CEITEC

Central European Institute of Technology
BRNO | CZECH REPUBLIC

Bi7420: Moderní metody pro analýzu genomu

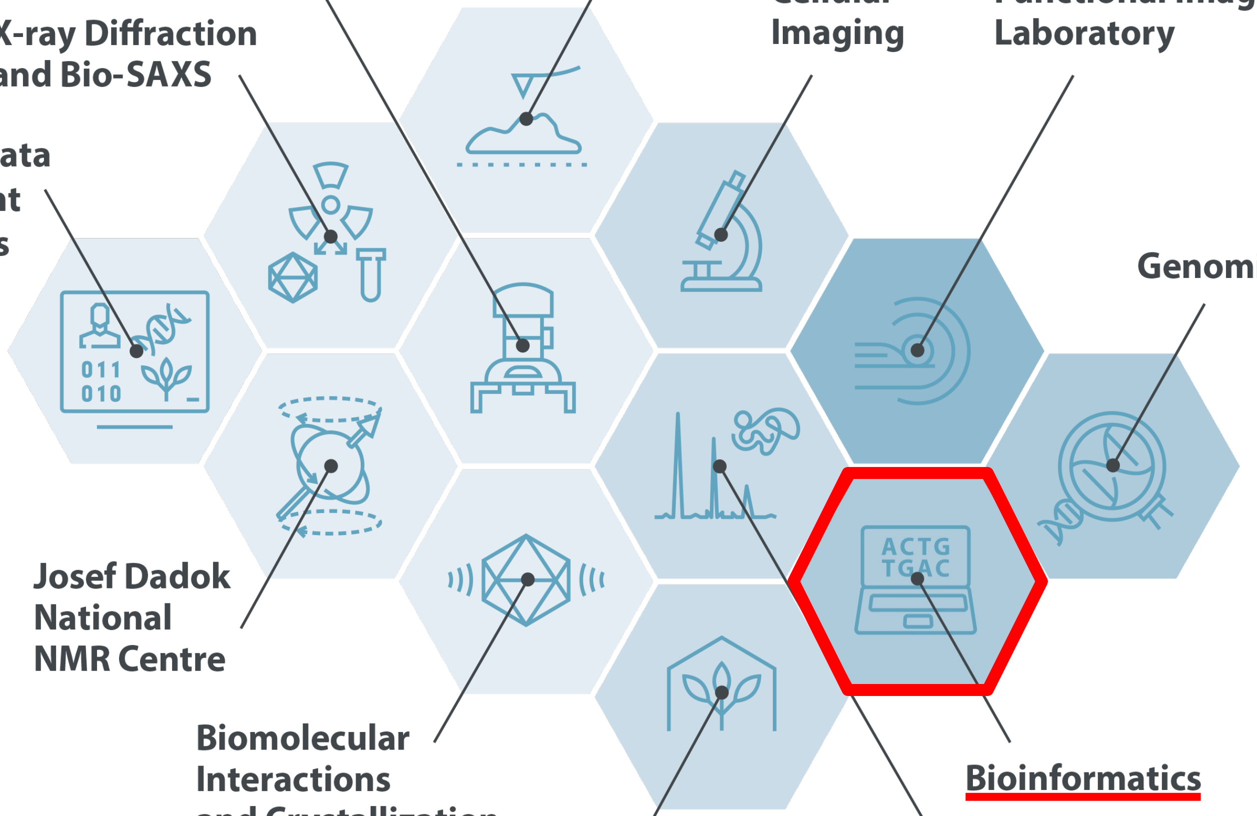
NGS data analysis introduction

Vojta Bystry
vojtech.bystry@ceitec.muni.cz

Plan for **Bi7420**

- Next generation sequencing methods overview
  - Focus on experiment planning and result interpretation

1. Introduction to NGS technology
2. Basic QC, DNA resequencing
3. DNA resequencing, Clinical genomics
4. miRNA, lncRNA in cancer - Marek Mráz
5. RNA-seq
6. RNA-seq, Single-cell RNA-seq, Spatial transcriptomics
7. Chip-seq (CLIP-seq), other methods



Cryo-electron Microscopy and Tomography

Nanobiotechnology

X-ray Diffraction and Bio-SAXS

Cellular Imaging

Multimodal and Functional Imaging Laboratory

Biological Data Management and Analysis

Genomics

Josef Dadok National NMR Centre

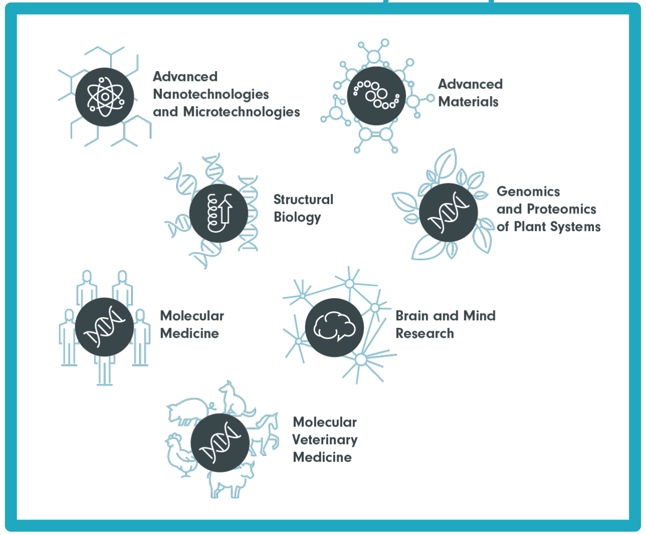
Biomolecular Interactions and Crystallization

Bioinformatics

Plant Sciences

Proteomics

Research groups



Advanced Nanotechnologies and Microtechnologies

Advanced Materials

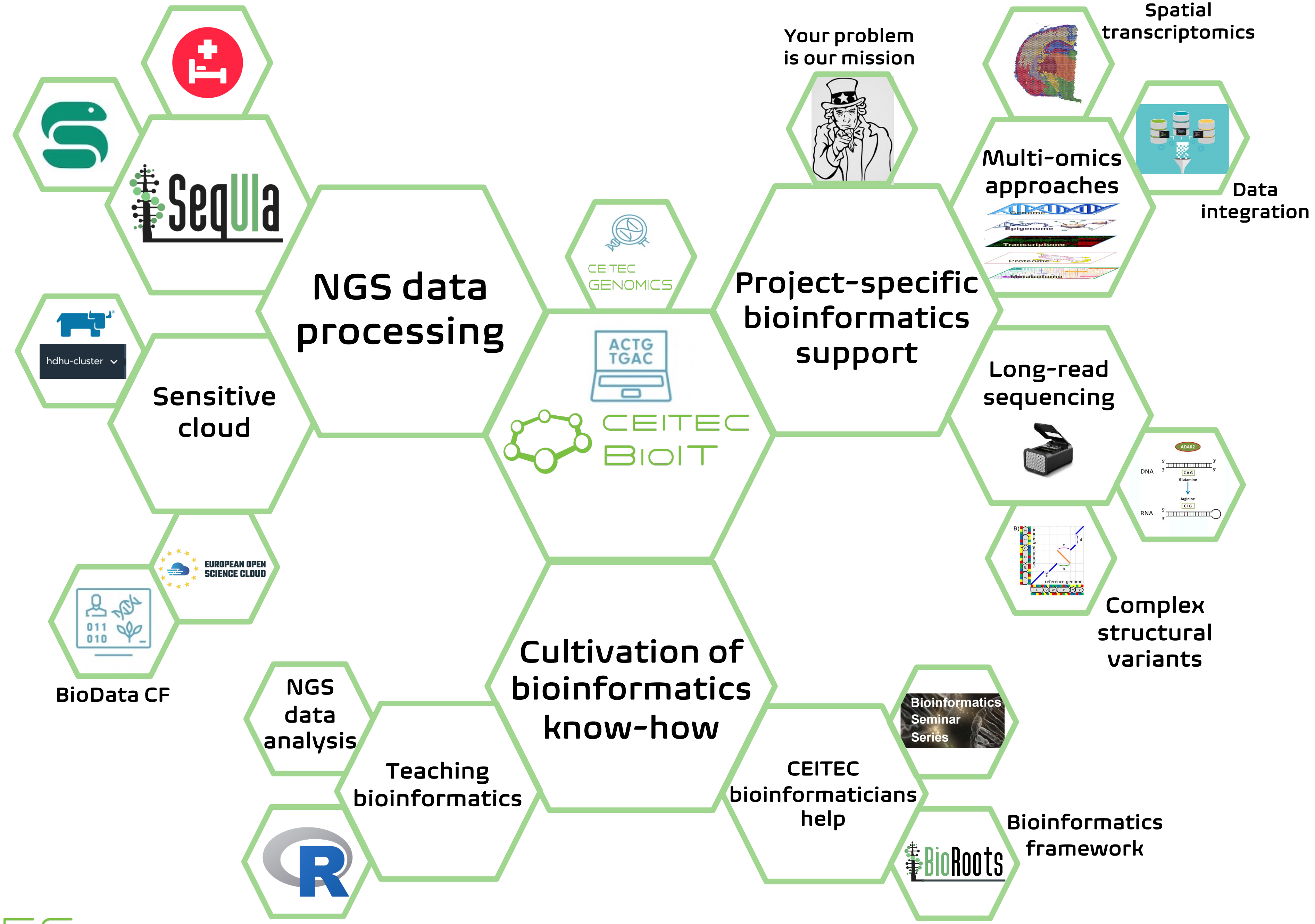
Structural Biology

Genomics and Proteomics of Plant Systems

Molecular Medicine

Brain and Mind Research

Molecular Veterinary Medicine



NGS data processing

Sensitive cloud

BioData CF

Project-specific bioinformatics support

Your problem is our mission

Spatial transcriptomics

Multi-omics approaches

Data integration

Long-read sequencing

Complex structural variants

Cultivation of bioinformatics know-how

Teaching bioinformatics

NGS data analysis

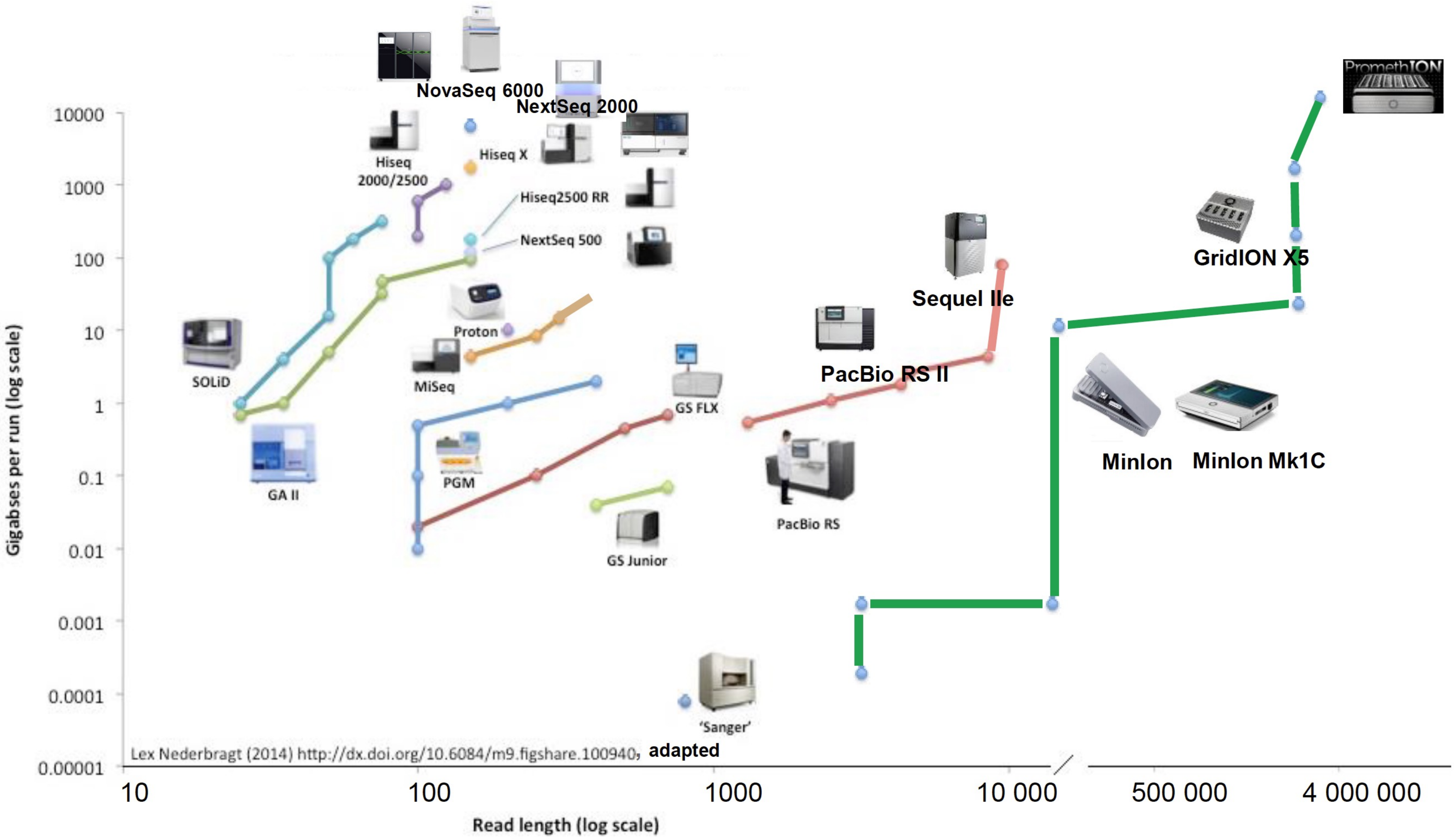
CEITEC bioinformaticians help

Bioinformatics framework

What is NGS?

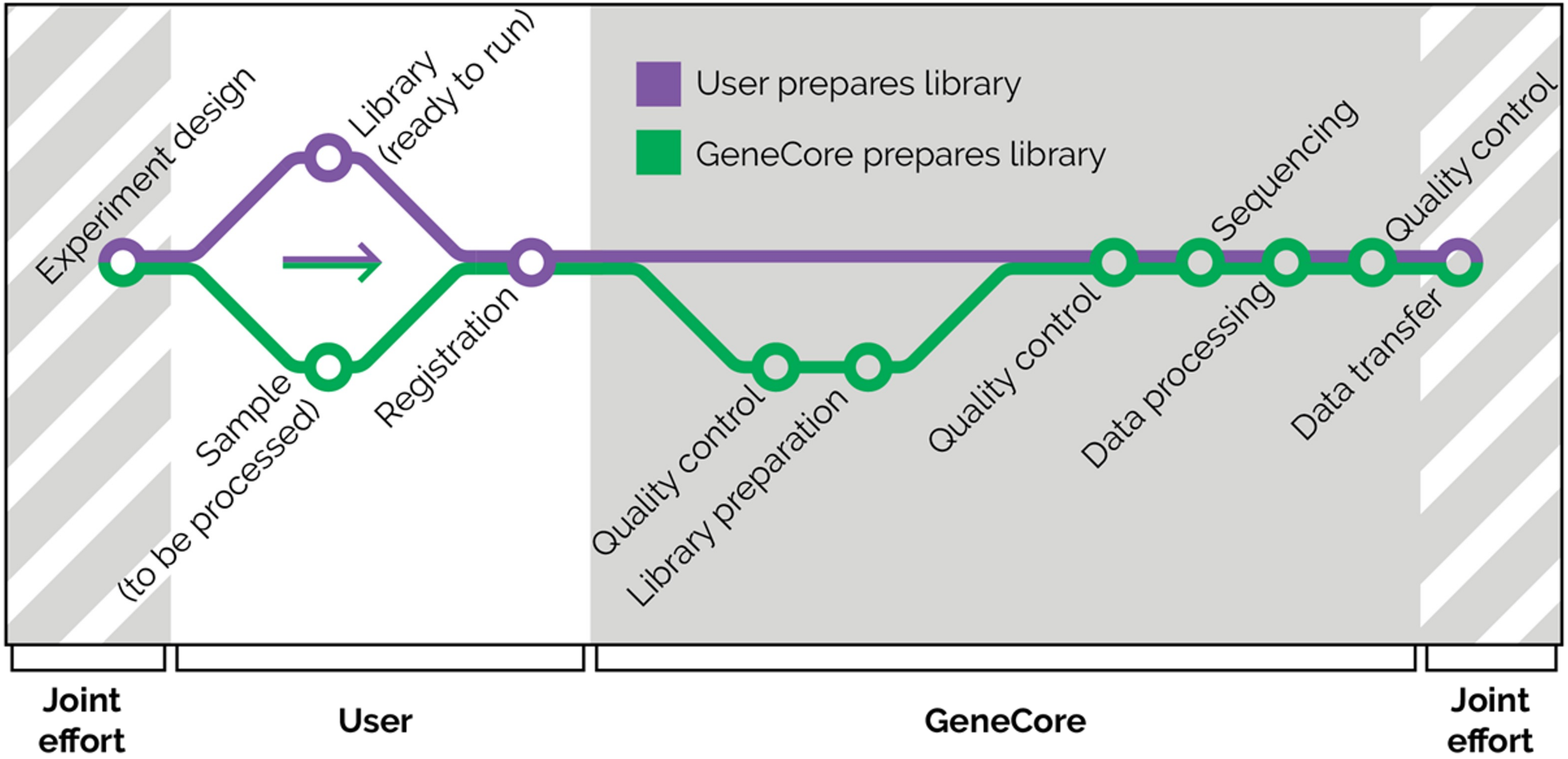
- Next generation sequencing
  - New generation sequencing
  - HTP = High throughput sequencing
  - Massively parallel sequencing
- Contrast to Sanger sequencing

What is NGS?



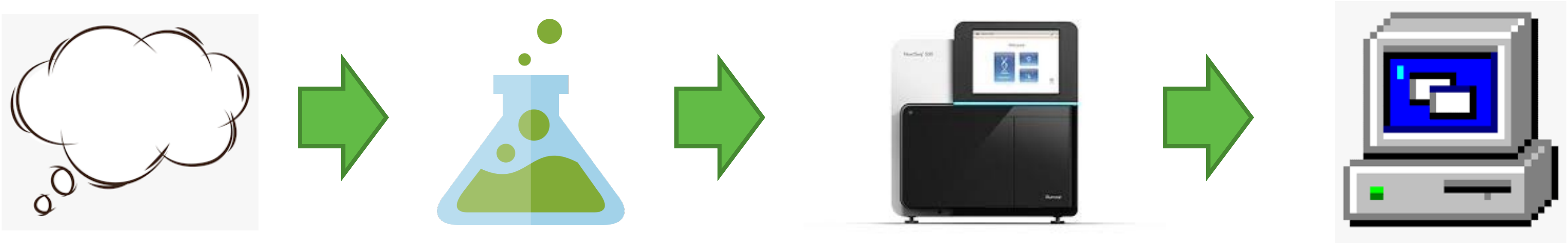
Lex Nederbragt (2014) http://dx.doi.org/10.6084/m9.figshare.100940, adapted

NGS experiment workflow



CEITEC

NGS experiment workflow



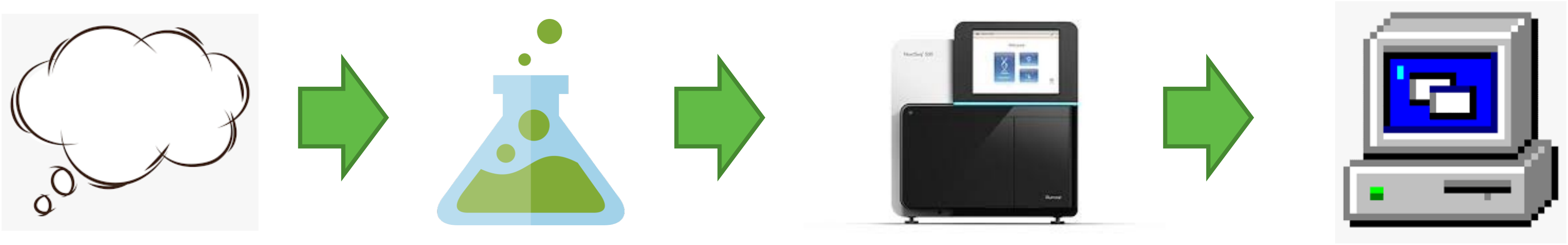
Experimental design

Library preparation

Sequencing

Data analysis

NGS experiment workflow



Experimental design

Library preparation

Sequencing

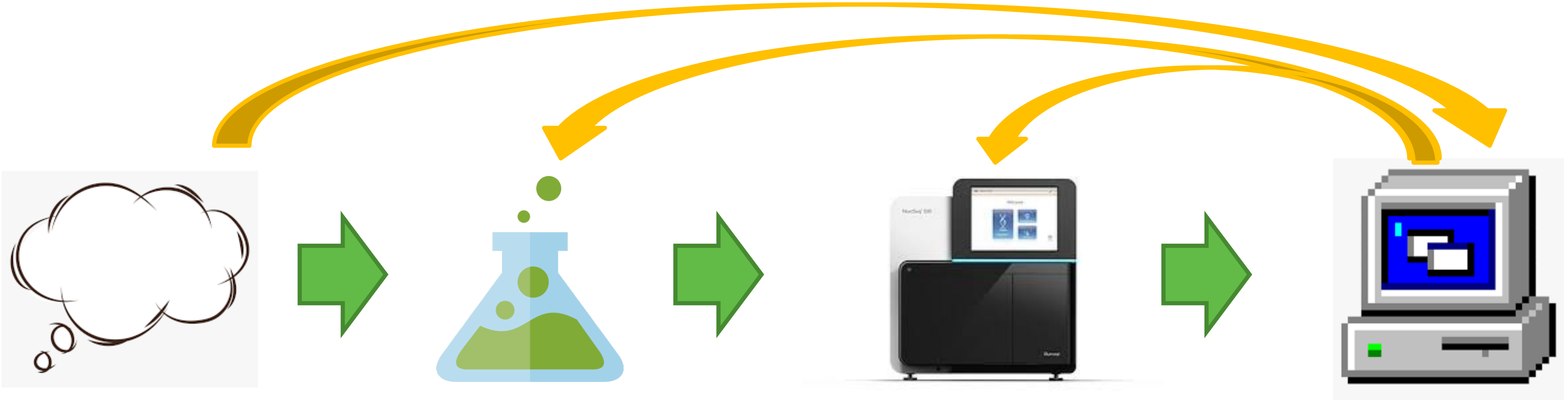
Data analysis

Why we sequence

What we sequence

How we sequence

NGS experiment workflow



Experimental design

Library preparation

Sequencing

Data analysis

Why we sequence

What we sequence

How we sequence

Consultation regarding data analysis is highly advisable.

Vocabulary

Library: Fragmented DNA with technical sequences attached

Pool: Mix of different libraries, that are sequenced in one run

Read: String of letters coming out of a sequencer

Depth: How many reads we have coming from a single region of our reference

Flow Cell: The glass slide where sequencing happens

Barcode / Index: Technical sequence used to differentiate samples

Adapter: Technical sequence used to anchor the template to the Flow Cell

NGS sequencing technologies

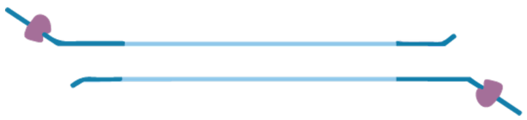
Currently provided sequencing technologies:

Illumina: NovaSeq, NextSeq 500, MiSeq
PacBio: Sequel IIe
Oxford Nanopore: GridION, PromethION P2 Solo

ONT

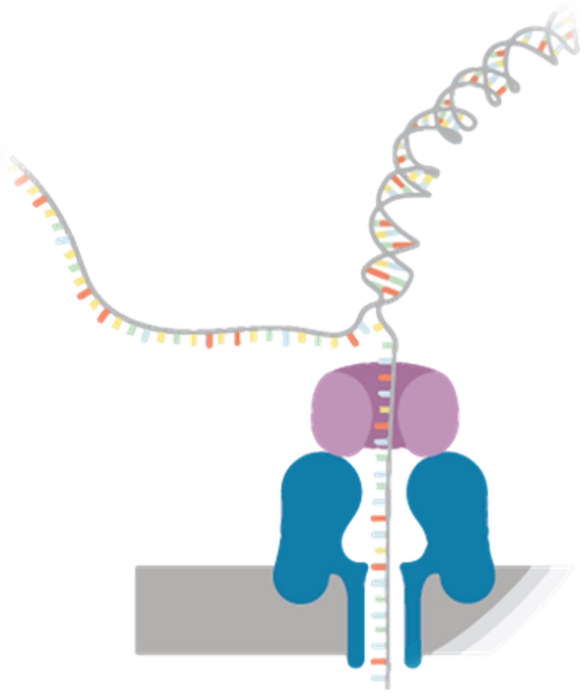
Input material
(example: DNA)

Library preparation



Nanopore
adapters with
motor proteins

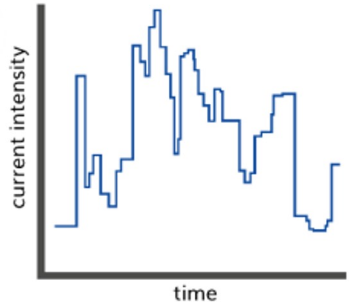
Flow cell loading



Nucleic acids fed
through pore,
generating current

Shifts in
electric current

Current is interpreted by
algorithms to generate
sequence - including base
modifications: squiggle



ONT

The ONT sequencers:

1. MinION/Flongle
2. GridION
3. PromethION P2 Solo (developer version)

MinION
1 Flowcell
512 channels/FC
10-15Gb
(~900 €/FC)

GridION
5 Flowcells
512 channels/FC
10-15Gb/FC
(~900 €/FC)

P2 Solo
2 Flowcells
2675 channels/FC
100-120Gb
(~1600€/FC)

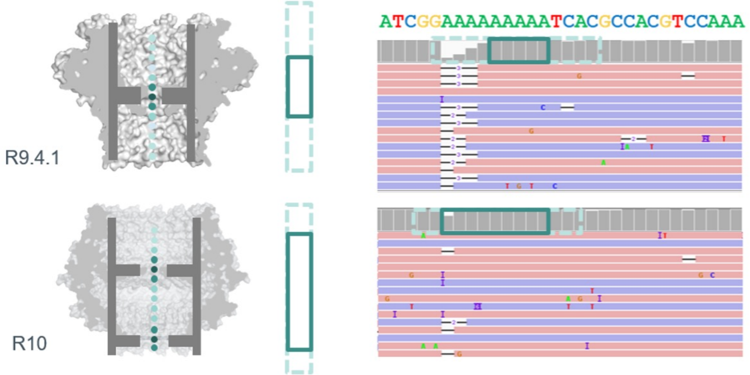
ONT - “news”

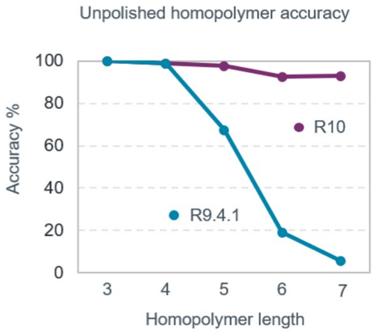
Chemistry V14

- New motor protein
- New buffer composition - lower pH
- Lower flowcell loading amounts

Pore R10.4.1

- Improved enzyme-pore docking
- Faster speeds
- Higher output yield

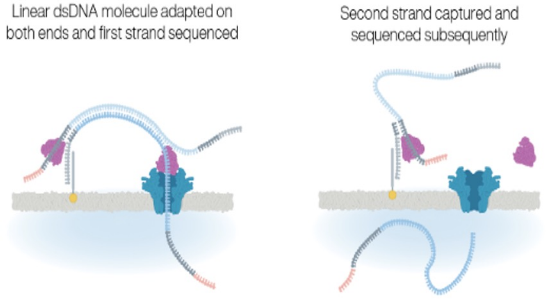




New instrument: P2 Solo

- Connected to GridION
- Two high-yield flowcells (100-200Gb)

Duplex reads for higher accuracy

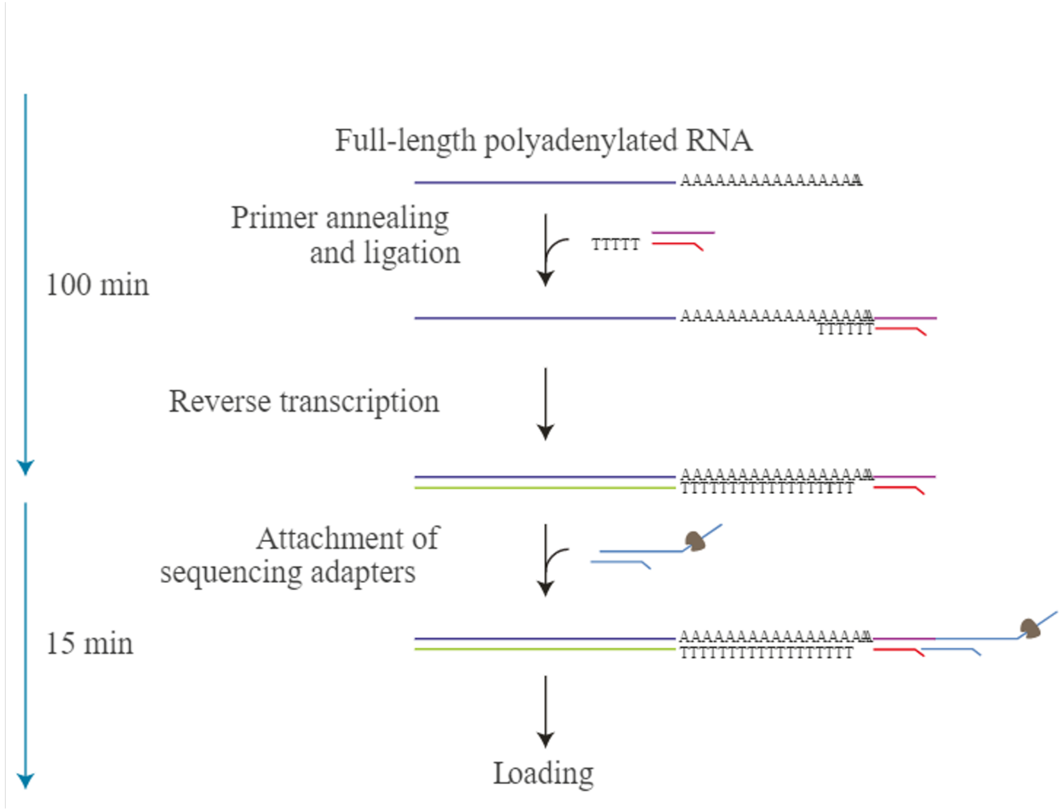

Linear dsDNA molecule adapted on both ends and first strand sequenced
Second strand captured and sequenced subsequently

POD5: New output file format

- Smaller file
- Faster file writing
- Incompatible with most current tools

ONT

Direct RNA

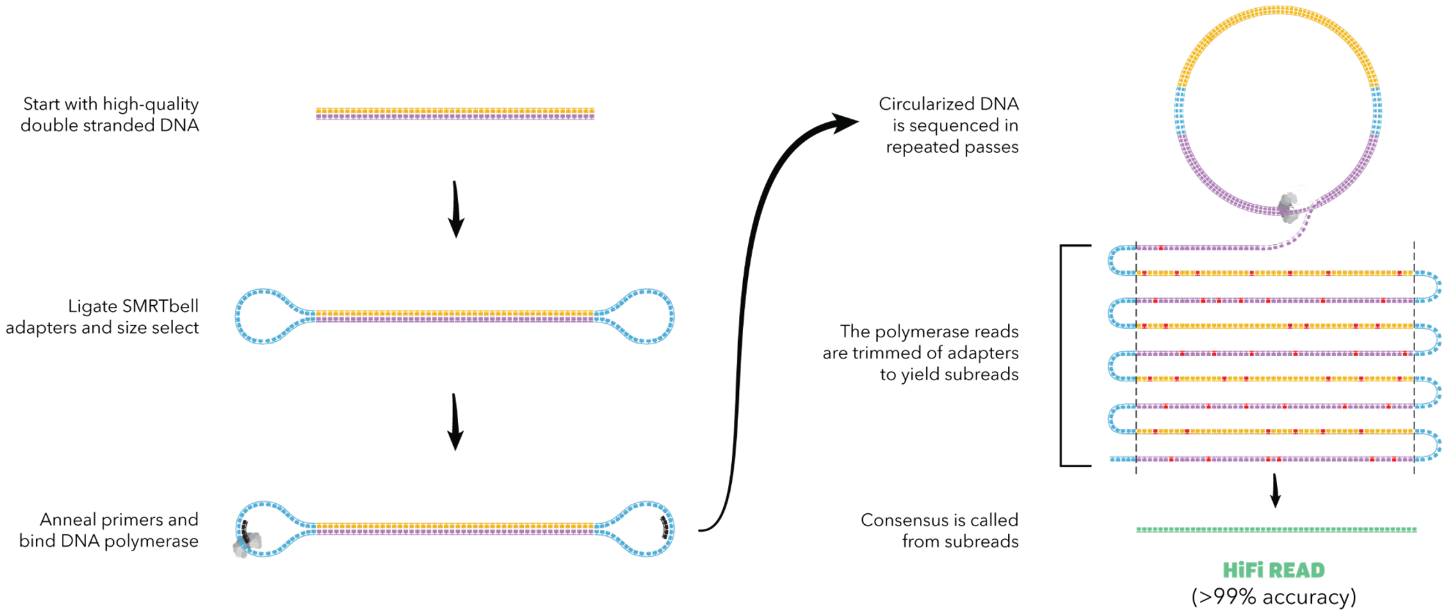

Full-length polyadenylated RNA
AAAAAAAAAAAAAAAA
Primer annealing and ligation
TTTTT
100 min
Reverse transcription
Attachment of sequencing adapters
15 min
Loading

Additional kits available:

- cDNA sequencing kit
  PCR full length transcripts
- 16S sequencing kit
  PCR
- PCR sequencing kit
  targeted amplicon

PacBio

SMRT Sequencing: Single Molecule Real-Time Sequencing



Start with high-quality double stranded DNA

Ligate SMRTbell adapters and size select

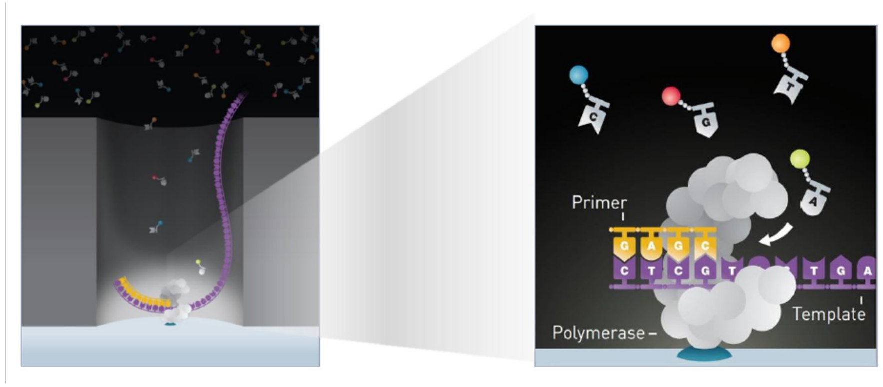
Anneal primers and bind DNA polymerase

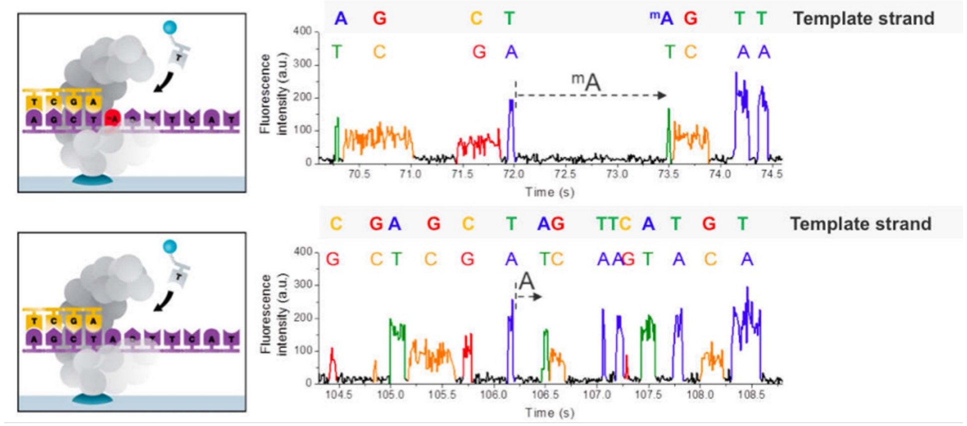
Circularized DNA is sequenced in repeated passes

The polymerase reads are trimmed of adapters to yield subreads

Consensus is called from subreads

HiFi READ
(>99% accuracy)





Warning

Keeping epigenetics information or not must be decided **prior** to the run!

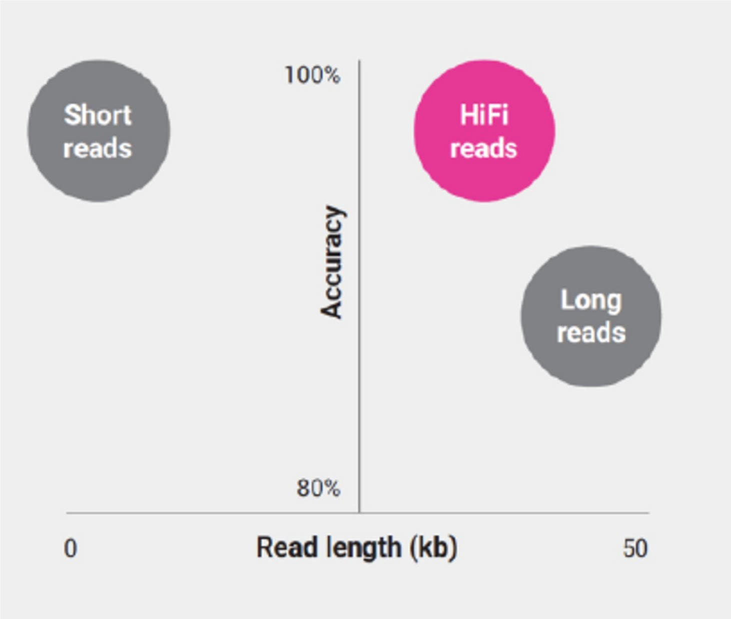
PacBio

- ❑ Generates ~**2.2-2.4 million HIFI reads** / 8M SMRTCell
- ❑ HiFi reads have **99.9%** accuracy*
- ❑ HiFi reads can reach between **18-25** kb*
- ❑ Movie times of **10-30h** → depends on library size

Warnings

* The longer the less accurate!

1514€ /SMRTCell w/o library preparation



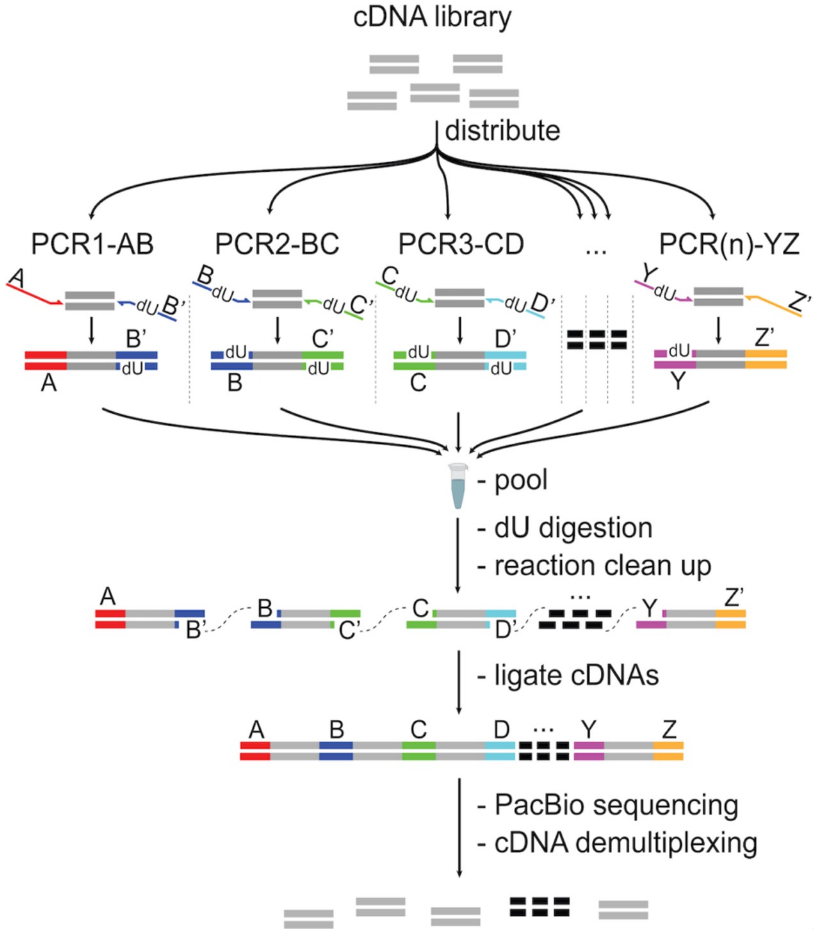
PacBio MAS-Seq (Multiplexed Arrays Sequencing)

10X Chromium Next GEM Single Cell 3’ kit v(3.1)

3,000 -10,000 cell

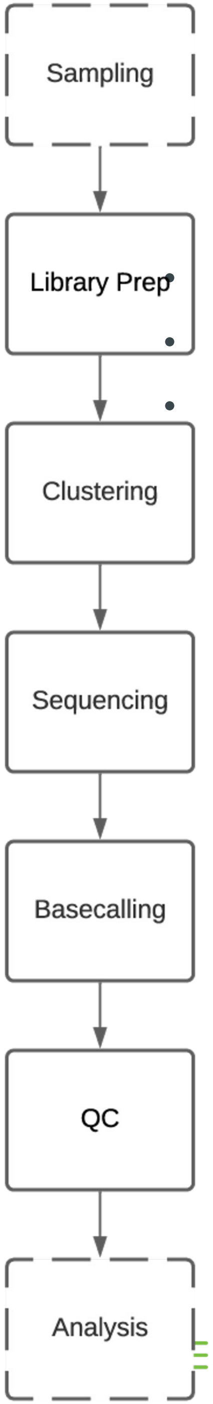
10X workflow for **cDNA** generation

15-75 ng of cDNA as input



- pool
- dU digestion
- reaction clean up
- ligate cDNAs
- PacBio sequencing
- cDNA demultiplexing

Preprint: High-throughput RNA isoform sequencing using programmable cDNA concatenation.
Aziz M. Al’Khafaji et al. 2021



Illumina Sequencing

Short reads ~ 30 - 300 bases

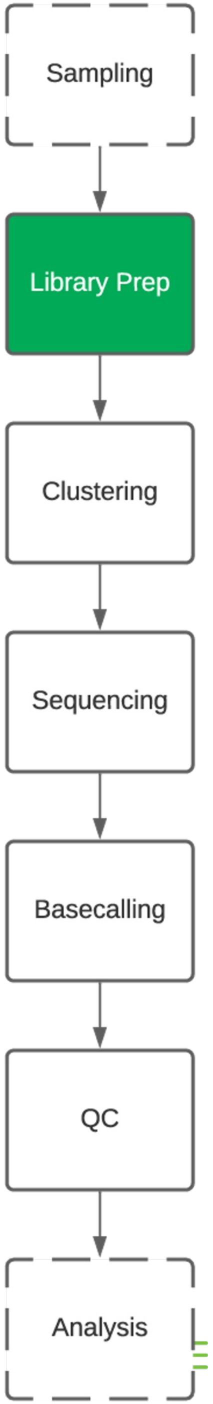
Random error, mostly mismatches

Usually quite good quality 99.9%

A lot of data produced

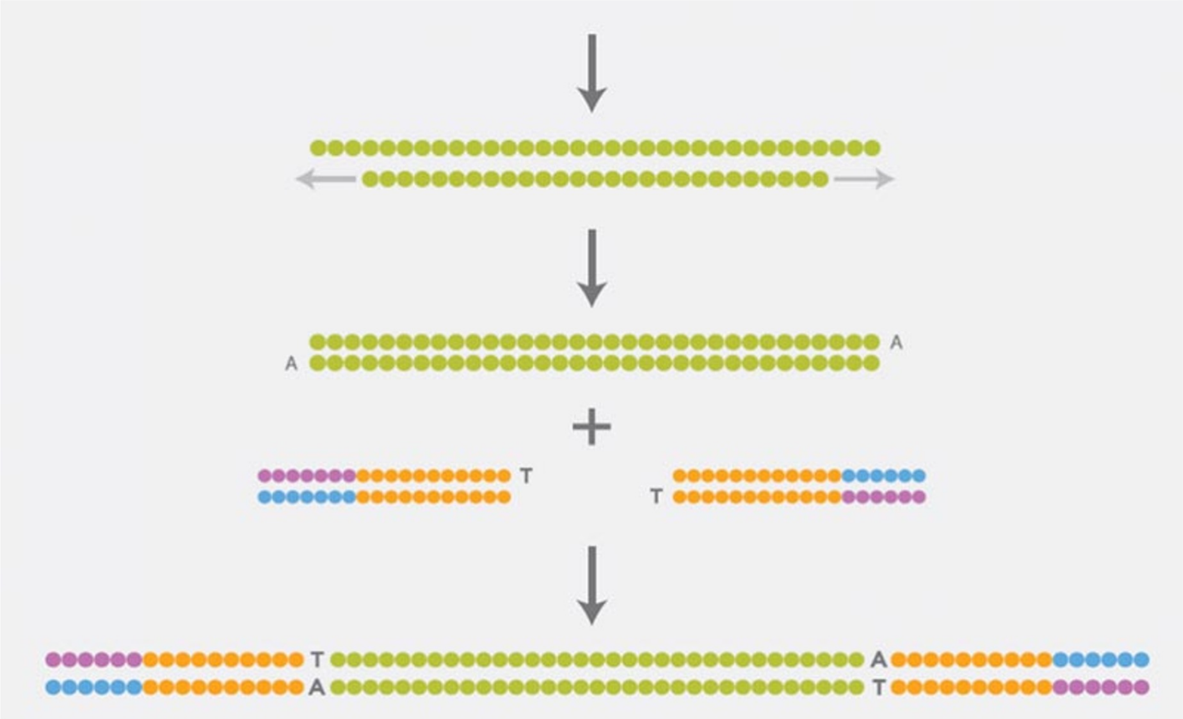
“Affordable”

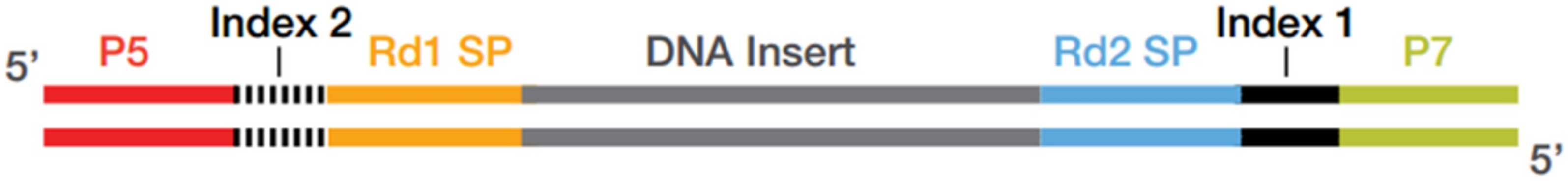
©2012 Illumina, Inc. All rights reserved.

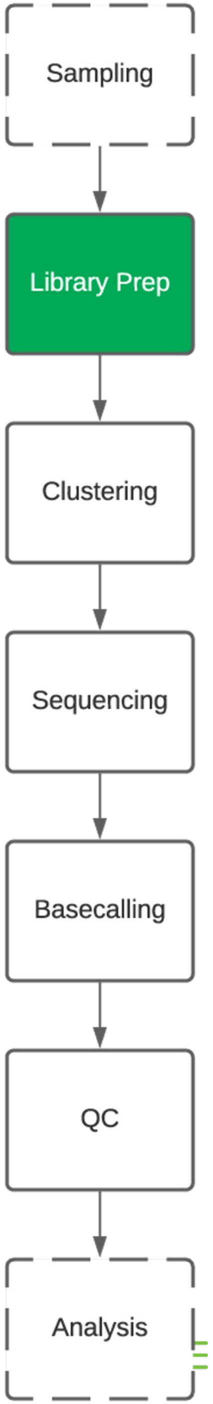


Illumina Sequencing - Library prep

- Hundreds of methods to select the desired molecular landscape
- Adapters necessary
- Barcodes to differentiate individual samples







Sampling

Library Prep

Clustering

Sequencing

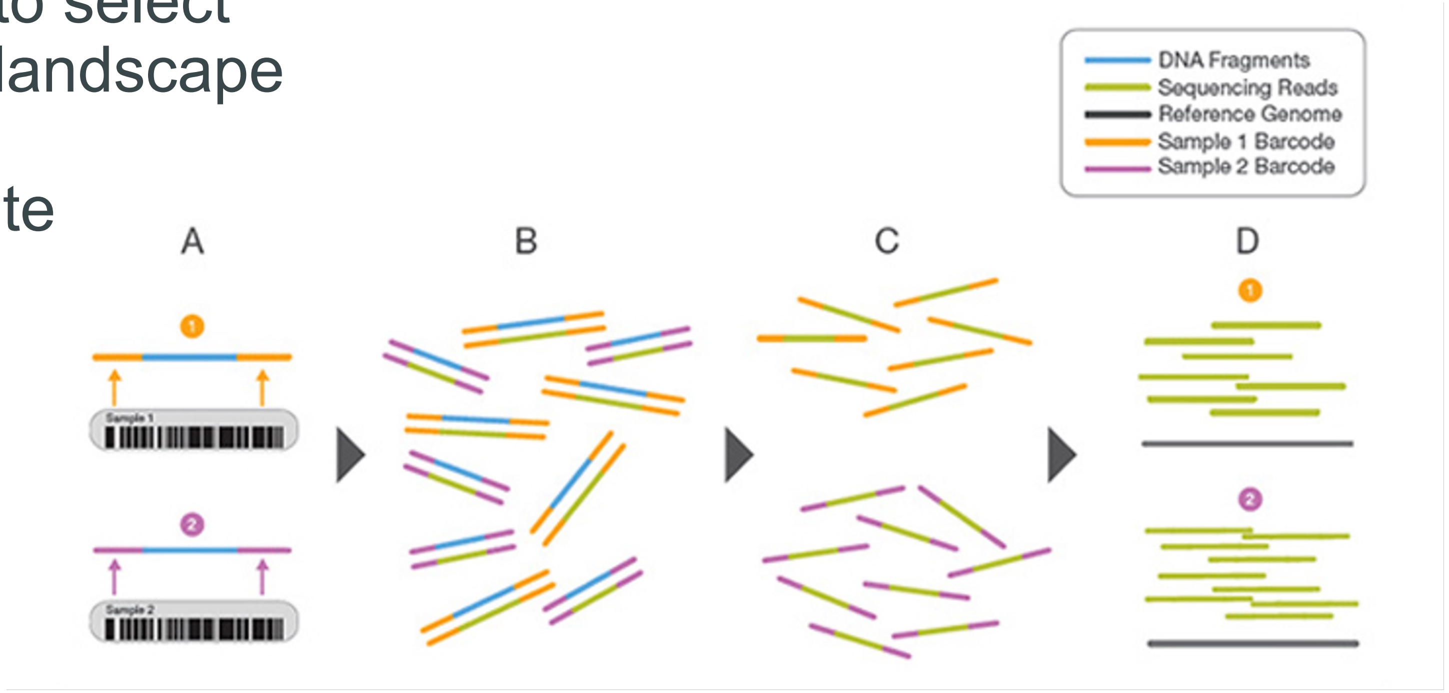
Basecalling

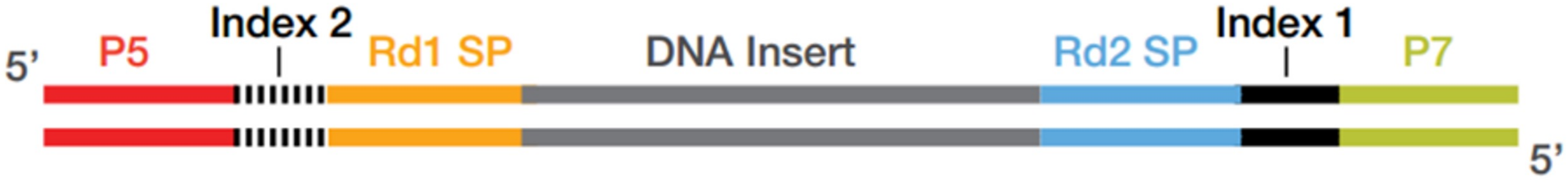
QC

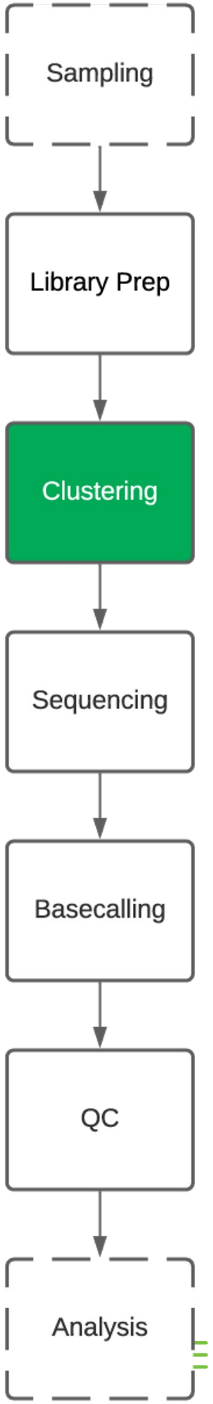
Analysis

Illumina Sequencing - Library prep

- Hundreds of methods to select the desired molecular landscape
- Adapters necessary
- Barcodes to differentiate individual samples






Sampling

Library Prep

Clustering

Sequencing

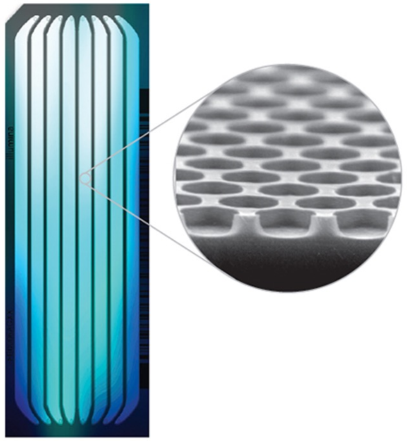
Basecalling

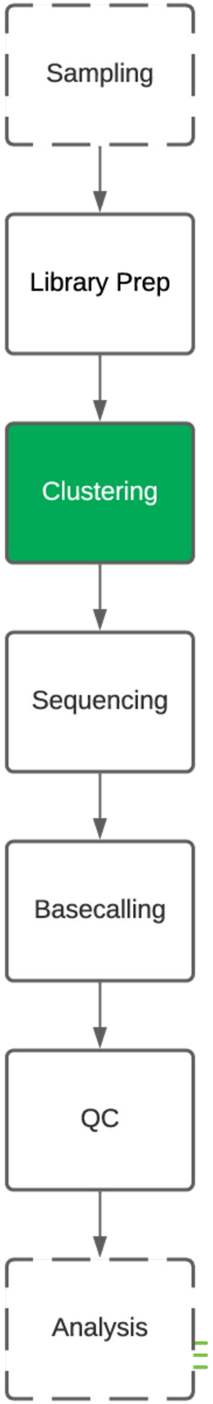
QC

Analysis

Illumina Sequencing - Clustering

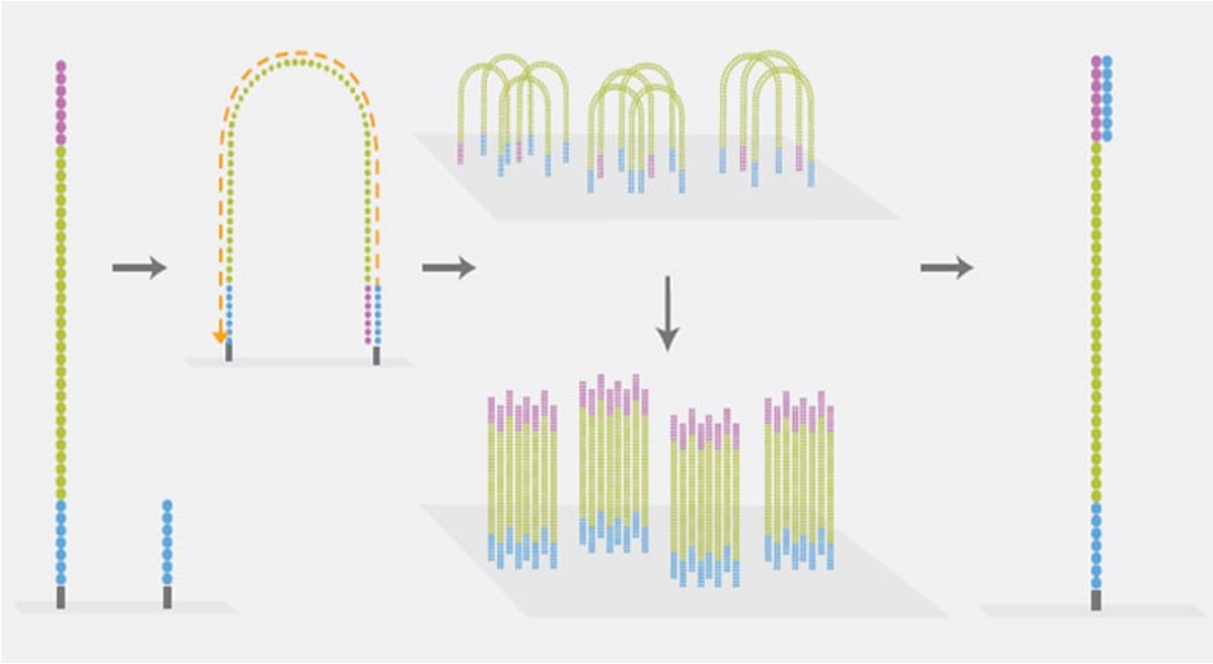
- Signal from a single DNA molecule is not enough to be detected

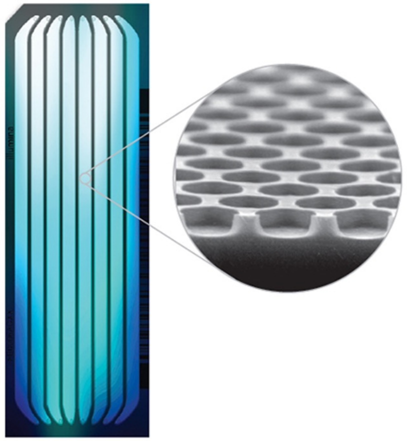




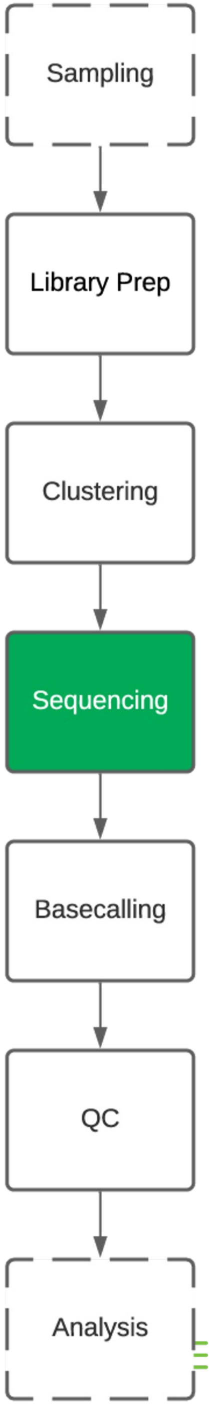
Illumina Sequencing - Clustering

- Signal from a single DNA molecule is not enough to be detected





Illumina Sequencing - Sequencing


Sampling

Library Prep

Clustering

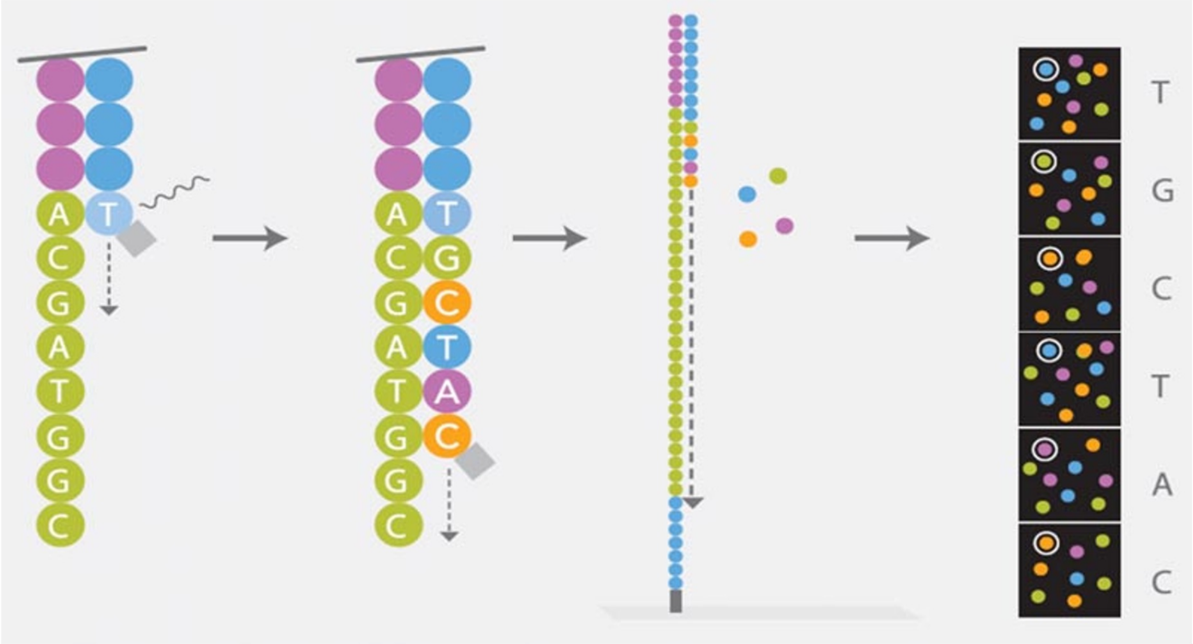
Sequencing

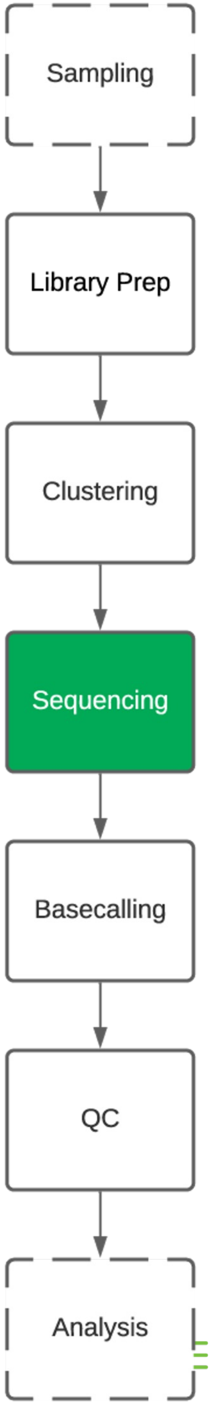
Basecalling

QC

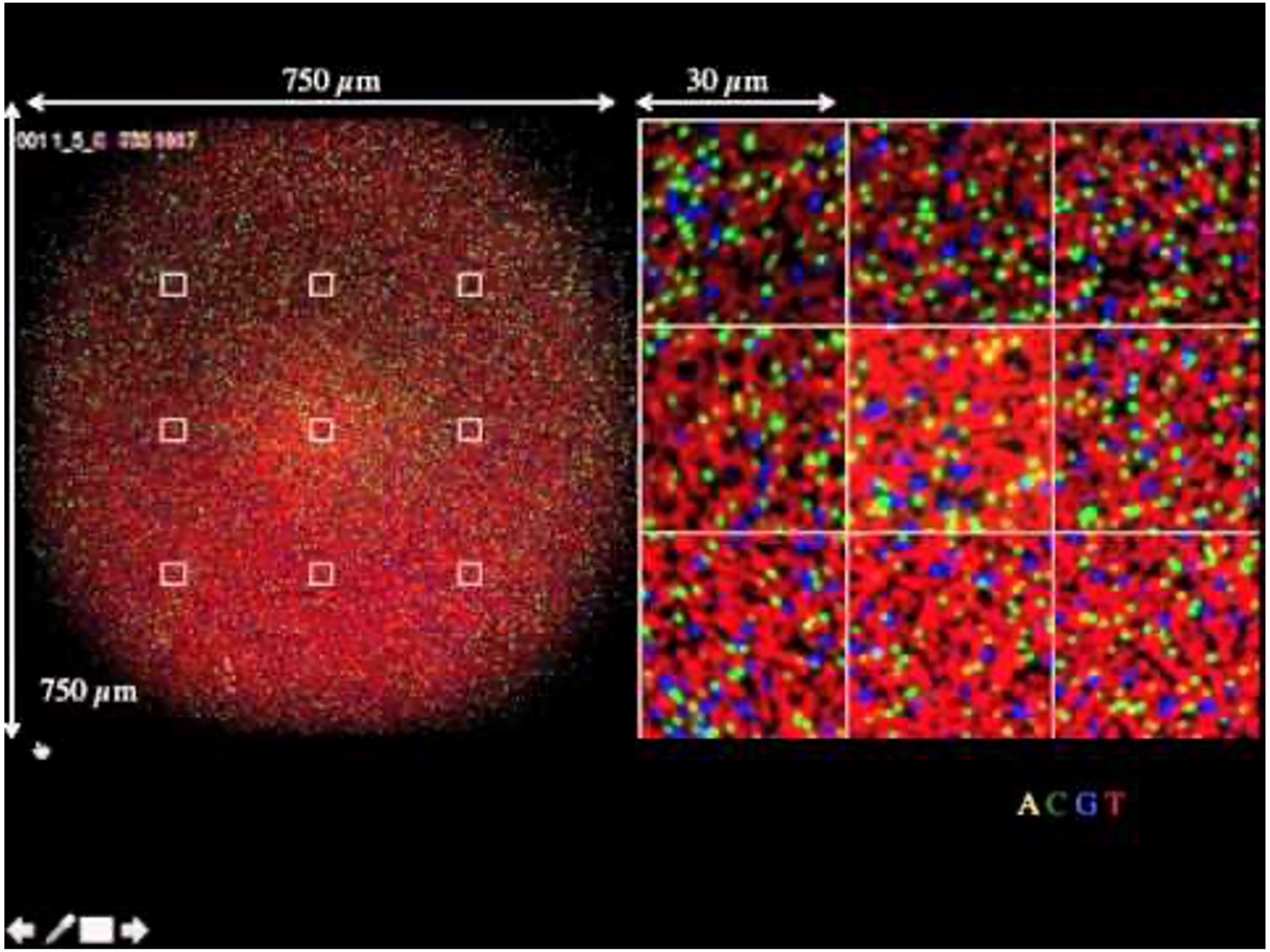
Analysis

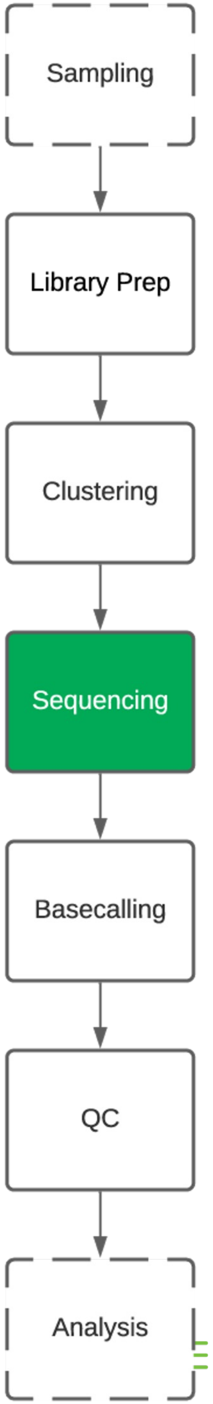
- Sequencing by synthesis
- Each cycle - 1 nucleotide read





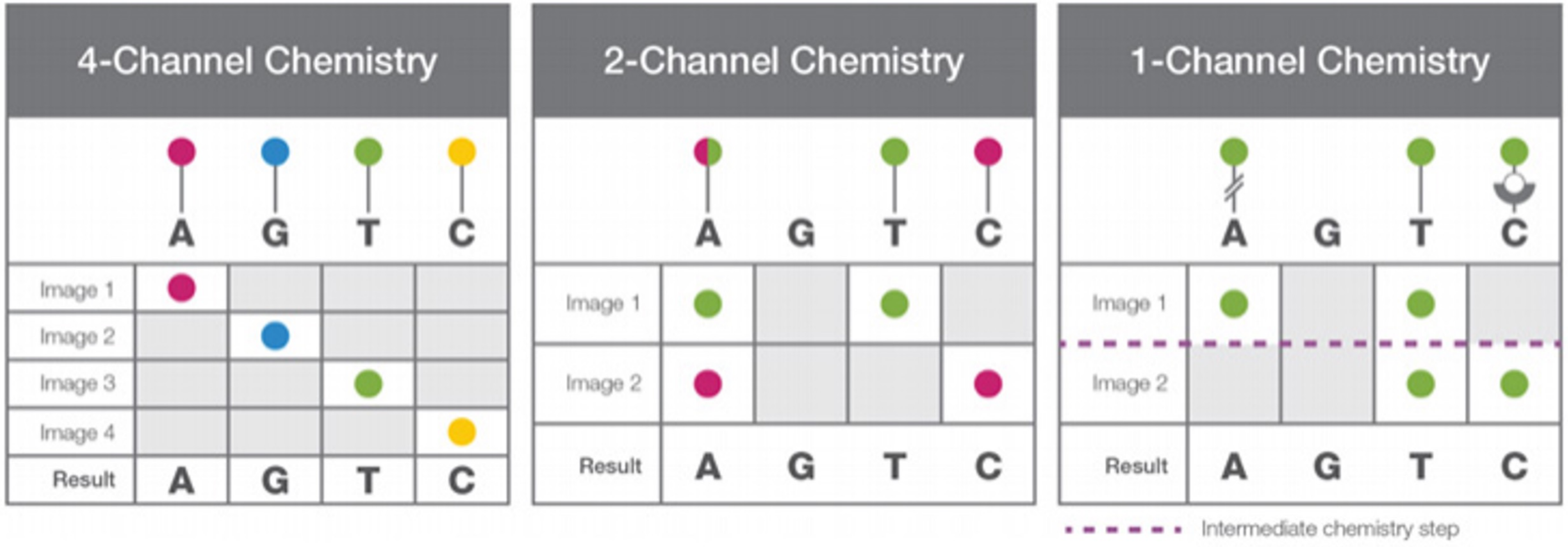
Illumina Sequencing - Sequencing





Illumina Sequencing - Sequencing

- Sequencing by synthesis
- Each cycle - 1 nucleotide read
- Readout is machine dependent
- Different error profiles



4-Channel Chemistry

	A	G	T	C
Image 1	●			
Image 2		●		
Image 3			●	
Image 4				●
Result	A	G	T	C

2-Channel Chemistry

	A	G	T	C
Image 1	●		●	
Image 2	●			●
Result	A	G	T	C

1-Channel Chemistry

	A	G	T	C
Image 1	●		●	
Image 2			●	●
Result	A	G	T	C

Intermediate chemistry step

NGS sequencing technologies

Currently provided sequencing technologies:

Illumina: NovaSeq, NextSeq 500, MiSeq
PacBio: Sequel IIe
Oxford Nanopore: GridION, PromethION P2 Solo

NGS sequencing technologies

Currently provided sequencing technologies:

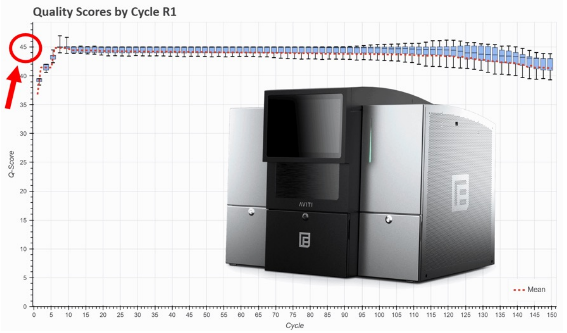
Illumina: NovaSeq, NextSeq 500, MiSeq
~~PacBio: Sequel IIe~~
~~Oxford Nanopore: GridION, PromethION P2 Solo~~

NGS sequencing technologies

Currently provided sequencing technologies:

Illumina: NovaSeq, NextSeq 500, MiSeq
~~PacBio: Sequel IIe~~
~~Oxford Nanopore: GridION, PromethION P2 Solo~~

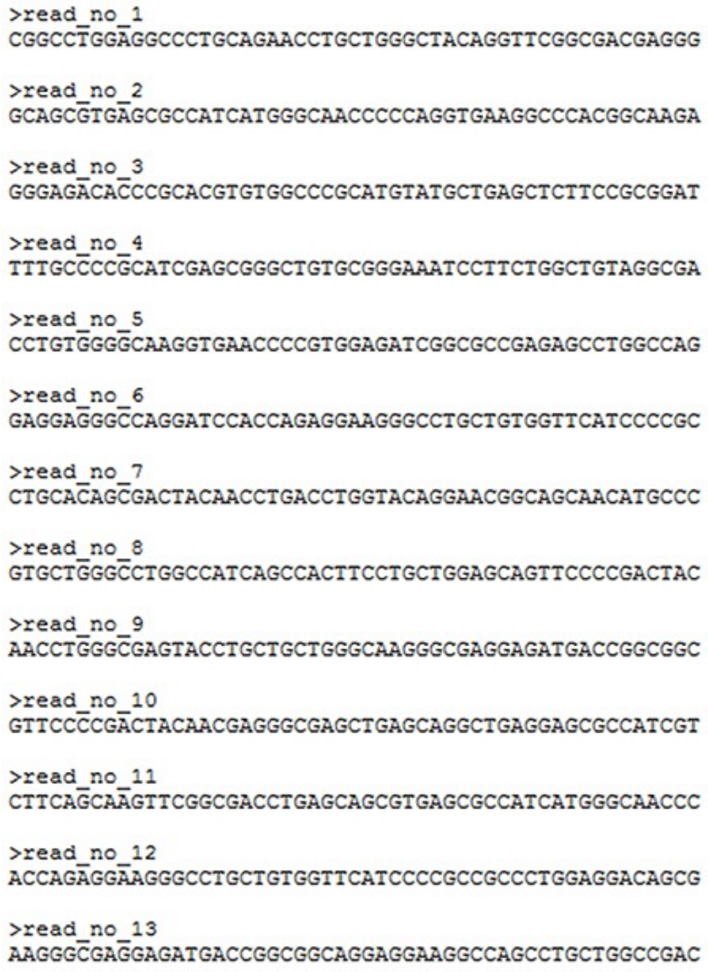
Element Biosciences



MGI

SINGULAR GENOMICS

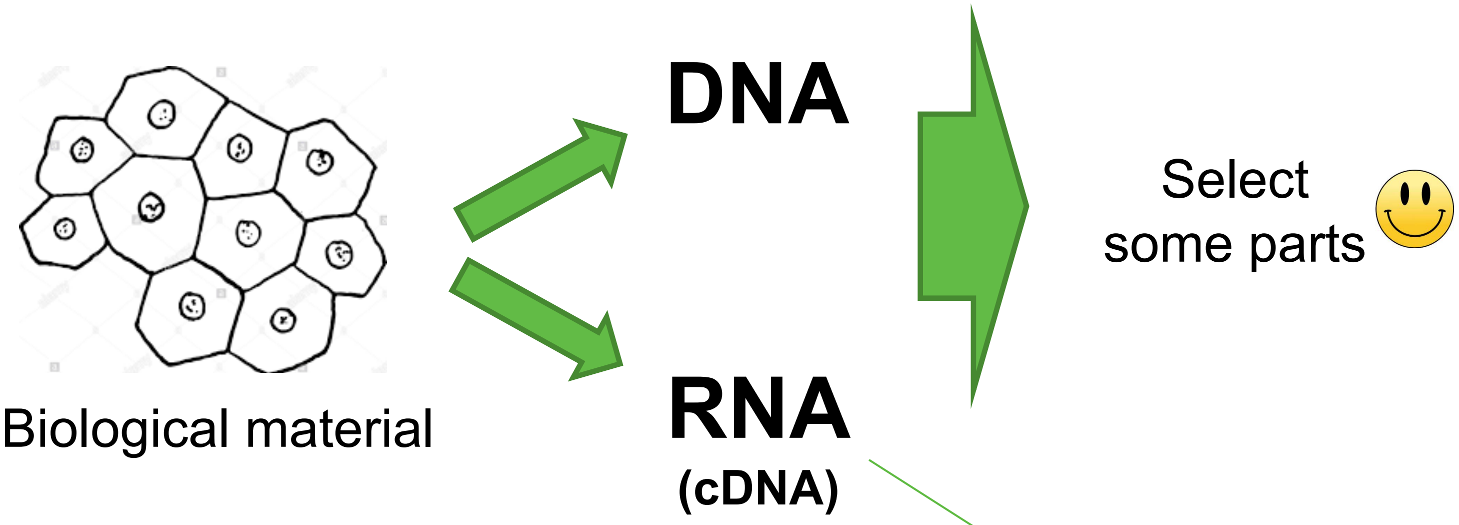
Short-read sequencing result



```
>read_no_1
CGGCCTGGAGGCCCTGCAGAACCTGCTGGGCTACAGGTTCGGCGACGAGGG

>read_no_2
GCAGCGTGAGCGCCATCATGGGCAACCCCCAGGTGAAGGCCCACGGCAAGA

>read_no_3
GGGAGACACCCGCACGTGTGGCCCGCATGTATGCTGAGCTCTTCCGCGGAT

>read_no_4
TTTGCCCCGCATCGAGCGGGCTGTGCGGGAAATCCTTCTGGCTGTAGGCGA

>read_no_5
CCTGTGGGGCAAGGTGAACCCCGTGGAGATCGGCGCCGAGAGCCTGGCCAG

>read_no_6
GAGGAGGGCCAGGATCCACCAGAGGAAGGGCCTGCTGTGGTTCATCCCCGC

>read_no_7
CTGCACAGCGACTACAACCTGACCTGGTACAGGAACGGCAGCAACATGCCC

>read_no_8
GTGCTGGGCCTGGCCATCAGCCACTTCCTGCTGGAGCAGTTCCCCGACTAC

>read_no_9
AACCTGGGCGAGTACCTGCTGCTGGGCAAGGGCGAGGAGATGACCGGCGGC

>read_no_10
GTTCCCCGACTACAACGAGGGCGAGCTGAGCAGGCTGAGGAGCGCCATCGT

>read_no_11
CTTCAGCAAGTTCGGCGACCTGAGCAGCGTGAGCGCCATCATGGGCAACCC

>read_no_12
ACCAGAGGAAGGGCCTGCTGTGGTTCATCCCCGCCGCCCTGGAGGACAGCG

>read_no_13
AAGGGCGAGGAGATGACCGGCGGCAGGAGGAAGGCCAGCCTGCTGGCCGAC
```

- 10^5 – 10^10 reads
- 75 – 300Bp
- Could be pair-end

NGS library preparation - What we sequence



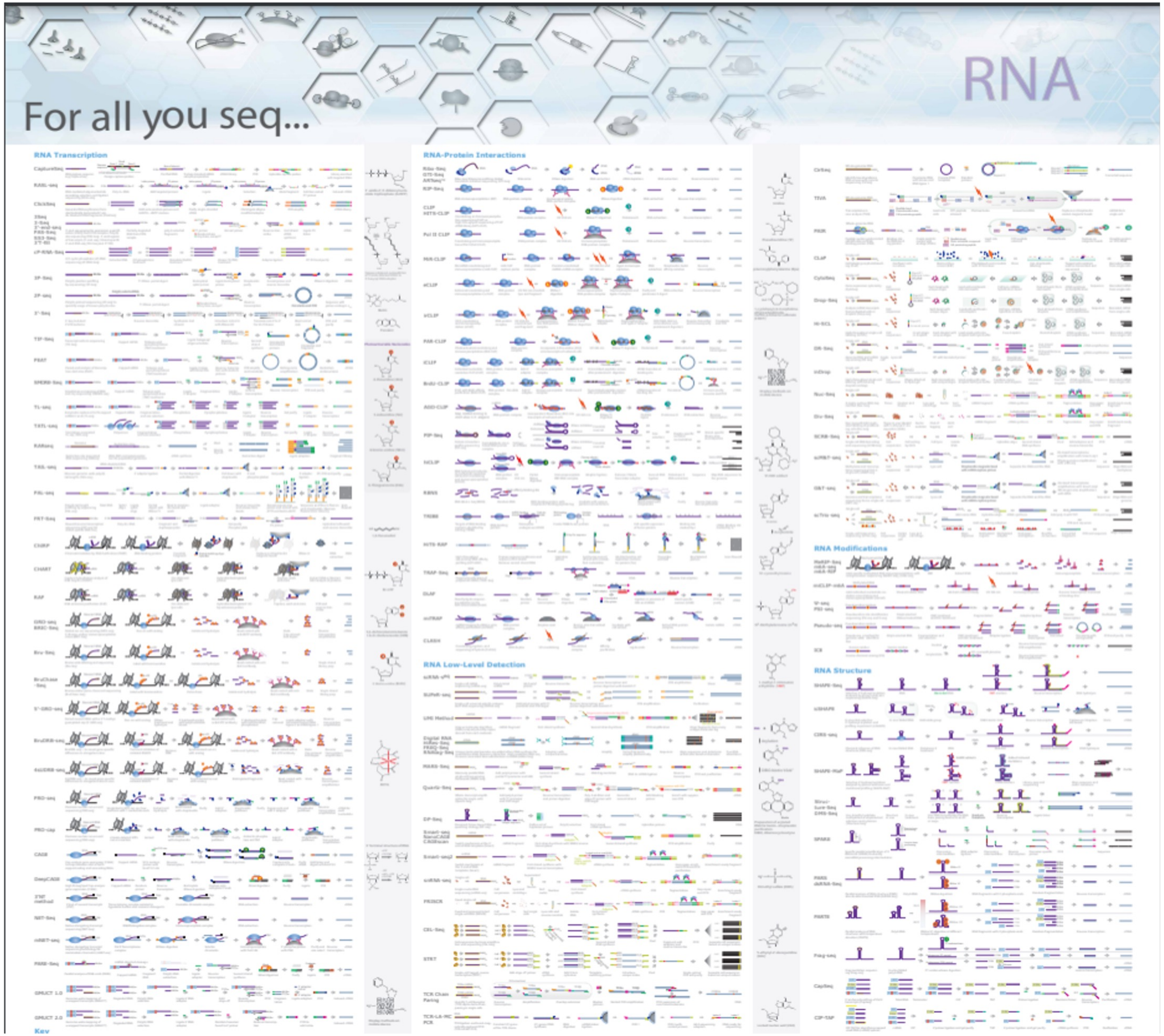
Biological material

DNA

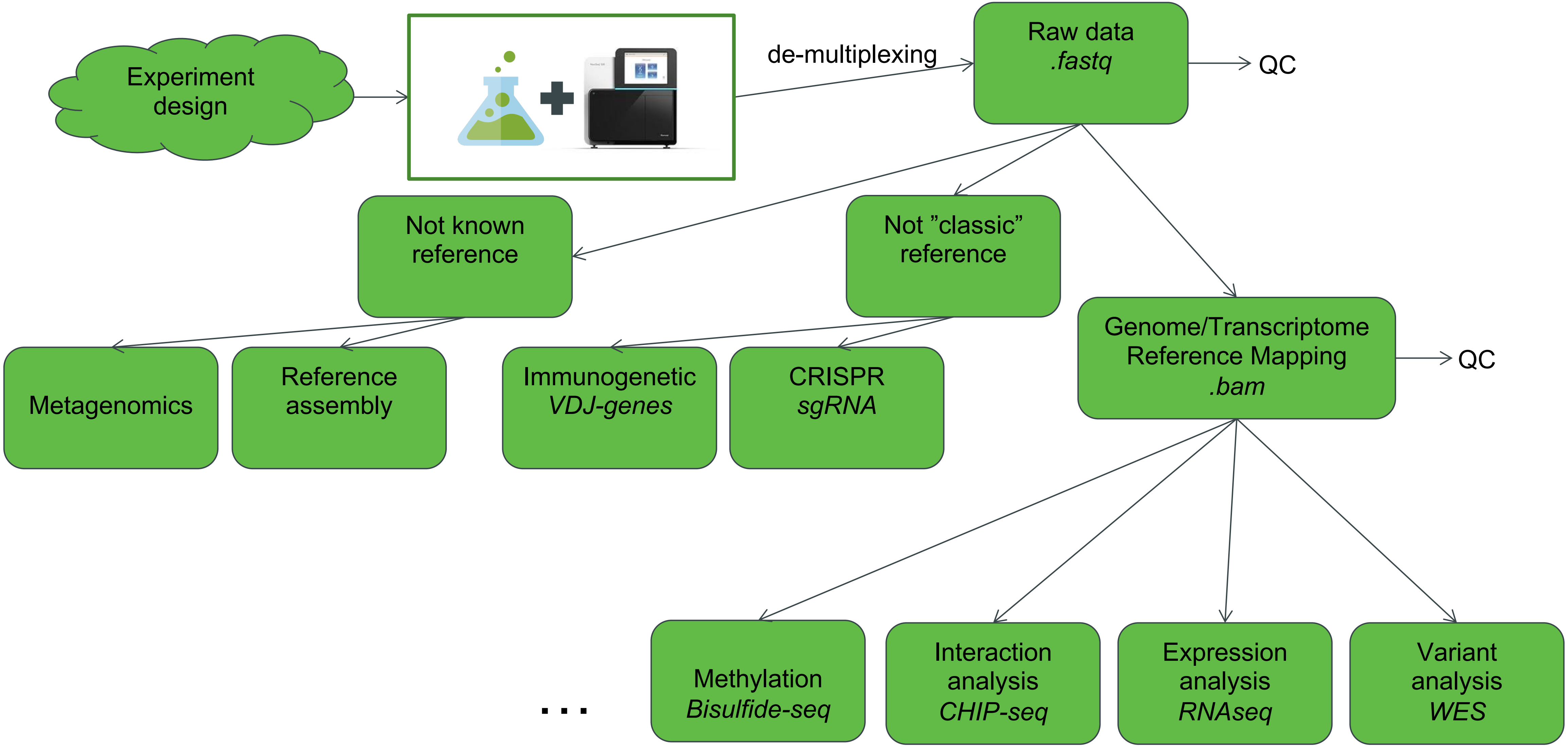
RNA
(cDNA)

Select
some parts 🙂

Note on a direct RNA sequencing using Oxford nanopore


For all you seq...
RNA
RNA Transcription
RNA-Protein Interactions
RNA Low-Level Detection
RNA Modifications
RNA Structure
Key

NGS data analysis



Experiment design

de-multiplexing

Raw data *.fastq*

QC

Not known reference

Not ”classic” reference

Genome/Transcriptome Reference Mapping *.bam*

QC

Metagenomics

Reference assembly

Immunogenetic *VDJ-genes*

CRISPR *sgRNA*

...

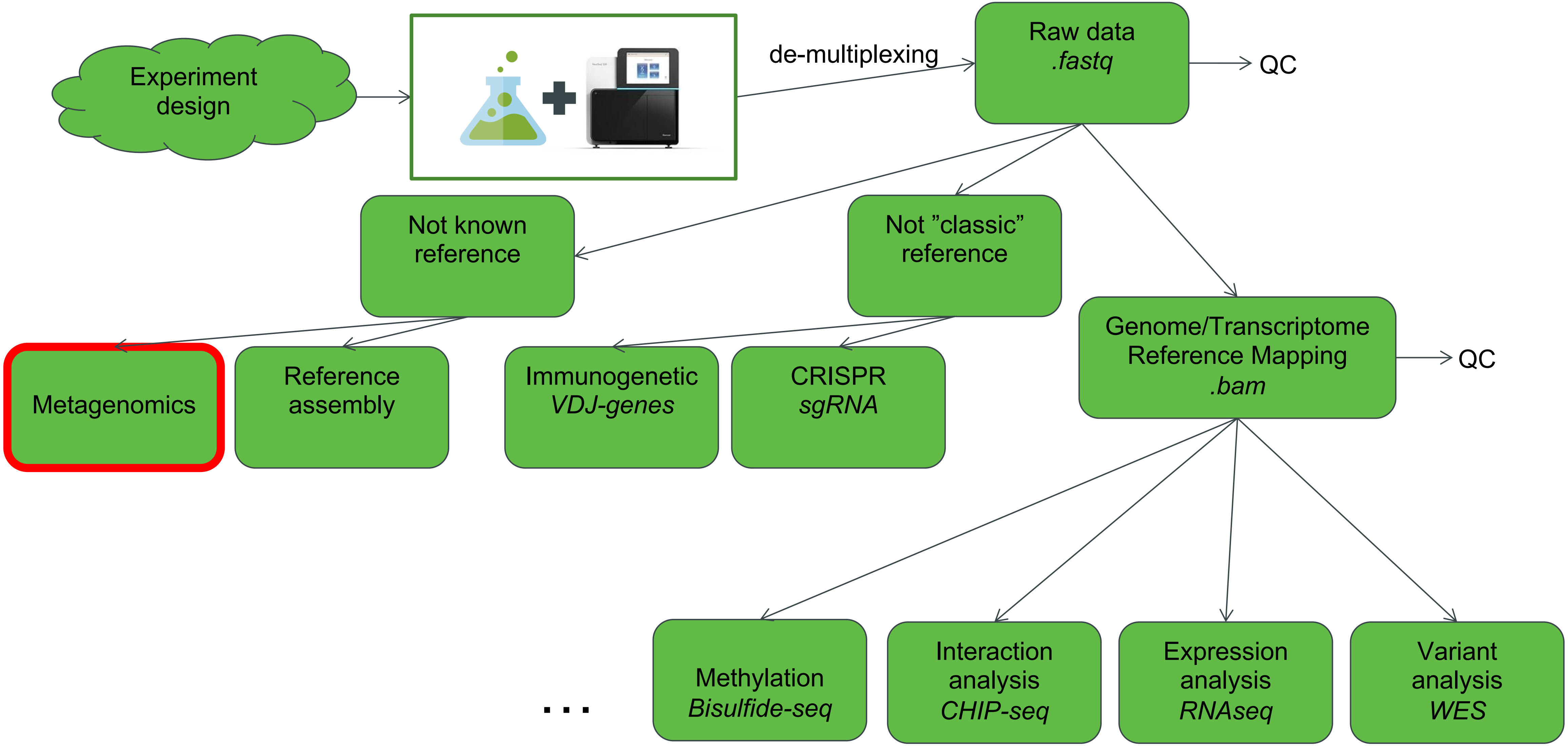
Methylation *Bisulfide-seq*

Interaction analysis *CHIP-seq*

Expression analysis *RNAseq*

Variant analysis *WES*

NGS data analysis



Experiment design

de-multiplexing

Raw data
.fastq

QC

Not known reference

Not ”classic” reference

Genome/Transcriptome Reference Mapping
.bam

QC

Metagenomics

Reference assembly

Immunogenetic
VDJ-genes

CRISPR
sgRNA

. . .

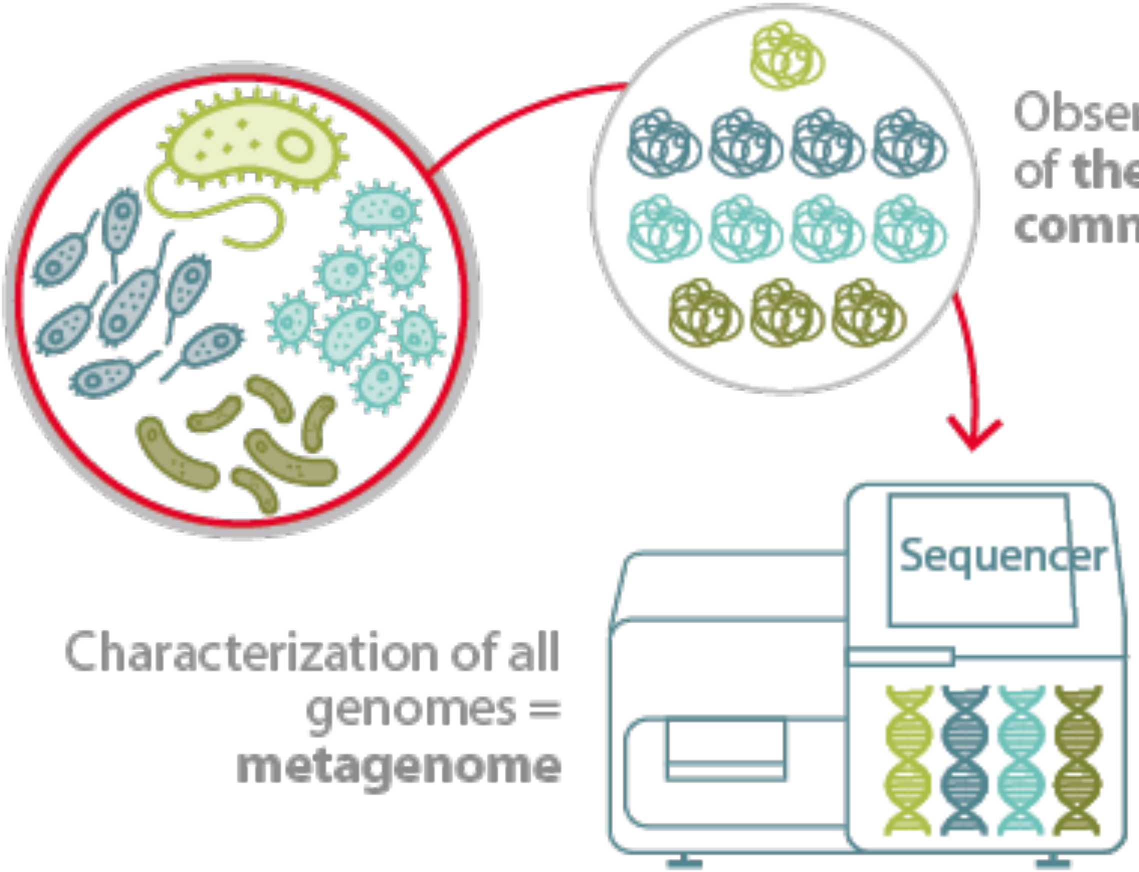
Methylation
Bisulfide-seq

Interaction analysis
CHIP-seq

Expression analysis
RNAseq

Variant analysis
WES

Metagenomics

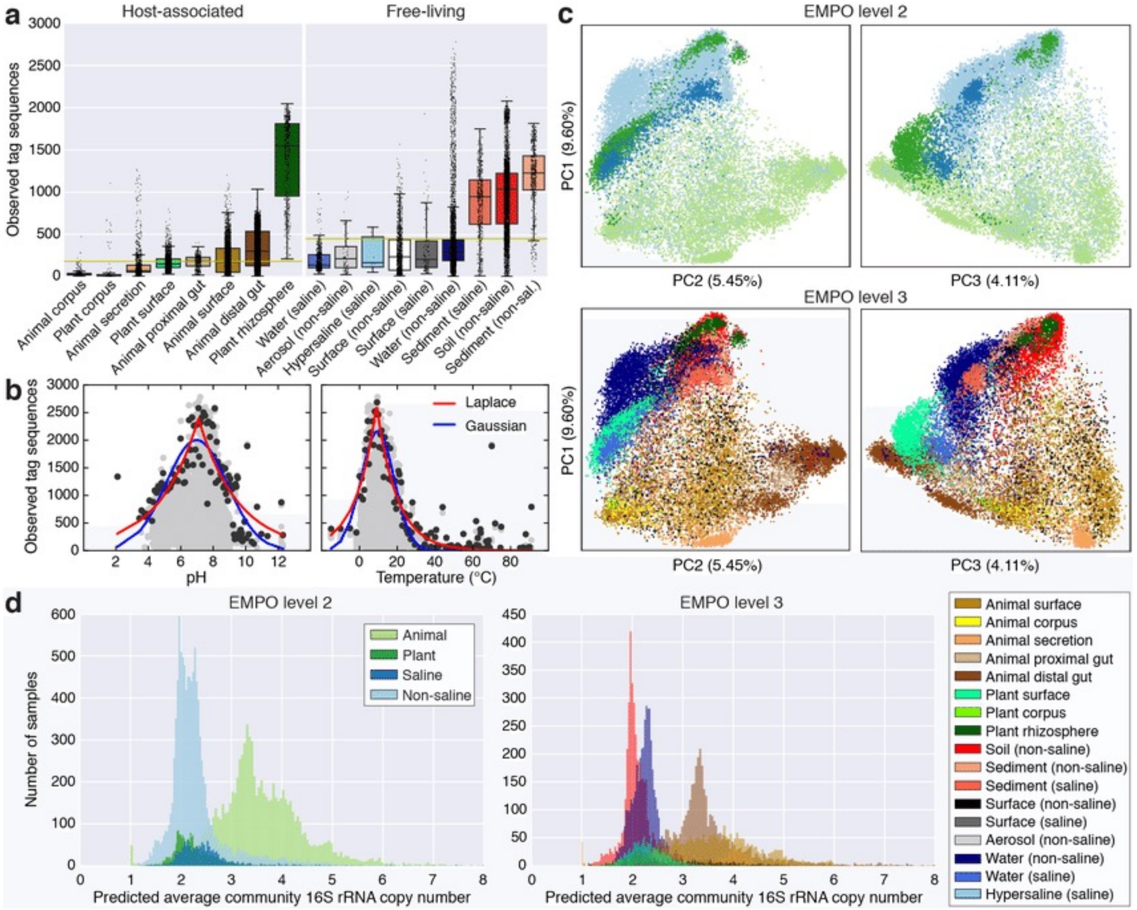


Observation of **the whole community**

Characterization of all genomes = **metagenome**

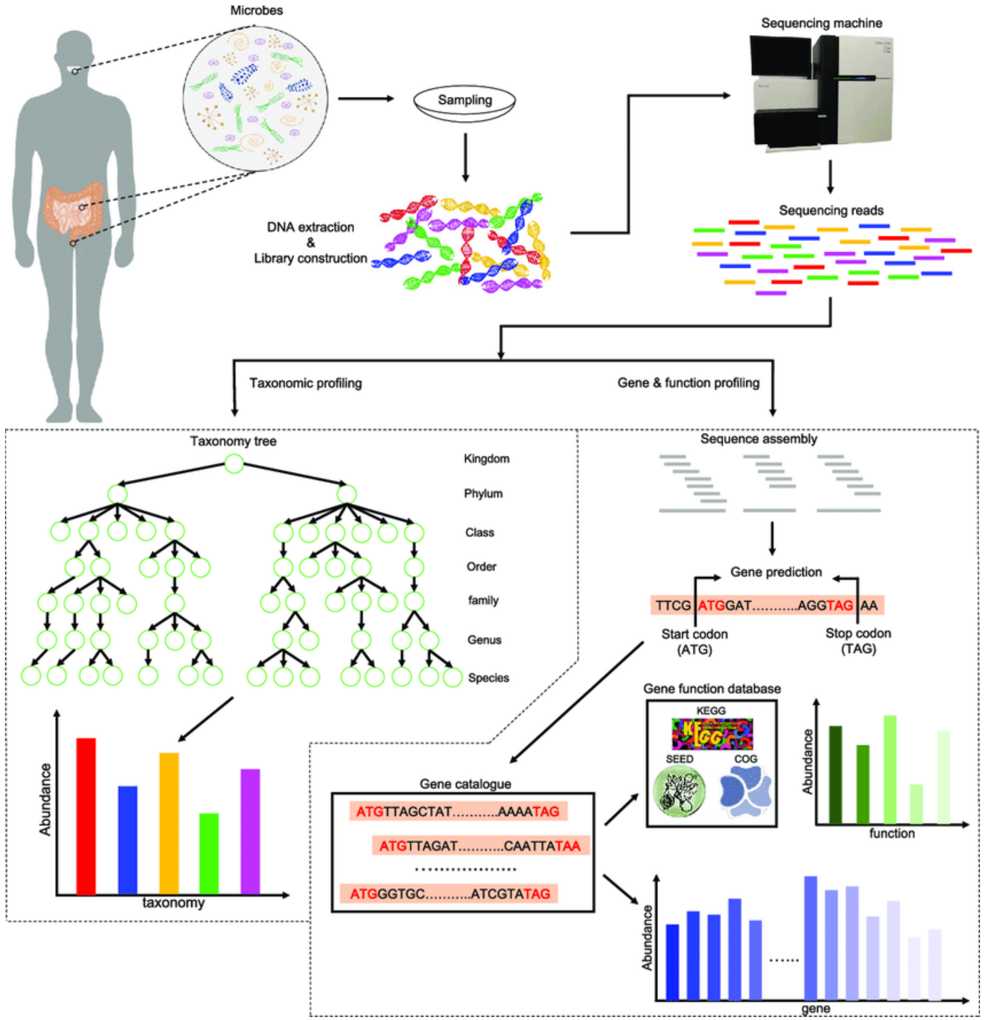
Metagenomics results

- Environmental statistics about populations
  - alpha, beta, gamma diversity



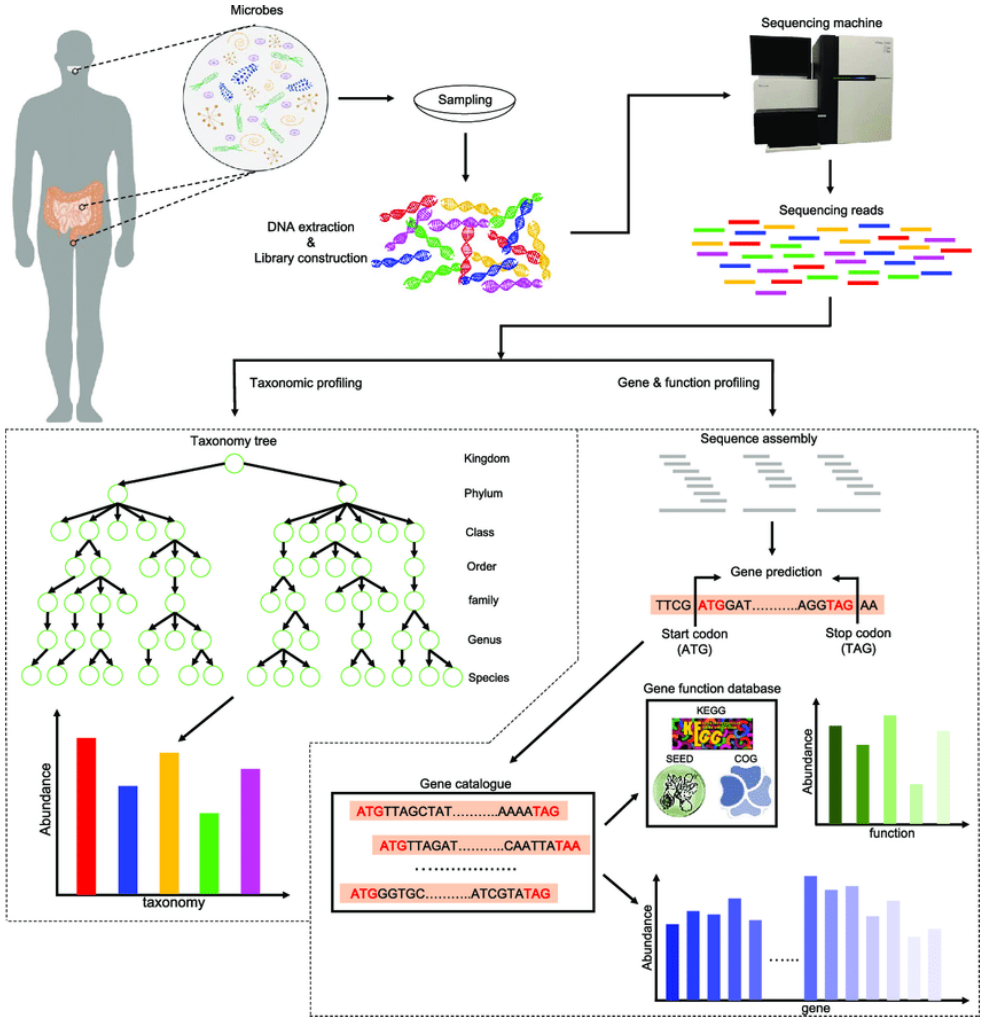
Metagenomics results

- Environmental statistics about populations
  - identify known bacterial species
    - taxonomy profiling
  - eventually functional profiling
    - E.g. antimicrobial resistance genes



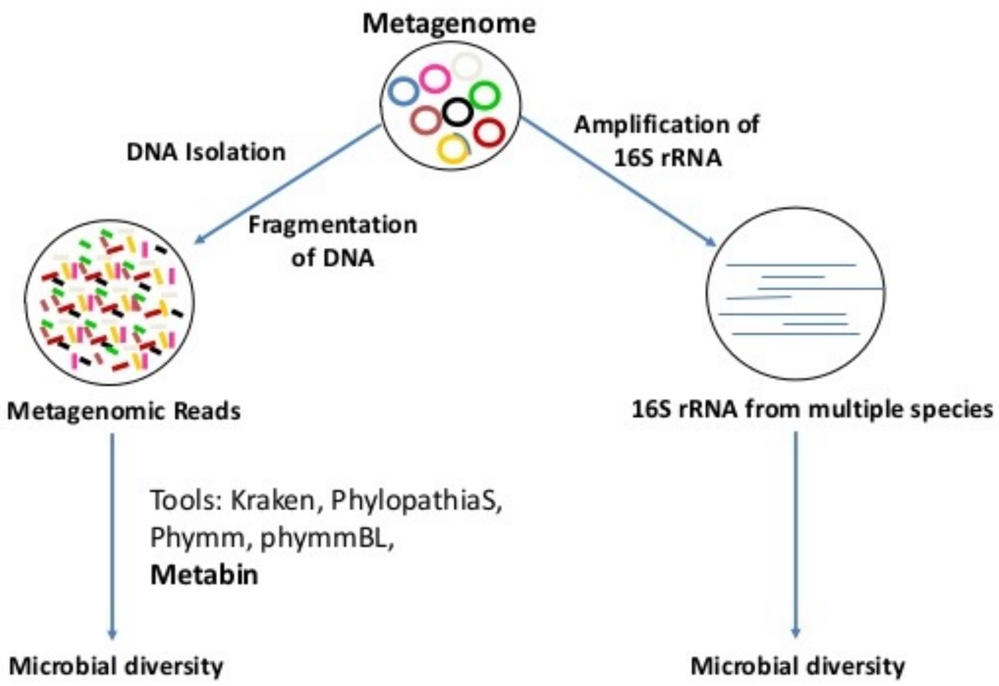
Metagenomics results

- Environmental statistics about populations
  - identify known bacterial species
    - taxonomy profiling
  - eventually functional profiling
    - E.g. antimicrobial resistance genes
- Sequencing techniques
  - 16S rRNA sequencing
  - Shotgun metagenomic sequencing



Metagenomics – 16S rRNA vs. Shotgun

Metagenomic reads vs 16S rRNA for microbial diversity identification


Metagenome
DNA Isolation
Fragmentation of DNA
Amplification of 16S rRNA
Metagenomic Reads
16S rRNA from multiple species
Tools: Kraken, PhylopathiaS, Phymm, phymmBL, Metabin
Microbial diversity
Microbial diversity

Factors	16S rRNA sequencing	Shotgun Metagenomic Sequencing
Cost	~$50 USD	Starting at ~$150 but price will depend on sequencing depth required
Sample preparation	Similar complexity to shotgun sequencing	Similar complexity to 16S rRNA sequencing
Functional profiling (profile microbial genes)	No (but ‘predicted’ functional profiling is possible)	Yes (but it only reveals information on functional potential)
Taxonomic resolution: Genus, species, strain?	Bacterial genus (sometimes species); dependent on region(s) targeted	Bacterial species (sometimes strains and single nucleotide variants, if sequencing is deep enough)
Taxonomic coverage	Bacteria and archaea	All taxa, including viruses
Bioinformatics requirements	Beginner to intermediate expertise	Intermediate to advanced expertise
Databases	Established, well-curated	Relatively new, still growing
Sensitivity to host DNA contamination	Low (but PCR success depends on the absence of inhibitors and the presence of a detectable microbiome)	High , varies with sample type (but this can be mitigated by calibrating the sequencing depth)
Bias	Medium to high (retrieved taxonomic composition is dependent on selected primers and targeted variable region)	Lower (while metagenomics is “untargeted”, experimental and analytical biases can be introduced at various stages)

Metagenomics – 16S rRNA vs. Shotgun

- Study Examples
  - **Assessment of the bacterial microbiome of Amazonian soil**

Metagenomics – 16S rRNA vs. Shotgun

- Study Examples
  - **Assessment of the bacterial microbiome of Amazonian soil**
    - 16S rRNA sequencing may provide more taxonomic resolution

Metagenomics – 16S rRNA vs. Shotgun

- Study Examples
  - **Assessment of the bacterial microbiome of Amazonian soil**
    - 16S rRNA sequencing may provide more taxonomic resolution
  - **Changes in microbiome composition and antimicrobial gene carriage following fecal transplant**

Metagenomics – 16S rRNA vs. Shotgun

- Study Examples
  - **Assessment of the bacterial microbiome of Amazonian soil**
    - 16S rRNA sequencing may provide more taxonomic resolution
  - **Changes in microbiome composition and antimicrobial gene carriage following fecal transplant**
    - shotgun sequencing to assess both compositional and functional differences

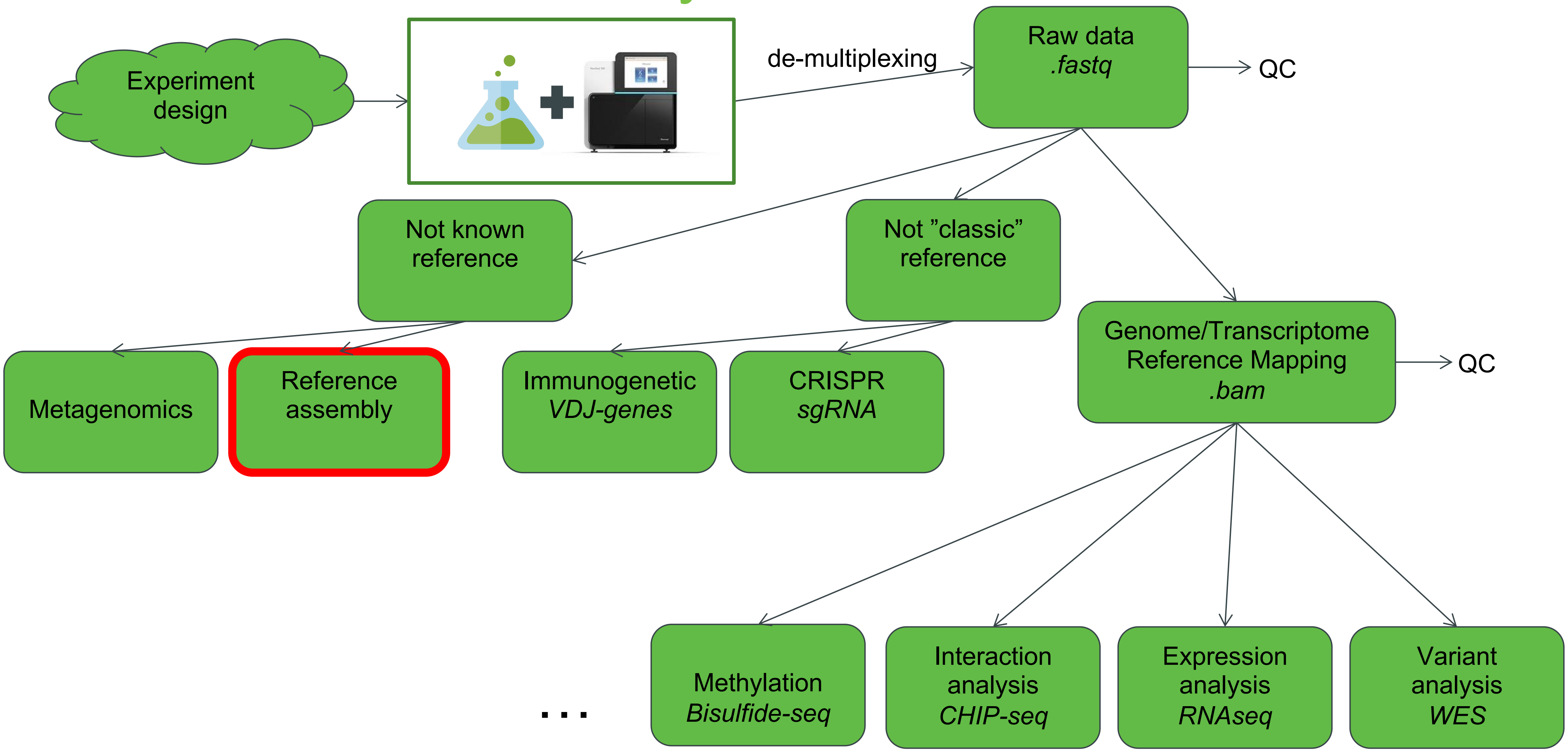
Metagenomics – 16S rRNA vs. Shotgun

- Study Examples
  - **Assessment of the bacterial microbiome of Amazonian soil**
    - 16S rRNA sequencing may provide more taxonomic resolution
  - **Changes in microbiome composition and antimicrobial gene carriage following fecal transplant**
    - shotgun sequencing to assess both compositional and functional differences
  - **Daily fluctuations in gut microbiome following 2 week dietary fiber intervention**

Metagenomics – 16S rRNA vs. Shotgun

- Study Examples
  - **Assessment of the bacterial microbiome of Amazonian soil**
    - 16S rRNA sequencing may provide more taxonomic resolution
  - **Changes in microbiome composition and antimicrobial gene carriage following fecal transplant**
    - shotgun sequencing to assess both compositional and functional differences
  - **Daily fluctuations in gut microbiome following 2 week dietary fiber intervention**
    - shotgun sequencing or 16S rRNA
      - assess both compositional and functional differences
      - cheaper and in this case can use ‘predicted’ functional profiling

NGS data analysis



Experiment design

de-multiplexing

Raw data
.fastq

QC

Not known reference

Not ”classic” reference

Genome/Transcriptome Reference Mapping
.bam

QC

Metagenomics

Reference assembly

Immunogenetic
VDJ-genes

CRISPR
sgRNA

. . .

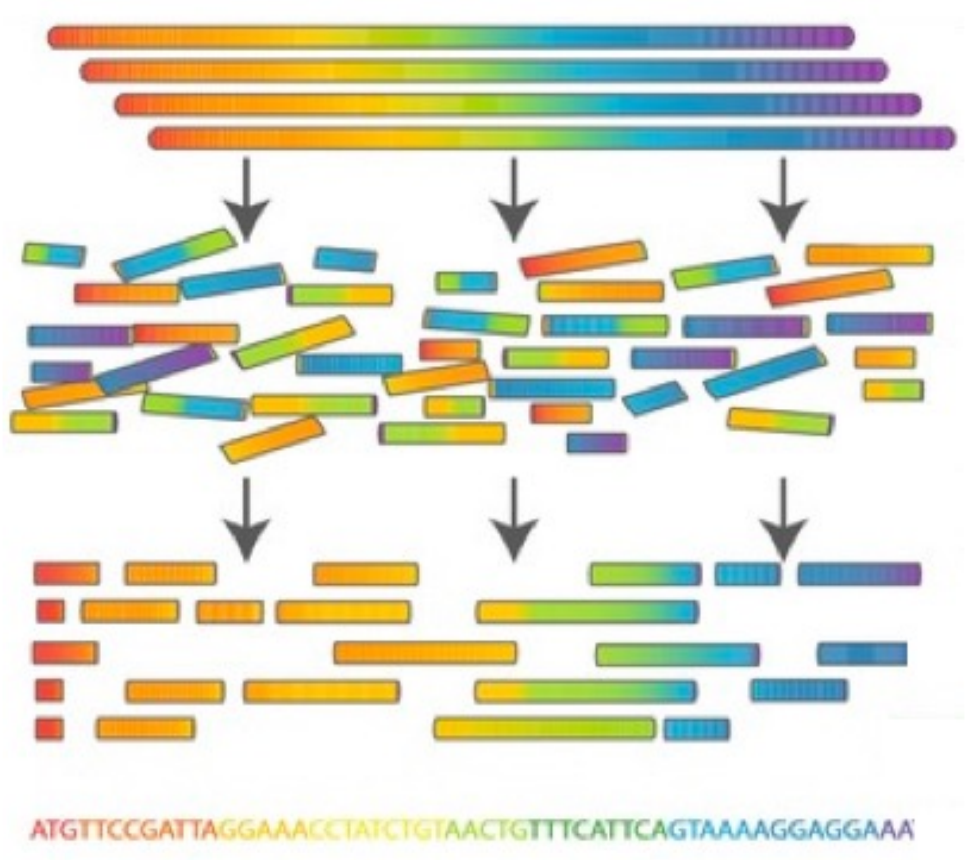
Methylation
Bisulfide-seq

Interaction analysis
CHIP-seq

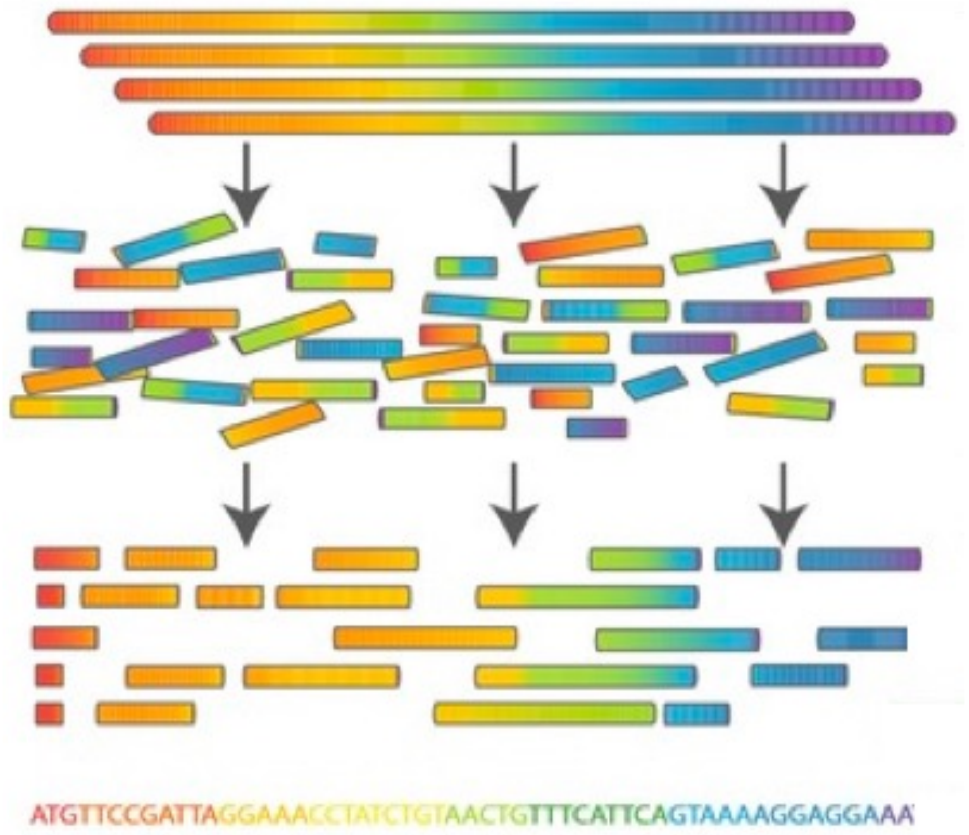
Expression analysis
RNAseq

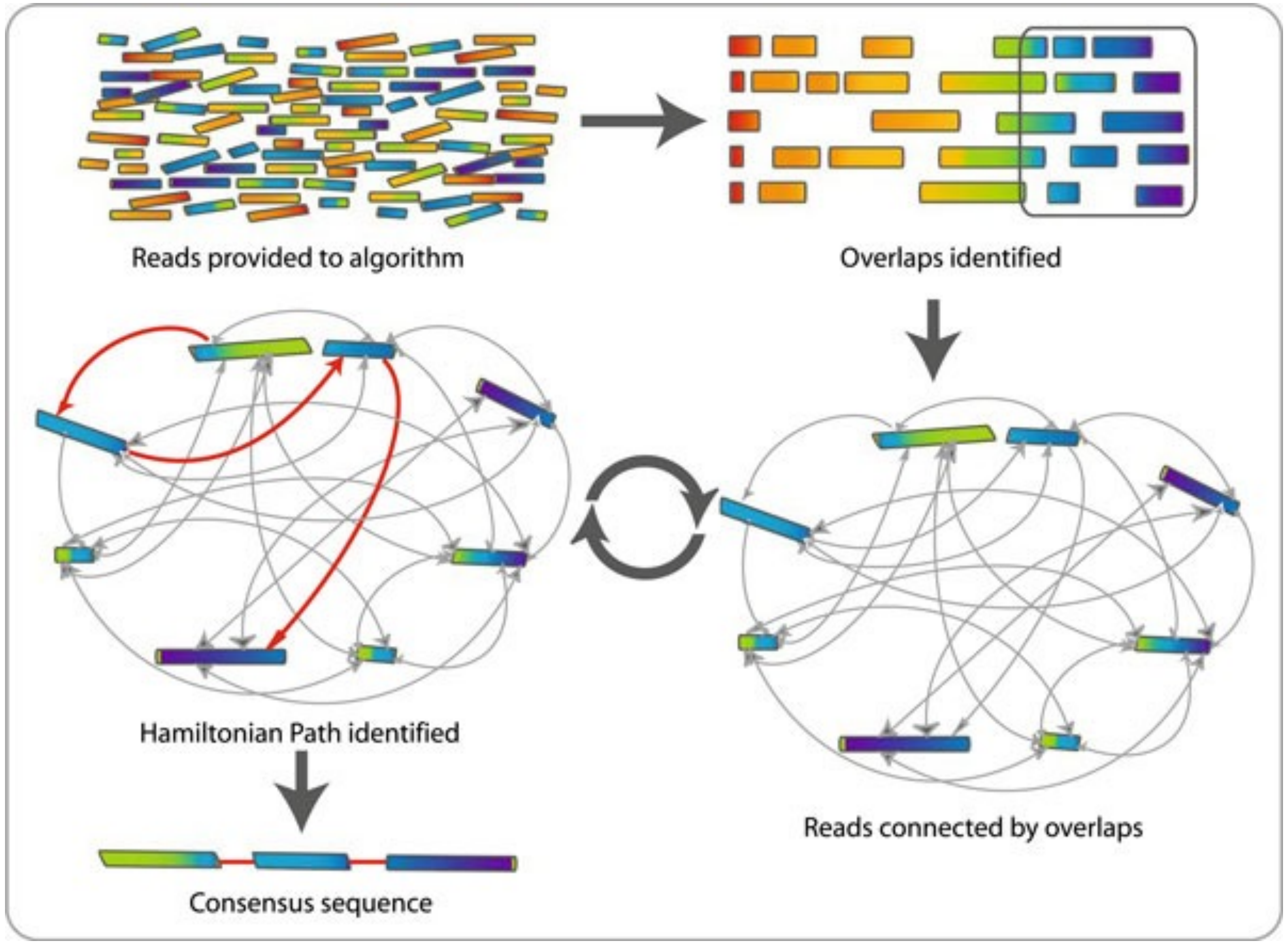
Variant analysis
WES

Reference Assembly

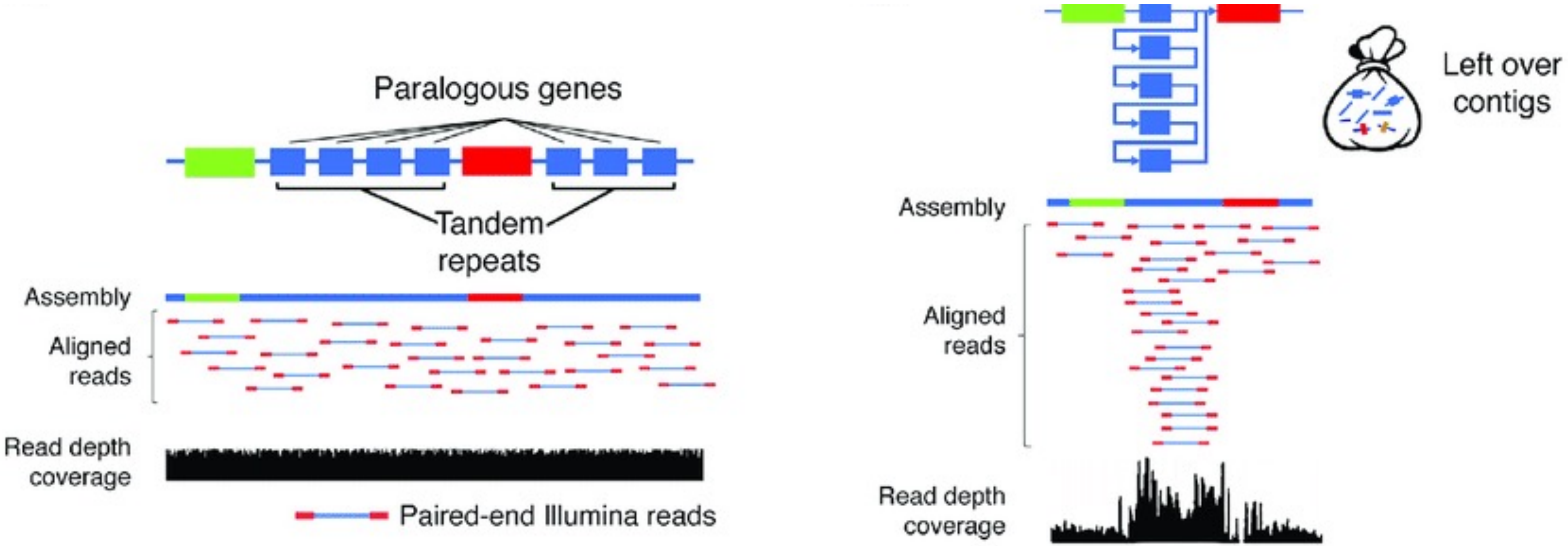


Reference Assembly



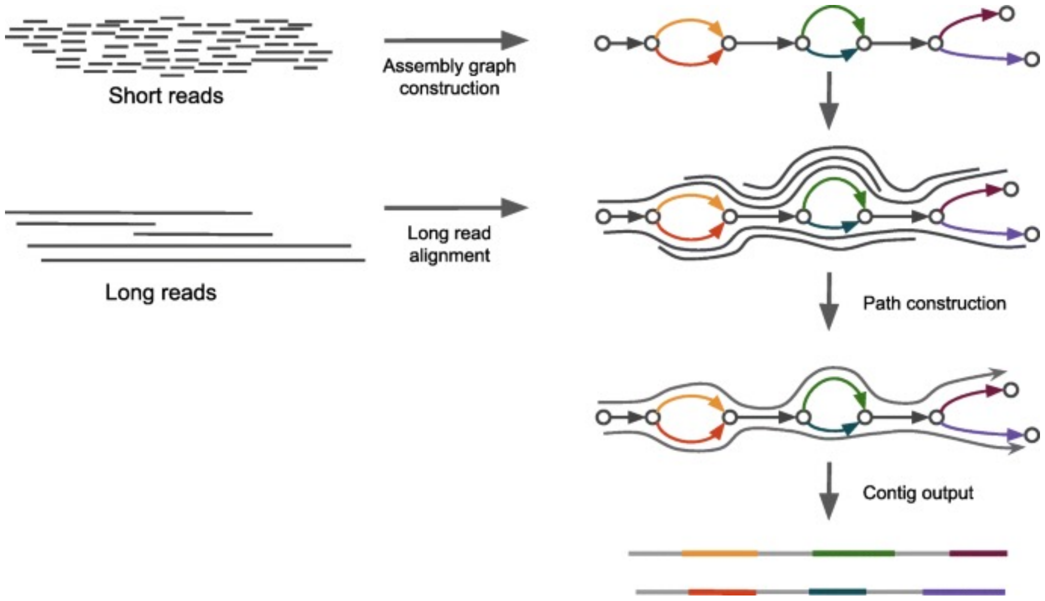

Reads provided to algorithm
Overlaps identified
Reads connected by overlaps
Hamiltonian Path identified
Consensus sequence

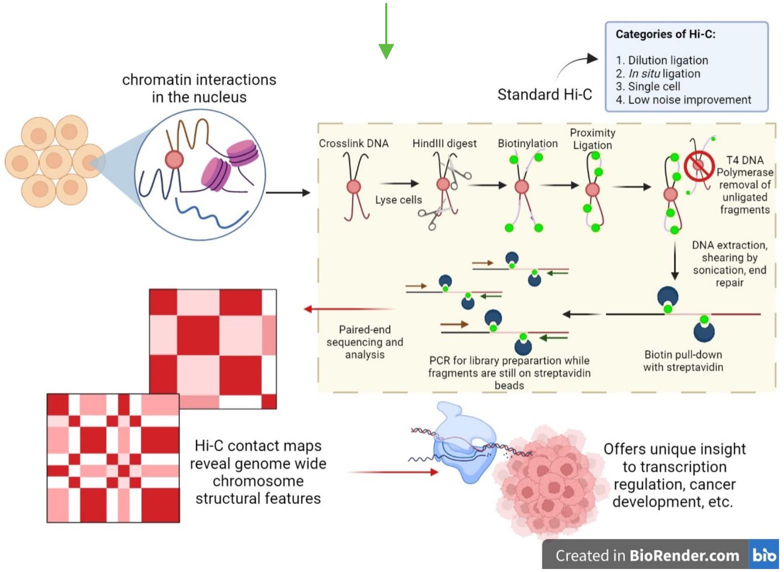
Reference Assembly problematic with short read



Genome Assembly

- Very hard and costly (in eukaryota)
- Multiple sequencing types needed
  - Pair-end short reads
  - Long reads
  - Mate-pairs (e.g. Hi-C)

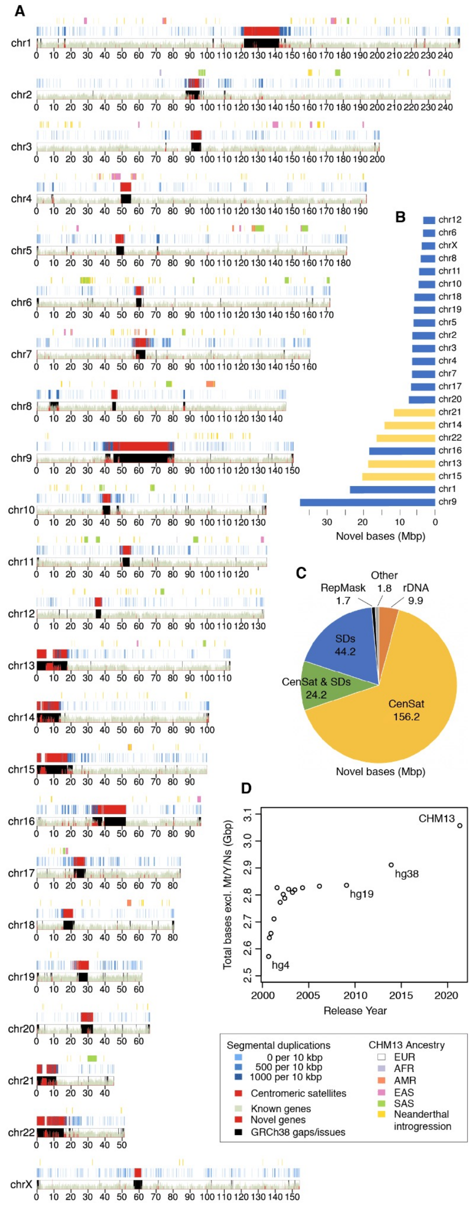



chromatin interactions in the nucleus
Standard Hi-C
Categories of Hi-C:
1. Dilution ligation
2. *In situ* ligation
3. Single cell
4. Low noise improvement
Crosslink DNA
Lyse cells
HindIII digest
Biotinylation
Proximity Ligation
T4 DNA Polymerase removal of unligated fragments
DNA extraction, shearing by sonication, end repair
Biotin pull-down with streptavidin
PCR for library preparartion while fragments are still on streptavidin beads
Paired-end sequencing and analysis
Hi-C contact maps reveal genome wide chromosome structural features
Offers unique insight to transcription regulation, cancer development, etc.
Created in BioRender.com bio

Genome Assembly

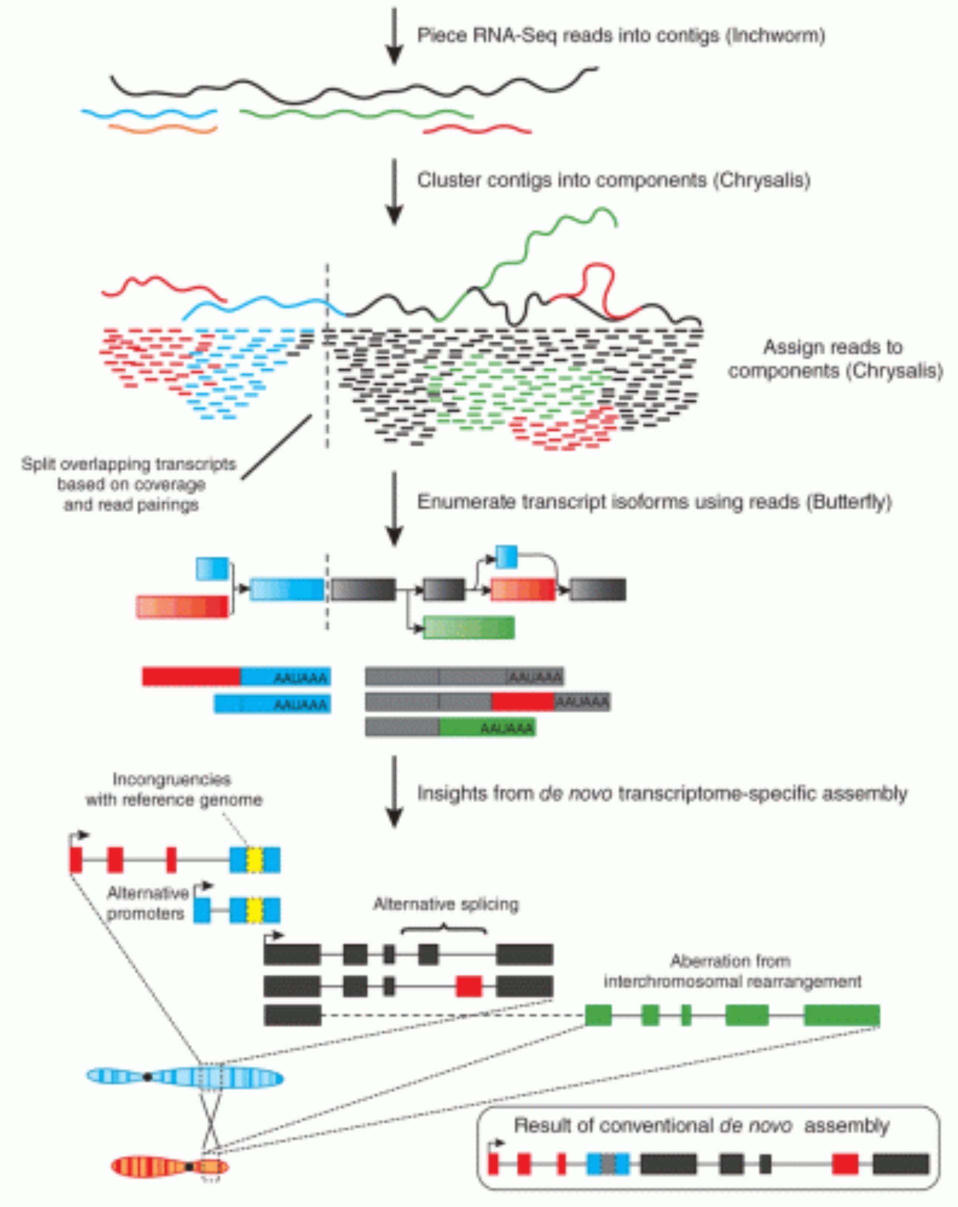
- Very hard and costly (in eukaryota)
- Multiple sequencing types needed
  - Pair-end short reads
  - Long reads
  - Mate-pairs (e.g. Hi-C)

T2T-CHM13

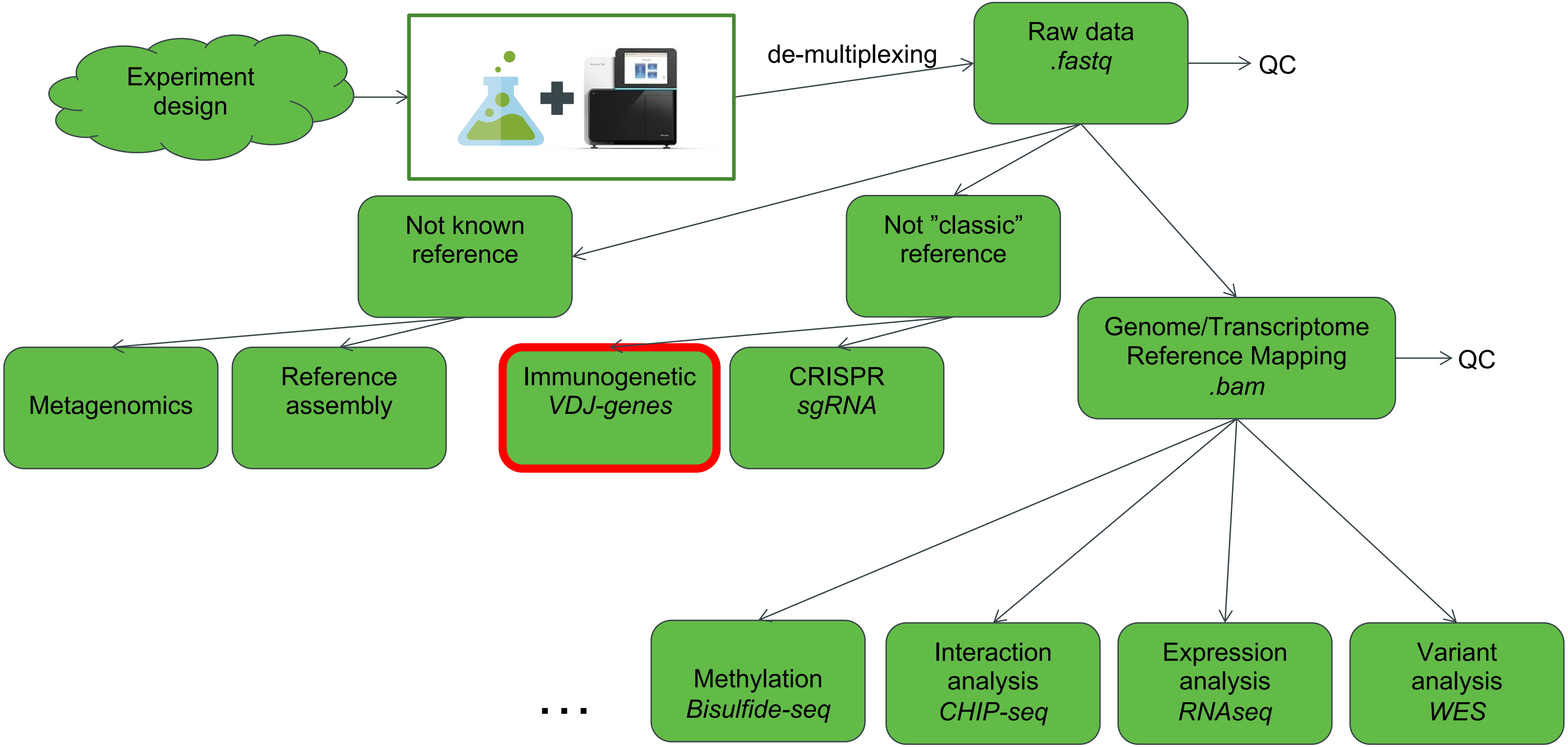


Transcriptome Assembly

- Assemble RNA fragments
  - Similar reference helpful
- Genome guided assembly
  - Good for poorly annotated organisms with known genomic reference

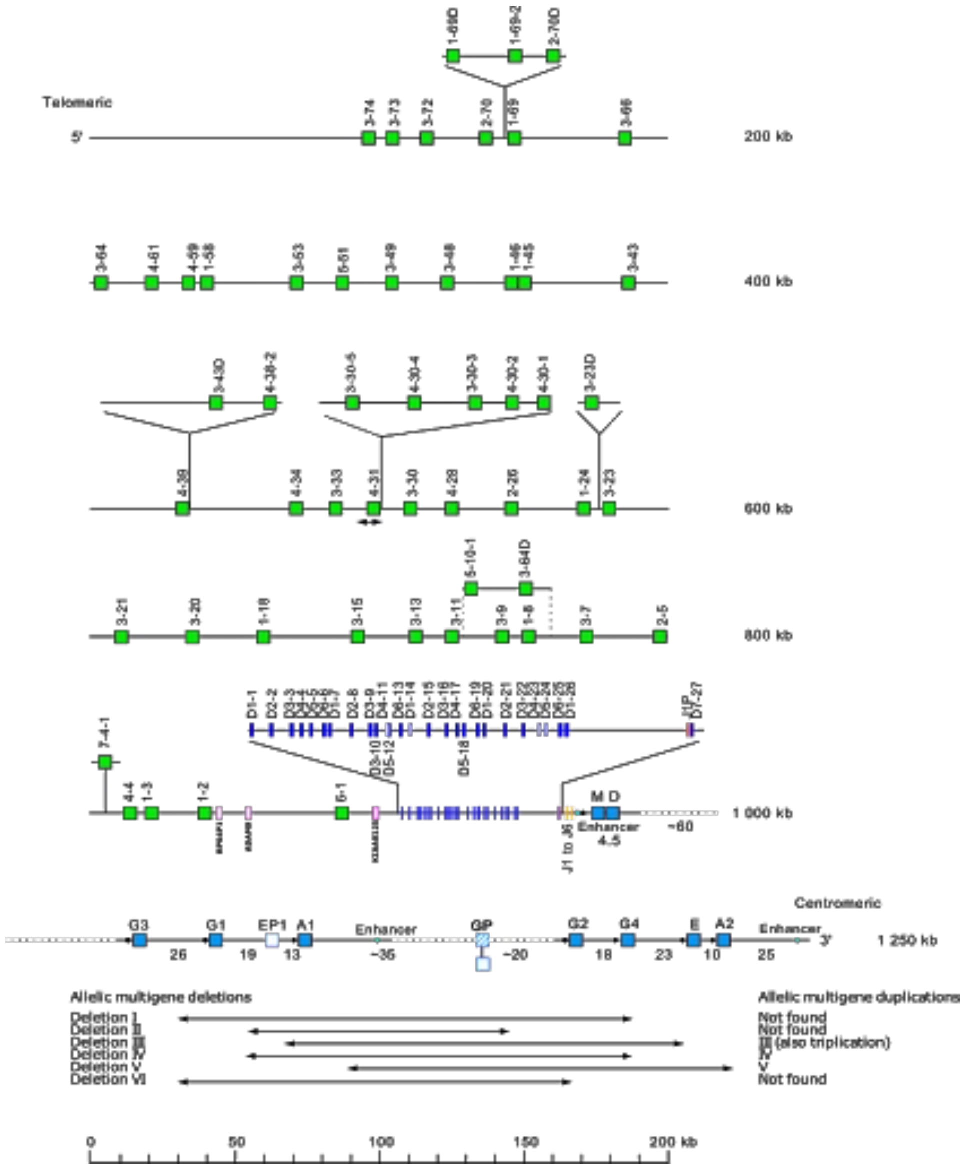

Piece RNA-Seq reads into contigs (Inchworm)
Cluster contigs into components (Chrysalis)
Assign reads to components (Chrysalis)
Split overlapping transcripts based on coverage and read pairings
Enumerate transcript isoforms using reads (Butterfly)
AAUAAA
AAUAAA
AAUAAA
AAUAAA
AAUAAA
Incongruencies with reference genome
Insights from *de novo* transcriptome-specific assembly
Alternative promoters
Alternative splicing
Aberration from interchromosomal rearrangement
Result of conventional *de novo* assembly

NGS data analysis



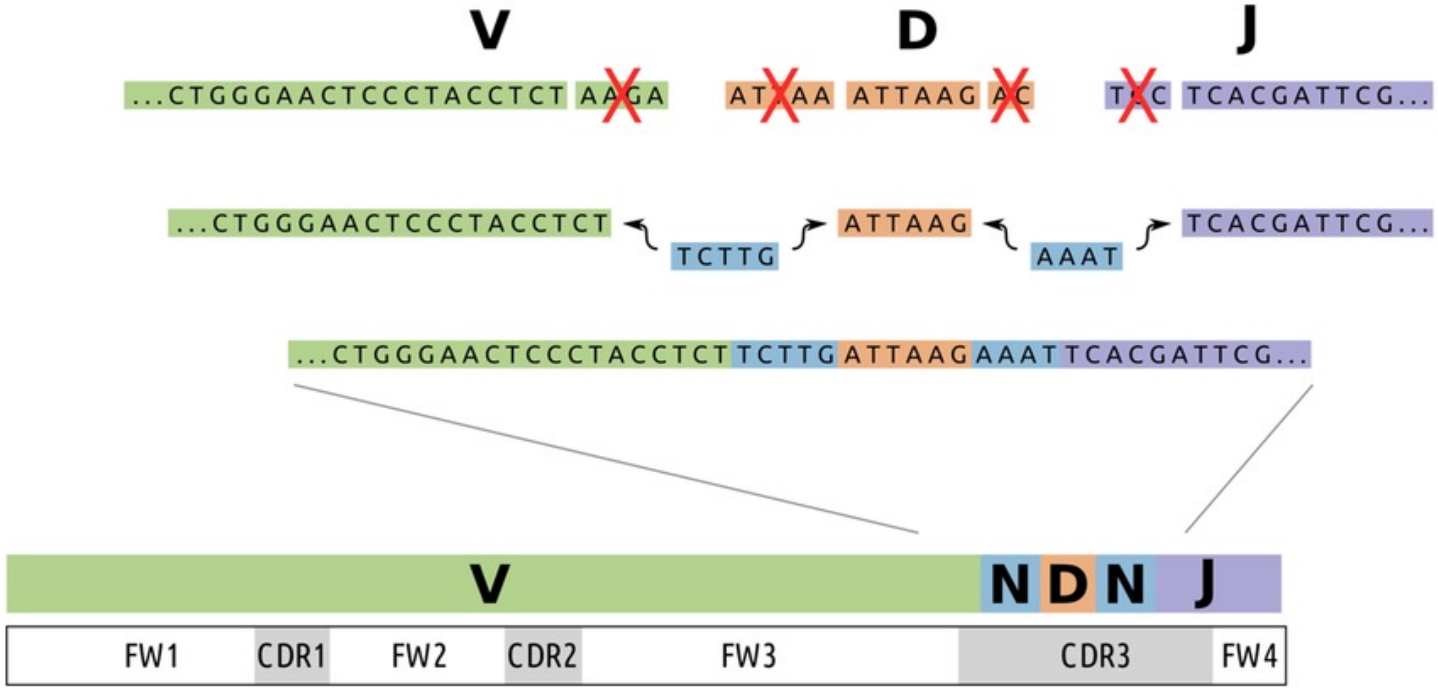
Immunogenetic

- T-cell receptor , Immunoglobulin – (B-cell)
- Gene rearrangement during cell maturation
  - VDJ recombination

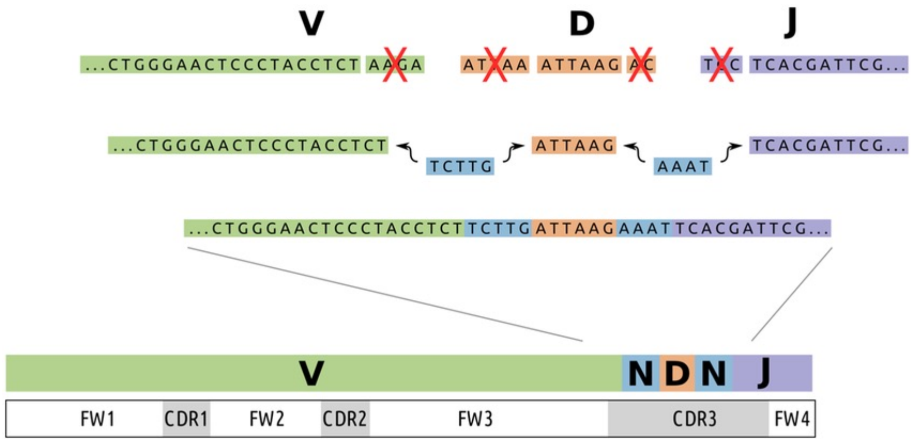


Immunogenetic

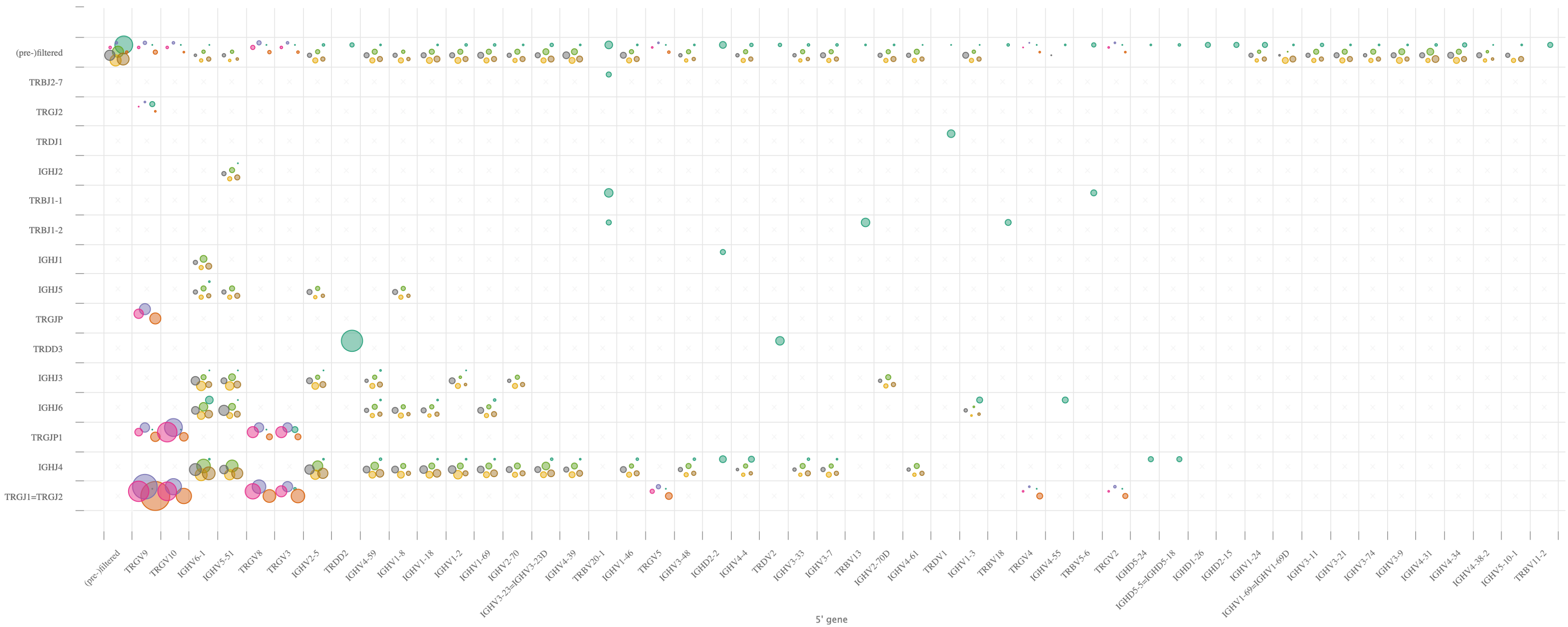
- T-cell receptor , Immunoglobulin – (B-cell)
- Gene rearrangement during cell maturation
  - VDJ recombination



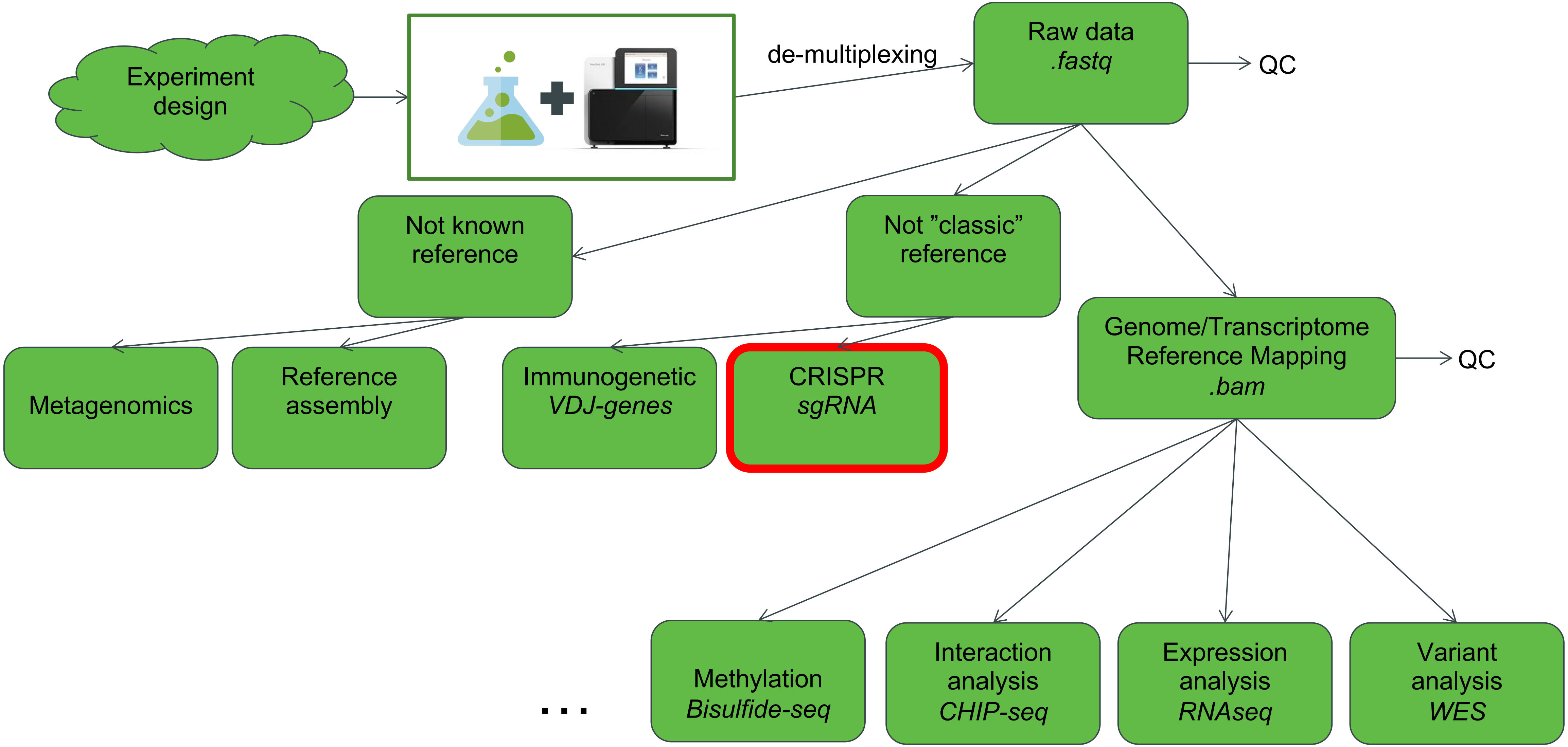
Immunogenetic



- Different cell populations
  - Clonal studies
  - Repertoire usage
- Main usage – blood malignancies (leukemias)



NGS data analysis



Experiment design

de-multiplexing

Raw data *.fastq*

QC

Not known reference

Not ”classic” reference

Genome/Transcriptome Reference Mapping *.bam*

QC

Metagenomics

Reference assembly

Immunogenetic *VDJ-genes*

CRISPR *sgRNA*

. . .

Methylation *Bisulfide-seq*

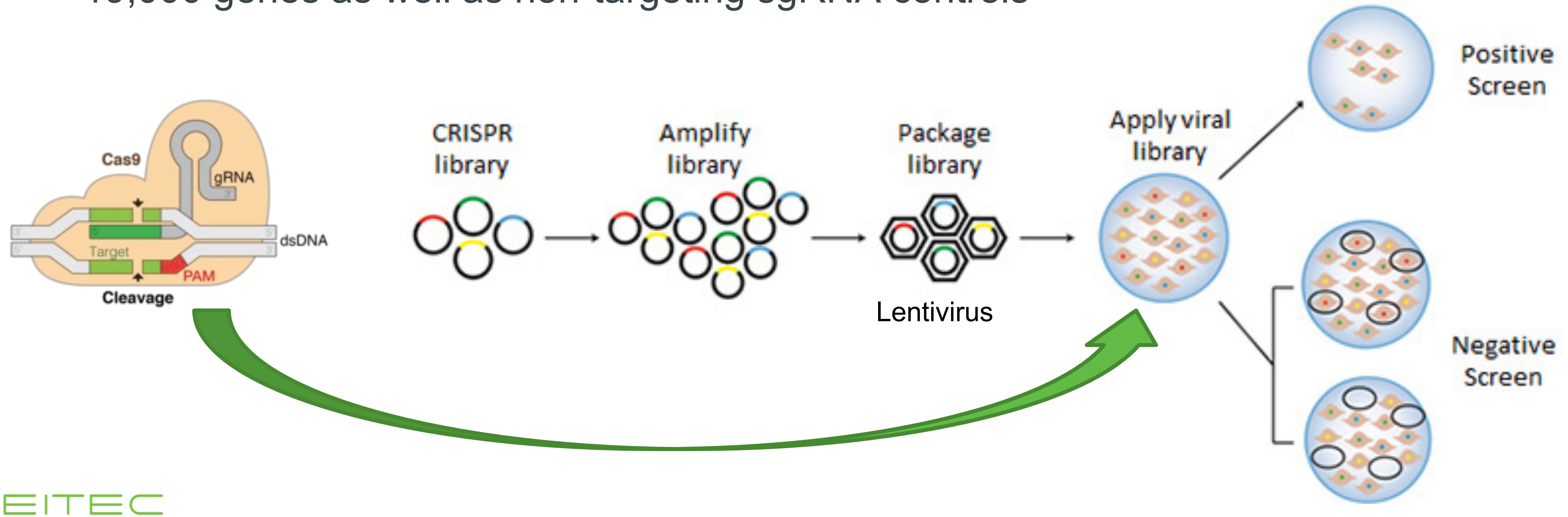
Interaction analysis *CHIP-seq*

Expression analysis *RNAseq*

Variant analysis *WES*

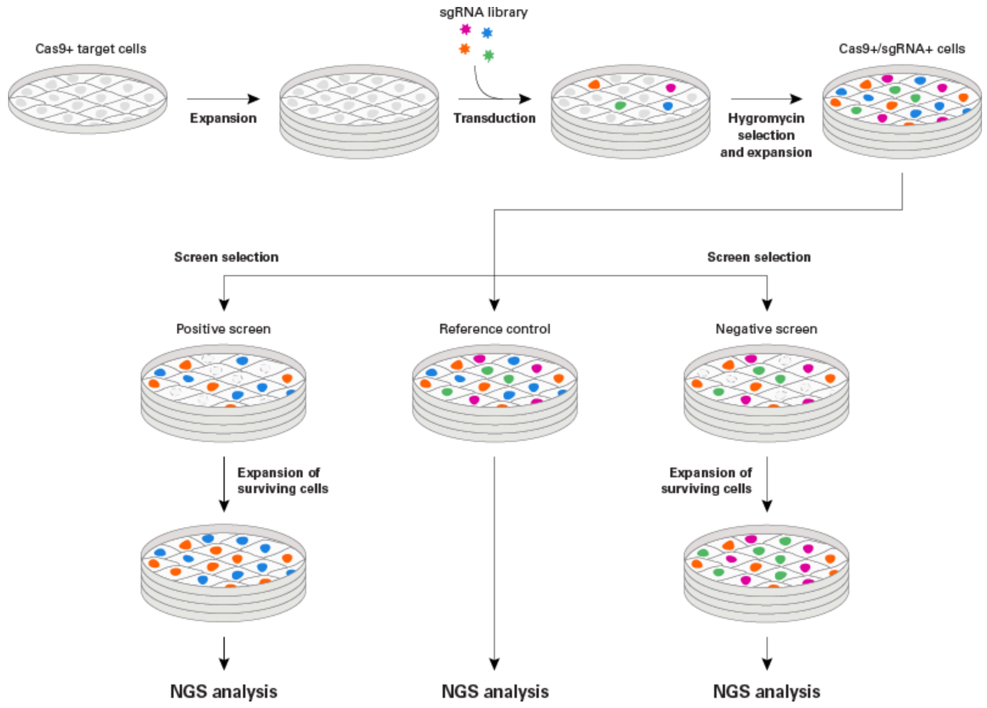
Genome-wide CRISPR-Cas9 knockout screens

- Cas9 (CRISPR associated protein 9) is a protein which plays a vital role in the immunological defense of certain bacteria against DNA viruses
- sgRNA libraries
  - Each sgRNA knockout specific gene
  - 76,000 guide RNAs (sgRNAs) with four highly active guides per gene, targeting about 19,000 genes as well as non-targeting sgRNA controls



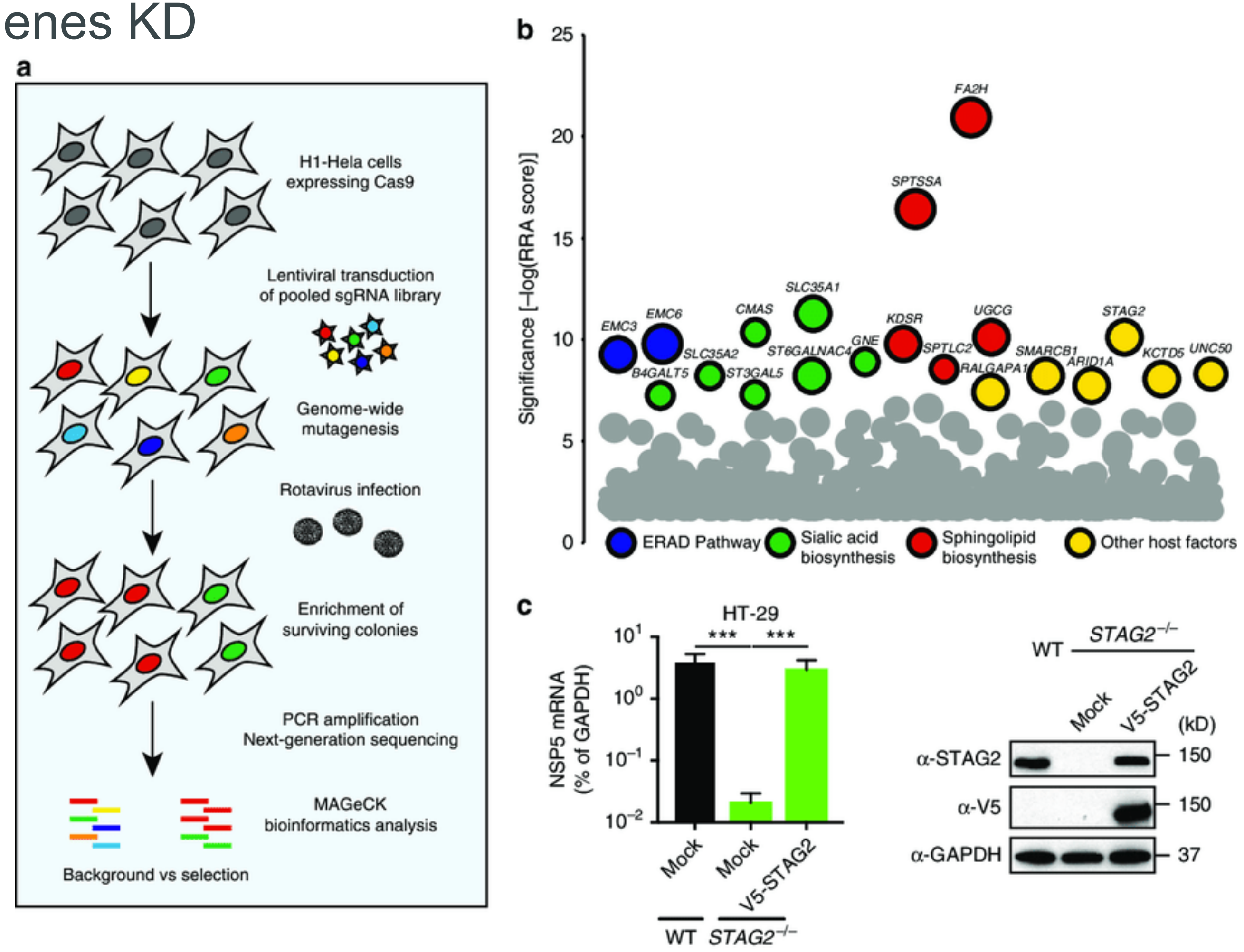
Genome-wide CRISPR-Cas9 knockout screens

- Screen selection + expansion/enrichment of surviving cells
- NGS sequencing



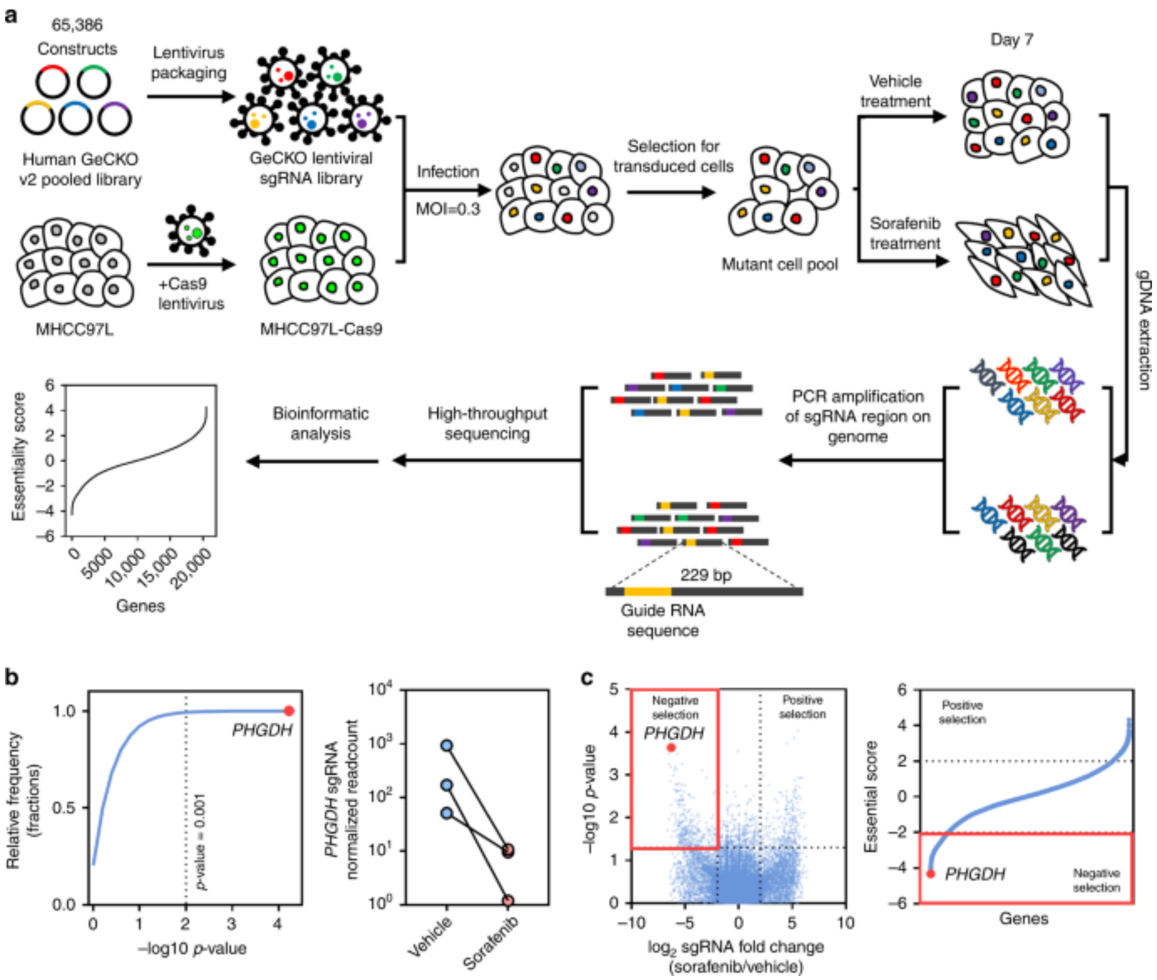
Genome-wide CRISPR-Cas9 knockout screens

- NGS data analysis
  - Counting cells with different genes KD
  - Counting sgRNA fragments
  - Compare conditions



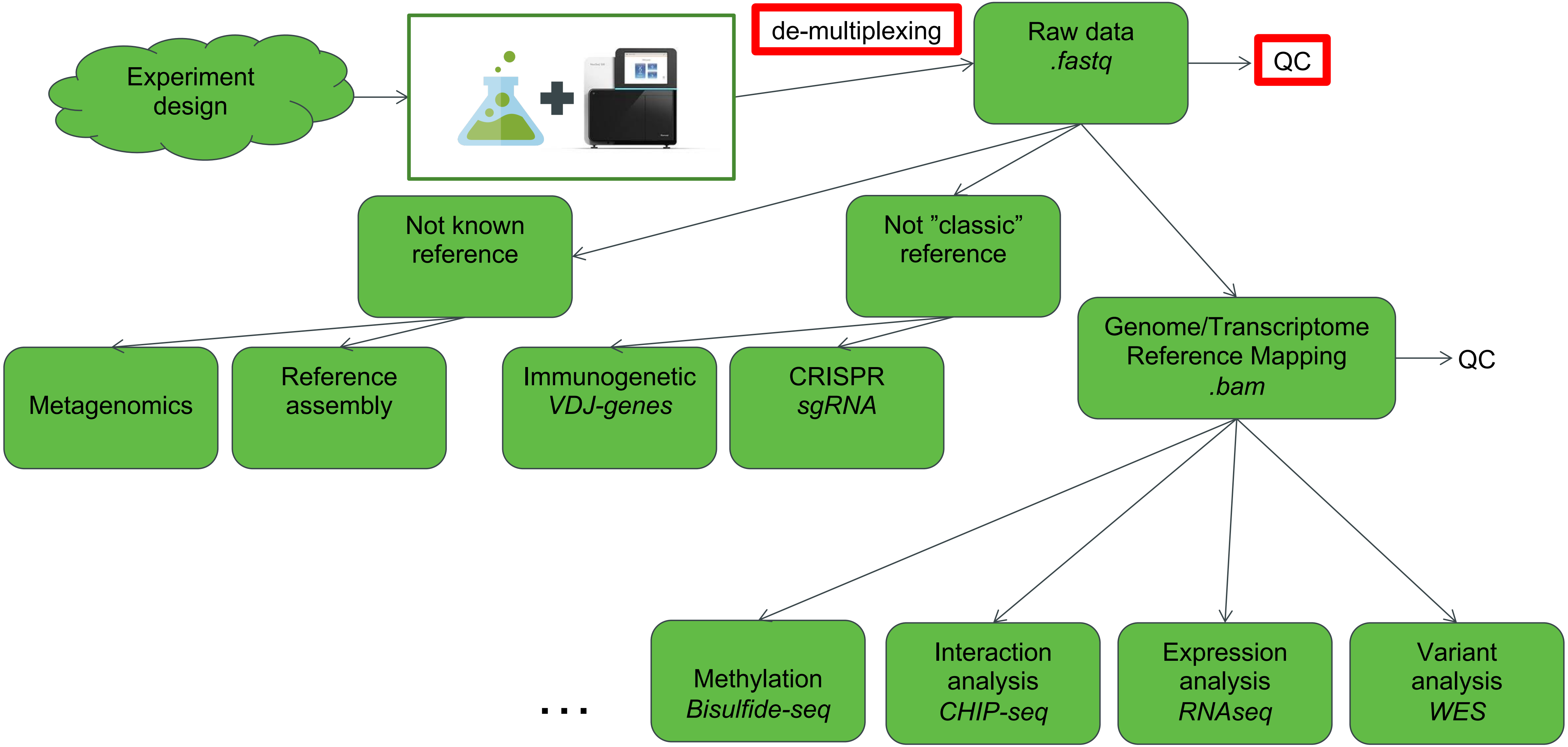
Genome-wide CRISPR-Cas9 knockout screens

- Example study



Wei, L., Lee, D., Law, CT. *et al.* Genome-wide CRISPR/Cas9 library screening identified PHGDH as a critical driver for Sorafenib resistance in HCC. *Nat Commun* **10,** 4681 (2019). https://doi.org/10.1038/s41467-019-12606-7

NGS data analysis



Experiment design

de-multiplexing

Raw data
.fastq

QC

Not known reference

Not ”classic” reference

Genome/Transcriptome Reference Mapping
.bam

QC

Metagenomics

Reference assembly

Immunogenetic
VDJ-genes

CRISPR
sgRNA

. . .

Methylation
Bisulfide-seq

Interaction analysis
CHIP-seq

Expression analysis
RNAseq

Variant analysis
WES

CEITEC

@CEITEC_Brno

Thank you for your attention!

Vojta Bystry
vojtech.bystry@ceitec.muni.cz