Analysis

Reads	ATGGCATTGCAATTTGACAT
	TGGCATTGCAATTTG
	AGATGGTATTG
	GATGGCATTGCAA
	GCATTGCAATTTGAC
	ATGGCATTGCAATT
	AGATGGCATTGCAATTTG
Reference Genome	AGATGGTATTGCAATTTGACAT

Reads are aligned to a reference sequence with bioinformatics software. After alignment, differences between the reference genome and the newly sequenced reads can be identified.

Illumina – Paired end reads + Multiplexing



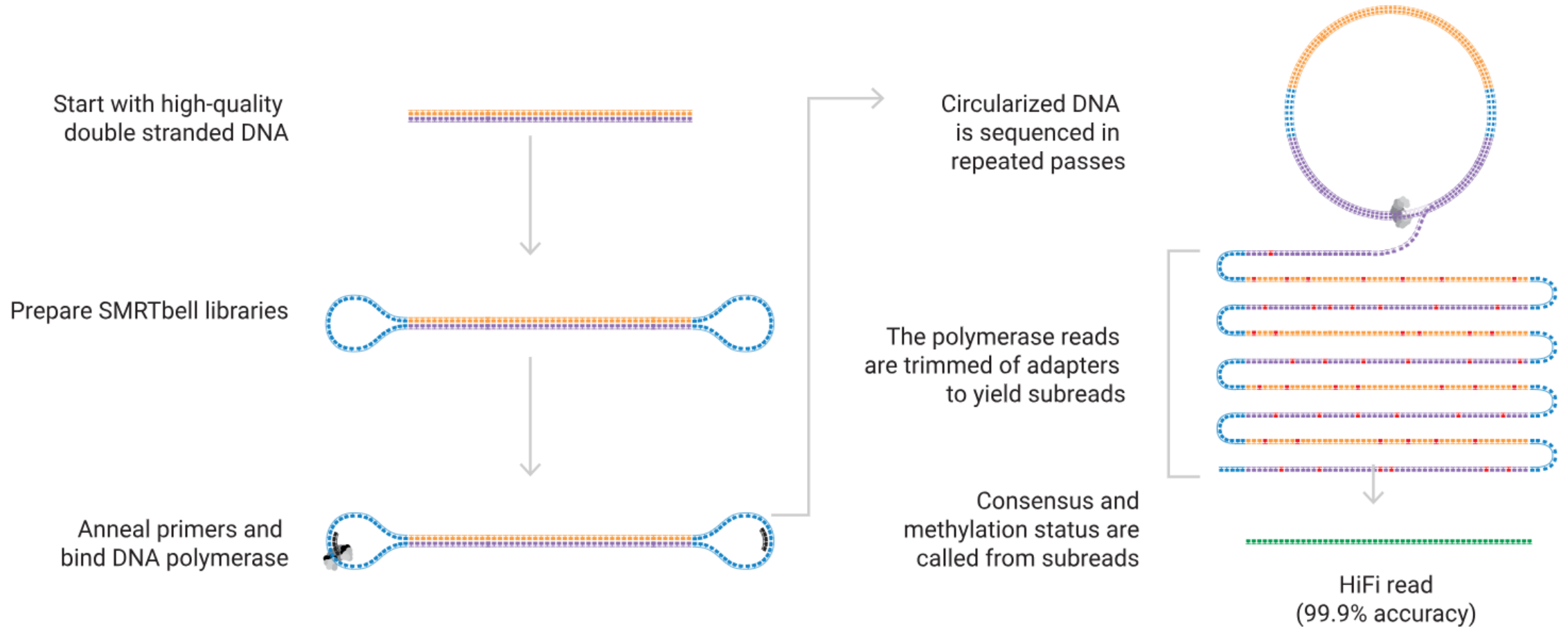
Illumina – Pros & Cons

- 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

<u>Illumina</u>
Cheap
High throughput (Population genomics, GWAS)
Low error rate (<1%)
Better recovery of binding sites within the telomeric repeats
Short reads (150-300 bp)
Long run (1-3.5 days)
Limited <i>de novo</i> genome assembly
Poor coverage of subtelomeric regions
No coverage of long repetitive regions (i.e., Ty elements, rDNA cluster, repeated genes)
mtDNA usually not assembled or analyzed

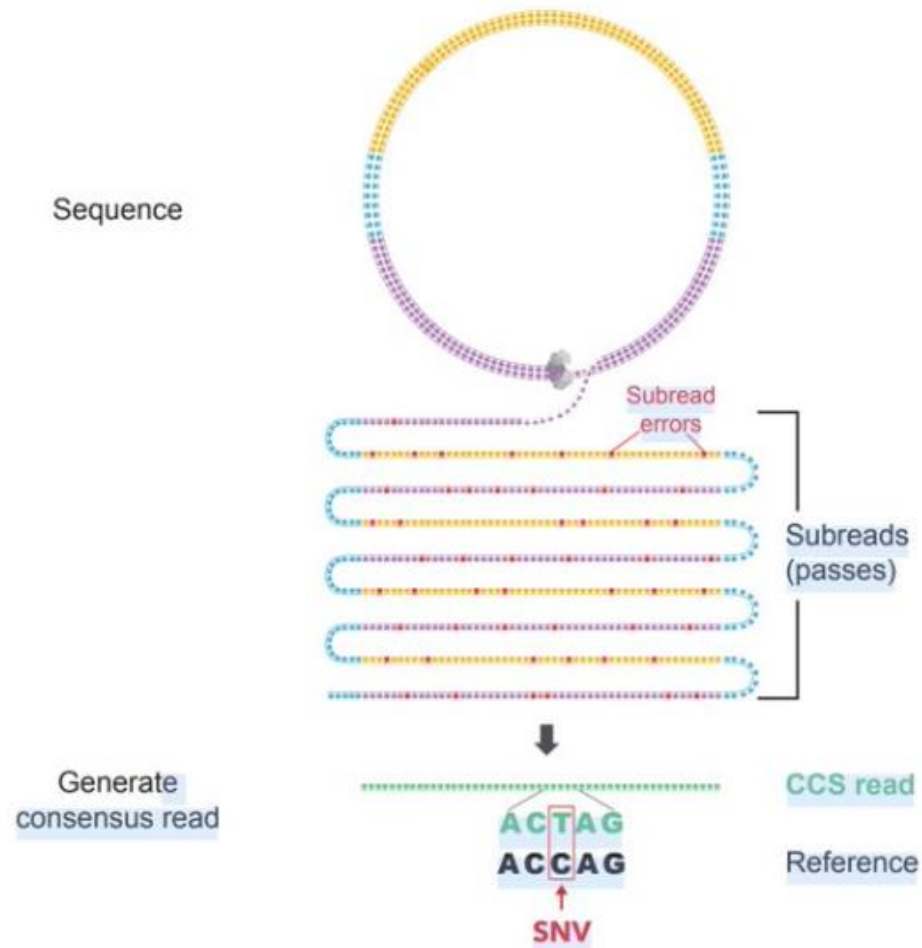
PacBio – SMRT single-molecule real-time (3rd gen)

How are HiFi reads generated?



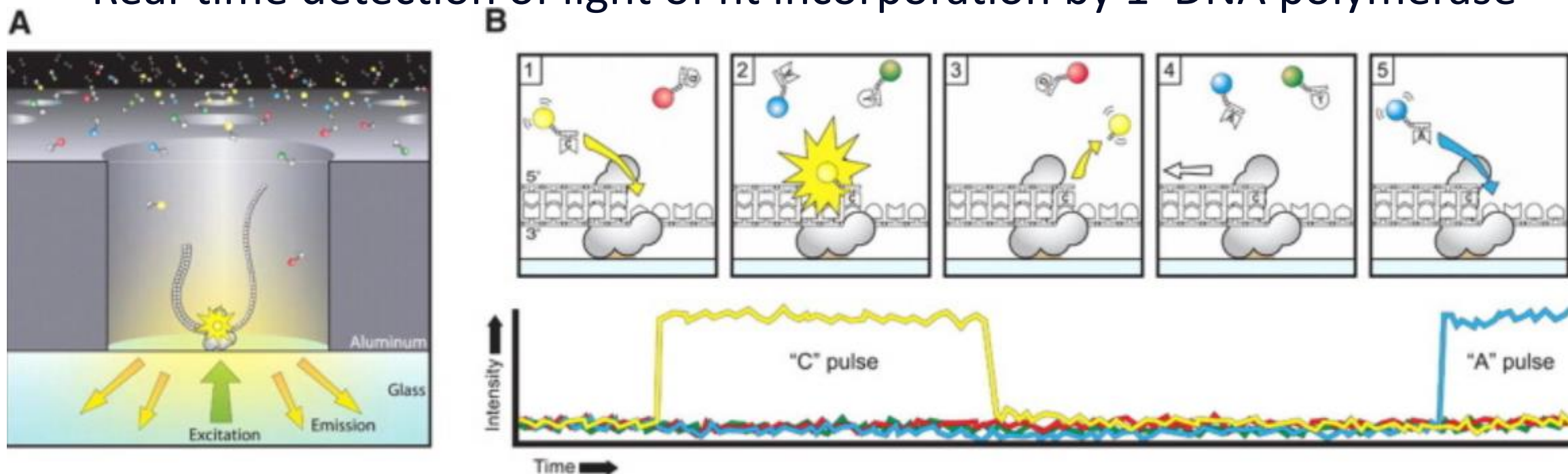
PacBio – Circular consensus reads

HOW DOES CIRCULAR CONSENSUS SEQUENCING WORK



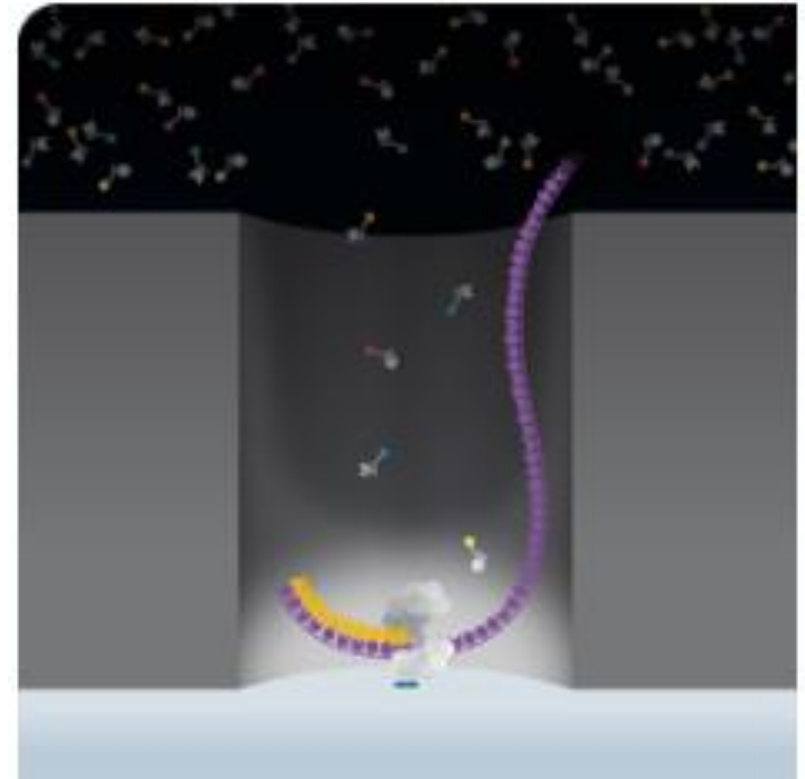
PacBio – 1 molecule/well, DNA poly at bottom

- 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

PacBio

Fast (10 hours)

Mostly complete *de novo* assembly
(Comparative Genomics)

Detects structural variation

Long reads in CLR method (30-250 Kb)

CC method with low error rate (<1%)

Recovery of most subtelomeric regions

Recovery of long repetitive regions

Study of DNA methylation

(4mC and 6mA)

Better recovery of mtDNA

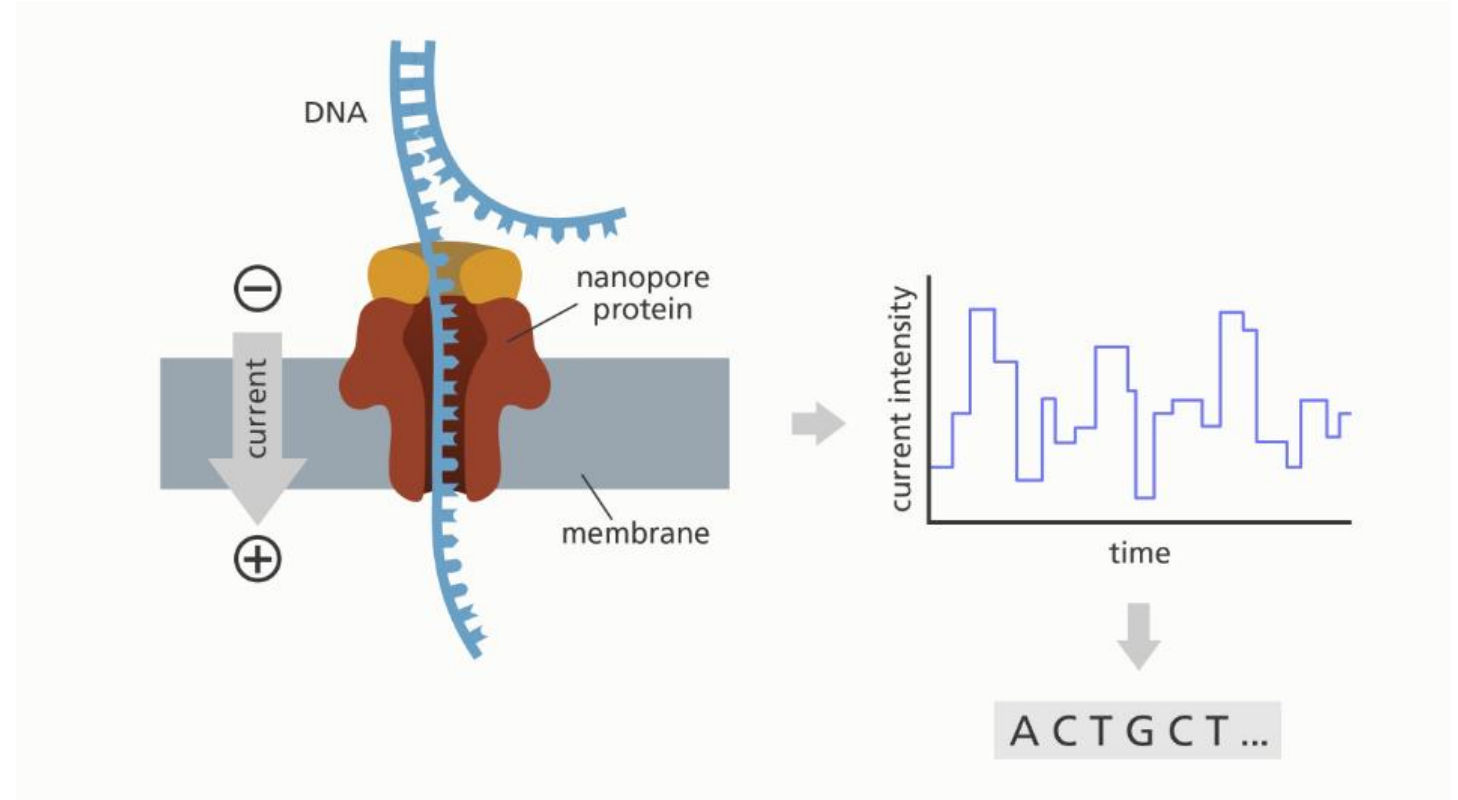
High error rate of CLR method (~13%)

Shorter reads in CC method (25 Kb)

Generates duplicated regions in mtDNA

Oxford Nanopore – Nanopore sequencing (3rd gen)

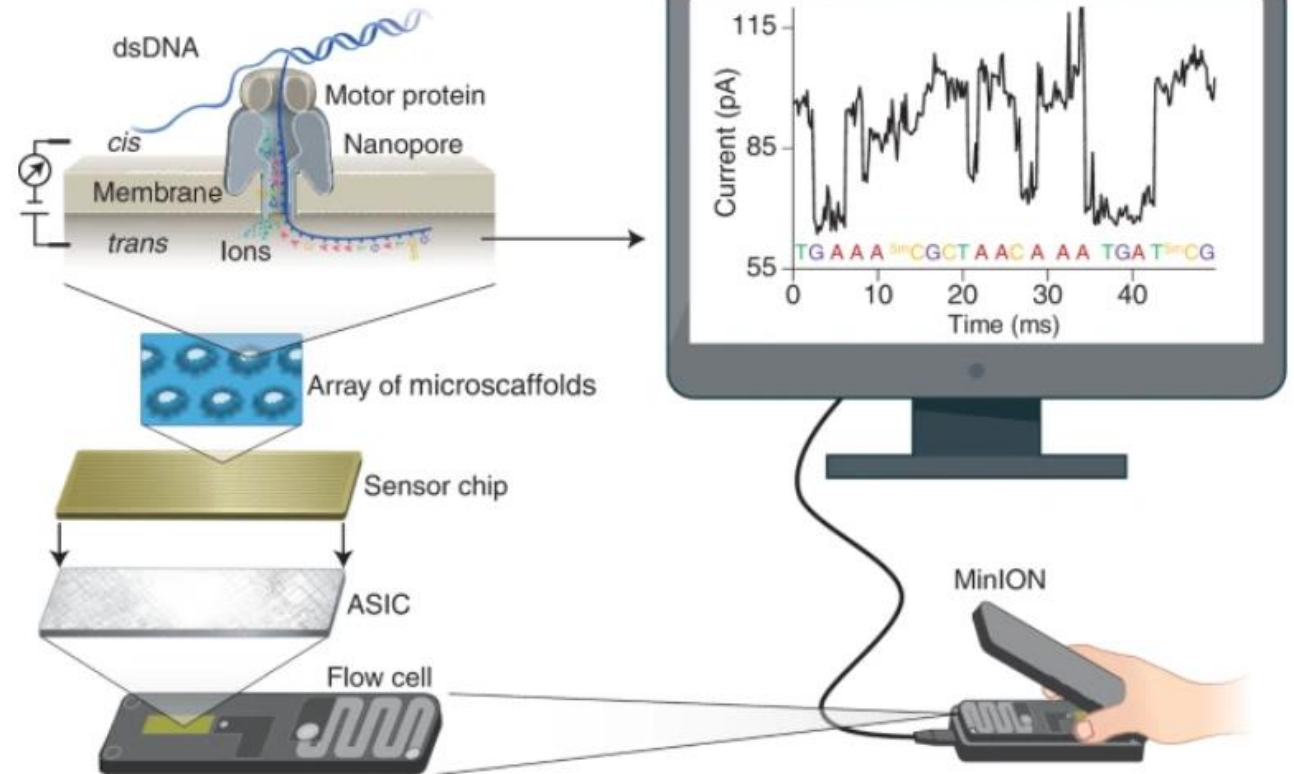
- Pore forming protein alpha-hemolysin
- Inner diameter 1 nm – single strand of DNA/RNA molecule
- Voltage to attract negatively charge DNA/RNA through nanopore -> current variation based on A/C/G/T -> sequence
- Motor protein on top
- No polymerase



Oxford Nanopore – Scalability

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

Fig. 1: Principle of nanopore sequencing.



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

Nanopore

The fastest run (1 hour)

Mostly complete *de novo* assembly
(Comparative Genomics)

Detects structural variation

1D² chemistry with low error rate (<4%)

Recovery of most subtelomeric regions

Recovery of long repetitive regions

Study of DNA methylation (5mC)

High error rate of 1D chemistry (~13%)
Limited coverage of subtelomeric regions

with excessive repeats

Homomer issues

1D² chemistry reduces throughput by half

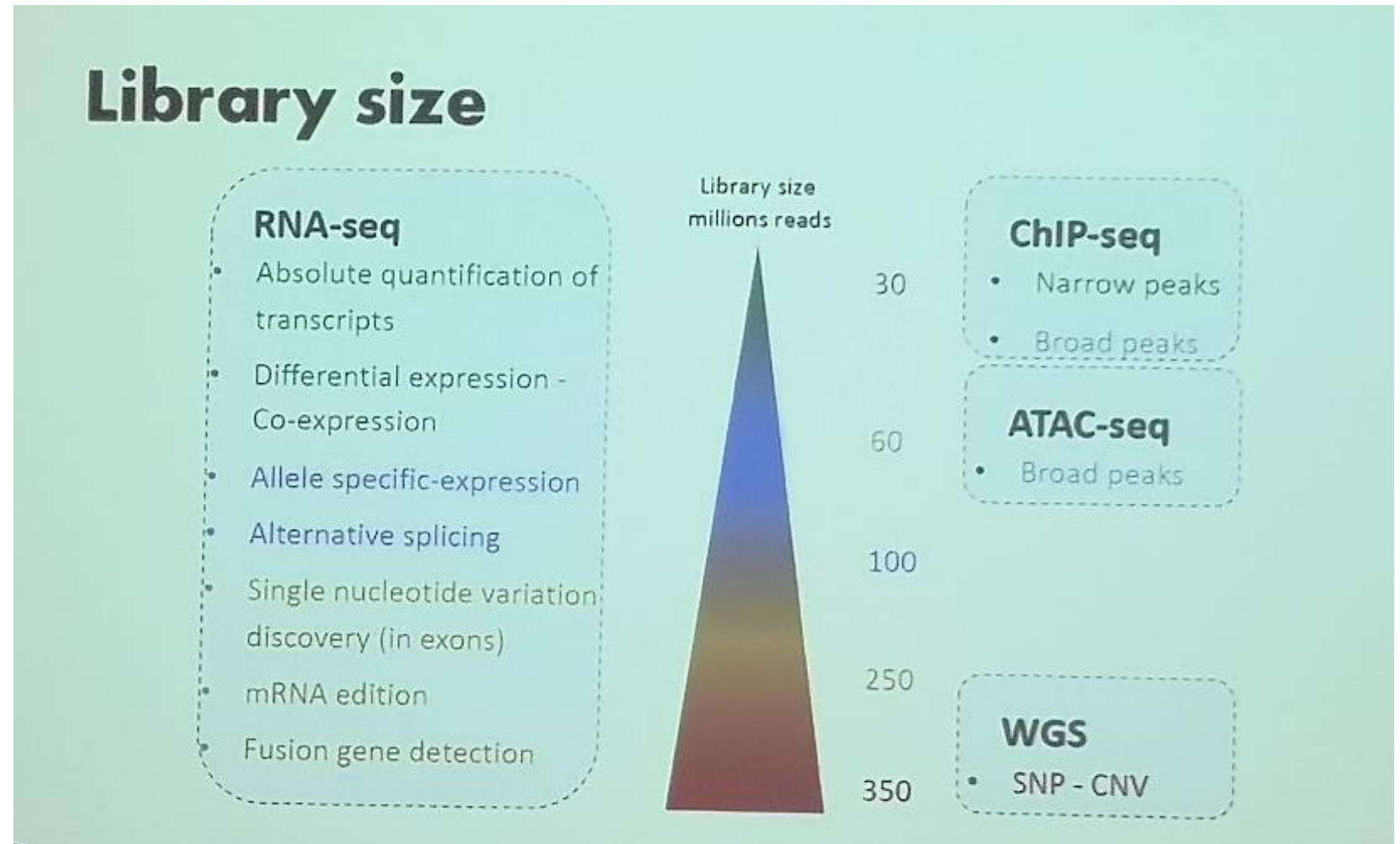
Generates duplicated regions in mtDNA
(using standard assembly methods)

Plan

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

Sequencing needed - Illumina

- Slide from IRCM bioinfo

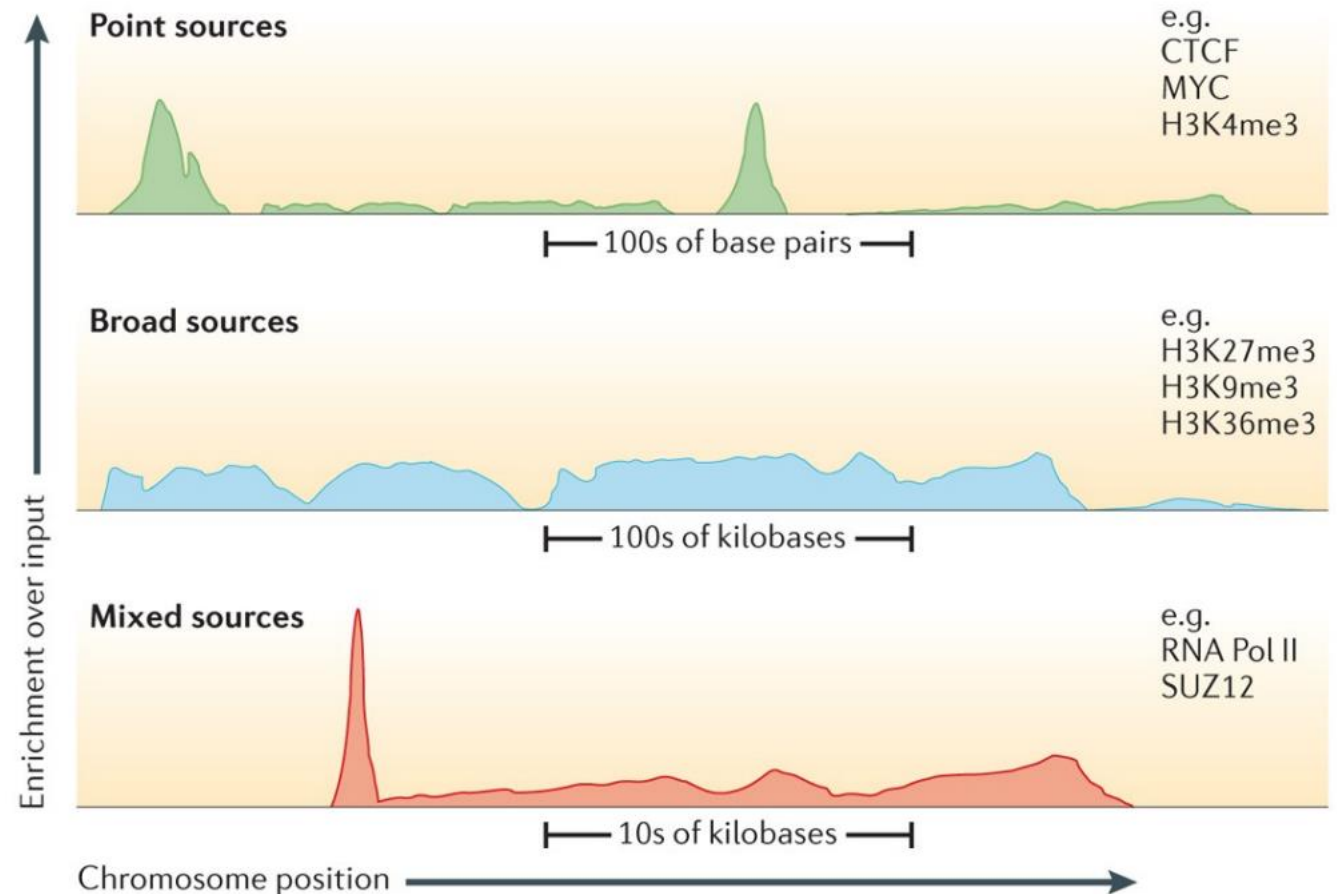


Sequencing needed – Design is key

- 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



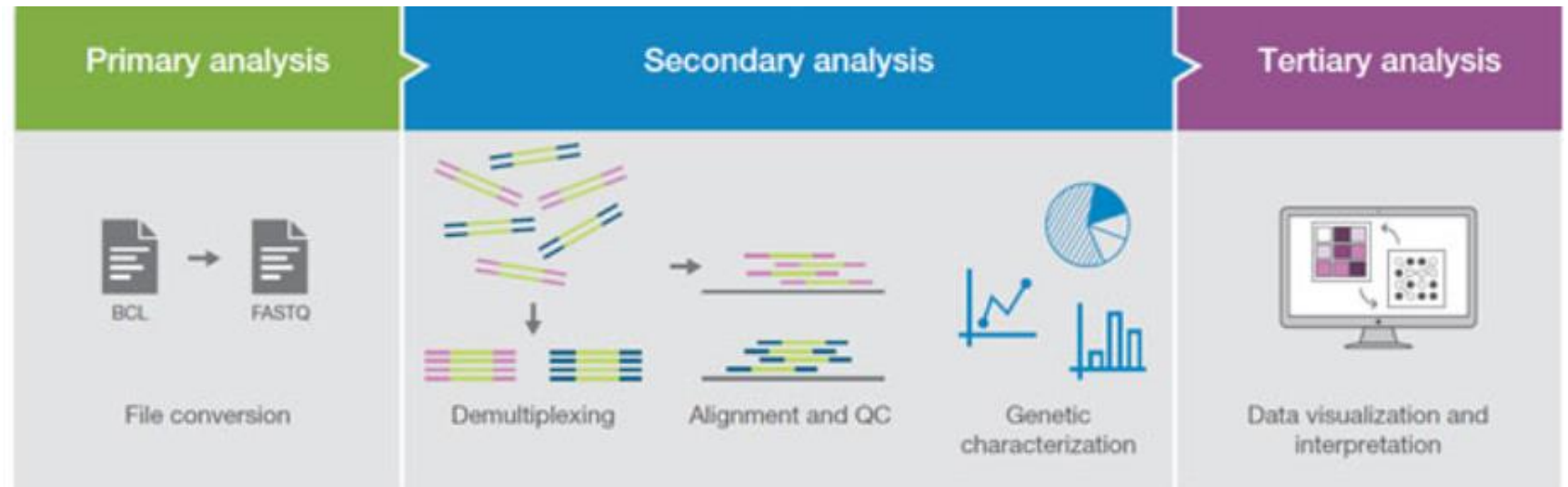
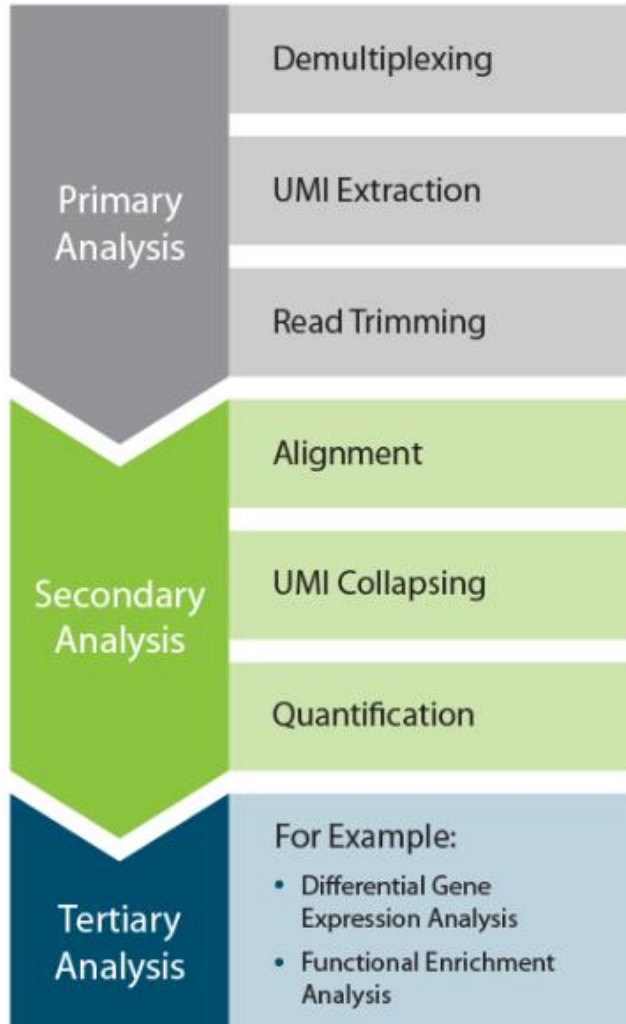
Sequencing needed – Other considerations

- 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

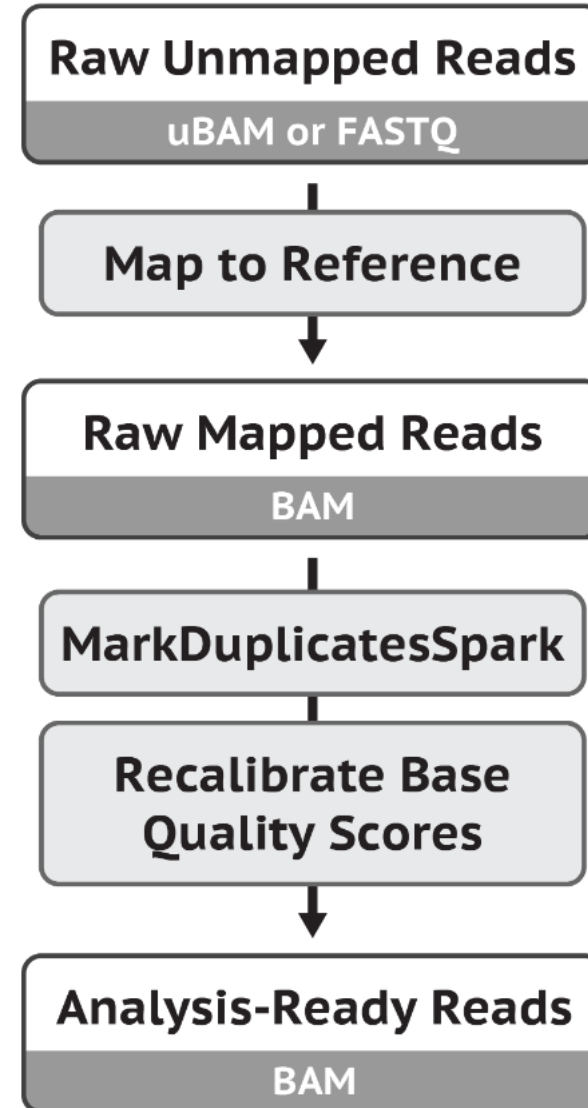


Primary – From image to fastq

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

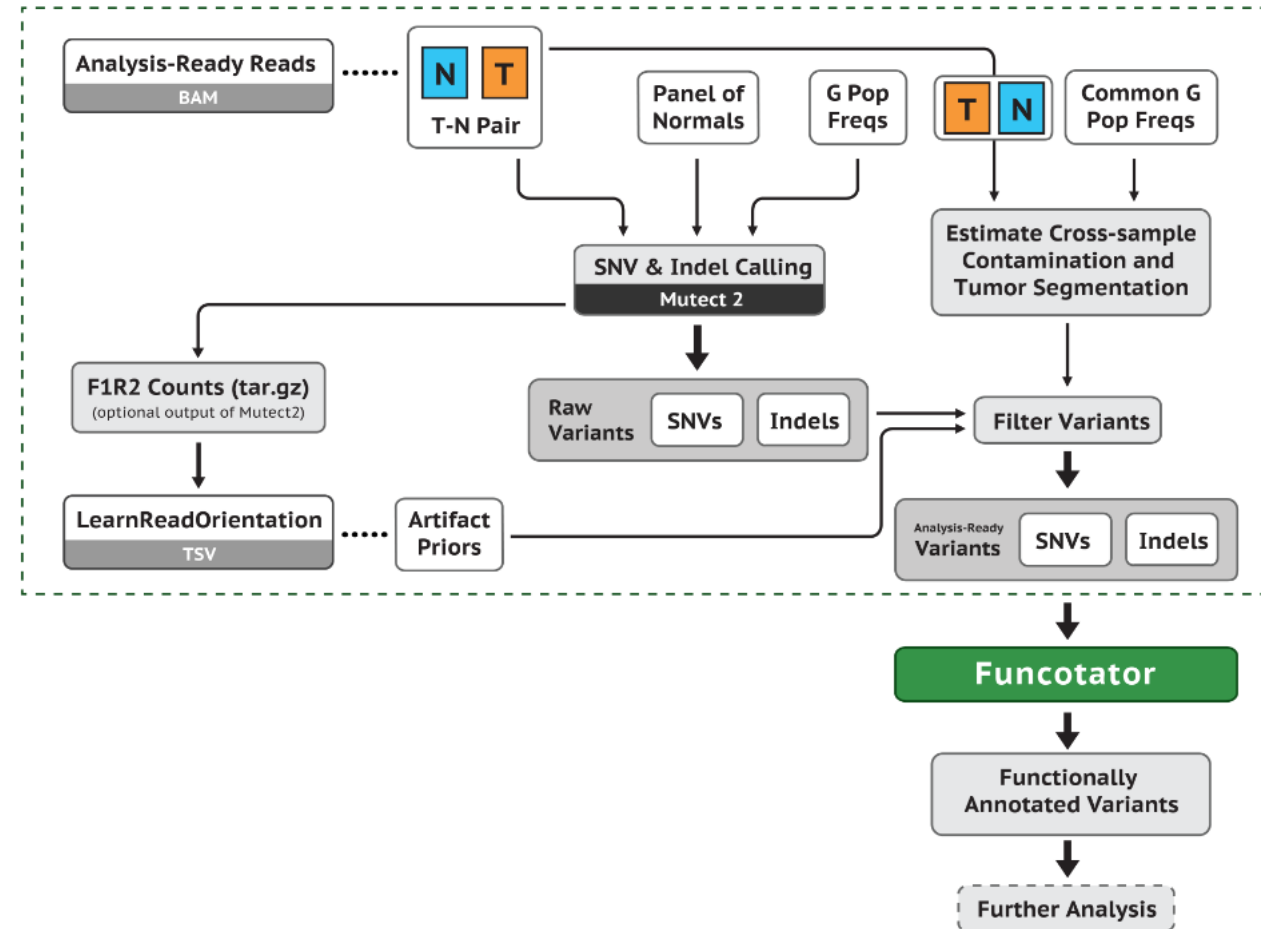
Secondary – From Fastq to counts/variants

- GATK (Genomic Analysis ToolKit)
- Pre-processing



Secondary – From Fastq to counts/variants

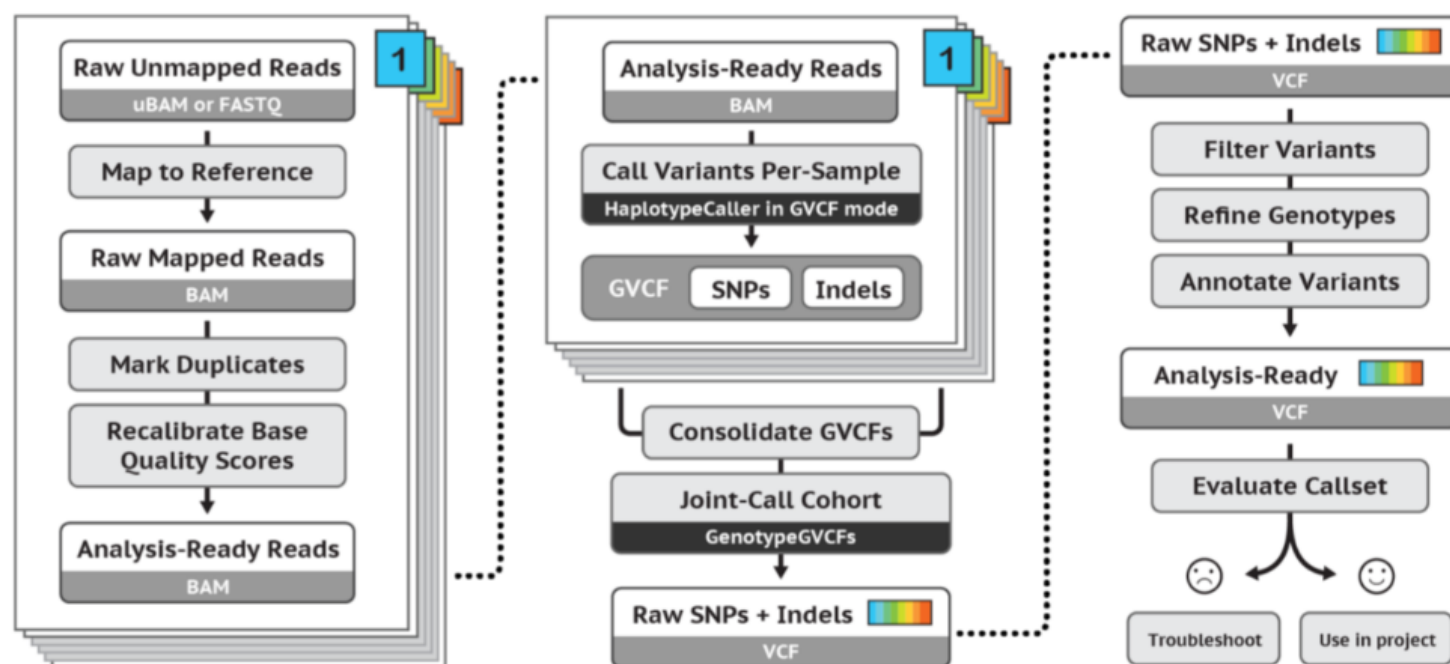
- GATK (Genomic Analysis ToolKit)
- Somatic short variants



Secondary – From Fastq to counts/variants

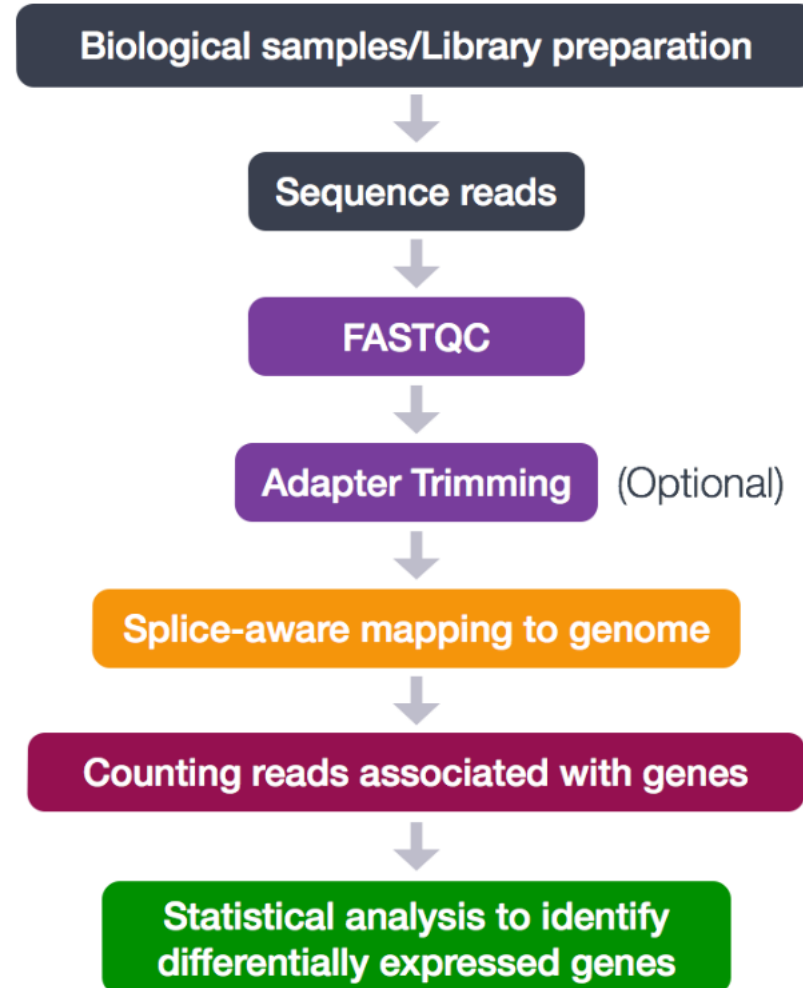
- GATK (Genomic Analysis ToolKit)
- Germline short variants

Main steps for Germline Cohort Data



Secondary – From Fastq to counts/variants

- RNA-Seq workflow (basic)

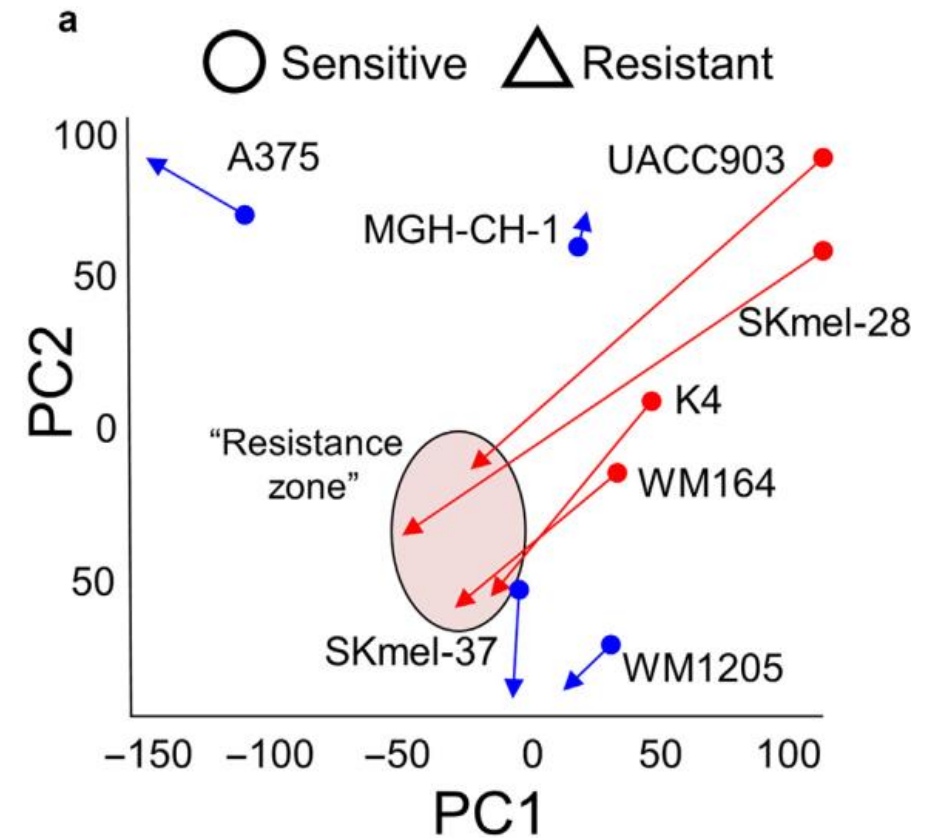


Tertiary - Examples

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

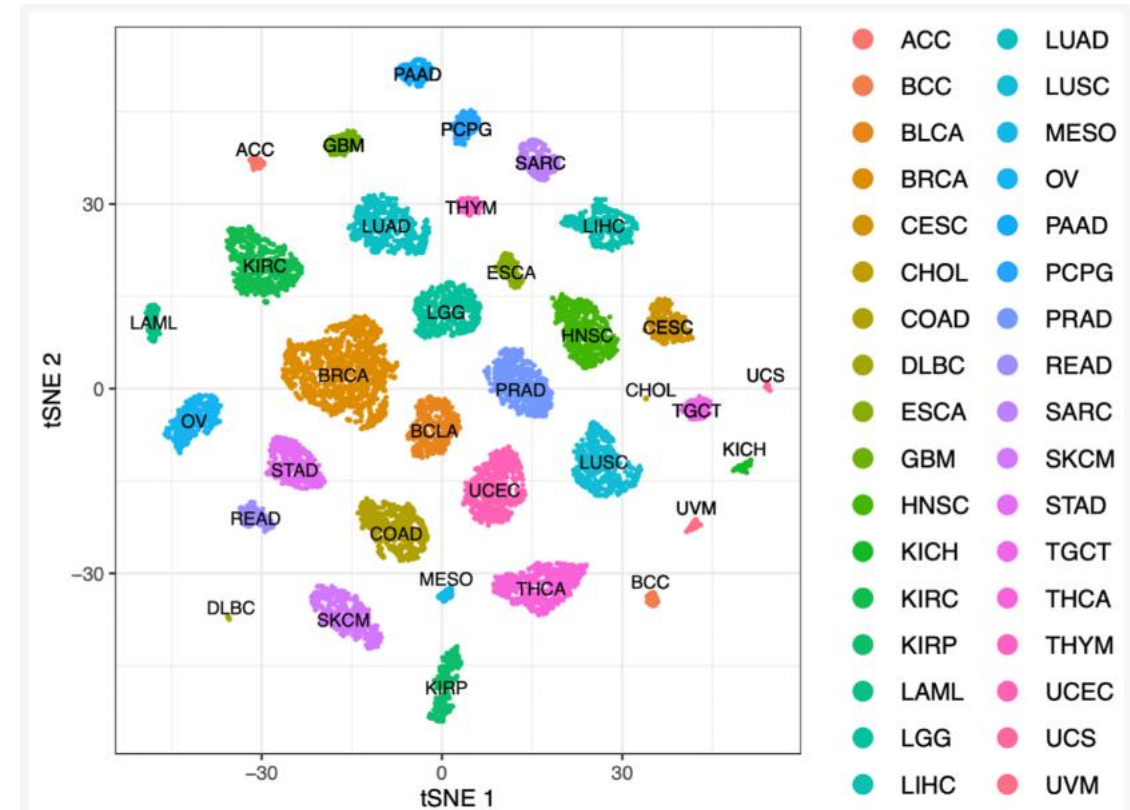
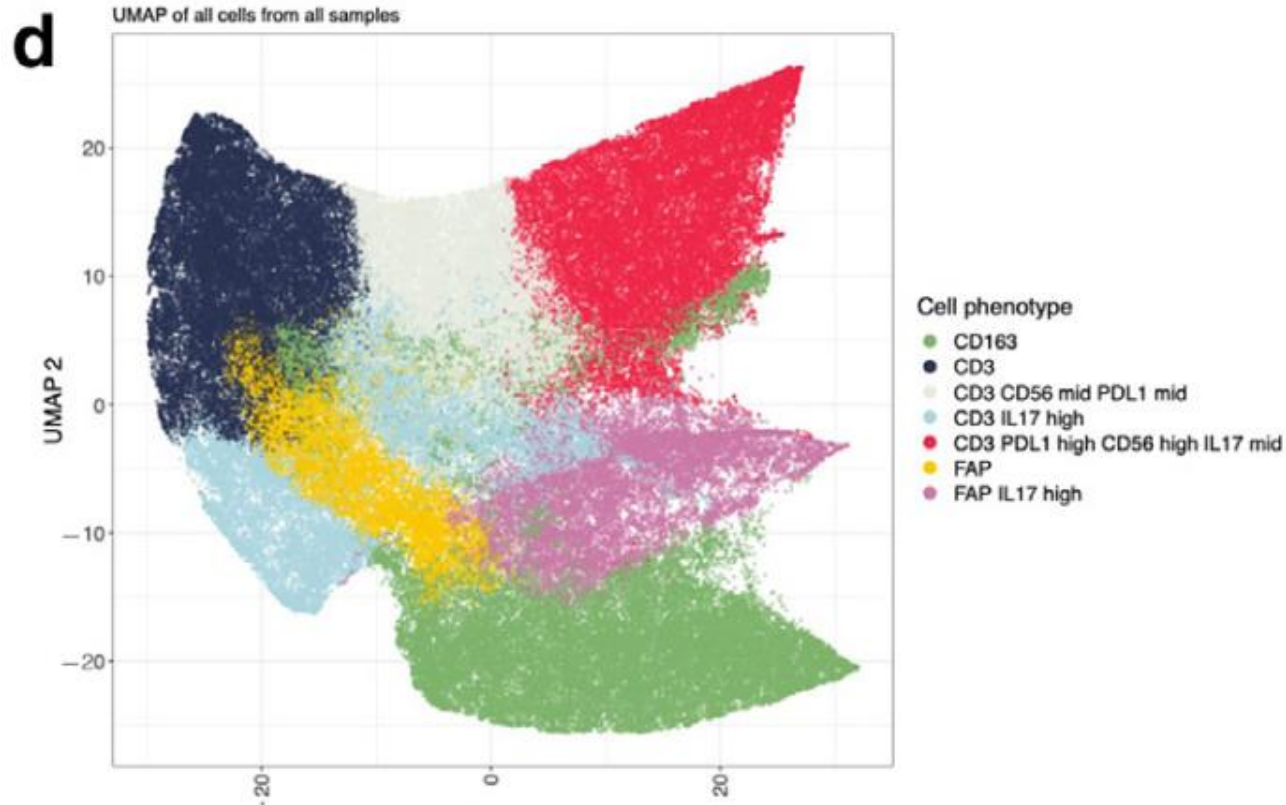
Tertiary – Key concepts

- 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



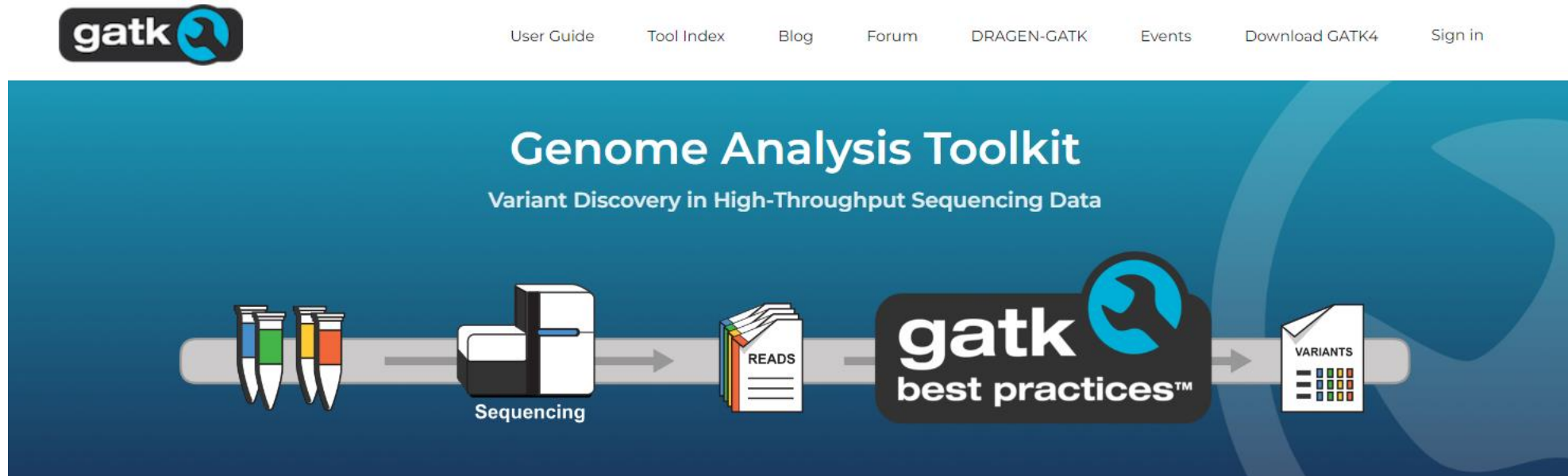
Digital Research
Alliance of Canada

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

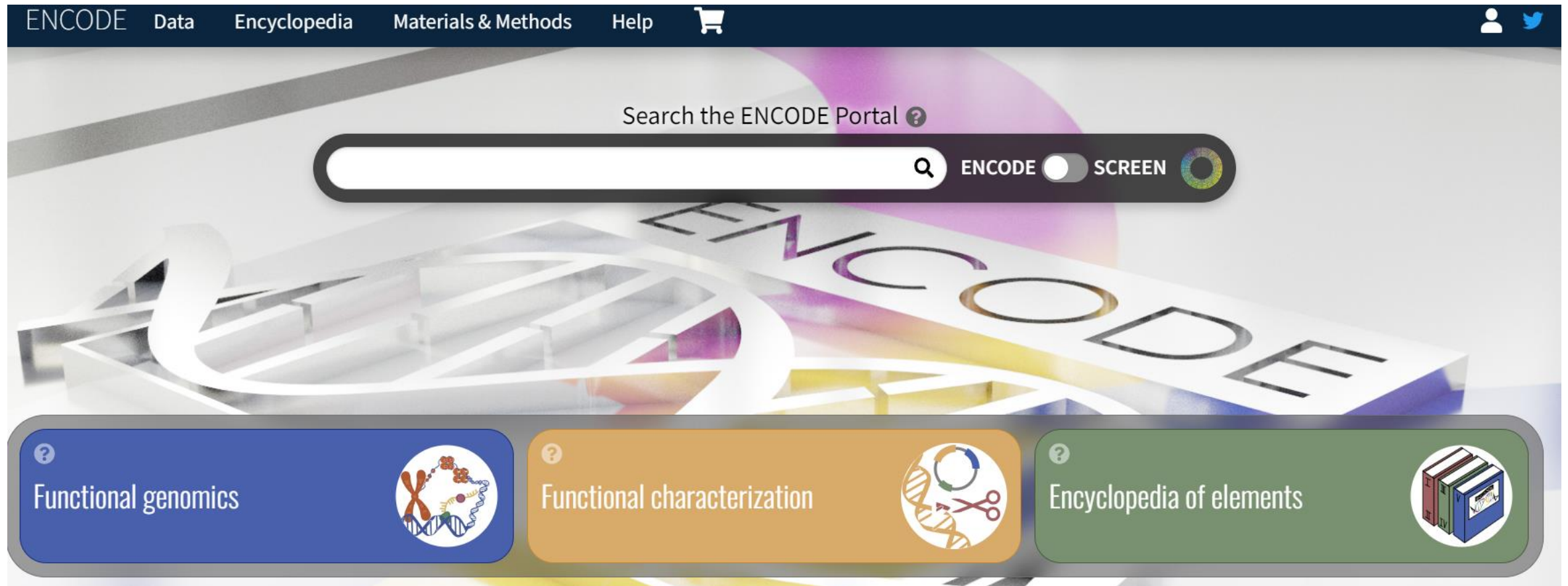
GATK (Genomic Analysis ToolKit)

- Standardized workflows for variant discovery



ENCODE

- For gene regulation, but highly integrated with standardized analyses



The image shows a screenshot of the ENCODE Portal website. At the top, there is a dark blue navigation bar with the following links: ENCODE, Data, Encyclopedia, Materials & Methods, Help, and a shopping cart icon. On the right side of the navigation bar, there are icons for a user profile and Twitter. Below the navigation bar is a large search bar with the text "Search the ENCODE Portal" and a magnifying glass icon. To the right of the search bar, there is a toggle switch labeled "ENCODE" and "SCREEN" with a circular color wheel icon. The background of the page features a 3D rendering of a DNA double helix and the word "ENCODE" in large, stylized letters. At the bottom of the page, there are three colored buttons with icons and text: a blue button for "Functional genomics" with a DNA and protein icon, an orange button for "Functional characterization" with a DNA and scissors icon, and a green button for "Encyclopedia of elements" with a book icon.

UseGalaxy

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


The screenshot displays the Galaxy web interface. On the left is the 'Tools' panel with a search bar and various tool categories like 'GENERAL TEXT TOOLS' and 'GENOMIC FILE MANIPULATION'. The central area features a yellow-highlighted text box containing information about laboratories for Ukrainian scientists, including the email ukraine@galaxyproject.org. Below this is a paragraph describing Galaxy as an open-source platform for biomedical research. On the right is the 'History' panel, which is currently empty and displays a message: 'This history is empty. You can load your own data or get data from an external source.' The top navigation bar includes 'Workflow', 'Visualize', 'Shared Data', 'Help', 'Login or Register', and a 'Using 0%' indicator. At the bottom, the version information is shown: 'Galaxy version 23.0.1.dev0, commit 6c4141bee122d505178160f9f1ce05a6563ef950'.

Biostars

- Forum with Q&As and tutorials for computational biology

The screenshot shows the top navigation bar of the Biostars website. It includes a menu with 'Latest' (selected), 'Open', 'Jobs', 'Tutorials', 'Tags', 'About', and 'FAQ'. Below this is the Biostars logo with the tagline 'BIOINFORMATICS EXPLAINED', and links for 'Community', 'Planet', and 'New Post'. A 'Log In' button is on the right. The search area features a search bar with 'Search ...' and a magnifying glass icon. Below the search bar, it displays 'Limit: all time', navigation arrows, '110,502 results • Page 1 of 2211', and a 'Sort: Rank' dropdown menu. On the right, a 'Recent Votes' section shows a thumbs-up icon and the text 'Comment: Limma returned only positive logFC values'.

Latest Open Jobs Tutorials Tags About FAQ

 **Biostars**
BIOINFORMATICS EXPLAINED

Community Planet New Post Log In

Search ...

Limit: all time << 110,502 results • Page 1 of 2211 >> Sort: Rank

Recent Votes

Comment: Limma returned only positive logFC values

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](#)
[D 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:



[Home](#)

[Install](#)

[Help](#)

[Developers](#)

[About](#)

Search:

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