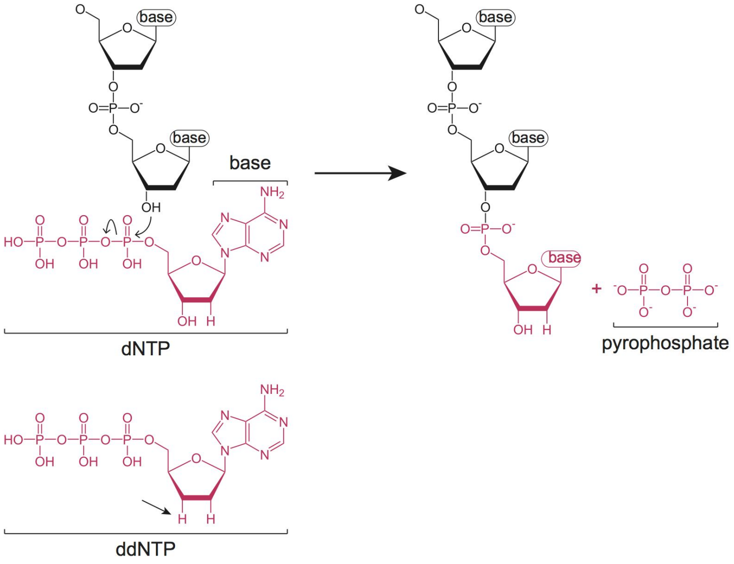
DNA sequencing

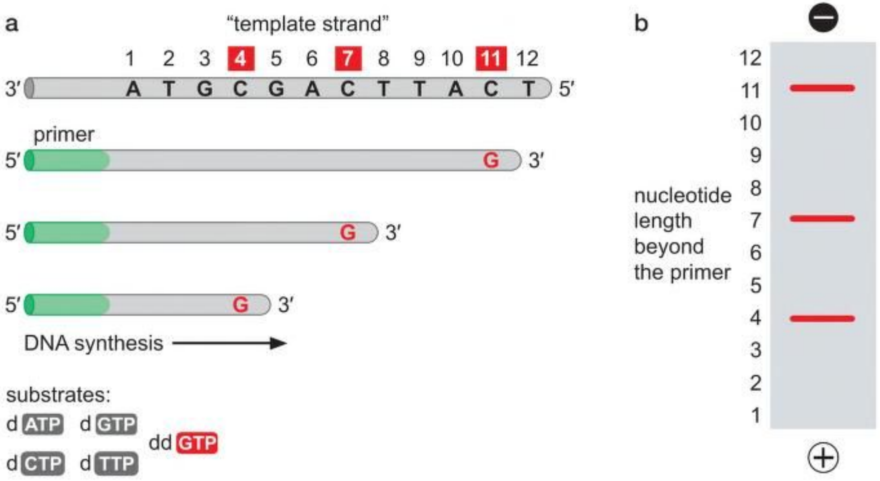
Anna Cuomo
EBI & University of Cambridge

Ximena Ibarra-Soria
Cancer Research UK

Sanger sequencing

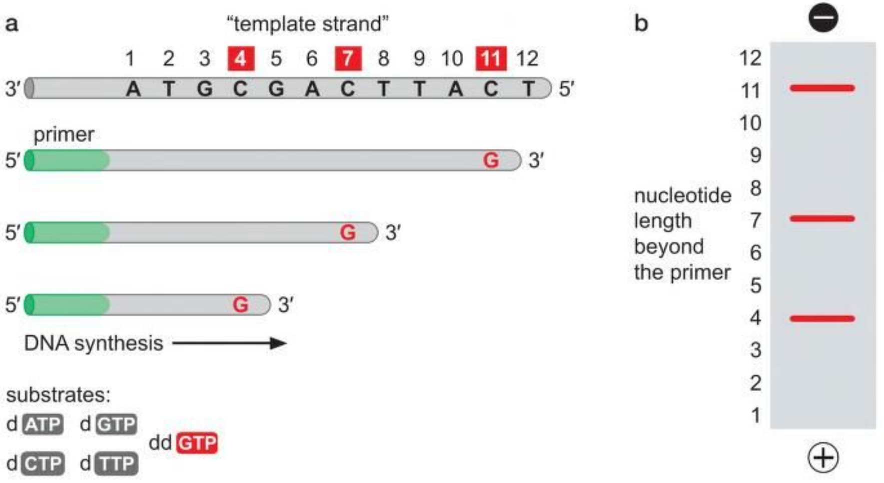
Process of determining the order of nucleotides in a DNA molecule.
Uses **chain-terminating nucleotides** to block extension at particular bases.

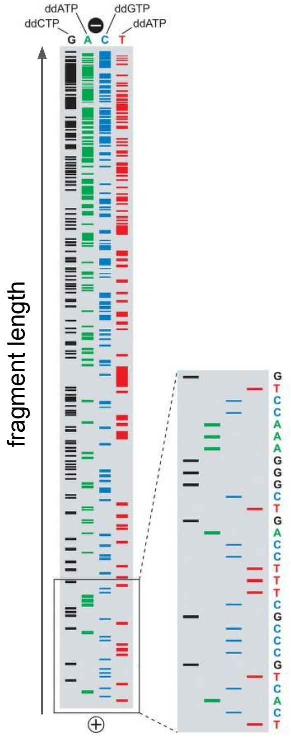




Watson et al., Molecular Biology of the Gene 7th Edition, Chapter 7.

Sanger sequencing





Produces fragment of around 600-1000 bp.

Very low error rate.

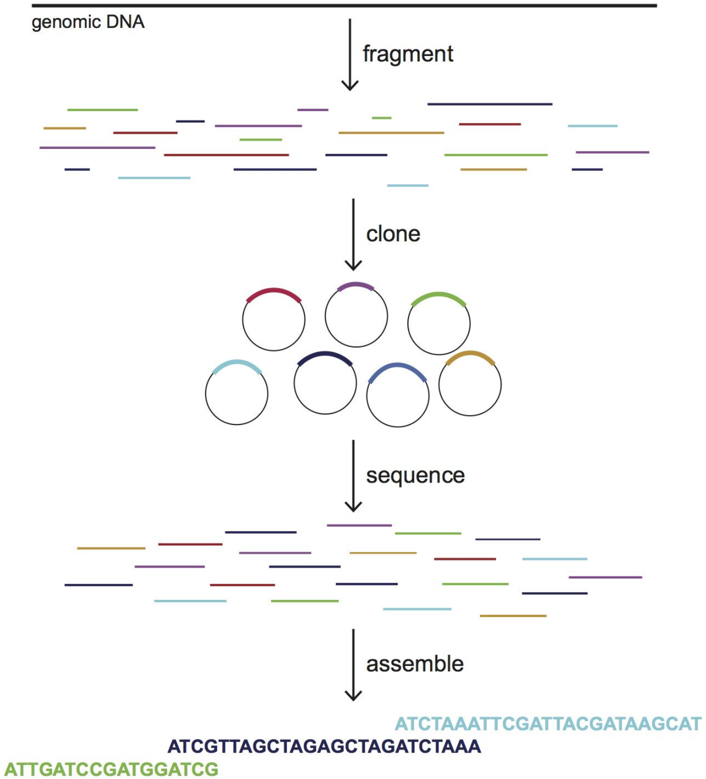
Watson et al., Molecular Biology of the Gene 7th Edition, Chapter 7.

Sanger sequencing

Shotgun sequencing: DNA is fragmented into small pieces that are cloned into plasmids, amplified and sequenced.

The resulting sequences are assembled based on overlapping segments.

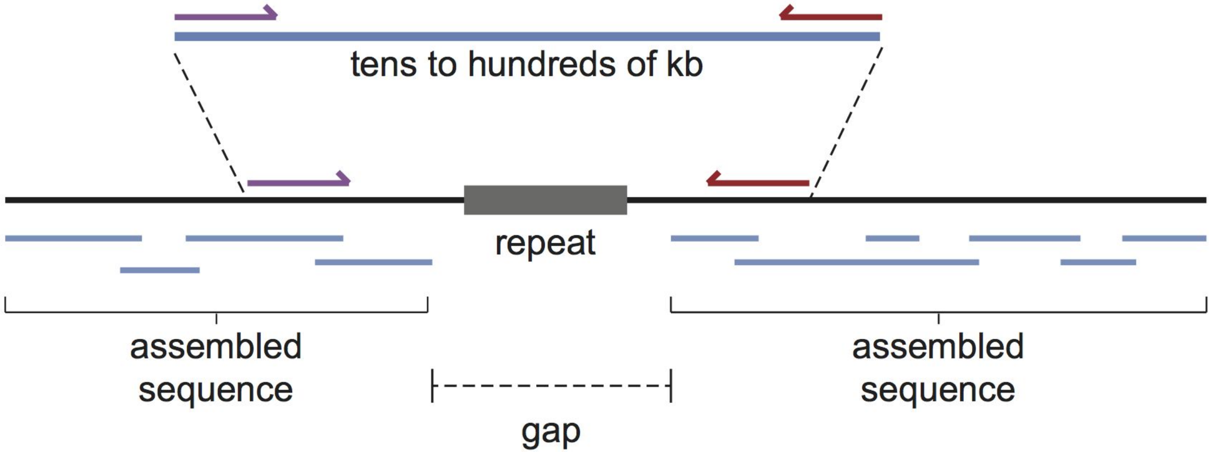
- A major challenge is the repetitive nature of eukaryotic genomes.



Shendure et al., *DNA sequencing at 40...*, Nature (2017).
doi:10.1038/nature24286

Paired-end sequencing

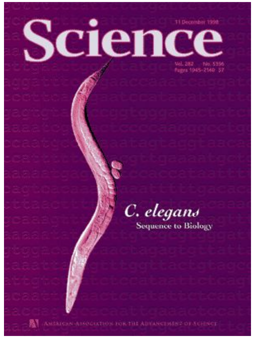
Sequence the ends of a large fragment.



Sequences matching the two short reads are now known to come from the same molecule and to be in *close* proximity.

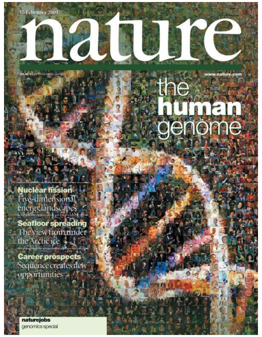
Applications of Sanger sequencing

Widely used for *de novo* sequencing of complete genomes.



1998

2000



2001

2002

2004

2005

Remains the gold standard.

- Used for validation.

Next generation sequencing

Human Genome -> 15 years to complete (published in 2004).
-> 3 billion US dollars.

Development of new sequencing technologies with **increased throughput**.
Known interchangeably as:

- next generation (NGS)
- second generation
- massively parallel
- high-throughput

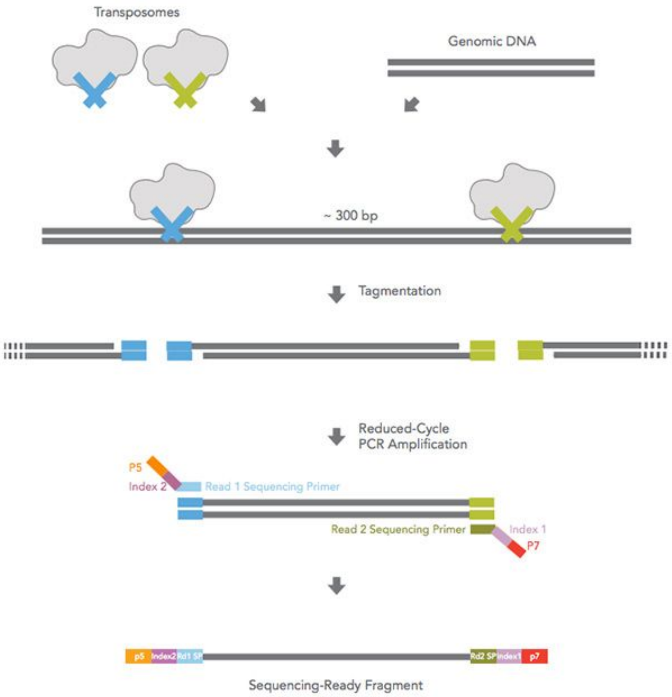
Next generation sequencing technologies

	Read length	Reads / run	Run time	Error rate	Cost per Gb
Pyrosequencing	400-700 bp	1 M	10-23 hours	<1%	US$19,500
Sequencing by ligation	50-75 bp	0.7-1.4 B	6-10 days	<0.1%	US$70-130
Sequencing by synthesis	36-150 bp	1.5-3 B	1-6 days	<0.1%	US$7-50

Illumina sequencing

Library preparation (Nextera).

- Tagmentation: a transposase randomly inserts into the DNA and ligates an adaptor.
- Barcodes and terminal sequences are added via PCR.
- Library is amplified and cleaned up.



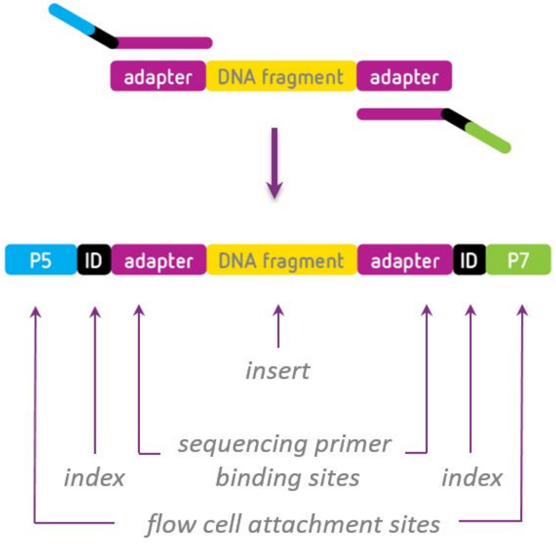
https://emea.illumina.com/content/dam/illumina-marketing/documents/products/illumina_sequencing_introduction.pdf

Illumina sequencing

Library preparation (Nextera).

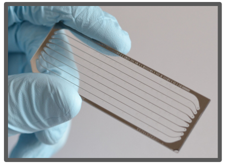
- Tagmentation: a transposase randomly inserts into the DNA and ligates an adaptor.
- Barcodes and terminal sequences are added via PCR.
- Library is amplified and cleaned up.

The index attached to each fragment is a barcode used to identify the sample: **multiplexing.**



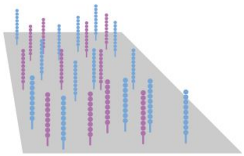
https://emea.illumina.com/content/dam/illumina-marketing/documents/products/illumina_sequencing_introduction.pdf

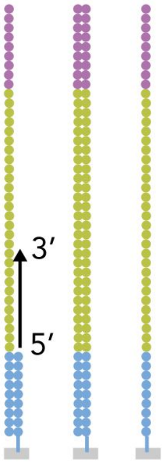
Illumina sequencing

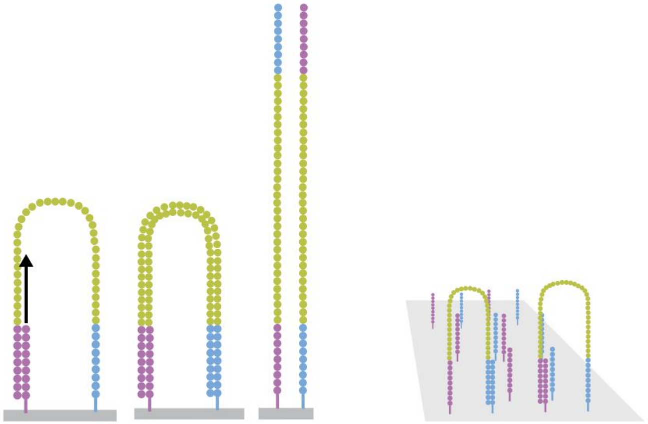


flowcell

glass slide with 8 lanes





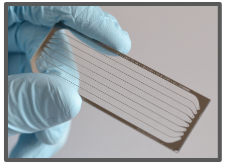


surface coated by millions of terminal sequences

bridge amplification

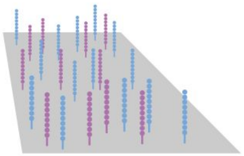
https://www.illumina.com/documents/products/datasheets/datasheet_cbot.pdf

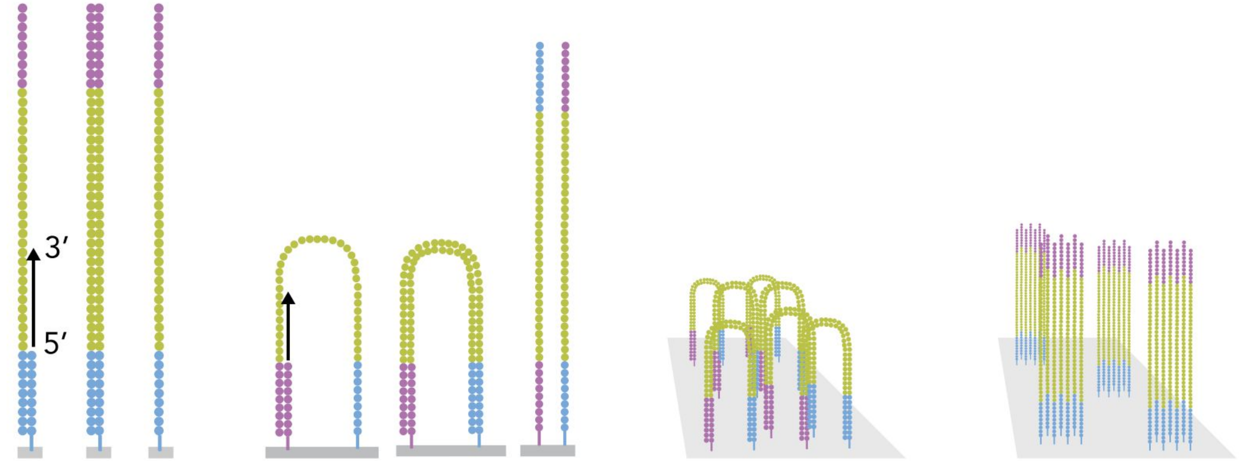
Illumina sequencing



flowcell

glass slide with 8 lanes





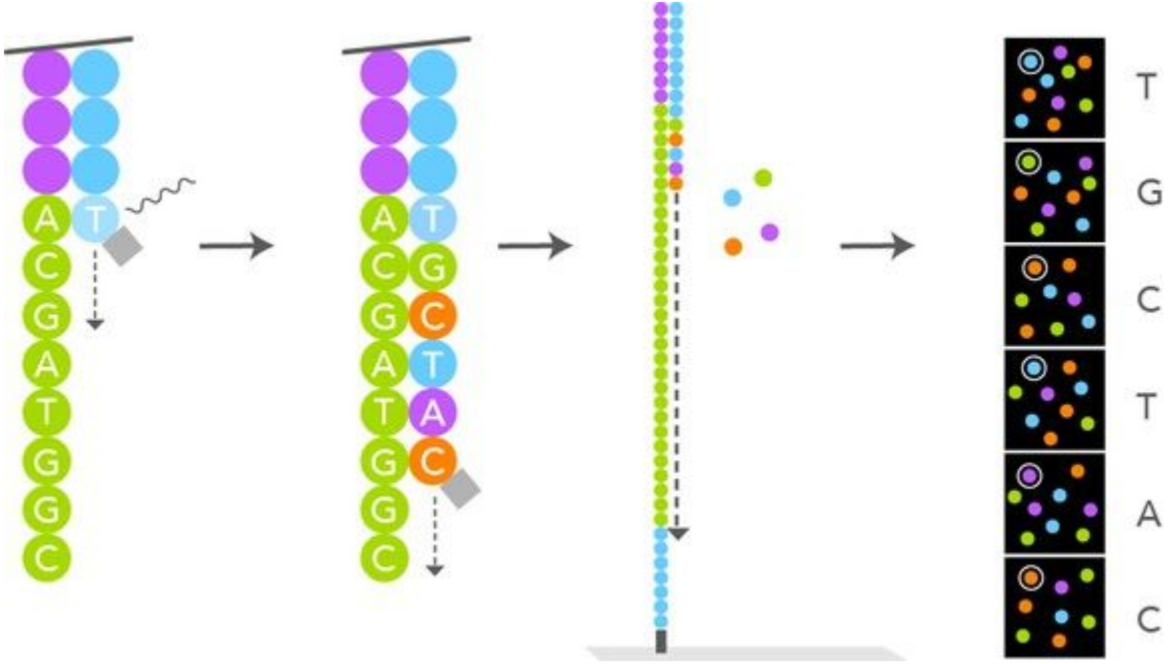
surface coated by millions of terminal sequences

bridge amplification

cluster generation

https://www.illumina.com/documents/products/datasheets/datasheet_cbot.pdf

Illumina sequencing



Each nucleotide is tagged with a different fluorophore.
Nucleotides are reversibly blocked => only one nucleotide can be added per cycle.

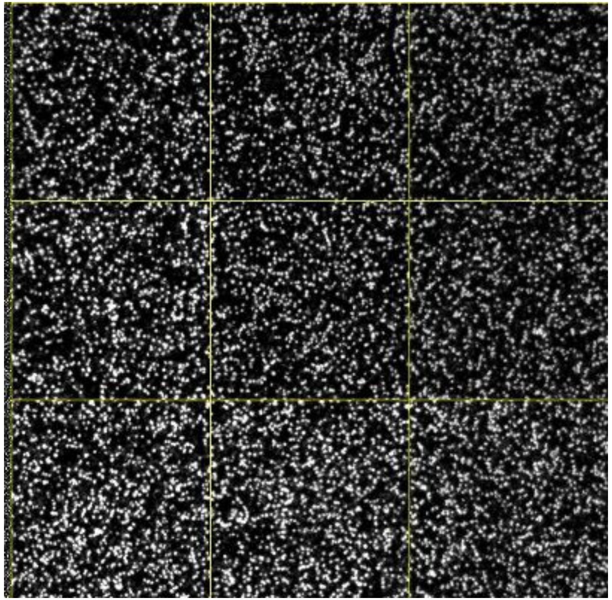
https://youtu.be/fCd6B5HRaZ8

Illumina sequencing

Each cycle the fluorescence is recorded across the flow cell, separately for each nucleotide: **TIFF files**.

Each image is analysed to identify clusters and quantify the intensity level.

A **base calling** algorithm uses cluster intensities and noise estimates to output a base for each cycle in each cluster, with an associated quality score: **BCL files**.

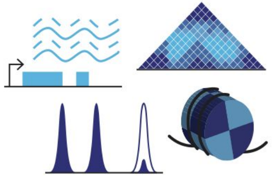


Applications of NGS

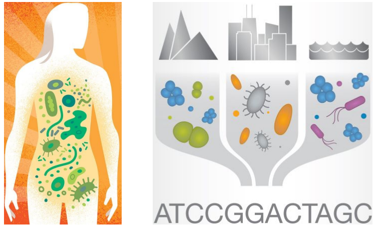
Resequencing: allows cataloguing variation among individuals of the same species.

Clinical applications: prenatal testing to identify trisomies. 

As a **molecular counter**: RNA expression, transcription factor binding, chromatin accessibility.



Metagenome sequencing: environmental or organism microbiomes.



Third generation sequencing

PACIFIC BIOSCIENCES®

https://www.pacb.com

Oxford NANOPORE Technologies

https://nanoporetech.com/

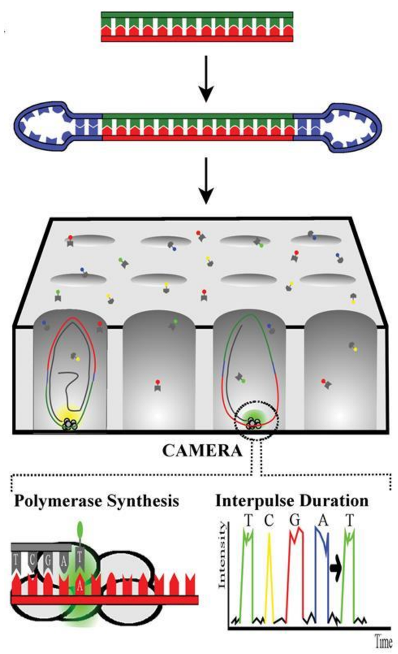
MinION

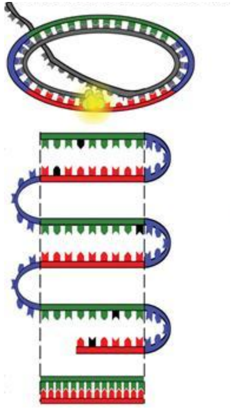
PromethION

Single-Molecule Real Time

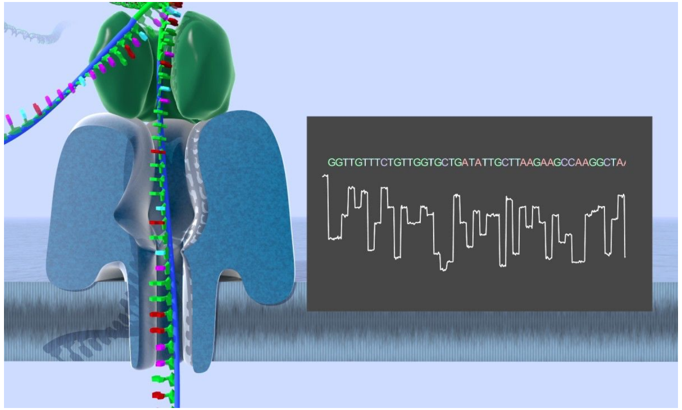
SMRT-sequencing

Nanopore sequencing





Error correction (1.7%)



- Long reads: tens to hundreds of kb.
- High error rates: ~13-15%.
- PCR-free.

https://www.youtube.com/watch?time_continue=29&v=WMZmG00uhwU

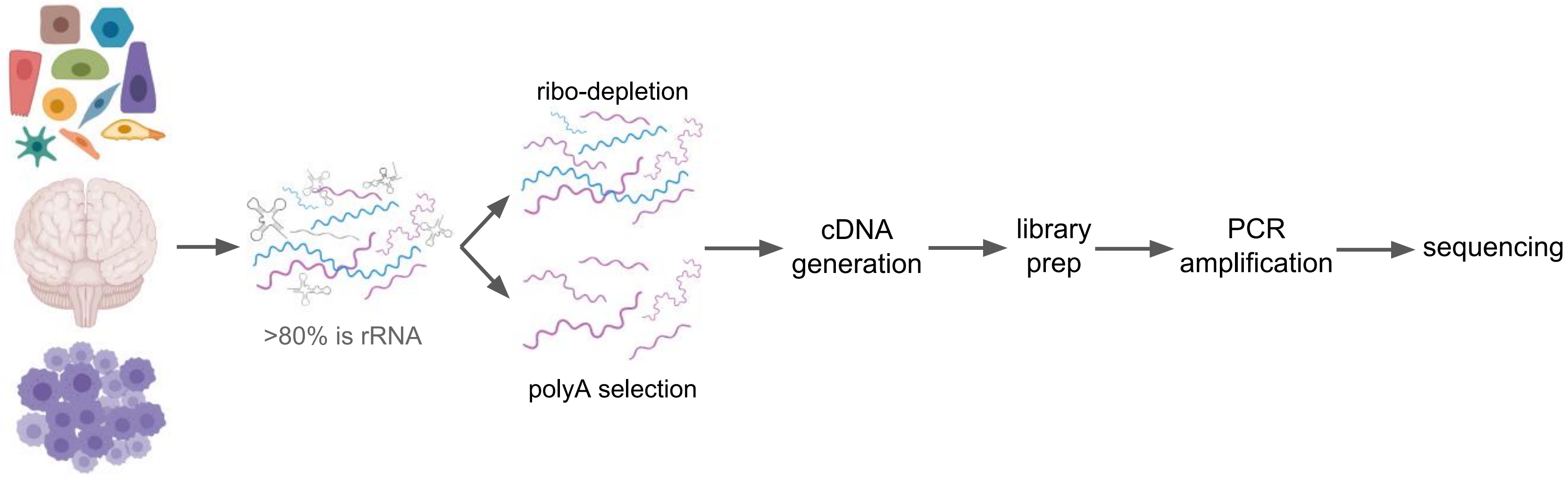
https://www.youtube.com/watch?v=E9-Rm5AoZGw

RNA sequencing

Anna Cuomo
EBI & University of Cambridge

Ximena Ibarra-Soria
Cancer Research UK

Experimental workflow



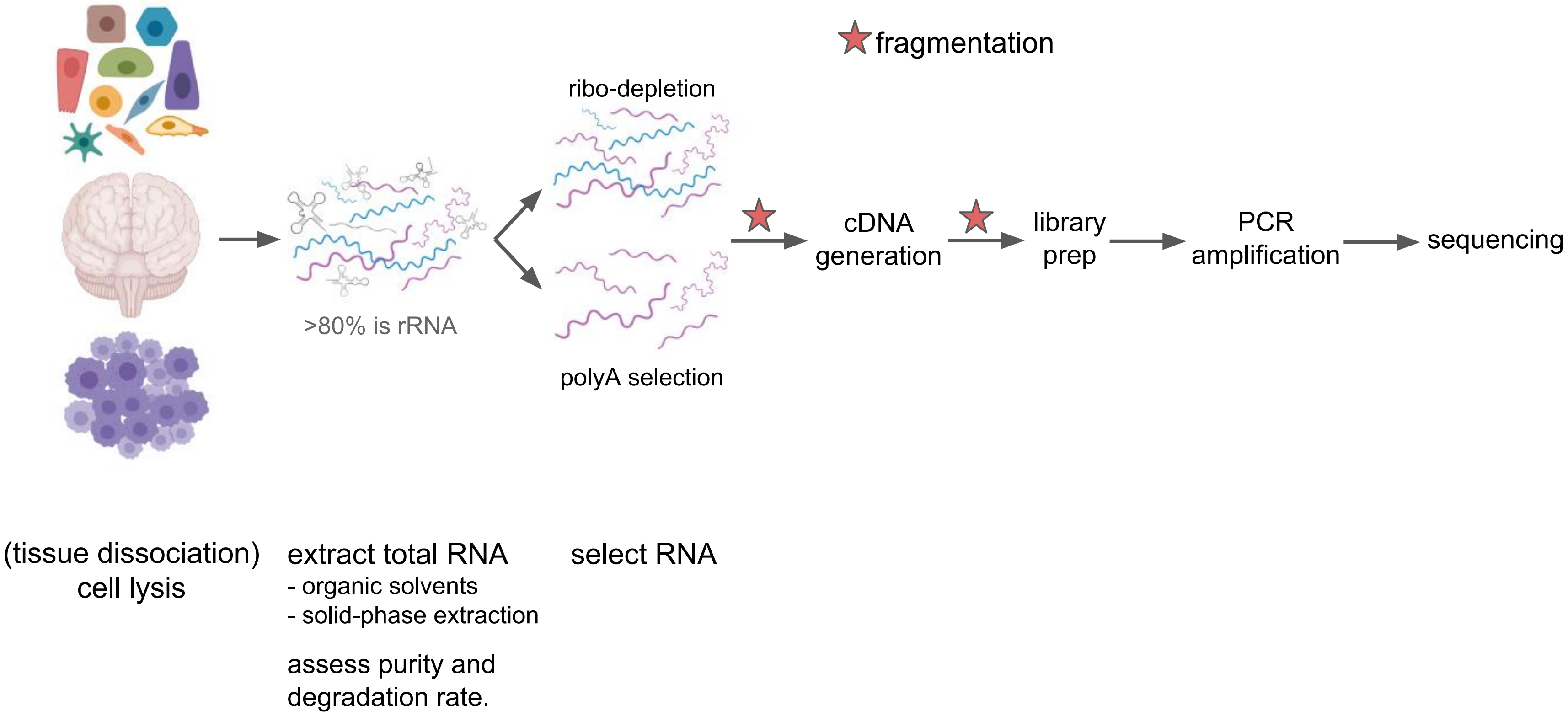
(tissue dissociation)
cell lysis

extract total RNA
- organic solvents
- solid-phase extraction

assess purity and
degradation rate.

select RNA

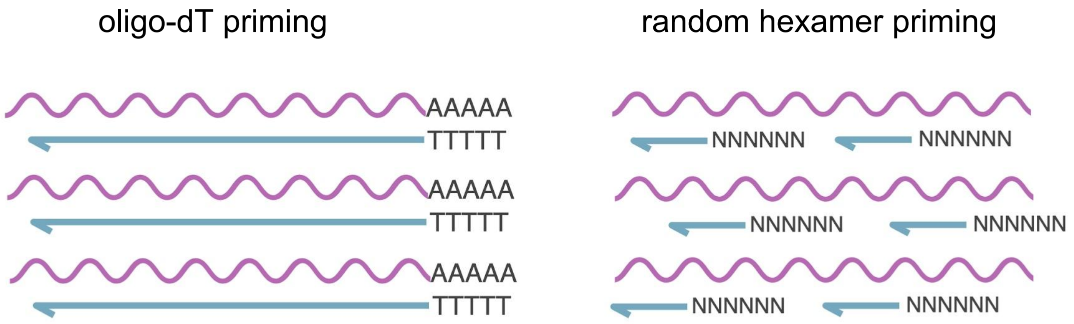
Experimental workflow



cDNA generation

RNA needs to be reverse-transcribed into cDNA which can then be sequenced with standard technologies.

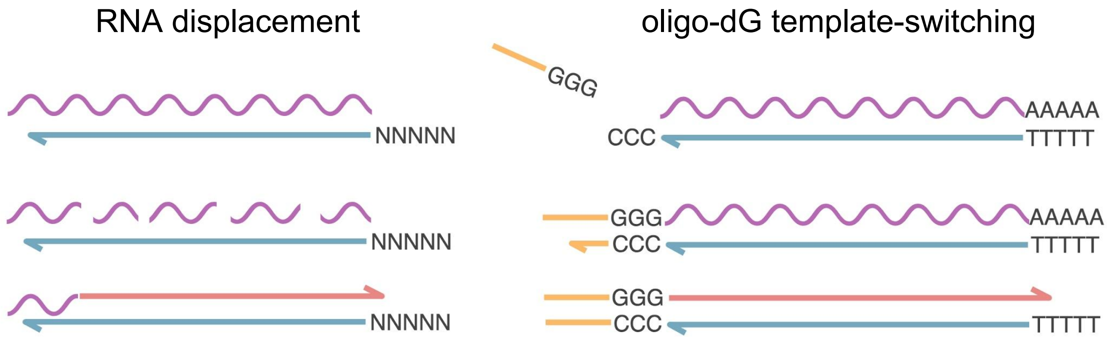
First-strand synthesis:



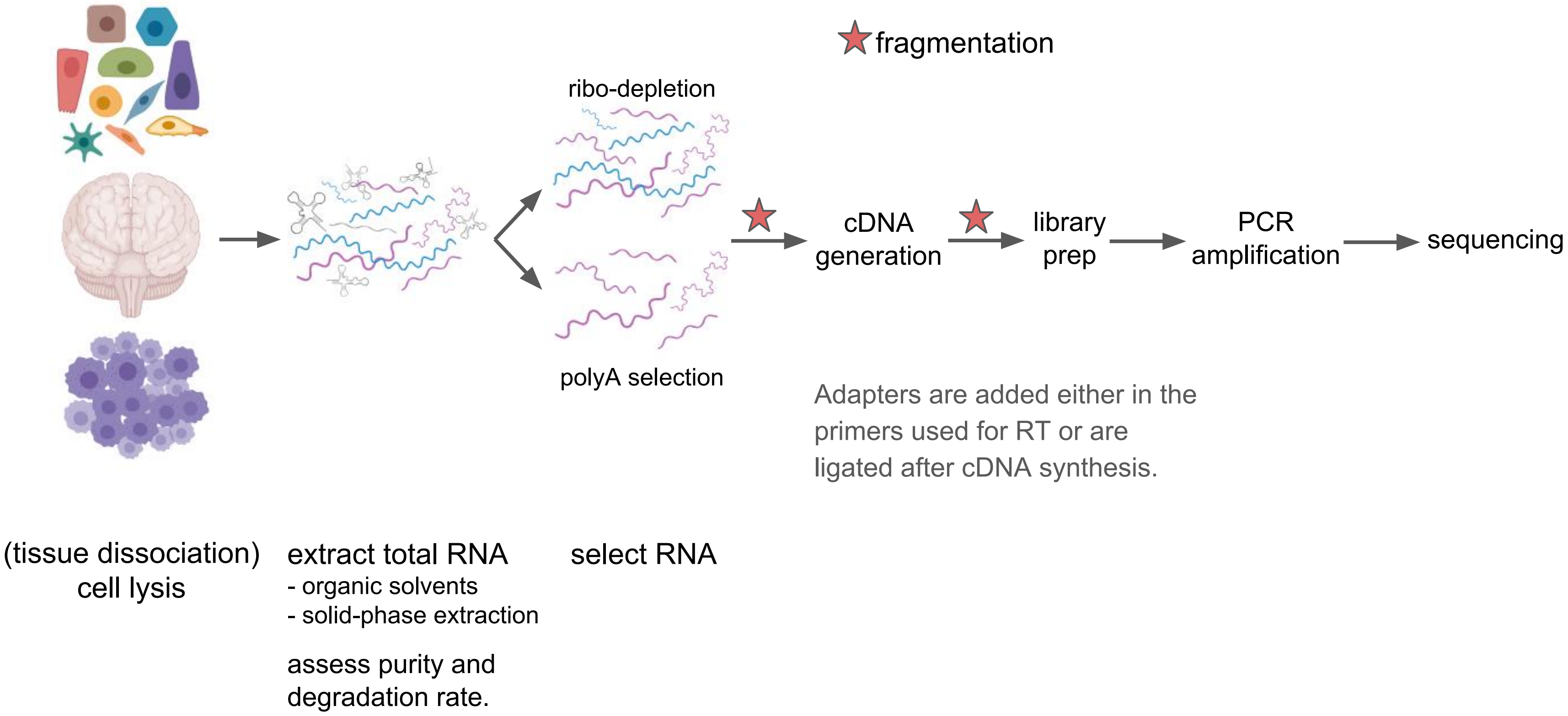
cDNA generation

RNA needs to be reverse-transcribed into cDNA which can then be sequenced with standard technologies.

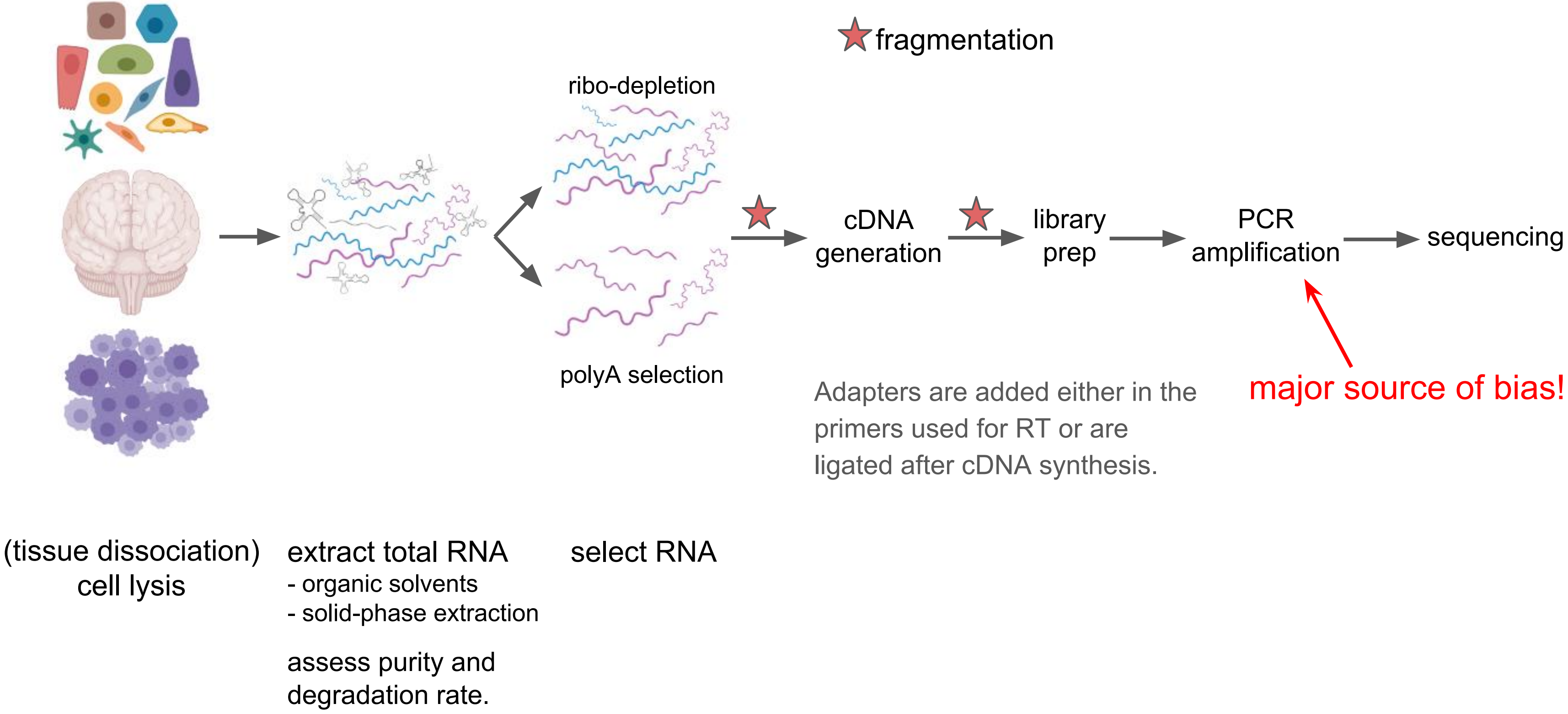
Second-strand synthesis:



Experimental workflow



Experimental workflow



PCR amplification bias

PCR amplification of the library introduces several biases.

Molecules with particular characteristics amplify with different efficiencies.

- Length.
- GC content.
- Secondary structure.

Plus, PCR has a stochastic component that affects more low-abundance species.

PCR amplification bias

PCR duplicates are normally defined as any group of reads with identical 5' mapping position.

Assumption: when DNA is randomly fragmented the probability of capturing two molecules starting at the same position is very low.

Only one alignment is retained.

MarkDuplicates from Picard tools. https://broadinstitute.github.io/picard/command-line-overview.html#MarkDuplicates

This **doesn't hold for RNA-seq** or when fragmentation is not random (restriction enzymes).

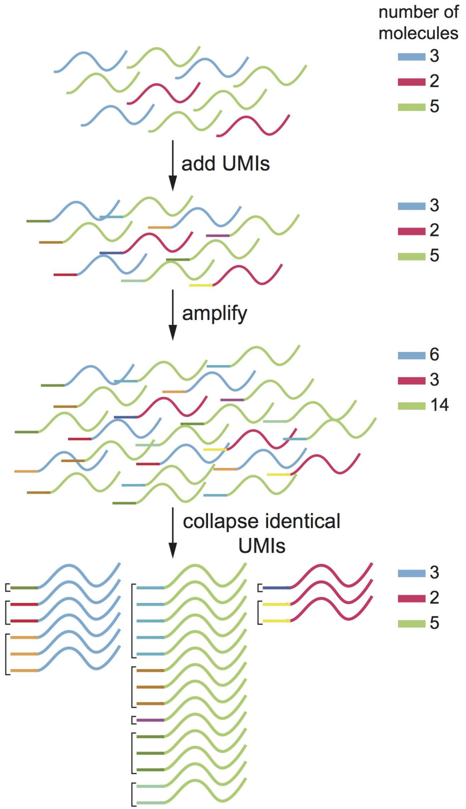
Unique Molecular Identifiers (UMIs)

To mitigate PCR biases, each molecule present in the initial sample needs to be made unique.

By adding a random barcode = unique molecular identifier (UMI).

Used for counting accurately.

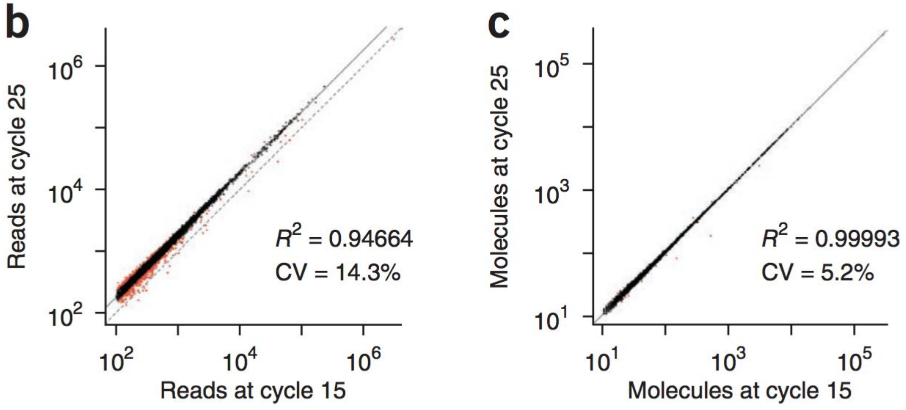
Full-transcript coverage is lost.
Only one end of the RNA is read.

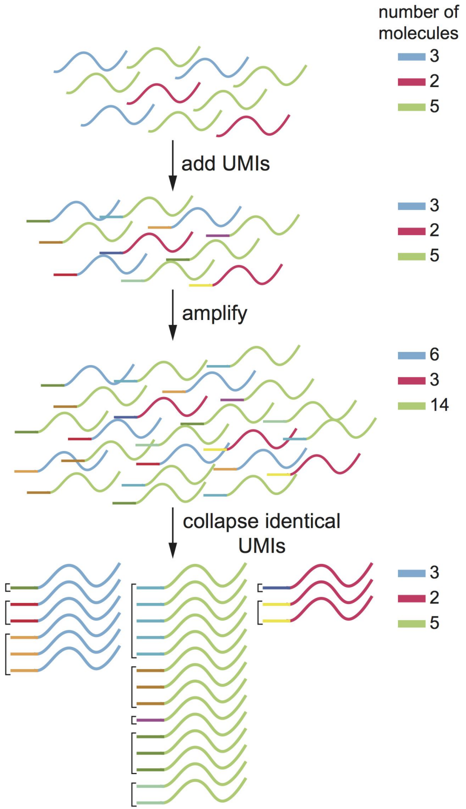


Unique Molecular Identifiers (UMIs)

To mitigate PCR biases, each molecule present in the initial sample needs to be made unique.

By adding a random barcode = unique





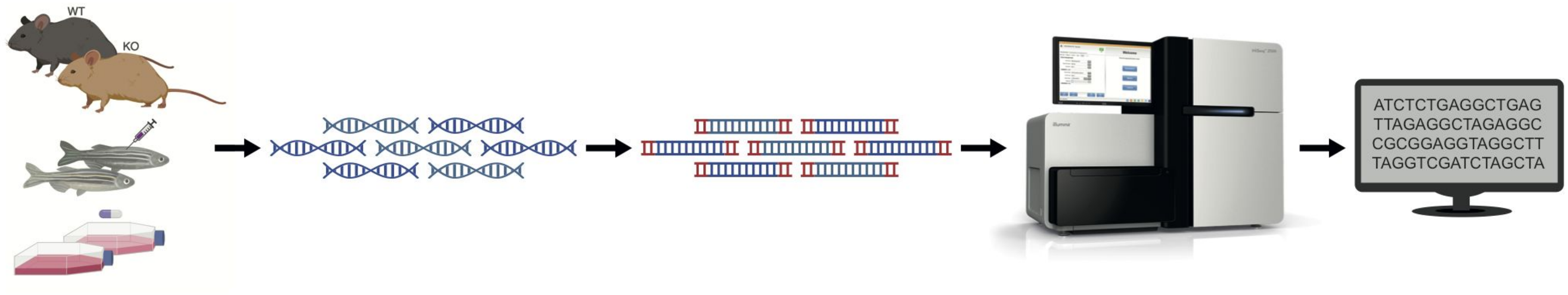
Kivioja et al., *Counting absolute numbers of molecules using unique molecular identifiers*, Nature Methods (2012). doi:10.1038/nmeth.1778

High-throughput sequencing experiments

Anna Cuomo
EBI & University of Cambridge

Ximena Ibarra-Soria
Cancer Research UK

High-throughput sequencing experiments



sample collection

extraction of DNA, RNA, chromatin...

library prep

sequencing

data analysis

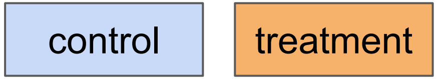
Experimental design

What is the question to answer.

control | treatment

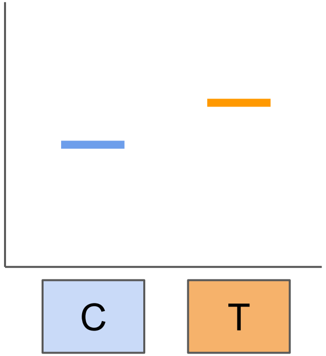
- Sources of variation.
  - Biological: gender, age, ethnicity, genetic background…
  - Technical: sample processing date, reagent’s batch, time of sample collection...
- To estimate variation we need **replicates**.

Experimental design



What is the question to answer.

- Sources of variation.
  - Biological: gender, age, ethnicity, genetic background…
  - Technical: sample processing date, reagent’s batch, time of sample collection...
- To estimate variation we need **replicates**.

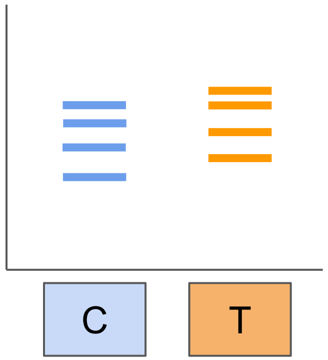


Experimental design

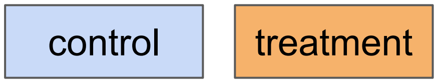
control treatment

What is the question to answer.

- Sources of variation.
    - Biological: gender, age, ethnicity, genetic background…
    - Technical: sample processing date, reagent’s batch, time of sample collection...
- To estimate variation we need **replicates**.

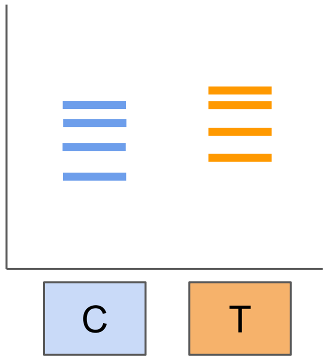


Experimental design

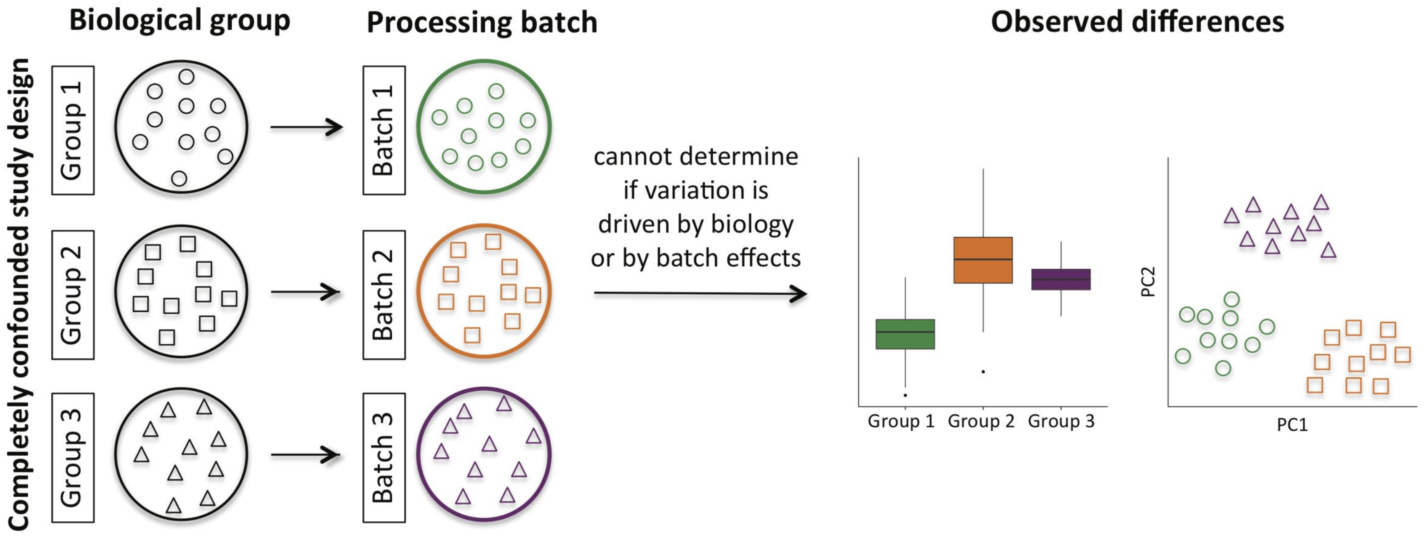


What is the question to answer.

- Sources of variation.
  - Biological: gender, age, ethnicity, genetic background…
  - Technical: sample processing date, reagent’s batch, time of sample collection...
- To estimate variation we need **replicates**.
- Power calculations.
  - Number of replicates needed to observe an effect.
- Type of information needed.
  - Sequencing platform.
  - Sequencing depth.
  - Single vs paired-end.

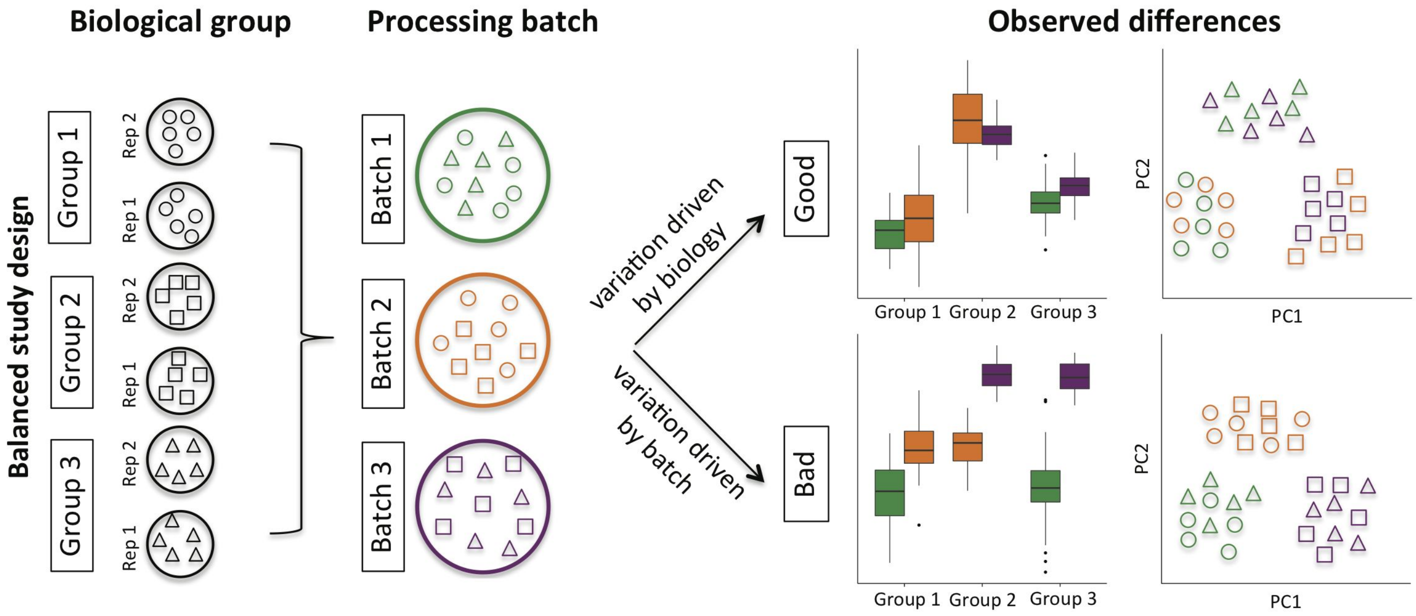


Experimental design



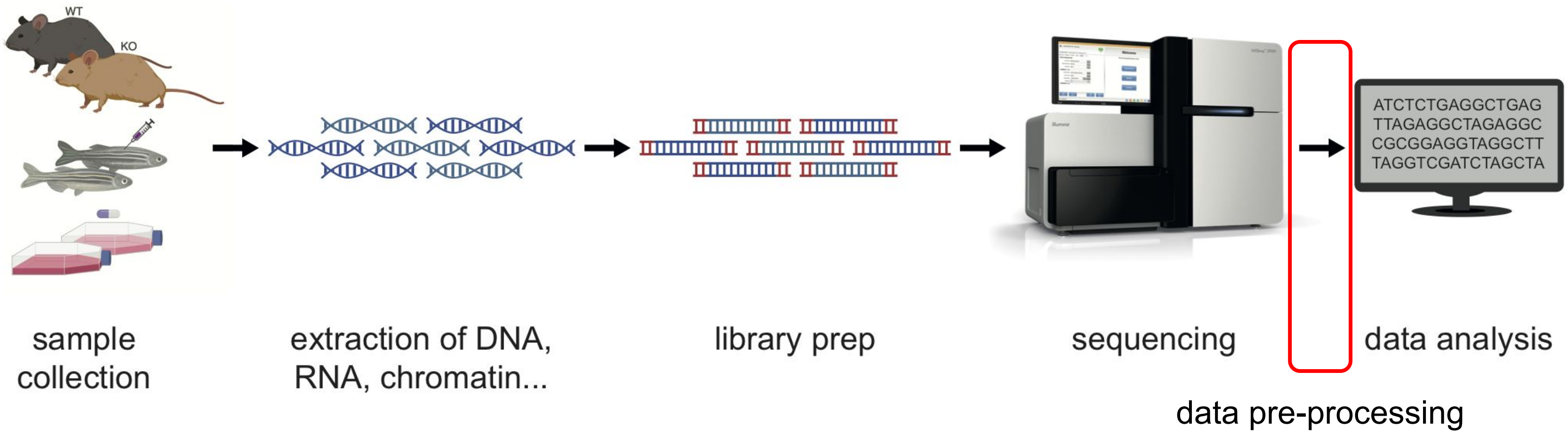
Adapted from Hicks et al., *On the widespread and critical impact...,* BioRxiv, version from December 27, 2015 - 06:15
doi: https://doi.org/10.1101/025528

Experimental design



Adapted from Hicks et al., *On the widespread and critical impact...*, BioRxiv, version from December 27, 2015 - 06:15
doi: https://doi.org/10.1101/025528

High-throughput sequencing experiments



Sequencing data: FASTQ files

BCL files -> FASTQ files (bcl2fastq conversion software (Illumina)).

Performs demultiplexing also.

FASTQ format: stores the nucleotide sequence with its associated quality.

header	@SEQ_ID
sequence	GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
	+
quality	!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65

Sequencing data: FASTQ files

BCL files -> FASTQ files (bcl2fastq conversion software (Illumina)).

Performs demultiplexing also.

FASTQ format: stores the nucleotide sequence with its associated quality.

header
sequence

quality

```
@SEQ_ID
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65
```

```
@K00252:342:HWCHVBBXX:6:1101:19715:1859 1:N:0:TAAGGCGA+TCTTACGC
```

instrument
run ID
flowcell ID
lane
tile
x coord
y coord
pair
index sequence

Sequencing data: FASTQ files

BCL files -> FASTQ files (bcl2fastq conversion software (Illumina)).

Performs demultiplexing also.

FASTQ format: stores the nucleotide sequence with its associated quality.

header	@SEQ_ID
sequence	GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
	+
quality	!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65

Quality scores indicate the probability (*p*) of the base call being wrong.

$Q = -10 \log_{10} p$ *Phred quality score*

They are encoded in ASCII, by adding 64 to the quality value.

Sequencing reads

Sequencing reads can have several quality issues.

- Adaptor contamination.
- Systematic failure at specific cycles.
- Substantially lower quality at the end of the read.

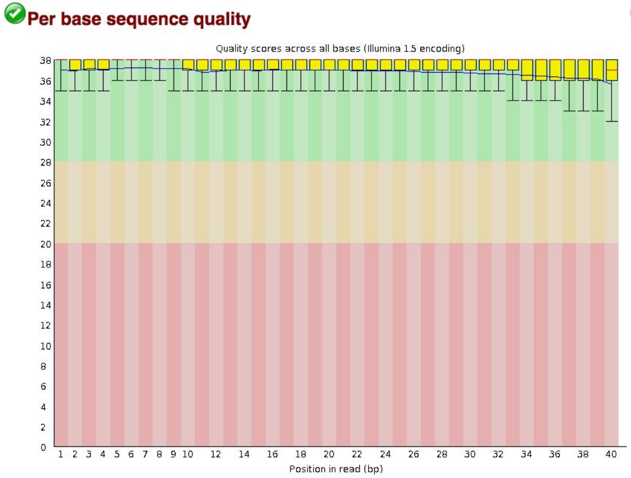
A sequencing library can also have quality issues that can be spotted from the sequencing data.

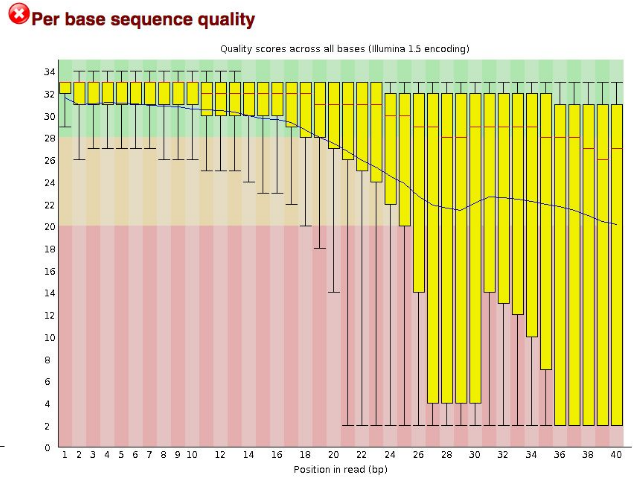
- Low complexity resulting in high number of PCR duplicates.

Sequencing reads

Initial QC is a good sanity check about data quality.

FastQC [http://www.bioinformatics.babraham.ac.uk/projects/fastqc/] reports on basic quality statistics.



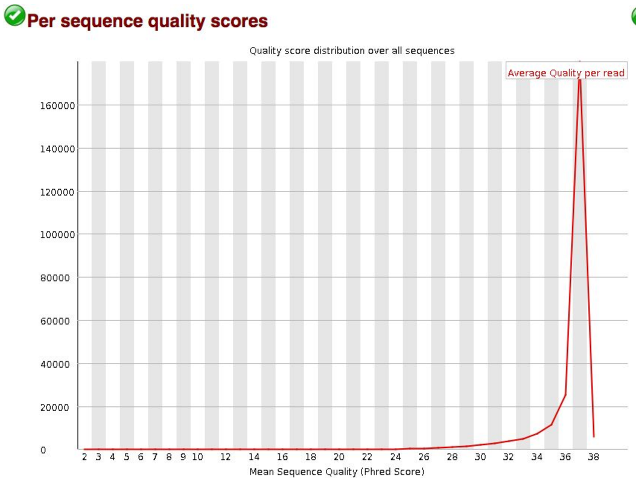


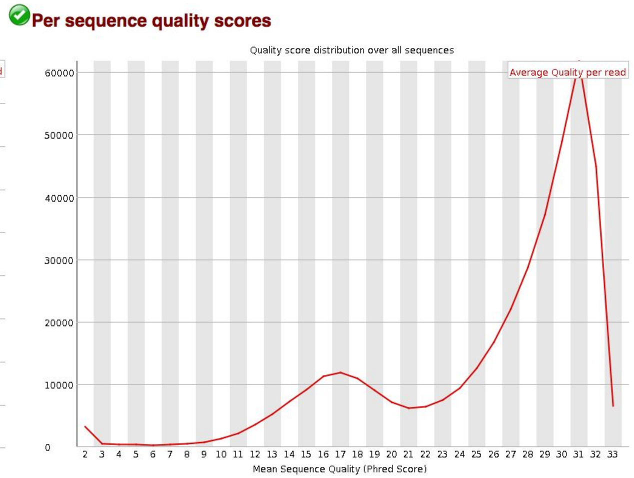
Consider trimming the reads to remove the low-quality portion.

Sequencing reads

Initial QC is a good sanity check about data quality.

FastQC [http://www.bioinformatics.babraham.ac.uk/projects/fastqc/] reports on basic quality statistics.

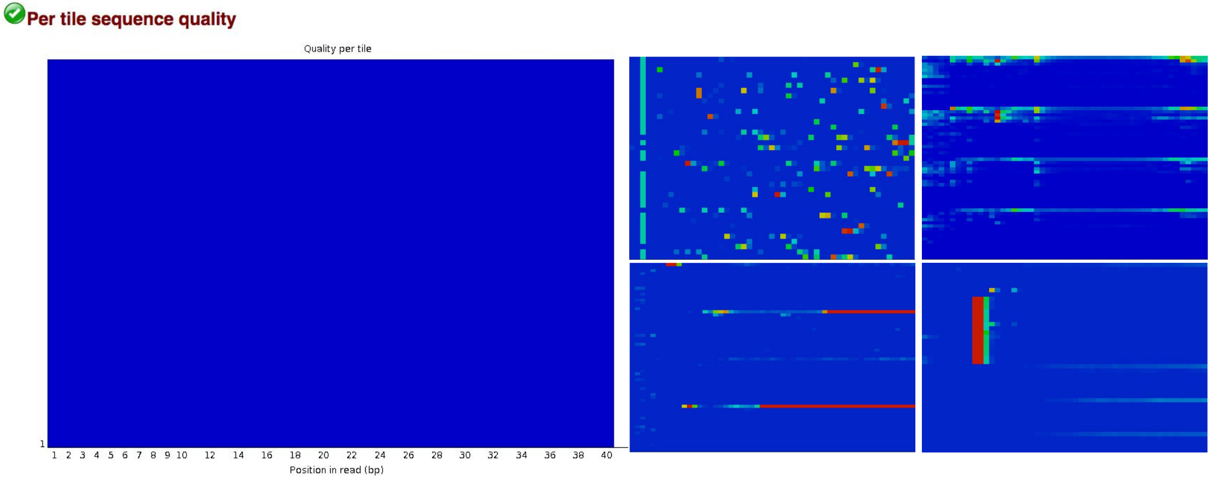




Sequencing reads

Initial QC is a good sanity check about data quality.

FastQC [http://www.bioinformatics.babraham.ac.uk/projects/fastqc/] reports on basic quality statistics.

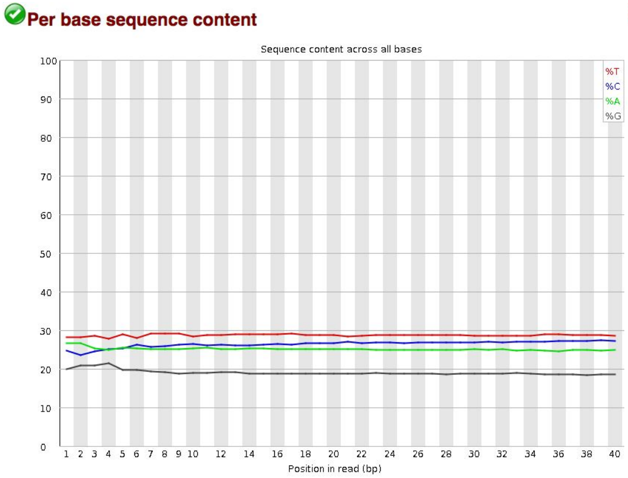


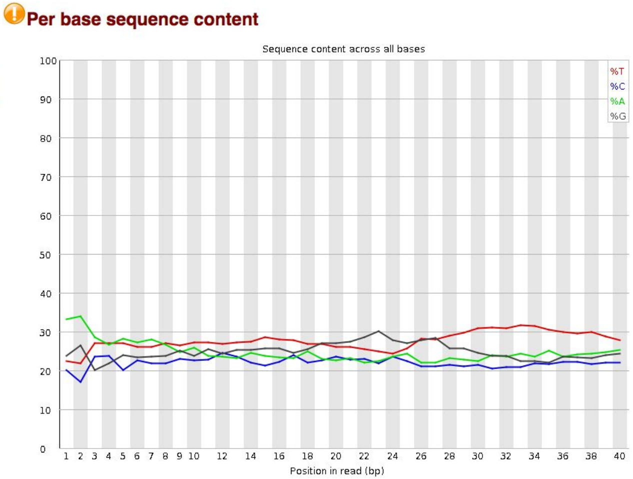
https://sequencing.qcfail.com/articles/position-specific-failures-of-flowcells/

Sequencing reads

Initial QC is a good sanity check about data quality.

FastQC [http://www.bioinformatics.babraham.ac.uk/projects/fastqc/] reports on basic quality statistics.

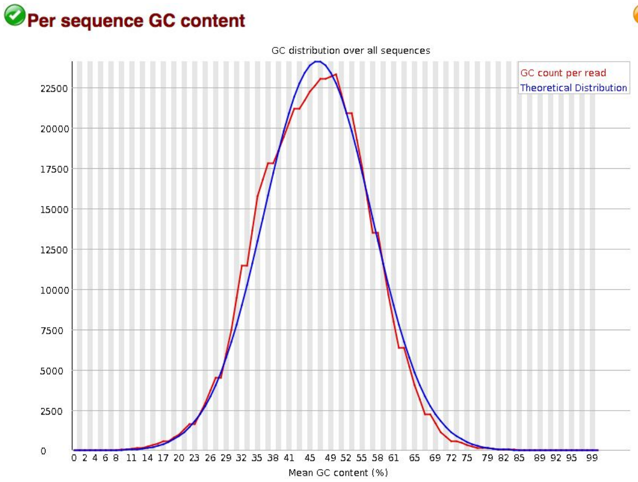


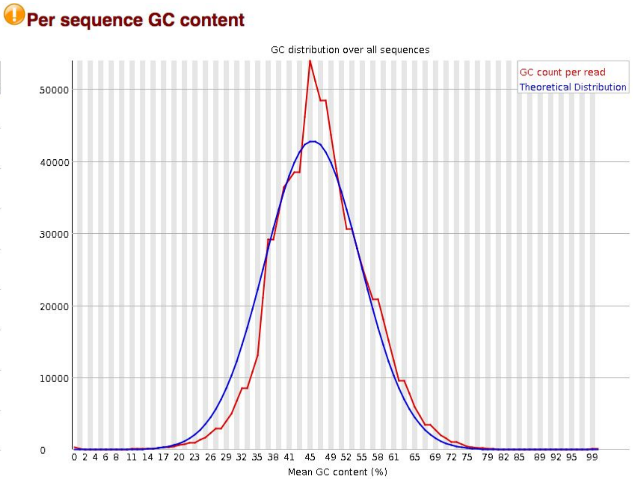


Sequencing reads

Initial QC is a good sanity check about data quality.

FastQC [http://www.bioinformatics.babraham.ac.uk/projects/fastqc/] reports on basic quality statistics.

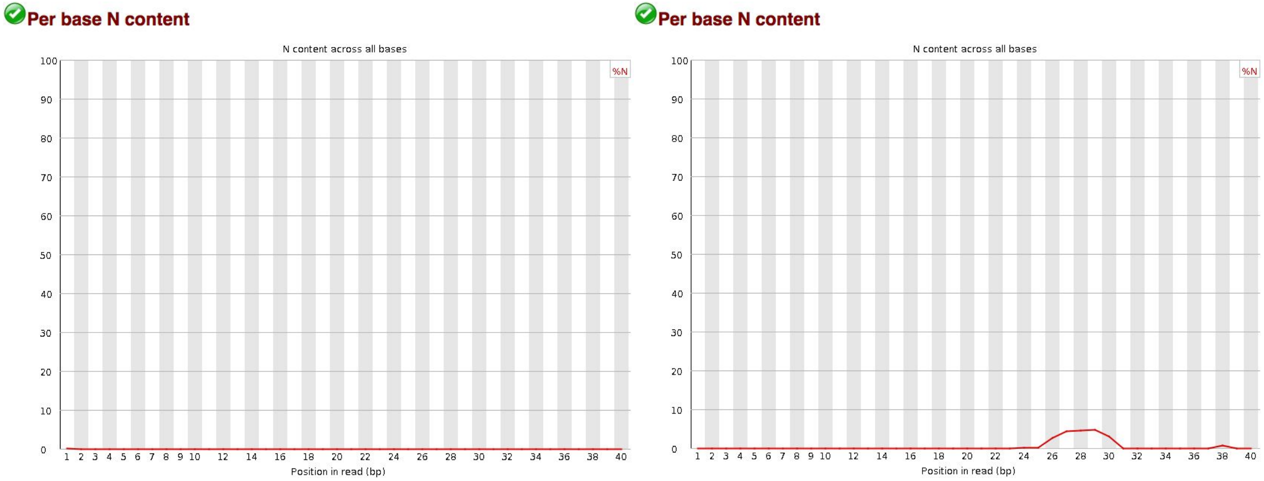




Sequencing reads

Initial QC is a good sanity check about data quality.

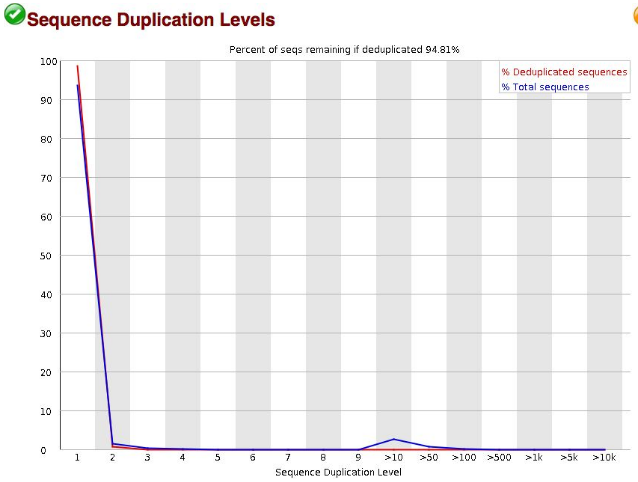
FastQC [http://www.bioinformatics.babraham.ac.uk/projects/fastqc/] reports on basic quality statistics.

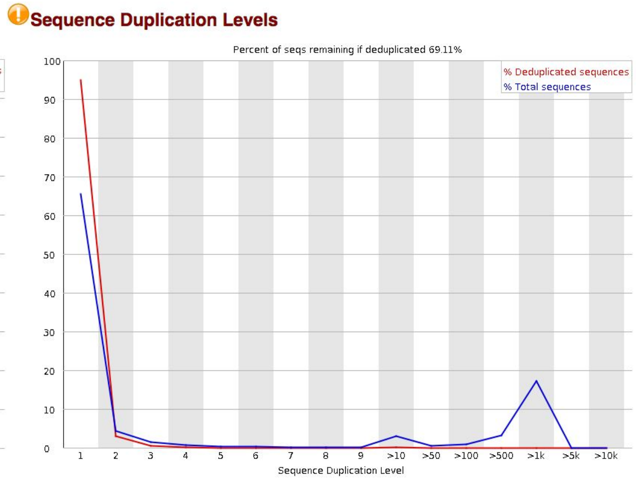


Sequencing reads

Initial QC is a good sanity check about data quality.

FastQC [http://www.bioinformatics.babraham.ac.uk/projects/fastqc/] reports on basic quality statistics.





Sequencing reads

Initial QC is a good sanity check about data quality.

FastQC [http://www.bioinformatics.babraham.ac.uk/projects/fastqc/] reports on basic quality statistics.

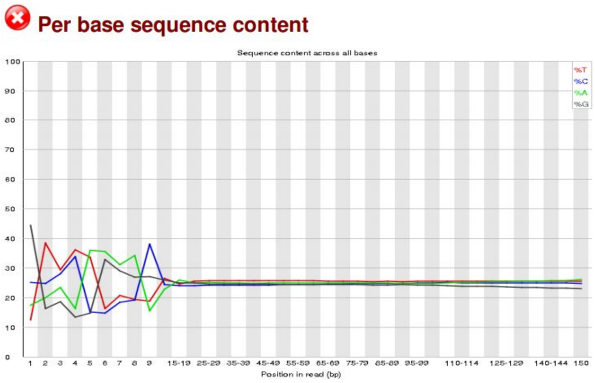
Overrepresented sequences

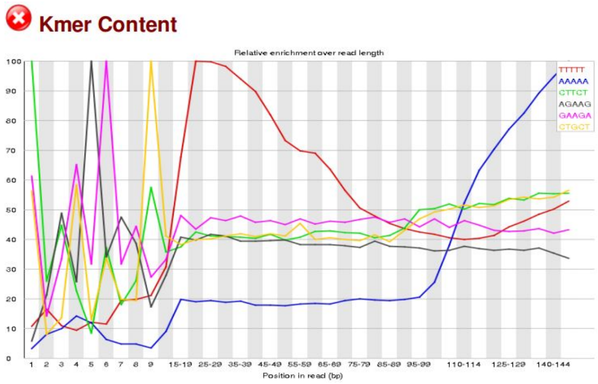
Sequence	Count	Percentage	Possible Source
GATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGATCT	8122	8.122	Illumina Paired End PCR Primer 2 (100% over 40bp)
GATCGGAAGAGCGGTTCAGCAGGAATGCCGAGATCGGAAG	5086	5.086	Illumina Paired End PCR Primer 2 (97% over 36bp)
AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTAC	1085	1.085	Illumina Single End PCR Primer 1 (100% over 40bp)
GATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGGAAG	508	0.508	Illumina Paired End PCR Primer 2 (97% over 36bp)
AATTATACGGCGACCACCGAGATCTACACTCTTTCCCTAC	242	0.242	Illumina Single End PCR Primer 1 (97% over 40bp)
GATCGGAAGAGCGGTTCAGCAGGAATGCCGAAGATCGGAA	235	0.23500000000000001	Illumina Paired End Adapter 2 (96% over 31bp)
GATCGGAAGAGCGGTTCAGCAGGAATGCGAGATCGGAAGA	228	0.22799999999999998	Illumina Paired End Adapter 2 (96% over 28bp)
GATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGGACG	205	0.20500000000000002	Illumina Paired End PCR Primer 2 (97% over 36bp)
GATCGGAAGAGCGGTTCAGCAGGAATGCCGAGGATCGGAA	183	0.183	Illumina Paired End Adapter 2 (100% over 32bp)
GATCGGAAGAGCGGTTCAGCAGGAATGCCGAGGTCGGAAG	183	0.183	Illumina Paired End Adapter 2 (100% over 32bp)
GATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGAACT	164	0.164	Illumina Paired End PCR Primer 2 (97% over 40bp)
GATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGGTCT	129	0.129	Illumina Paired End PCR Primer 2 (97% over 40bp)
AATTATACTTCTACCACCTATATCTACACTCTTTCCCTAC	123	0.123	No Hit
GATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGGACT	122	0.122	Illumina Paired End PCR Primer 2 (97% over 36bp)
CGGTTCAGCAGGAATGCCGAGATCGGAAGAGCGGTTCAGC	113	0.11299999999999999	Illumina Paired End PCR Primer 2 (96% over 25bp)

Sequencing reads - RNA-seq

RNA-seq data has a few particular characteristics not shared with DNA-seq data.

Random hexamer priming introduces biased nucleotide composition in the first ~13 nucleotides of the reads.





Hansen et al., *Biases in Illumina transcriptome sequencing caused by random hexamer priming*, Nucleic Acids Res (2010)
doi: https://doi.org/10.1093/nar/gkq224

Alignment to a reference genome

Objective: find the true location in the genome where a sequencing read *came from*.

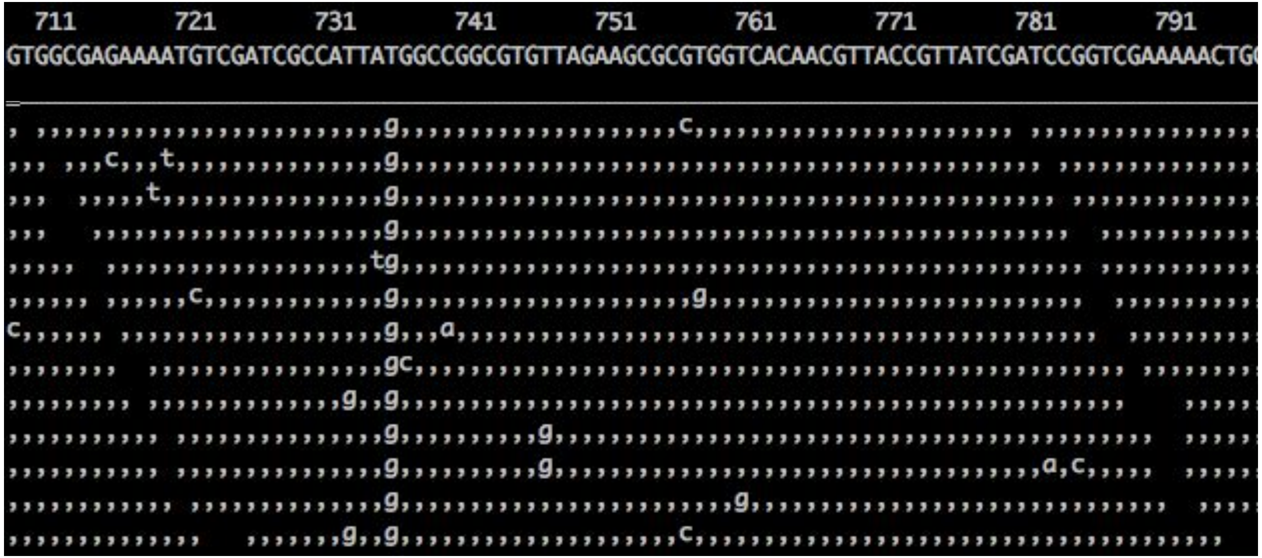
Challenges:

- Millions of short reads.
- Large search space.
  - Human haploid genome: 3,234.83 Mb
  - Mouse haploid genome: 2,653.99 Mb
- Matching needs to allow errors.

Alignment to a reference genome

reference

sequencing reads



should be able to handle mismatching bases and gaps → - PCR and sequencing errors
- genetic variation

Alignment to a reference genome

To address the large input size problem (millions of reads and a large reference):

- **Filtering:** quickly exclude large reference regions where matches cannot be found.
  - Take a substring of the read (**seed**) and find perfect matches.

- **Indexing:** involves pre-processing the reference to speed-up matching without scanning the whole reference.
  - Hash tables.
  - Suffix trees.
  - FM indices with Burrows-Wheeler transform.

Reinert et al., *Alignment of Next-Generation Sequencing Reads*, Annu. Rev. Genomics Hum. Genet. (2015).
doi: https://doi.org/110.1146/annurev-genom-090413-025358

Alignment to a reference genome

Each seed has a list of candidate matches in the genome.

The region around each is examined to determine if a high-scoring alignment exists.

Mapping quality: measures the confidence of the alignment by considering all possible locations discovered.

p_{cor} = probability alignment is correct

$Q = -10 \log_{10} (1-p_{cor})$

Q = 30 1 in 1000 chance the alignment is wrong.

Reinert et al., *Alignment of Next-Generation Sequencing Reads*, Annu. Rev. Genomics Hum. Genet. (2015).
doi: https://doi.org/110.1146/annurev-genom-090413-025358

Alignment to a reference genome

The biggest problem for aligners comes from the high repeat content of most eukaryotic genomes.

multi-mapping reads: reads that align equally well at two or more loci.

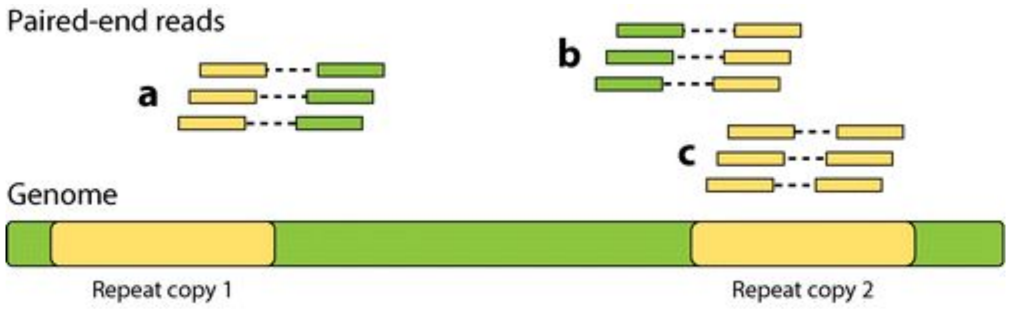
Reinert et al., *Alignment of Next-Generation Sequencing Reads*, Annu. Rev. Genomics Hum. Genet. (2015).
doi: https://doi.org/110.1146/annurev-genom-090413-025358

Alignment to a reference genome

The biggest problem for aligners comes from the high repeat content of most eukaryotic genomes.

multi-mapping reads: reads that align equally well at two or more loci.

paired-end reads help reducing multimappers.



Reinert et al., *Alignment of Next-Generation Sequencing Reads*, Annu. Rev. Genomics Hum. Genet. (2015).
doi: https://doi.org/110.1146/annurev-genom-090413-025358

Alignment to a reference genome

The biggest problem for aligners comes from the high repeat content of most eukaryotic genomes.

multi-mapping reads: reads that align equally well at two or more loci.

paired-end reads help reducing multimappers.

When the sample comes from a genome *substantially* different to the reference, the alignment becomes less accurate and there is information loss.

Relaxing the stringency of the alignments might be necessary.

If known, consider imputing the variable positions.

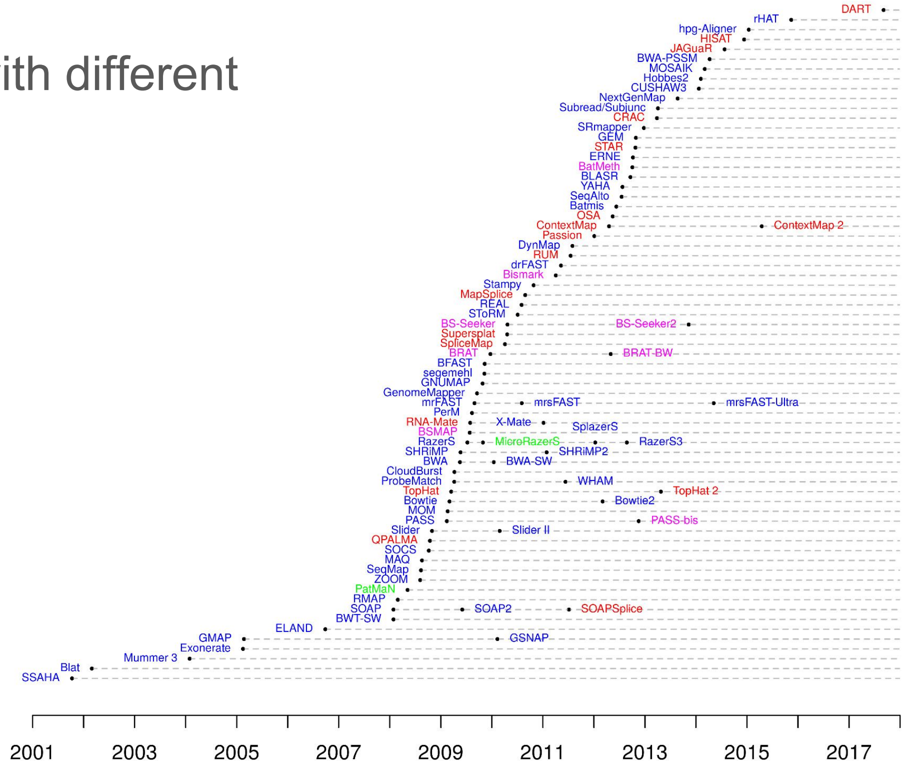
https://www.sanger.ac.uk/science/data/mouse-genomes-project

Reinert et al., *Alignment of Next-Generation Sequencing Reads*, Annu. Rev. Genomics Hum. Genet. (2015).
doi: https://doi.org/110.1146/annurev-genom-090413-025358

NGS aligners

There are dozens of different aligners with different

- indexing methods.
- scoring criteria.
- memory requirements.
- speed.
- ...



https://www.ebi.ac.uk/~nf/hts_mappers/

NGS aligners

Hash tables.

GSNAP, MAQ, RMAP, subread*.

* Can be used from within R with the **Rsubread** package.
https://bioconductor.org/packages/release/bioc/html/Rsubread.html

Burrows-Wheeler Transform (BWT).

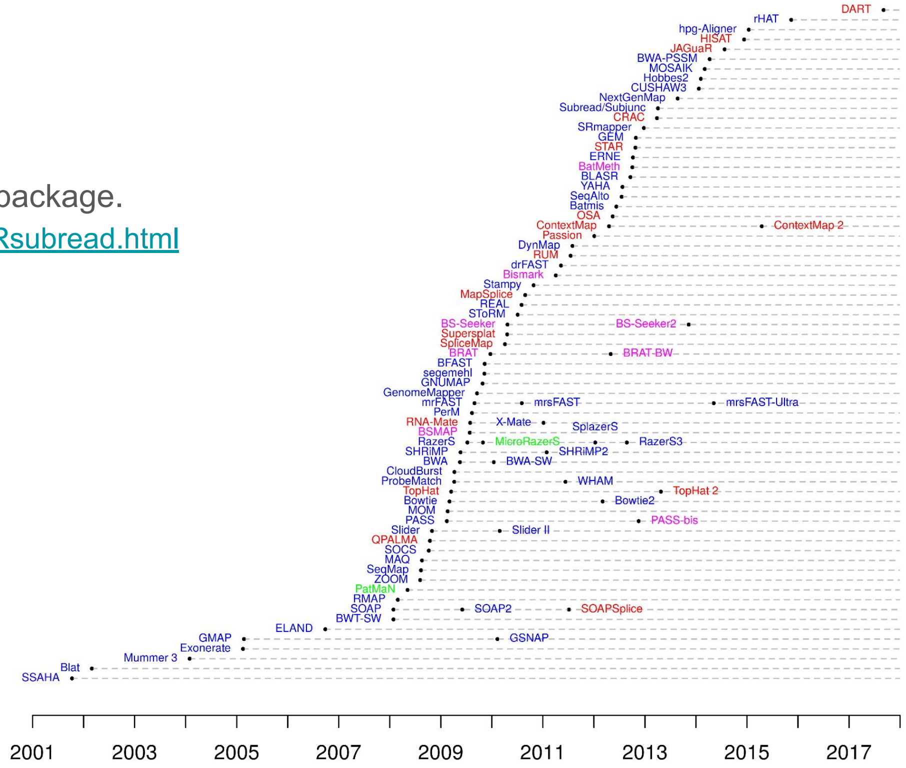
Bowtie, BWA, SOAP2.

There is no *best* aligner.

Each is suited to different types of data.

Adjust the parameters to reflect this.

Keep it consistent.



https://www.ebi.ac.uk/~nf/hts_mappers/

Alignment files: SAM/BAM/CRAM

The **Sequence Alignment/Map (SAM) format** is a tab-delimited text format to store genomic alignments. Contains two sections.

Header section:

Header lines start with @.

Information is encoded by TAG:VALUE entries.

@HD	**header line**. Version, sorting/grouping of alignments.
@SQ	**reference sequence dictionary**. Sequence name and length. Genome assembly, species...
@RG	**read group**. Barcode identifying the sample. Sequencing centre, date, platform, median insert size...
@PG	**program**. Program name, version, command line.
@CO	**comment line**.

Alignment files: SAM/BAM/CRAM

The **Sequence Alignment/Map (SAM) format** is a tab-delimited text format to store genomic alignments. Contains two sections.

Header section: information is encoded by TAG:VALUE entries.

header

reference sequence dictionary

read group

program

comment line

```
@HD   VN:1.5      SO:coordinate
@SQ   SN:1        LN:195471971
@SQ   SN:10       LN:130694993
@SQ   SN:11       LN:122082543
@SQ   SN:12       LN:120129022
...
@SQ   SN:JH584292.1    LN:14945
@SQ   SN:JH584295.1    LN:1976
@RG   ID:1  PL:illumina      PU:1  LB:do9029  SM:do9029  CN:CRI
@PG   ID:bwa-E39E2AF   PN:bwa     VN:0.7.12-r1039  CL:bwa samse mm10.fa - wt1.fq
@PG   ID:MarkDuplicates     PN: MarkDuplicates    VN:1.139   CL:MarkDuplicates
INPUT=[wt1.bam] OUTPUT=temp.bam METRICS_FILE=metric.txt
REMOVE_DUPLICATES=false...
@CO   [optional]
```

https://samtools.github.io/hts-specs/SAMv1.pdf

Alignment files: SAM/BAM/CRAM

The **Sequence Alignment/Map (SAM) format** is a tab-delimited text format to store genomic alignments. Contains two sections.

Alignment section:

11 mandatory fields; always in the same order.

<table>
<tr><th>Col</th><th>Field</th><th>Type</th><th>Regexp/Range</th><th>Brief description</th></tr>
<tr><td>1</td><td>QNAME</td><td>String</td><td>[!-?A-~]{1,254}</td><td>Query template NAME</td></tr>
<tr><td>2</td><td>FLAG</td><td>Int</td><td>$[0, 2^{16}-1]$</td><td>bitwise FLAG</td></tr>
<tr><td>3</td><td>RNAME</td><td>String</td><td>*|[!-()+-<>-~][!-~]*</td><td>Reference sequence NAME</td></tr>
<tr><td>4</td><td>POS</td><td>Int</td><td>$[0, 2^{31}-1]$</td><td>1-based leftmost mapping POSition</td></tr>
<tr><td>5</td><td>MAPQ</td><td>Int</td><td>$[0, 2^{8}-1]$</td><td>MAPping Quality</td></tr>
<tr><td>6</td><td>CIGAR</td><td>String</td><td>*|([0-9]+[MIDNSHPX=])+</td><td>CIGAR string</td></tr>
<tr><td>7</td><td>RNEXT</td><td>String</td><td>*|=|[!-()+-<>-~][!-~]*</td><td>Ref. name of the mate/next read</td></tr>
<tr><td>8</td><td>PNEXT</td><td>Int</td><td>$[0, 2^{31}-1]$</td><td>Position of the mate/next read</td></tr>
<tr><td>9</td><td>TLEN</td><td>Int</td><td>$[-2^{31}+1, 2^{31}-1]$</td><td>observed Template LENgth</td></tr>
<tr><td>10</td><td>SEQ</td><td>String</td><td>*|[A-Za-z=.]+</td><td>segment SEQuence</td></tr>
<tr><td>11</td><td>QUAL</td><td>String</td><td>[!-~]+</td><td>ASCII of Phred-scaled base QUALity+33</td></tr>
</table>

https://samtools.github.io/hts-specs/SAMv1.pdf

Alignment files: SAM/BAM/CRAM

The **Sequence Alignment/Map (SAM) format** is a tab-delimited text format to store genomic alignments. Contains two sections.

Alignment section:

11 mandatory fields

Col	Field
1	QNAME
2	FLAG
3	RNAME
4	POS
5	MAPQ
6	CIGAR
7	RNEXT
8	PNEXT
9	TLEN
10	SEQ
11	QUAL

TRUE/FALSE for pre-defined criteria.

Bit		Description
1	0x1	template having multiple segments in sequencing
2	0x2	each segment properly aligned according to the aligner
4	0x4	segment unmapped
8	0x8	next segment in the template unmapped
16	0x10	SEQ being reverse complemented
32	0x20	SEQ of the next segment in the template being reverse complemented
64	0x40	the first segment in the template
128	0x80	the last segment in the template
256	0x100	secondary alignment
512	0x200	not passing filters, such as platform/vendor quality controls
1024	0x400	PCR or optical duplicate
2048	0x800	supplementary alignment

Explain SAM flags: https://broadinstitute.github.io/picard/explain-flags.html https://samtools.github.io/hts-specs/SAMv1.pdf

Alignment files: SAM/BAM/CRAM

The **Sequence Alignment/Map (SAM) format** is a tab-delimited text format to store genomic alignments. Contains two sections.

Alignment section:

11 mandatory fields; always in the same order.

<table>
<tr><th>Col</th><th>Field</th><th>Type</th><th>Regexp/Range</th><th>Brief description</th></tr>
<tr><td>1</td><td>QNAME</td><td>String</td><td>[!-?A-~]{1,254}</td><td>Query template NAME</td></tr>
<tr><td>2</td><td>FLAG</td><td>Int</td><td>$[0,2^{16}-1]$</td><td>bitwise FLAG</td></tr>
<tr><td>3</td><td>RNAME</td><td>String</td><td>*|[!-()+-<>-~][!-~]*</td><td>Reference sequence NAME</td></tr>
<tr><td>4</td><td>POS</td><td>Int</td><td>$[0,2^{31}-1]$</td><td>1-based leftmost mapping POSition</td></tr>
<tr><td>5</td><td>MAPQ</td><td>Int</td><td>$[0,2^{8}-1]$</td><td>MAPping Quality</td></tr>
<tr><td>6</td><td>CIGAR</td><td>String</td><td>*|([0-9]+[MIDNSHPX=])+</td><td>CIGAR string</td></tr>
<tr><td>7</td><td>RNEXT</td><td>String</td><td>*|=|[!-()+-<>-~][!-~]*</td><td>Ref. name of the mate/next read</td></tr>
<tr><td>8</td><td>PNEXT</td><td>Int</td><td>$[0,2^{31}-1]$</td><td>Position of the mate/next read</td></tr>
<tr><td>9</td><td>TLEN</td><td>Int</td><td>$[-2^{31}+1,2^{31}-1]$</td><td>observed Template LENgth</td></tr>
<tr><td>10</td><td>SEQ</td><td>String</td><td>*|[A-Za-z=.]+</td><td>segment SEQuence</td></tr>
<tr><td>11</td><td>QUAL</td><td>String</td><td>[!-~]+</td><td>ASCII of Phred-scaled base QUALity+33</td></tr>
</table>

https://samtools.github.io/hts-specs/SAMv1.pdf

Alignment files: SAM/BAM/CRAM

The **Sequence Alignment/Map (SAM) format** is a tab-delimited text format to store genomic alignments. Contains two sections.

Alignment section:

11 mandatory fields; always in the same order.

Col	Field
1	QNAME
2	FLAG
3	RNAME
4	POS
5	MAPQ
6	CIGAR
7	RNEXT
8	PNEXT
9	TLEN
10	SEQ
11	QUAL

Op	BAM	Description
M	0	alignment match (can be a sequence match or mismatch)
I	1	insertion to the reference
D	2	deletion from the reference
N	3	skipped region from the reference
S	4	soft clipping (clipped sequences present in SEQ)
H	5	hard clipping (clipped sequences NOT present in SEQ)
P	6	padding (silent deletion from padded reference)
=	7	sequence match
X	8	sequence mismatch

https://samtools.github.io/hts-specs/SAMv1.pdf

Alignment files: SAM/BAM/CRAM

The **Sequence Alignment/Map (SAM) format** is a tab-delimited text format to store genomic alignments. Contains two sections.

Alignment section:

11 mandatory fields; always in the same order.

Col	Field	Type	Regexp/Range	Brief description
1	QNAME	String	[!-?A-~]{1,254}	Query template NAME
2	FLAG	Int	$[0,2^{16}-1]$	bitwise FLAG
3	RNAME	String	*|[!-()+-<>-~][!-~]*	Reference sequence NAME
4	POS	Int	$[0,2^{31}-1]$	1-based leftmost mapping POSition
5	MAPQ	Int	$[0,2^{8}-1]$	MAPping Quality
6	CIGAR	String	*|([0-9]+[MIDNSHPX=])+	CIGAR string
7	RNEXT	String	*|=|[!-()+-<>-~][!-~]*	Ref. name of the mate/next read
8	PNEXT	Int	$[0,2^{31}-1]$	Position of the mate/next read
9	TLEN	Int	$[-2^{31}+1,2^{31}-1]$	observed Template LENgth
10	SEQ	String	*|[A-Za-z=.]+	segment SEQuence
11	QUAL	String	[!-~]+	ASCII of Phred-scaled base QUALity+33

https://samtools.github.io/hts-specs/SAMv1.pdf

Alignment files: SAM/BAM/CRAM

The **Sequence Alignment/Map (SAM) format** is a tab-delimited text format to store genomic alignments. Contains two sections.

Alignment section:

- 11 mandatory fields; always in the same order.
  - 0 or * if information is unavailable.
- Optional fields encoded as TAG:TYPE:VALUE.
  - Edit distance, number of total alignments, alignment score, string of mismatching positions, read group, information of mate's alignment...

https://samtools.github.io/hts-specs/SAMv1.pdf

Alignment files: SAM/BAM/CRAM

The **Sequence Alignment/Map (SAM) format** is a tab-delimited text format to store genomic alignments. Contains two sections.

Alignment section:

11 mandatory fields; always in the same order.

0 or * if information is unavailable.

Optional fields encoded as TAG:TYPE:VALUE.

Edit distance, number of total alignments, alignment score, string of mismatching positions, read group, information of mate's alignment...

```
K00252:349:HWT3WBBXX:6:2123:2301:12269     99    10    3101416    57    44M106S    =     3101416
      43    TCCTTCTCCAGTGCGCTTCATCTTTTTGTGTGTAGTCT...
AAFFFJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJFJJFJJ...   XA:Z:chr10,-7460382,106S44M,1;   MC:Z:107S43M
MD:Z:44     PG:Z:MarkDuplicates    RG:Z:10     NM:i:0      MQ:i:57     AS:i:44     XS:i:39
```

https://samtools.github.io/hts-specs/SAMv1.pdf

Alignment files: SAM/BAM/CRAM

The **Sequence Alignment/Map (SAM) format** is a tab-delimited text format to store genomic alignments. Contains two sections.

Alignment section:

11 mandatory fields; always in the same order.

0 or * if information is unavailable.

Optional fields encoded as TAG:TY

Edit distance, number of total align ng of mismatching positions, read group, information of mate's a

read paired (0x1)
read mapped in proper pair (0x2)
mate reverse strand (0x20)
first in pair (0x40)

```
K00252:349:HWT3WBBXX:6:2123:2301:12269      99    10    3101416    57    44M106S    =     3101416
      43    TCCTTCTCCAGTGCGCTTCATCTTTTTGTGTGTAGTCT...
AAFFFJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJFJJFJJ...   XA:Z:chr10,-7460382,106S44M,1;   MC:Z:107S43M
MD:Z:44     PG:Z:MarkDuplicates    RG:Z:10     NM:i:0      MQ:i:57     AS:i:44     XS:i:39
```

https://samtools.github.io/hts-specs/SAMv1.pdf

Alignment files: SAM/BAM/CRAM

The **Sequence Alignment/Map (SAM) format** is a tab-delimited text format to store genomic alignments. Contains two sections.

Alignment section:

11 mandatory fields; always in the same order.

0 or * if information is unavailable.

Optional fields encoded as TAG:TYPE:VALUE.

Edit distance, number of total alignments, alignment score, string of mismatching positions, read group, information of mate's alignment...

```
K00252:349:HWT3WBBXX:6:2123:2301:12269      99    10    3101416    57    44M106S    =     3101416
      43    TCCTTCTCCAGTGCGCTTCATCTTTTTGTGTGTAGTCT...
AAFFFJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJFJJFJJ...   XA:Z:chr10,-7460382,106S44M,1;   MC:Z:107S43M
MD:Z:44     PG:Z:MarkDuplicates    RG:Z:10     NM:i:0      MQ:i:57     AS:i:44     XS:i:39
```

https://samtools.github.io/hts-specs/SAMv1.pdf

Alignment files: SAM/BAM/CRAM

BAM file: binary (compressed) version of SAM file.

Can be **indexed** -> allows fast retrieval of specific regions of the genome.

- Requires the BAM file to be sorted by position.
- The index file is named by appending .bai to the bam file name.

CRAM file: further compressed version of a BAM file.

- Uses a reference-based compression.
- Only bases differing from the reference need to be stored.

http://samtools.github.io/hts-specs/SAMv1.pdf

Alignment files: SAM/BAM/CRAM

SAM/BAM/CRAM files can be manipulated with **SAMtools**.

Sorting, merging, indexing and generating alignments in a per-position format.

Rsamtools provides an interface to the 'samtools', 'bcftools', and 'tabix' utilities for manipulating SAM, FASTA, BCF and tabix files.

https://bioconductor.org/packages/release/bioc/html/Rsamtools.html

Picard tools is also useful.

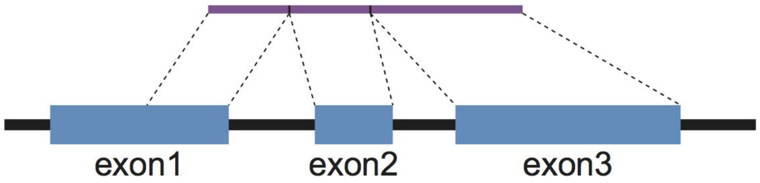
Marking duplicate reads, collecting metrics, fix mate information (paired-end reads)

http://samtools.sourceforge.net/
http://htslib.org/
https://broadinstitute.github.io/picard/

Alignment of RNA-seq data

RNA-seq sequencing reads come from spliced mRNAs.

Their alignment in the genome is interrupted by introns.

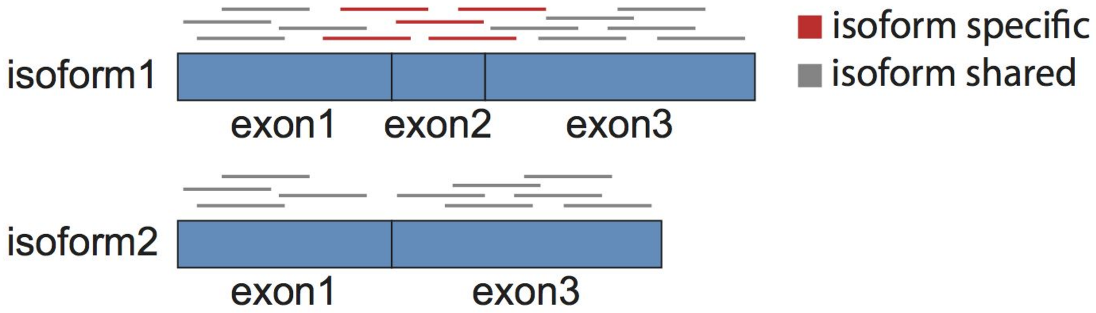


Two solutions:

- Map reads to the transcriptome instead of the genome.
- Allow gapped alignments.

Map reads to the transcriptome

Reads in exons that are shared across transcript isoforms will map multiple times.



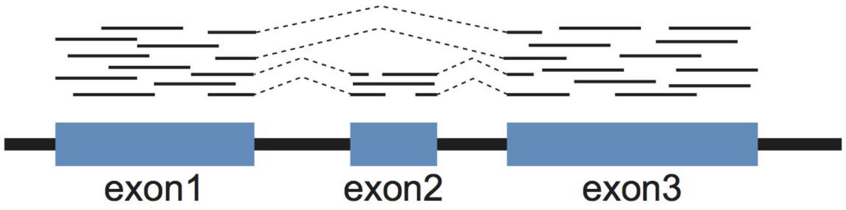
Requires good annotation.

- Any novel genes or isoforms will be lost.

Splice-aware aligners

Map to the genome but allow large gaps.

Intron size ranges from 10^2 to ~10^5.



Allows gene and isoform discovery.

Greatly enhanced by paired-end reads.

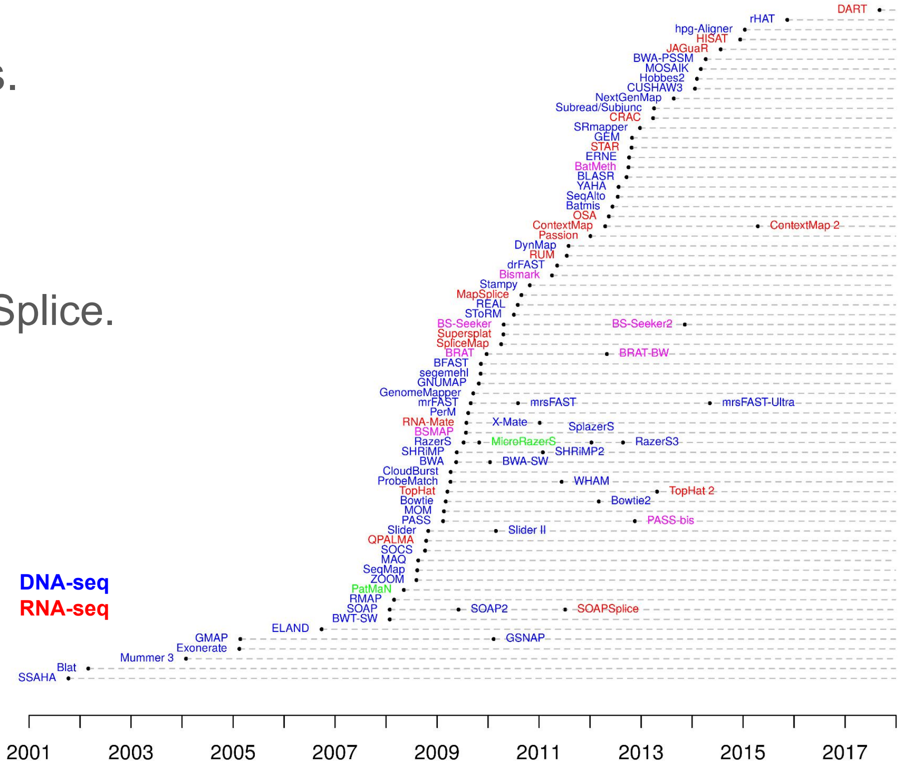
Splice-aware aligners

Map to the genome but allow large gaps.

Intron size ranges from 10^2 to ~10^5.

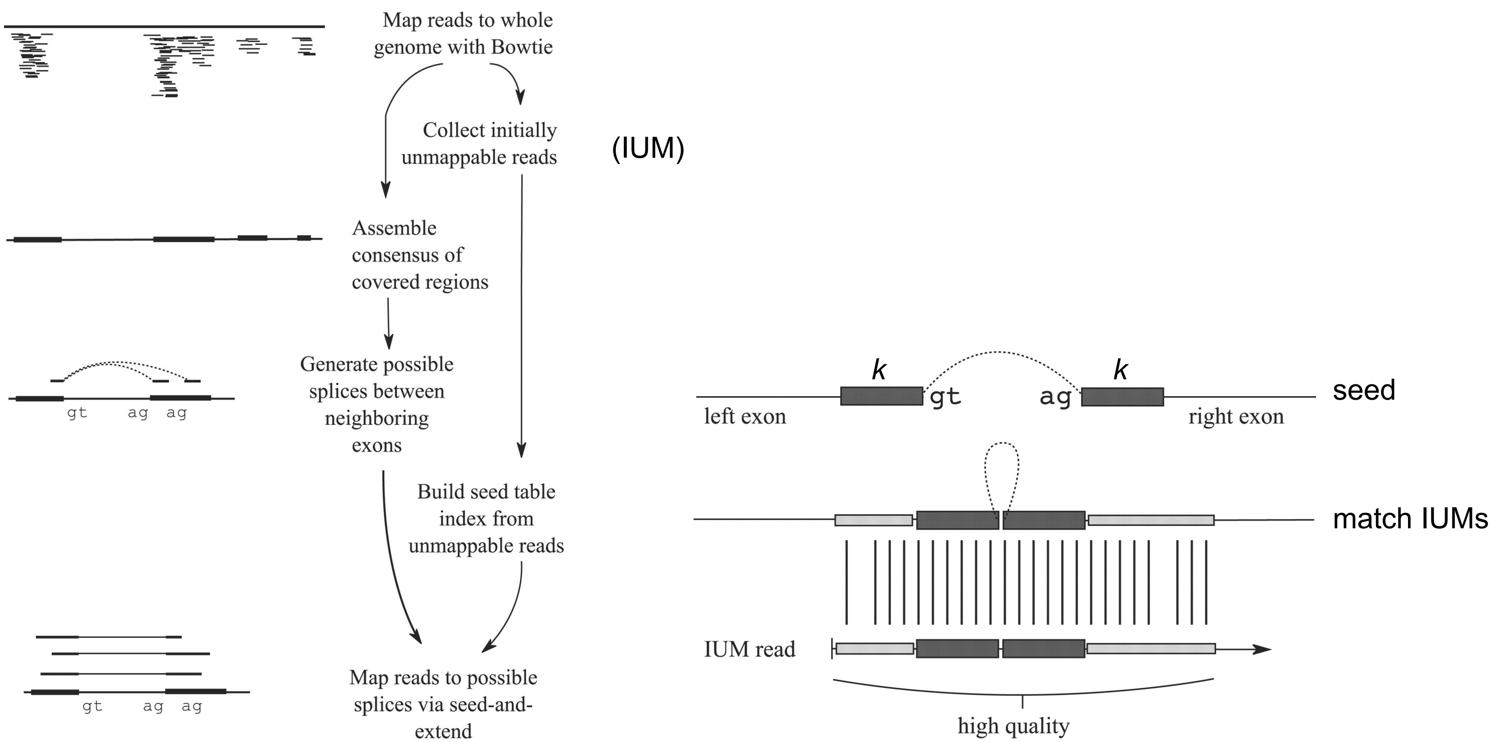
Many different mappers.

Tophat, **STAR**, GSNAP, subread, MapSplice.



https://www.ebi.ac.uk/~nf/hts_mappers/

TopHat

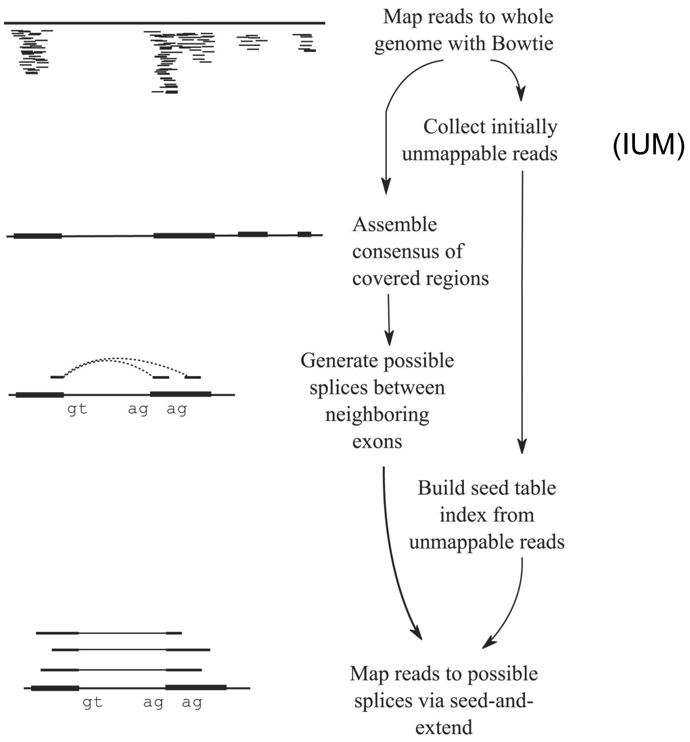


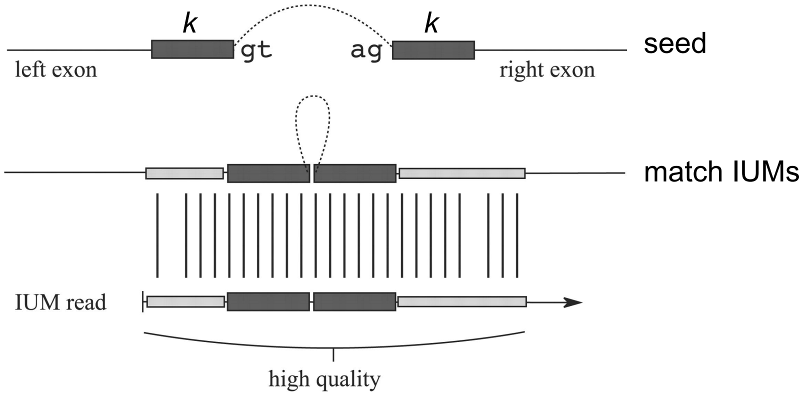
Trapnell et al., *TopHat: discovering splice junctions with RNA-Seq*, Bioinformatics (2009).
doi: https://doi.org/10.1093/bioinformatics/btp120

https://ccb.jhu.edu/software/tophat/index.shtml

TopHat

Kim et al., *TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions*, Genome Biology (2013).
doi: https://doi.org/10.1186/gb-2013-14-4-r36



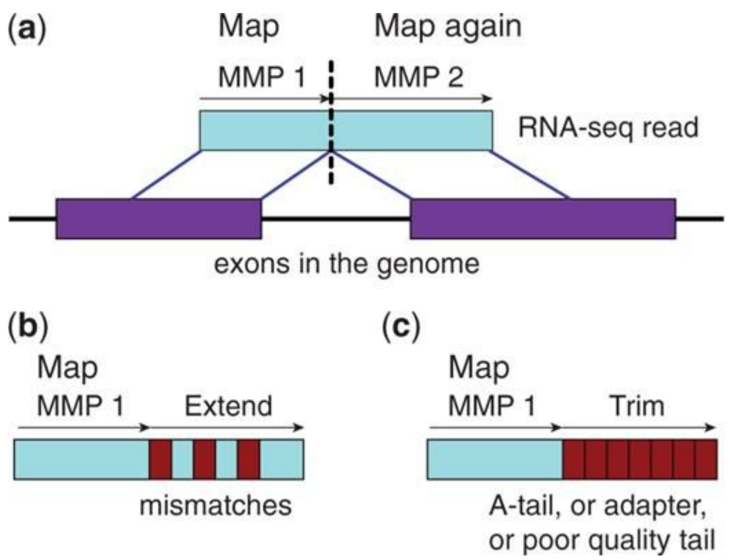


Trapnell et al., *TopHat: discovering splice junctions with RNA-Seq*, Bioinformatics (2009).
doi: https://doi.org/10.1093/bioinformatics/btp120

https://ccb.jhu.edu/software/tophat/index.shtml

STAR

1. Find **seeds** that align perfectly.
2. **Cluster** seeds mapping within a confined region.
3. **Stitch** them together.
   Using local alignment allowing mismatches and gaps.
4. **Score** all possible alignments and chose best.



MMP = maximal mappable prefix

https://github.com/alexdobin/STAR

Dobin et al., *STAR: ultrafast universal RNA-seq aligner*, Bioinformatics (2012).
doi: https://doi.org/10.1093/bioinformatics/bts635

Splice-aware aligners

Map to the genome but allow large gaps.

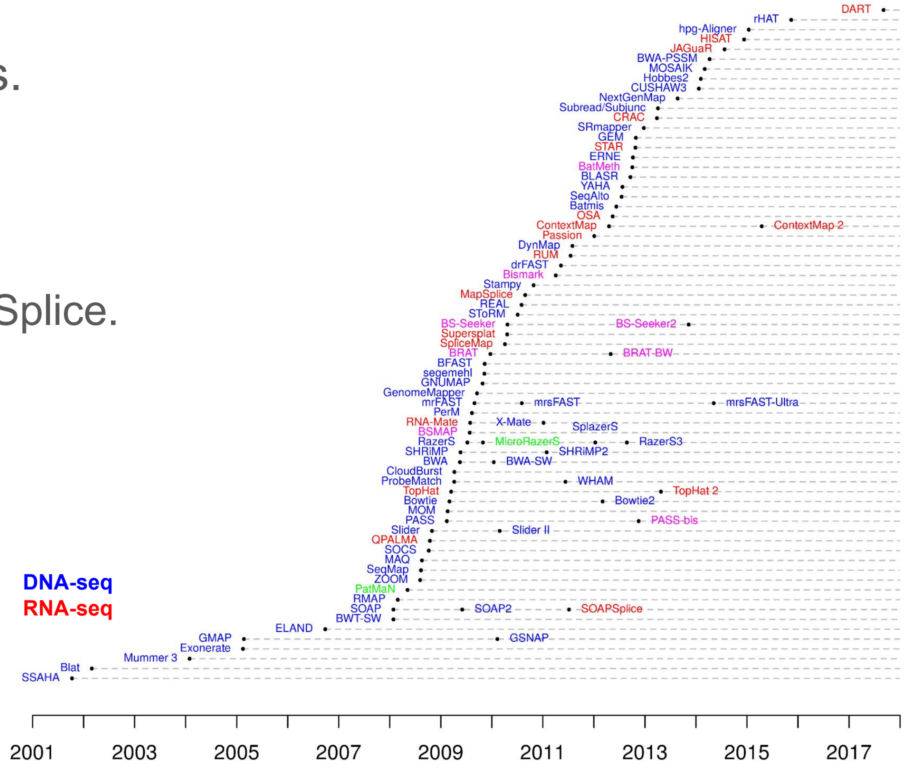
Intron size ranges from 10^2 to $\sim10^5$.

Many different mappers.

Tophat, **STAR**, GSNAP, subread, MapSplice.

Again, there is no best aligner.

- Speed.
- Memory usage.
- Accuracy of found exon junctions.



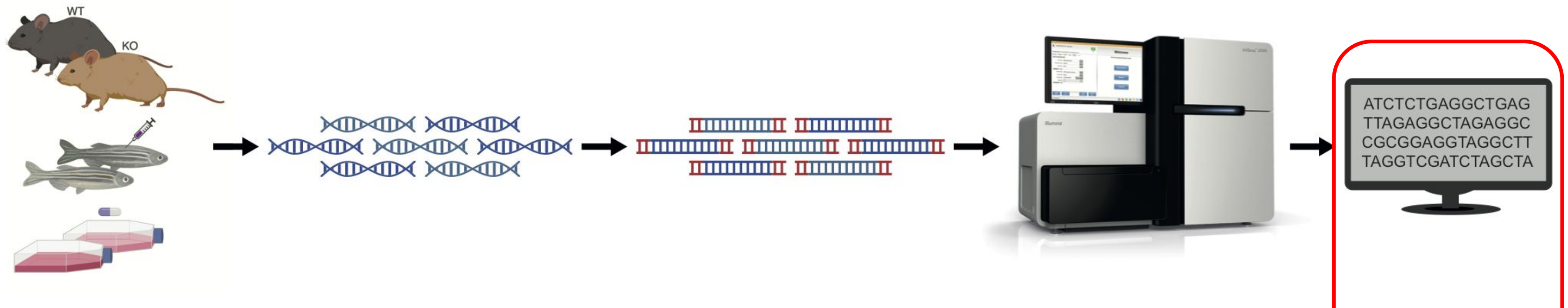
https://www.ebi.ac.uk/~nf/hts_mappers/

Analysis of RNA-seq data

Anna Cuomo
EBI & University of Cambridge

Ximena Ibarra-Soria
Cancer Research UK

High-throughput sequencing experiments



sample collection → extraction of DNA, RNA, chromatin... → library prep → sequencing → data analysis

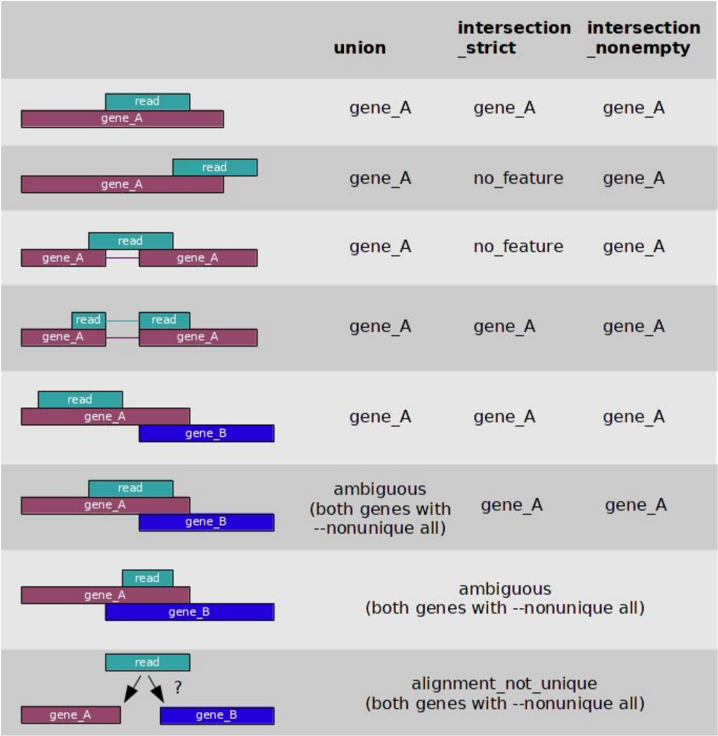
Quantify gene expression

Take a BAM file with aligned reads and a set of features of interest and count the number of reads overlapping each feature.

HTSeq, featureCounts, STAR.

Other programs have more complex algorithms to try and

- quantify transcript abundance.
- correct multimapping reads.
- correct known biases.



https://htseq.readthedocs.io/en/release_0.9.1/count.html

Pseudo - aligners

Kallisto, Salmon (Sailfish in a previous version)

- Alignment + quantification
- Maps k-mers (does not allow for mismatch)
    - Extremely fast and memory efficient
    - But only transcript quantification, not suitable for defining gene structure

```
Read:  ATCCCGGGTTAT
       ATCCCGG
        TCCCGGG
         CCCGGGT
          CCGGGTT
           CGGGTTA
            GGGTTAT
```

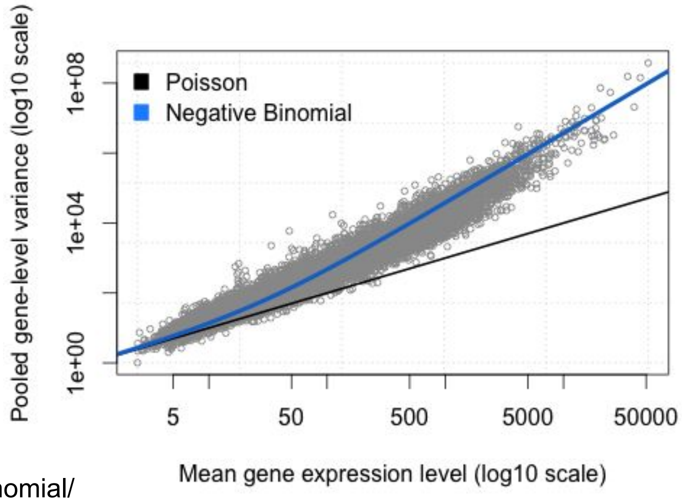
Kallisto: Bray et al, *Nat Biotechnology* 2016 (doi: https://doi.org/10.1038/nbt.3519)
Salmon: Patro et al, *Nat Methods* 2017 (doi: 10.1038/nmeth.4197)

Negative Binomial (NB) distribution

- RNA-seq data is count data: number of reads mapped to a gene. Discrete, not continuous.
- Poisson distribution is designed for modelling count data.
  - Sampling from large pool (~million reads per sample), small chance (10-100k counts per gene)
- Poisson assumes $\sigma^2 = \mu$
  - But data clearly shows higher variance
- NB is an extension of Poisson, with an extra

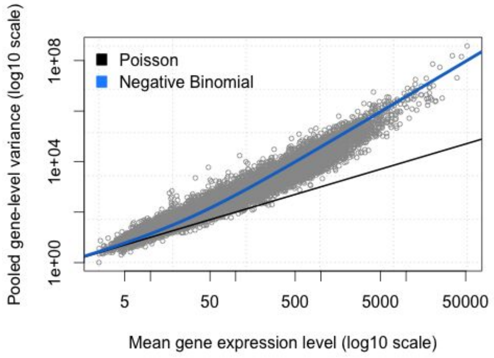
parameter, called overdispersion (alpha)

  - $\sigma^2 = \mu + \alpha\mu^2$



https://bioramble.wordpress.com/2016/01/30/why-sequencing-data-is-modeled-as-negative-binomial/

Mean - variance relationship



Because variance is a function of mean (and the other way around)

For downstream analyses we want to apply some form of variance stabilization

E.g.

- defining highly variable genes,
- performing differential expression analysis

DESeq2 provides two different functions for this, vst and rlog

Spike-in transcripts

- ERCC spike-ins are commonly used to estimate the RNA content of the cell.
    - 92 single-exon transcripts.
    - 250 – 2,000 nucleotides in length.
    - Variable GC content.
    - 10^6-fold concentration range.
- The same amount is added to every cell.
    - [spike-in] / [endogenous RNA] is an indication of the initial RNA content.

https://www.thermofisher.com/order/catalog/product/4456740

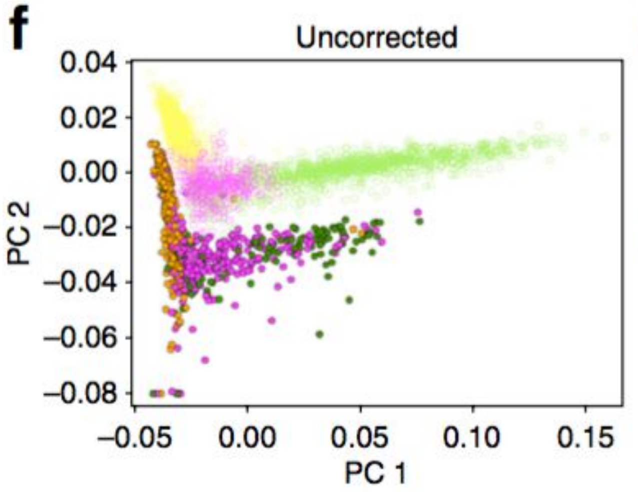
Analysis of sc-RNA-seq data

Anna Cuomo
EBI & University of Cambridge

Ximena Ibarra-Soria
Cancer Research UK

Batch correction: MNN

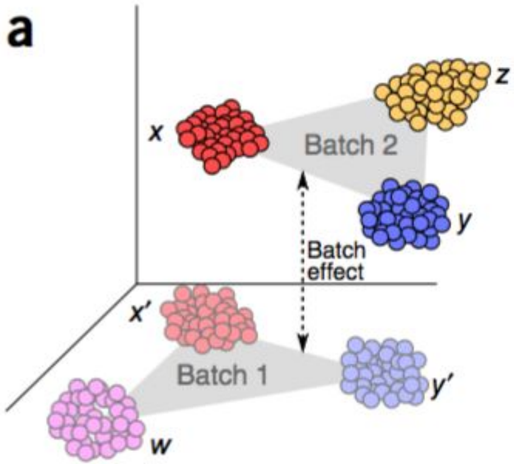
Find mutual nearest neighbours (MNNs) in the different batches that represent *equivalent* cell types. Model and remove the technical effects.



Haghverdi et al., *Batch effects in single-cell RNA-sequencing data are corrected by matching...*, Nat Biotechnol (2018)
doi: https://doi.org/10.1038/nbt.4091

Batch correction

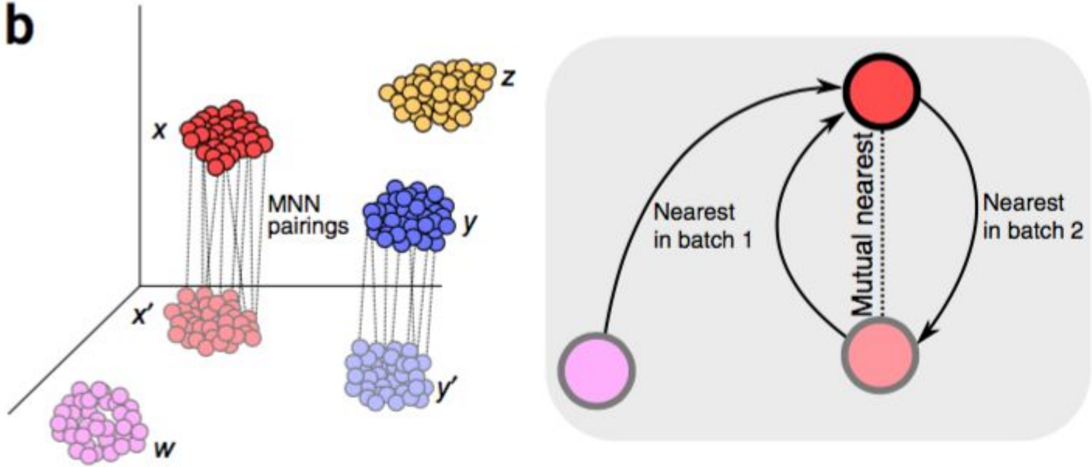
Find mutual nearest neighbours (MNNs) in the different batches that represent *equivalent* cell types. Model and remove the technical effects.



Haghverdi et al., *Batch effects in single-cell RNA-sequencing data are corrected by matching...*, Nat Biotechnol (2018)
doi: https://doi.org/10.1038/nbt.4091

Batch correction: MNN

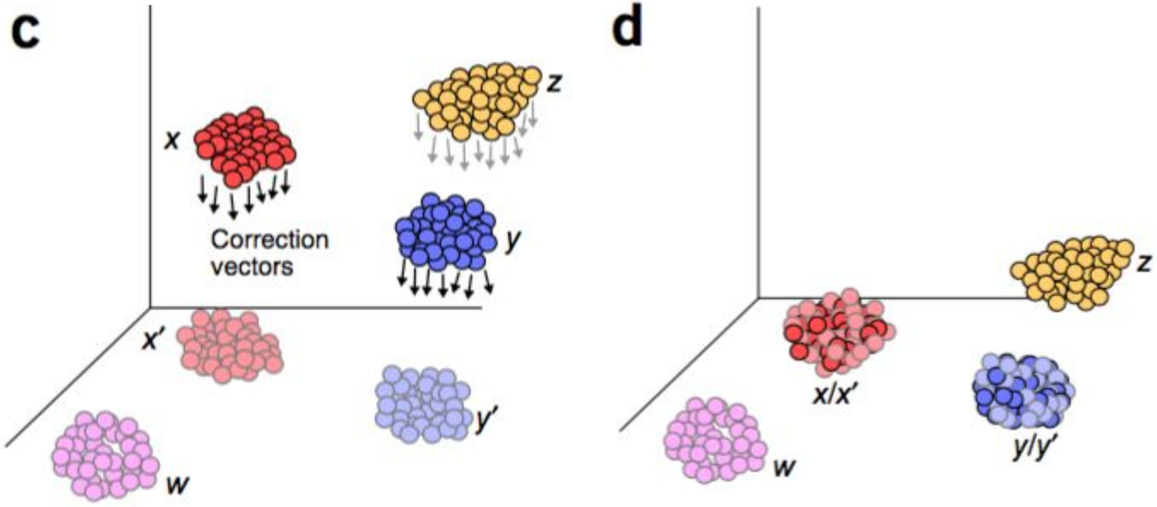
Find **mutual nearest neighbours** (MNNs) in the different batches that represent *equivalent* cell types. Model and remove the technical effects.



Haghverdi et al., *Batch effects in single-cell RNA-sequencing data are corrected by matching...*, Nat Biotechnol (2018)
doi: https://doi.org/10.1038/nbt.4091

Batch correction: MNN

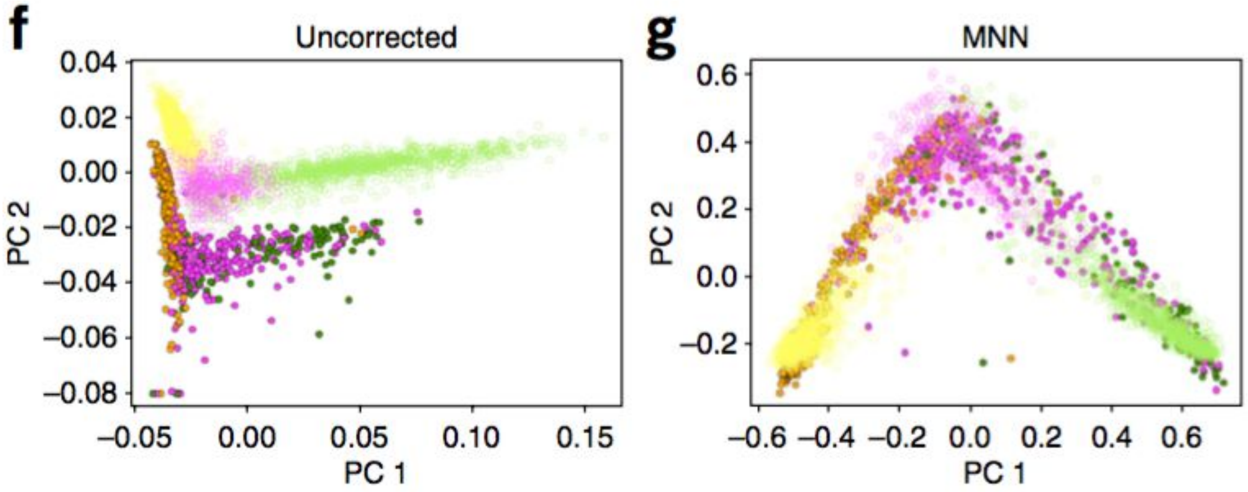
Find mutual nearest neighbours (MNNs) in the different batches that represent *equivalent* cell types. Model and remove the technical effects.



Haghverdi et al., *Batch effects in single-cell RNA-sequencing data are corrected by matching...*, Nat Biotechnol (2018)
doi: https://doi.org/10.1038/nbt.4091

Batch correction: MNN

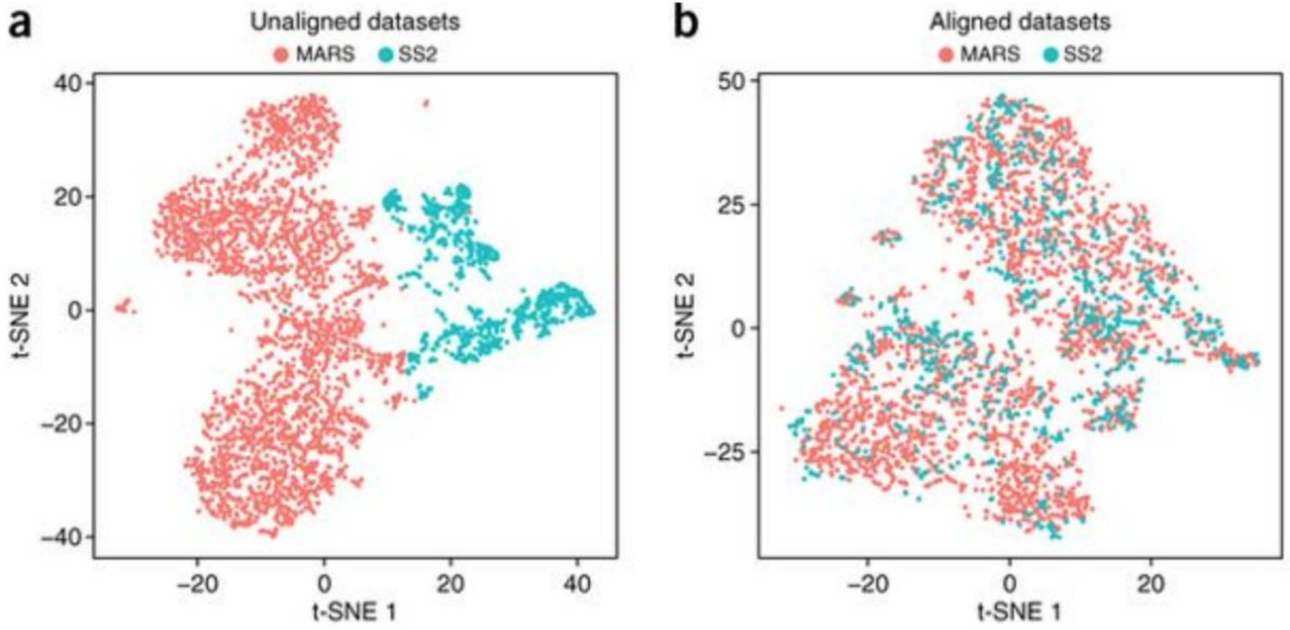
Find mutual nearest neighbours (MNNs) in the different batches that represent *equivalent* cell types. Model and remove the technical effects.



Haghverdi et al., *Batch effects in single-cell RNA-sequencing data are corrected by matching...*, Nat Biotechnol (2018)
doi: https://doi.org/10.1038/nbt.4091

Batch correction: CCA

Canonical correlation analysis

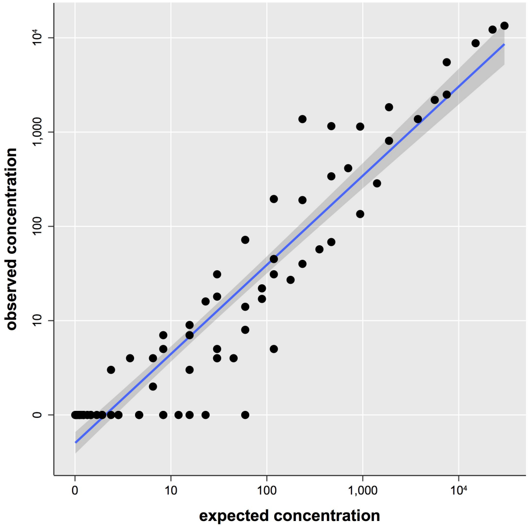


Butler et al., *Integrating single-cell transcriptomic data across different conditions, technologies, and species*, Nat Biotechnol (2018)
doi: https://doi.org/10.1038/nbt.4096

Technical noise estimation

One way to estimate technical noise is to **spike-in** a known concentration of RNA.

- ERCC spike-ins are the most commonly used.
  - 92 single-exon transcripts.
  - 250 – 2,000 nucleotides in length.
  - Variable GC content.
  - 10^6-fold concentration range.
- The same amount is added to every cell.
- Affected only by technical noise.

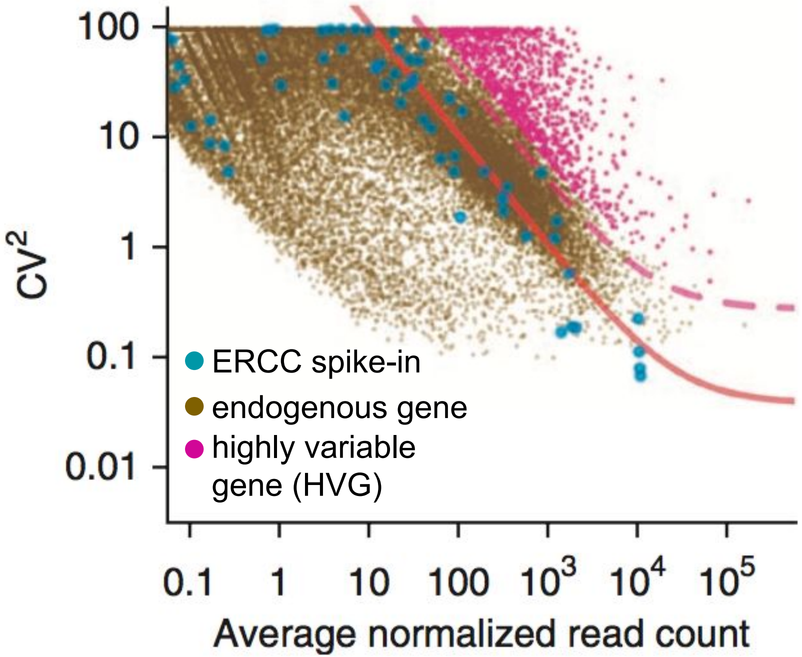


Spike-ins also allow estimating the RNA content of the cell.

https://www.thermofisher.com/order/catalog/product/4456740

Highly variable gene detection

To identify the genes that are variable across cells, it is necessary to account for the technical noise.



Technical variance can be estimated from spike-ins.

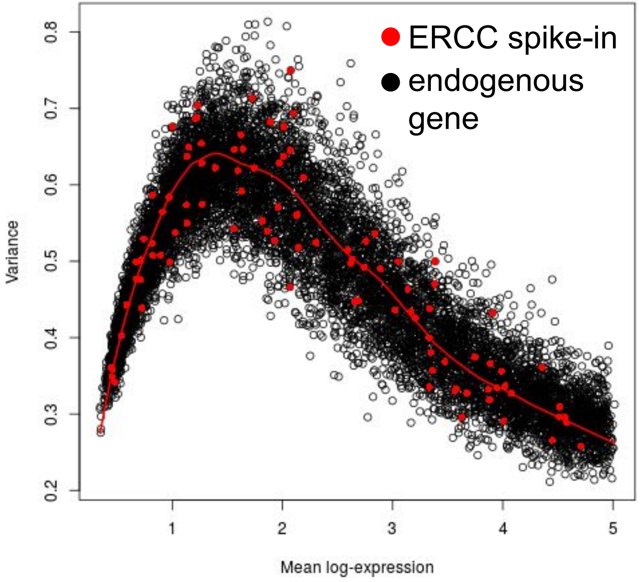
HVGs are those that have significantly higher variance than expected by noise only.

Brennecke et al., *Accounting for technical noise in single-cell RNA-seq experiments*, Nature Methods (2013)
doi: https://doi.org/10.1038/nmeth.2645

Highly variable gene detection

To identify the genes that are variable across cells, it is necessary to account for the technical noise.

A different approach is to fit the mean-variance trend and subtract that from total variance, thus retaining only the biological component.



https://bioconductor.org/packages/release/bioc/html/scran.html

Miscellaneous

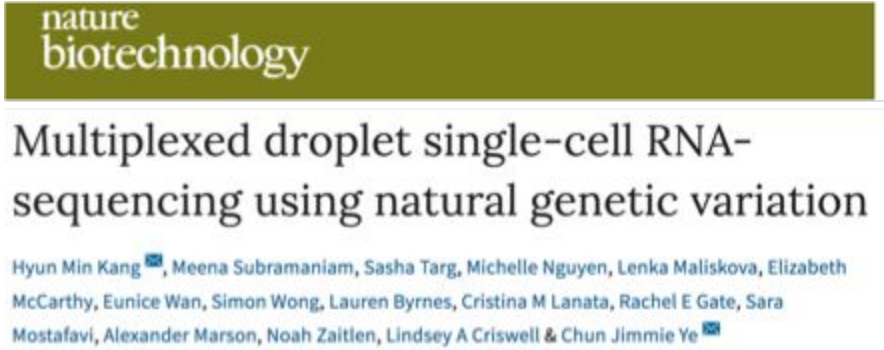
Anna Cuomo
EBI & University of Cambridge

Ximena Ibarra-Soria
Cancer Research UK

Doublets

Can be inferred when there are two types of cells.

- male and female.
- mouse and human.
- diverse genetic background.


nature biotechnology

Multiplexed droplet single-cell RNA-sequencing using natural genetic variation

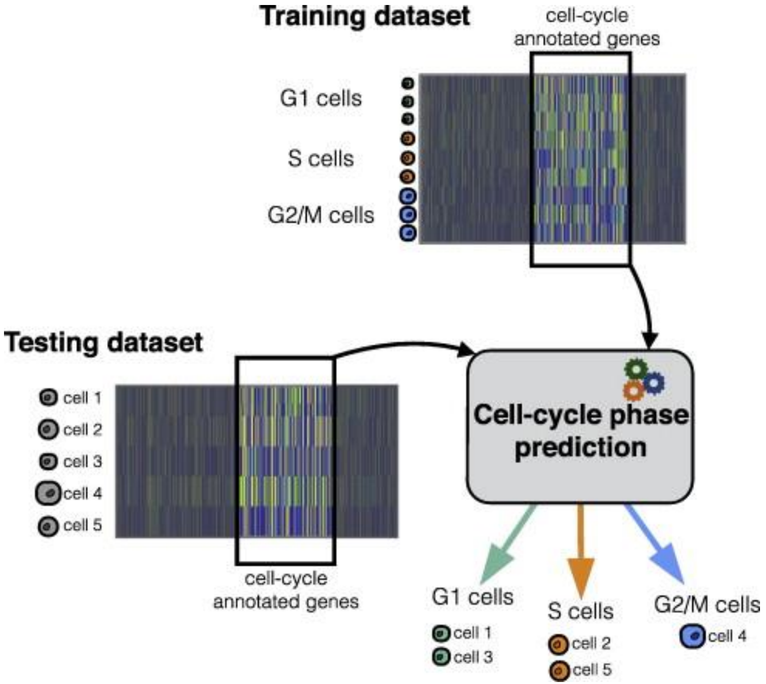
Hyun Min Kang, Meena Subramaniam, Sasha Targ, Michelle Nguyen, Lenka Maliskova, Elizabeth McCarthy, Eunice Wan, Simon Wong, Lauren Byrnes, Cristina M Lanata, Rachel E Gate, Sara Mostafavi, Alexander Marson, Noah Zaitlen, Lindsey A Criswell & Chun Jimmie Ye

scran::doubletCells

Cell Cycle

Cell cycle phase can be a confounder

- f-scLVM (doi.org/10.1186/s13059-017-1334-8)
- CCA (doi.org/10.1038/nbt.4096)
- cyclone (implemented in scran; doi.org/10.1016/j.ymeth.2015.06.021)



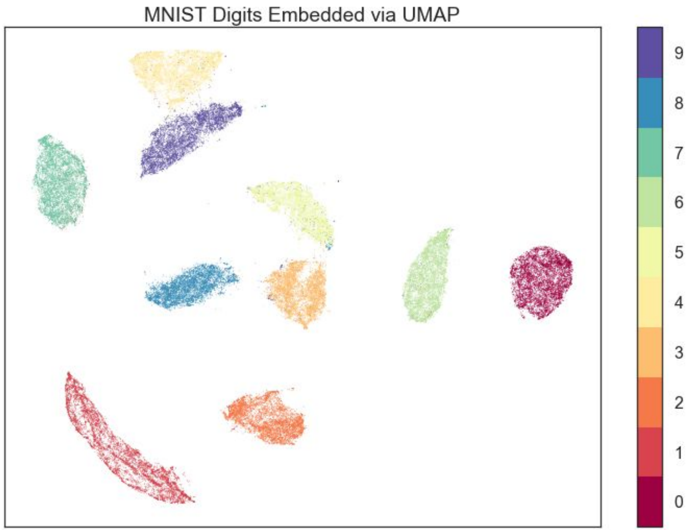
Scialdone et al., *Computational assignment of cell-cycle stage from single-cell transcriptome data*, Methods (2015)
doi: http://doi.org/10.1016/j.ymeth.2015.06.021

scran::cyclone

UMAP

UMAP (Uniform Manifold Approximation and Projection for Dimension Reduction) is another increasingly popular dimensionality reduction / visualization tool, often compared to t-SNE

Evaluation of UMAP as an alternative to t-SNE for single-cell data
https://www.biorxiv.org/content/early/2018/04/10/298430



McInnes, L, Healy, J, *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction*, ArXiv e-prints 1802.03426, 2018

Cancer single cell rna seq approaches

Clonealign: https://www.biorxiv.org/content/early/2018/06/11/344309

HoneyBADGER: https://jef.works/HoneyBADGER/

Cardelino: https://github.com/PMBio/cardelino

CONICS: https://academic.oup.com/bioinformatics/advance-article/doi/10.1093/bioinformatics/bty316/4979546

Additional resources

scrNA-seq data workflows:

- http://bioconductor.org/packages/release/workflows/html/simpleSingleCell.html (Lun et al.)
- https://hemberg-lab.github.io/scRNA.seq.course/index.html (Hemberg lab)
- http://hms-dbmi.github.io/scw/ (Harvard single cell workshop)

About t-SNE: https://distill.pub/2016/misread-tsne/

Additional packages (for scRNA-seq data analysis)

R/Bioconductor (other than SingleCellExperiment/scater/scran)

- Seurat, MAST
- Monocle, SLICER (Pseudotime / diffusion maps analysis)
- SC3 (clustering)
- edgeR, DESeq2 (differential expression)
- iSEE (visualisation)
- Honeybadger (CNVs)
- BASiCS (differential expression, differential variability)
- Slalom (see f-scLVM, R implementation)
- scDD, Splatter (simulation of scRNA-seq data)

Python

- Scanpy
- f-scLVM (factor single cell latent variable model)
- MOFA (multi omics factor analysis) (has R implementation too)

Comprehensive list of scRNA-seq data analysis tools: https://github.com/seandavi/awesome-single-cell

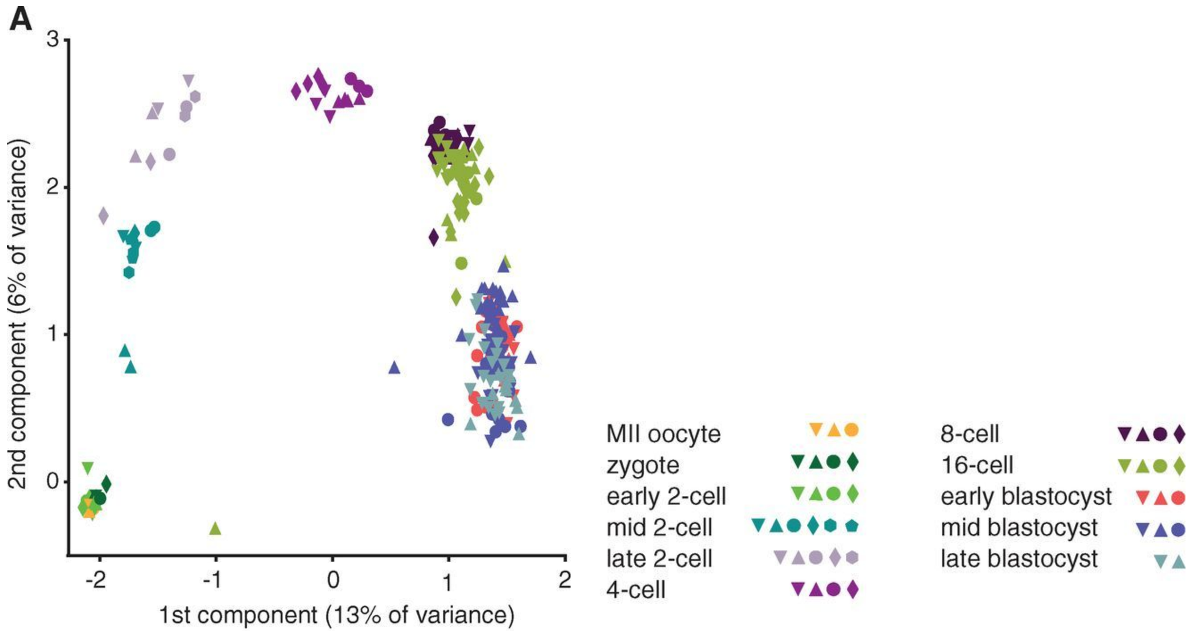
Link to repository of papers with available data

http://imlspenticton.uzh.ch:3838/conquer/

Possible datasets for projects (Deng et al.)

Early mouse embryo development (zygote -> late blastocyst)

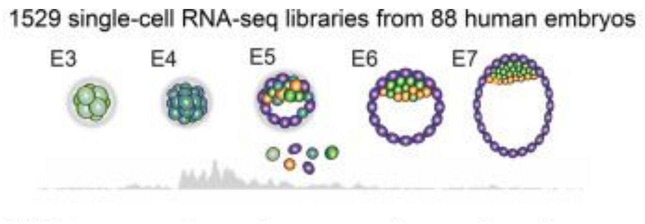
http://www.sciencemag.org/cgi/pmidlookup?view=long&pmid=24408435

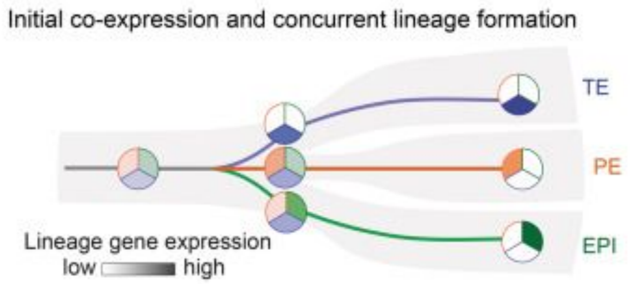


Possible datasets for projects (Petropoulos et al.)

Early human embryo development

https://www.sciencedirect.com/science/article/pii/S009286741630280X?via%3Dihub





3 main cell types of mature blastocyst:

trophectoderm (TE)

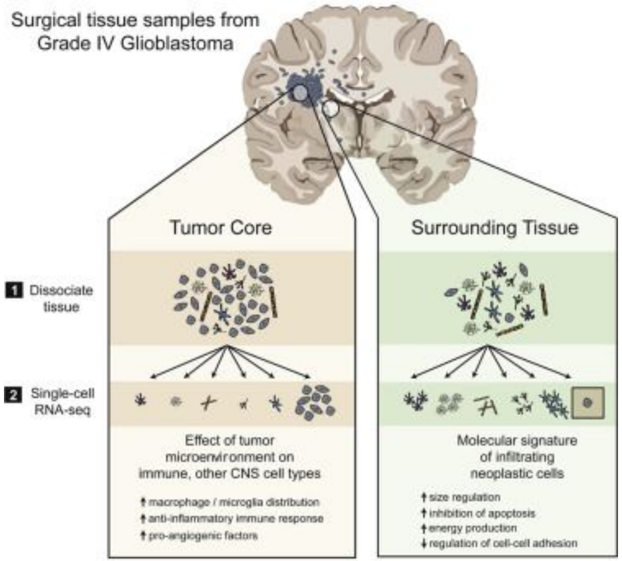
primitive endoderm (PE)

epiblast (EPI)

Possible datasets for projects (Darmanis et al.)

Heterogeneity of glioblastoma tumour cells, and surrounding tissue

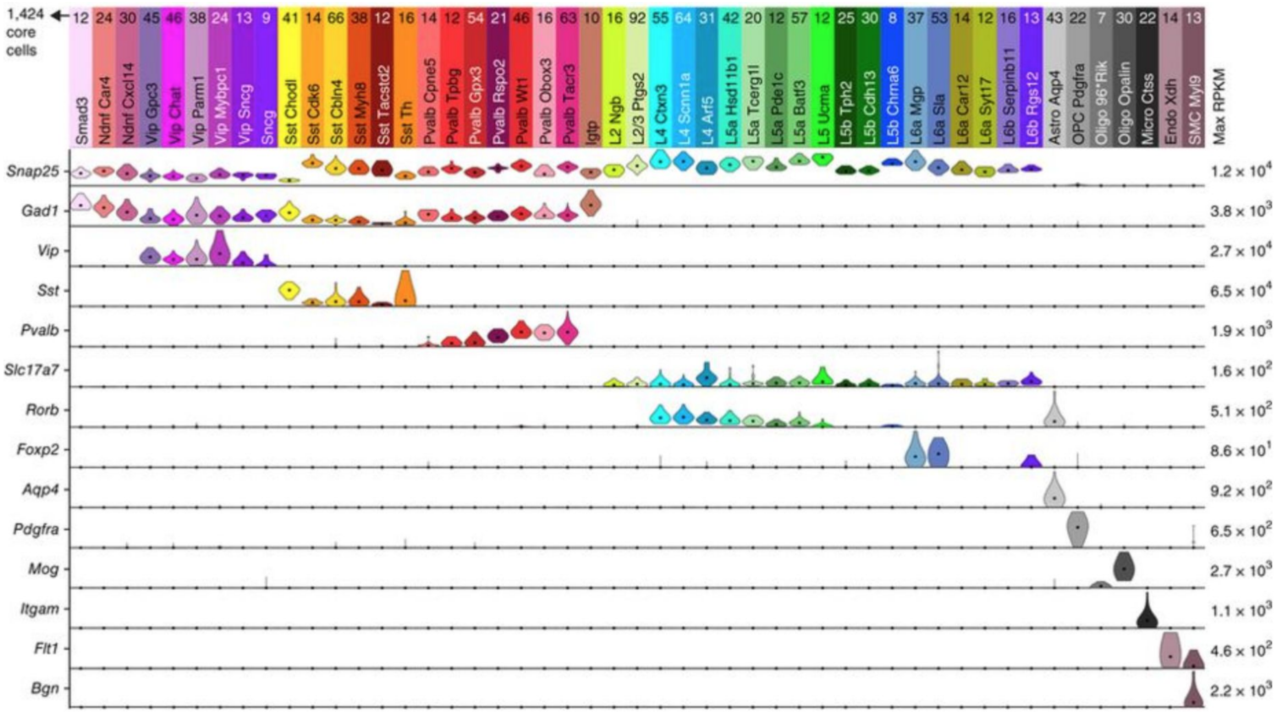
https://www.sciencedirect.com/science/article/pii/S2211124717314626?via%3Dihub



Possible datasets for projects (Tasic et al.)

Cellular diversity in the mouse primary visual cortex.

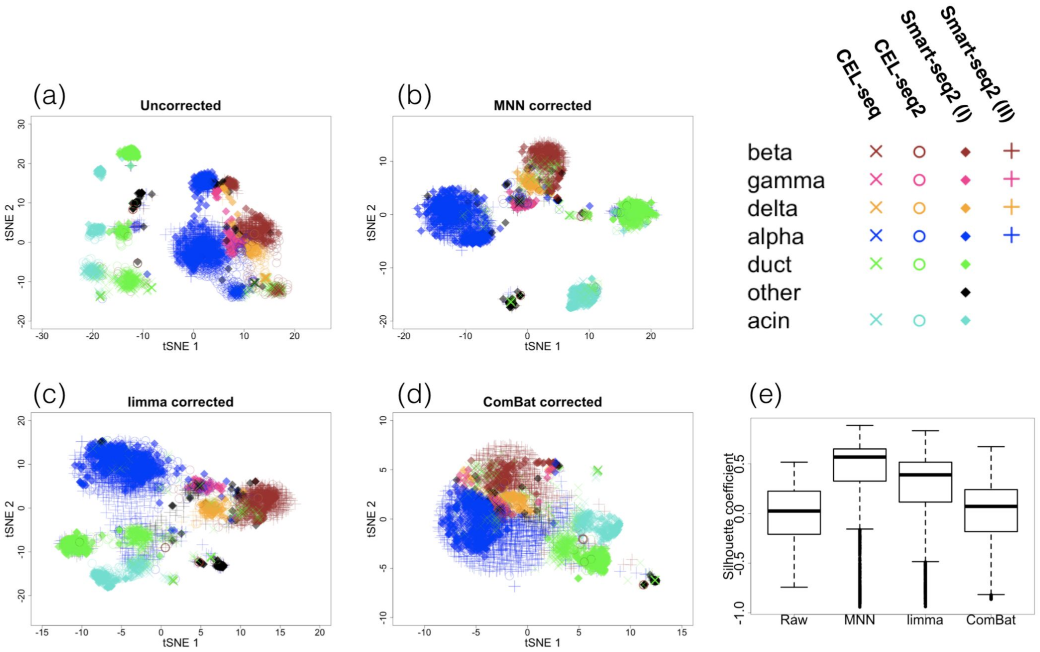
https://www.nature.com/articles/nn.4216



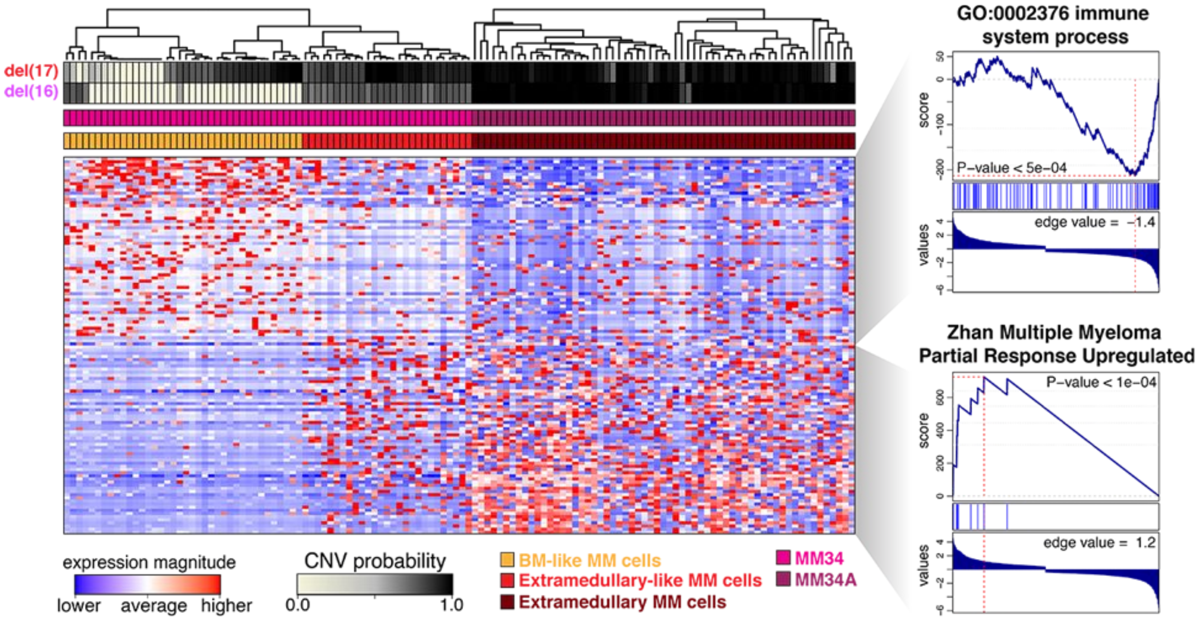
Possible datasets for projects (MNN - correct)

One of the application of the MNN batch correction method described in the paper (https://doi.org/10.1038/nbt.4091) is a comparison of pancreatic cells across different studies:

1. CEL-seq, Grun et al, 2016
2. CEL-seq2, Muraro et al, 2016
3. Smart-seq2, Lawlor et al. 2017
4. Smart-seq2, Segerstolpe et al, 2016



Possible datasets for projects (Fan et al.)

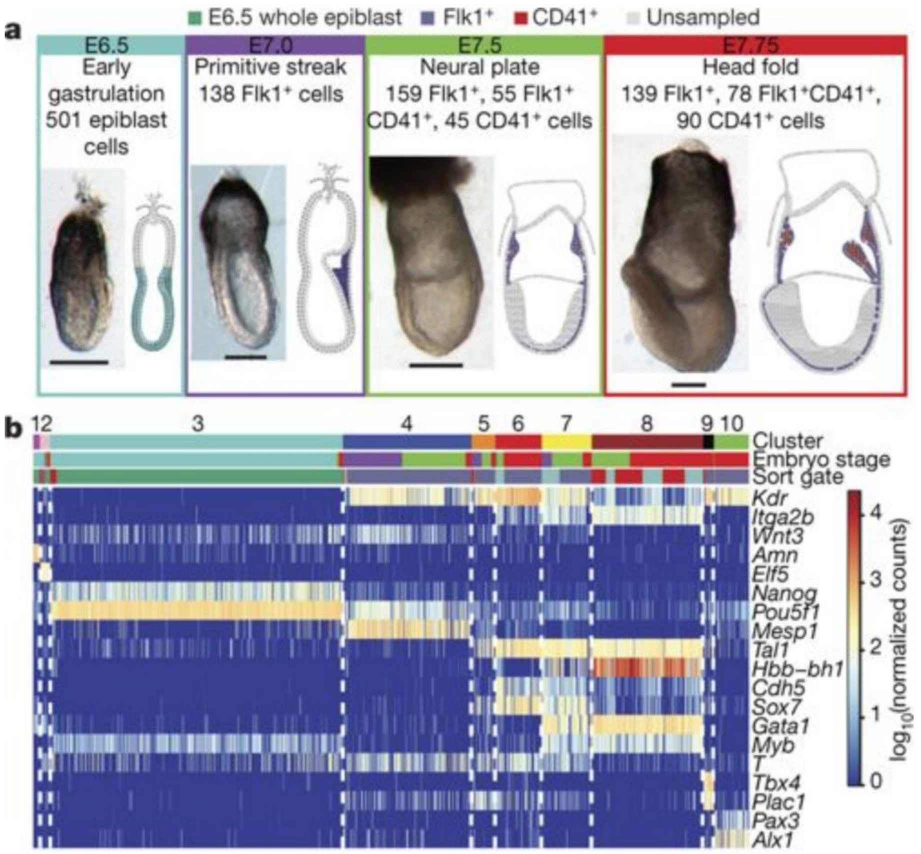


HoneyBADGER identifies and infers the presence of CNV and LOH events in single cells and reconstructs subclonal architecture using allele and expression information from single-cell RNA-sequencing data.

https://genome.cshlp.org/content/early/2018/06/13/gr.228080.117.full.pdf+html

https://github.com/JEFworks/HoneyBADGER

Possible datasets for projects (Scialdone et al.)



Mouse early embryonic development.

https://www.nature.com/articles/nature18633

Possible datasets for projects (Halpern et al.)

Single-cell spatial reconstruction reveals global division of labour in the mammalian liver

https://www.nature.com/articles/nature21065