Was leisten moderne Sequenzierverfahren – und wie analysieren wir diese Daten?

Benedikt Brors, Div. Applied Bioinformatics
DKFZ, DKTK & NCT Heidelberg

DKTK German Cancer Consortium

NCT NATIONAL CENTER FOR TUMOR DISEASES

Disclaimer

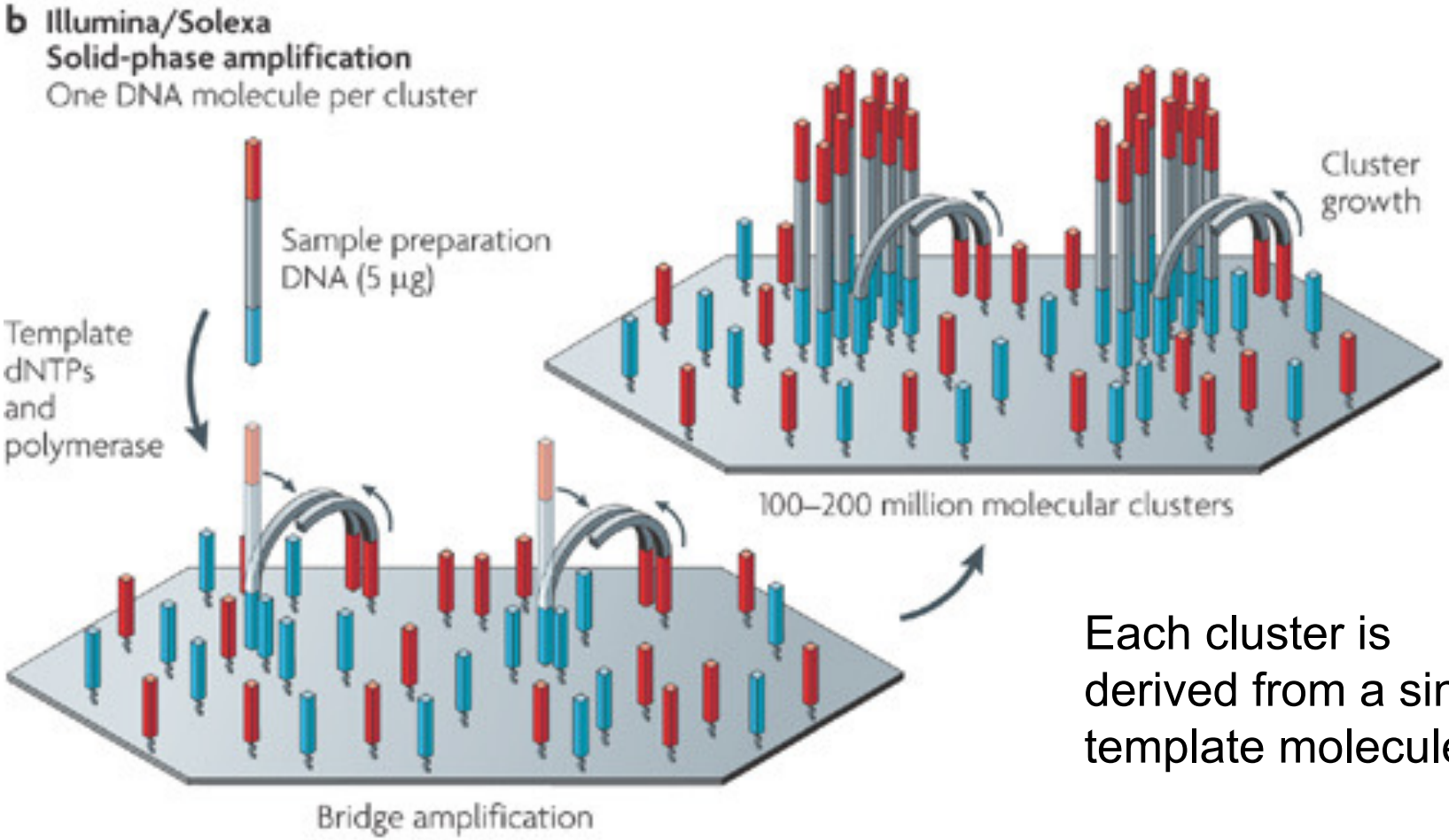
- Based on material from:
  - Matthias Schlesner
  - Charles Imbusch
  - Naveed Ishaque
  - Siao-Han Wong
  - Lena Voithenberg
  - Malte Simon
  - Niklas Beumer
  - Tobias Bauer
  - Matthias Bieg
  - Carl Herrmann
  - et al.

Next-generation sequencing

- Ultra-hochparalleles Verfahren
- DNA wird in kleine Abschnitte gespalten (Ultraschall)
- Fragmente werden an Oberflächen gebunden und lokal amplifiziert
- durch „Sequencing by synthesis“ werden schrittweise fluoreszent markierte Nukleotide eingebaut
- Können durch Bildanalyse sequentiell erfasst werden
  - => Sequenzen der einzelnen „Cluster“
- Es gibt 2- und 4-Farb-Verfahren
- Sehr hoher Durchsatz (TBasen in wenigen Stunden)

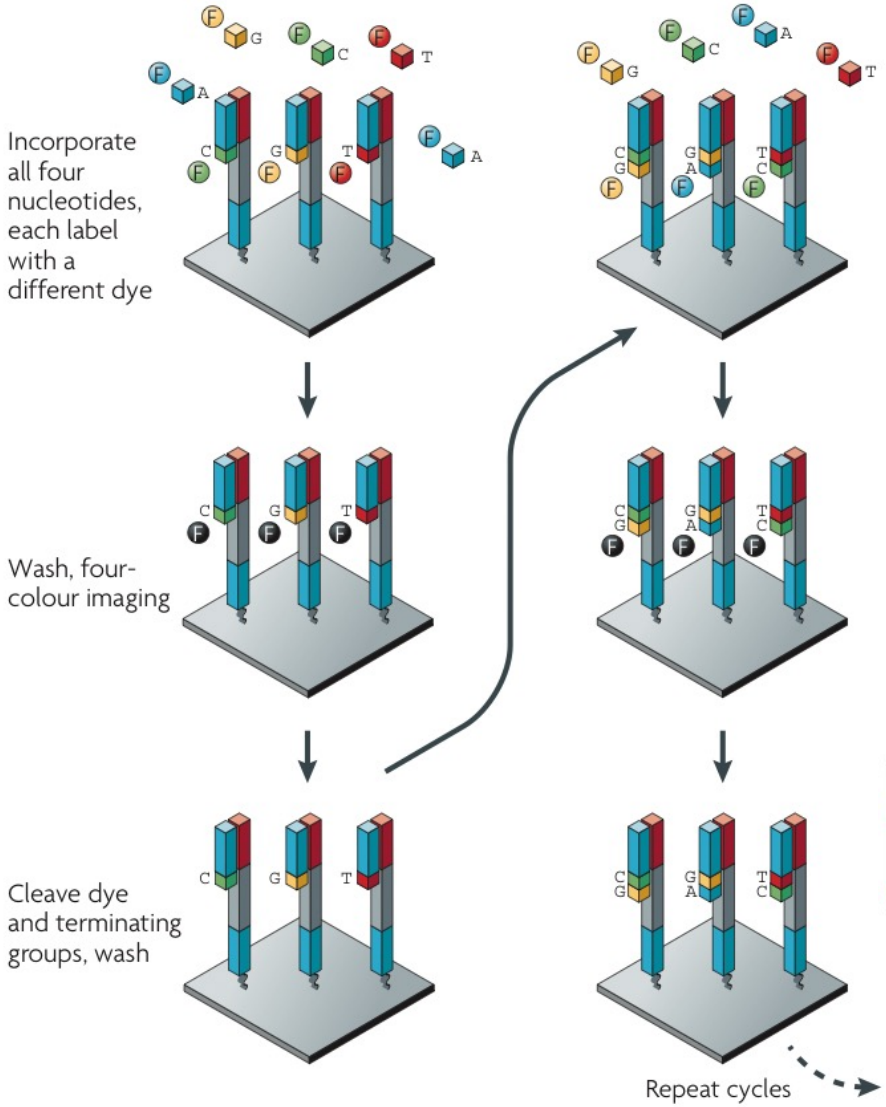
- Alternativen: Einzelmolekülsequenzierung (lange Sequenzen)

Illumina: Solid-Phase Amplification

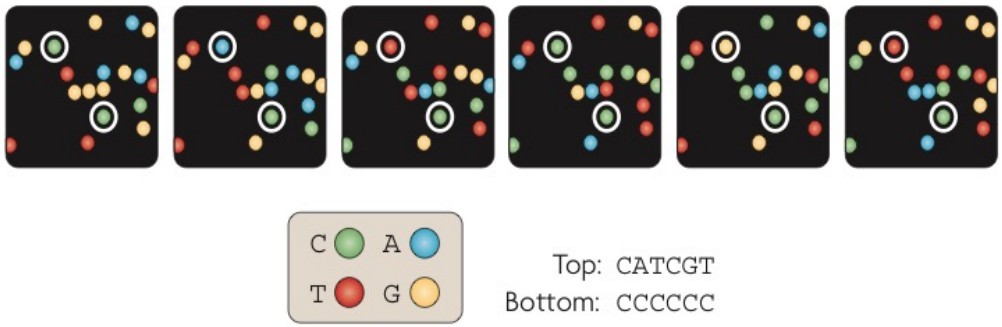


Each cluster is derived from a single template molecule

Illumina: Sequencing by Synthesis



- Reversible terminator chemistry
- Cycles:
  - Add labeled nucleotides
  - Excite and detect light emission
  - Remove dye and blocking group
- Sequence of images yields DNA sequence



Verwirrende Vielfalt

DNA-Sequenzierung

- whole-genome
- whole-exome
- targeted sequencing
- ChIP-seq
- ATAC-seq
- DNaseI-seq
- Cut&Run
- 3C, 4C, 5C, Hi-C

DNA-Methylierung

- Bisulfite-seq
- RRBS
- MeDIP-seq
- NOMe-seq
- NMT-seq

RNA-Sequenzierung

- mRNA-seq
- totalRNA-seq
- smallRNA-seq
- CAGE

Einzelzell-Sequenzierung

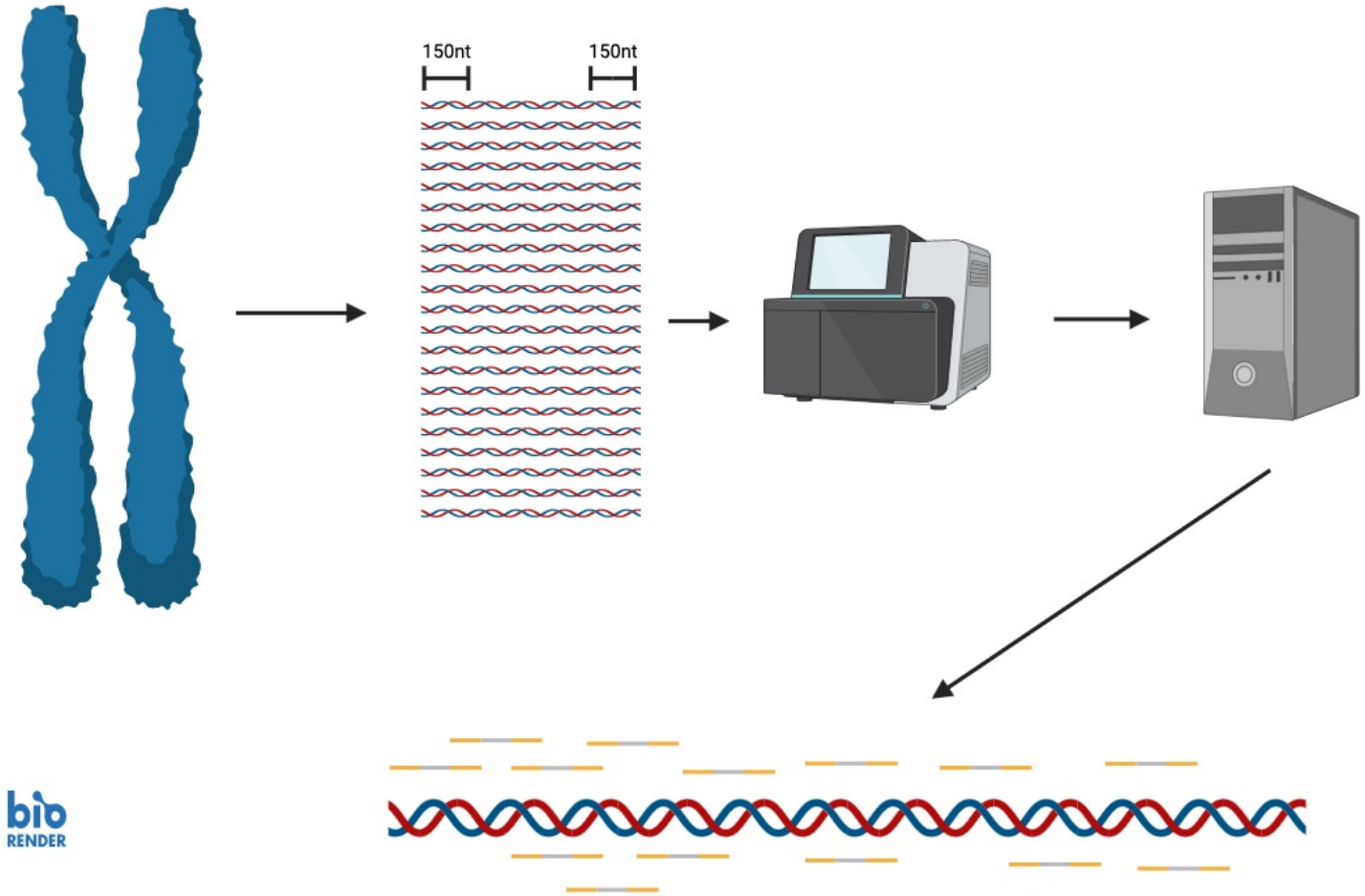
- scRNA-seq (3‘, 5‘, 10x, SMART-seq)
- scATAC-seq
- scNMT-seq
- scRRBS
- CITE-seq
- V(D)J-seq

Genomsequenzierung

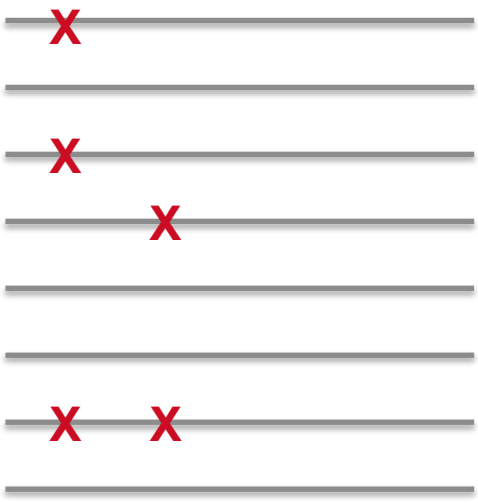
- Genom ist noch nicht bekannt: de-novo Sequenzierung
  - Herausforderung: Genom assemblieren

- Genom ist schon bekannt: Resequenzierung
  - genetische Polymorphismen
    - Einzelnukleotid-Polymorphismen: SNP
    - Kopienzahl-Polymorphismen (STR)
    - strukturelle Polymorphismen
  - somatische Mutationen (z.B. bei Krebs)
  - Immunphänotypen (TCR/BCR Sequenzen, HLA-Typing)
  - Mikrobiom: Vielfalt der Organismen in einer bestimmten Umgebung
    - ribo-tag: angereichert für rRNA
    - shotgun

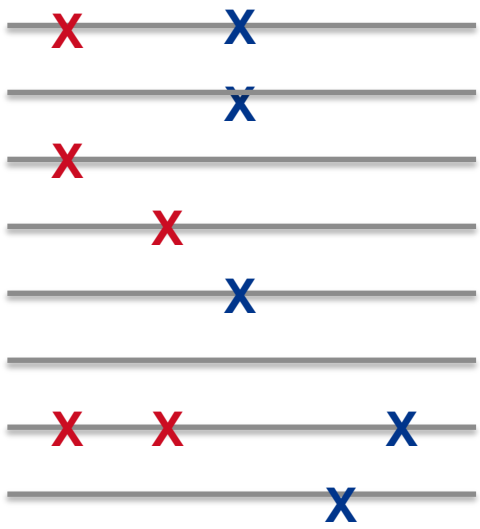
Mapping: Alignment zum Referenzgenom



Variantenanalyse

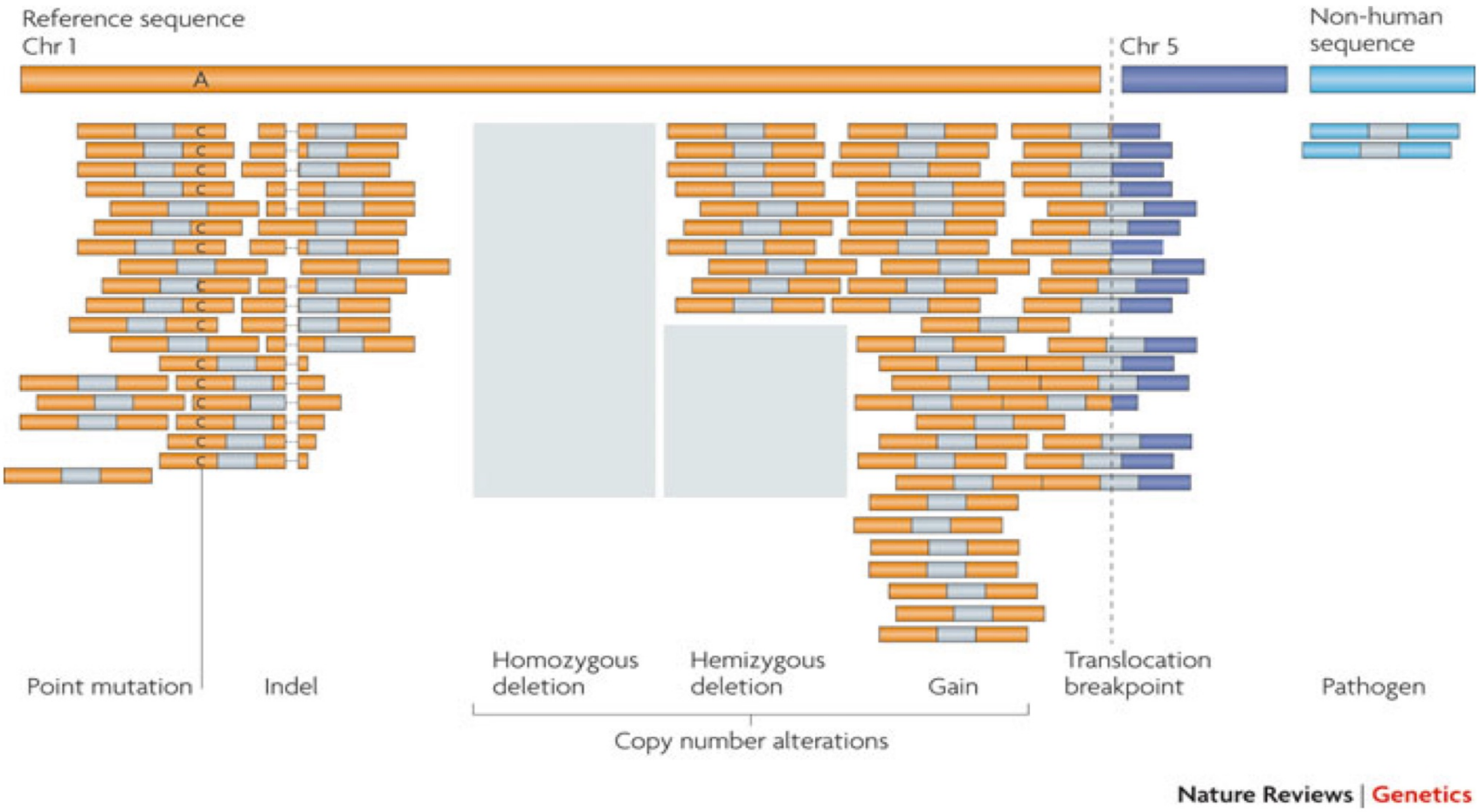


“Keimbahn“-Genom



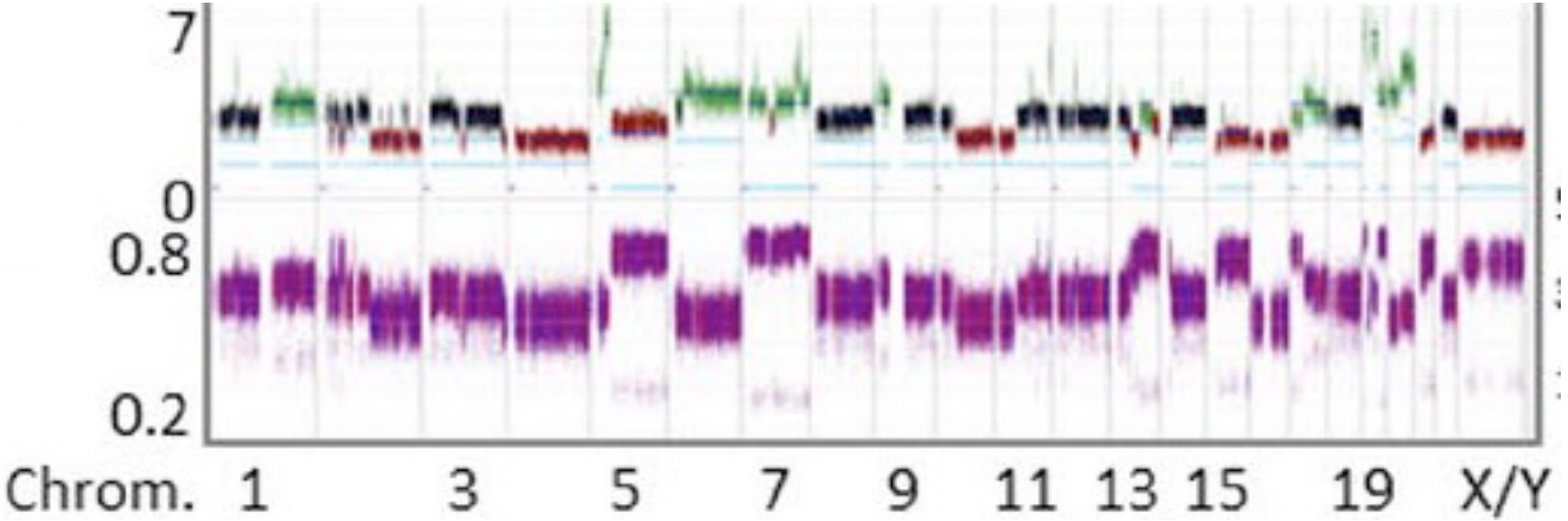
Tumor-Genom

Cancer genome sequencing



Meyerson, Nat Rev Genet 2010

Kopienzahlprofil



- Y-Achse: log2(reads_tumor / reads_normal)
- darunter: Allelfrequenzen

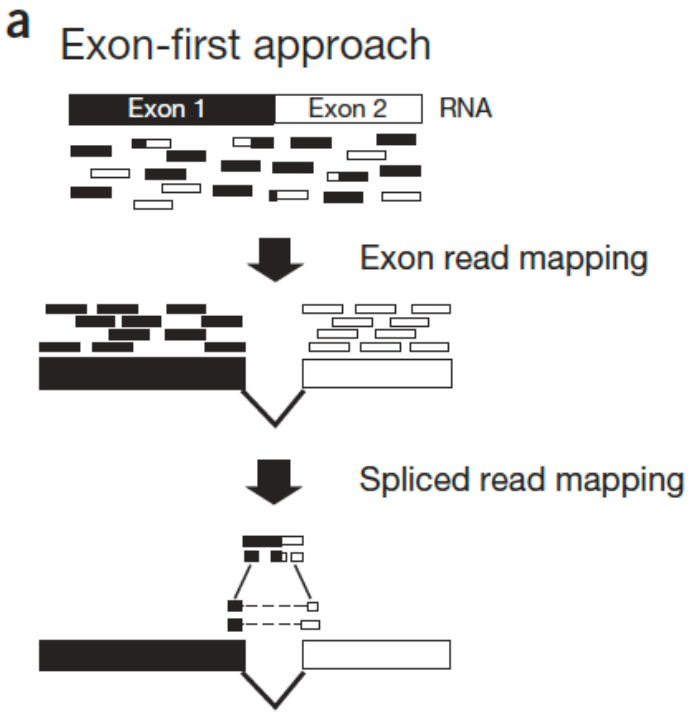
Giessler, Kleinheinz et al. J Exp Med 2017

RNA-Sequenzierung

- Mapping auf: Transkriptom
  - Problem: nicht alle mRNAs bekannt (z.B. seltenes Transkript, das nur an Tag 3,5 der Embryonalentwicklung exprimiert wird)
  - Isoformen durch alternatives Spleißen

- Mapping auf Genom
  - Problem: Exon/Intron-Struktur bei Eukaryonten
  - Sequenzfragmente mappen nur teilweise auf Genom
  - einfach: Reads, die vollständig in Exome mappen
  - schwierig: Intron-überspannende Reads
    - z.B. drei Basen Überhang, die 150kb entfernt aligniert werden müssten

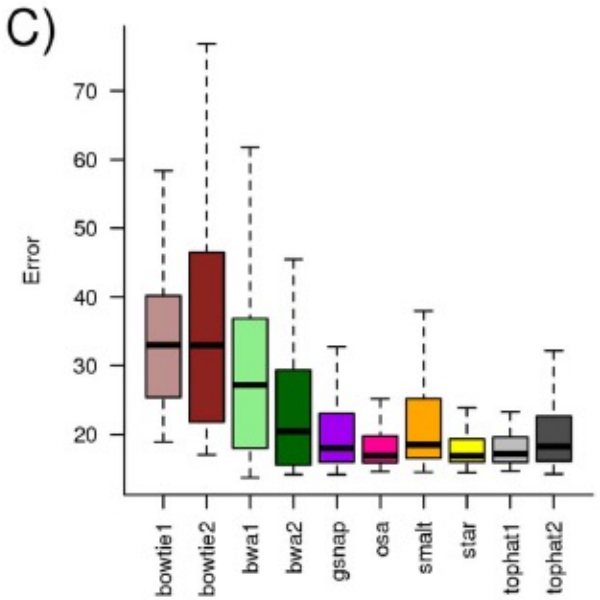
Spliced aligners



Garber et al. Nat Methods 2011

Spliced read aligners are better

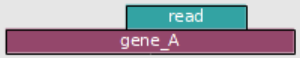
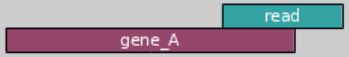
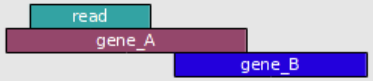
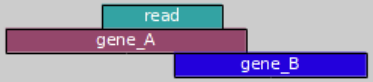
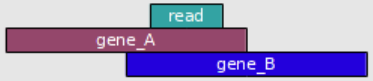
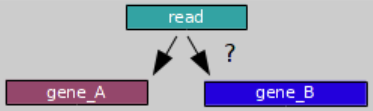
- Performance of spliced read aligners (e.g. Tophat, Tophat2, STAR) is superior to those of general aligners
- Mapping against genome preferable
- Transcriptome model (Refseq, ENSEMBL, GENCODE) can improve accuracy of alignments



Fonseca et al.
PLOS ONE 2014

Counting

- Different counting modes
- e.g., htseq-count
  - Union
  - Intersection strict
  - Intersection non-empty

	union	intersection_strict	intersection_nonempty
	gene_A	gene_A	gene_A
	gene_A	no_feature	gene_A
	gene_A	no_feature	gene_A
	gene_A	gene_A	gene_A
	gene_A	gene_A	gene_A
	ambiguous (both genes with --nonunique all)	gene_A	gene_A
	ambiguous (both genes with --nonunique all)		
	alignment_not_unique (both genes with --nonunique all)		

Quantification: at which level?

- Quantification can be on
  - Gene level
  - Exon level
  - Transcript level

- Transcript level would be optimal but needs transcript reconstruction (unstable)
- Gene level: easy to work with, but can mask differentially spliced transcripts
- Exon level: can at least detect differential exon usage

Expression measures: RPKM and FPKM

- RPKM: Reads per kilobase per million mapped reads
- FPKM: Fragments per kilobase per million mapped reads (for paired-end data)

$$\mathrm{RPKM}_g = \frac{r_g \times 10^9}{\mathrm{fl}_g \times R}$$

r_g: reads mapped for each gene
R: total number of mapped reads for the sample
fl_g: feature length of each gene

Explanation:

Normalize for gene length (”reads per kilobase”): $\frac{r_g}{\mathrm{fl}_g} 10^3$

Normalize for total number of reads (“per million Mapped reads”): $\frac{R}{10^6}$

$$\mathrm{RPKM}_g = \frac{\frac{r_g 10^3}{\mathrm{fl}_g}}{\frac{R}{10^6}} = \frac{r_g \times 10^9}{\mathrm{fl}_g \times R}$$

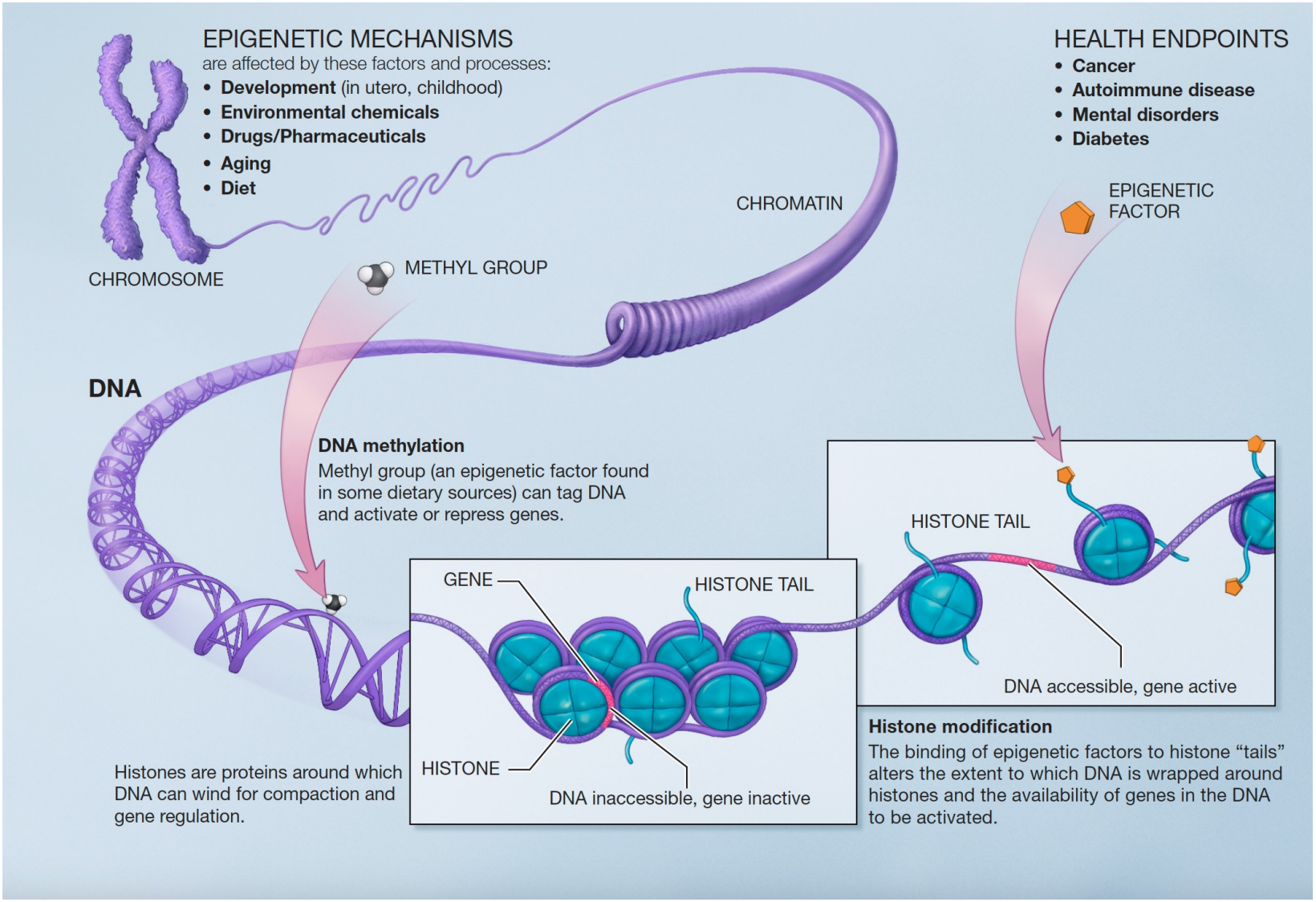
TPM: transcripts per million

$$\mathrm{TPM} = \frac{r_g \times \mathrm{rl} \times 10^6}{\mathrm{fl}_g \times T} \qquad T = \sum_{g \in G} \frac{r_g \times \mathrm{rl}}{\mathrm{fl}_g}$$

- r_g: number of reads for gene g
- l_r: read length
- fl_g : length of gene/transcript/exon
- T: total number of transcripts sampled in a sequencing run

- Proportional to RPKM, but with a sample-specific scaling factor; T estimate for #transcripts derived from #mapped reads per gene normalized by length of gene

Epigenetik



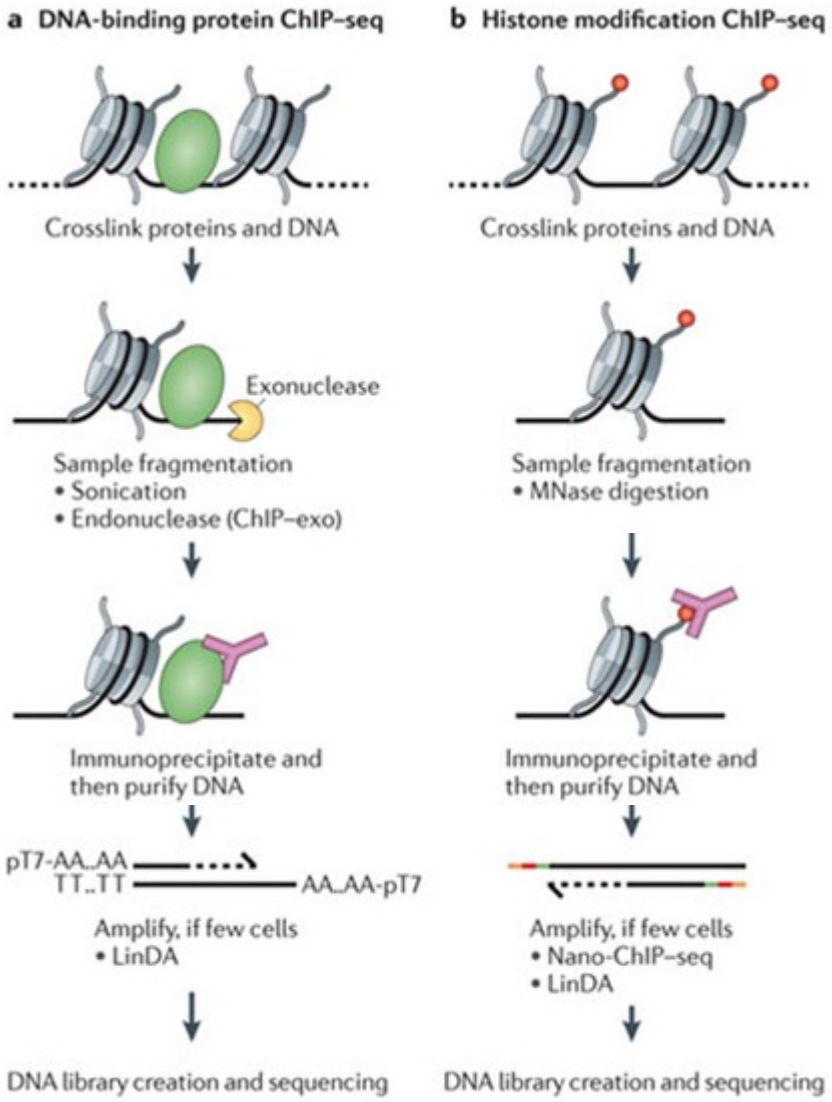
https://commonfund.nih.gov/epigenomics/figure

ChIP-seq

- Immunpräzipitation von DNA-bindenden Proteinen
- für alle DNA-bindenden Proteine, beispielsweise:
  - Transkriptionsfaktoren
  - Chromatin-Modifikatoren (z.B. Histonacetylasen)
  - spezifische Antikörper für Histonmodifikatoren verfügbar (H3K4me3, H3K27ac, ...)
- Geben Aufschluss über Bindestellen im Genom (Promoterarchitektur, Genregulation)
- Dichte der sequenzierten Fragmente => ‘Peaks‘

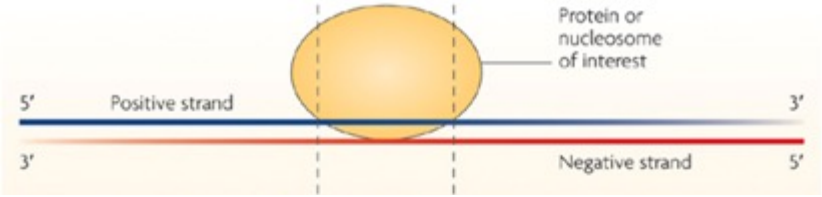
ChIPseq experimental protocol

- (Formaldehyde) Crosslinking
  - Fix DNA bound proteins
- Fragment DNA
  - Physical fragmentation
  - Enzymatic digestion
- Immunoprecipitate and purify
  - Using antibody against protein of interest
- Library preparation



Adapted from Park P.J., Nat Rev Genet 2009

ChIPseq – what do we see in the ‘reads’?



Peak identification - theory

- 5’ ends of fragments are sequenced

- Sequenced reads are aligned to reference genome

- Distribution of forward and reverse oriented tags is computed

- Profile is generated from fragment extension or read shifting

- ChIP signal is generated

Adapted from Park P.J., Nat Rev Genet 2009

ChIPseq downstream analysis

- Differential analysis

- Peak annotation
  - Genomic region
    - TSS, introns, exons, intergenic regions etc
  - Over represented motifs in peak regions

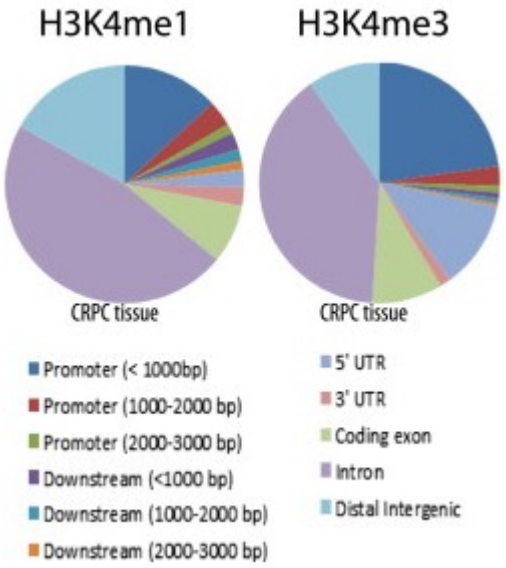
- Gene profile analysis
  - e.g. histone modification profiles over the gene body

- Profile clustering analysis

- Chromatin segmentation

ChIPseq peak annotation

- Example: different histone modifications occur at different genomic regions (using CEAS)



- Example: over enriched sequence motifs in peak regions (using HOMER)

Homer de novo Motif Results

Homer *de novo* Motif Results

Known Motif Enrichment Results
Gene Ontology Enrichment Results
If Homer is having trouble matching a motif to a known motif, try copy/pasting the matrix file into STAMP
More information on motif finding results: HOMER | Description of Results | Tips
Total target sequences = 37301
Total background sequences = 35962
* - possible false positive

Rank	Motif	P-value	log P-pvalue	% of Targets	% of Background	STD(Bg STD)	Best Match/Details	Motif File
1		1e-12661	-2.915e+04	70.91%	15.19%	40.5bp (65.1bp)	Foxa2(Forkhead)/Liver-Foxa2-ChIP-Seq/Homer More Information | Similar Motifs Found	motif file (matrix)
2		1e-578	-1.332e+03	27.14%	16.52%	54.0bp (65.5bp)	NF1-halfsite(CTF)/LNCaP-NF1-ChIP-Seq/Homer More Information | Similar Motifs Found	motif file (matrix)
3		1e-384	-8.860e+02	17.77%	10.53%	53.9bp (62.1bp)	Unknown/Homeobox /Limb-p300-ChIP-Seq/Homer More Information | Similar Motifs Found	motif file (matrix)
4		1e-164	-3.783e+02	3.17%	1.28%	52.2bp (62.9bp)	PH0048.1_Hoxa13 More Information | Similar Motifs Found	motif file (matrix)
5		1e-151	-3.485e+02	3.38%	1.47%	50.2bp (65.4bp)	NF-E2(bZIP)/K562-NFE2-ChIP-Seq/Homer More Information | Similar Motifs Found	motif file (matrix)
6		1e-107	-2.485e+02	1.21%	0.35%	56.3bp (69.7bp)	CTCF(Zf)/CD4+-CTCF-ChIP-Seq/Homer More Information | Similar Motifs Found	motif file (matrix)
7		1e-72	-1.671e+02	2.10%	1.02%	55.1bp (58.5bp)	MA0029.1_Evi1 More Information | Similar Motifs Found	motif file (matrix)

Adapted from Sharma N.L. et al, Cancer Cell 2012

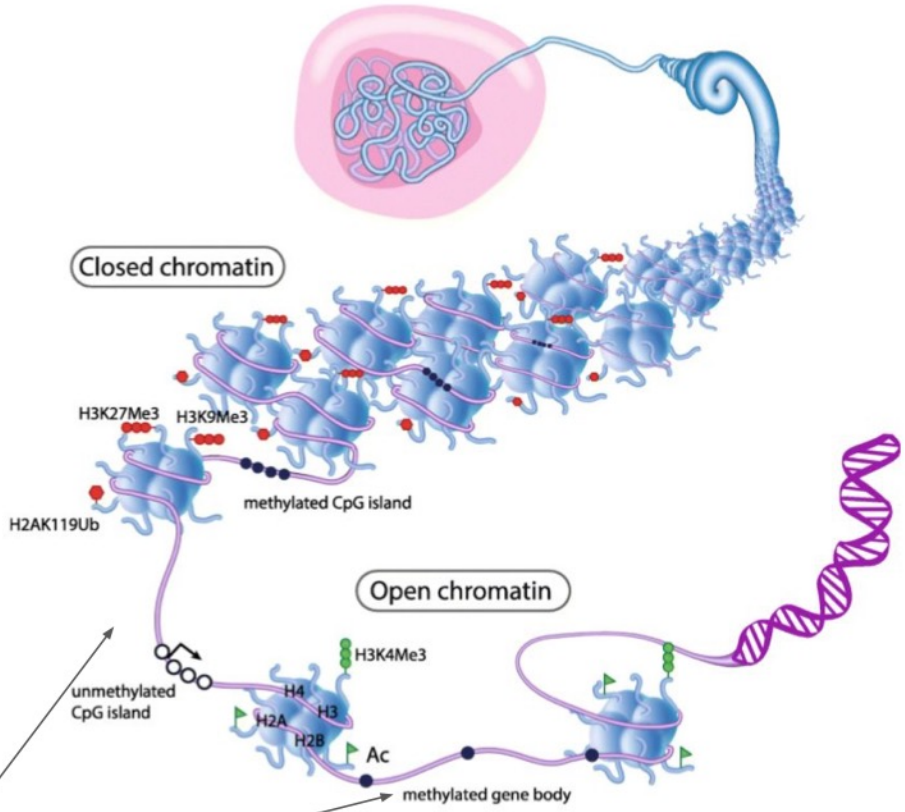
http://biowhat.ucsd.edu/homer/ngs/peakMotifs.html

ATAC-seq / DNaseI seq

- zeigen an, wo ‘offenes‘ oder ‘geschlossenes‘ Chromatin vorliegt
- ATAC-seq: Transposase inseriert Sequenzen präferentiell in offenes Chromatin (z.B. TFBS, aktiv transkribierte Regionen); Dichte der sequenzierten Fragmente => ‘Peaks‘
- DNaseI: zersetzt präferentiell offene Chromatin-Bereiche

ATAC-seq overview

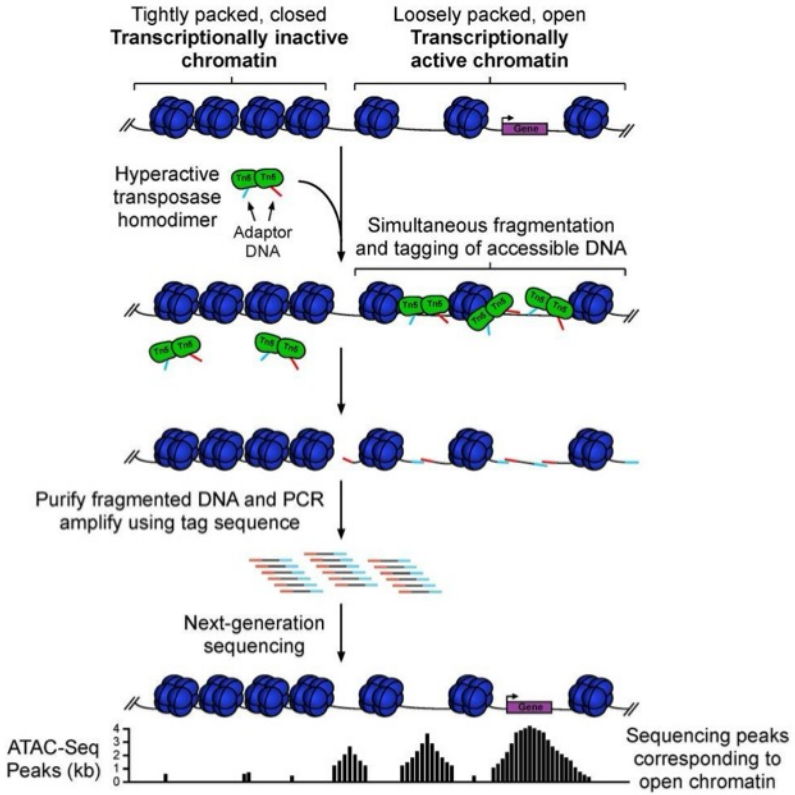
- **A**ssay for **T**ransposase-**A**ccessible **C**hromatin using **seq**uencing
- sequence DNA from accessible chromatin
- applications:
  - identify promoters, enhancers and silencers
  - reveal regulators of gene expression (such as transcription factors controlling accessibility of a set of genes)
  - map nucleosome positioning
  - ...

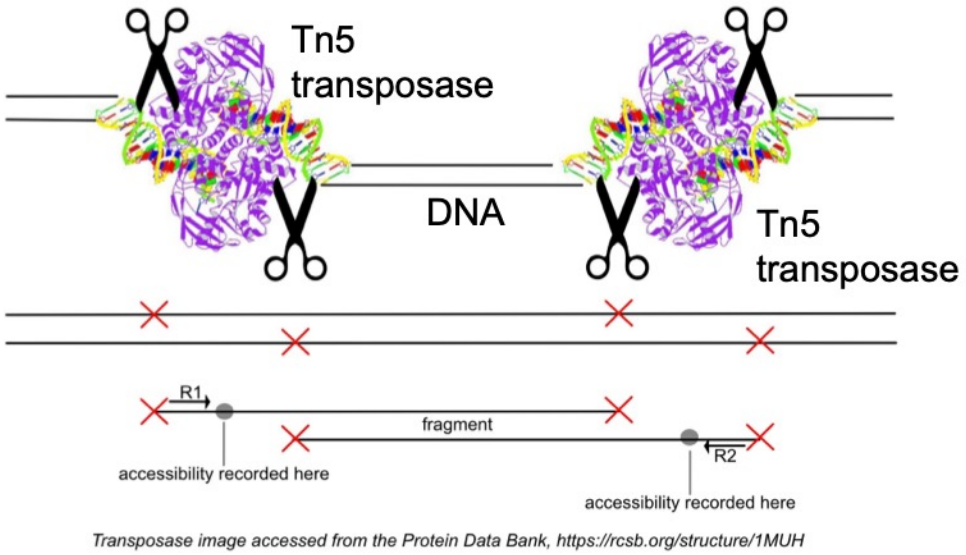


accessible DNA
that can be sequenced with ATAC-seq

Holliday et al., 2018

DNA transposition with Tn5 transposase

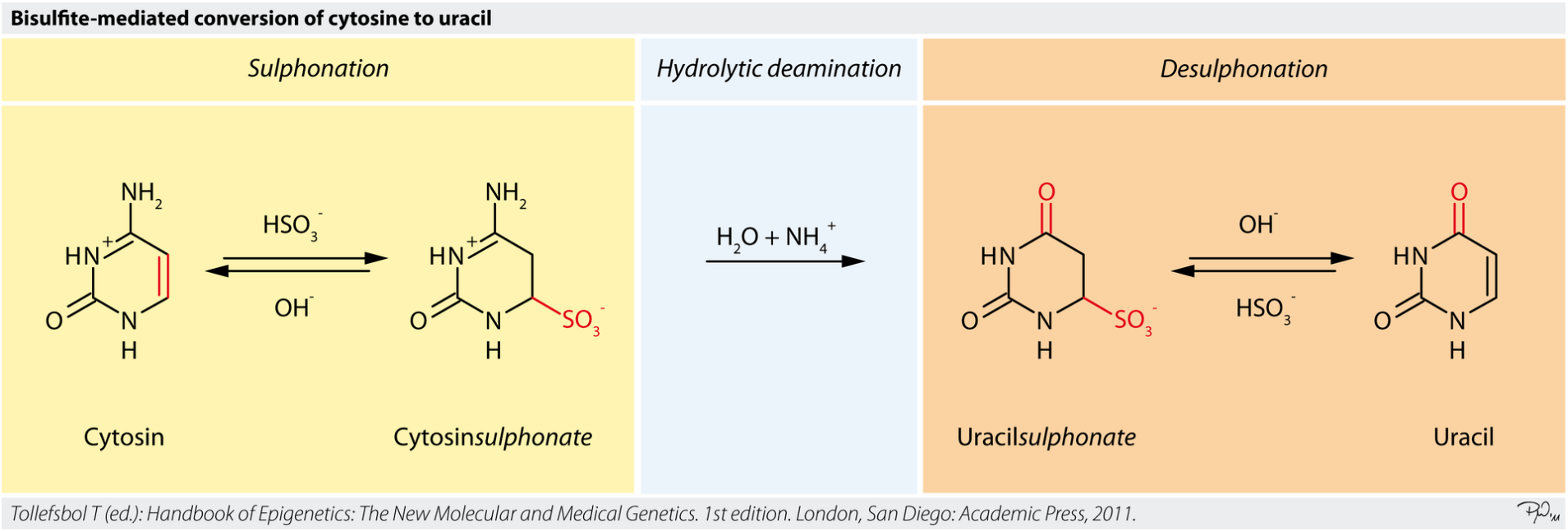




Transposase image accessed from the Protein Data Bank, https://rcsb.org/structure/1MUH

https://en.wikipedia.org/wiki/ATAC-seq
https://support.10xgenomics.com/single-cell-atac/software/pipelines/latest/algorithms/overview

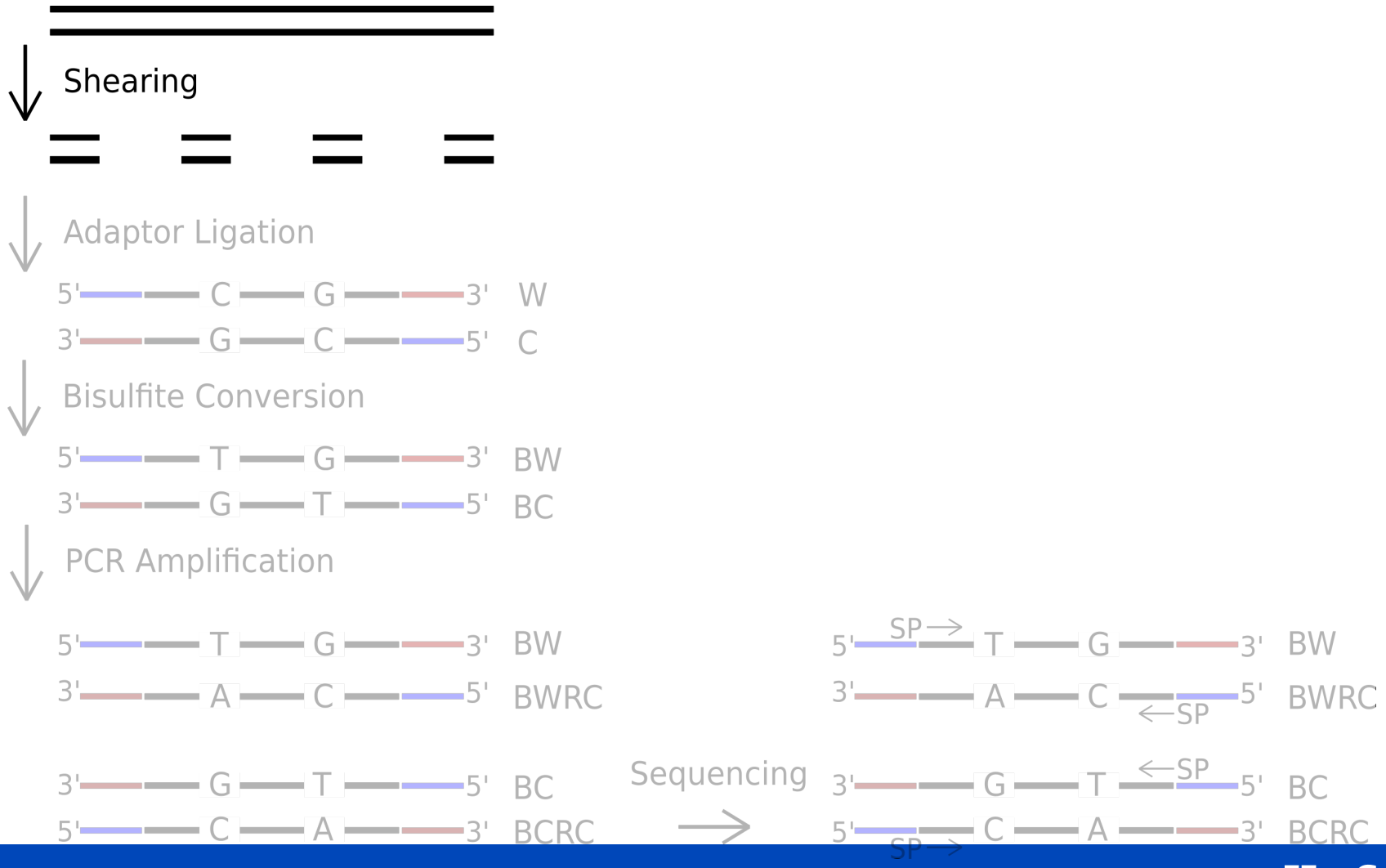
Bisulfite Sequencing



Tollefsbol T (ed.): Handbook of Epigenetics: The New Molecular and Medical Genetics. 1st edition. London, San Diego: Academic Press, 2011.

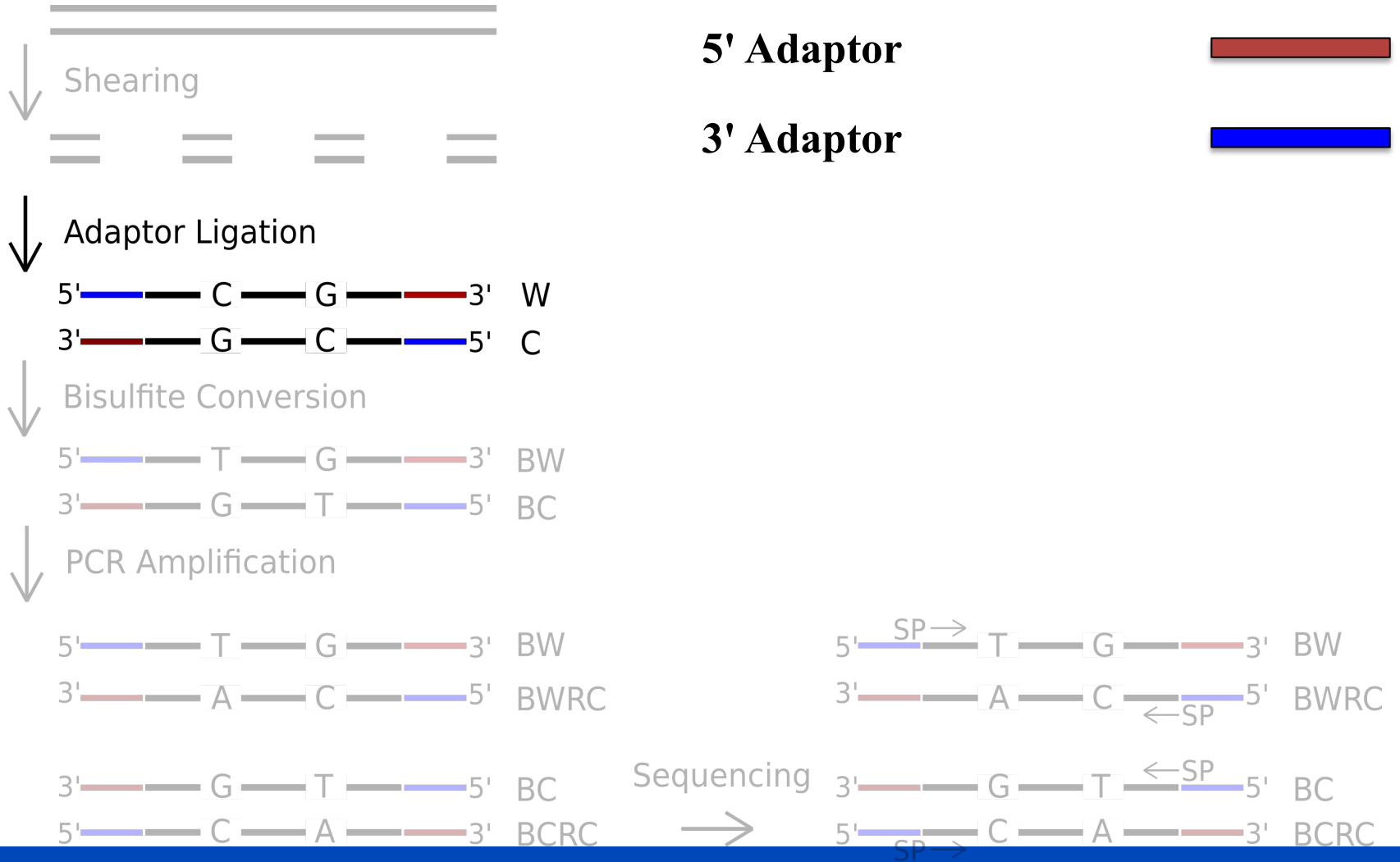
Bisulfite Sequencing

Undirectional protocol



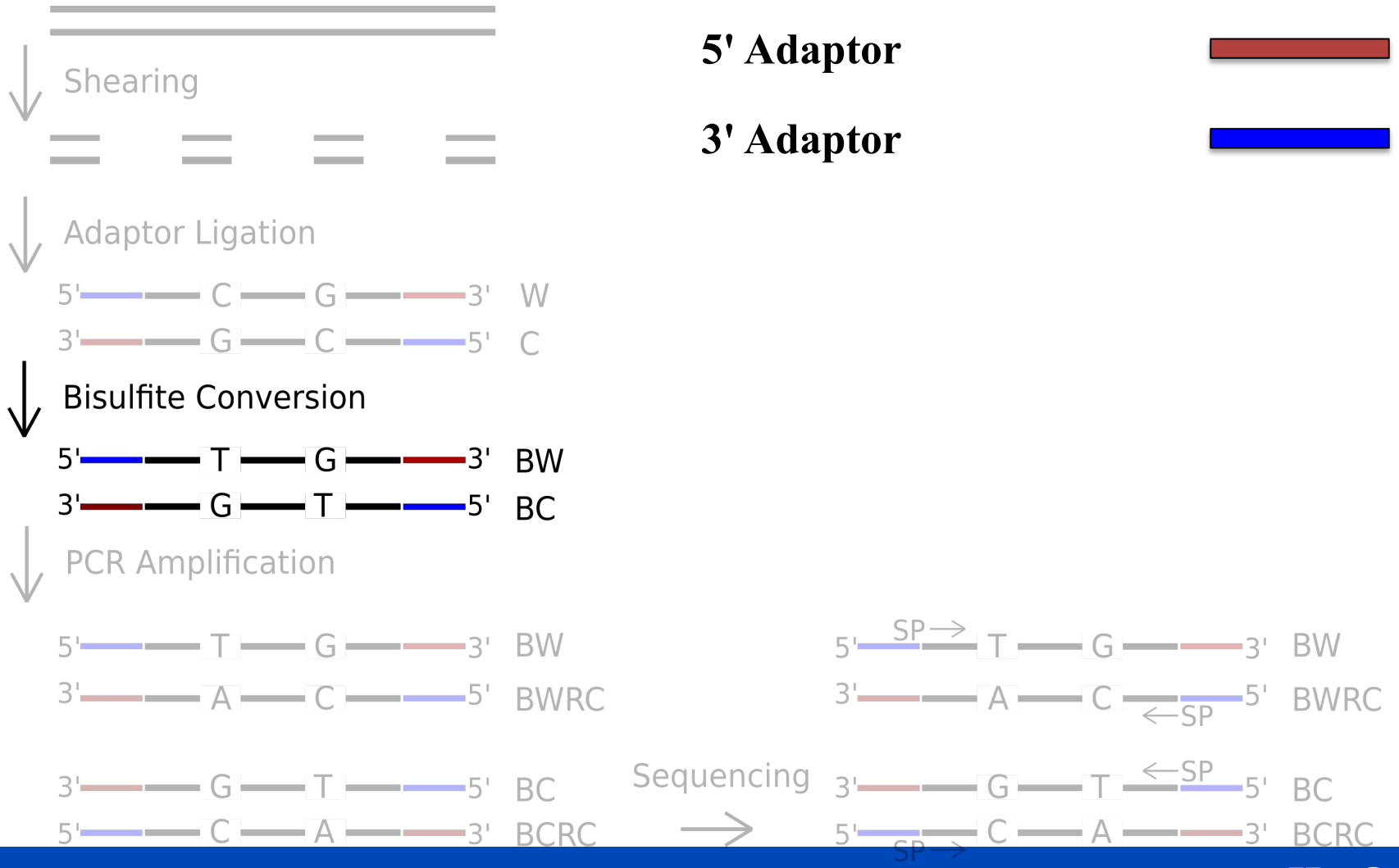
Bisulfite Sequencing

Undirectional protocol



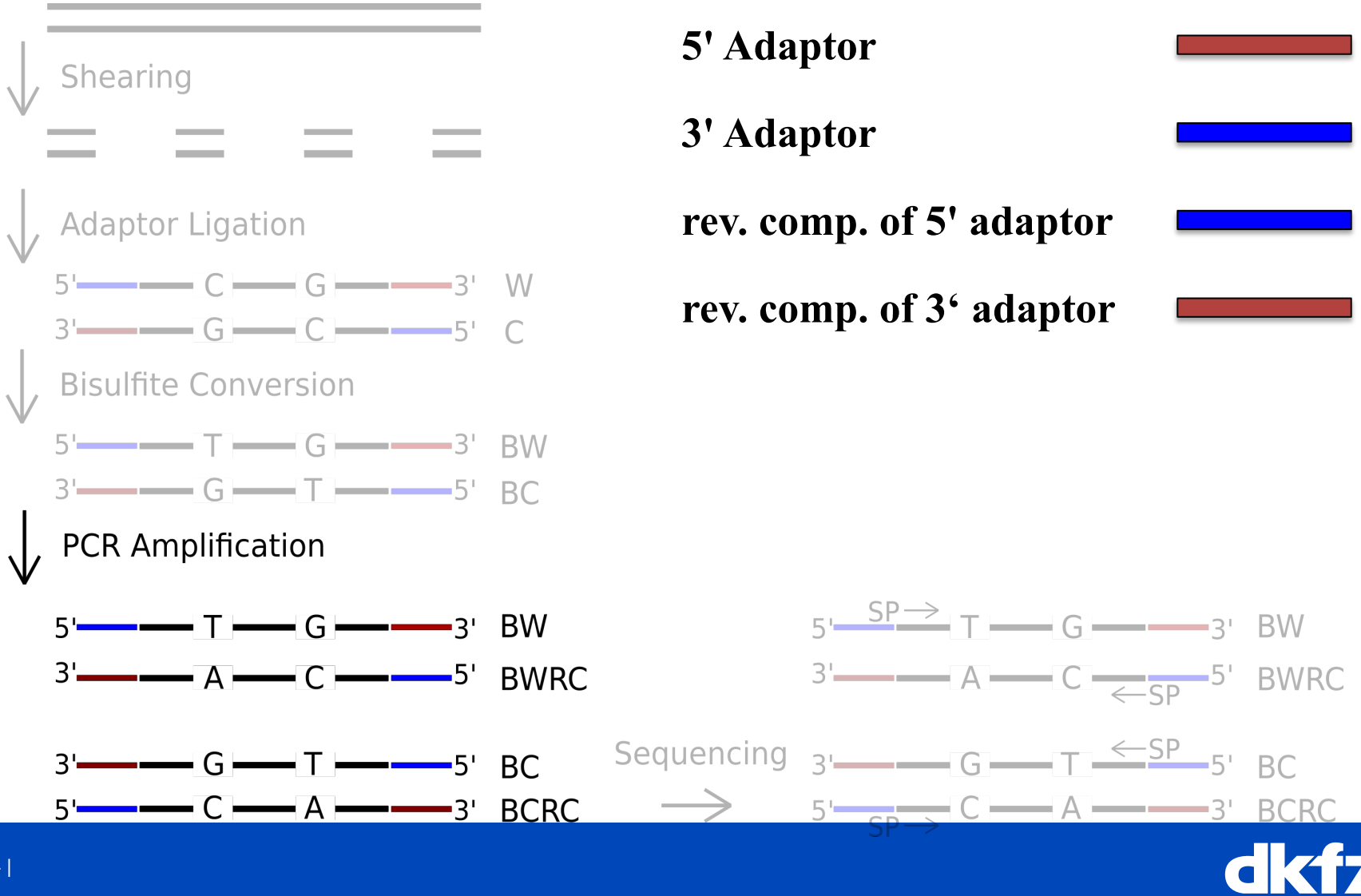
Bisulfite Sequencing

Undirectional protocol

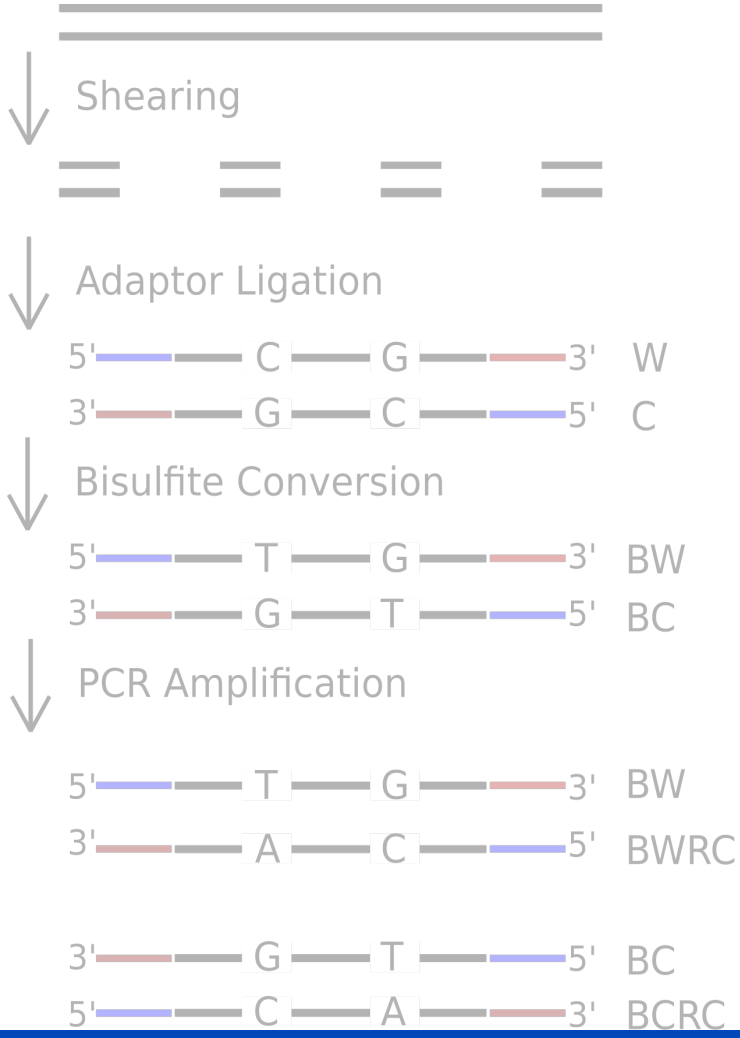


Bisulfite Sequencing

Undirectional protocol



Bisulfite Sequencing



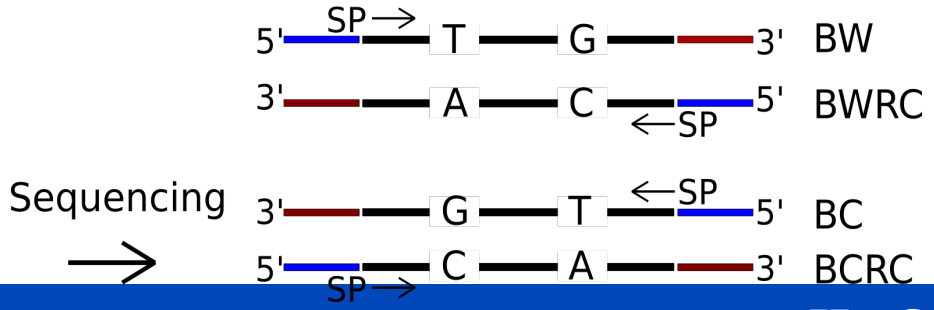
5' Adaptor

3' Adaptor

rev. comp. of 5' adaptor

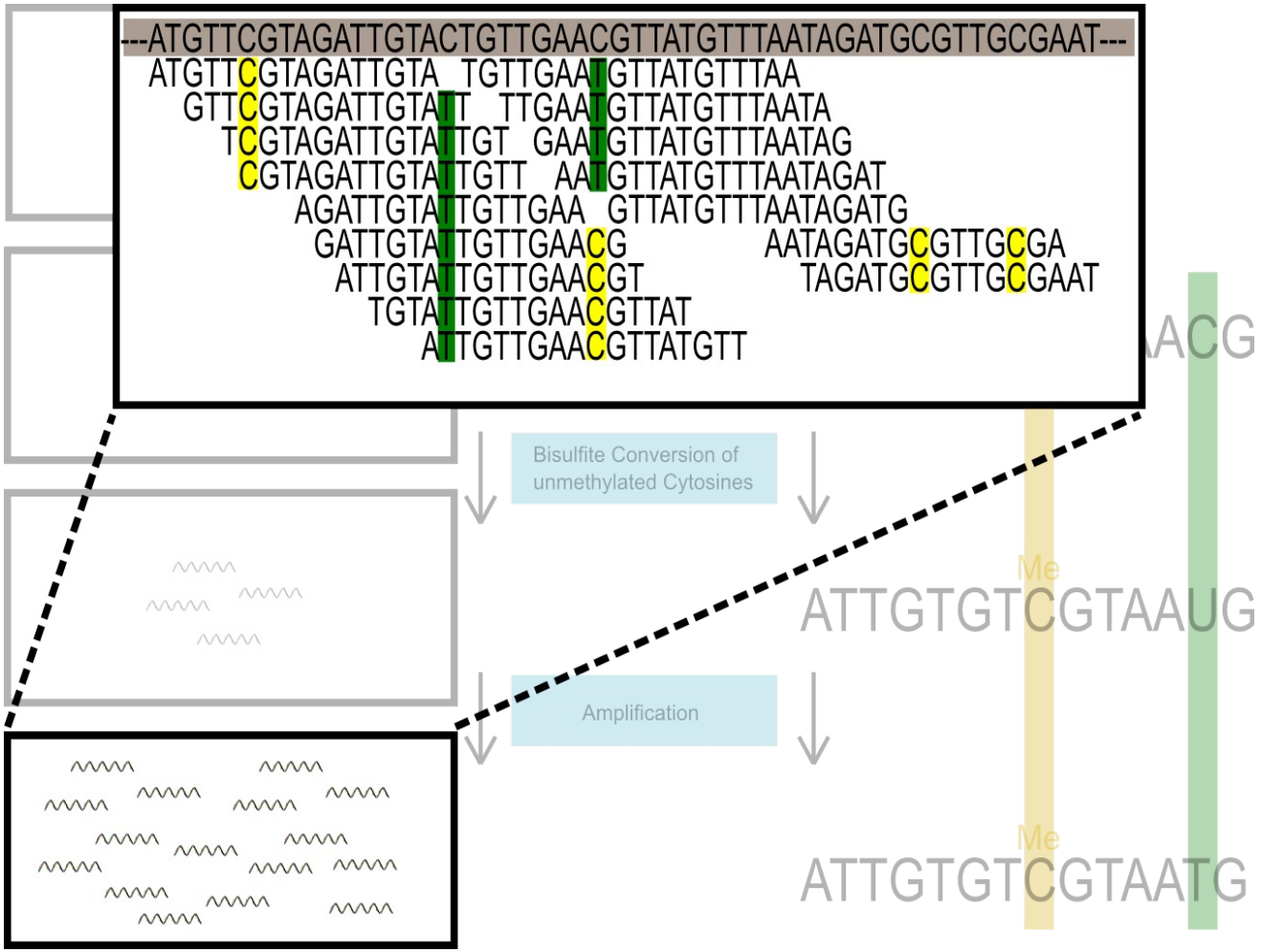
rev. comp. of 3‘ adaptor

- Sequencing Primer (SP) is part of the 5' Adaptor
- → Sequencing of BW, BWRC, BC, BCRC



Bisulfite Sequencing

Workflow (III) Sequencing/Mapping



Chromatin-Interaktionen

- 3C, 4C, 5C:
  - cross-linking von DNA
  - am häufigsten mit direkt benachbarten Regionen
  - seltener mit entfernten DNA-Abschnitten, aber dort wo physische Kontakte sind
  - z.B. Enhancer-Promoter-Beziehungen
  - gehen jeweils von Kandidaten-Region aus

- Hi-C
  - genomweit: alle wechselseitigen Kontakte in einem Genom
  - benötigt hohe Sequenziertiefe (da entfernte Kontakte selten sind)

Why single cell sequencing?

smoothie

bulk data (mix of cells)

What are you drinking?

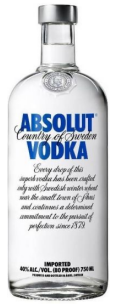
?

Macrophage

?

T-cell

?



Cancer cell

And how much?

?

?

?

Why single cell sequencing?

smoothie

bulk data (mix of cells)

Deconvolution
(infer cell types and proportions computationally but has drawbacks)

What are you drinking?

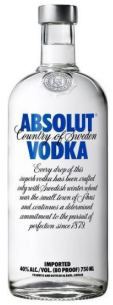
?

Makrophage

?

T-cell

?



Cancer cell

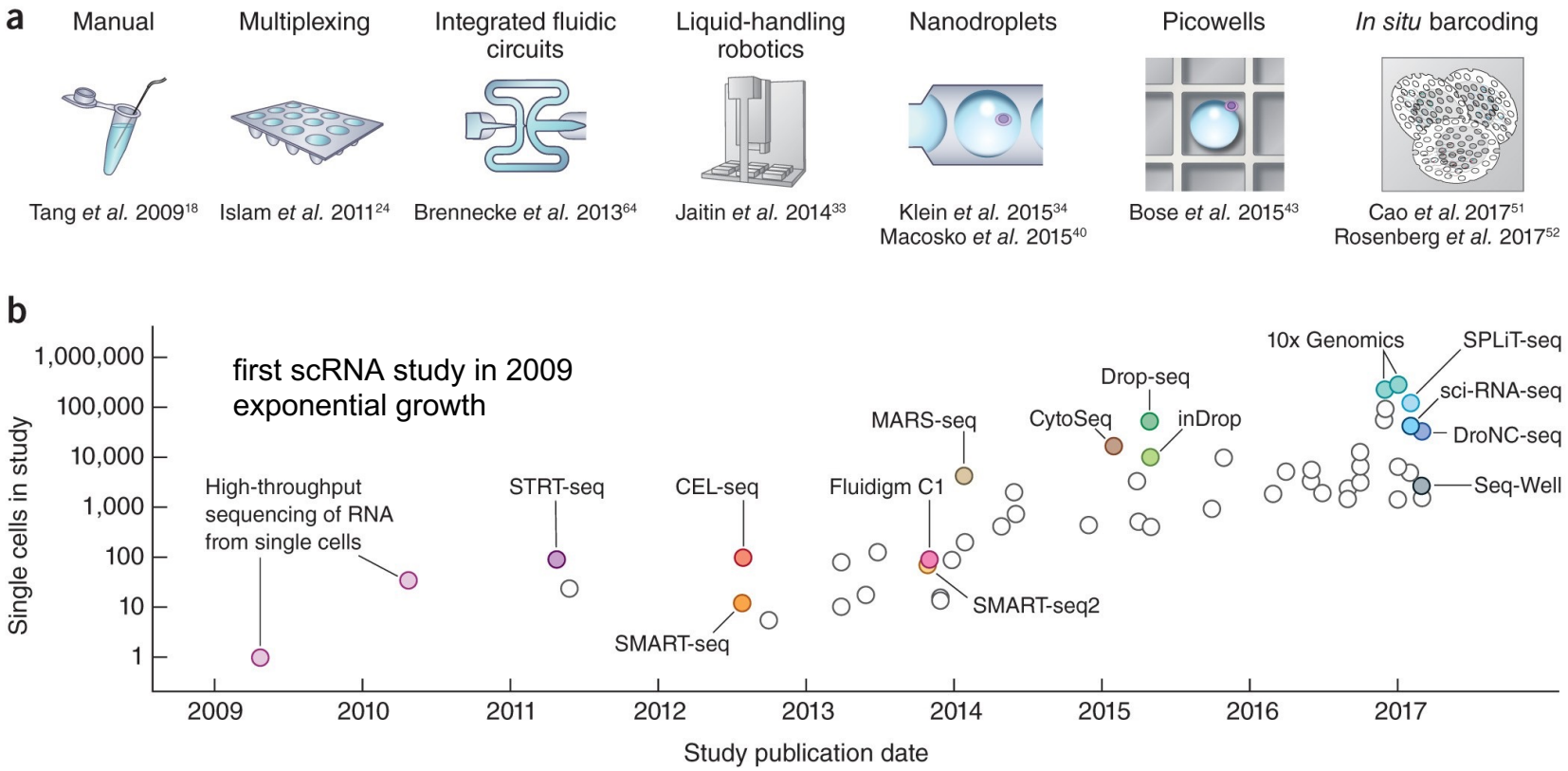
And how much?

?

?

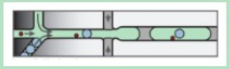
?

Single cell sequencing technologies



Exponential scaling of single-cell RNA-seq in the past decade (Svensson et al, 2018)

Single cell sequencing technologies

	inDrops	10x	Drop-seq	Seq-well	SMART-seq
Cell capture efficiency	~70-80%	~50-65%	~10%	~80%	~80%
Time to capture 10k cells	~30min	10min	1-2 hours	5-10min	--
Encapsulation type	Droplet 	Droplet 	Droplet 	Nanolitre well 	Plate-based 
Library prep	CEL-seq Linear amplification by IVT	SMART-seq Exponential PCR based amplification	SMART-seq Exponential PCR based amplification	SMART-seq Exponential PCR based amplification	SMART-seq Exponential PCR based amplification
Commercial	Yes	Yes	--	--	Yes
Cost (~$ per cell)	~0.06	~0.2	~0.06	--	1
Strengths	• Good cell capture • Cost-effective • Real-time monitoring • Customizable	• Good cell capture • Fast and easy to run • Parallel sample collection • High gene / cell counts	• Cost-effective • Customizable	• Good cell capture • Cost-effective • Real-time monitoring • Customizable	• Good cell capture • Good mRNA capture • Full-length transcript • No UMI
Weaknesses	Difficult to run	Expensive	Difficult to run & low cell capture efficiency	Still new!	Expensive

https://hbctraining.github.io/In-depth-NGS-Data-Analysis-Course/sessionIV/slides/Single_Cell_Sept_2018_final.pdf

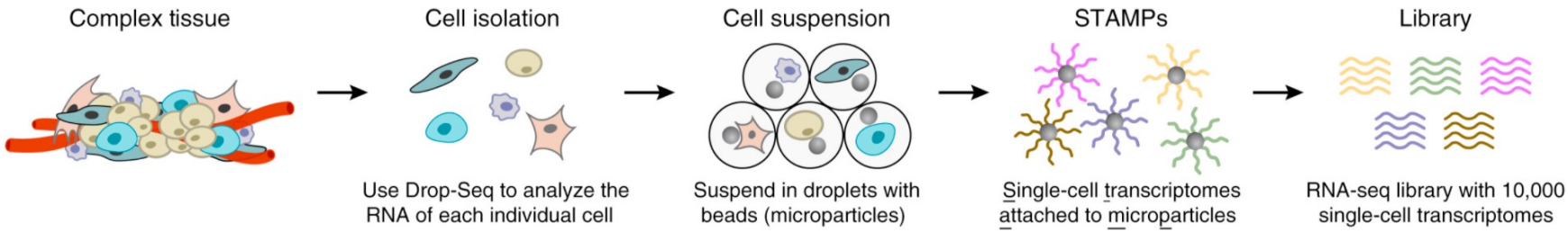
Single cell sequencing technologies

- there are many different single cell technologies available
- unimodal / multimodal technologies exist
- number of cells captured are highly variable for each technology

Single cell sequencing technologies

- there are many different single cell technologies available
- unimodal / multimodal technologies exist
- number of cells captured are highly variable for each technology
- number of transcripts captured per cell can vary for each technology
- price/cell varies for each technology
- can use UMIs or not
- other differences in protocols (5’ or 3’, …)

From complex tissue to sequencing library using droplet based microfluidics single cell technology

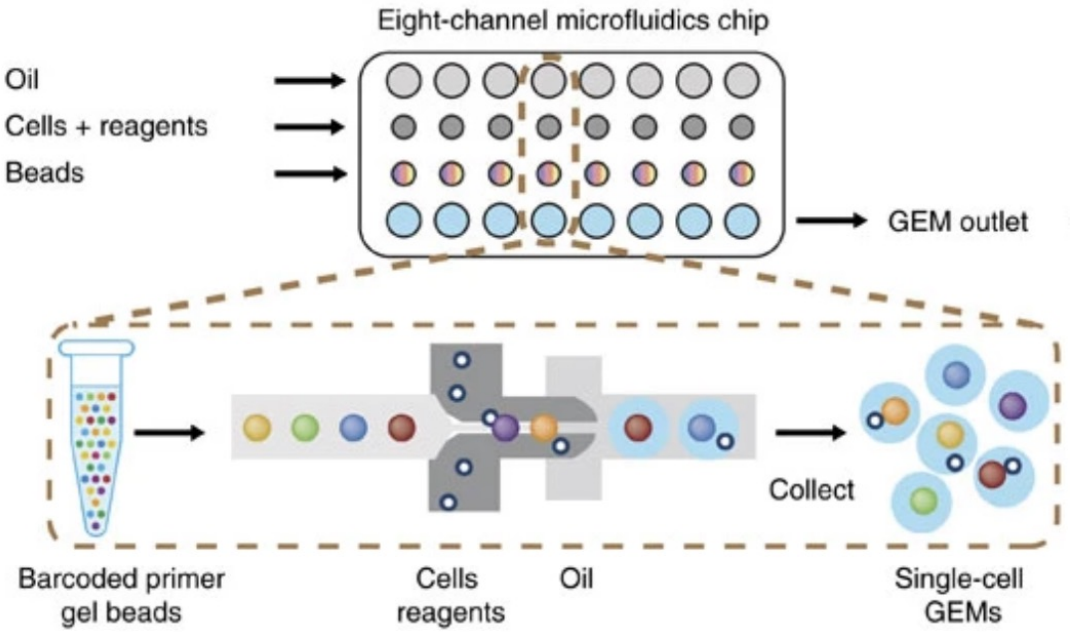


single-cell dissociation: digest tissue
single-cell isolation:
capture cell in microfluidic droplet (droplet-based methods)
empty droplets / **doublets** or **multiplets** possible
capture cell in wells (not shown)
library construction: each droplet contains necessary chemicals to break the cell membranes and perform library construction. During library construction mRNA is captured, reverse-transcribed to cDNA and amplified.

Macosko et al (2015) Luecken et al (2019)

10xGenomics Chromium: a droplet microfluidics single cell technology

example: 10xGenomics for scRNA

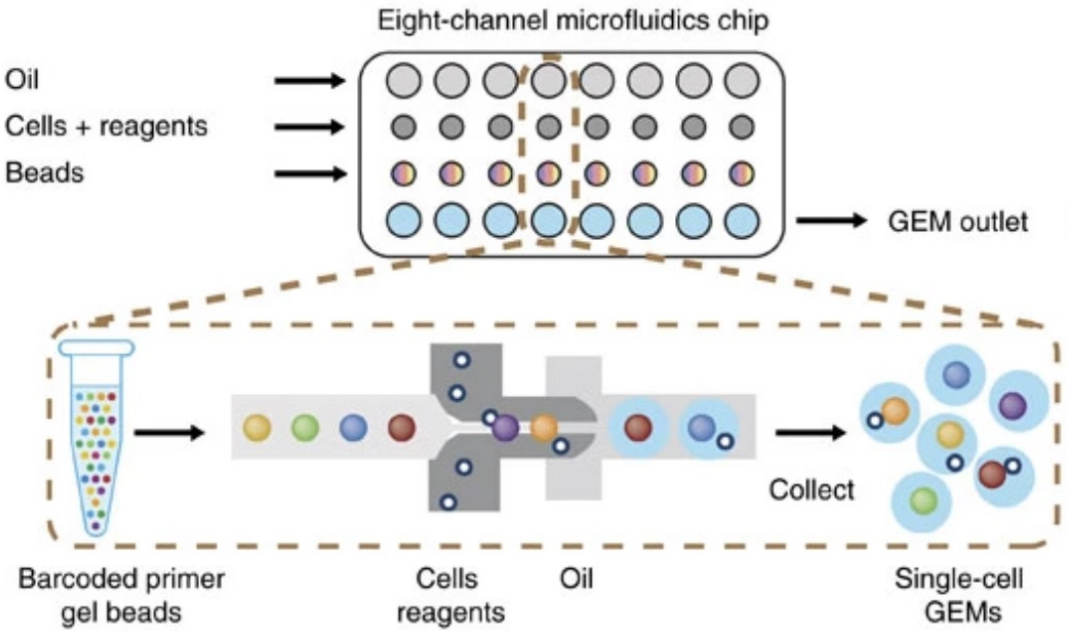


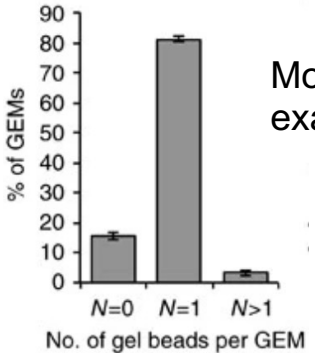
GEM = Gel bead in EMulsion

Massively parallel digital transcriptional profiling of single cells (Zheng et al, 2017)

10xGenomics Chromium: a droplet microfluidics single cell technology

example: 10xGenomics for scRNA





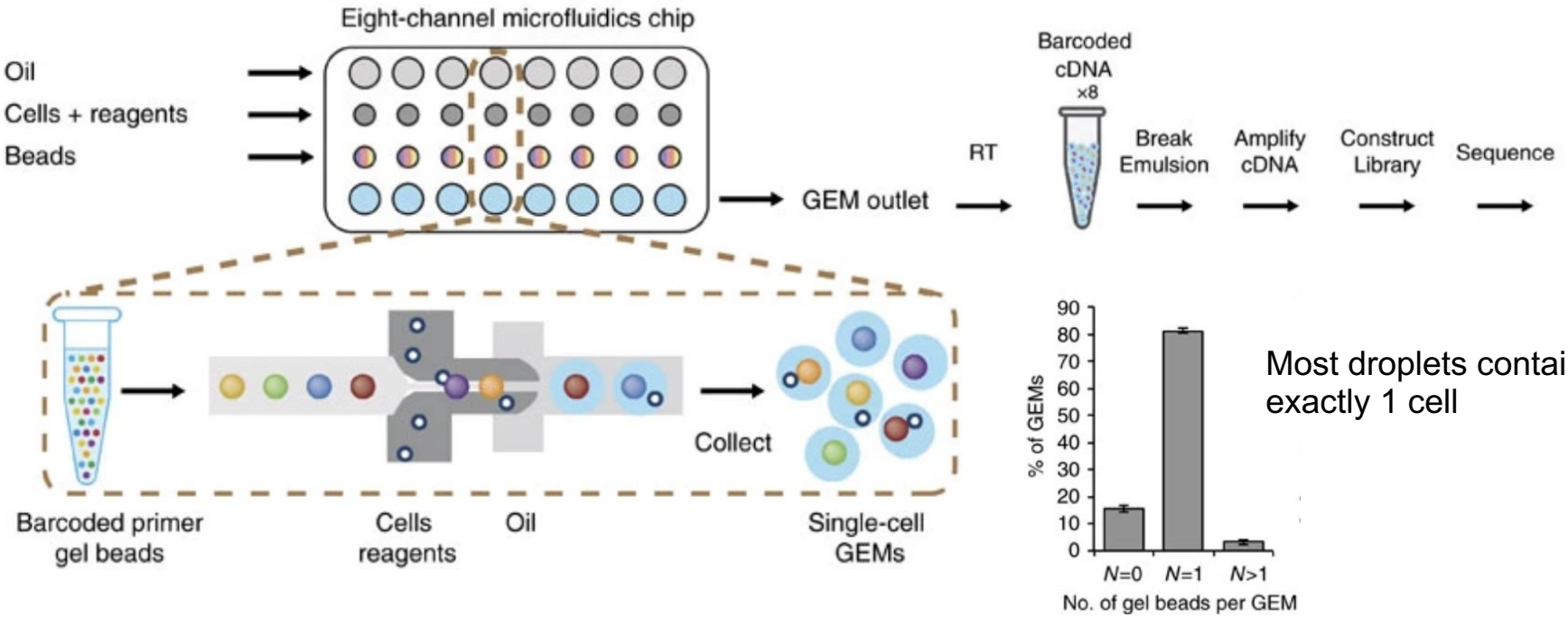
Most droplets contain exactly 1 cell

GEM = Gel bead in EMulsion

Massively parallel digital transcriptional profiling of single cells (Zheng et al, 2017)

10xGenomics Chromium: a droplet microfluidics single cell technology

example: 10xGenomics for scRNA

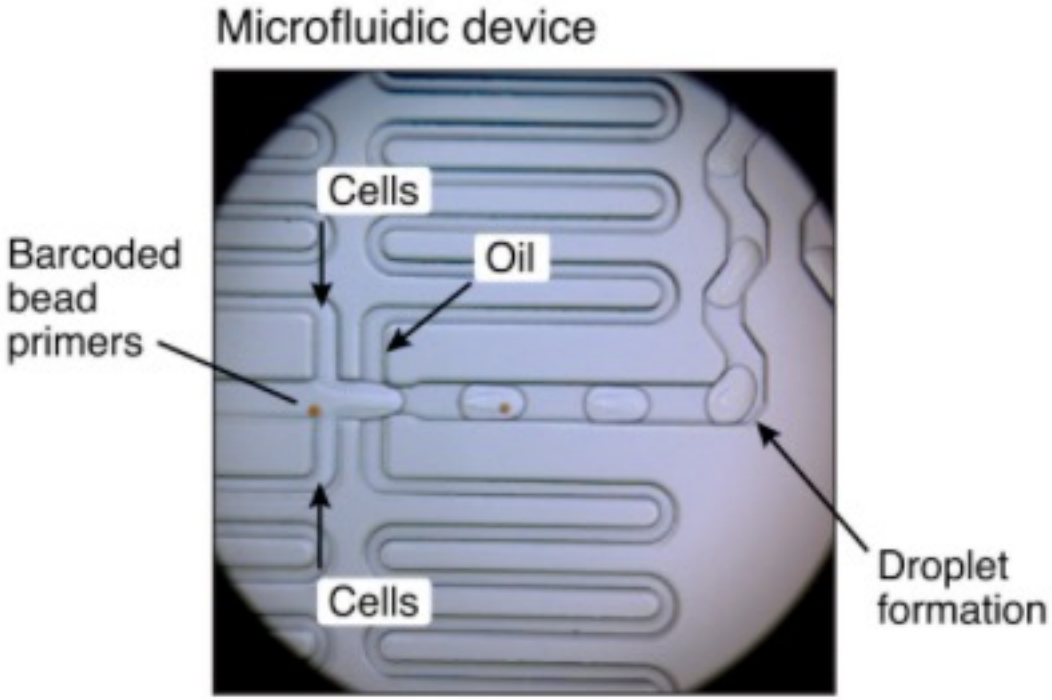


Most droplets contain exactly 1 cell

GEM = Gel bead in EMulsion

Massively parallel digital transcriptional profiling of single cells (Zheng et al, 2017)

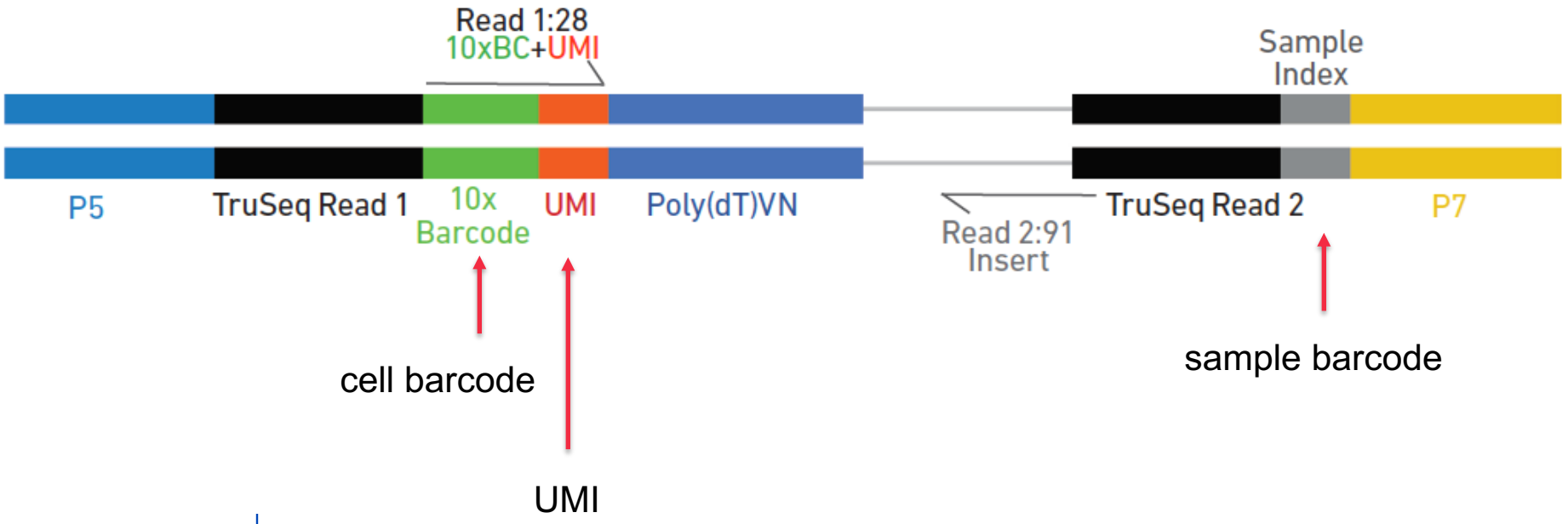
Droplet microfluidics single cell technology



Macosko et al, 2015

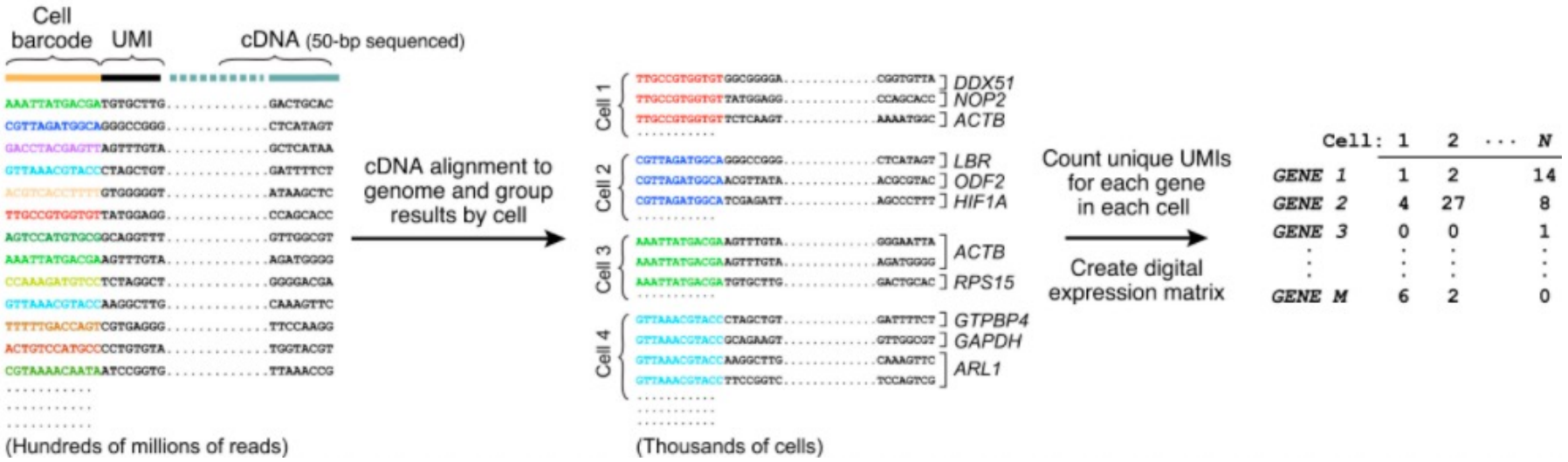
barcodes, barcodes, barcodes…

Single Cell 3' v3 Gene Expression Library (10xGenomics)



Pre-processing

example scRNA



Unique Molecular Identifiers (UMI)

- in the **PCR** amplification step some sequences are preferentially amplified which results in a **PCR amplification bias**

- as the PCR amplification bias highly **influences quantification** of each transcript therefore one wants to **identify** and **remove**
PCR duplicates

Unique Molecular Identifiers (UMI)

- in the **PCR** amplification step some sequences are preferentially amplified which results in a **PCR amplification bias**

- as the PCR amplification bias highly **influences quantification** of each transcript therefore one wants to **identify** and **remove**
PCR duplicates

- thus after alignment to the reference genome PCR duplicates are **identified** by the definition of **same start** and **end point** (**Picard**)
and then **removed** or **marked** (classical way, when using no UMIs)

- prior to alignment, UMIs are removed from the read sequence, not to influence alignment with artificial sequence

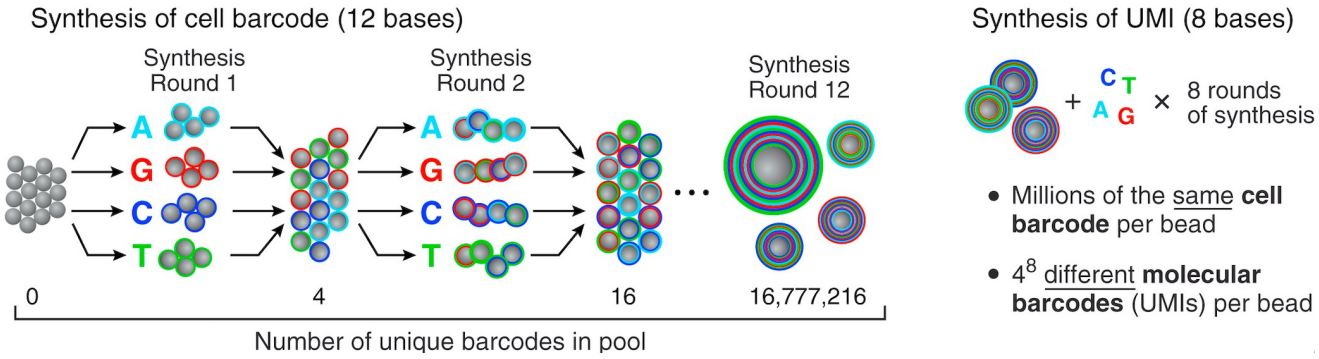
Unique Molecular Identifiers (UMI)

- in the **PCR** amplification step some sequences are preferentially amplified which results in a **PCR amplification bias**

- as the PCR amplification bias highly **influences quantification** of each transcript therefore one wants to **identify** and **remove**
PCR duplicates

- thus after alignment to the reference genome PCR duplicates are **identified** by the definition of **same start** and **end point** (**Picard**)
and then **removed** or **marked** (classical way, when using no UMIs)

- prior to alignment, UMIs are removed from the read sequence, not to influence alignment with artificial sequence



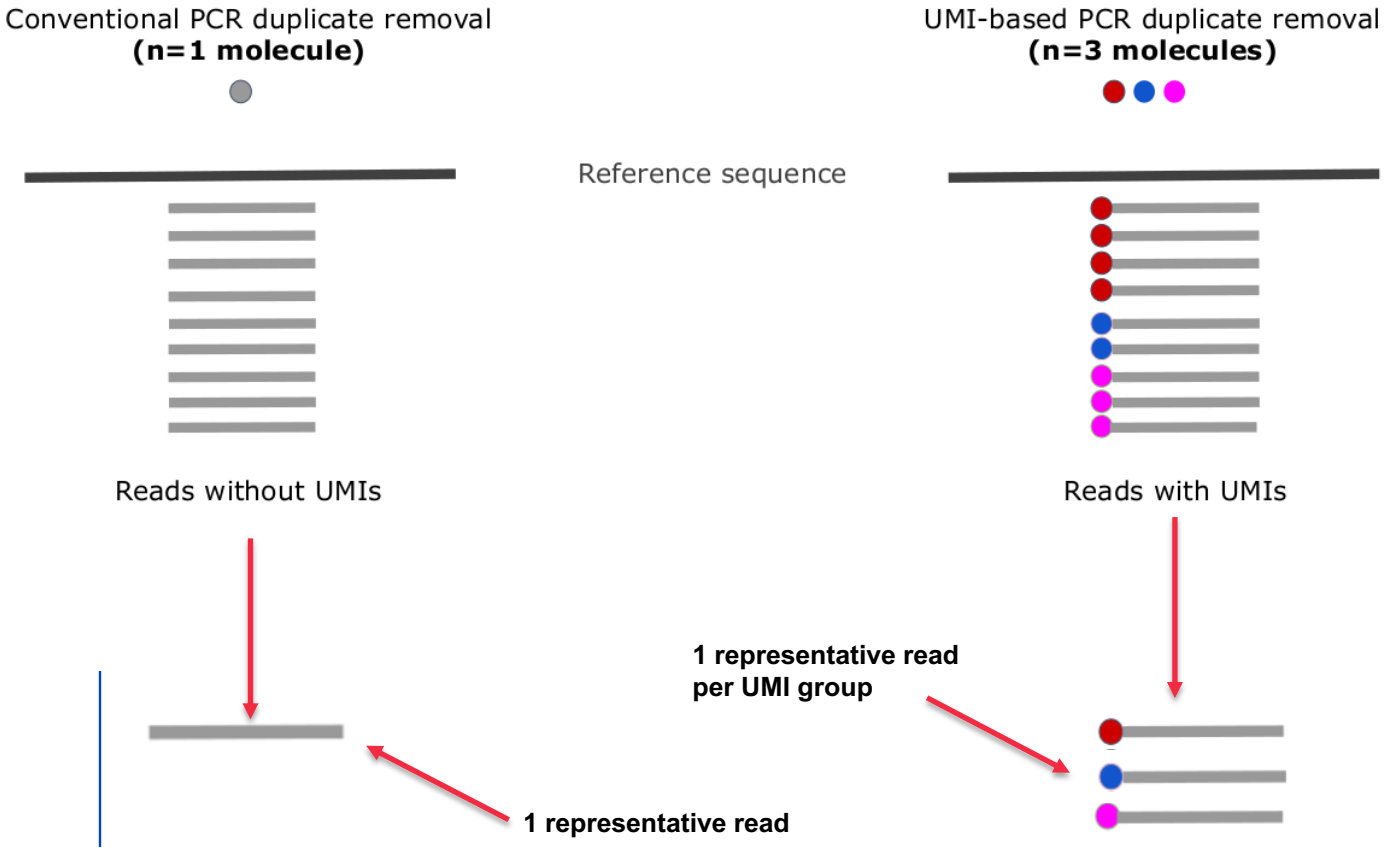
Unique Molecular identifiers (UMI)

- **Unique Molecular Sequences (UMIs)** are **random sequences** of bases used to tag each molecule / fragment

- UMIs are ligated to fragments **before PCR** to help in identification of PCR duplicates

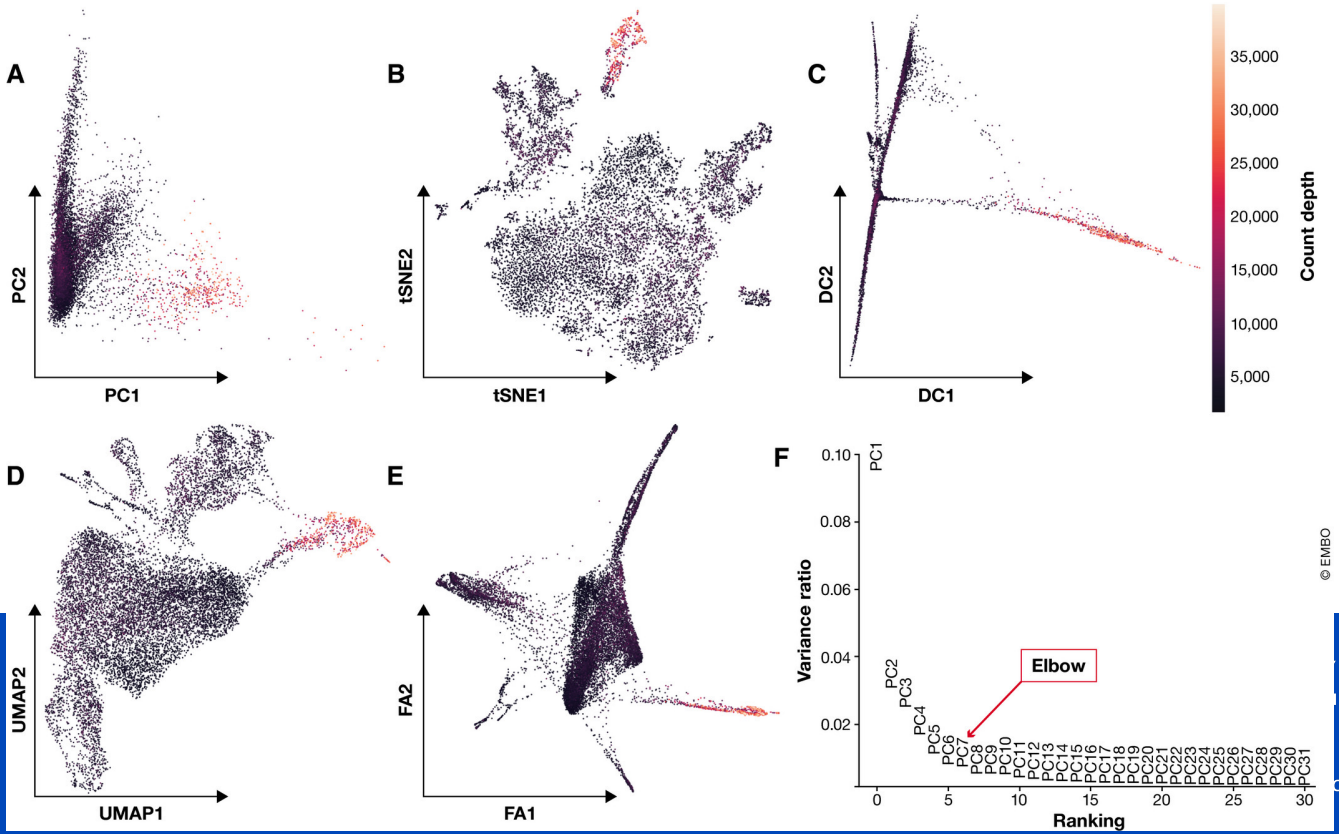
- if two (or multiple) reads **align** to the **same genomic / transcriptomic location** and have the **same UMI** it is highly likely they originate from the **same DNA/RNA fragment** and are thus PCR duplicates

Unique Molecular identifiers (UMI)



Dimensionality reduction

- describe ~15k genes with fewer components
- main objectives:
  - **summarization** (for example 40 PCA components), **PCA** (linear) or **diffusion maps** (non-linear)
  - **vizualisation** (for example 2 to 3 components)
  - popular is for **summarisation PCA** and **visualisation UMAP** based on PCA components



A-E
different
visualisation techniques

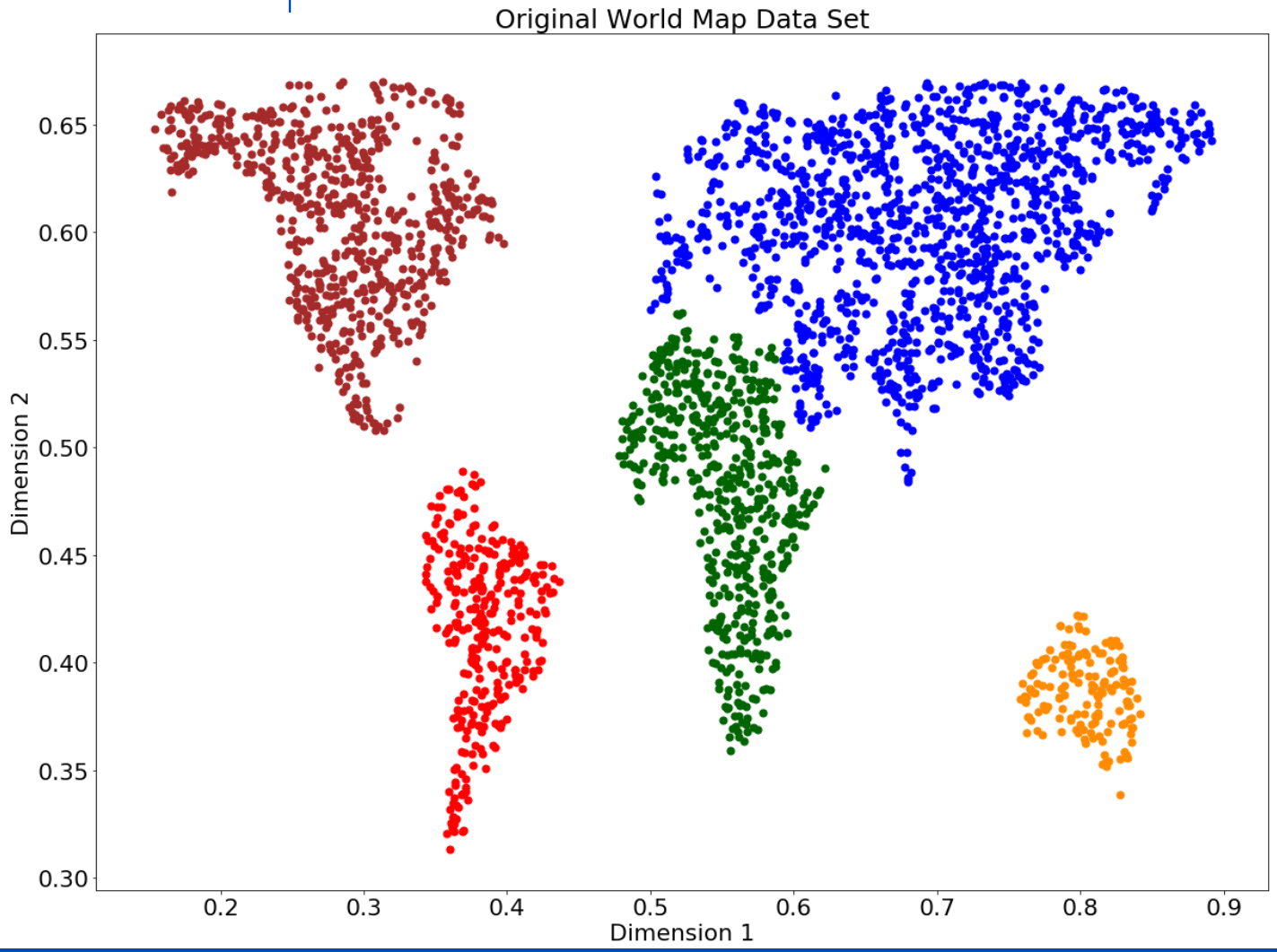
F **elbow plot based on PCA** to determine cutoff of components to be used for summarisation

Luecken et al (2019)

Vizualization
UMAP: Uniform Manifold Approximation and Projection

- manifold learning technique for dimension reduction
- similar to tSNE
- PCA: linear algorithm
- tSNE & UMAP non-linear manifold learners
- PCA, tSNE, UMAP ?

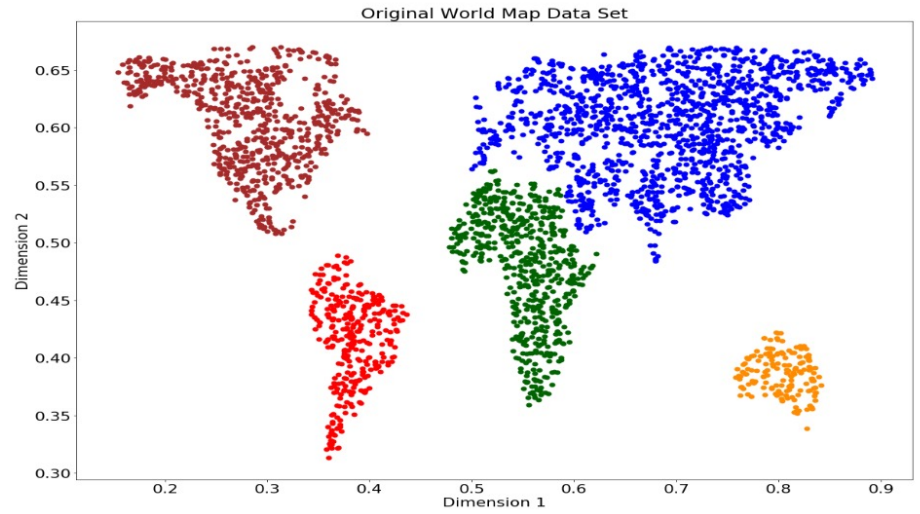
UMAP: Uniform Manifold Approximation and Projection

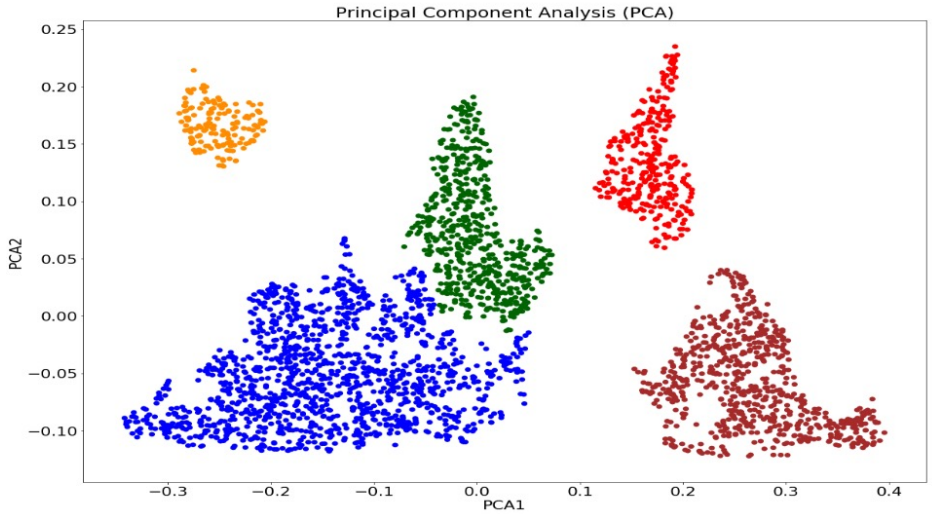


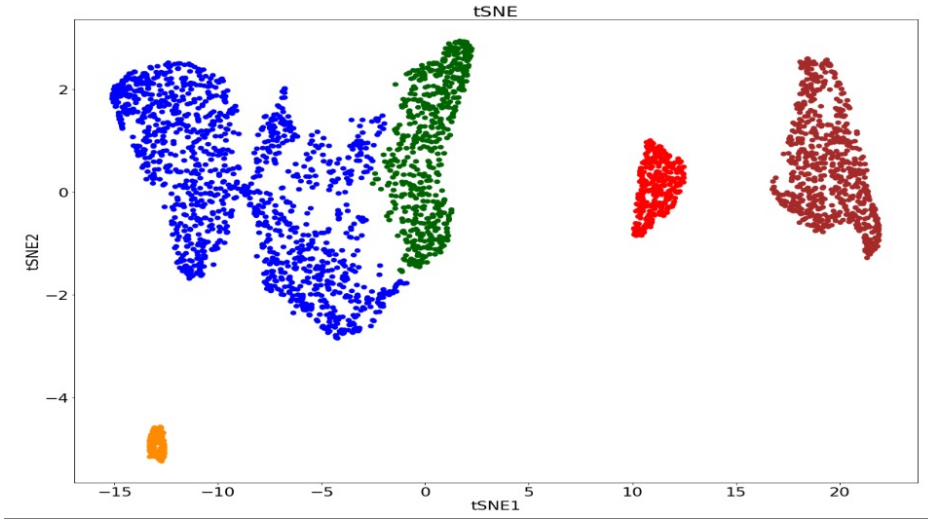
toy example

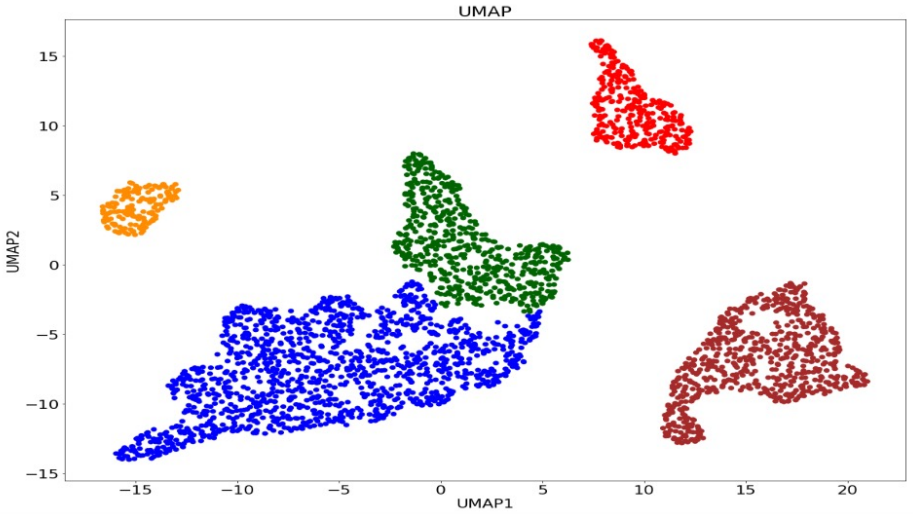
https://towardsdatascience.com/tsne-vs-umap-global-structure-4d8045acba17

UMAP: Uniform Manifold Approximation and Projection



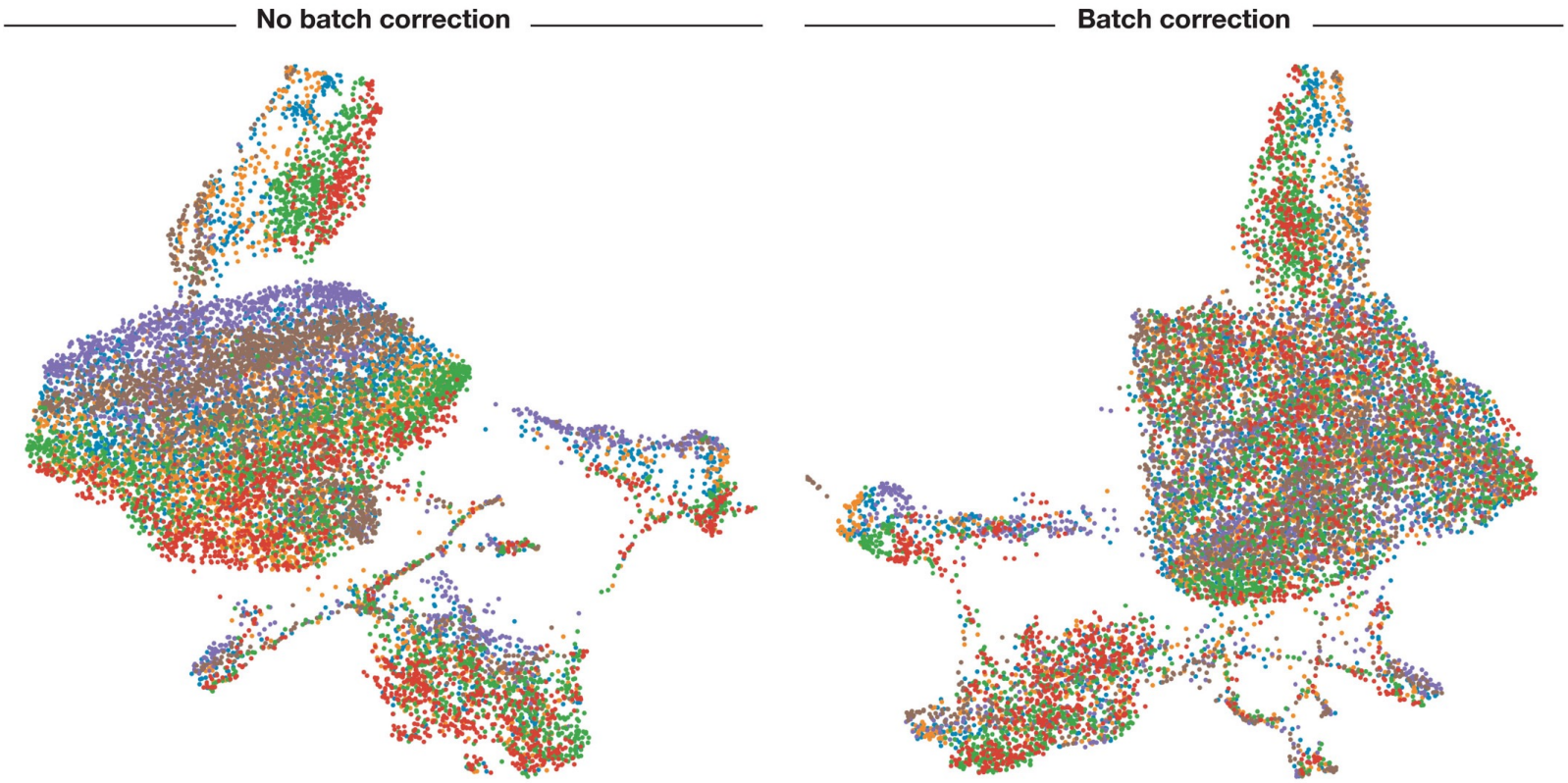






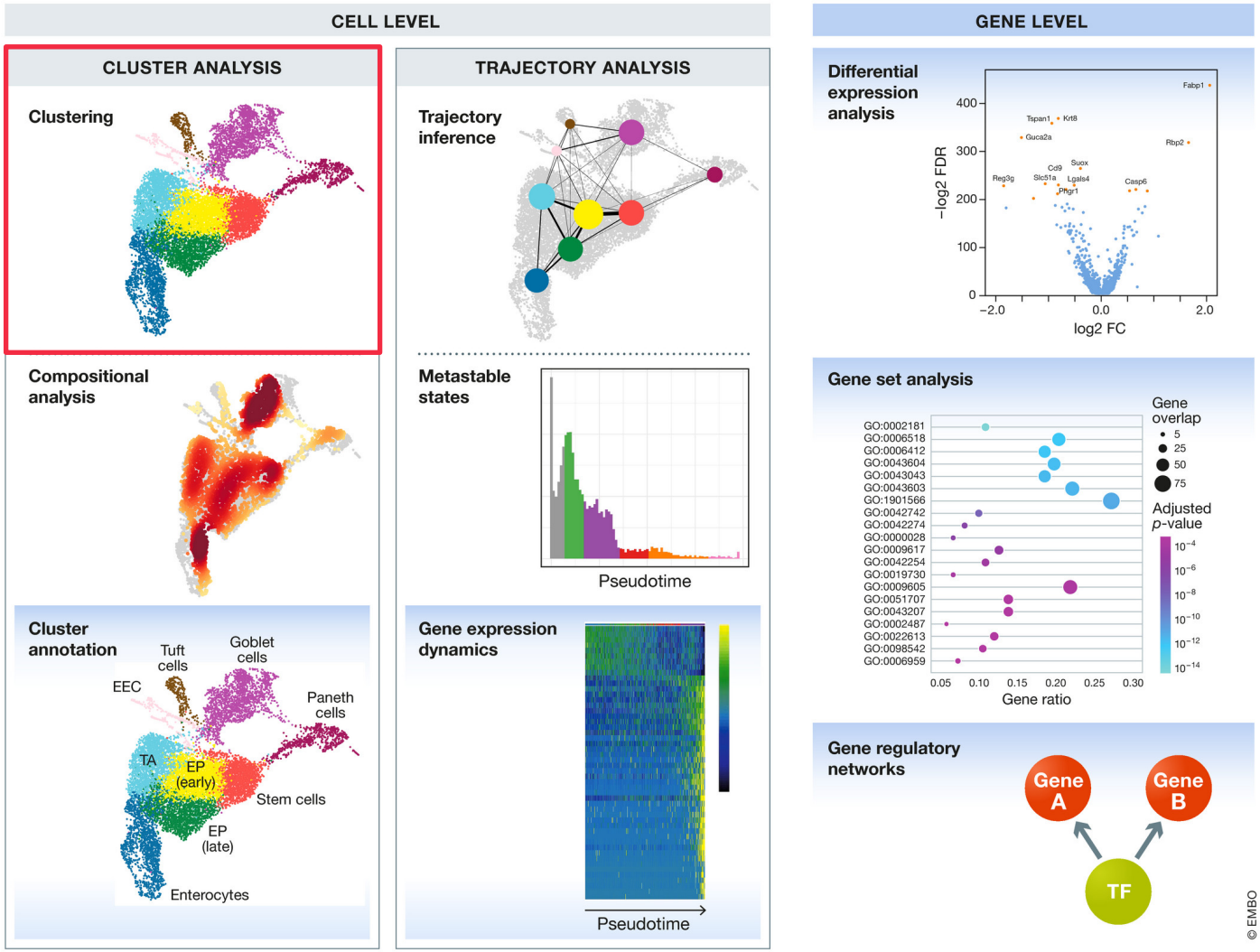
https://towardsdatascience.com/tsne-vs-umap-global-structure-4d8045acba17

Batch effects and data integration



Luecken et al (2019)

Downstream analysis



Luecken et al (2019)