HELMHOLTZ MUNICH

Sequencing methods

Mauro Tutino
VSS 2024

Genetic profiling

1. Genotyping array
2. NGS sequencing
3. Third-generation sequencing
4. RNA-sequencing
5. Single-cell sequencing

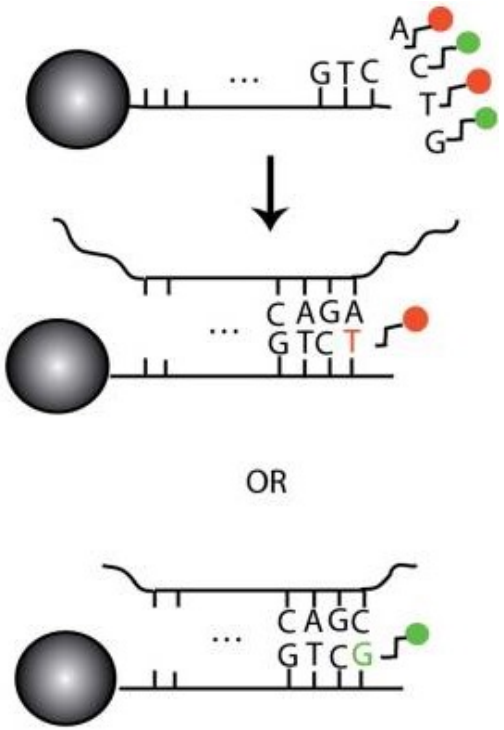
1

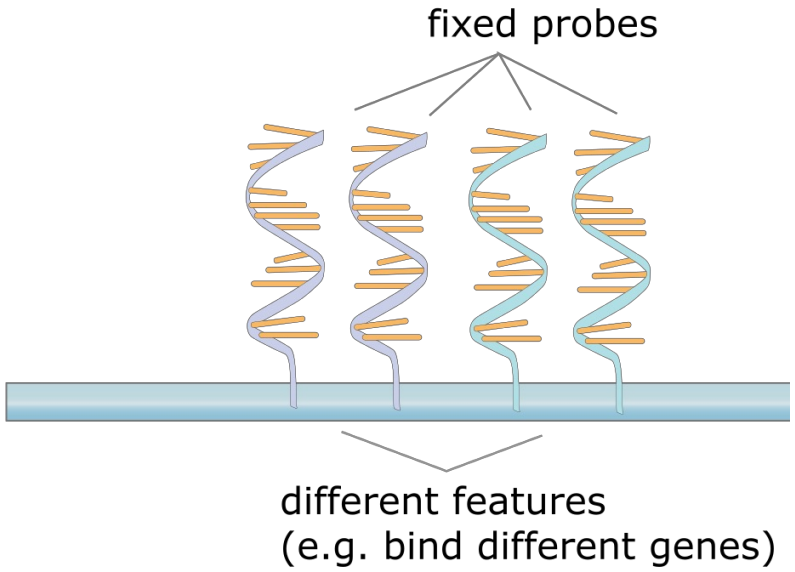
Genotyping

Genotyping arrays

Limitations:

- *SNP frequencies are very different from one population to another*
  - *Every significantly different population will need a custom array*
- *Rare variants*
  - *Difficult to genotype*
  - *Poorly imputed*
- *De novo mutations*

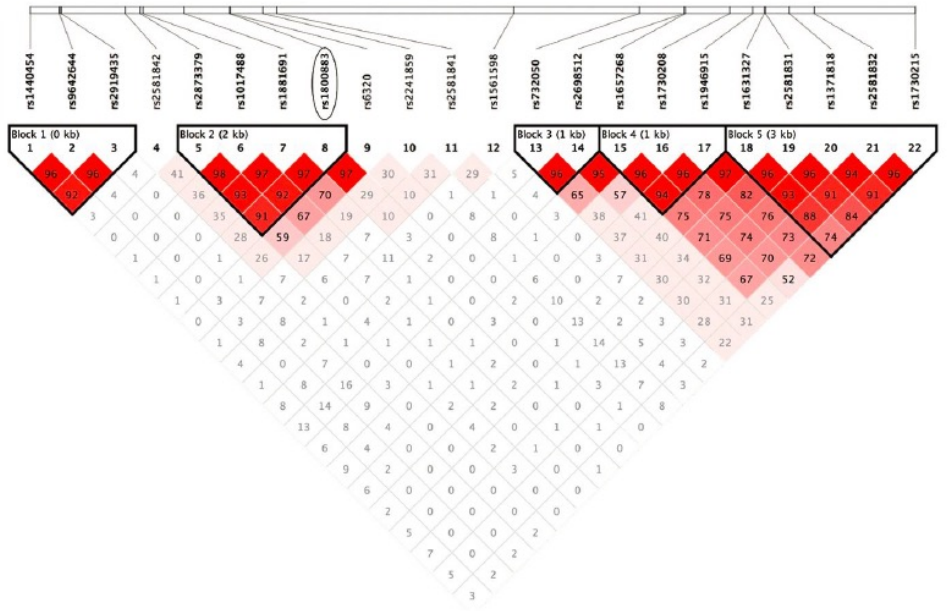




https://en.wikipedia.org/wiki/DNA_microarray#/media/File:NA_hybrid.svg

Imputation – LD

- Goal = get genotypes at untyped positions in the target dataset
- Based on the linkage disequilibrium (LD) between variants
  - “Haplotypes” = group of alleles inherited together

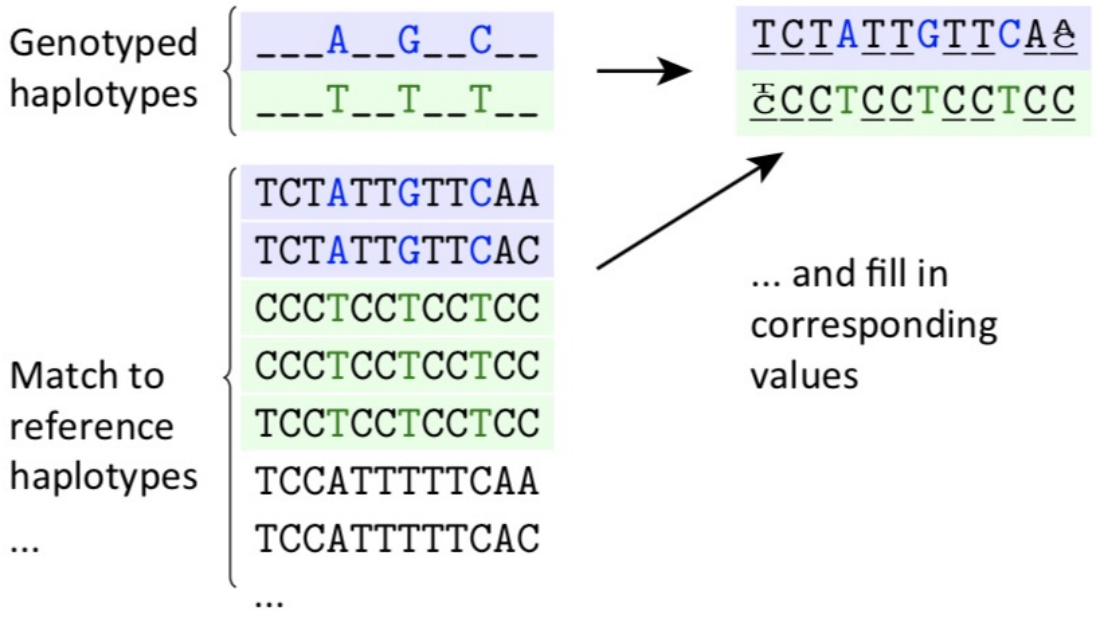


Imputation

- Use the correlations between the SNPs to ‘guess’ the genotypes at untyped positions

- Reference datasets such as www.1000genomes.org, www.uk10k.org and www.haplotype-reference-consortium.org
  - Importance to match the population
  - Panels for non-European populations:

- Example of software used for imputation: + REF



Genotyping arrays

Pros

- Cheap technology
- Easy to do
- Standard pipelines for QC

Cons

- Interrogate only known variations
- Limitations when comparing different studies
- Imputation is limited for rare variants as low LD

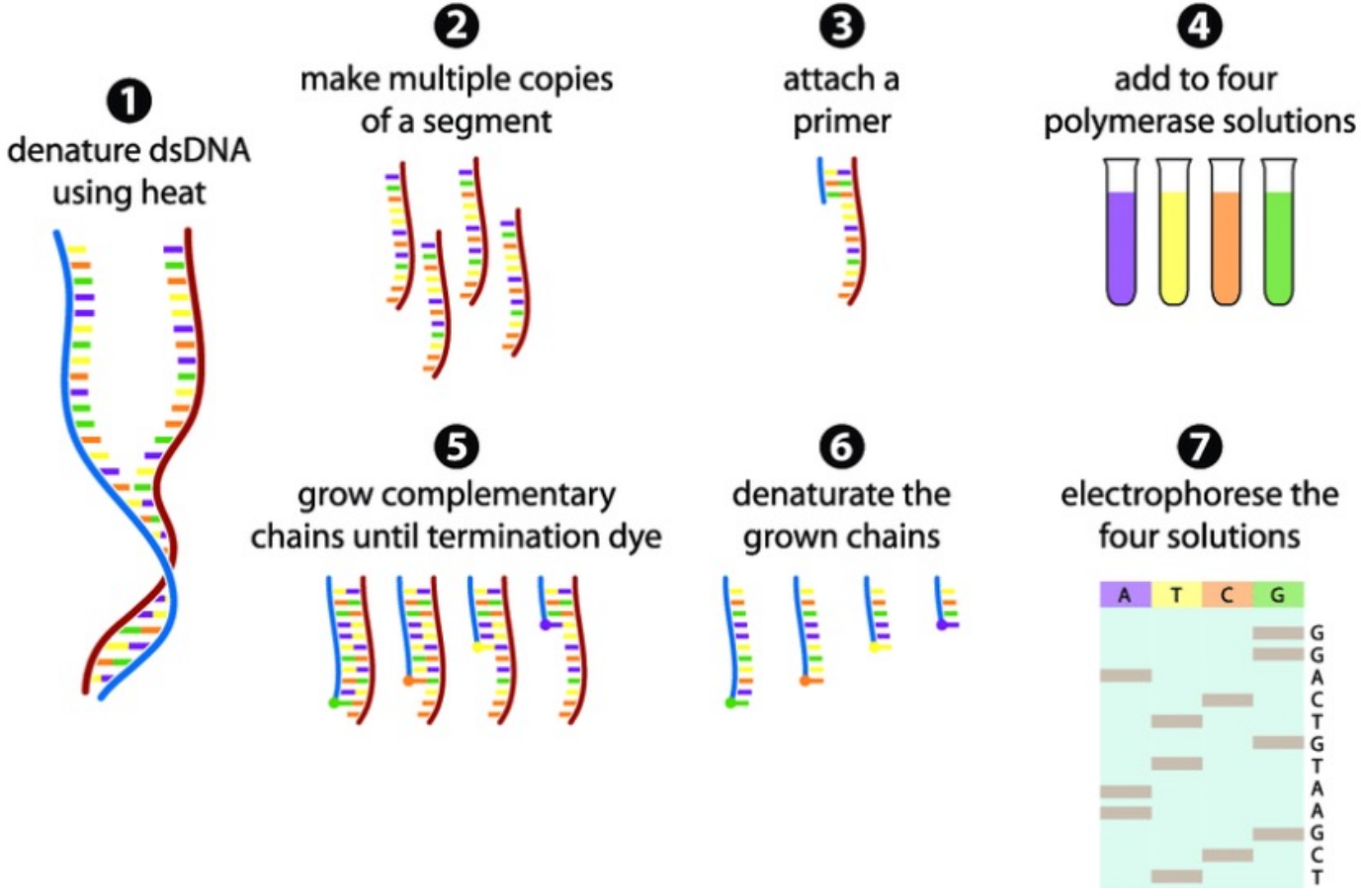
2

NGS sequencing

Sanger sequencing

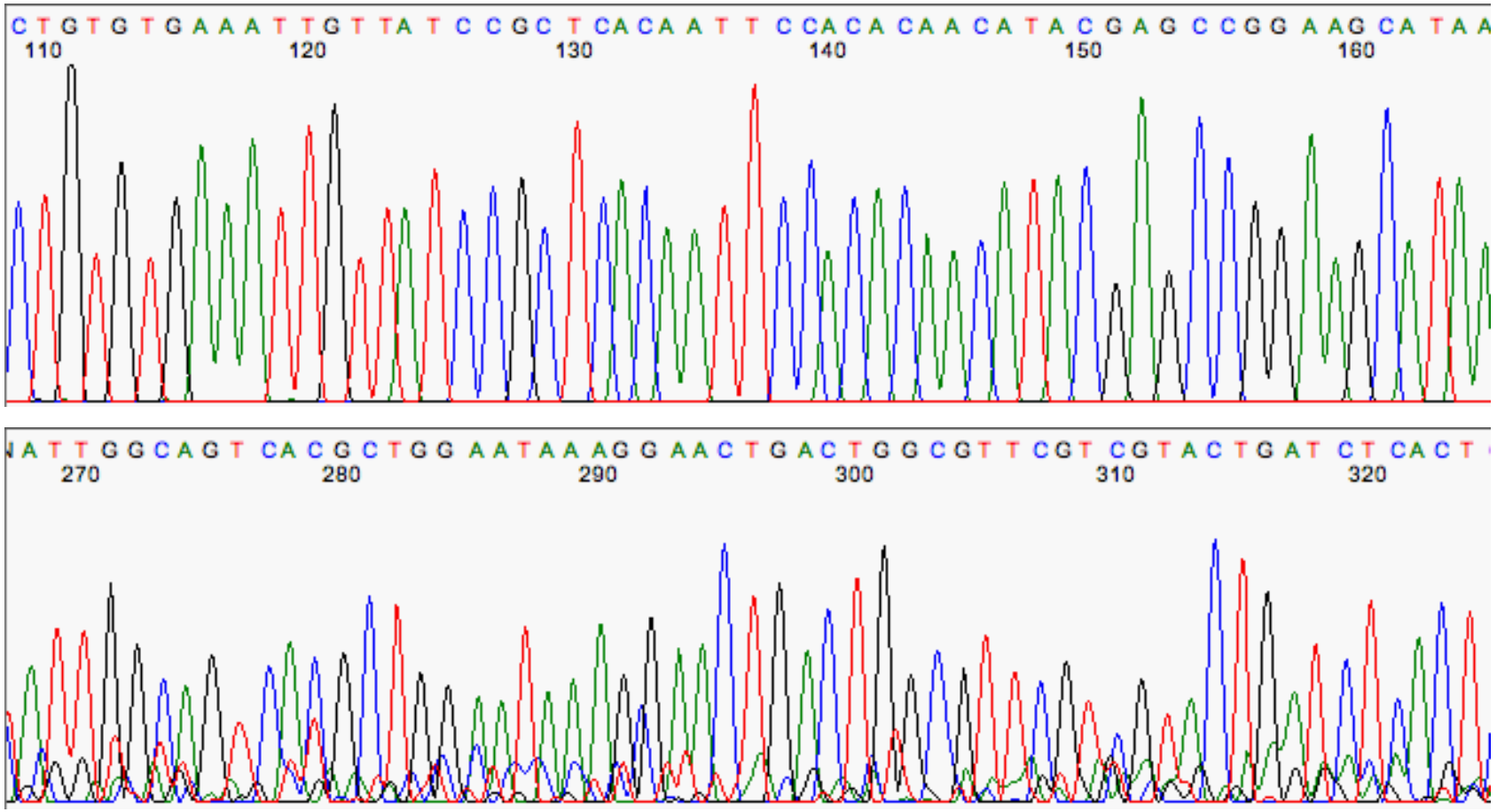
1st generation: 1977 – Reads ~ 700 bp

- Sequencing = get genotype information for every position in the genome including unknown variation
- 1st sequencing approach = Sanger sequencing, reads of ~700bp



Sanger sequencing

1st generation: 1977 – Reads ~ 700 bp



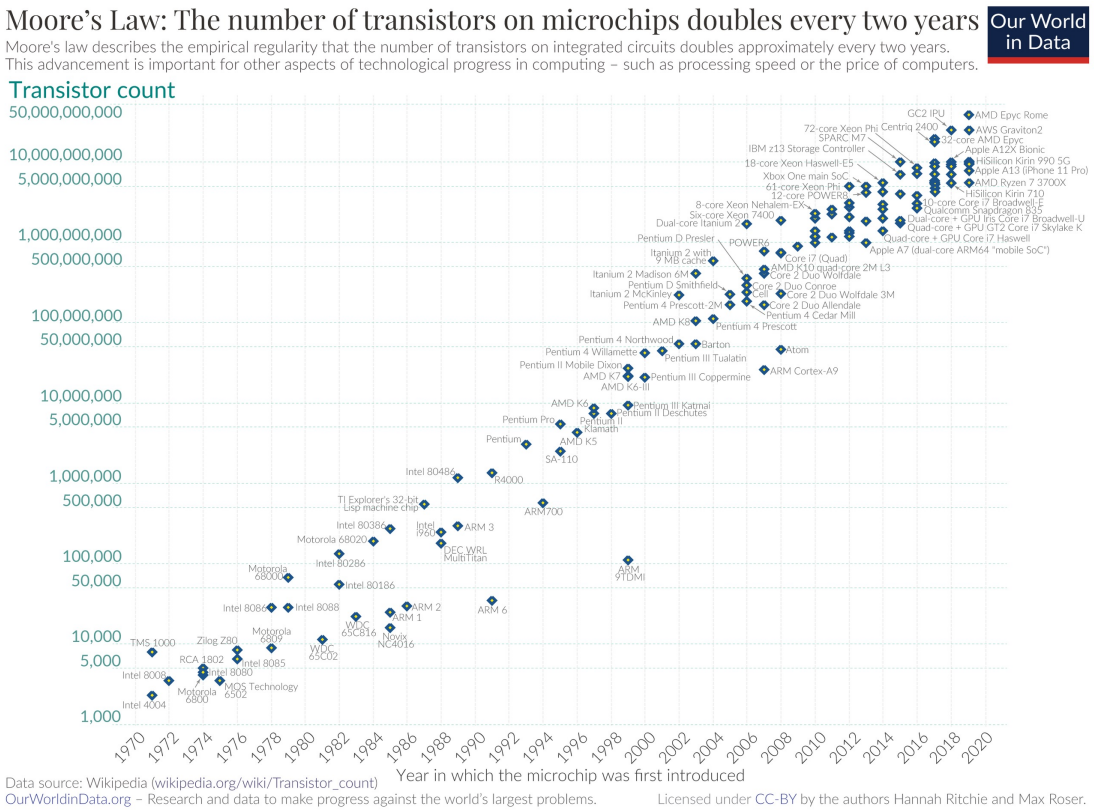
NGS Sequencing

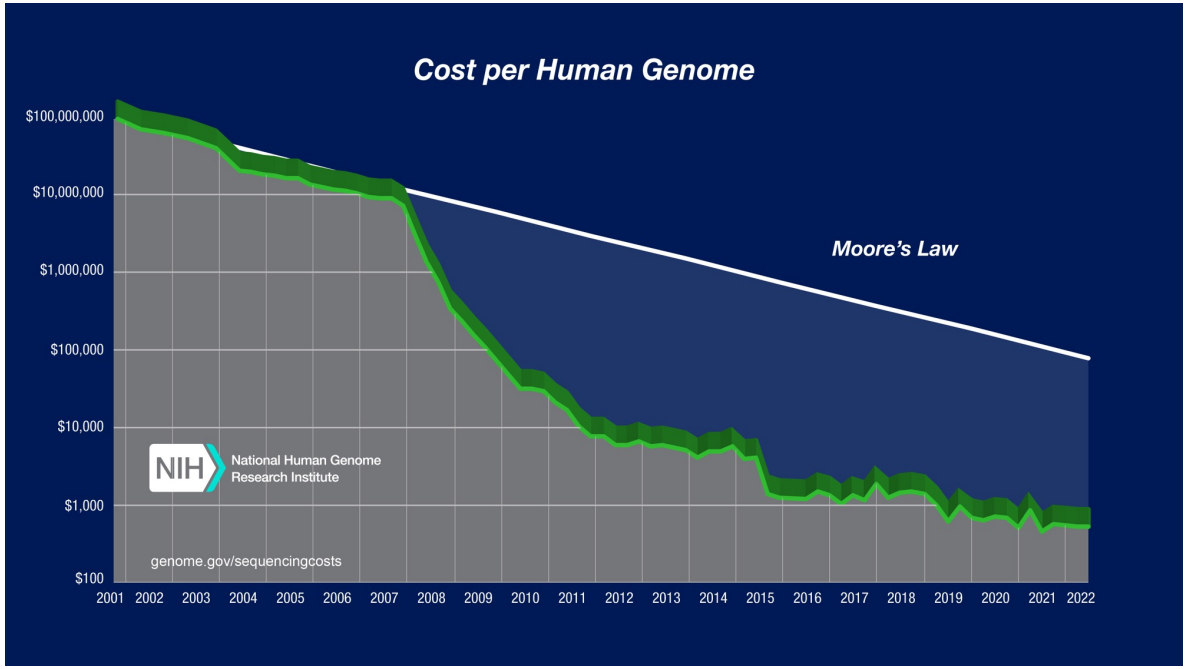
- NGS = Next Generation Sequencing
  - More efficient technologies for sequencing, especially through the use of multiplexing
  - Smaller reads

- 2 main technologies:
  - IonTorrent: transistor sequencing
  - **Illumina: Sequencing-By-Synthesis (SBS)**

- Illumina is now the market leader due to very high throughput and low prices

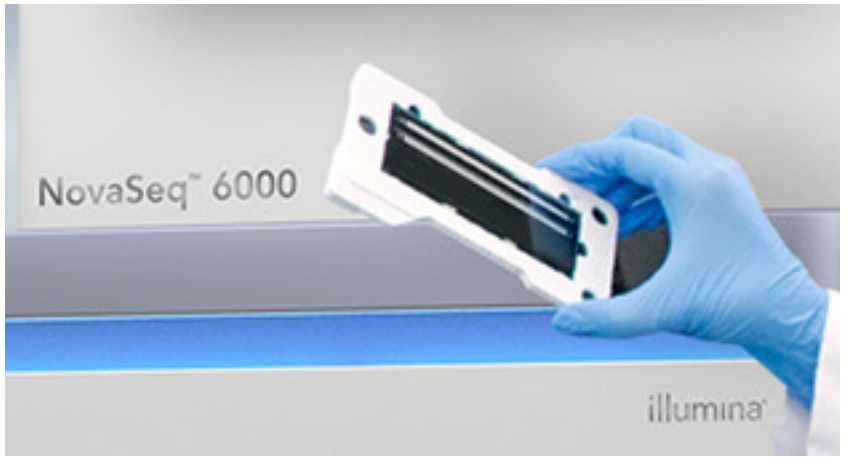
NGS price is getting lower and lower

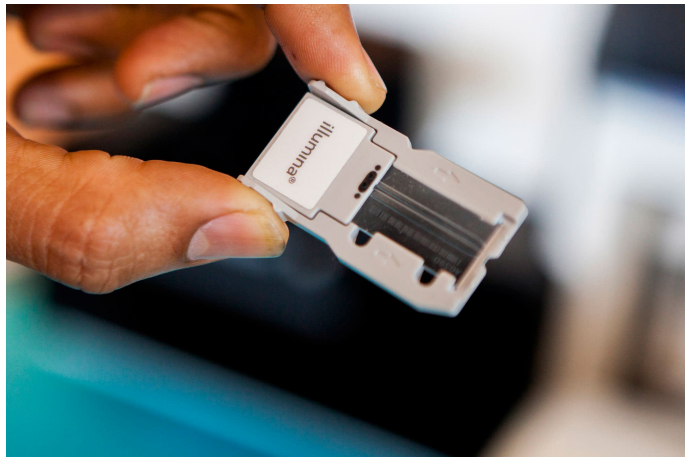




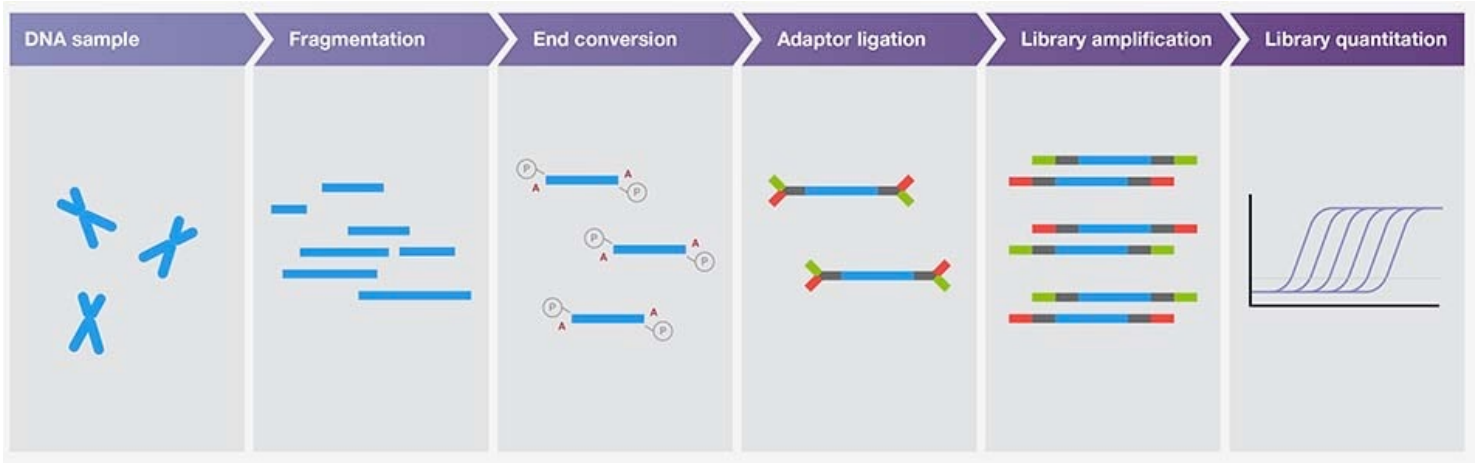
HELMHOLTZ MUNICH

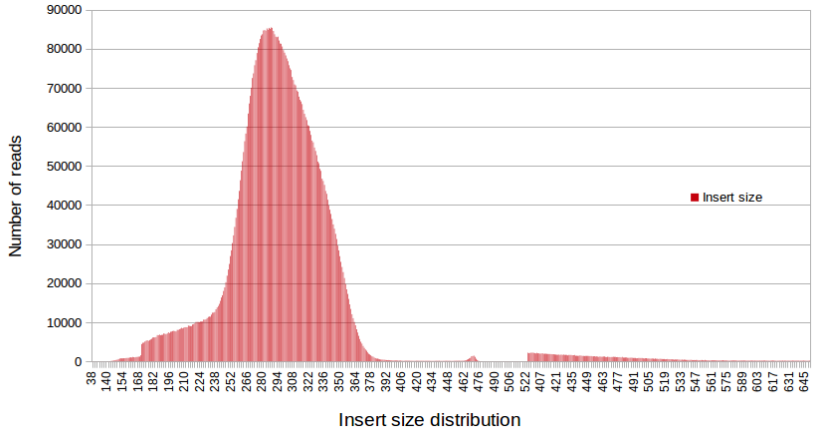
Illumina sequencer

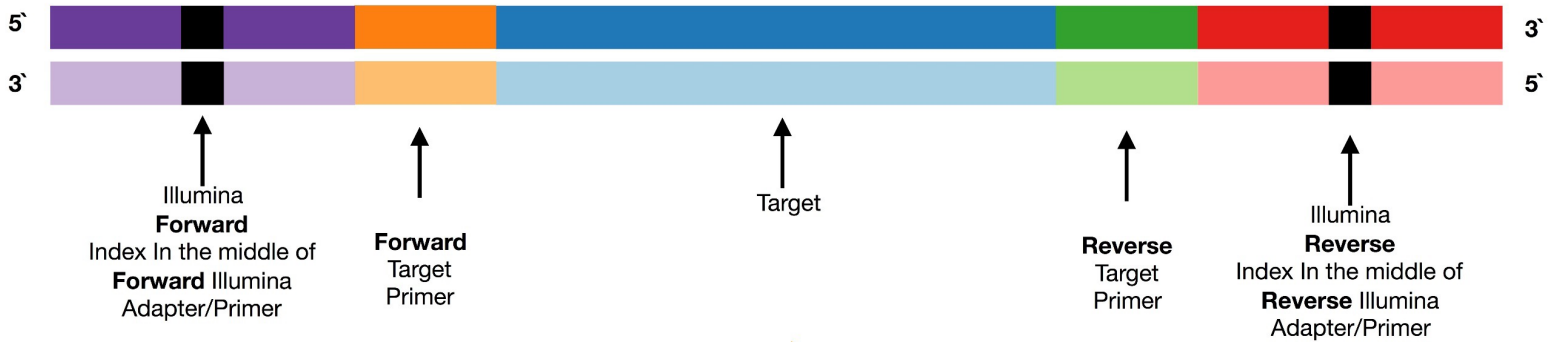




Illumina sequencing reads

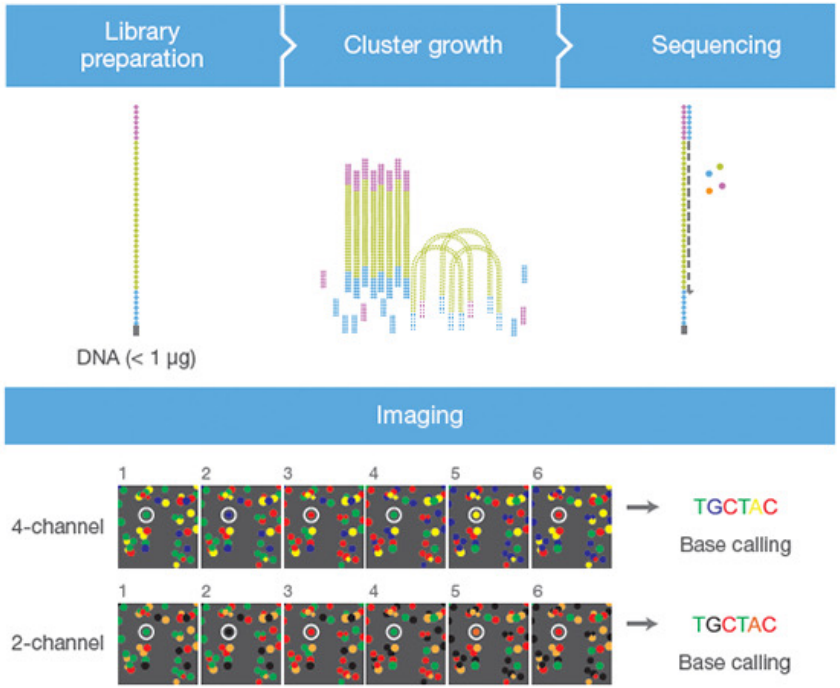


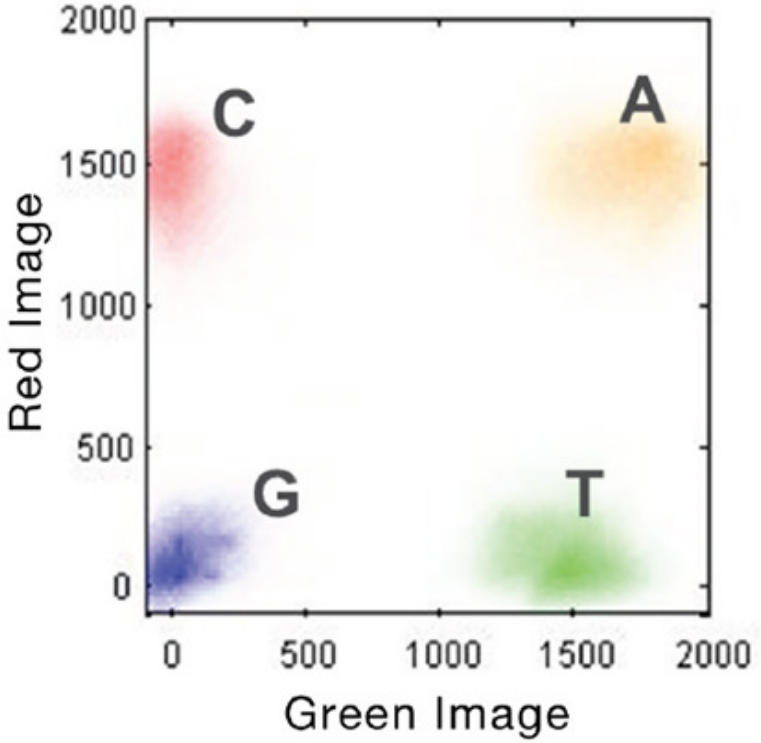


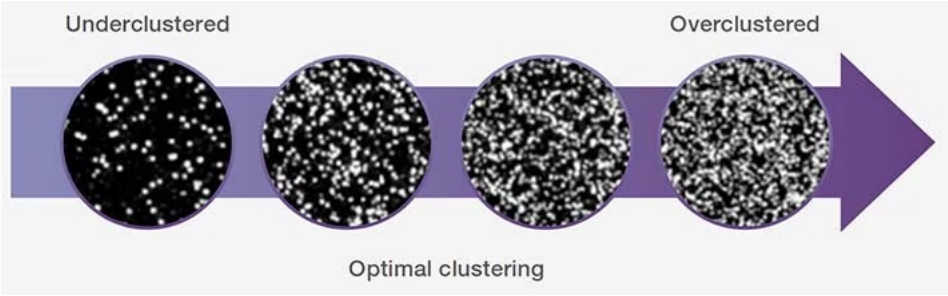


HELMHOLTZ MUNICH

Illumina sequencing (SBS)







FASTQ format

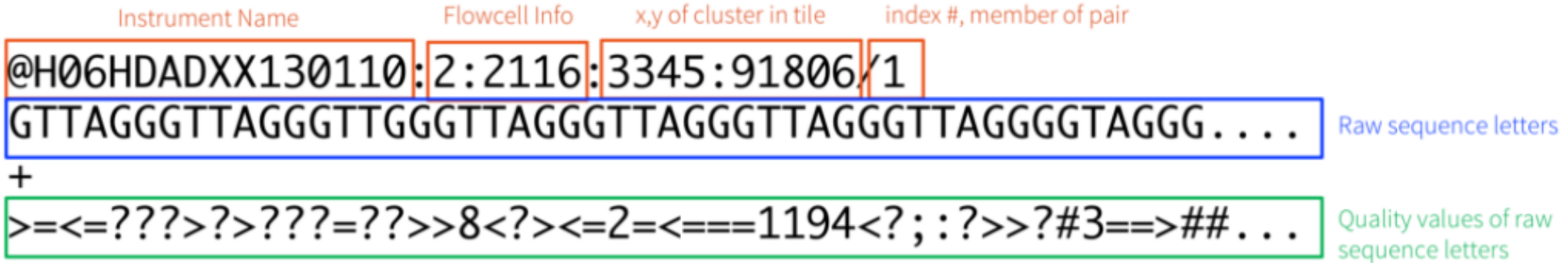
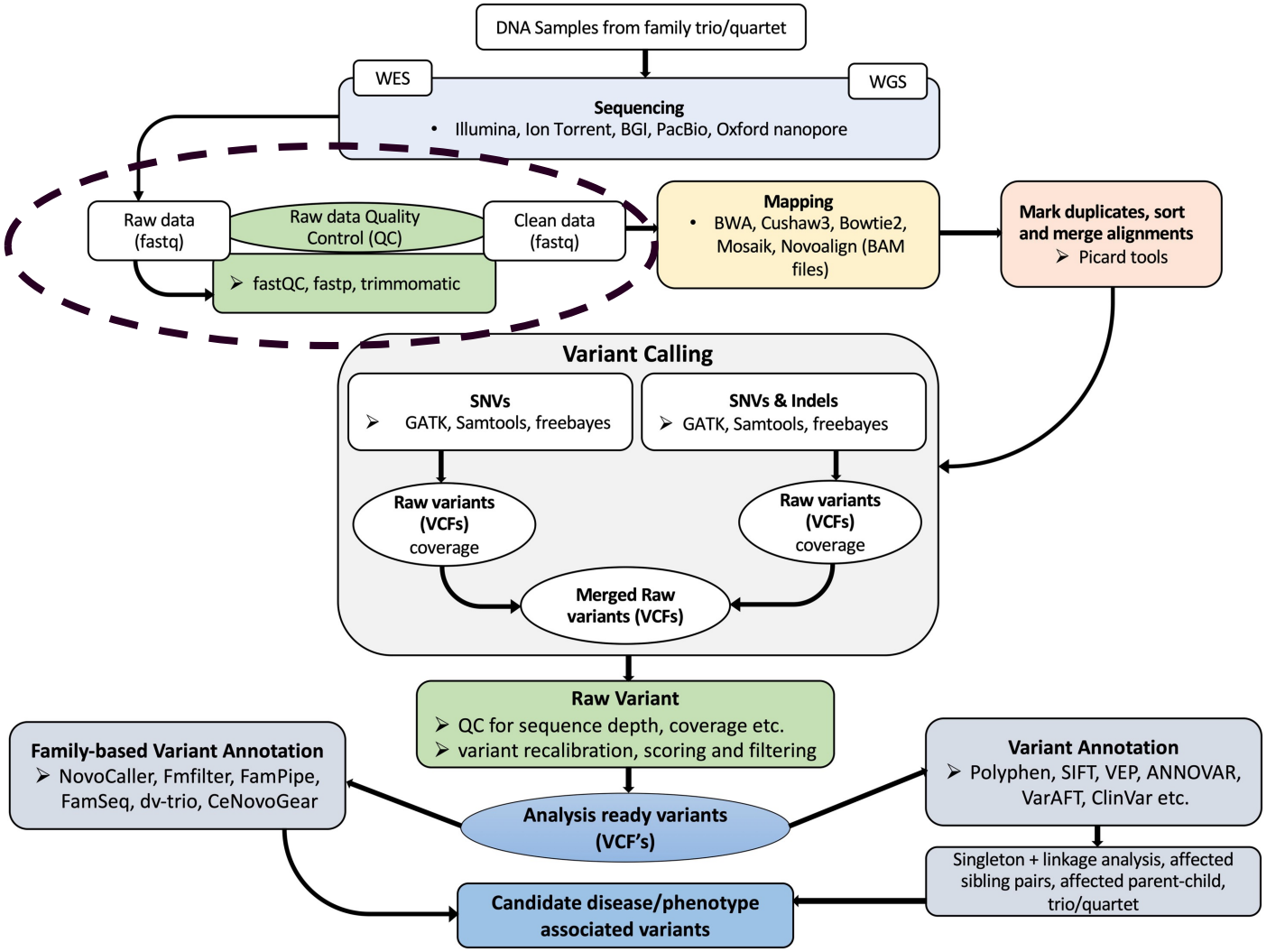


Table 1 ASCII Characters Encoding Q-scores 0–40

Symbol	ASCII Code	Q-Score	Symbol	ASCII Code	Q-Score	Symbol	ASCII Code	Q-Score
!	33	0	/	47	14	=	61	28
"	34	1	0	48	15	>	62	29
#	35	2	1	49	16	?	63	30
$	36	3	2	50	17	@	64	31
%	37	4	3	51	18	A	65	32
&	38	5	4	52	19	B	66	33
'	39	6	5	53	20	C	67	34
(	40	7	6	54	21	D	68	35
)	41	8	7	55	22	E	69	36
*	42	9	8	56	23	F	70	37
+	43	10	9	57	24	G	71	38
,	44	11	:	58	25	H	72	39
-	45	12	;	59	26	I	73	40
.	46	13	<	60	27			

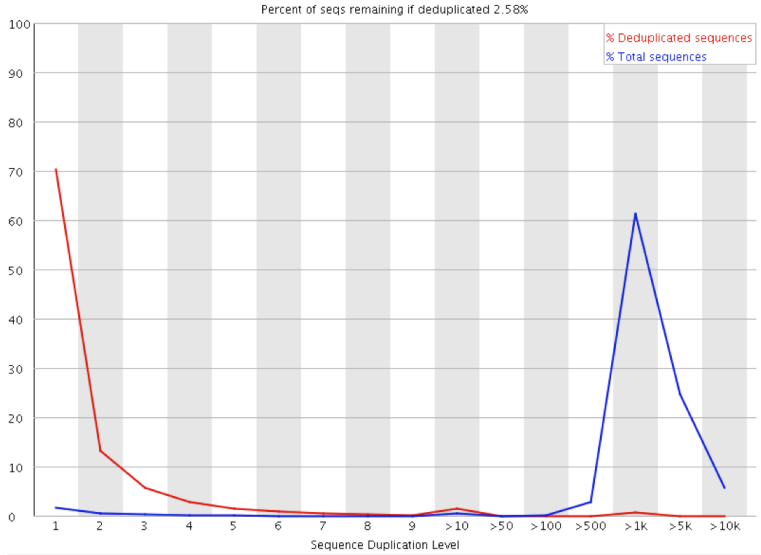
NGS variant calling bioinformatic pipeline

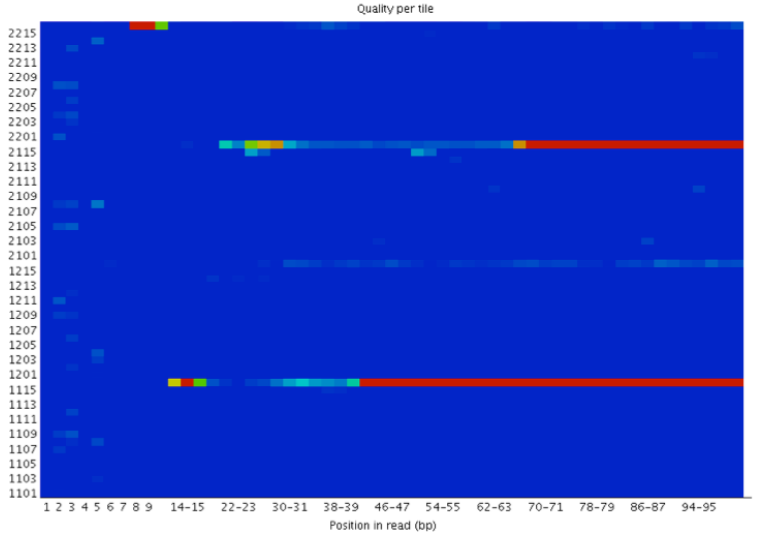


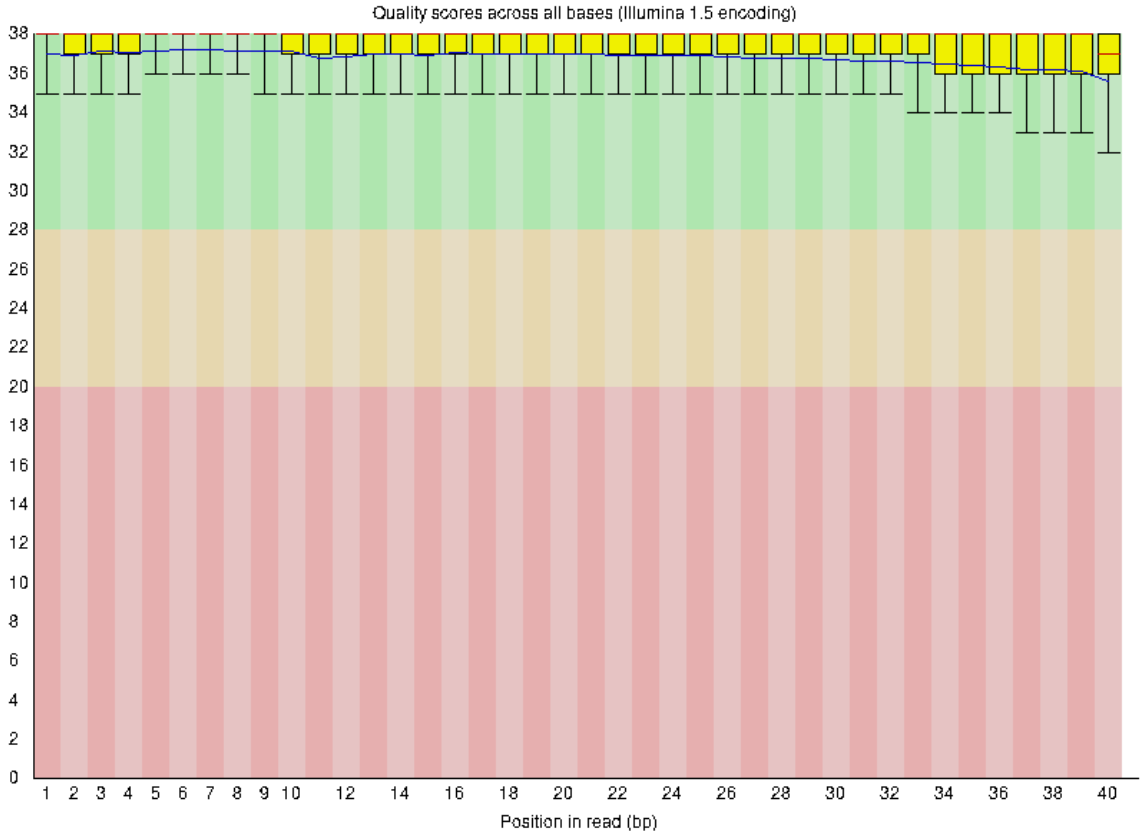
HELMHOLTZ MUNICH

https://doi.org/10.3389/fgene.2020.544162

FASTQ QC

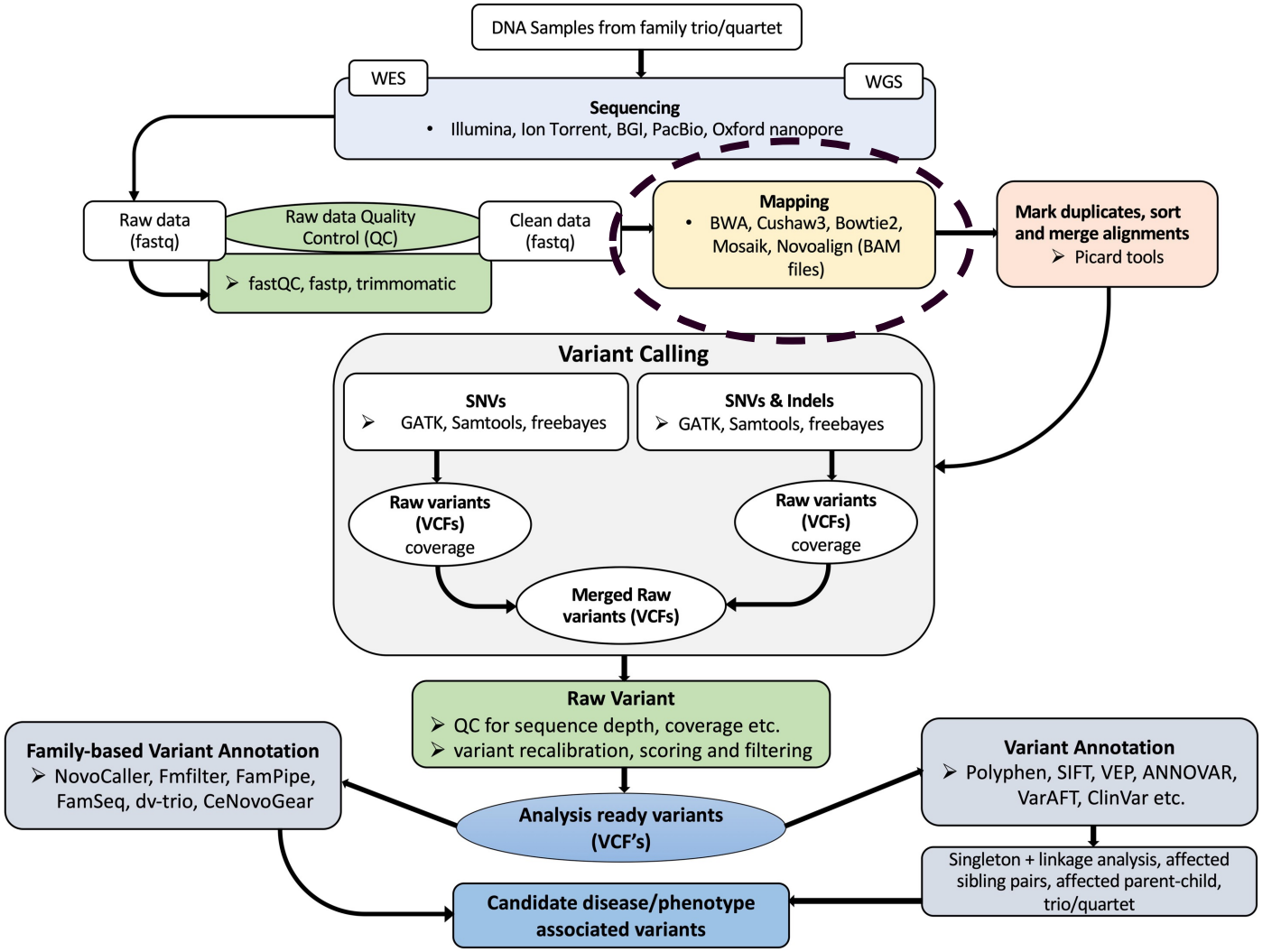






HELMHOLTZ MUNICH

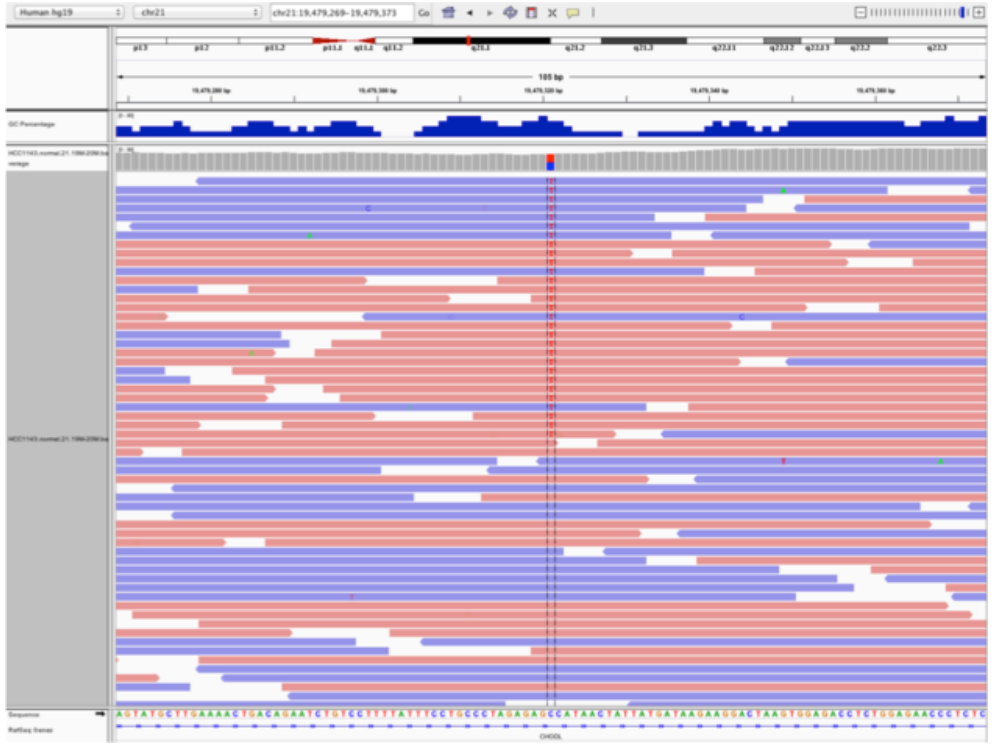
NGS variant calling bioinformatic pipeline



HELMHOLTZ MUNICH

https://doi.org/10.3389/fgene.2020.544162

Alignment or assembly

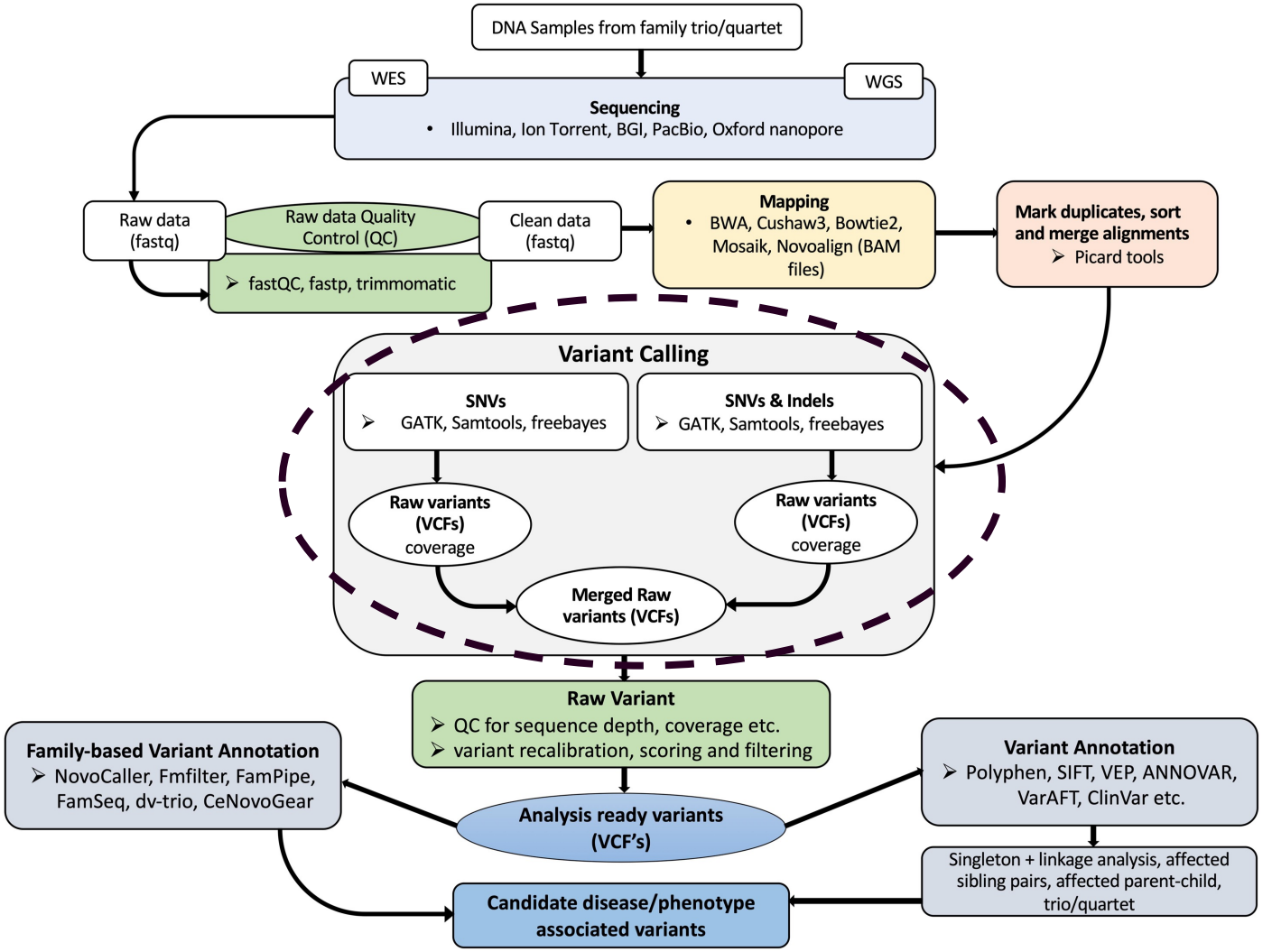


- Need to know where the reads map to the genome
- Takes into account insert size, read orientation and sequence similarity
- If the read does not match entirely: is it at the wrong place? A sequencing error? A SNP?

  → Probabilistic process → MAPQ
- Alignment requires a reference genome to align to

  → Available for humans but not for all the species
- Example of software: BWA, Bowtie2

NGS variant calling bioinformatic pipeline



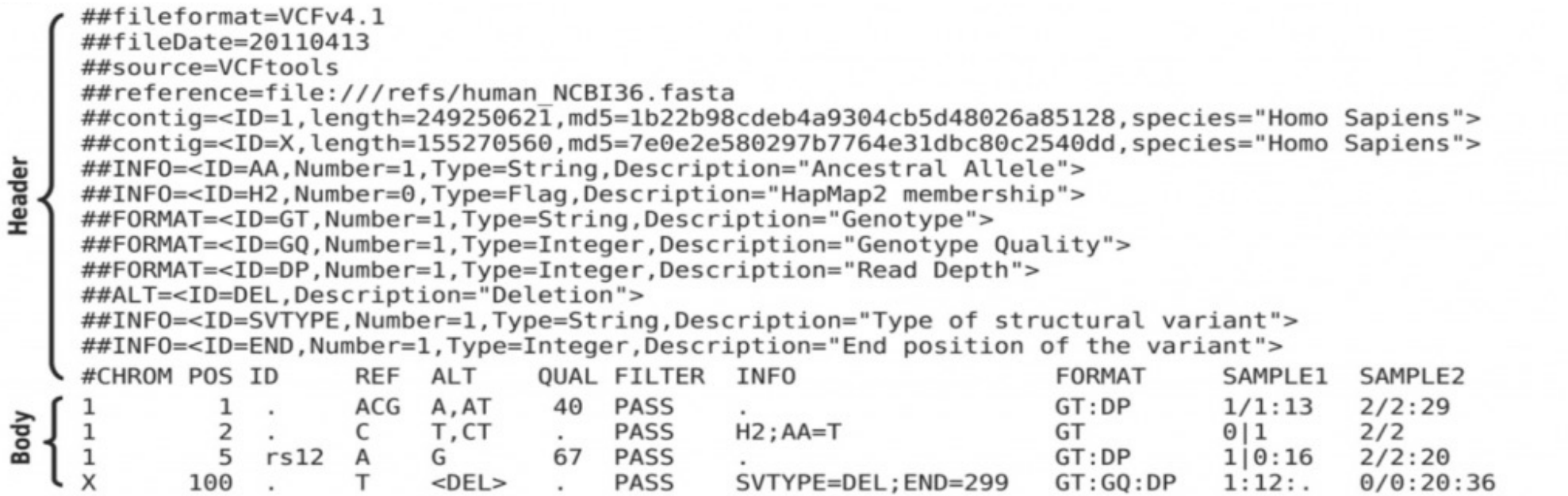
HELMHOLTZ MUNICH

https://doi.org/10.3389/fgene.2020.544162

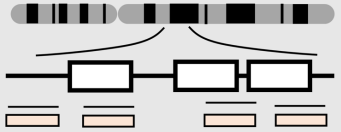
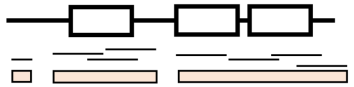
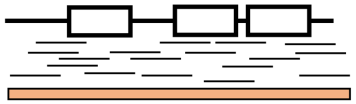
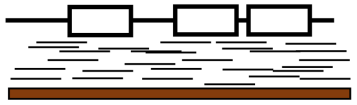
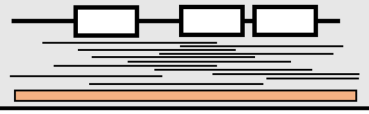
Variant calling

- Step to determine the genotypes of the individuals
- Output format = VCF which will be used for association analyses
- Quality measures:
  - QualByDepth (QD)
  - FisherStrand (FS)
  - StrandOddsRatio (SOR)
  - RMSMappingQuality (MQ)
  - MappingQualityRankSumTest (MQRankSum)
  - ReadPosRankSumTest (ReadPosRankSum)
- Variant Quality Score Recalibration (VQSR) – ML to flag poor quality SNV
- Measure of overall quality → Ti/Tv

VCF file format



Sequencing approaches

Sequencing Strategy		Pros	Cons
	Ultra low depth <0.5x	Can use off-target reads from exome sequencing to provide frequencies in understudied populations	Can't reliably provide individual-level genotypes
	Very low depth 0.5-2x	Works best in related samples	Heavy computational requirements
	Low depth 2-8x	Can provide haplotypes for imputation and variant-site discovery	Requires computing and imputation to estimate individual-level calls
	Medium depth (8-25x)	Reasonably high genotype accuracy and capture of rare variants relative to cost	Incomplete capture of indels and very rare variants (e.g., singletons)
	Exome (>20x at target)	Highly accurate genotype calls in covered coding regions - the most directly interpretable portion of the genome	Little coverage outside of coding regions, small fraction of genes missed, harder to call indels and structural variants. Some challenges comparing across capture methods
	High depth (>25x)	Best coverage of sequencable genome, rare variants, and high accuracy of genotype calls	More expensive, doesn't detect structural variants (as previous approaches)
	Long-read	Detection of structural variants, better assembly, mapping and phasing	Data processing, most expensive approach

Box 1: Overview of the different sequencing techniques currently available. White boxes correspond to coding exons and thin black lines to sequencing reads. The sequencing depth is represented at the bottom of each graphic by brown shades. Pros and cons of the different depths and genome coverage are highlighted.

HELMHOLTZ MUNICH

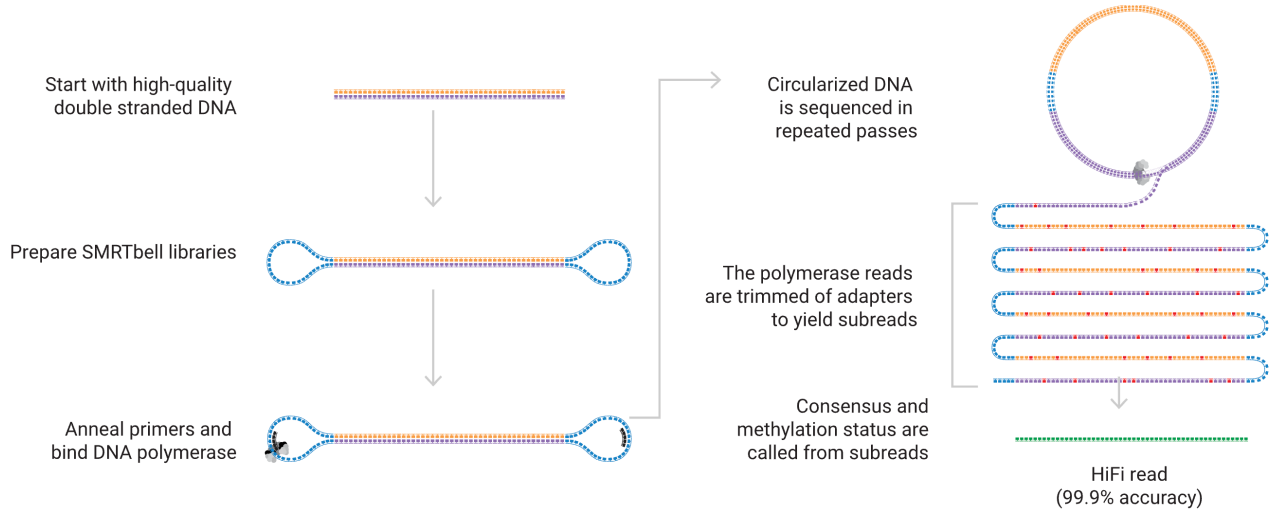
Bocher et al., 2023, Nat Comm

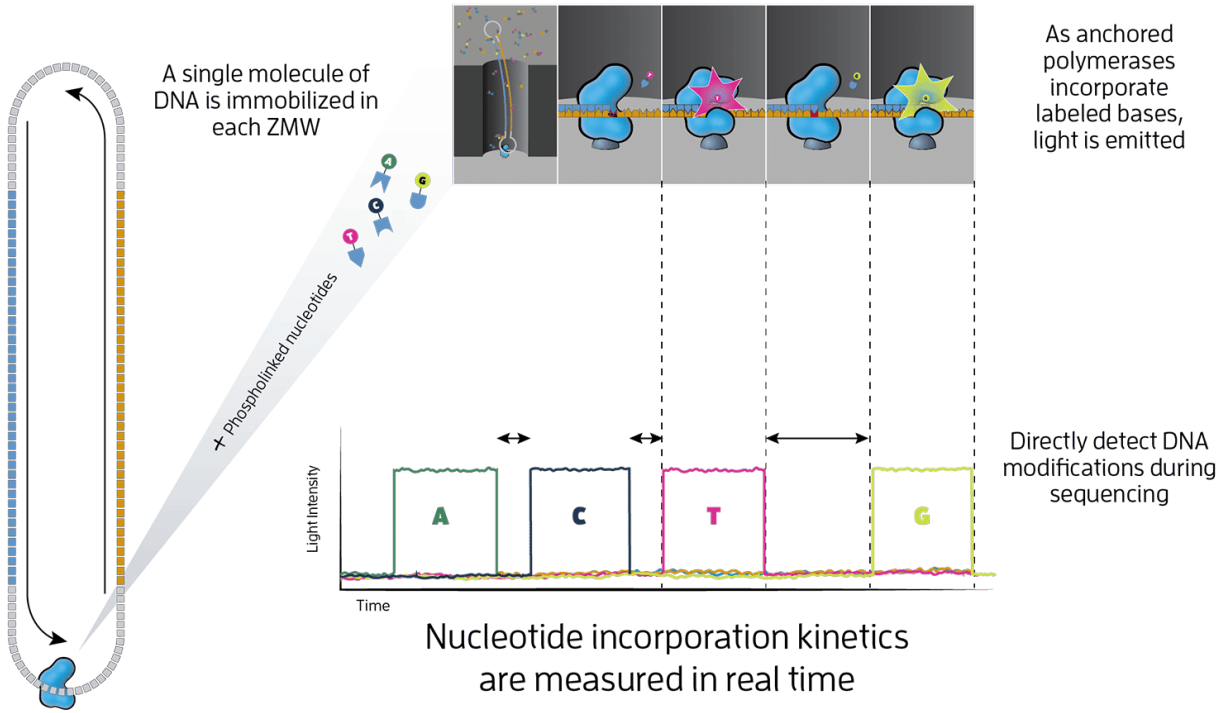
3

Third-generation sequencing

Pacific Biosciences (PacBio): Long-read sequencing

- Reads are > 1kb long
- Accuracy is > 99.9%, approaching Illumina and Sanger sequencing quality
- Particularly suited for *de novo* genome assembly or difficult to map regions





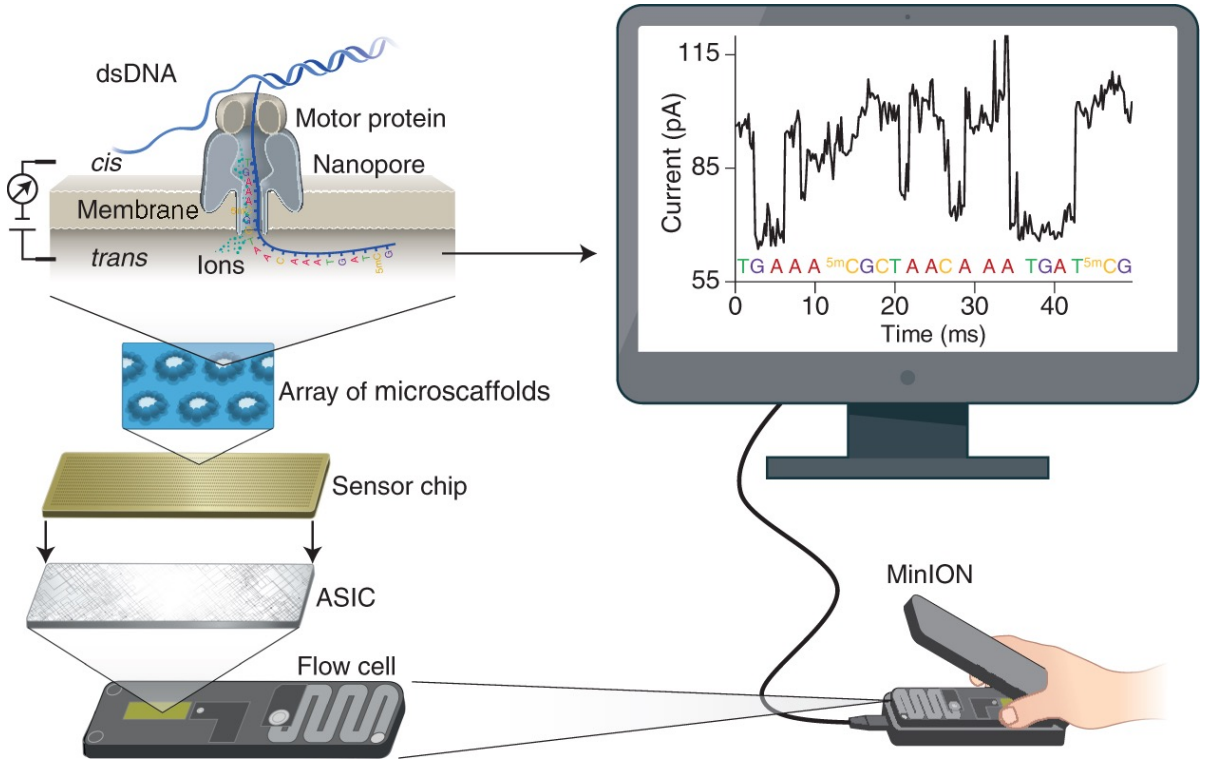
HELMHOLTZ MUNICH

Nanopore

As a nucleotide passes through the pore, it disrupts a current that has been applied to the nanopore. Each nucleotide provides a characteristic electronic signal that is recorded as a current disruption event.

Particularly suited to sequence environmental and metagenomic samples, and repetitive sequences.

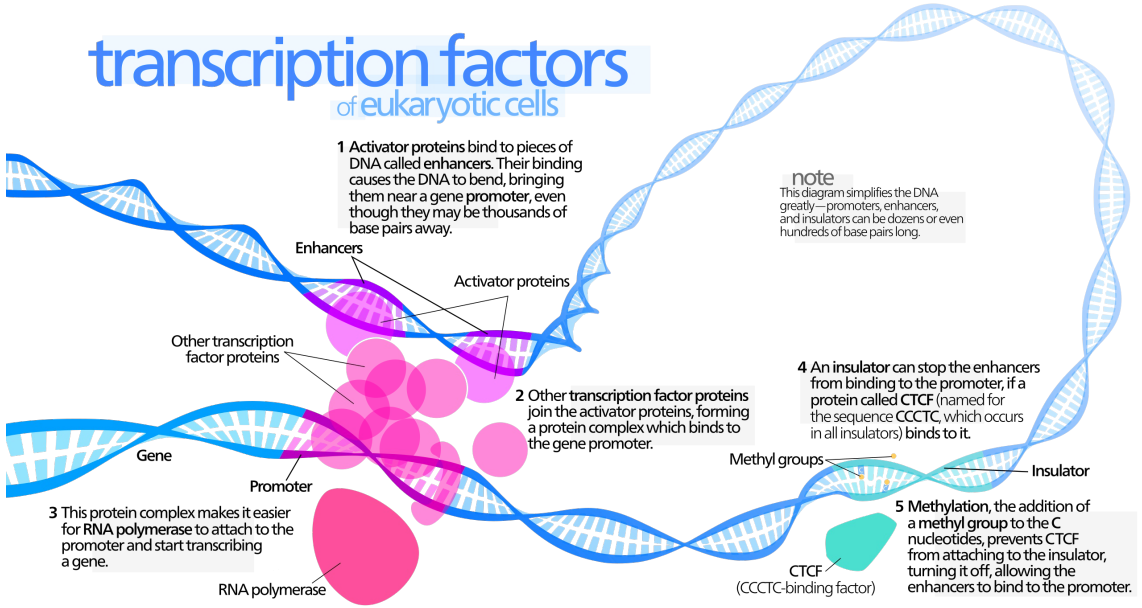
Also, now used to study RNA splicing and DNA methylation



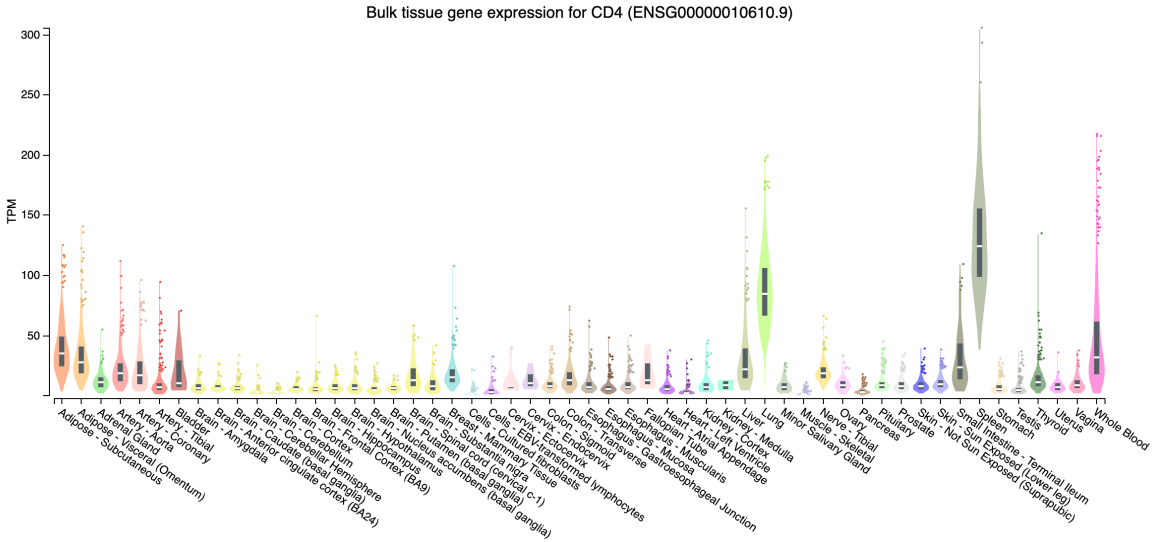
4

RNA-sequencing

Why using RNA-sequencing



- Previously, we wanted to assess DNA variation, which is common to the vast majority of all cells
- All genes are not expressed the same in all cells
- Each cell type is defined by an RNA expression profile
- Some diseases are characterised by a change of RNA expression patterns in affected tissue
- RNA measurements are cell-specific

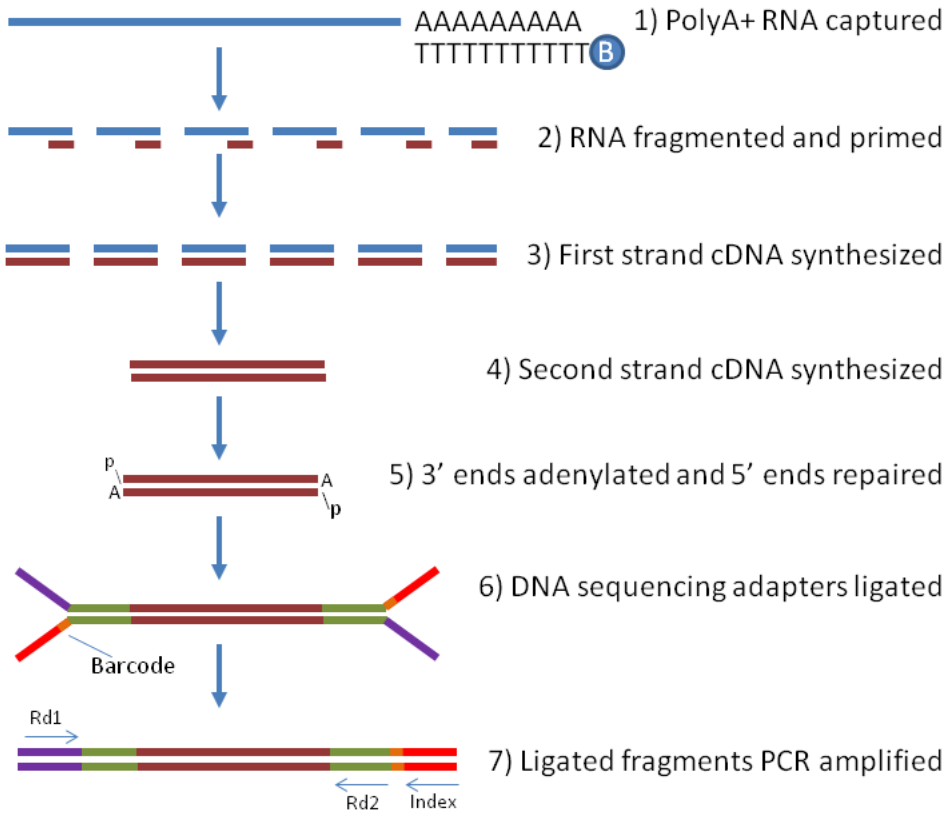


HELMHOLTZ MUNICH

RNA-sequencing

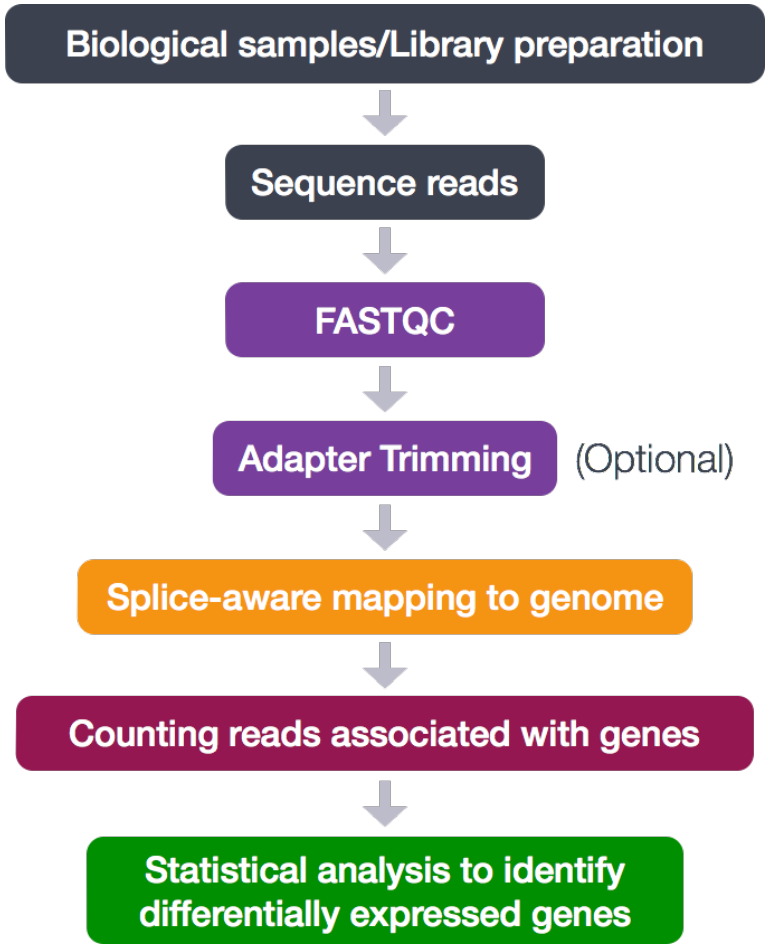
RNA selection

- Removing genomic DNA with DNAse
- We may be primarily interested in mRNA (protein-coding)
- But they only account for 1-5% of RNA pool (rest is ribosomal DNA and ncRNA)
- mRNA and most lncRNA contain a poly-A tail
- Selection through hybridisation
- If focus on noncoding RNA or degraded sample:
- Removal of ribosomal DNA by hybridisation to sequence-specific probes



RNA-sequencing pipeline

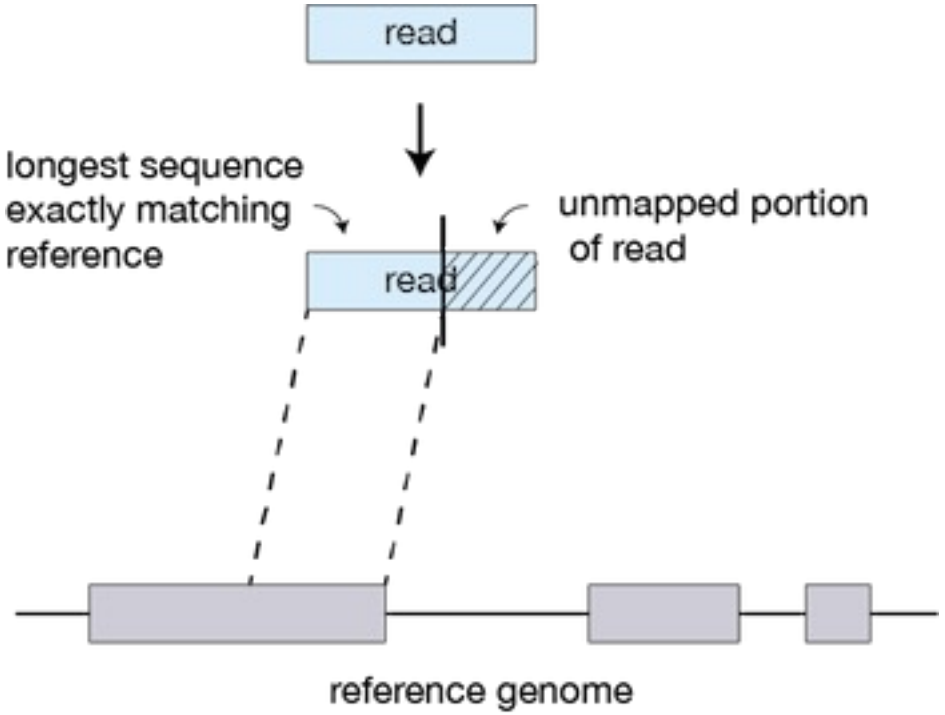
- Splice-aware alignment – STAR2
- Feature counts - HTSeq
- We are interested in comparing read count per gene isoform between tissues (Differential analysis – DESeq2, Limma, edgeR)
- But many factors influence the total number of reads mapping to a gene
  - GC content
  - Gene/isoform length
  - Experiment-specific variables (total number of reads)
  - …
- Complex read count normalisation is needed

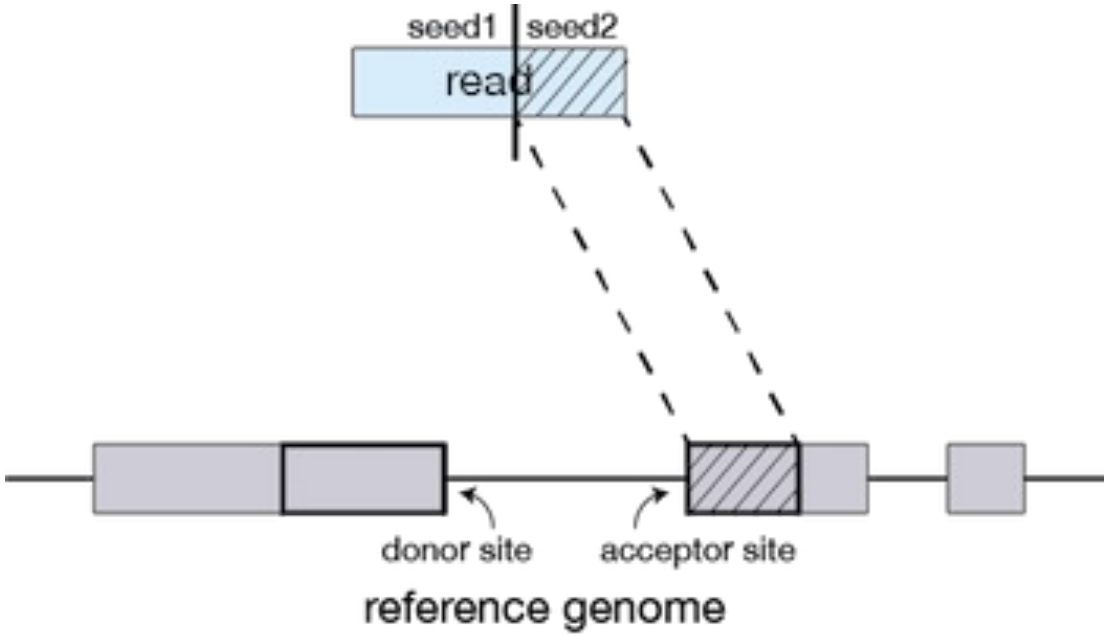


HELMHOLTZ MUNICH

https://hbctraining.github.io/Intro-to-rnaseq-hpc-O2/lessons/03_alignment.html

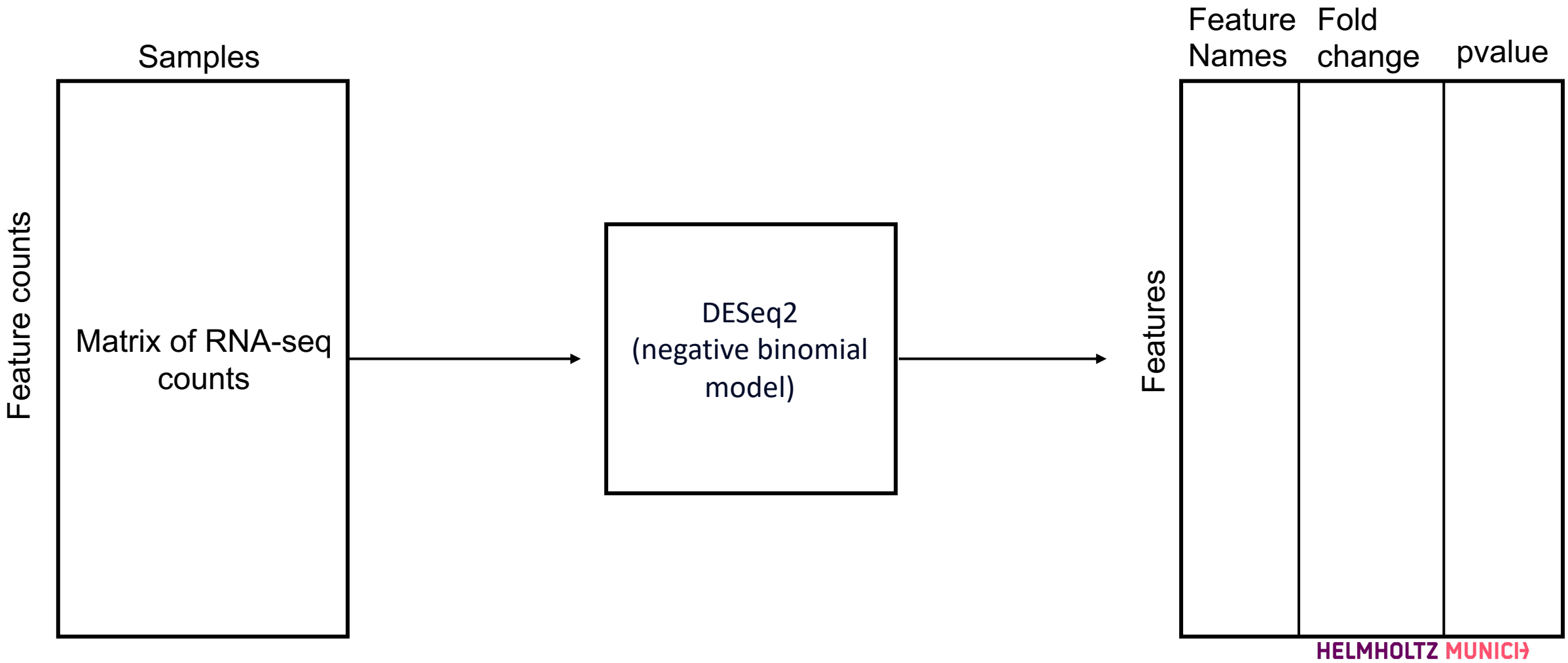
Splice-aware alignment – STAR2





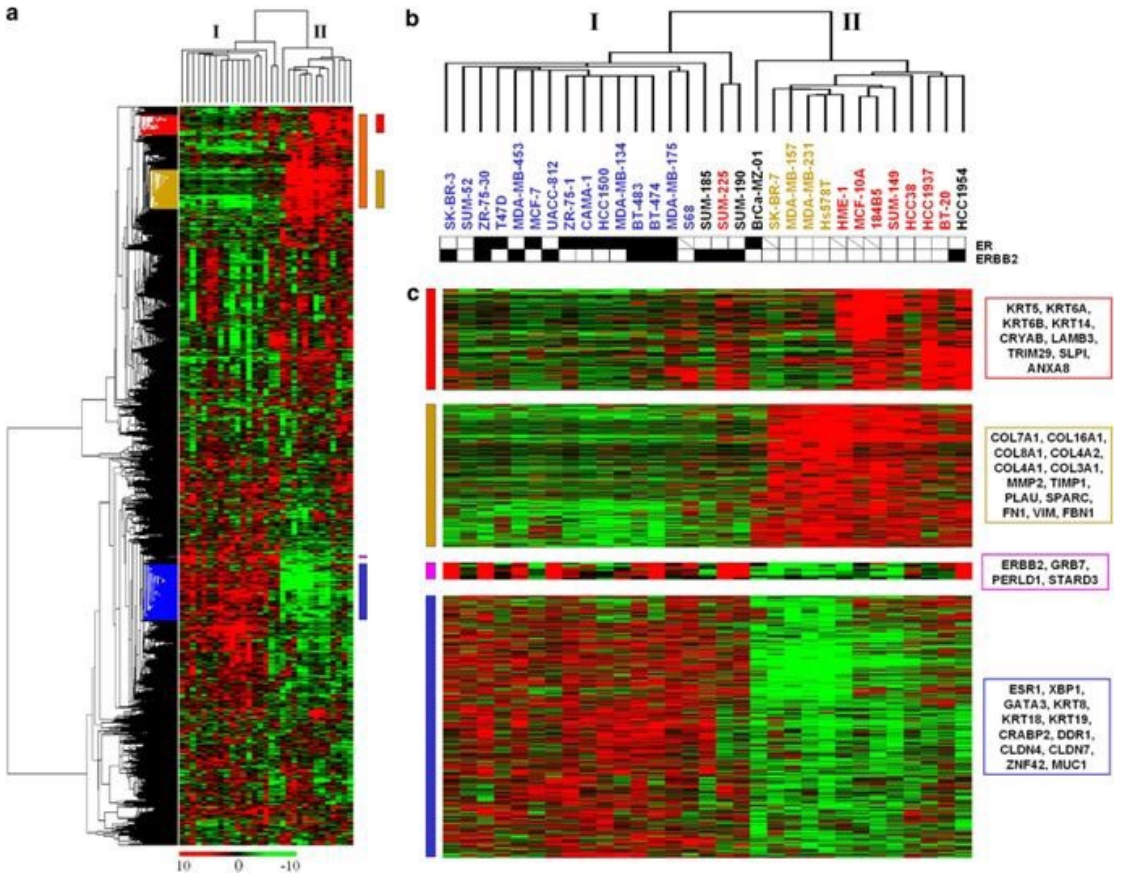
https://hbctraining.github.io/Intro-to-rnaseq-hpc-O2/lessons/03_alignment.html

Differential gene expression (DGE) analysis, typical input and output



Use genes identified from DGE for clustering

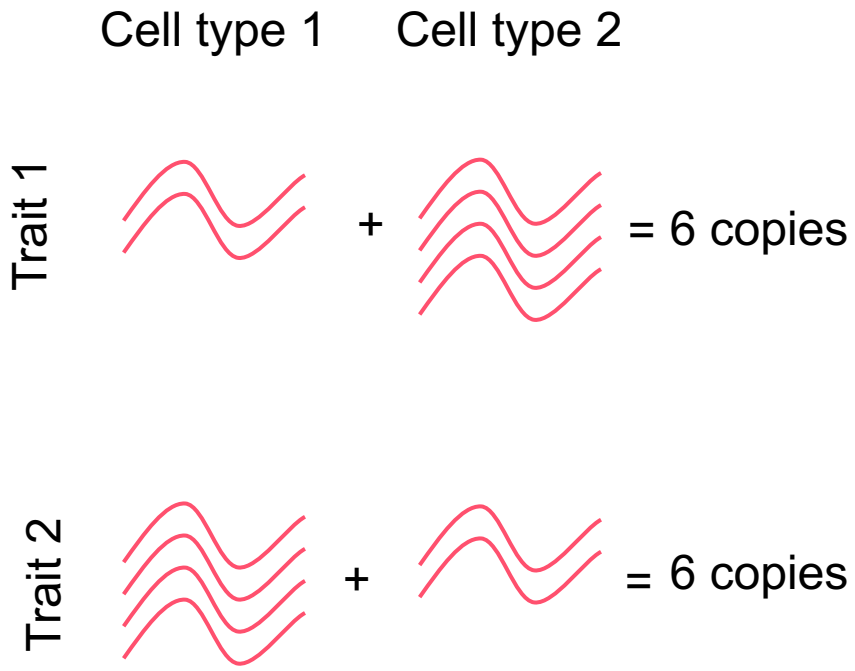
This is a typical use case scenario. The genes identified by DGE analysis are visualised with a heatmap. Hierarchical clustering, or other follow-up analyses, on the gene expression can also be used for separating the groups based on the identified genes.



5

Single-cell sequencing

Why single-cell sequencing

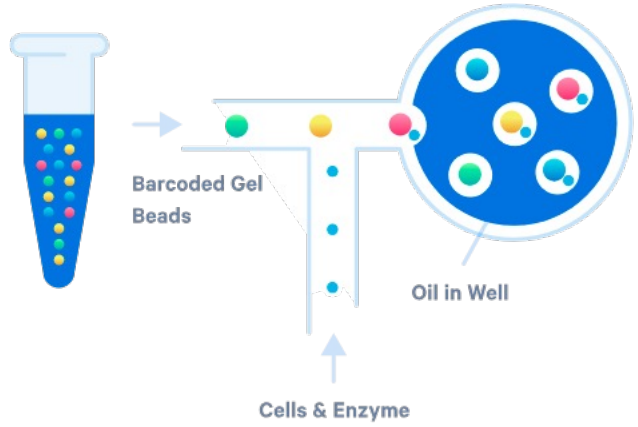


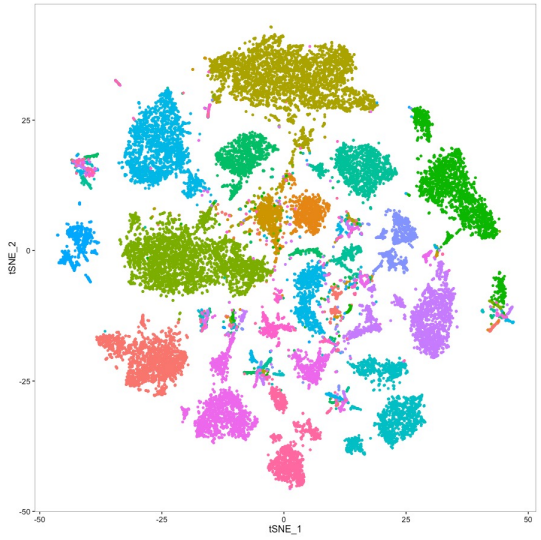
- Bulk RNA-sequencing is measuring the average number of copies of mixed cell population
- The disease causing gene might only be altered in a specific cell type
- For DNA sequencing, only a specific cell type might carry the mutation, for example cancer vs healthy tissue

Single-cell (RNA) sequencing: 10X genomics

Main differences

- Single cells are partitioned into Gel Beads-in-emulsion (GEMs) inside the microfluidic chip.
- In the chip, barcoded gel beads, cells and partitioning oil are combined.
- To establish single cell resolution and minimize doublets (two cells in a GEM), a limiting cell dilution is used. By doing this, only 1-10% of GEMs will contain a cell, whereas 90-99% of GEMs will remain empty. Because of this low percentage, it is necessary to load an accurate number of cells.





Limitations:

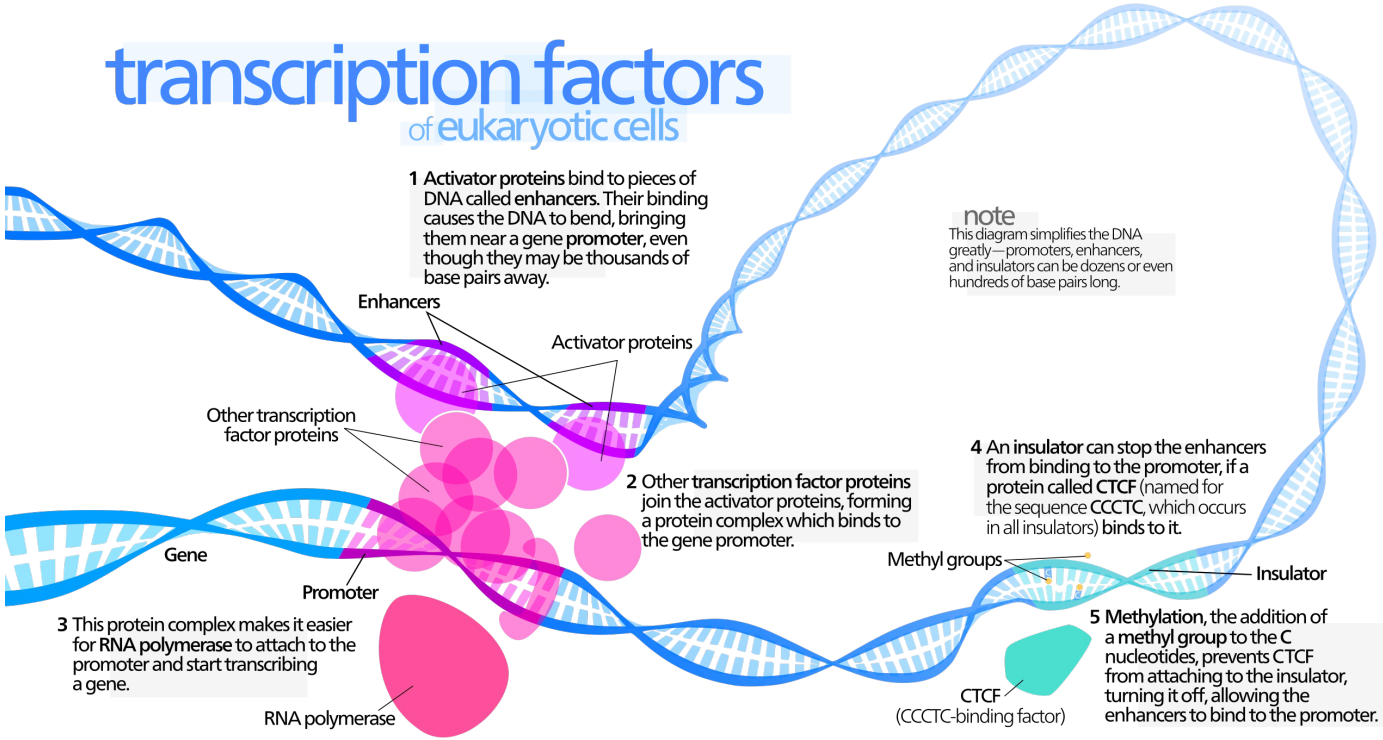
- Gene coverage is limited
- Complex bioinformatic pipeline
- Expensive
- Accurate cell type annotation requires manual curation

Functional genomics

We looked at how to study DNA variation and gene expression. NGS can be applied to the study of all other functional modifications:

- DNA methylation (bisulfite sequencing)
- Histone modification (ChIP-seq)
- Chromatin conformation (Hi-C)
- Open Chromatin (ATAC-seq)
- Protein levels (Proteomics)

The aim is to combine them to understand the context specific modifications that lead to a specific phenotype, for example our disease of interest.



Thank you.