Introduction to single cell genomics

Raheleh Rahbari

Raheleh.Rahbari@sanger.ac.uk

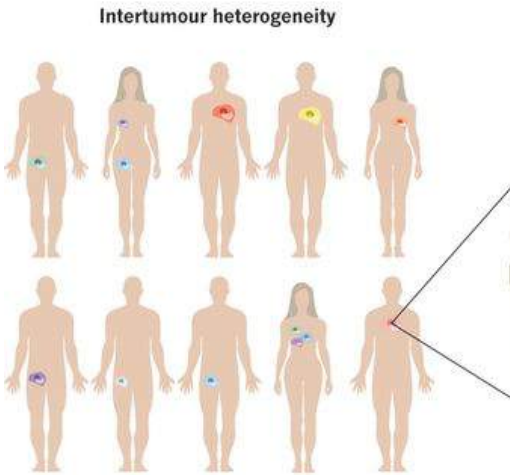
wellcome sanger institute

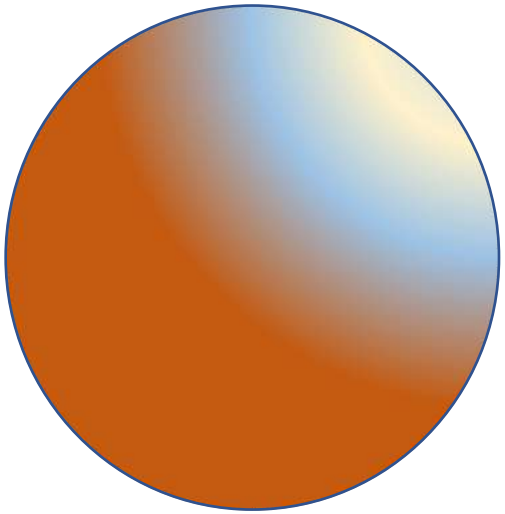
Learning objectives

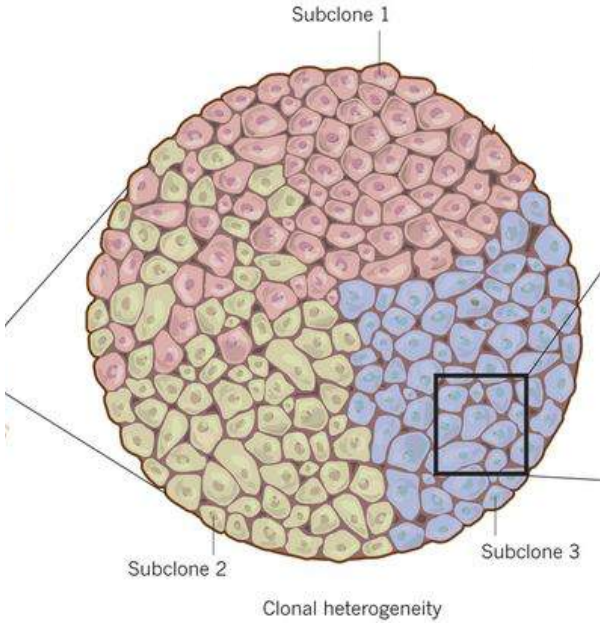
- Why single cell sequencing?
- Methods for single cell sequencing
- Analysis of single cell data
- Single cell multi-omics approach – A case study

wellcome sanger institute

Low resolution of clonal dynamics
Bulk tissue Vs. Single cell







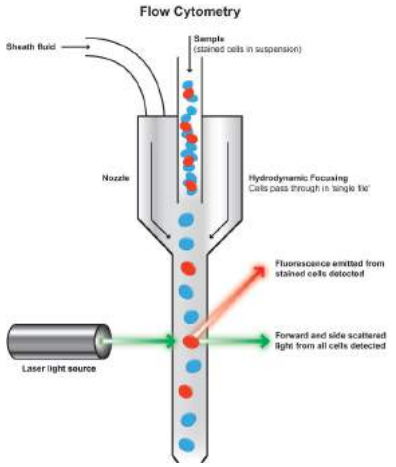
wellcome sanger institute

Methods for single cell isolation

Micromanipulation

- Inefficient
- Prone to contamination
- Needs a good microscopy set up for single molecule facility

Flow Cytometry based methods

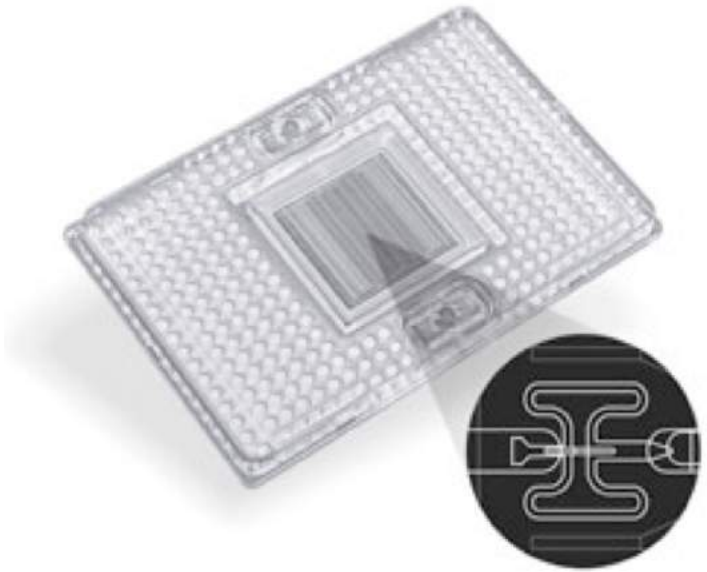


- Inefficient
- Prone to contamination as it is in an open space
- Needs a large volume, decreases the uniformity of amplification

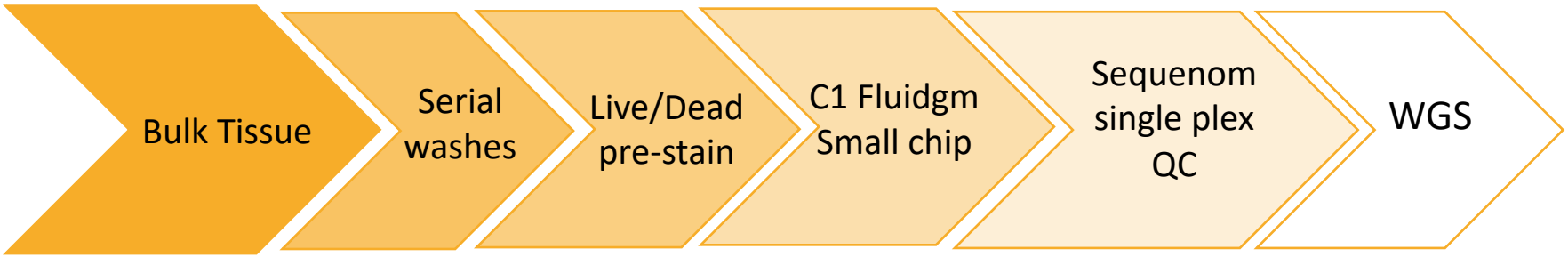
Microfluidics methods

- Not reliable
- Reproducibility issues between different chip batches
- Issues with isolating single cells

C1– Automating Single-Cell Isolation and Genome Amplification



C1– Automating Single-Cell Isolation and Genome Amplification



Methods for single cell isolation

- Micromanipulation
- Flow Cytometry methods
- Microfluidics methods
- Single nuclei methods 

wellcome sanger institute

Nuclear isolation workflow

Frozen tissue → Homogenize → Gradient Centrifuge → Stain & marker → Single Nucleus Per well

Single-cell amplification methods

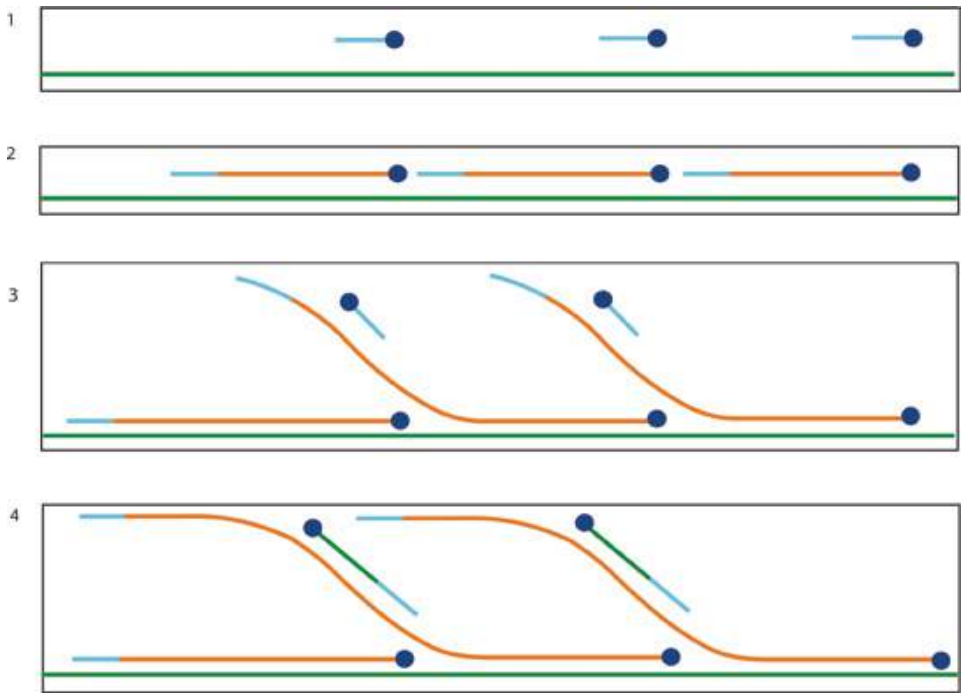
wellcome sanger institute

Whole genome amplification methods

wellcome sanger institute

Isothermal amplification

Whole genome Multiple Displacement Amplification (MDA)

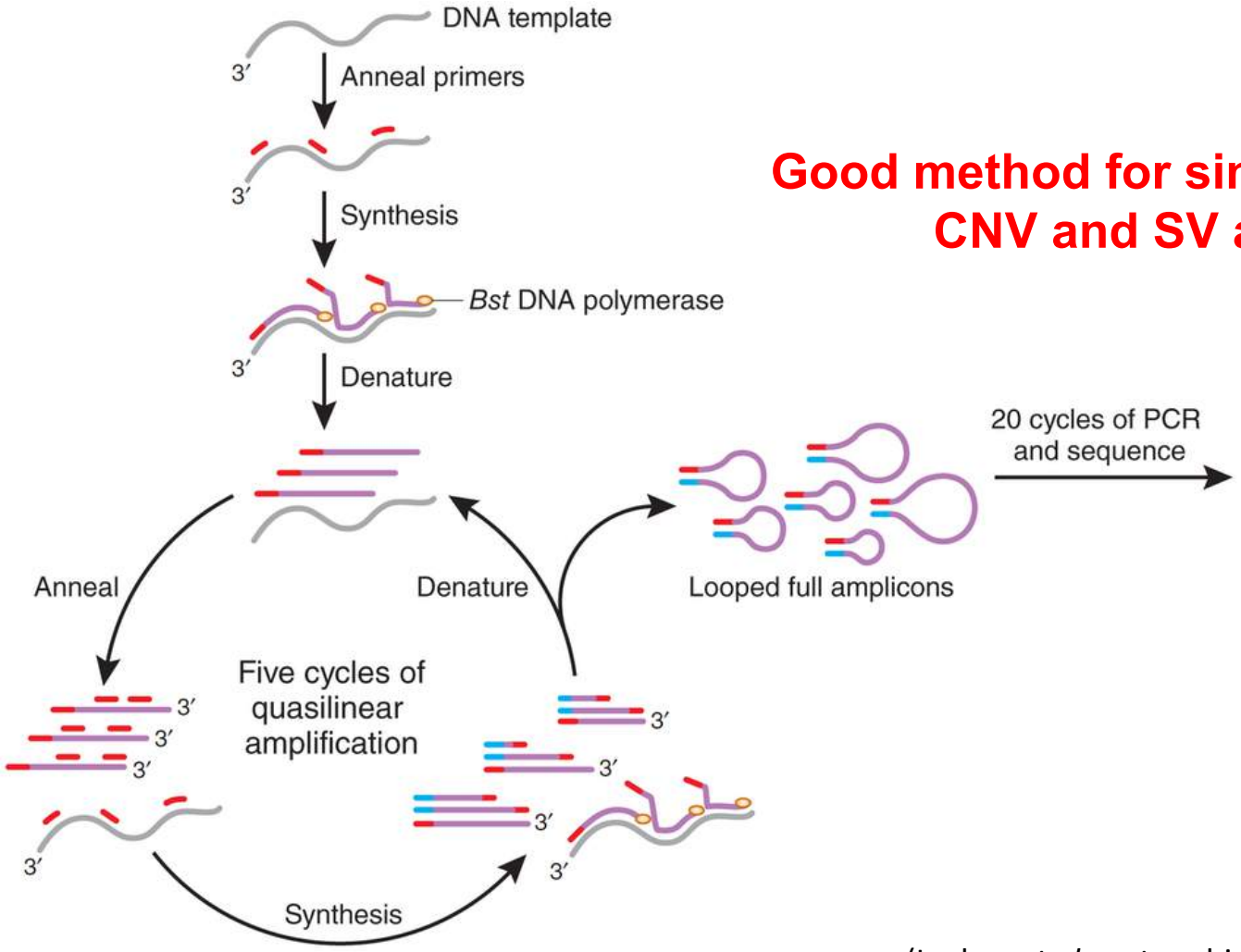


Good method for single cell genomics SNV analysis!!!

(Spits *et al*., nature protocol 2006)

PCR-based Genome amplification

Multiple annealing and looping-based amplification cycle
(eg. Picoplex, MALBAC)



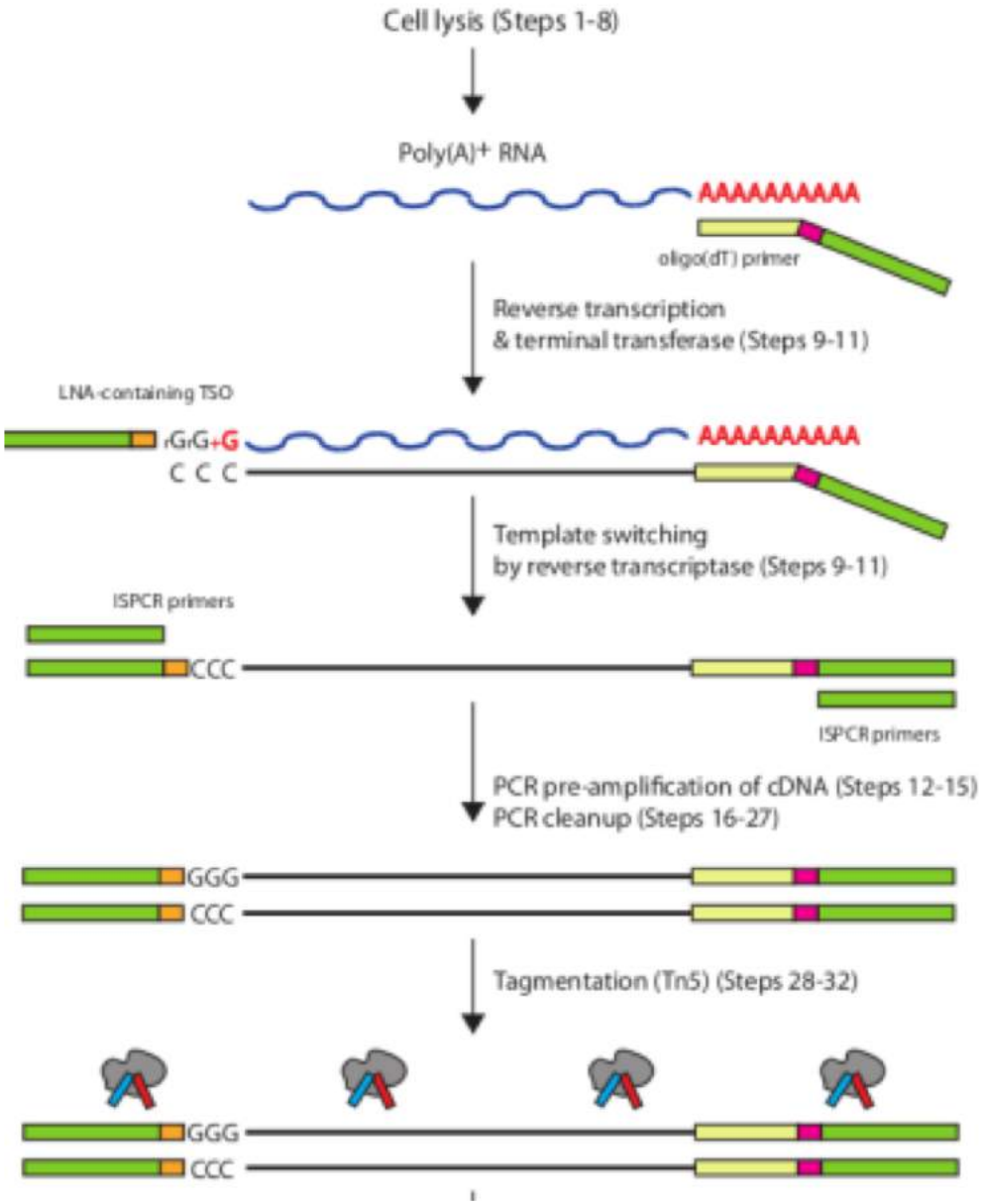
Good method for single cell genomic CNV and SV analysis!!!

(Lasken *et al.*, nature biotechnology 2013)

Whole genome transcriptome amplification

wellcome sanger institute

SMART-Seq2

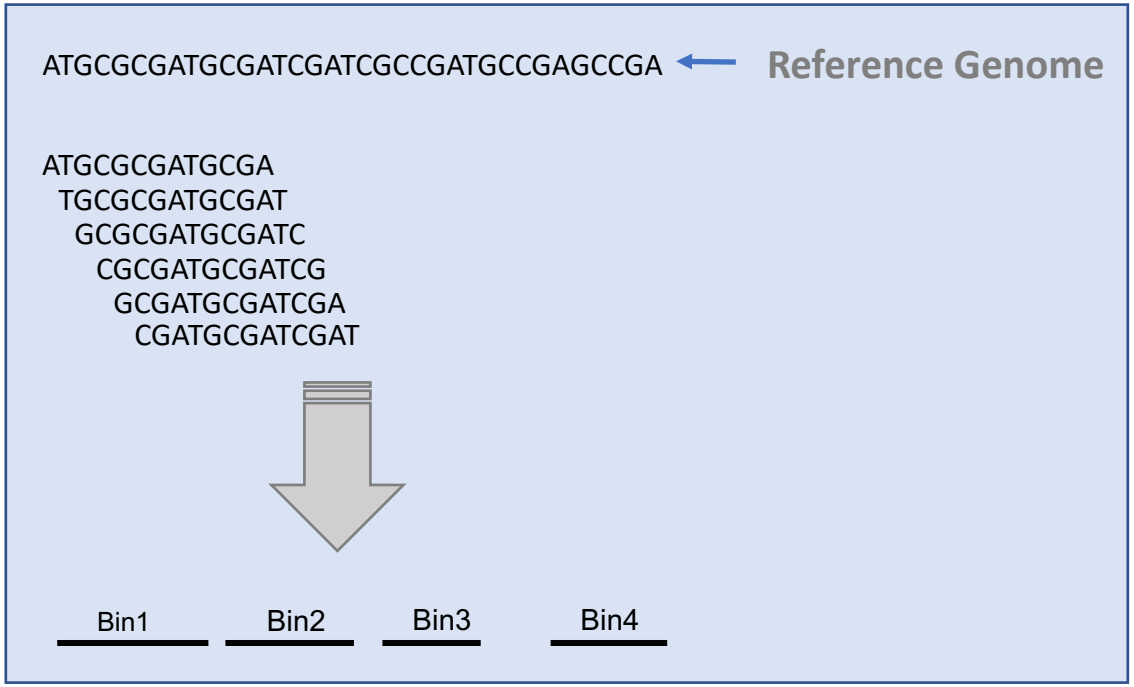

Cell lysis (Steps 1-8)
Poly(A)+ RNA
AAAAAAAAAA
oligo(dT) primer
Reverse transcription
& terminal transferase (Steps 9-11)
LNA-containing TSO
rGrG+G
C C C
AAAAAAAAAA
Template switching
by reverse transcriptase (Steps 9-11)
ISPCR primers
CCC
ISPCR primers
PCR pre-amplification of cDNA (Steps 12-15)
PCR cleanup (Steps 16-27)
GGG
CCC
Tagmentation (Tn5) (Steps 28-32)
GGG
CCC

Single cell CNV profiling

wellcome sanger institute

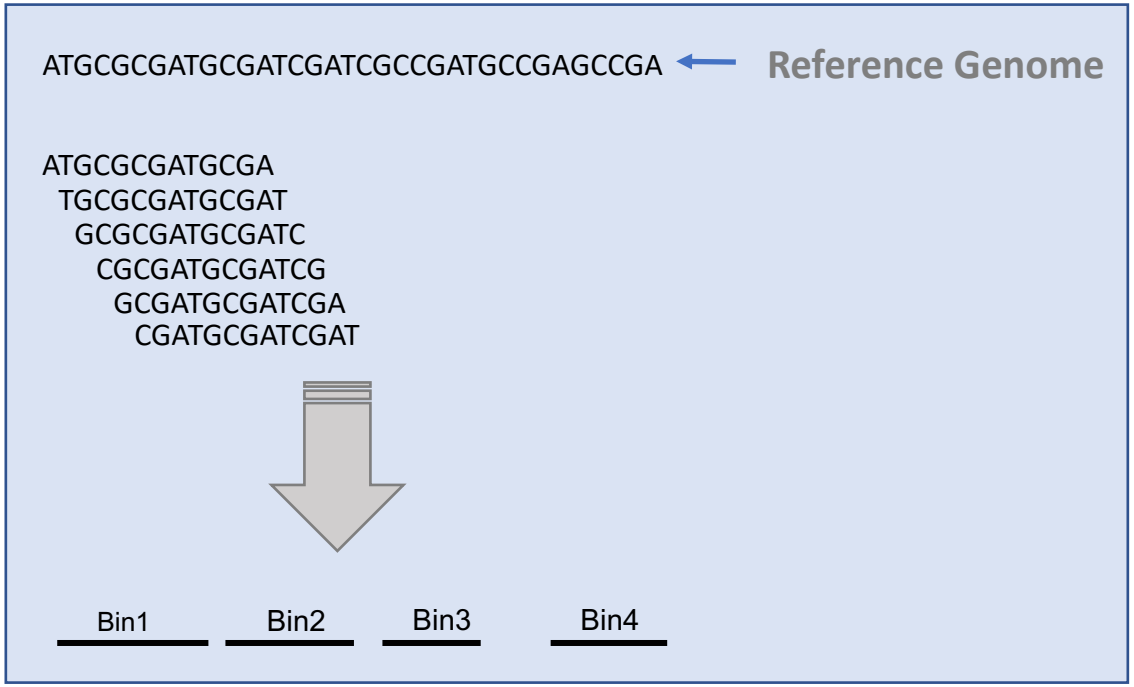
Define genomic bins

Generating artificial reads equal in length to the single-cell trimmed reads from every base in the human genome and mapping them back to the reference genome



Define genomic bins

Generating artificial reads equal in length to the single-cell trimmed reads from every base in the human genome and mapping them back to the reference genome

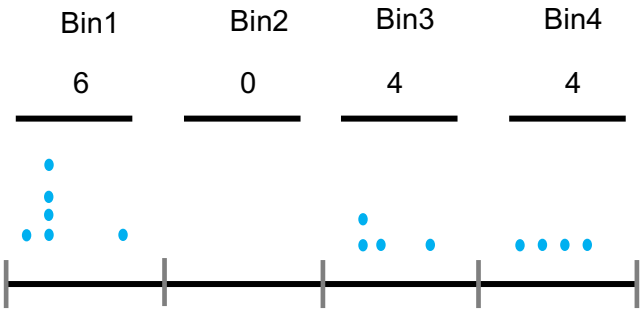


Subsequently, the human genome is divided into **non-overlapping bins of 500,000 uniquely mappable positions**, resulting in physical bin sizes of 514 kb on average (s.d. = 28 kb when 1% of the top bins was removed)

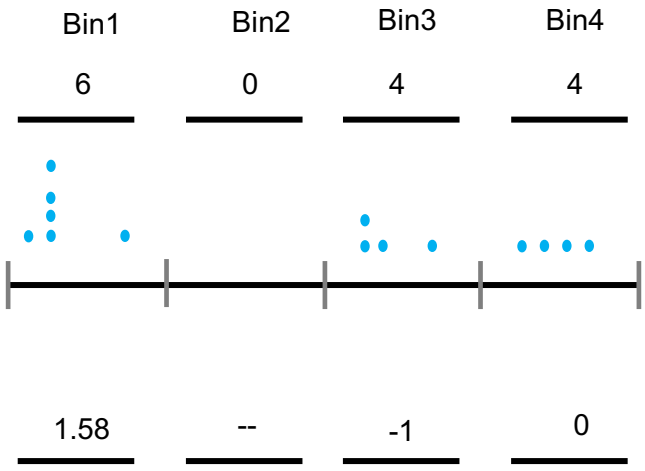
wellcome sanger institute

Focal read-depth analysis

The uniquely mapped reads of the cells were counted in these bins and bins with a %GC content of less than 28% were discarded.

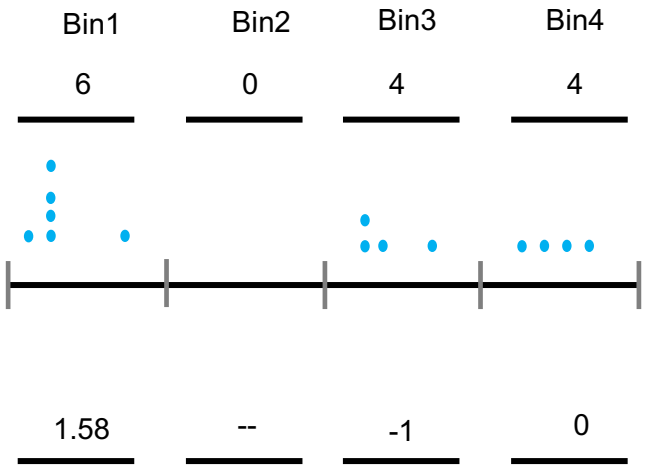


Focal read-depth analysis

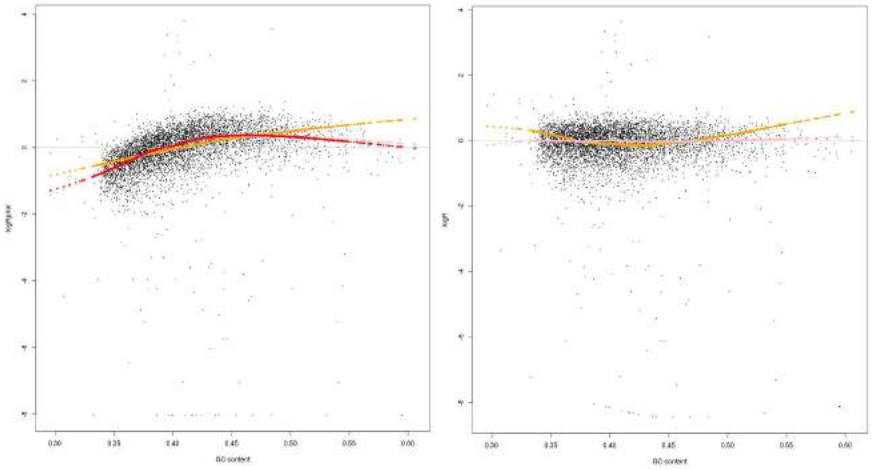


$\log_2$ ratio per bin by dividing the read count of a given bin by the average read count of the bins genome-wide.

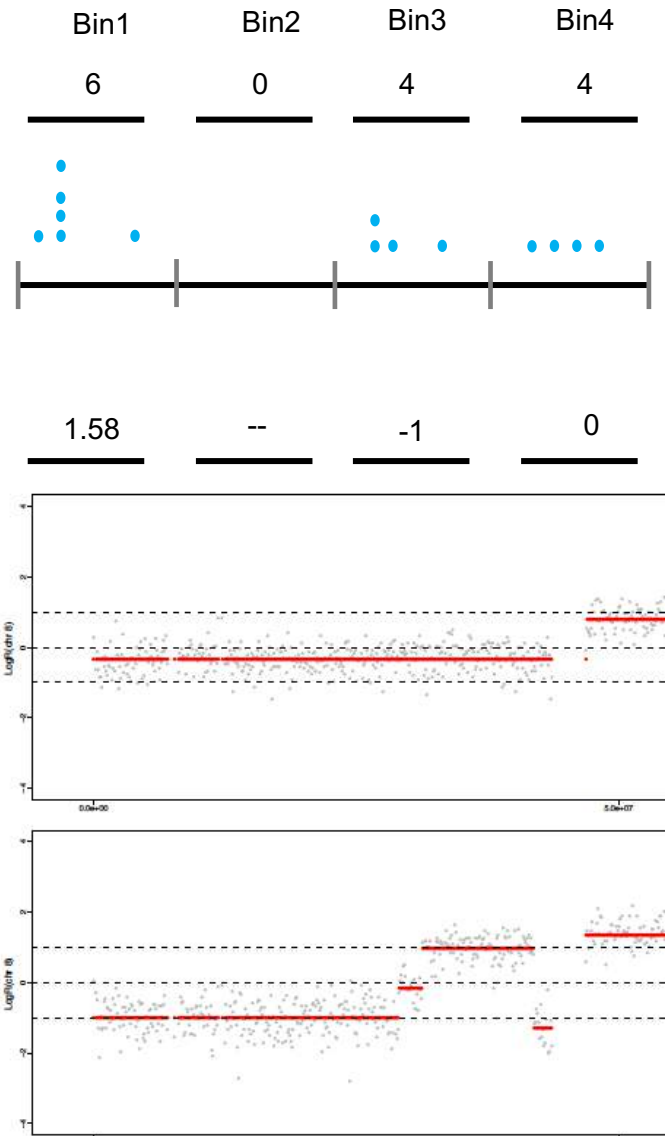
Focal read-depth analysis



The logR values were corrected for %GC bias using a Loess fit and were normalized according to the median of the genome-wide logR values.



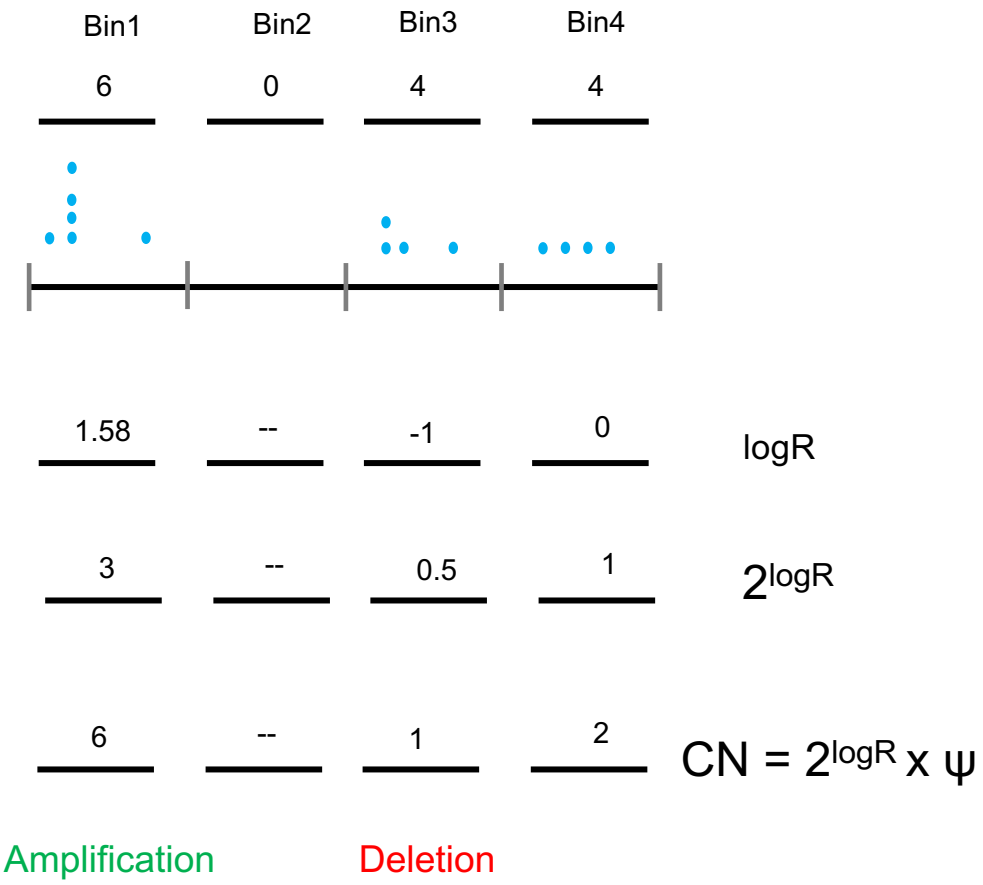
Focal read-depth analysis



Corrected logR values were segmented using PCF (γ= 15)

Segmentation with piecewise constant fitting:
the process of finding abrupt changes (steps, jumps, shifts) in the mean level of a time series or signal.

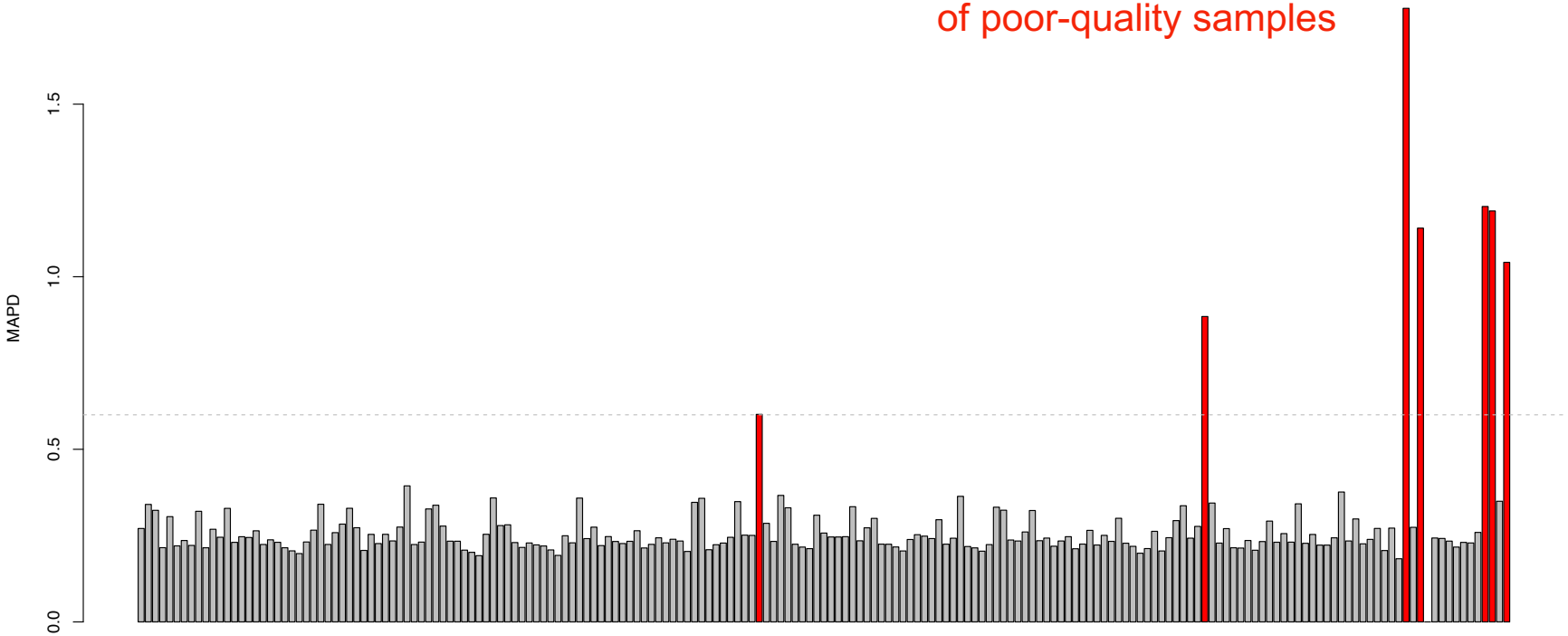
Focal read-depth analysis



Compute integer DNA copy number: $\mathbf{2^{logR} \times \Psi}$, where the average ploidy of the cell, Ψ, was estimated based on the logR value of a large reference region with known DNA copy number.

Mean absolute pairwise difference

High MAPD values result from greater noise, which is characteristic of poor-quality samples



$$\text{MAPD} = \text{median}\left(\left|\ \log R \text{ of } bin_{k+1} - \log R \text{ of } bin_{k}\ \right|\right)$$

k represents a specific bin on a specific chromosome across the genome

wellcome sanger institute

Introduction to Single cell RNA analysis

wellcome sanger institute

ScRNA-seq

•Did not gain widespread popularity until 2014 when new protocols and lower sequencing costs made it more accessible

•Measures the **distribution of expression levels** for each gene across a population of cells

•Allows to study new biological questions in which **cell-specific changes in transcriptome are important**, e.g. cell type identification, heterogeneity of cell responses, stochasticity of gene expression, inference of gene regulatory networks across the cells.

•Currently there are several different protocols in use, e.g. SMART-seq2 (Picelli et al. 2013), CELL-seq and Drop-seq

•There are also commercial platforms available, including the Fluidigm C1, Wafergen ICELL8 and the 10X Genomics Chromium

wellcome sanger institute

ScRNA-seq

•Did not gain widespread popularity until 2014 when new protocols and lower sequencing costs made it more accessible

•Measures the **distribution of expression levels** for each gene across a population of cells

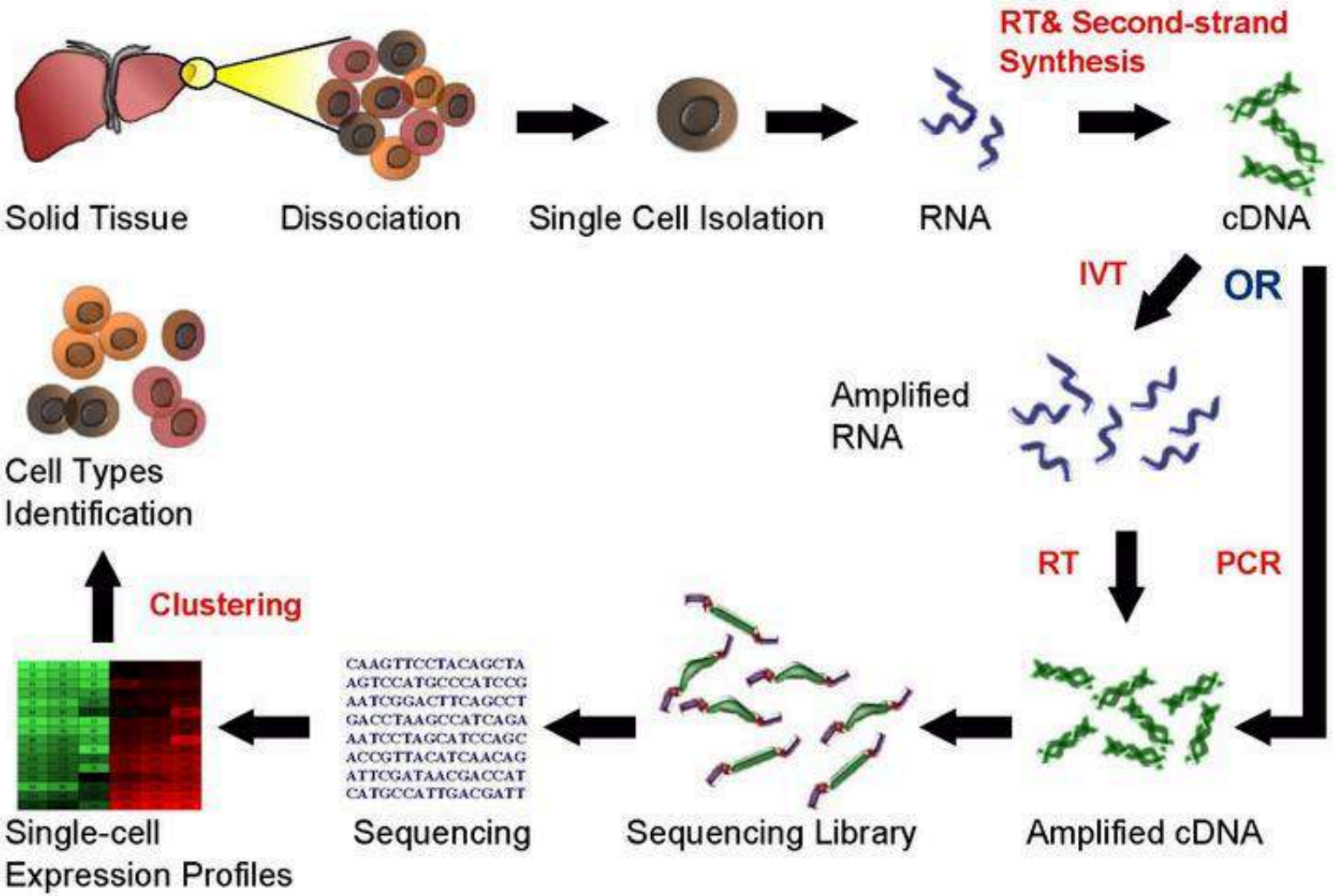
•Allows to study new biological questions in which **cell-specific changes in transcriptome are important**, e.g. cell type identification, heterogeneity of cell responses, stochasticity of gene expression, inference of gene regulatory networks across the cells.

•Currently there are several different protocols in use, e.g. SMART-seq2 (Picelli et al. 2013), CELL-seq and Drop-seq

•There are also commercial platforms available, including the Fluidigm C1, Wafergen ICELL8 and the 10X Genomics Chromium

wellcome sanger institute

Single Cell RNA Sequencing Workflow



Challenges

The main difference between bulk and single cell RNA-seq is that each sequencing library represents a single cell, instead of a population of cells. Therefore, significant attention has to be paid to comparison of the results from different cells (sequencing libraries). The main sources of discrepancy between the libraries are:

•**Amplification** (up to 1 million fold)

•**Gene 'dropouts'** in which a gene is observed at a moderate expression level in one cell but is not detected in another cell

In both cases the discrepancies are introduced due to low starting amounts of transcripts since the RNA comes from one cell only.

However, it is possible to alleviate some of these issues through proper normalization and corrections.

wellcome sanger institute

What platform to use for my experiment?

The most suitable platform depends on the biological question at hand.

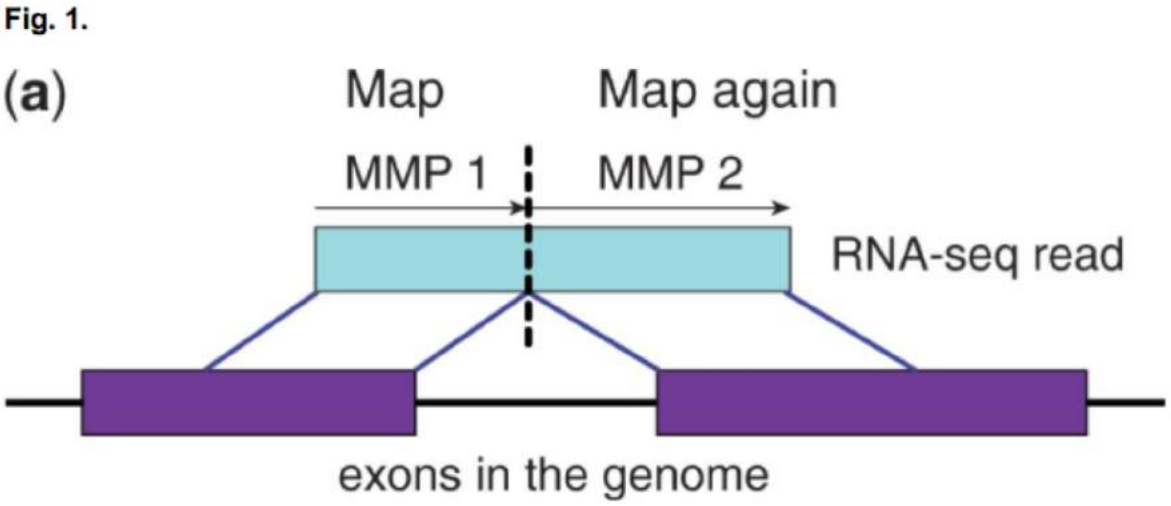
- If one is interested in characterizing the composition of a tissue, then a droplet-based method which will allow a very large number of cells to be captured is likely to be the most appropriate.

- On the other hand, if one is interesting in characterizing a rare cell-population for which there is a known surface marker, then it is probably best to enrich using FACS and then sequence a smaller number of cells.

wellcome sanger institute

Using STAR to Align Reads

For each read in our reads data, STAR tries to find the longest possible sequence which matches one or more sequences in the reference genome.

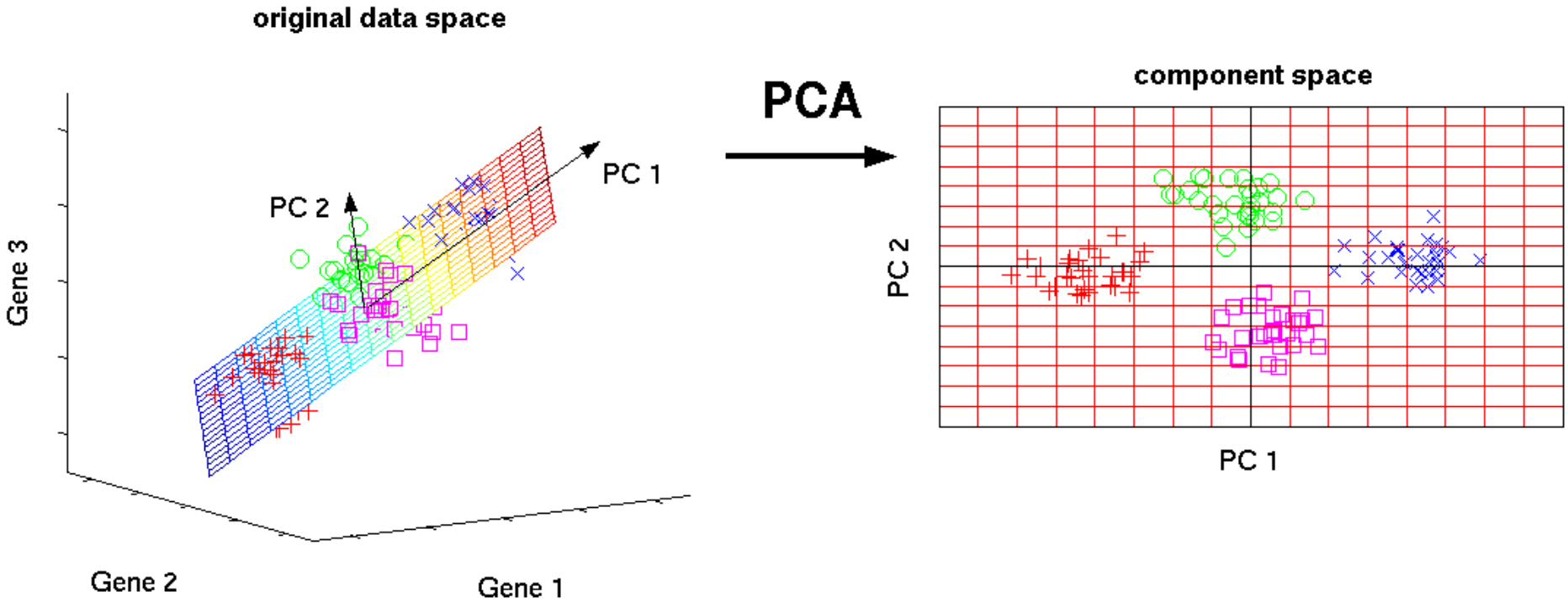


Blue read above, spans two exons and an alternative splicing junction (purple).

STAR finds that the first part of the read is the same as the sequence of the first exon, whilst the second part of the read matches the sequence in the second exon. Because STAR is able to recognise splicing events in this way, it is described as a ‘splice aware’ aligner.

Principal Component Analysis (PCA)

PCA is a statistical procedure that uses a transformation to convert a set of observations into a set of values of linearly uncorrelated variables called principal components (PCs).



Example of Principal Component Analysis