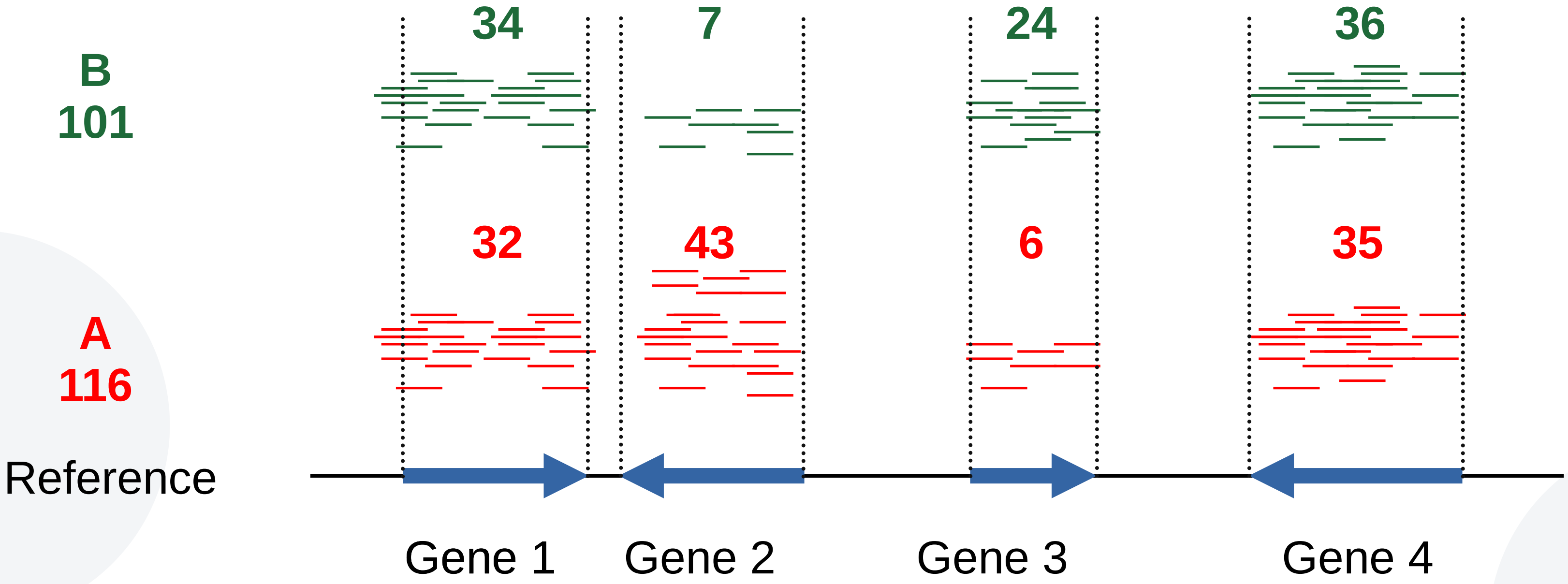
Biological data analysis

Single cell data analysis

Juan D. Montenegro

January 15th, 2024

Summarising read counts



	A	B	C
Gene 1	34	32	36
Gene 2	2	0	16
Gene 3	0	2	0
Gene 4	4	6	10
Gene 5	2	28	32
Gene 6	7	32	33
Gene 7	90	16	17
Gene 8	13	0	13
TOTAL	**152**	**116**	**157**

Illustrations by Pixeltrue on
icons8

Read assignment

Process that compares read mapping position with gene locations and assigns reads to genes.

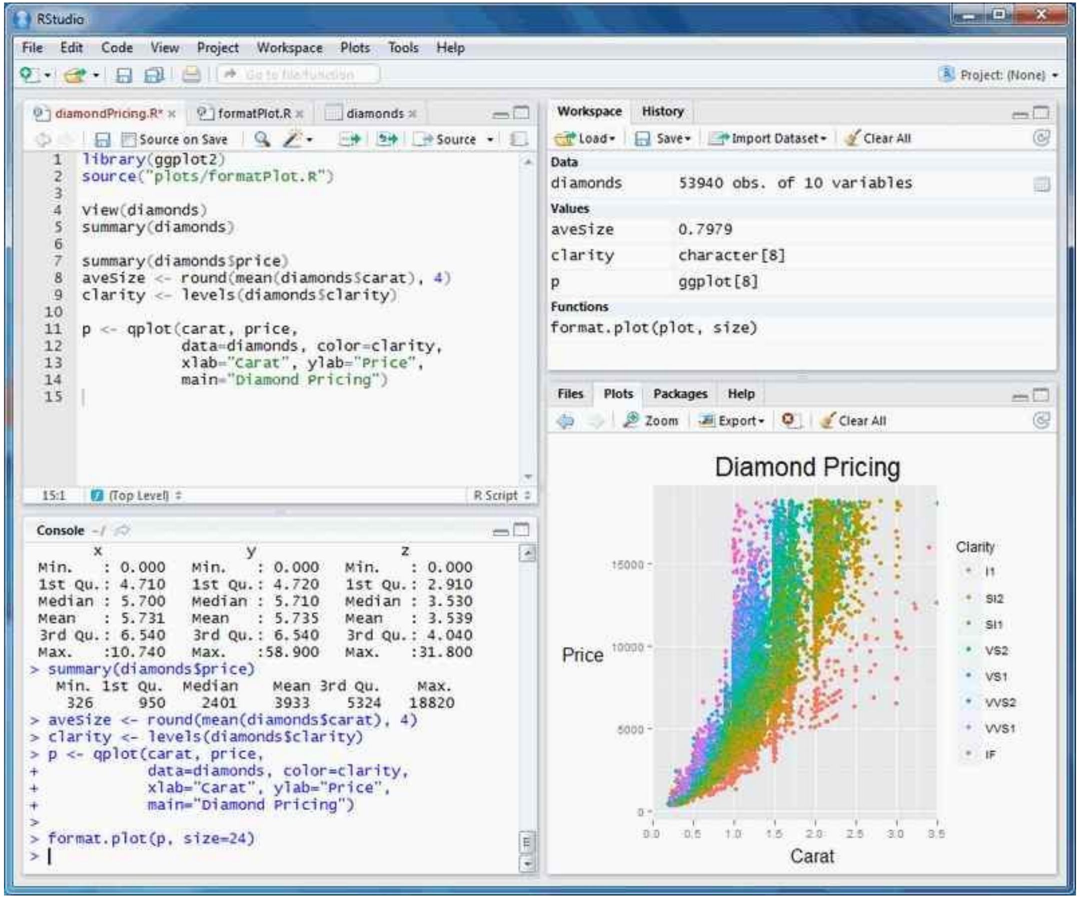
Reads assigned to genes can be of three types:

- CDS
- UTR
- Intron

Seventh Exercise

- Run featureCounts:
- Package Subread/ command featureCounts
- Write a slurm script and submit it.
- Review results:
  - Summary
  - Count Matrix

Analising read counts Using R



```
library(ggplot2)
source("plots/formatPlot.R")

View(diamonds)
summary(diamonds)

summary(diamonds$price)
aveSize <- round(mean(diamonds$carat), 4)
clarity <- levels(diamonds$clarity)

p <- qplot(carat, price,
           data=diamonds, color=clarity,
           xlab="Carat", ylab="Price",
           main="Diamond Pricing")

```

```
Data
diamonds      53940 obs. of 10 variables
Values
aveSize       0.7979
clarity       character[8]
p             ggplot[8]
Functions
format.plot(plot, size)
```

```
       x                y                z
 Min.   : 0.000   Min.   : 0.000   Min.   : 0.000
 1st Qu.: 4.710   1st Qu.: 4.720   1st Qu.: 2.910
 Median : 5.700   Median : 5.710   Median : 3.530
 Mean   : 5.731   Mean   : 5.735   Mean   : 3.539
 3rd Qu.: 6.540   3rd Qu.: 6.540   3rd Qu.: 4.040
 Max.   :10.740   Max.   :58.900   Max.   :31.800
> summary(diamonds$price)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    326     950    2401    3933    5324   18820
> aveSize <- round(mean(diamonds$carat), 4)
> clarity <- levels(diamonds$clarity)
> p <- qplot(carat, price,
+            data=diamonds, color=clarity,
+            xlab="Carat", ylab="Price",
+            main="Diamond Pricing")
>
> format.plot(p, size=24)
> 
```

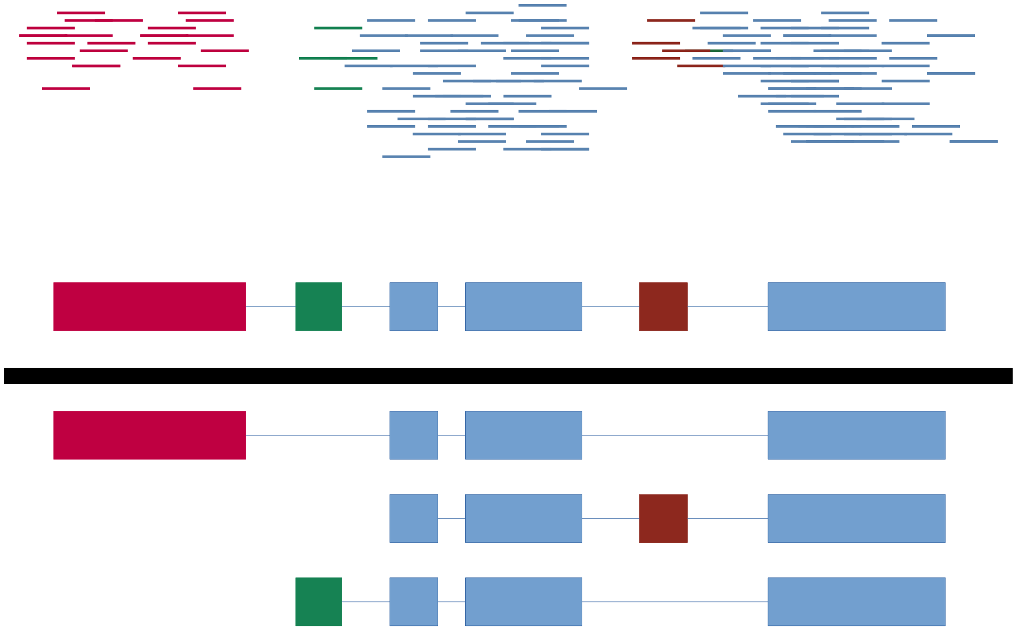
Eighth Exercise

- Open Rstudio on the LiSC
  - http://rstudio.lisc.univie.ac.at
- Basic commands:
  - getwd() / setwd()
  - c()
  - read.table() / read_tsv()
- Main libraries:
  - tidyverse
  - ggplot2

Eighth Exercise

- Plot basic statistics from the featureCounts results
- Write an R script in Rstudio, save it and share it on GitHub.
- Perform PCA analysis to identify biological replicates

Mapping reads to transcriptome



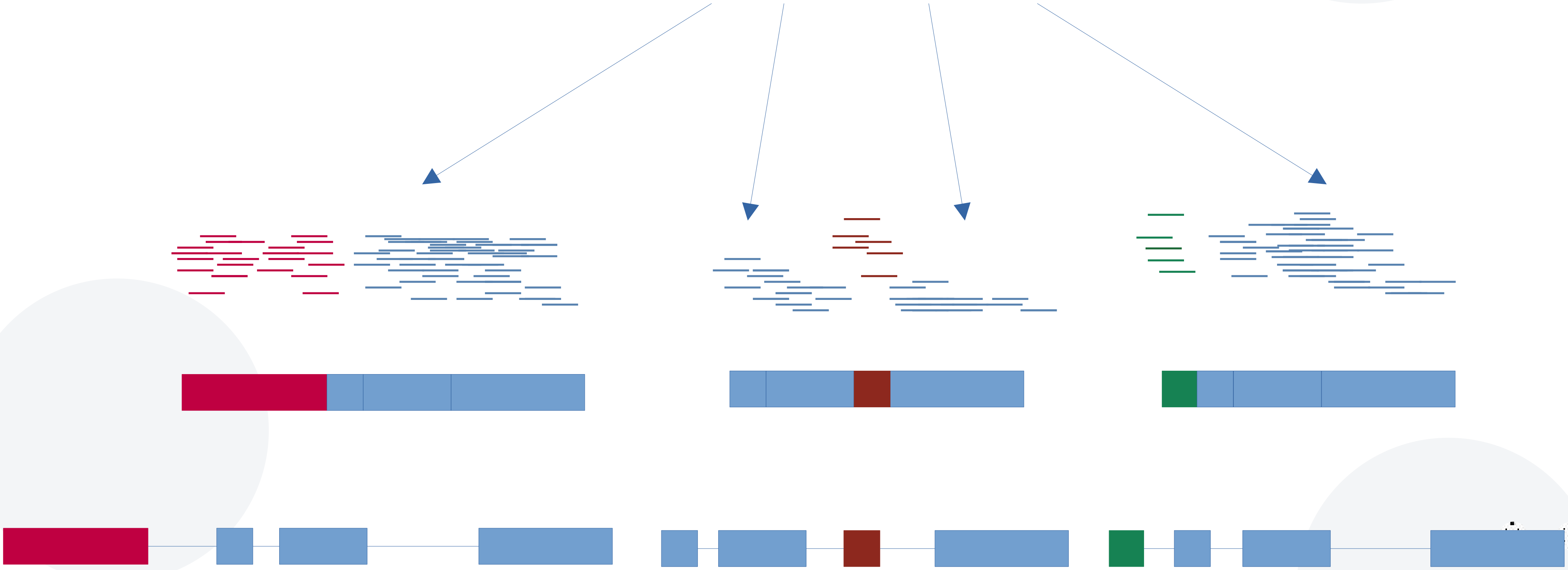
Reads map to unique place in the genome

Gene model

Reference

Isoforms

Reads map to multiple places in the transcriptome

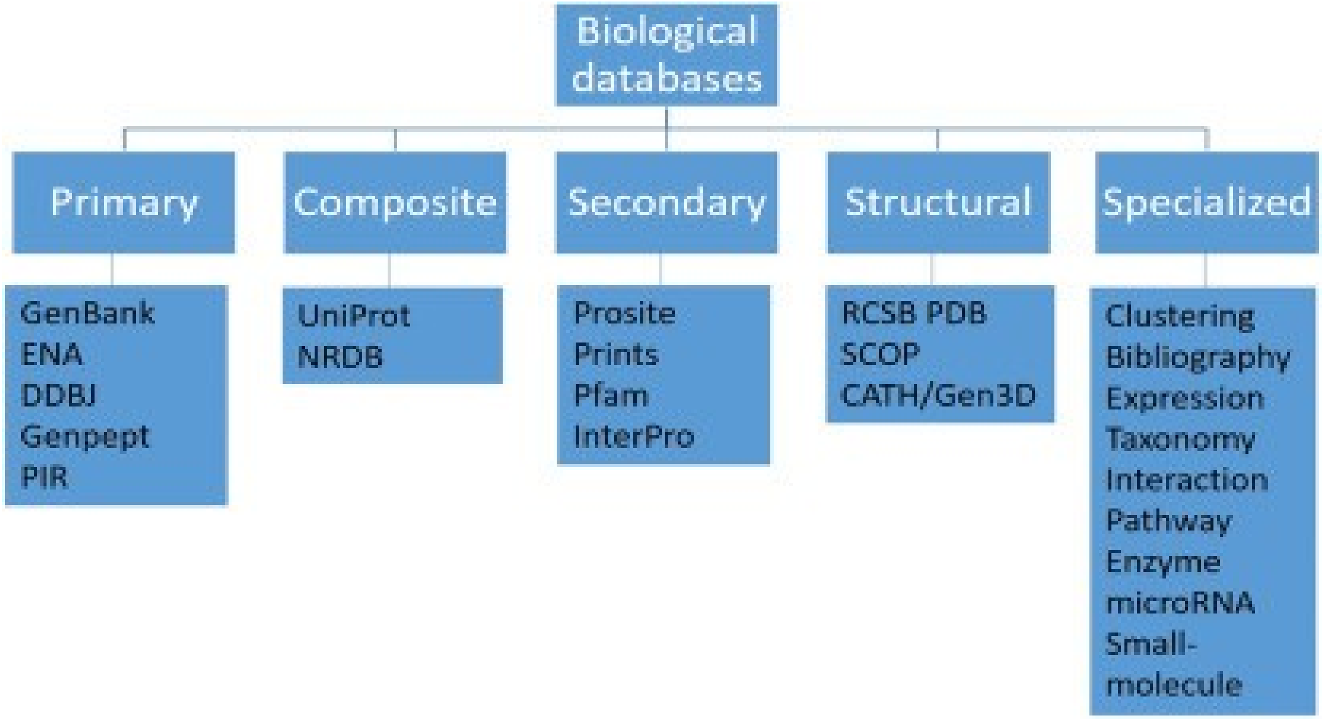


Eigth Exercise

- Download a reference transcriptome of Nematostella vectensis from the European Nucleotide Archive (ENA)
- Align the reads to the reference transcriptome
- Assign reads to genes
- Compare mapping efficiency and assignment efficiency between transcriptome and genome using R

Functional annotation of genes

Process of assigning functions to genes
Relies on high throughput comparison to large annotated databases



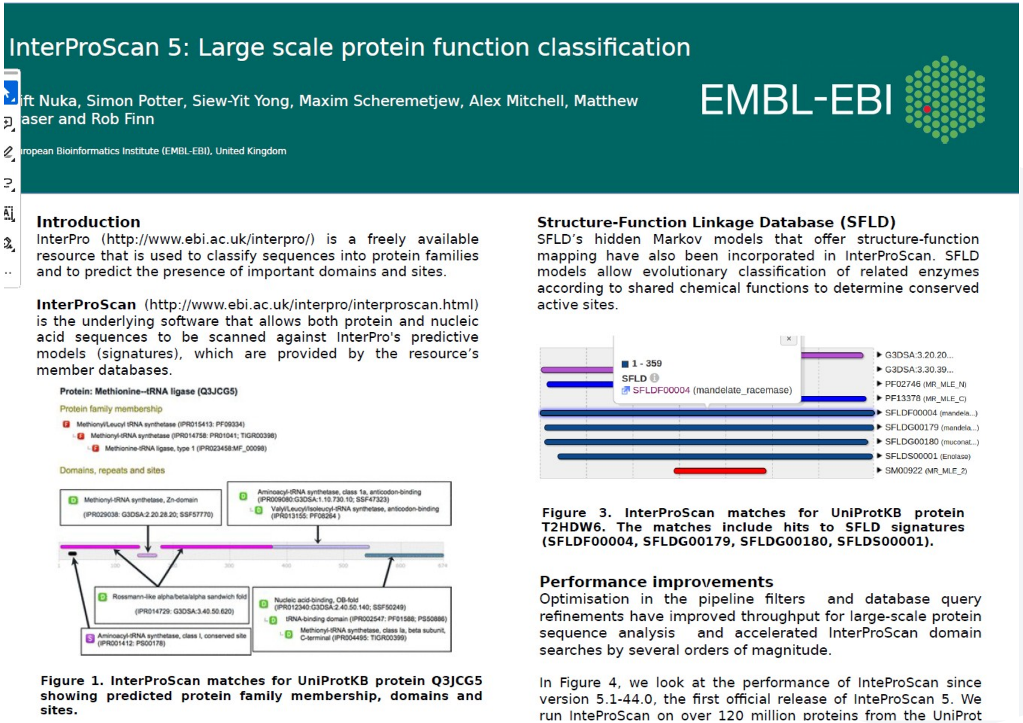
Biological databases

- Primary
  - GenBank
  - ENA
  - DDBJ
  - Genpept
  - PIR
- Composite
  - UniProt
  - NRDB
- Secondary
  - Prosite
  - Prints
  - Pfam
  - InterPro
- Structural
  - RCSB PDB
  - SCOP
  - CATH/Gen3D
- Specialized
  - Clustering
  - Bibliography
  - Expression
  - Taxonomy
  - Interaction
  - Pathway
  - Enzyme
  - microRNA
  - Small-molecule

Sharma & Yadav (2022) Biological databases and their application; Bioinformatics: Methods and applications:17-31

EggNogMapper

InterProScan



InterProScan 5: Large scale protein function classification

ift Nuka, Simon Potter, Siew-Yit Yong, Maxim Scheremetjew, Alex Mitchell, Matthew
aser and Rob Finn

ropean Bioinformatics Institute (EMBL-EBI), United Kingdom

EMBL-EBI

Introduction

InterPro (http://www.ebi.ac.uk/interpro/) is a freely available resource that is used to classify sequences into protein families and to predict the presence of important domains and sites.

InterProScan (http://www.ebi.ac.uk/interpro/interproscan.html) is the underlying software that allows both protein and nucleic acid sequences to be scanned against InterPro's predictive models (signatures), which are provided by the resource’s member databases.

Protein: Methionine--tRNA ligase (Q3JCG5)
Protein family membership
F Methionyl/Leucyl tRNA synthetase (IPR015413: PF09334)
F Methionyl-tRNA synthetase (IPR014758: PR01041; TIGR00398)
F Methionine-tRNA ligase, type 1 (IPR023458:MF_00098)
Domains, repeats and sites
D Methionyl-tRNA synthetase, Zn-domain (IPR029038: G3DSA:2.20.28.20; SSF57770)
D Aminoacyl-tRNA synthetase, class 1a, anticodon-binding (IPR009080:G3DSA:1.10.730.10; SSF47323)
D Valyl/Leucyl/Isoleucyl-tRNA synthetase, anticodon-binding (IPR013155: PF08264)
D Rossmann-like alpha/beta/alpha sandwich fold (IPR014729: G3DSA:3.40.50.620)
D Nucleic acid-binding, OB-fold (IPR012340:G3DSA:2.40.50.140; SSF50249)
D tRNA-binding domain (IPR002547: PF01588; PS50886)
D Methionyl-tRNA synthetase, class Ia, beta subunit, C-terminal (IPR004495: TIGR00399)
S Aminoacyl-tRNA synthetase, class I, conserved site (IPR001412: PS00178)

Figure 1. InterProScan matches for UniProtKB protein Q3JCG5 showing predicted protein family membership, domains and sites.

Structure-Function Linkage Database (SFLD)

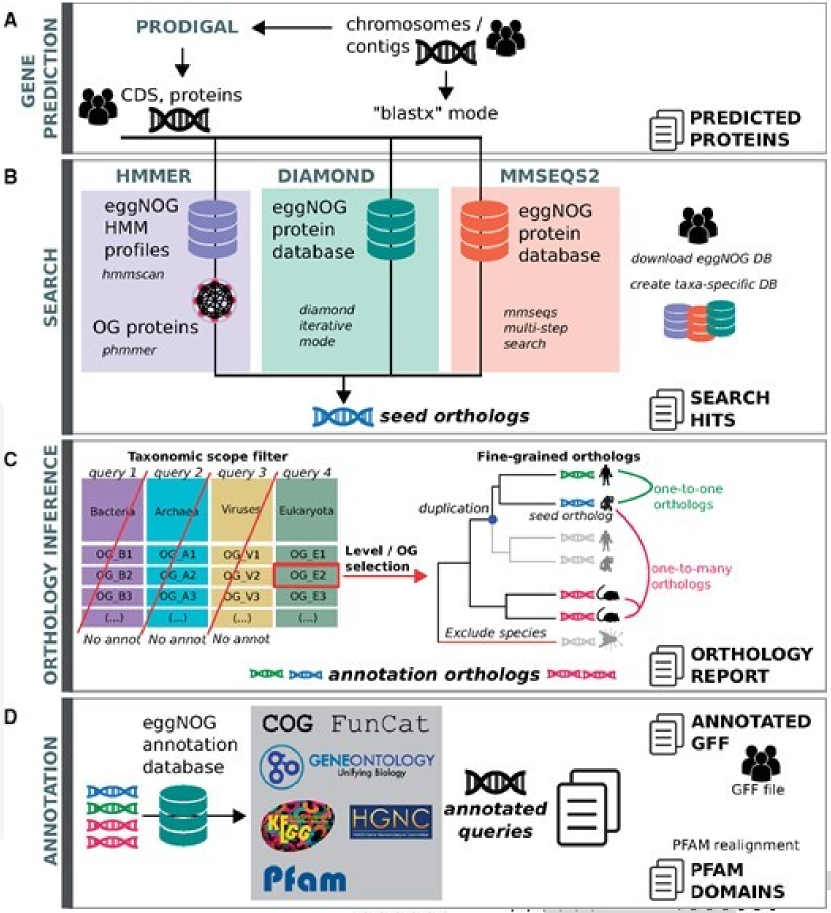
SFLD’s hidden Markov models that offer structure-function mapping have also been incorporated in InterProScan. SFLD models allow evolutionary classification of related enzymes according to shared chemical functions to determine conserved active sites.

Figure 3. InterProScan matches for UniProtKB protein T2HDW6. The matches include hits to SFLD signatures (SFLDF00004, SFLDG00179, SFLDG00180, SFLDS00001).

Performance improvements

Optimisation in the pipeline filters and database query refinements have improved throughput for large-scale protein sequence analysis and accelerated InterProScan domain searches by several orders of magnitude.

In Figure 4, we look at the performance of InteProScan since version 5.1-44.0, the first official release of InteProScan 5. We run InteProScan on over 120 million proteins from the UniProt



Ninth Exercise

- Use gffread to extract protein sequences from genome
- Use blastp to align proteins to uniref
- Use interProScan and EggNogMapper to add GO term annotation and identify conserved motifs
-

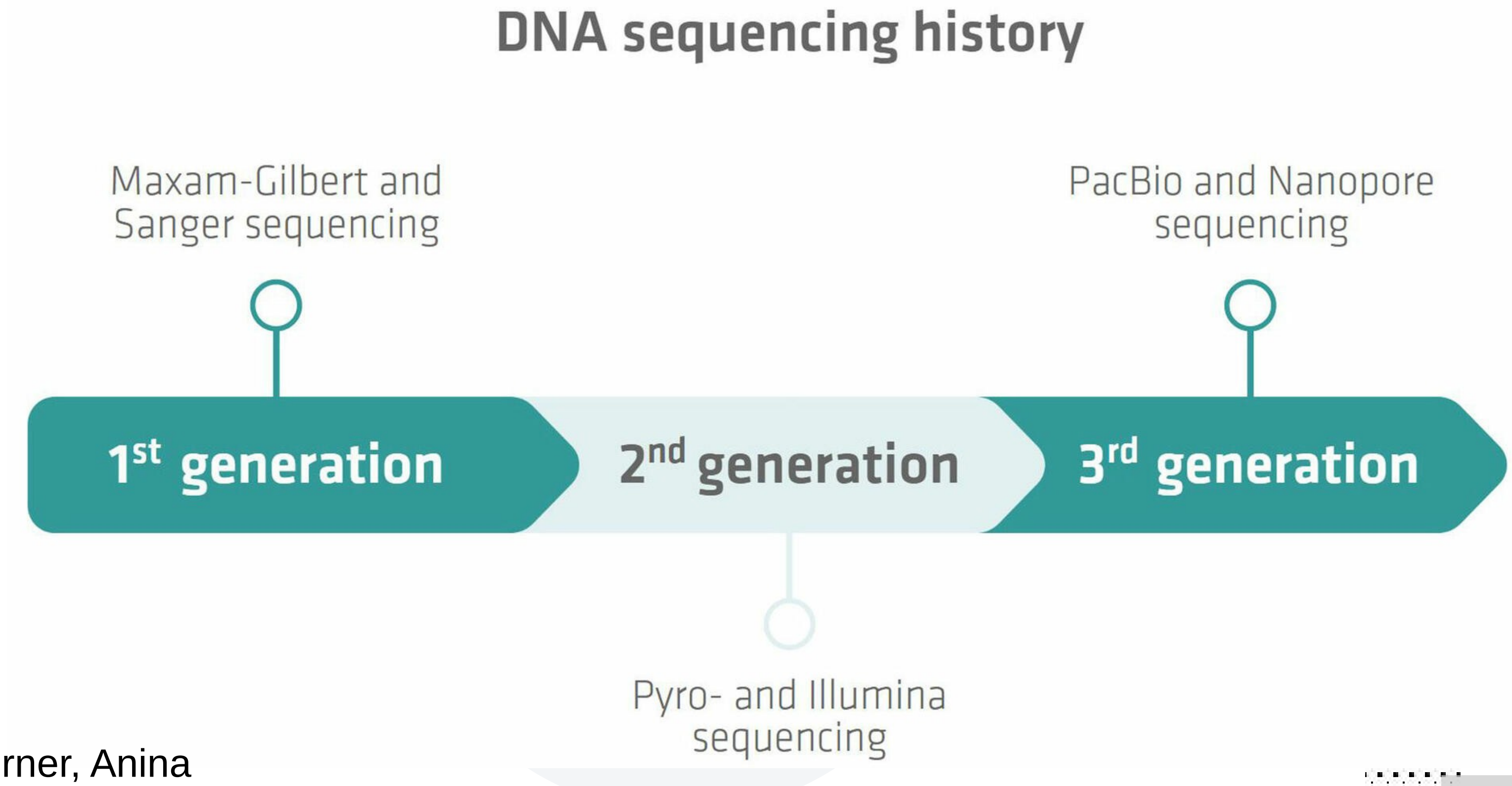
Assessing the mapping tool

Using external and internal clues to determine how useful a mapping tool is.

1) Contiguity: Fragmentation, N50, NG50, AuN curve
2) Coherence: Size estimation, proper mapping of reads
3) Completeness: Mapping efficiency of different data sources (cloned genes, ESTs, RNAseq, BAC ends, proteins)
4) Correctness: DNA sequence variation compared to actual known sequences

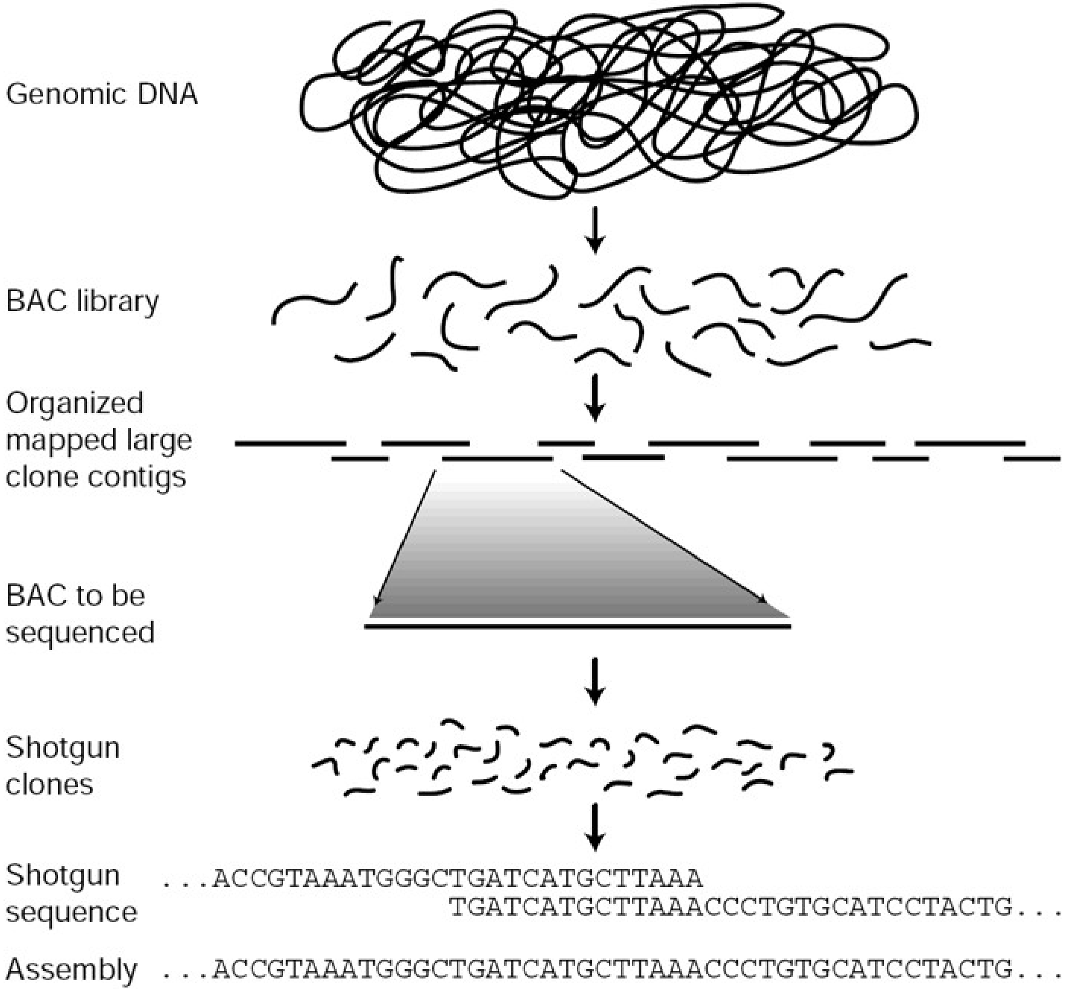
Contiguity:

Assembly fragmentation depends on technology


DNA sequencing history
Maxam-Gilbert and Sanger sequencing
PacBio and Nanopore sequencing
1st generation
2nd generation
3rd generation
Pyro- and Illumina sequencing

Creator: Werner, Anina
https://www.integra-biosciences.com/canada/en/blog/article/dna-sequencing-methods-sanger-ngs

Hierarchical shotgun sequencing



Genomic DNA

BAC library

Organized mapped large clone contigs

BAC to be sequenced

Shotgun clones

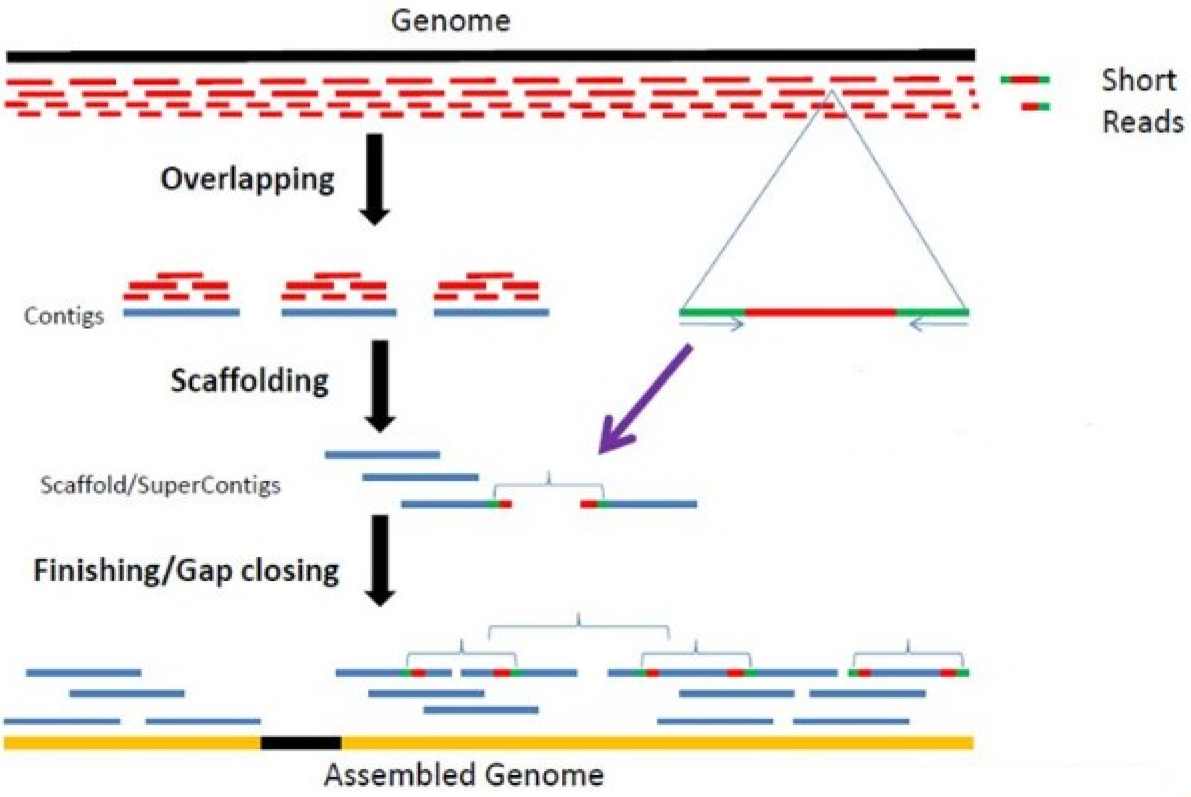
Shotgun sequence

```
...ACCGTAAATGGGCTGATCATGCTTAAA
                   TGATCATGCTTAAACCCTGTGCATCCTACTG...
```

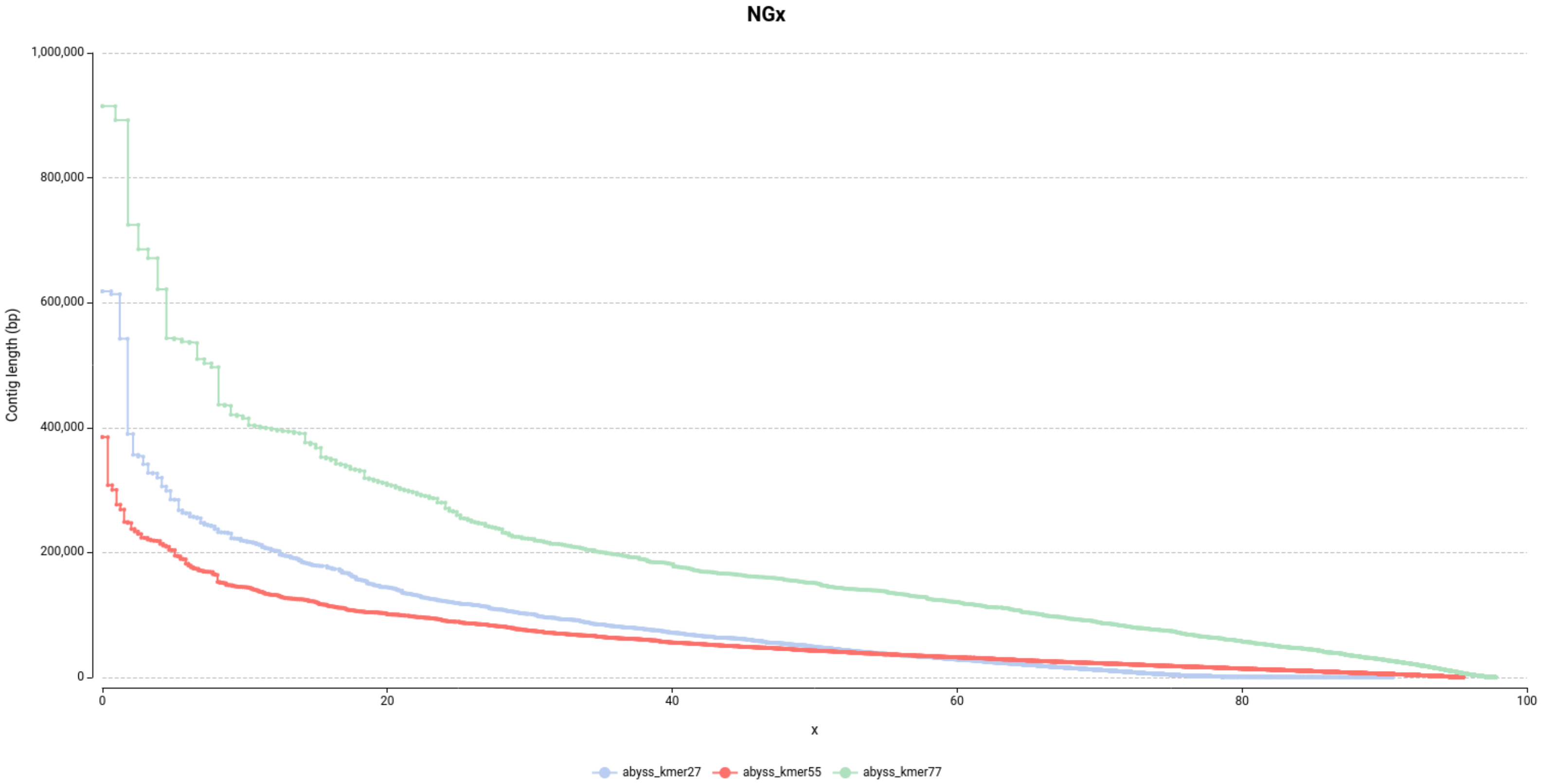
Assembly

```
...ACCGTAAATGGGCTGATCATGCTTAAACCCTGTGCATCCTACTG...
```

Whole genome shotgun sequencing



Level of fragmentation (AuN)

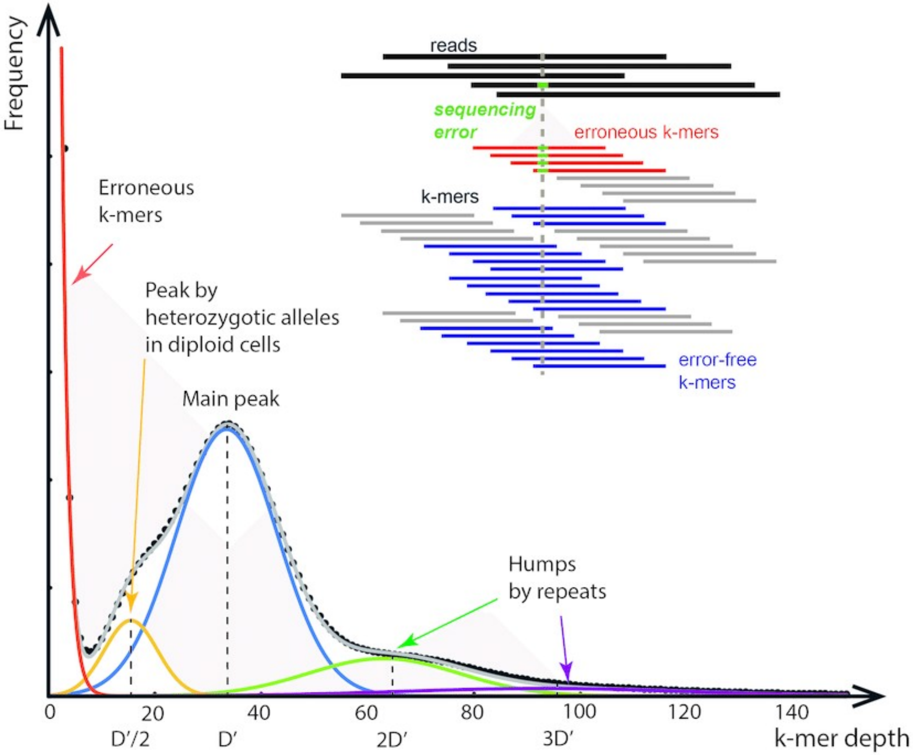


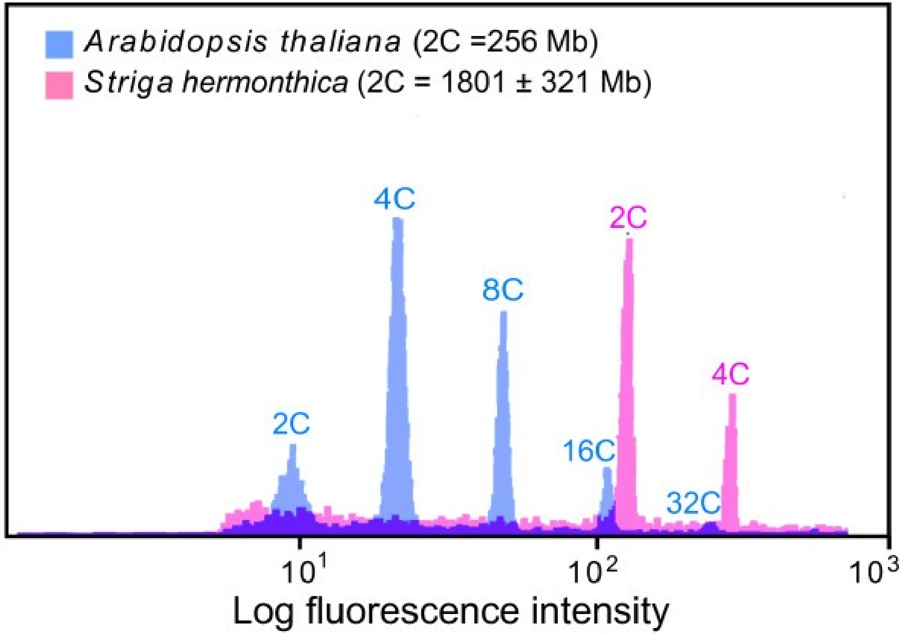
Coherence:

Estimating genome size

In silico: kmer analysis

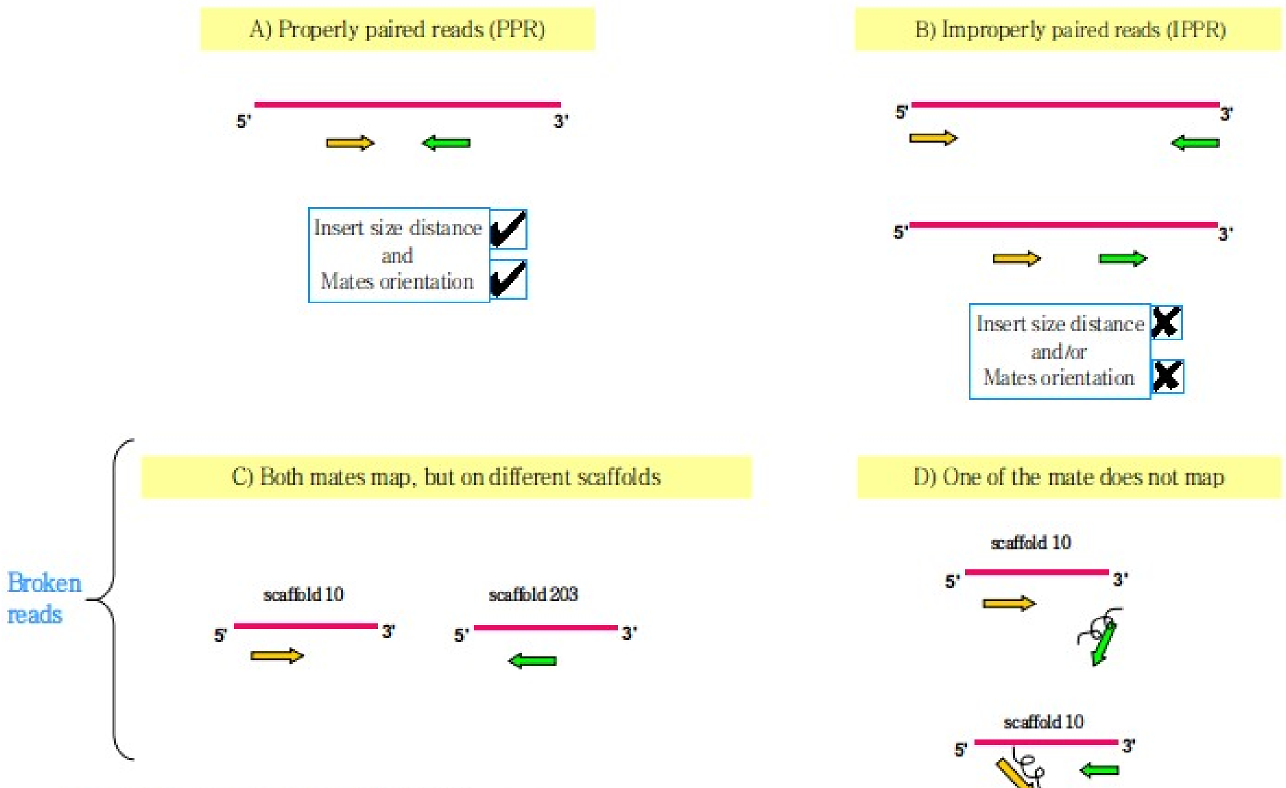
Flow cytometry





Coherence:
Read distance and orientation

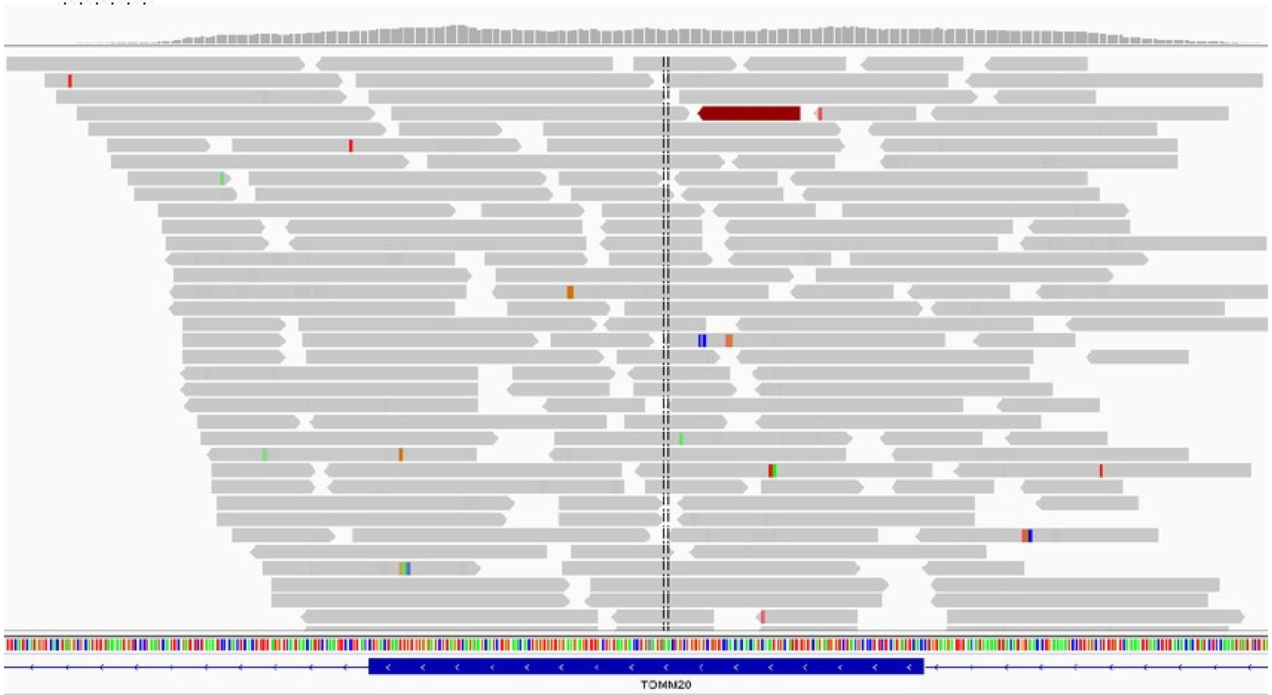
Mapped reads


A) Properly paired reads (PPR)
5'
3'
Insert size distance and Mates orientation ✓ ✓
B) Improperly paired reads (IPPR)
5'
3'
5'
3'
Insert size distance and/or Mates orientation ✗ ✗
Broken reads
C) Both mates map, but on different scaffolds
scaffold 10
5'
3'
scaffold 203
5'
3'
D) One of the mate does not map
scaffold 10
5'
3'
scaffold 10
5'
3'

Category B – potential misassembled regions
Category C – scaffolding purposes (if read coverage is high)

Completeness

Mapping efficiency

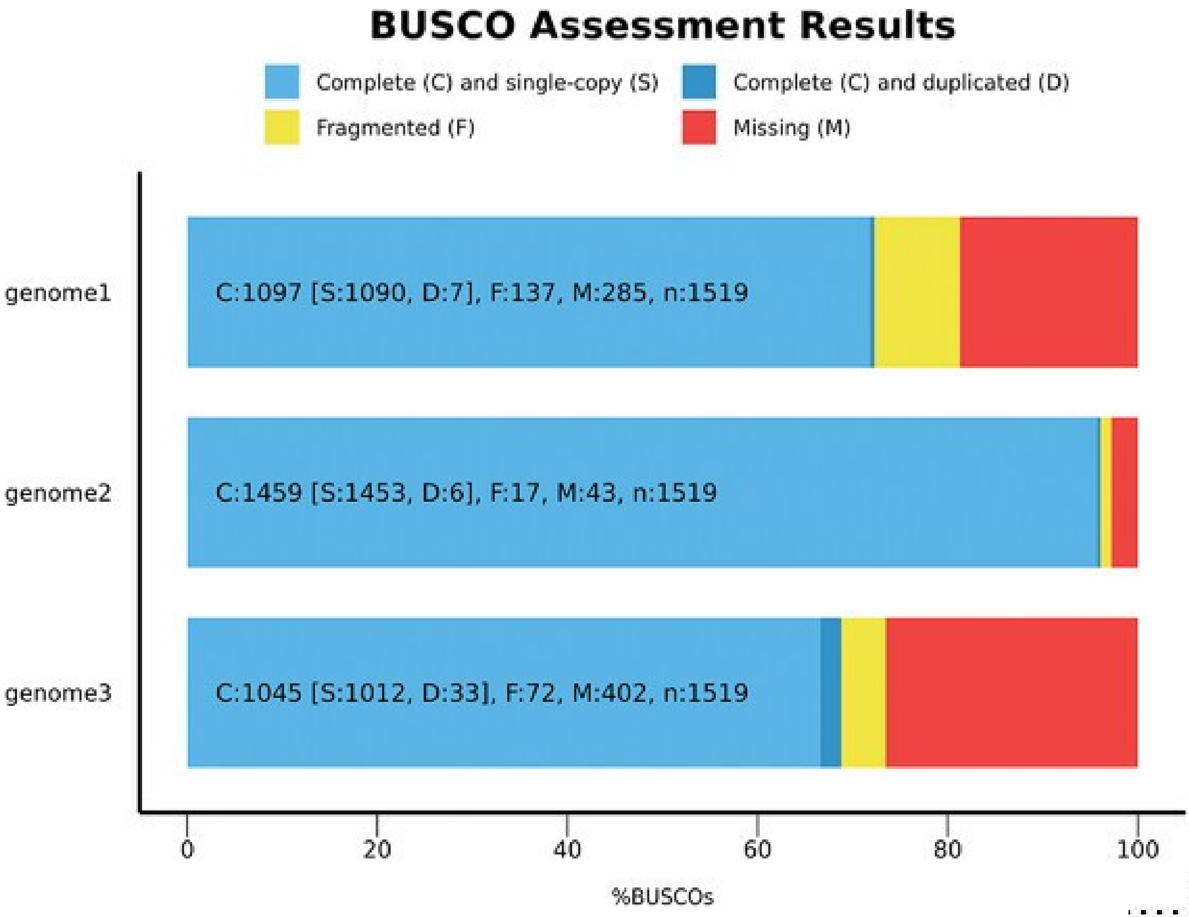


Alignment of:

- Assembly reads
- RNAseq
- Cloned genes
- Genomic surveillance sequences (EST, BES, Fosmids, flcDNA)

Completeness

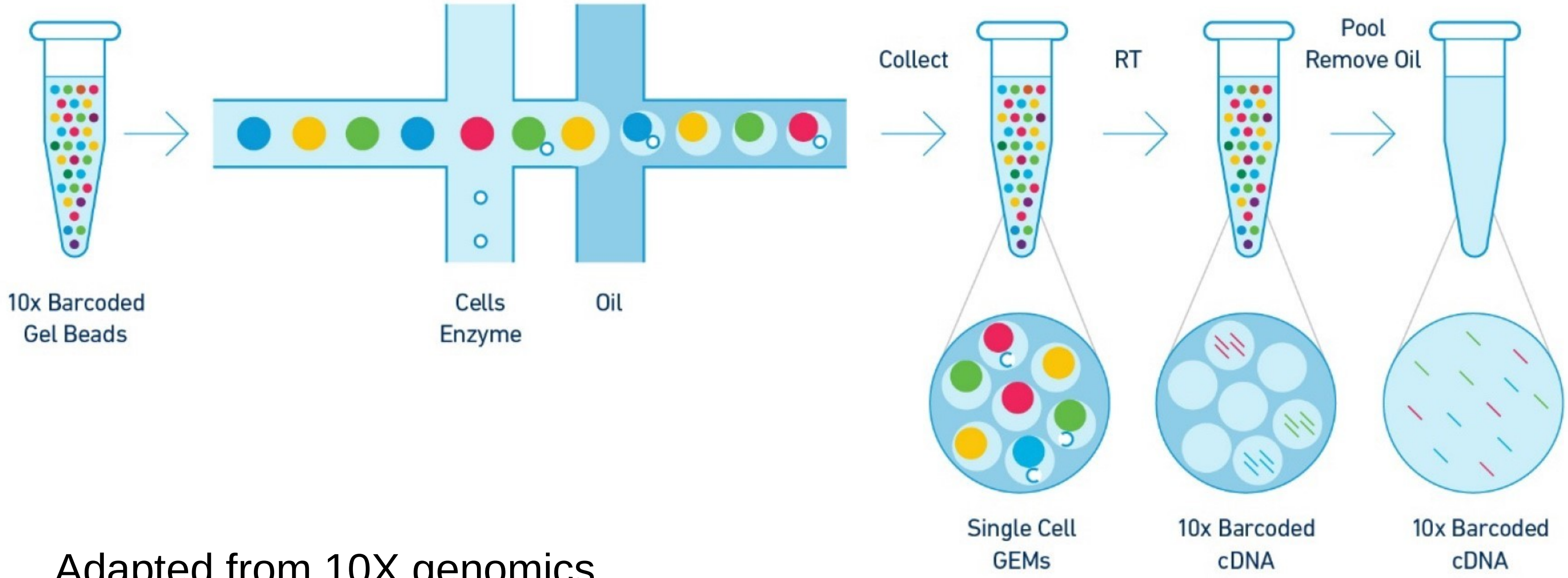
Gene Space



Single cell data analysis

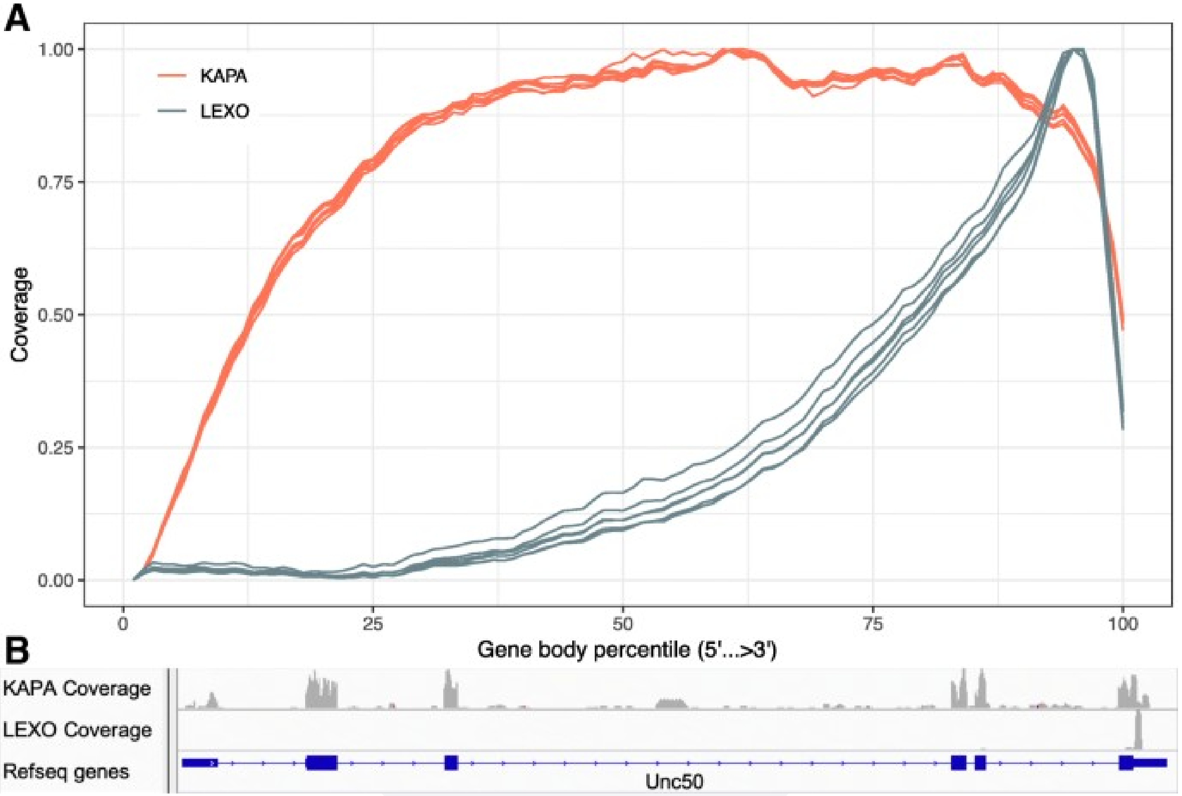
Library preparation
Indexing and alignment
Exploring the results
Data analysis

Library preparation

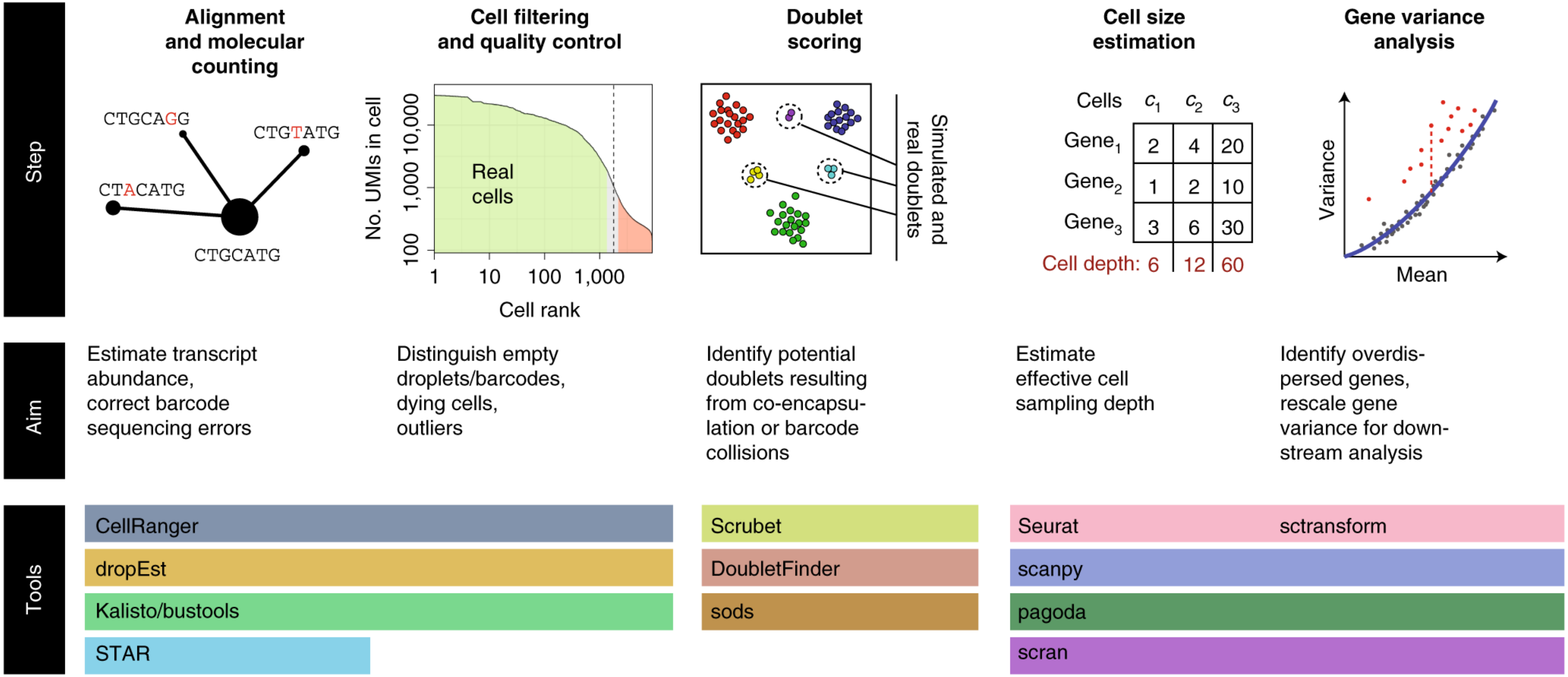


Adapted from 10X genomics

sc-RNAseq are enriched in the 3’ end



Ma et al. (2019). BMC Genomics 20(1)



Cells	c_1	c_2	c_3
Gene_1	2	4	20
Gene_2	1	2	10
Gene_3	3	6	30
Cell depth:	6	12	60

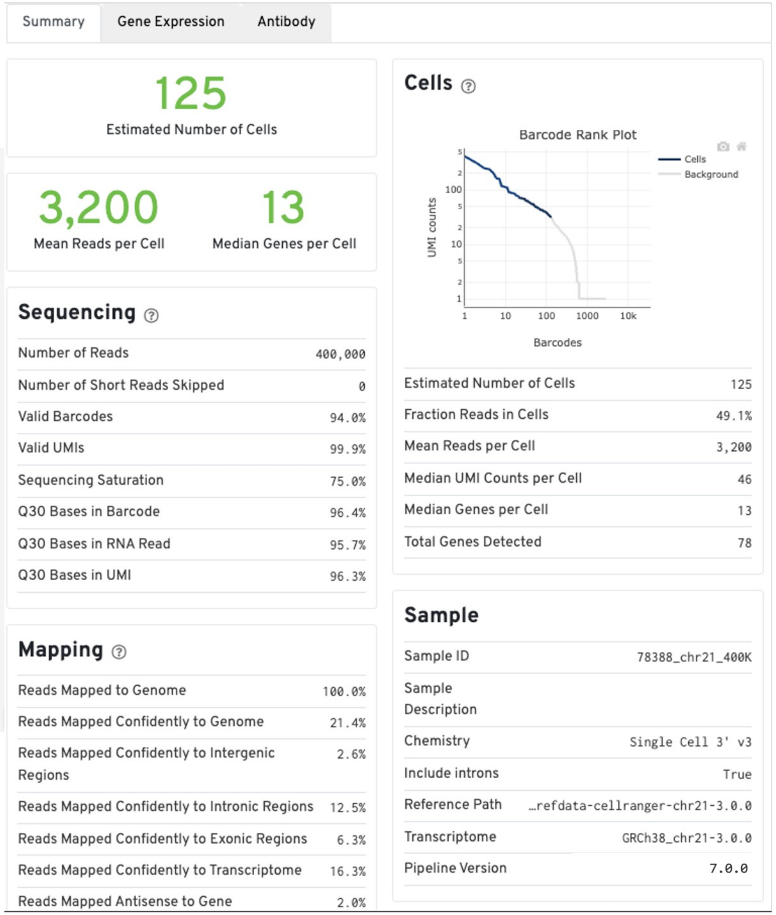
Step	Alignment and molecular counting	Cell filtering and quality control	Doublet scoring	Cell size estimation	Gene variance analysis
Aim	Estimate transcript abundance, correct barcode sequencing errors	Distinguish empty droplets/barcodes, dying cells, outliers	Identify potential doublets resulting from co-encapsulation or barcode collisions	Estimate effective cell sampling depth	Identify overdispersed genes, rescale gene variance for downstream analysis
Tools	CellRanger dropEst Kalisto/bustools STAR	CellRanger dropEst Kalisto/bustools	Scrubet DoubletFinder sods	Seurat scanpy pagoda scran	sctransform scanpy pagoda scran

Karchenko (2021) Nature Methods 18:723–732

Tenth Exercise

- Find and load cellranger software
- Index the reference genome
- Align reads to the reference genome
- Understand the results

Cellranger html reports



125
Estimated Number of Cells

3,200
Mean Reads per Cell

13
Median Genes per Cell

Sequencing

Number of Reads	400,000
Number of Short Reads Skipped	0
Valid Barcodes	94.0%
Valid UMIs	99.9%
Sequencing Saturation	75.0%
Q30 Bases in Barcode	96.4%
Q30 Bases in RNA Read	95.7%
Q30 Bases in UMI	96.3%

Mapping

Reads Mapped to Genome	100.0%
Reads Mapped Confidently to Genome	21.4%
Reads Mapped Confidently to Intergenic Regions	2.6%
Reads Mapped Confidently to Intronic Regions	12.5%
Reads Mapped Confidently to Exonic Regions	6.3%
Reads Mapped Confidently to Transcriptome	16.3%
Reads Mapped Antisense to Gene	2.0%

Cells

Estimated Number of Cells	125
Fraction Reads in Cells	49.1%
Mean Reads per Cell	3,200
Median UMI Counts per Cell	46
Median Genes per Cell	13
Total Genes Detected	78

Sample

Sample ID	78388_chr21_400K
Sample Description	
Chemistry	Single Cell 3' v3
Include introns	True
Reference Path	…refdata-cellranger-chr21-3.0.0
Transcriptome	GRCh38_chr21-3.0.0
Pipeline Version	7.0.0

End of the second week

Thank you for your attention