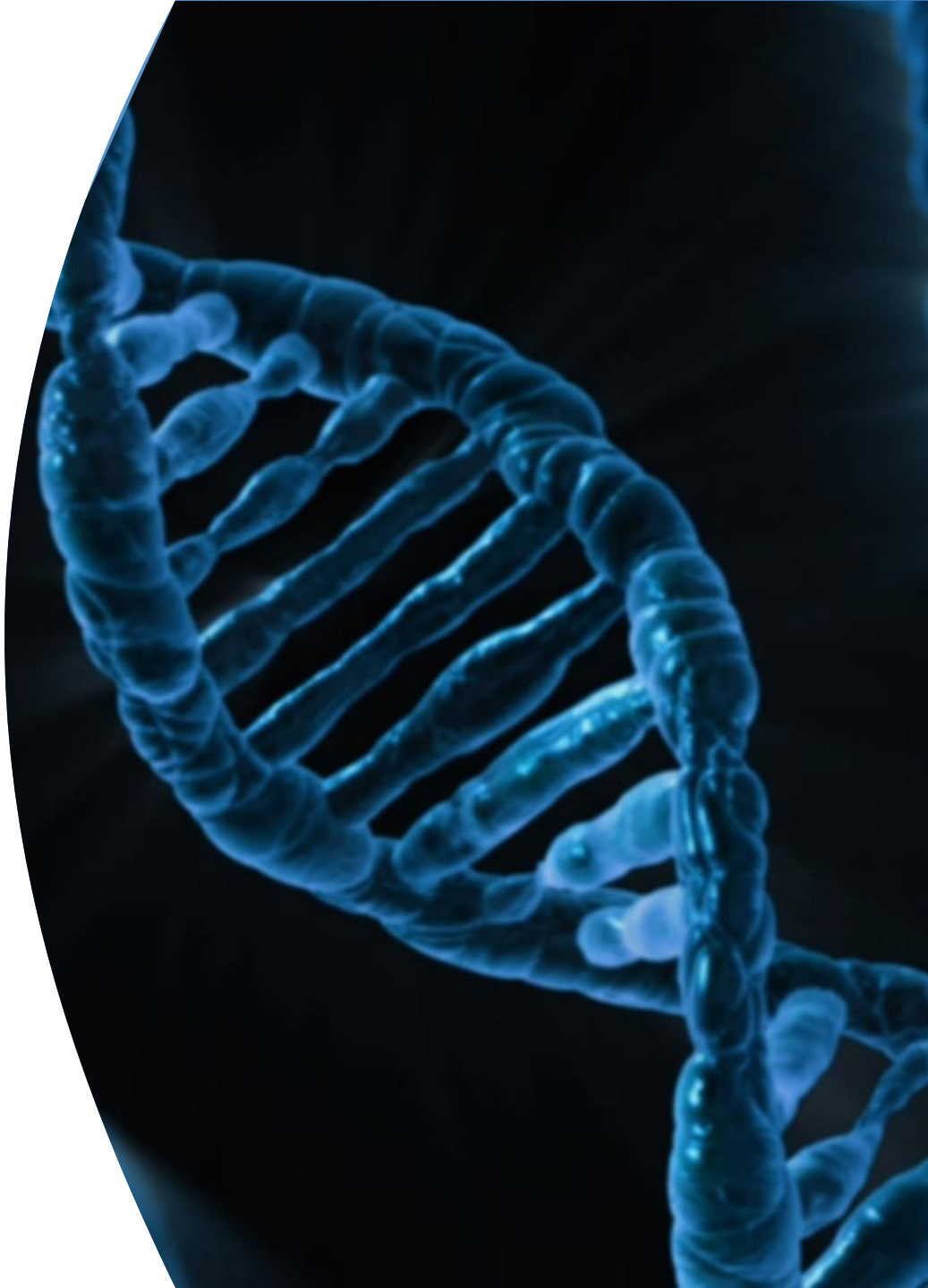
Introduction to Bioinformatics

Cambridge Makerere Summer School 2024

Ashley D Sawle



CANCER RESEARCH UK | Cambridge Institute

Together we will beat cancer

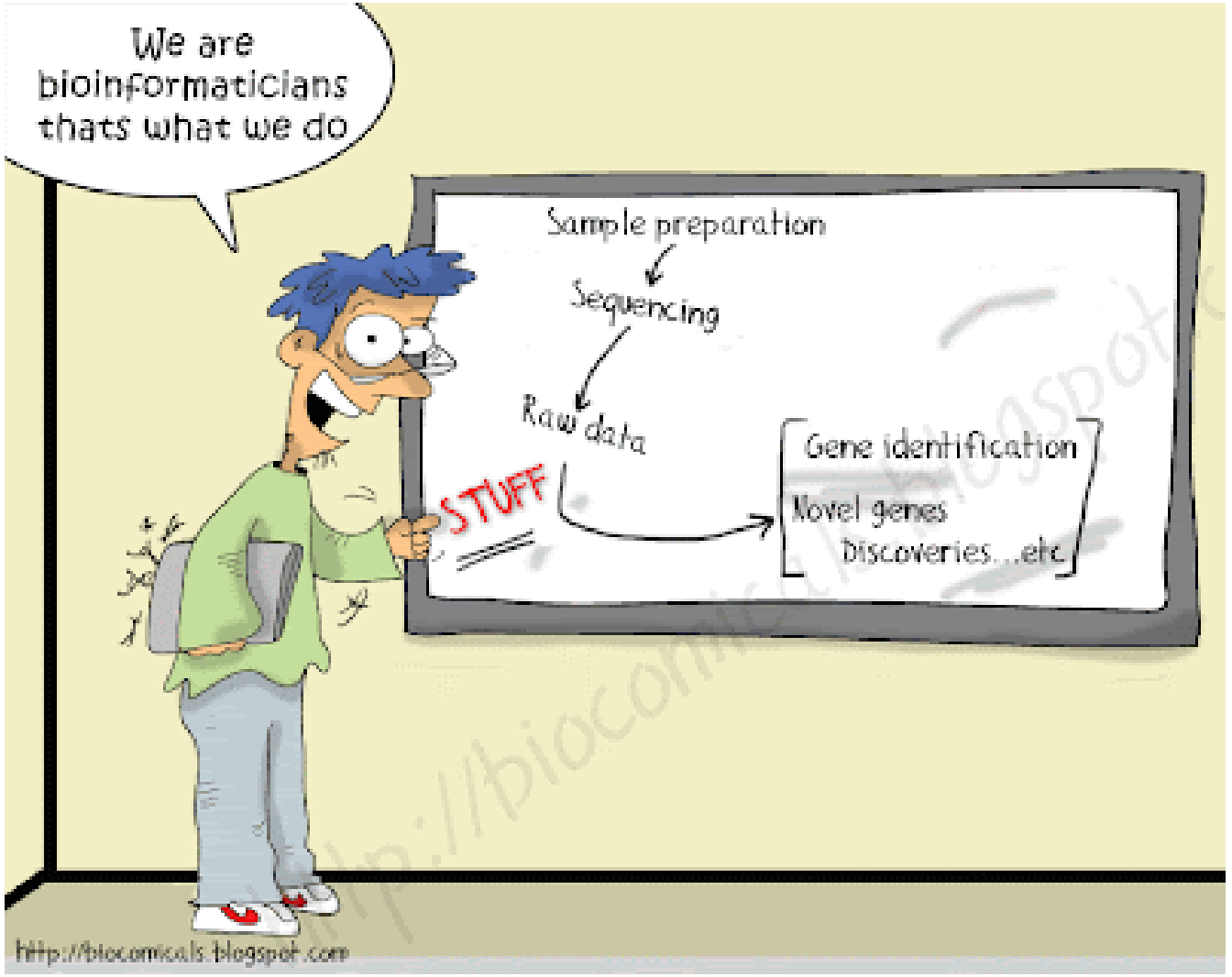
Overview

- A Brief Overview of Bioinformatics
- Bioinformatic Analysis of Next Generation Sequencing Data

Overview

- **A Brief Overview of Bioinformatics**
- Bioinformatic Analysis of Next Generation Sequencing Data

What is Bioinformatics?



What is Bioinformatics?

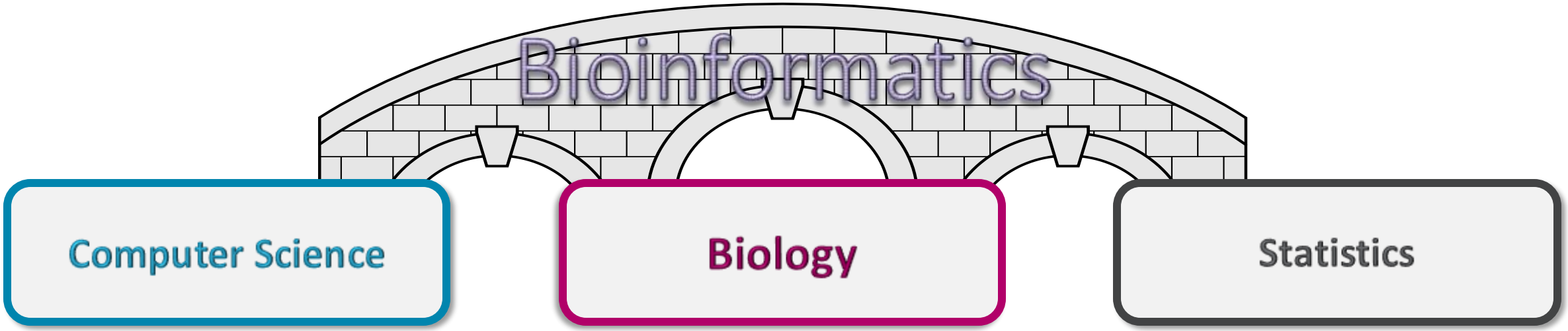
- Bioinformatics is a relatively new and evolving discipline that combines skills and technologies from computer science and biology to help us better understand and interpret biological data.
- Bioinformatics, as related to genetics and genomics, is a scientific subdiscipline that involves using computer technology to collect, store, analyze and disseminate biological data and information
- The mathematical, statistical and computing methods that aim to solve biological problems using DNA and amino acid sequences and related information.

What is Bioinformatics?

- Bioinformatics is a relatively **new and evolving** discipline that combines skills and technologies from computer science and biology to help us better understand and interpret biological data.
- Bioinformatics, as related to genetics and genomics, is a scientific subdiscipline that involves **using computer technology to collect, store, analyze and disseminate biological data and information**
- The mathematical, statistical and computing methods that aim **to solve biological problems using DNA and amino acid sequences** and related information.

What is Bioinformatics?

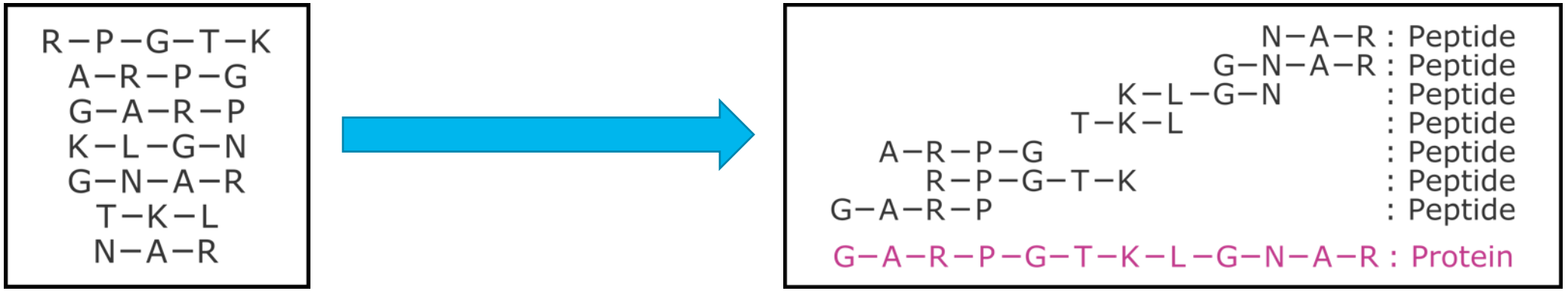
Bioinformatics is an **interdisciplinary** field of science that develops methods and software tools for understanding biological data, especially when the data sets are large and complex.



It all started with proteins

- 1951 – Frederick Sanger sequenced the amino acid structure of insulin

It all started with proteins



It all started with proteins …

- 1951 – Frederick Sanger sequenced the amino acid structure of insulin
- 1962 – Margaret Dayhoff and Robert Ledley publish COMPROTEIN
- 1965 – Margaret Dayhoff published the book “Atlas of Protein Sequence and Structure”
- 1970 – Needleman-Wunsch algorithm for sequence alignment published
- 1970s - Peter Chou and Gerald Fasman develop first protein structure prediction algorithm
- 1971 – The Protein Data Bank

… and it continues with proteins

The Nobel Prize in Chemistry 2024

David Baker

"for computational protein design"

David Baker. Ill. Niklas Elmehed © Nobel Prize Outreach

Demis Hassabis

"for protein structure prediction"

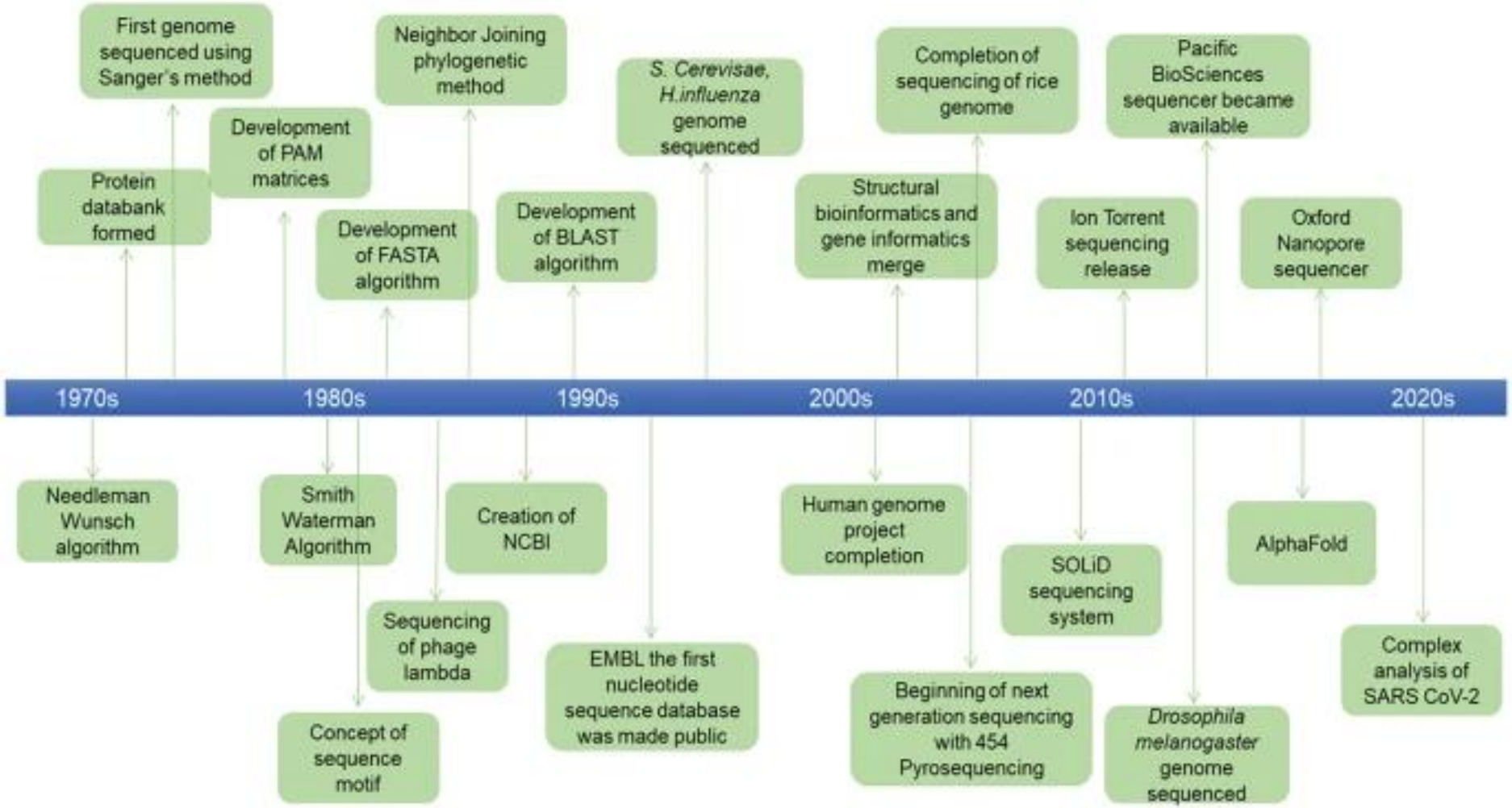
Demis Hassabis. Ill. Niklas Elmehed © Nobel Prize Outreach

John Jumper

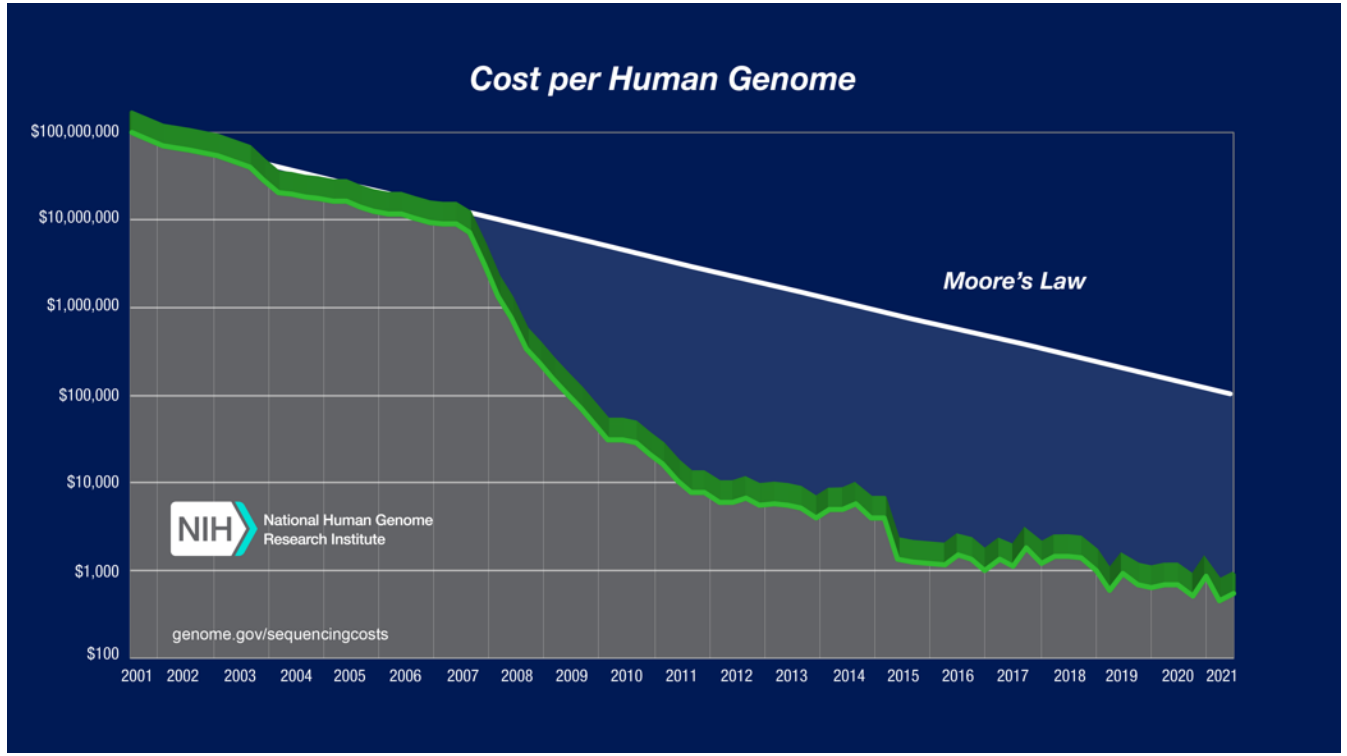
"for protein structure prediction"

John Jumper. Ill. Niklas Elmehed © Nobel Prize Outreach

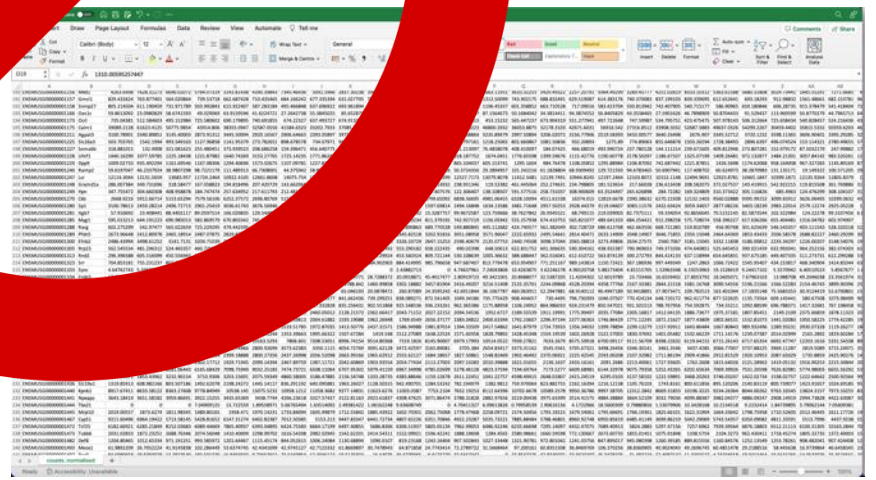
A history of nucleotide sequencing



The era of big data



The era of big data



Data storage and Access

- Verification of results presented in papers requires access to the data used to generate them
- Data generated in one study can still be useful to other
- Open Data rather than siloed inaccessible data

Open Access | Published: 15 March 2016

The FAIR Guiding Principles for scientific data management and stewardship

Mark D. Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie

Scientific Data **3**, Article number: 160018 (2016) | Cite this article

Findable **A**ccessible **I**nteroperable **R**eusable

Data Repositories

ENA European Nucleotide Archive

EUROPEAN GENOME-PHENOME ARCHIVE

UCSC

RefSeqGene

swissprot UniProt

SRA

GEO Gene Expression Omnibus

GENCODE

e!Ensembl

ENCODE

PRIDE Archive PRoteomics IDEntifications Database

gnomAD Genome Aggregation Database

reactome

KEGG Kyoto Encyclopedia of Genes and Genomes

MSigDB Molecular Signatures Database

GENEONTOLOGY Unifying Biology

FlyBase

WormBase

MGI

TCGA

ICGC

COSMIC Catalogue Of Somatic Mutations In Cancer

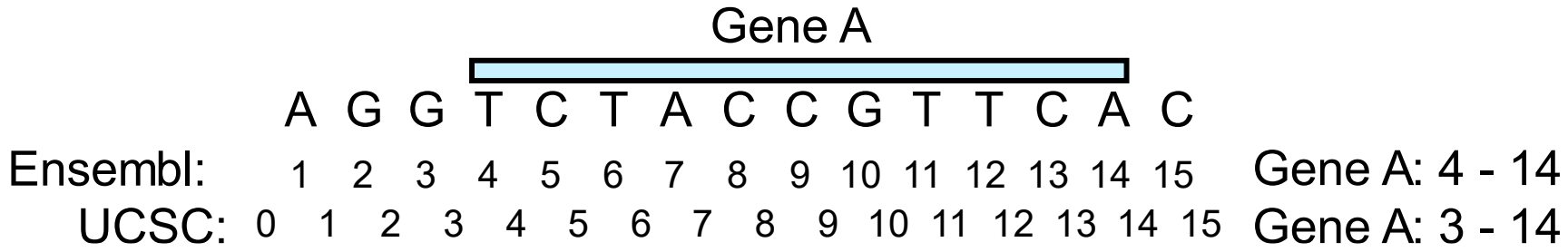
cBioPortal FOR CANCER GENOMICS

A word on gene names

Genomic Sequence Data

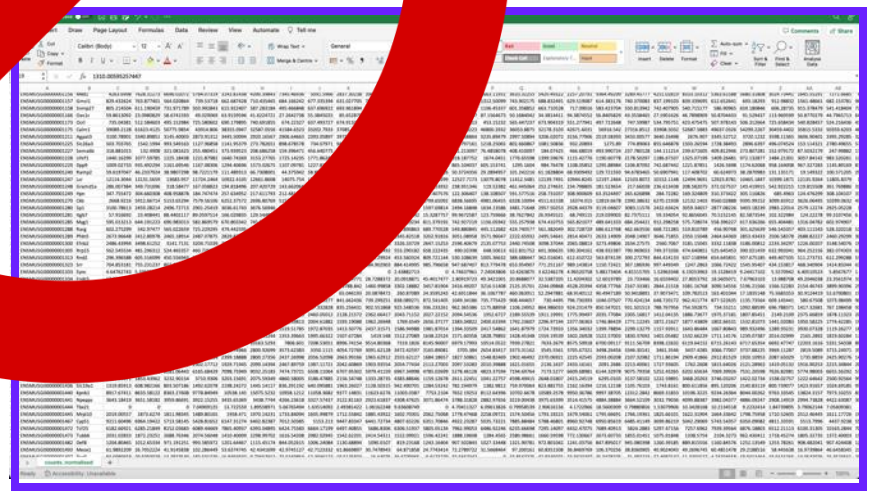
UCSC/NCBI versus Ensembl/Gencode/EBI

- Ensembl uses a one-based coordinate system, whereas UCSC uses a zero-based coordinate system.



- Ensembl/Gencode name sequences : 1, 2, 3 … 22, X, Y, MT
- UCSC/NCBI name sequences: chr1, chr2, chr3 … chr22, chrY, chrX, chrM
- Gene annotations differ significantly
- Gene IDs are different and do not map 1:1

The era of big data



Programming languages in Bioinformatics

- Bioinformatics creates huge quantities of data, and programming gives the means to analyse and interpret that data.
- The two most popular languages are **Python** and **R**
- Both are open source meaning they are freely available
- Both have large communities of users and developers
- Both have a wide range of bioinformatics resources and methods
- Both are cross-platform
- Others: Perl, Java, Ruby, Rust, Julia

- R is a language and environment for **statistical computing and graphics**
- R is available as Free Software under the terms of the Free Software Foundation's **GNU General Public License**
- **RStudio** provides a well developed integrated development environment (IDE) for R
- The Comprehensive R Archive Network (**CRAN**) repository features 19877 available general usage packages
- The **Bioconductor** project has 2230 bioinformatics specific packages

python™

- Python is a **general-purpose**, open-source programming language used in various software domains, including data science, web development, and gaming
- Python is developed under an **OSI-approved open source license**, making it freely available
- Various IDEs are available, e.g. **Jupyter** or Spyder
- 100,000s of packages available via the **Python Package Index**
- Virtual environments easily managed with **Conda**
- Python is more suitable for **deep learning** applications

BIOCONDA®

Excel tries to be helpful

Comment | Open Access | Published: 23 August 2016

Gene name errors are widespread in the scientific literature

Mark Ziemann, Yotam Eren & Assam El-Osta

Genome Biology **17**, Article number: 177 (2016) | Cite this article

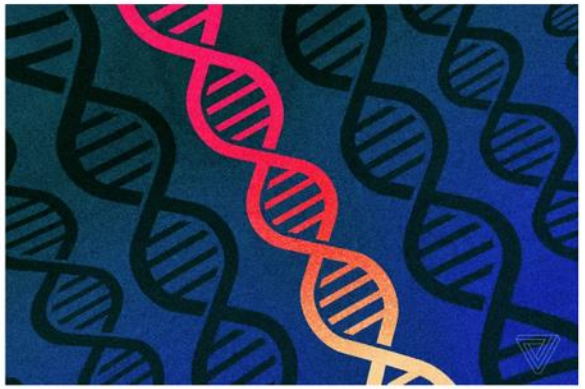
NEWS | 13 August 2021 | Correction 25 August 2021

Autocorrect errors in Excel still creating genomics headache

Despite geneticists being warned about spreadsheet problems, 30% of published papers contain mangled gene names in supplementary data.

The Verge CIENCE / TECH / MICROSOFT

Scientists rename human genes to stop Microsoft Excel from misreading them as dates



/ Sometimes it's easier to rewrite genetics than update Excel

Illustration by Alex Castro / The Verge

By James Vincent, a senior reporter who has covered AI, robotics, and more for eight years at The Verge.

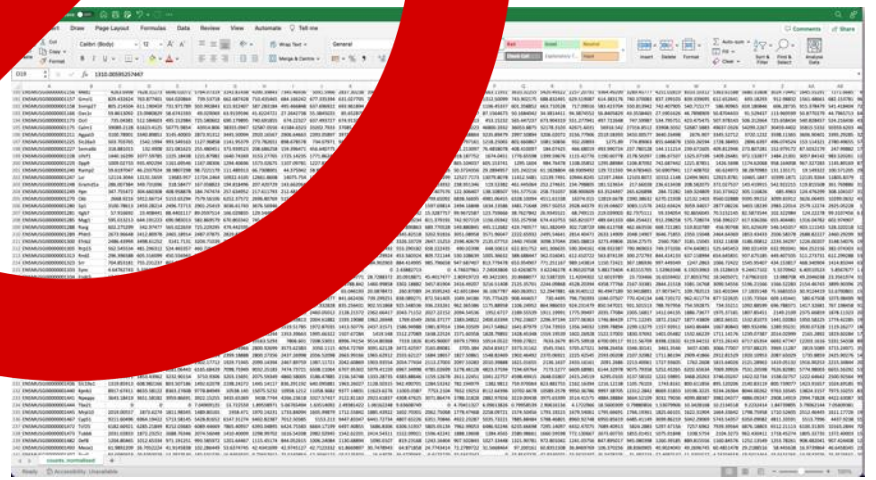
Aug 6, 2020, 1:44 PM GMT+1 | 0 Comments / 0 New

If you buy something from a Verge link, Vox Media may earn a commission. See our ethics statement.

There are tens of thousands of genes in the human genome: minuscule twists of DNA and RNA that combine to express all of the traits and characteristics that make each of us unique. Each gene is given a name

MARCH1 → MARCHF1
SEPT1 → SEPTIN1

The era of big data

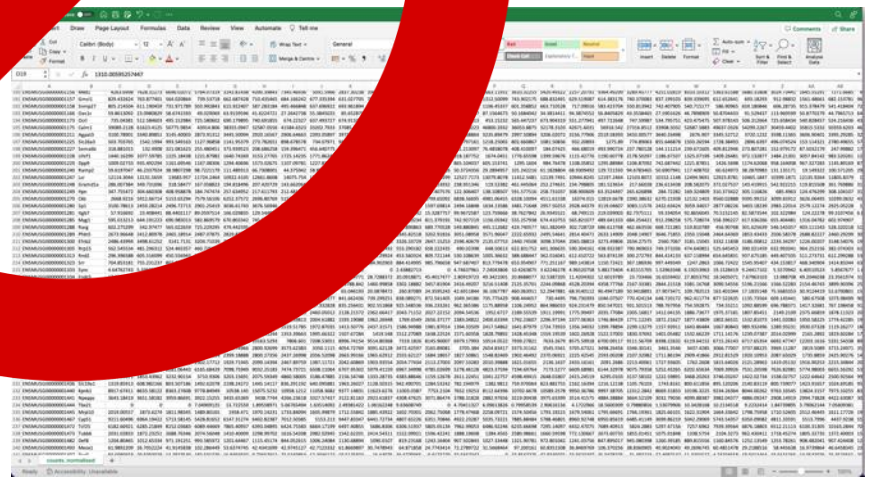


The need for more power

- Massive data sets require a lot memory to store and process
- Complex algorithms such as alignment need powerful processors to run in a reasonable time frame
- Having a lot of processors available means that many jobs can be run in parallel

- High-Performance Cluster – HPC (supercomputer)
- Cloud solution – Google Cloud, Amazon Web Services (AWS), Microsoft Azure
- Graphics Processing Units required for Deep Learning

The era of big data



Differing statistical approaches

- Single/Few measures in a simple experimental design – t-test
- More complex studies – Linear Model
  - Micro-arrays – Simple Linear Model with Normal Distribution
  - RNAseq data – Generalised Linear Model with Negative Binomial Distribution
- 10X Single Cell RNAseq – needs additional solutions to overcome missing data

Complex Data Requires New Solutions

- Sequence alignment algorithms
- Clustering algorithms
- Hidden Markov Models (HMMs)
- Phylogenetic tree construction algorithms
- Molecular modelling algorithms
- Variant/Mutation calling algorithms
- Deep learning, large language models and machine learning

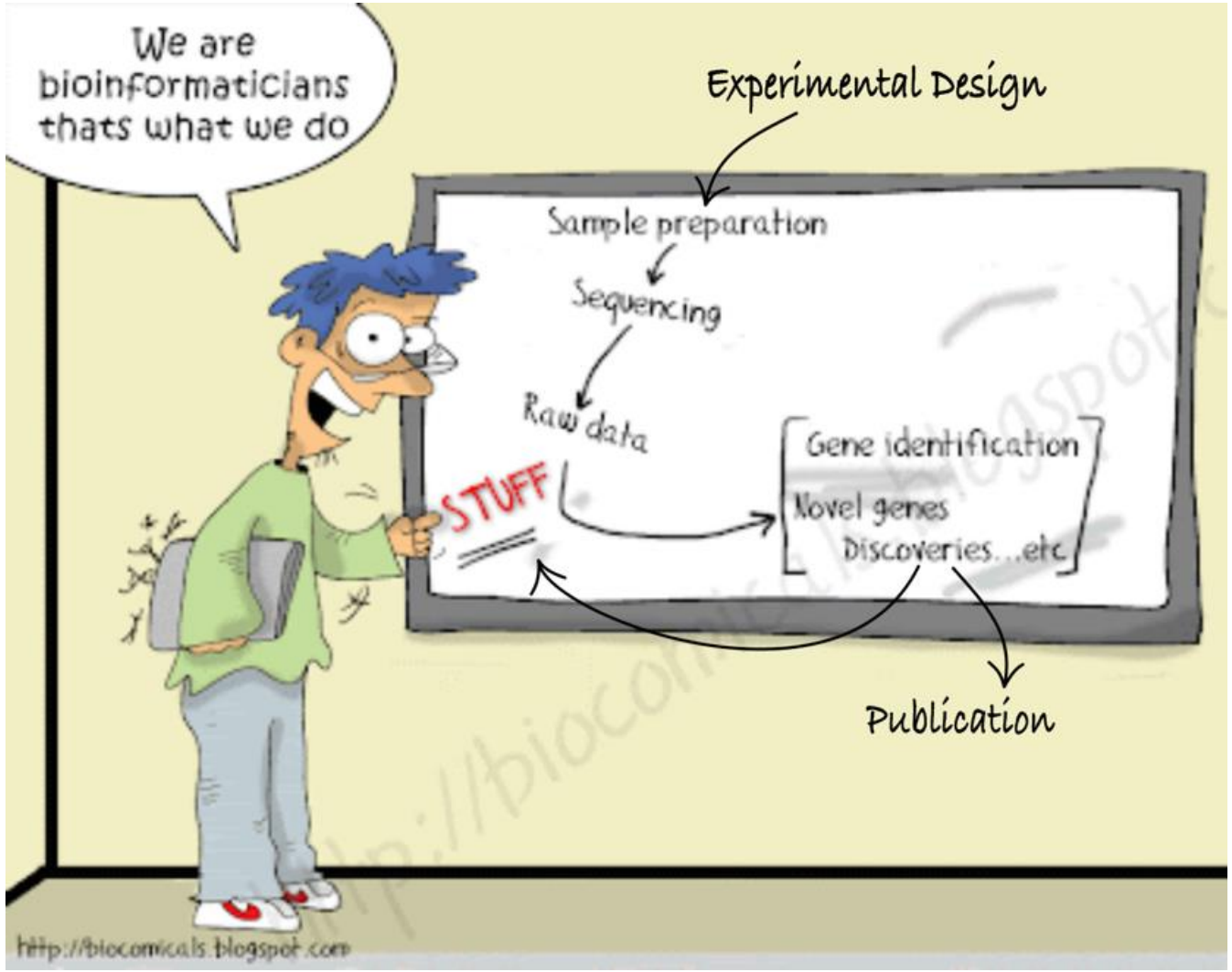
Overview

- A Brief Overview of Bioinformatics
- **Bioinformatic Analysis of Next Generation Sequencing Data**

Data literacy

The ability to not only carry out statistical analysis on real-world problems, but also to understand and critique any conclusions drawn by others on the basis of statistics.

What we do…


We are bioinformaticians thats what we do
Experimental Design
Sample preparation
Sequencing
Raw data
STUFF
Gene identification
Novel genes
Discoveries...etc
Publication
http://biocomicals.blogspot.com

The PPDAC cycle

- Understanding and defining the problem.
- How do we go about answering this question?

PROBLEM

- What to measure and how?
- Study design ?

PLAN

DATA

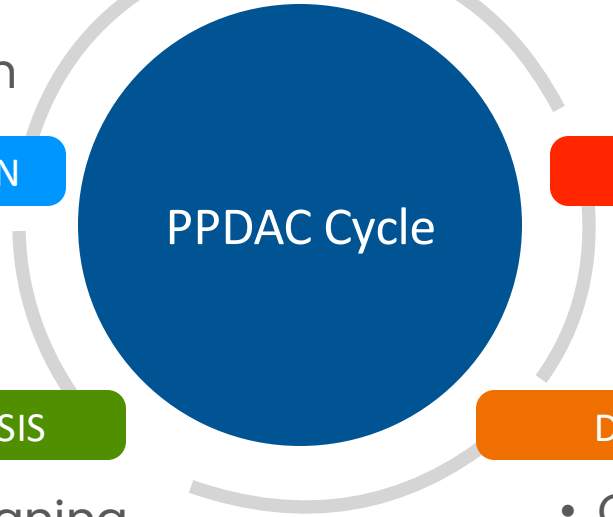
- Collection
- Management

ANALYSIS

- Data Cleaning
- Statistical Analysis
- Construct table, graphs
- Look for patterns

CONCLUSION

- Interpretation
- Conclusions
- Communication
- Hypothesis generation



Consequences of Poor Experimental Design

Inability to answer the questions we would like to answer

- Cost of experimentation.
- Limited & Precious material, esp. clinical samples.
- Immortalization of data sets in public databases and methods in the literature. Our bad science begets more bad science.
- Ethical concerns of experimentation: animals and clinical samples.

A Well-Designed Experiment

Should have:

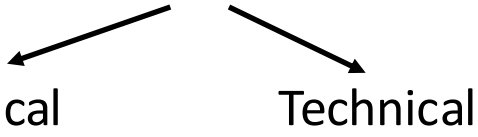
- Clear objectives
- Focus and simplicity
- Sufficient power
- Randomised comparisons

And be:

- Precise
- Unbiased
- **Amenable to statistical analysis**
- Reproducible

Sources of Variation

dependent variable = f (independent variable) + noise



Biological "noise":

- Biological processes are inherently stochastic
- Single cells, cell populations, individuals, organs, species….
- Timepoints, cell cycle, synchronized vs. unsynchronized

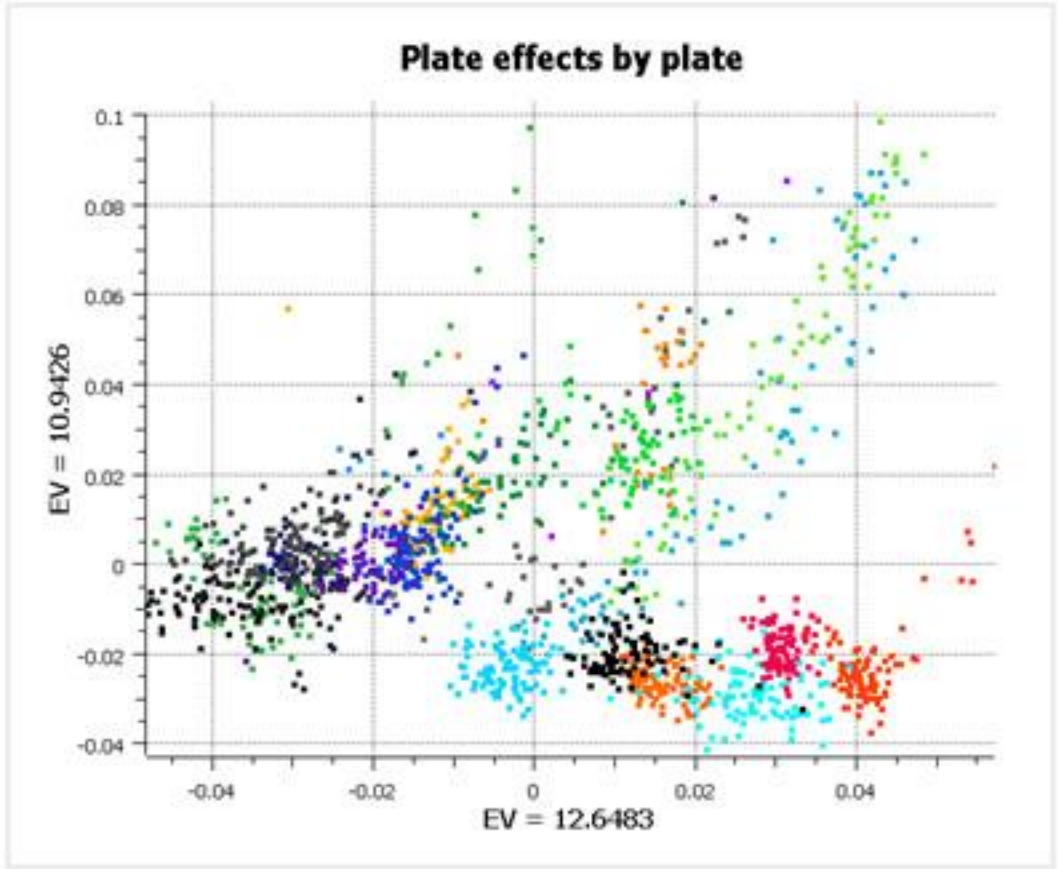
Technical noise:

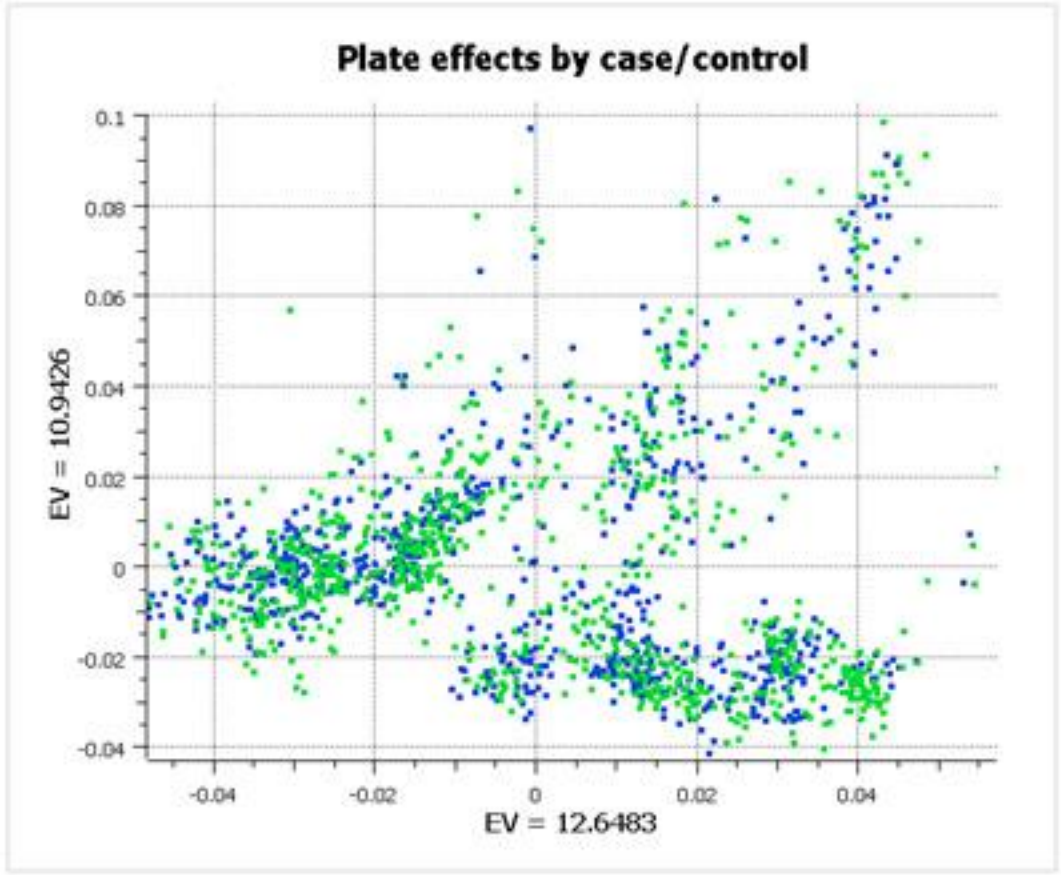
- Reagents, antibodies, temperatures, pollution
- Platforms, runs, operators

Replication is required to capture variance

Randomisation overcomes technical variation

Randomisation





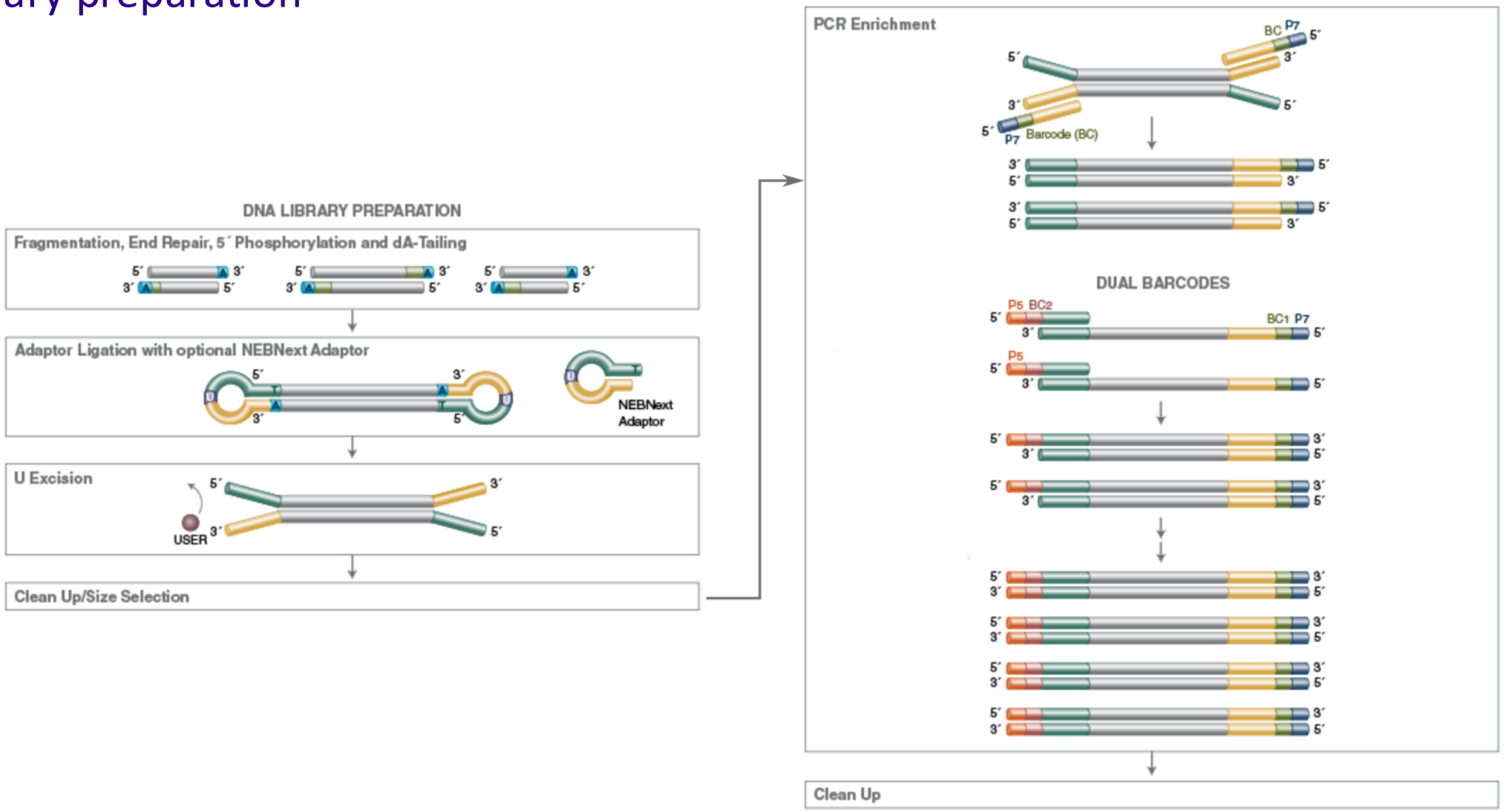
https://www.goldenhelix.com/blog/stop-ignoring-experimental-design-or-my-head-will-explode/

Next Generation Sequencing

- We'll focus on Illumina short read sequencing as this is the most commonly used method at the moment

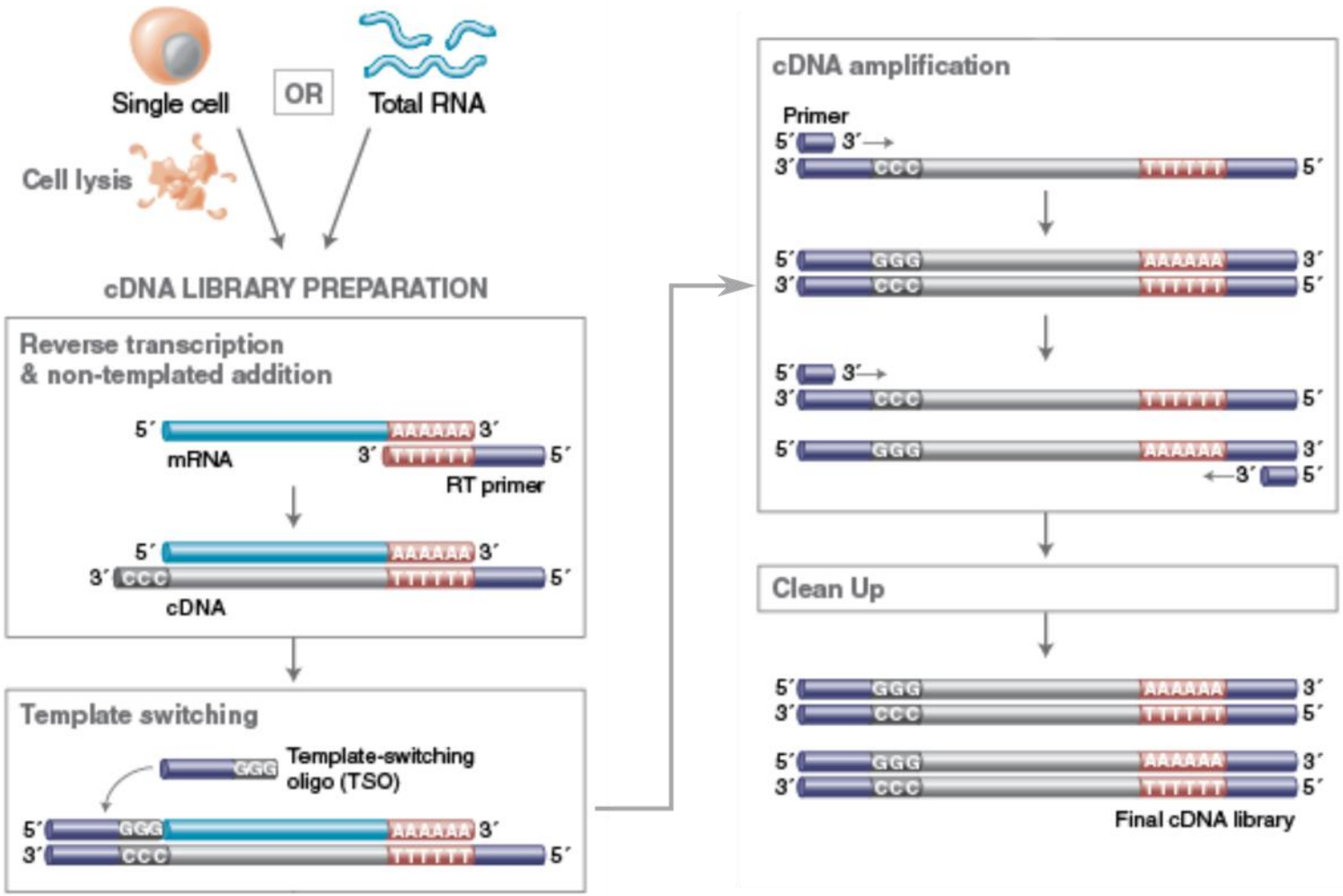
Next Generation Sequencing

Library preparation



Next Generation Sequencing

Library preparation



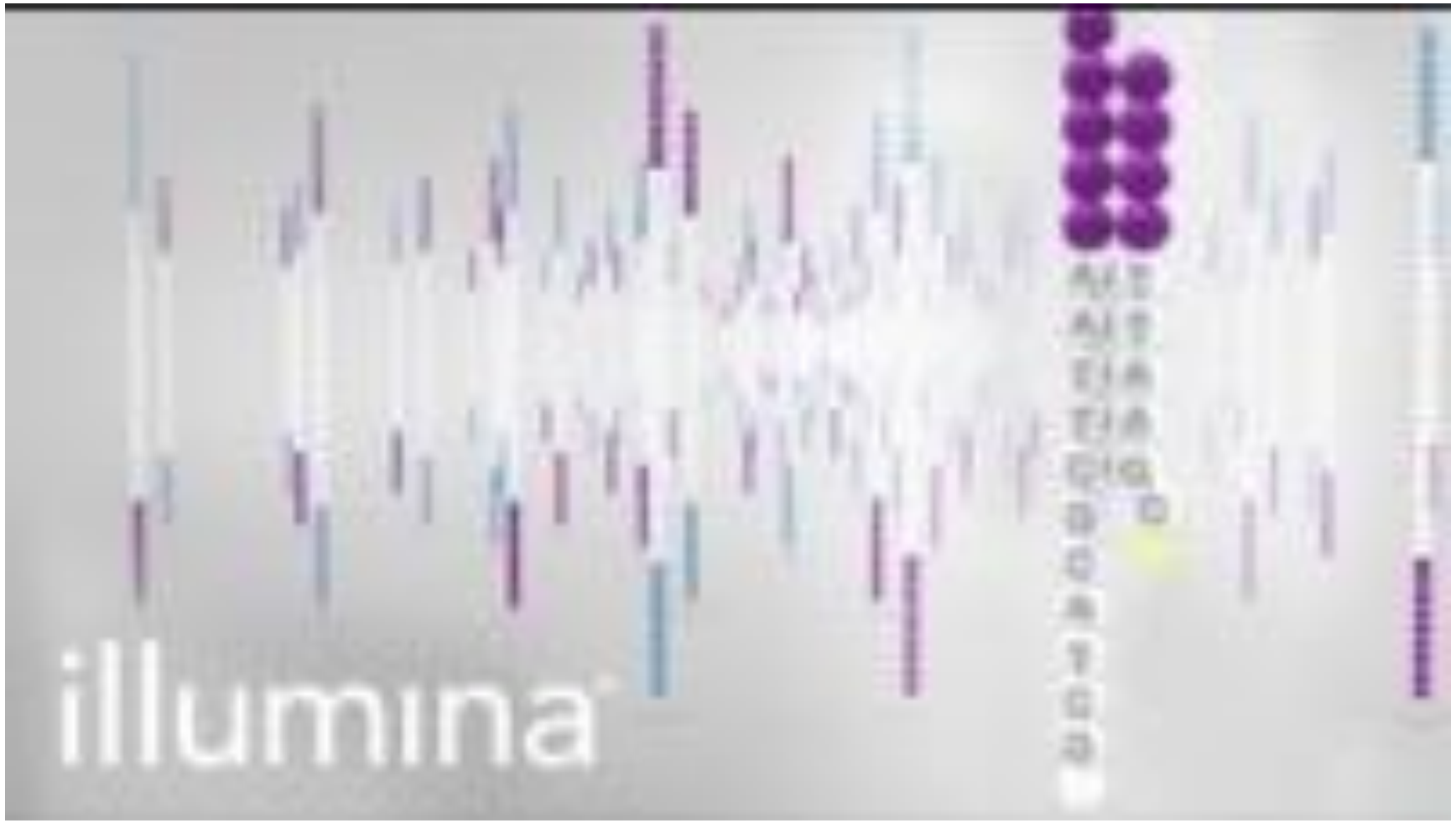
Next Generation Sequencing

Sequencing by synthesis

- A complimentary strand is synthesized using the cDNA fragment as template.
- Each nucleotide includes a fluorescent tag and as the new strand is synthesized, the colour of the fluorescence indicates which base is being added.
- The sequencer records the order of these flashes of light and translates them to a base sequence.

Next Generation Sequencing

Sequencing by synthesis



Fastq file format

```
@LH00417:64:22MCGJLT3:3:1101:12206:1064 1:N:0:GATCAAGGCA+ATTAACAAGG
CNTTGACATTGTACTCGGCAGAAGATGTGCGGGCCAGGCTGCTCTGCCACTGGGCCTCAGGGATAGAGGTGCTGAGCCATCGAGCCGAGCGCGGCCGTGG
+
9#IIIII9II9I-9IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
@LH00417:64:22MCGJLT3:3:1101:12410:1064 1:N:0:GATCAAGGCA+ATTAACAAGG
TNCCAAACAACCAAACAAAAATCCCCAGCATTTTTCAGTGCATGAATATCTGAAAATCCTTCCTTGTGCAAGTATATTTCTTTGAAGAAGAAAAGGGGAT
+
9#9IIIIIII-IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
@LH00417:64:22MCGJLT3:3:1101:12798:1064 1:N:0:GATCAAGGCA+ATTAACAAGG
TNTGACGATGTTAGCGGTTACTCTTCATTGCTTCTTTAGCTGCCAGGATCAGTGGCGTCAAGCTAGACAATGGTTTGGCCAGTTTCAGGAAAGGATGCTG
+
9#IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
@LH00417:64:22MCGJLT3:3:1101:13963:1064 1:N:0:GATCAAGGCA+ATTAACAAGG
CNTTCCTTTCAGCCAGTGCCTGCACAGATGAAGCTGACATCTGATCCTTCATGTCCCTGACAATTTGCCGAGTACTTGTGATGAAGGCTTTTCTTATGCT
+
9#IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII9I
@LH00417:64:22MCGJLT3:3:1101:14000:1064 1:N:0:GATCAAGGCA+ATTAACAAGG
GNAGCTTCTCTTTGATGGTCCGGTGTGGGTCGATCTCAATGGGATAATAGTGGTGAAAGAGCTCGGTGAGCTCTTTGCGACAATCCTCACTGATAATTTT
+
9#9I9IIII9IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII-IIIIIIIIIIIIIIIIIIIIIII9IIIIIIIIII9
@LH00417:64:22MCGJLT3:3:1101:14740:1064 1:N:0:GATCAAGGCA+ATTAACAAGG
ANTCTTCTCACTGGTGCCAACAGAAAGTCTGTACTCGATGTCTGCCATTTTGGTCAATAAGTGTATCCGAACTGAAGATGGAAAGTCAACTCTGTGTACA
+
I#IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
@LH00417:64:22MCGJLT3:3:1101:15295:1064 1:N:0:GATCAAGGCA+ATTAACAAGG
TNAGGTCTCTCACCTCCGGGGCCGTAGCGGACGCCTTAGCAGCGGCATTCTGGGACAATCCCCGGCCGCGGGGCACCAGCTTCCAATACCTAACTGTCTC
+
9#IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
@LH00417:64:22MCGJLT3:3:1101:16312:1064 1:N:0:GATCAAGGCA+ATTAACAAGG
CNTGGCCTGCACACTCAAAGAGCAGCTCTCCATCACCGAGCCTCCACACGAAGGCCTTGTCGTCTTCGCCCCCTGTCACTGCCAAGGTGTTGGTTTTGGG
+
I#IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
@LH00417:64:22MCGJLT3:3:1101:16404:1064 1:N:0:GATCAAGGCA+ATTAACAAGG
CNAACTATGAACTGAGCTCTGCCCATAGCTTCCAGAGAGCAGAGGAAATGGTTCTAAGCTAAACACCCACACTACGTGGTGGCAATGAGCCCGTTATCAA
+
-#I9IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
```

Fastq file format - Headers

```
@LH00417:64:22MCGJLT3:3:1101:12206:1064 1:N:0:GATCAAGGCA+ATTAACAAGG
CNTTGACATTGTACTCGGCAGAAGATGTGCGGGCCAGGCTGCTCTGCCACTGGGCCTCAGGGATAGAGGTGCTGAC
+
9#IIIII9II9I-9IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
@LH00417:64:22MCGJLT3:3:1101:12410:1064 1:N:0:GATCAAGGCA+ATTAACAAGG
TNCCAAACAACCAAACAAAAATCCCCAGCATTTTTCAGTGCATGAATATCTGAAAATCCTTCCTTGTGCAAGTATA
+
9#9IIIIIII-IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
@LH00417:64:22MCGJLT3:3:1101:12798:1064 1:N:0:GATCAAGGCA+ATTAACAAGG
TNTGACGATGTTAGCGGTTACTCTTCATTGCTTCTTTAGCTGCCAGGATCAGTGGCGTCAAGCTAGACAATGGTTT
+
9#IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
@LH00417:64:22MCGJLT3:3:1101:13963:1064 1:N:0:GATCAAGGCA+ATTAACAAGG
CNTTCCTTTCAGCCAGTGCCTGCACAGATGAAGCTGACATCTGATCCTTCATGTCCCTGACAATTTGCCGAGTACT
+
9#IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
```

Fastq file format - Sequence

```
@LH00417:64:22MCGJLT3:3:1101:12206:1064 1:N:0:GATCAAGGCA+ATTAACAAGG
CNTTGACATTGTACTCGGCAGAAGATGTGCGGGCCAGGCTGCTCTGCCACTGGGCCTCAGGGATAGAGGTGCTGA
+
9#IIIII9II9I-9IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
@LH00417:64:22MCGJLT3:3:1101:12410:1064 1:N:0:GATCAAGGCA+ATTAACAAGG
TNCCAAACAACCAAACAAAAATCCCCAGCATTTTTCAGTGCATGAATATCTGAAAATCCTTCCTTGTGCAAGTAT
+
9#9IIIIIII-IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
@LH00417:64:22MCGJLT3:3:1101:12798:1064 1:N:0:GATCAAGGCA+ATTAACAAGG
TNTGACGATGTTAGCGGTTACTCTTCATTGCTTCTTTAGCTGCCAGGATCAGTGGCGTCAAGCTAGACAATGGTT
+
9#IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
@LH00417:64:22MCGJLT3:3:1101:13963:1064 1:N:0:GATCAAGGCA+ATTAACAAGG
CNTTCCTTTCAGCCAGTGCCTGCACAGATGAAGCTGACATCTGATCCTTCATGTCCCTGACAATTTGCCGAGTAC
+
9#IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
```

Fastq file format – Line 3

```
@LH00417:64:22MCGJLT3:3:1101:12206:1064 1:N:0:GATCAAGGCA+ATTAACAAGG
CNTTGACATTGTACTCGGCAGAAGATGTGCGGGCCAGGCTGCTCTGCCACTGGGCCTCAGGGATAGAGGTGCTGAC
+
9#IIIII9II9I-9IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
@LH00417:64:22MCGJLT3:3:1101:12410:1064 1:N:0:GATCAAGGCA+ATTAACAAGG
TNCCAAACAACCAAACAAAAATCCCCAGCATTTTTCAGTGCATGAATATCTGAAAATCCTTCCTTGTGCAAGTATA
+
9#9IIIIIII-IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
@LH00417:64:22MCGJLT3:3:1101:12798:1064 1:N:0:GATCAAGGCA+ATTAACAAGG
TNTGACGATGTTAGCGGTTACTCTTCATTGCTTCTTTAGCTGCCAGGATCAGTGGCGTCAAGCTAGACAATGGTTT
+
9#IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
@LH00417:64:22MCGJLT3:3:1101:13963:1064 1:N:0:GATCAAGGCA+ATTAACAAGG
CNTTCCTTTCAGCCAGTGCCTGCACAGATGAAGCTGACATCTGATCCTTCATGTCCCTGACAATTTGCCGAGTACT
+
9#IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
```

Fastq file format - Headers

```
@LH00417:64:22MCGJLT3:3:1101:12206:1064 1:N:0:GATCAAGGCA+ATTAACAAGG
CNTTGACATTGTACTCGGCAGAAGATGTGCGGGCCAGGCTGCTCTGCCACTGGGCCTCAGGGATAGAGGTGCTGAC
+
9#IIIII9II9I-9IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
@LH00417:64:22MCGJLT3:3:1101:12410:1064 1:N:0:GATCAAGGCA+ATTAACAAGG
TNCCAAACAACCAAACAAAAATCCCCAGCATTTTTCAGTGCATGAATATCTGAAAATCCTTCCTTGTGCAAGTATA
+
9#9IIIIIII-IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
@LH00417:64:22MCGJLT3:3:1101:12798:1064 1:N:0:GATCAAGGCA+ATTAACAAGG
TNTGACGATGTTAGCGGTTACTCTTCATTGCTTCTTTAGCTGCCAGGATCAGTGGCGTCAAGCTAGACAATGGTTT
+
9#IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
@LH00417:64:22MCGJLT3:3:1101:13963:1064 1:N:0:GATCAAGGCA+ATTAACAAGG
CNTTCCTTTCAGCCAGTGCCTGCACAGATGAAGCTGACATCTGATCCTTCATGTCCCTGACAATTTGCCGAGTACT
+
9#IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
```

(Phred) Quality Scores

Sequence quality scores are transformed and translated p-values

Sequence bases are called after image processing (base calling):

- Each base in a sequence has a p-value associated with it
- p-values range from 0-1 (e.g.: 0.05, 0.01, 1e-30)
- p-value of 0.01 implies 1 in 100 chance that called base is wrong

(Phred) Quality Scores …

How do we assign p-values to bases in the fastq file?

- P-vales can be many characters long (e.g.:0.000005)
- Transform to Phred quality scores - Q = −10 x log10(pvalue) :

  If p = 0.01 → log10(0.01) = -2 → Q = 20
- Translate Q values to ASCII characters (adding 33):

  Q value of 2 = #, Q value of 40 = I
- This gives us a single digit quality score code for each base that fits nicely in the fastq format

Dec	Hex	Chr	Dec	Hex	Chr	Dec	Hex	Chr	Dec	Hex	Chr
0	00	NUL	32	20	Space	64	40	@	96	60	`
1	01	SOH	33	21	!	65	41	A	97	61	a
2	02	STX	34	22	"	66	42	B	98	62	b
3	03	ETX	35	23	#	67	43	C	99	63	c
4	04	EOT	36	24	$	68	44	D	100	64	d
5	05	ENQ	37	25	%	69	45	E	101	65	e
6	06	ACK	38	26	&	70	46	F	102	66	f
7	07	BEL	39	27	'	71	47	G	103	67	g
8	08	BS	40	28	(	72	48	H	104	68	h
9	09	HT	41	29	)	73	49	I	105	69	i
10	0A	LF	42	2A	*	74	4A	J	106	6A	j
11	0B	VT	43	2B	+	75	4B	K	107	6B	k
12	0C	FF	44	2C	,	76	4C	L	108	6C	l
13	0D	CR	45	2D	-	77	4D	M	109	6D	m
14	0E	SO	46	2E	.	78	4E	N	110	6E	n
15	0F	SI	47	2F	/	79	4F	O	111	6F	o
16	10	DLE	48	30	0	80	50	P	112	70	p
17	11	DC1	49	31	1	81	51	Q	113	71	q
18	12	DC2	50	32	2	82	52	R	114	72	r
19	13	DC3	51	33	3	83	53	S	115	73	s
20	14	DC4	52	34	4	84	54	T	116	74	t
21	15	NAK	53	35	5	85	55	U	117	75	u
22	16	SYN	54	36	6	86	56	V	118	76	v
23	17	ETB	55	37	7	87	57	W	119	77	w
24	18	CAN	56	38	8	88	58	X	120	78	x

QC is important

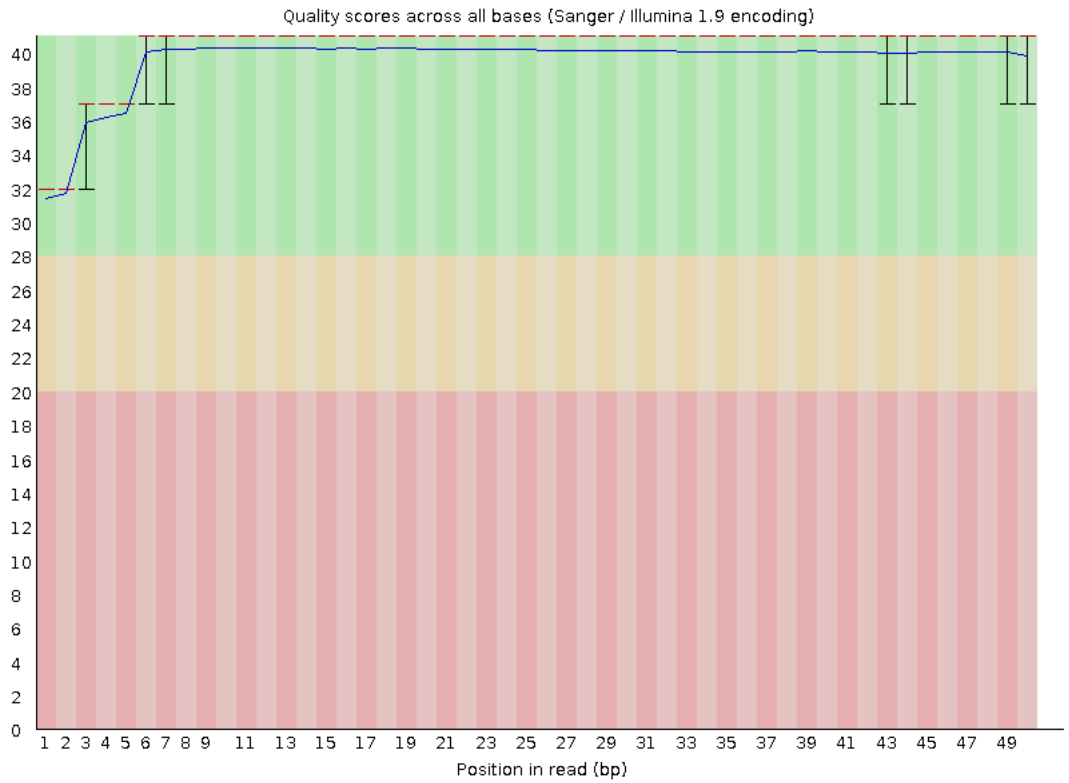
Check for any problems before we put time and effort into analysing potentially bad data

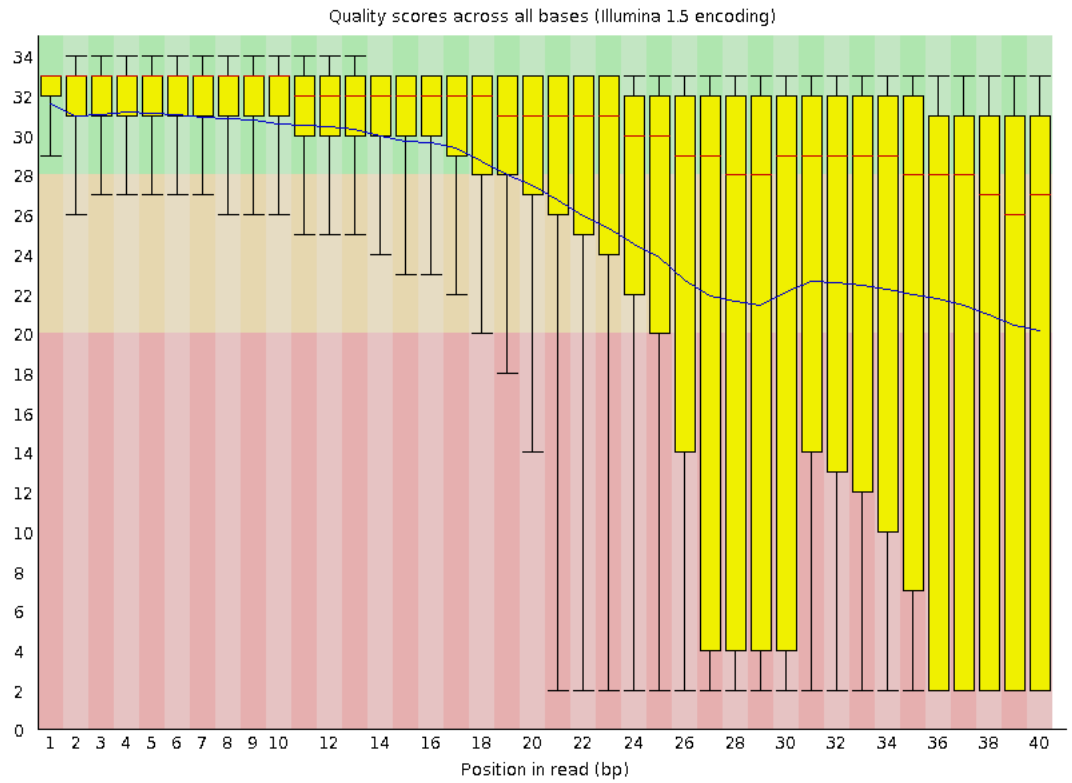
Start with FastQC:

- Quick
- Outputs an easy-to-read html report

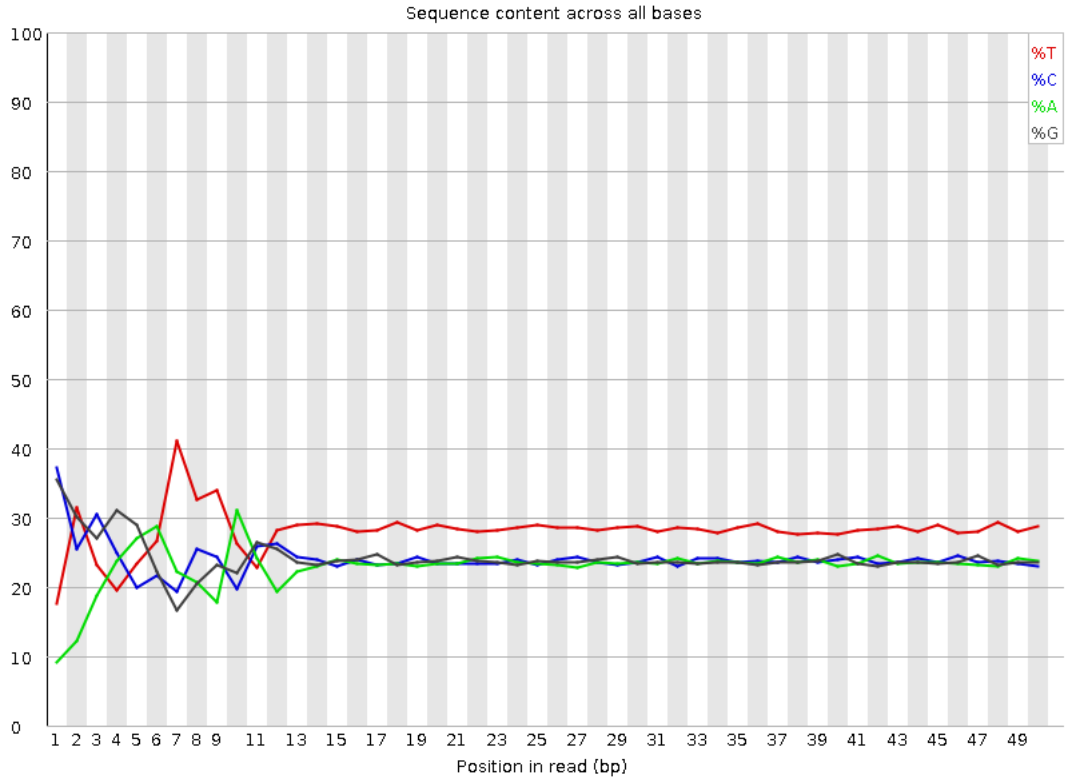
https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

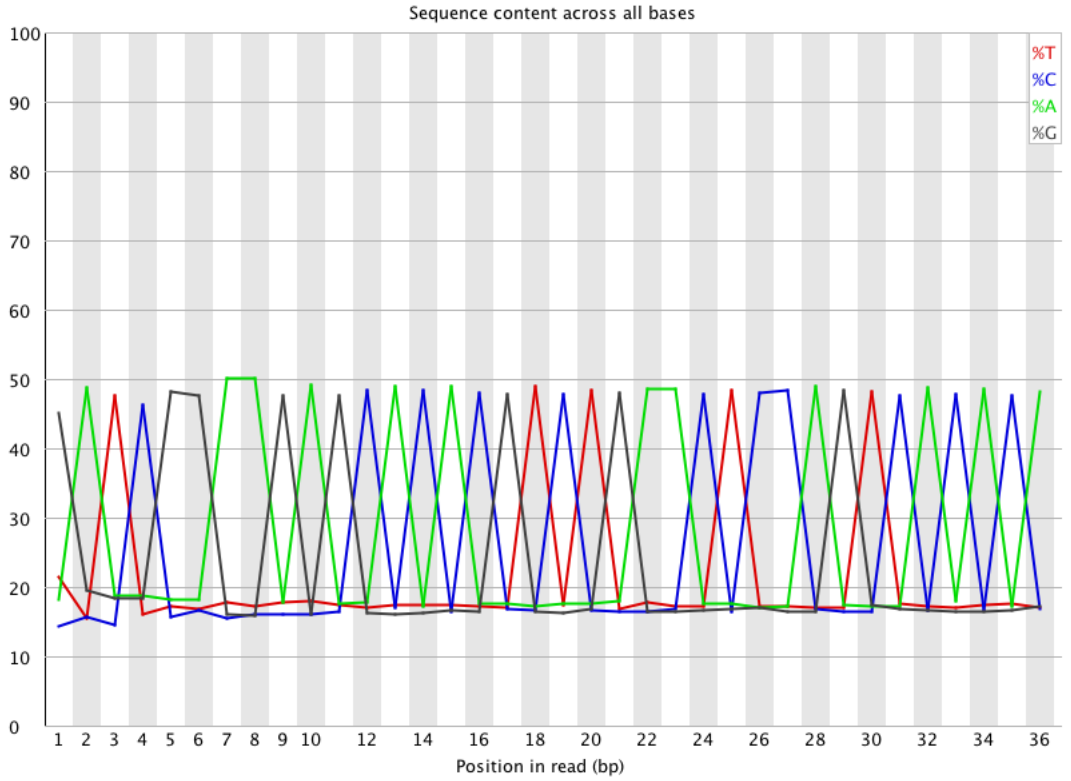
Per base sequence quality



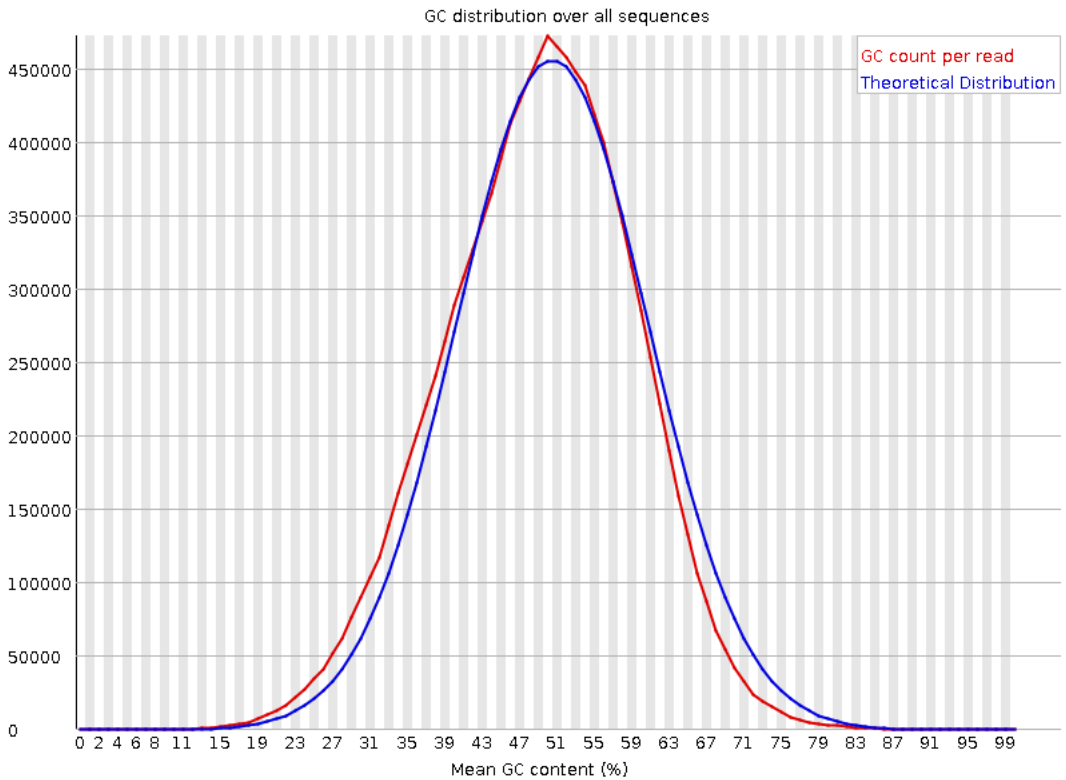


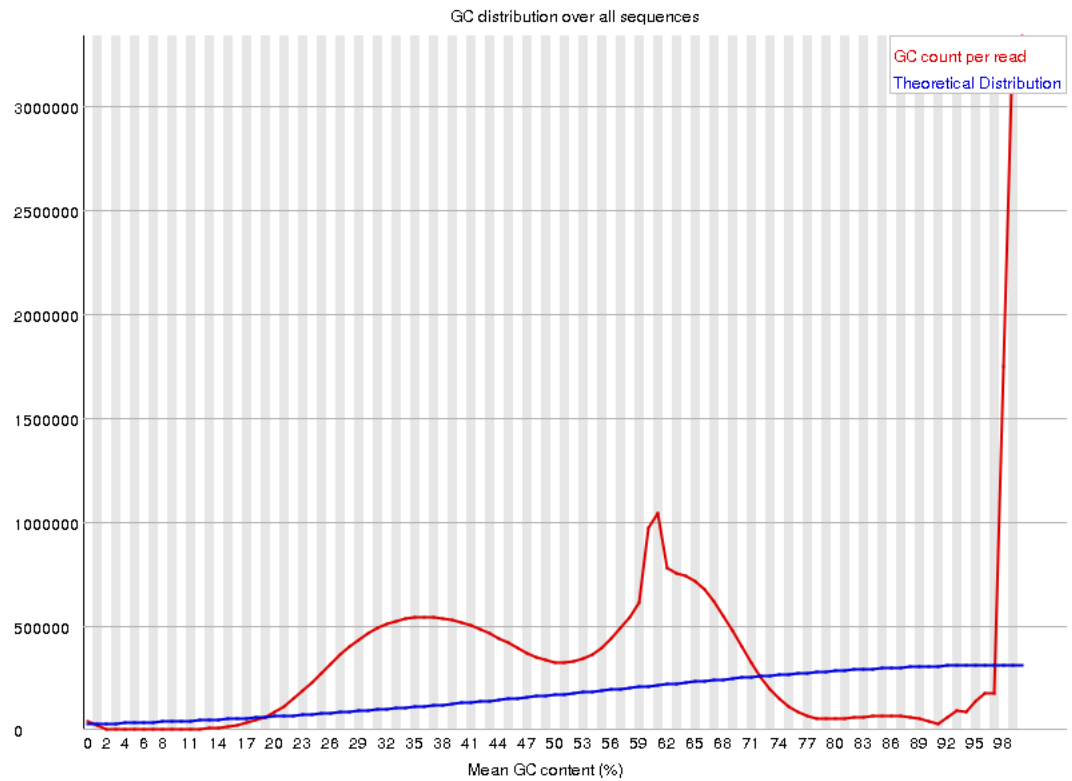
Per base sequence content





Per sequence GC content





Read Alignment

AIM: Given a reference sequence and a set of short reads, align each read to the reference sequence finding the most likely origin of the read sequence.

Reference: ...GCTGATGTGCCGCCTCACTTCGGTGGTACGCT...

Read Alignment

- 1970 – Needleman-Wunsch algorithm for sequence alignment published
- The Smith-Waterman algorithm was the first alignment algorithm to include the concept of gaps
- Faster less computationally intensive alignment algorithms have been developed
- Aligners also use indexes of the genome to speed up alignment

Initialize the scoring matrix

		T	G	T	T	A	C	G	G
	0	0	0	0	0	0	0	0	0
G	0								
G	0								
T	0								
T	0								
G	0								
A	0								
C	0								
T	0								
A	0								

Substitution matrix: $S(a_i, b_j) = \begin{cases} +3, & a_i = b_j \\ -3, & a_i \neq b_j \end{cases}$

Gap penalty: $W_k = kW_1$, $W_1 = 2$

Short Read Aligners

There are a lot of short read aligners, which use a variety of indexing algorithms and alignment algorithms.

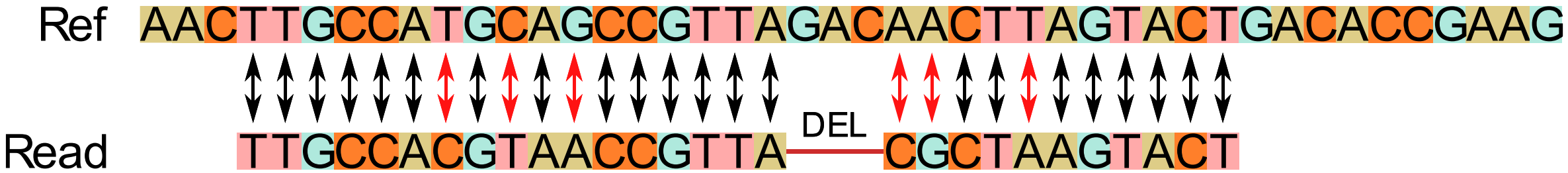
The most popular are probably:

- The Burrows-Wheeler Aligner (BWA) - https://github.com/lh3/bwa
- Bowtie2 - https://github.com/BenLangmead/bowtie2
- STAR - https://github.com/alexdobin/STAR
- HISAT2 - http://daehwankimlab.github.io/hisat2/

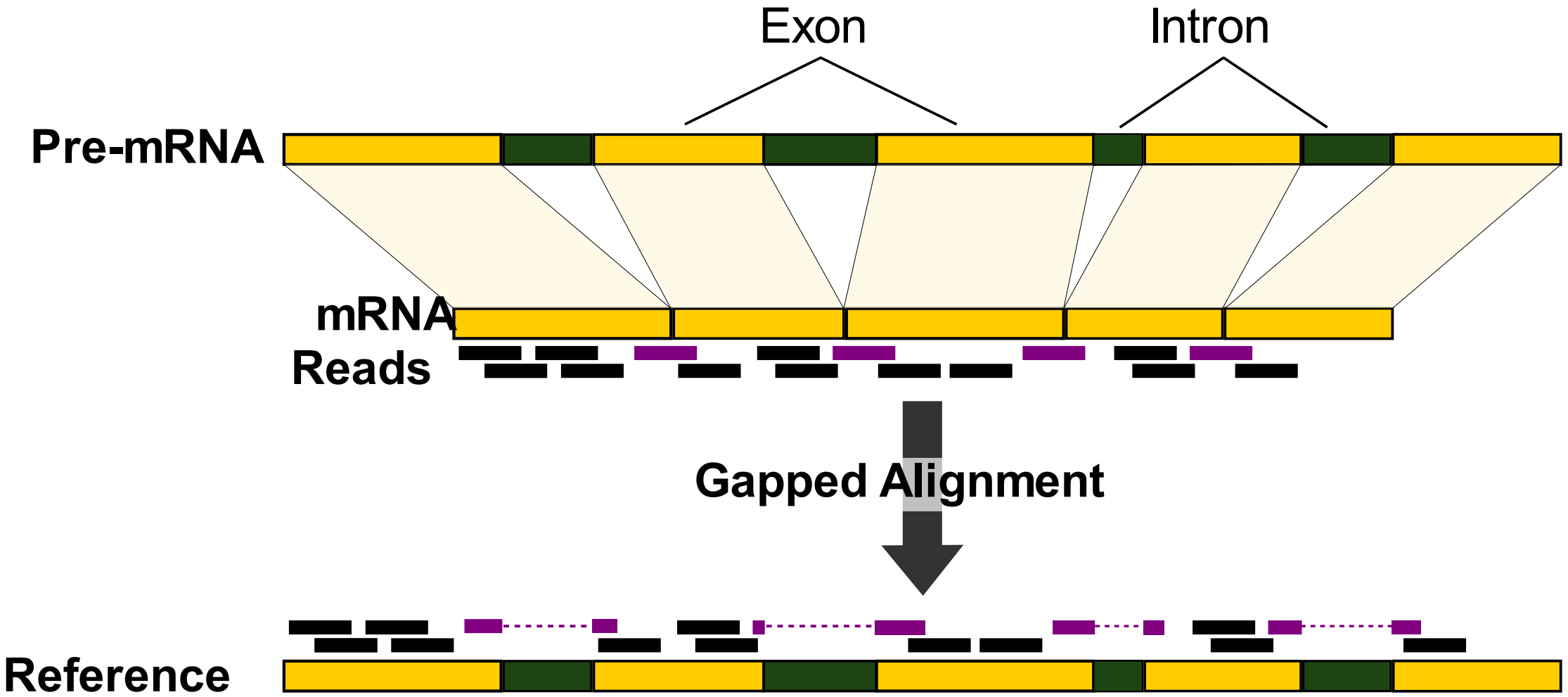
Alignment


Ref AACTTGCCATGCAGCCGTTAGACAACTTAGTACTGACACCGAAG
Read TTGCCACGTAACCGTTACGCTAAGTACT

Alignment


Ref
AACTTGCCATGCAGCCGTTAGACAACTTAGTACTGACACCGAAG
Read
TTGCCACGTAACCGTTA
DEL
CGCTAAGTACT

mRNAseq Alignment



Short Read Aligners

There are a lot of short read aligners, which use a variety of indexing algorithms and alignment algorithms.

The most popular are probably:

- The Burrows-Wheeler Aligner (BWA) - DNA
- Bowtie2 - DNA
- STAR - Either
- HISAT2 - Either

SAM format

Sequence Alignment/Map (SAM) format is the standard format for files containing aligned reads.

Two main parts:

- Header - meta data (source of the reads, reference genome, aligner, etc.)
- Alignment section:
  - 1 line for each alignment
  - contains details of alignment position, mapping, base quality etc.
  - 11 required fields, but other content may vary depending on aligner and other tools used to create the file

SAM format

```
LH00417:64:22MCGJLT3:3:1101:24078:1064  163 X   20546017    60  1S5M1N89M236N5M =   20547099
 1416    ATGTTGATGACAGCCGTCTTGAGGAGCTCAAAGCCACATTGCCCAGCCCAGACAAGTTACCCGGATTTAAGATGTACCCCATTGAT
TTTGAGAAGGATGA    -IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
IIIIIIIIIIIIIIIIIIII    NH:i:1  HI:i:1  AS:i:200    nM:i:0
LH00417:64:22MCGJLT3:3:1101:24078:1064  83  X   20547099    60  12M236N86M2S    =   20546017
 -1416   GATTTTGAGAAGGATGATGACAGCAATTTCCACATGGATTTCATTGTGGCTGCATCCAATCTTCGGGCCGAAAACTATGATATTTC
CCCTGCAGACCGNG    IIIIIIIIIIIIIIII-9III9IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
IIIIIIIIIIIIIIIIII#9    NH:i:1  HI:i:1  AS:i:200    nM:i:0
LH00417:64:22MCGJLT3:3:1101:25280:1064  163 1   34964132    60  53D21M2D26M =   4964187 155 AAG
AACAGAAAATTGCCTACTTCAGGGGCCTAGATACTATTGCCATTGAGTACAGAGCCGAGATTTTTTTTTTTAAGTGAGGGGTGTGTGTGTCTG
TG    -IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII-IIII9I-I-
I9IIII-IIII    NH:i:1  HI:i:1  AS:i:175    nM:i:0
LH00417:64:22MCGJLT3:3:1101:25280:1064  83  1   34964187    60  21M2D77M2S  =   4964132 -155
 AGAGCCGAGATTTTTTTTTTAAGTGAGGGGTGTGTGTGTCTGTGTGTGTGTGTGCGTGCGTGCGTGTGTGCGTGTGTGCGTGTGTGTGT
GTGTNG    IIIIII-IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
IIIIIIIIIIII#9    NH:i:1  HI:i:1  AS:i:175    nM:i:0
```

SAM format

```
LH00417:64:22MCGJLT3:3:1101:24078:1064  163 X   20546017    60  1S5M1N89M236N5M =   20547099
 1416    ATGTTGATGACAGCCGTCTTGAGGAGCTCAAAGCCACATTGCCCAGCCCAGACAAGTTACCCGGATTTAAGATGTACCCCATTGAT
TTTGAGAAGGATGA    -IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
IIIIIIIIIIIIIIIIIIII    NH:i:1  HI:i:1  AS:i:200    nM:i:0
LH00417:64:22MCGJLT3:3:1101:24078:1064  83  X   20547099    60  12M236N86M2S    =   20546017
 -1416   GATTTTGAGAAGGATGATGACAGCAATTTCCACATGGATTTCATTGTGGCTGCATCCAATCTTCGGGCCGAAAACTATGATATTTC
CCCTGCAGACCGNG    IIIIIIIIIIIIIIII-9III9IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
IIIIIIIIIIIIIIIIII#9    NH:i:1  HI:i:1  AS:i:200    nM:i:0
LH00417:64:22MCGJLT3:3:1101:25280:1064  163 1   34964132    60  53D21M2D26M =   4964187 155 AAG
AACAGAAAATTGCCTACTTCAGGGGCCTAGATACTATTGCCATTGAGTACAGAGCCGAGATTTTTTTTTTTAAGTGAGGGGTGTGTGTGTCTG
TG    -IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII-IIII9I-I-
I9IIII-IIII    NH:i:1  HI:i:1  AS:i:175    nM:i:0
LH00417:64:22MCGJLT3:3:1101:25280:1064  83  1   34964187    60  21M2D77M2S  =   4964132 -155
 AGAGCCGAGATTTTTTTTTTAAGTGAGGGGTGTGTGTGTCTGTGTGTGTGTGTGCGTGCGTGCGTGTGTGCGTGTGTGCGTGTGTGTGT
GTGTNG    IIIIII-IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
IIIIIIIIIIII#9    NH:i:1  HI:i:1  AS:i:175    nM:i:0
```

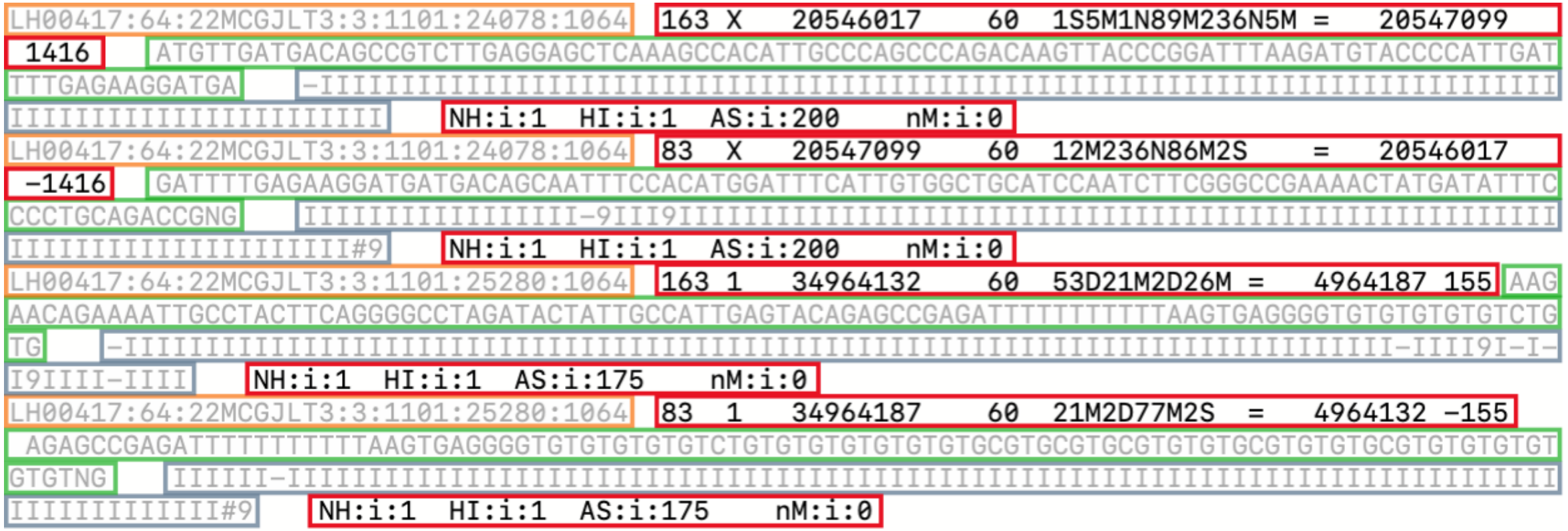
SAM format

```
LH00417:64:22MCGJLT3:3:1101:24078:1064  163 X   20546017    60  1S5M1N89M236N5M =   20547099
 1416    ATGTTGATGACAGCCGTCTTGAGGAGCTCAAAGCCACATTGCCCAGCCCAGACAAGTTACCCGGATTTAAGATGTACCCCATTGAT
TTTGAGAAGGATGA    -IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
IIIIIIIIIIIIIIIIIIII    NH:i:1  HI:i:1  AS:i:200    nM:i:0
LH00417:64:22MCGJLT3:3:1101:24078:1064  83  X   20547099    60  12M236N86M2S    =   20546017
 -1416   GATTTTGAGAAGGATGATGACAGCAATTTCCACATGGATTTCATTGTGGCTGCATCCAATCTTCGGGCCGAAAACTATGATATTTC
CCCTGCAGACCGNG    IIIIIIIIIIIIIIII-9III9IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
IIIIIIIIIIIIIIIIII#9    NH:i:1  HI:i:1  AS:i:200    nM:i:0
LH00417:64:22MCGJLT3:3:1101:25280:1064  163 1   34964132    60  53D21M2D26M =   4964187 155 AAG
AACAGAAAATTGCCTACTTCAGGGGCCTAGATACTATTGCCATTGAGTACAGAGCCGAGATTTTTTTTTTTAAGTGAGGGGTGTGTGTGTCTG
TG    -IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII-IIII9I-I-
I9IIII-IIII    NH:i:1  HI:i:1  AS:i:175    nM:i:0
LH00417:64:22MCGJLT3:3:1101:25280:1064  83  1   34964187    60  21M2D77M2S  =   4964132 -155
 AGAGCCGAGATTTTTTTTTTAAGTGAGGGGTGTGTGTGTCTGTGTGTGTGTGTGCGTGCGTGCGTGTGTGCGTGTGTGCGTGTGTGTGT
GTGTNG    IIIIII-IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
IIIIIIIIIIII#9    NH:i:1  HI:i:1  AS:i:175    nM:i:0
```

SAM format

```
LH00417:64:22MCGJLT3:3:1101:24078:1064  163 X   20546017    60  1S5M1N89M236N5M =   20547099
 1416    ATGTTGATGACAGCCGTCTTGAGGAGCTCAAAGCCACATTGCCCAGCCCAGACAAGTTACCCGGATTTAAGATGTACCCCATTGAT
TTTGAGAAGGATGA    -IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
IIIIIIIIIIIIIIIIIIII    NH:i:1  HI:i:1  AS:i:200    nM:i:0
LH00417:64:22MCGJLT3:3:1101:24078:1064  83  X   20547099    60  12M236N86M2S    =   20546017
 -1416    GATTTTGAGAAGGATGATGACAGCAATTTCCACATGGATTTCATTGTGGCTGCATCCAATCTTCGGGCCGAAAACTATGATATTTC
CCCTGCAGACCGNG    IIIIIIIIIIIIIIIII-9III9IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
IIIIIIIIIIIIIIIIIII#9    NH:i:1  HI:i:1  AS:i:200    nM:i:0
LH00417:64:22MCGJLT3:3:1101:25280:1064  163 1   34964132    60  53D21M2D26M =   4964187 155 AAG
AACAGAAAATTGCCTACTTCAGGGGCCTAGATACTATTGCCATTGAGTACAGAGCCGAGATTTTTTTTTTTAAGTGAGGGGTGTGTGTGTGTCTG
TG    -IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII-IIII9I-I-
I9IIII-IIII    NH:i:1  HI:i:1  AS:i:175    nM:i:0
LH00417:64:22MCGJLT3:3:1101:25280:1064  83  1   34964187    60  21M2D77M2S  =   4964132 -155
 AGAGCCGAGATTTTTTTTTTAAGTGAGGGGTGTGTGTGTGTCTGTGTGTGTGTGTGTGCGTGCGTGCGTGTGTGCGTGTGTGCGTGTGTGTGT
GTGTNG    IIIIII-IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
IIIIIIIIIIII#9    NH:i:1  HI:i:1  AS:i:175    nM:i:0
```

SAM format


LH00417:64:22MCGJLT3:3:1101:24078:1064 163 X 20546017 60 1S5M1N89M236N5M = 20547099
1416 ATGTTGATGACAGCCGTCTTGAGGAGCTCAAAGCCACATTGCCCAGCCCAGACAAGTTACCCGGATTTAAGATGTACCCCATTGAT
TTTGAGAAGGATGA -IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
IIIIIIIIIIIIIIIIIIII NH:i:1 HI:i:1 AS:i:200 nM:i:0
LH00417:64:22MCGJLT3:3:1101:24078:1064 83 X 20547099 60 12M236N86M2S = 20546017
-1416 GATTTTGAGAAGGATGATGACAGCAATTTCCACATGGATTTCATTGTGGCTGCATCCAATCTTCGGGCCGAAAACTATGATATTTC
CCCTGCAGACCGNG IIIIIIIIIIIIIIIII-9III9IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
IIIIIIIIIIIIIIIIIII#9 NH:i:1 HI:i:1 AS:i:200 nM:i:0
LH00417:64:22MCGJLT3:3:1101:25280:1064 163 1 34964132 60 53D21M2D26M = 4964187 155 AAG
AACAGAAAATTGCCTACTTCAGGGGCCTAGATACTATTGCCATTGAGTACAGAGCCGAGATTTTTTTTTTAAGTGAGGGGTGTGTGTGTCTG
TG -IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII-IIII9I-I-
I9IIII-IIII NH:i:1 HI:i:1 AS:i:175 nM:i:0
LH00417:64:22MCGJLT3:3:1101:25280:1064 83 1 34964187 60 21M2D77M2S = 4964132 -155
AGAGCCGAGATTTTTTTTTTAAGTGAGGGGTGTGTGTGTCTGTGTGTGTGTGTGCGTGCGTGCGTGTGTGCGTGTGTGCGTGTGTGTGT
GTGTNG IIIIII-IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
IIIIIIIIIIII#9 NH:i:1 HI:i:1 AS:i:175 nM:i:0

SAM format

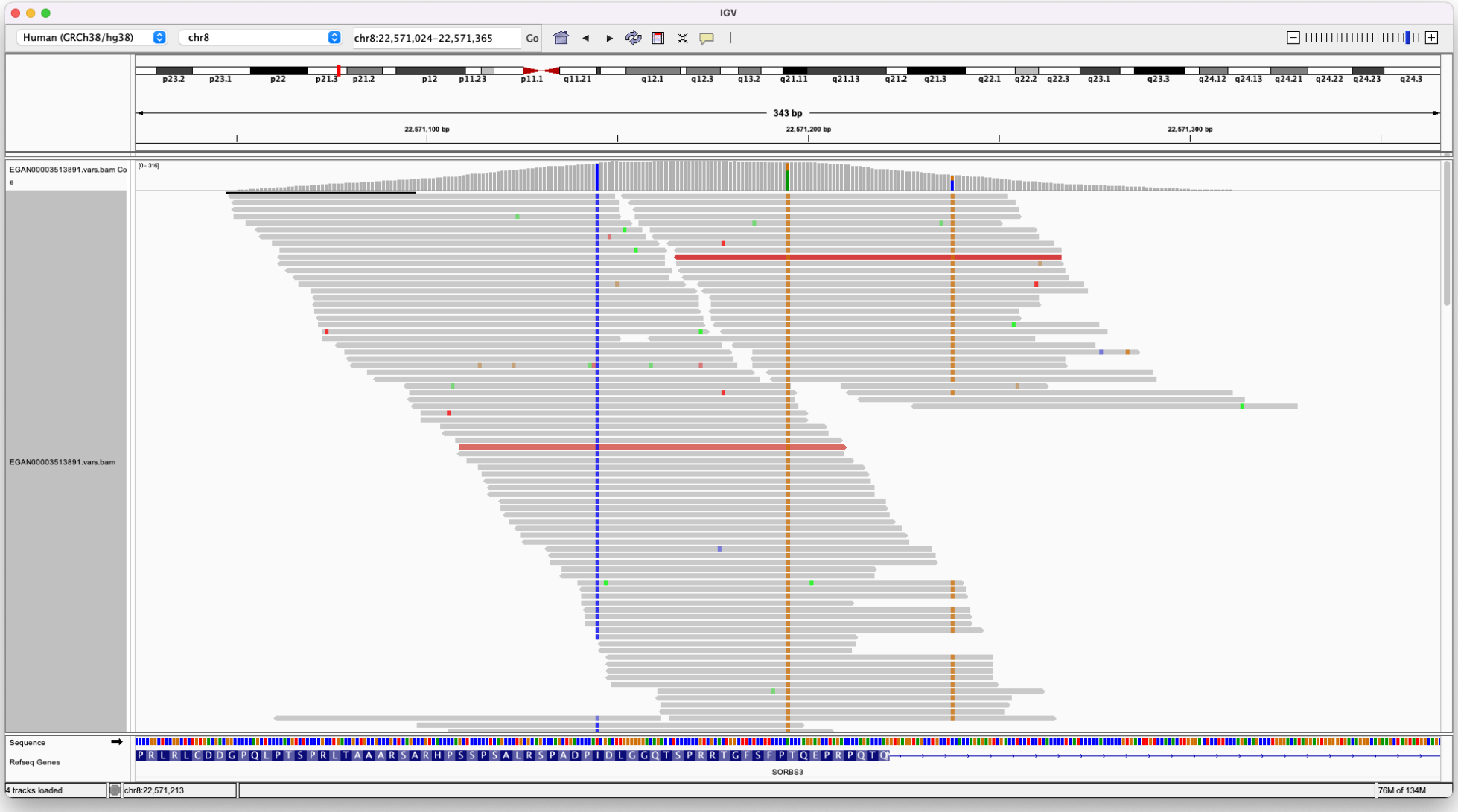
QNAME	LH00417:64:22MCGJLT3:3:1101:24078:1064
FLAG	163
RNAME	X
POS	20546017
MAPQ	60
CIGAR	1S5M1N89M236N5M
RNEXT	=
PNEXT	20547099
TLEN	1416
SEQ	ATGTTGATGACAGCCGTCTTGAGGAGCTC...
QUAL	-IIIIIIIIIIIIIIIIIIIIIIIIIII...
	NH:i:1
	HI:i:1
	AS:i:200
	nM:i:0

DNAseq – Somatic Variant Calling

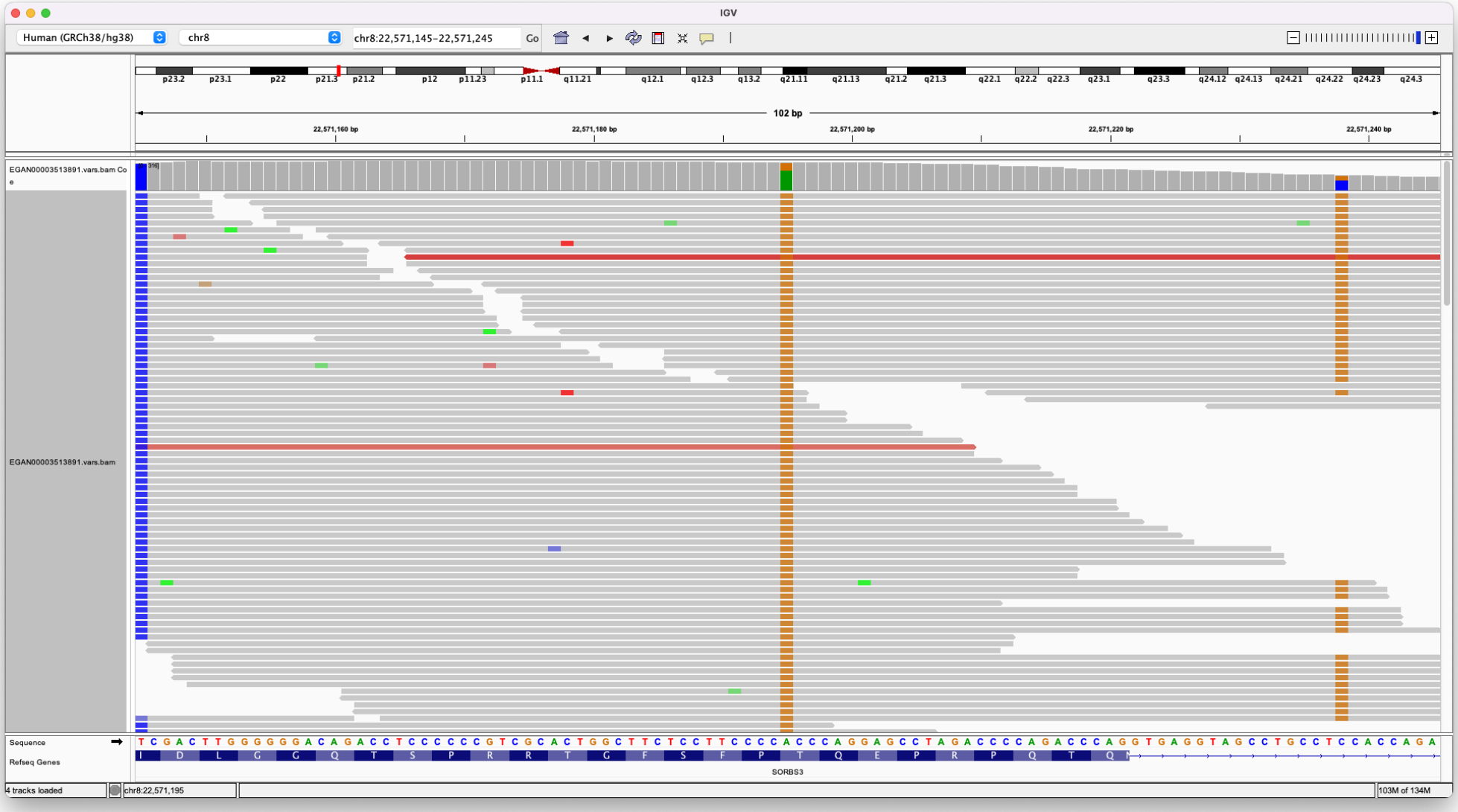
AIM: To identify mutations in the genome

- Whole Genome Sequencing, Whole Exome Sequencing, Targeted Panels
- Sequence paired end with long reads (>=150 bp)
- Remove or mark duplicate reads
- Align to genome including viral decoy sequences, e.g. human cytomegalovirus (CMV), Epstein-Barr virus (EBV)
- Call Single Nucleotide Variants (SNV) and small Insertions/Deletions (Indels)
- Many somatic mutations may have very low variant allele frequency - sequence to high depth – 100x coverage
- Differentiate germline variants by including both a tumour sample and a normal sample

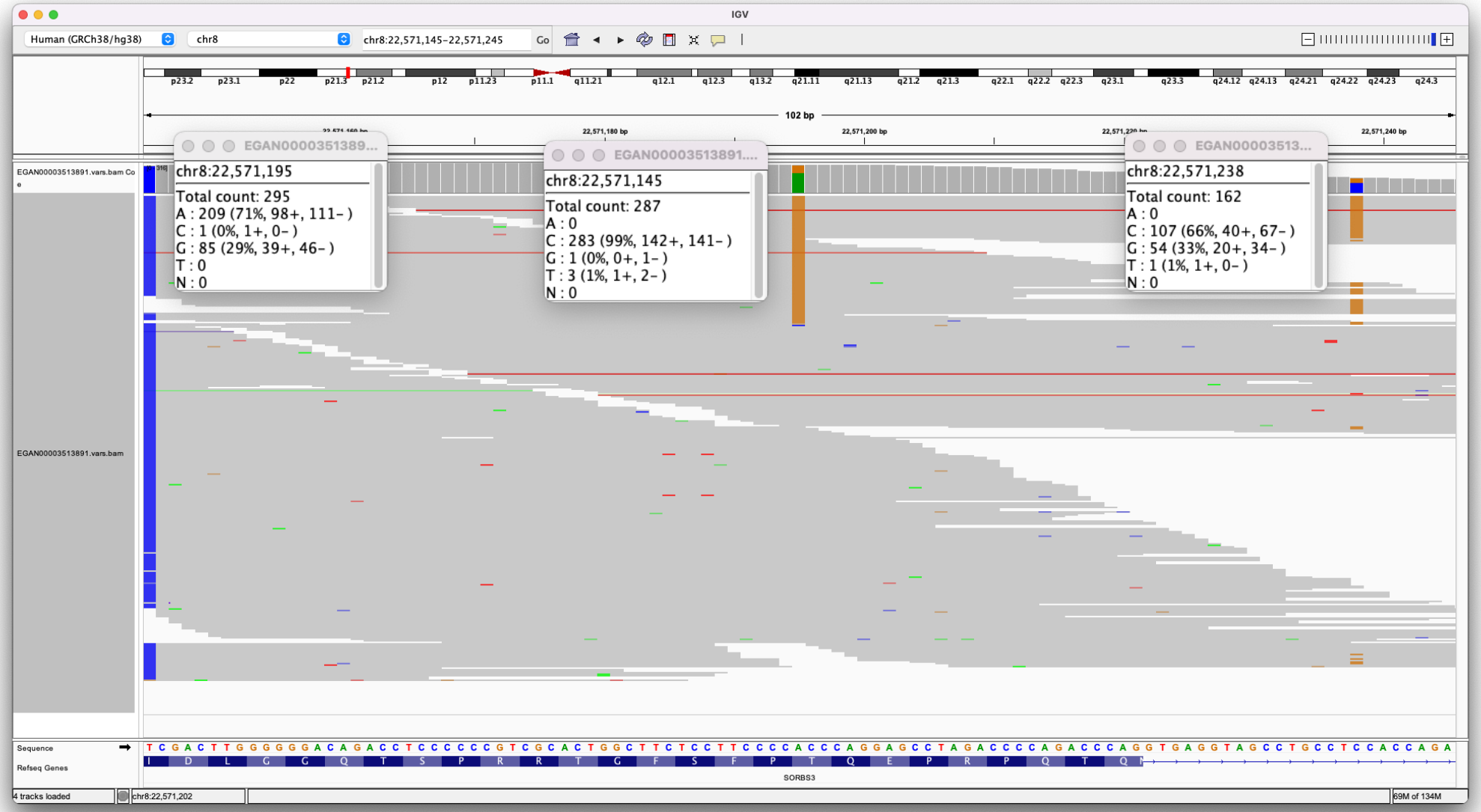
DNAseq – Somatic Variant Calling



DNAseq – Somatic Variant Calling



DNAseq – Somatic Variant Calling



Several factors complicate somatic SNV calling

- Low cellularity (tumour DNA content)
- Intra-tumour heterogeneity in which multiple tumour cell populations (subclones) exist
- Aneuploidy
- Unbalanced structural variation (deletions, duplications, etc.)
- Matched normal contaminated with cancer DNA
  - adjacent normal tissue may contain residual disease or early tumour-initiating somatic mutations
  - circulating tumour DNA in blood normals
- Sequencing errors
- Alignment artefacts

Mwenifumbo & Marra, Nat Rev Genet. 2013

DNAseq – Somatic Variant Calling

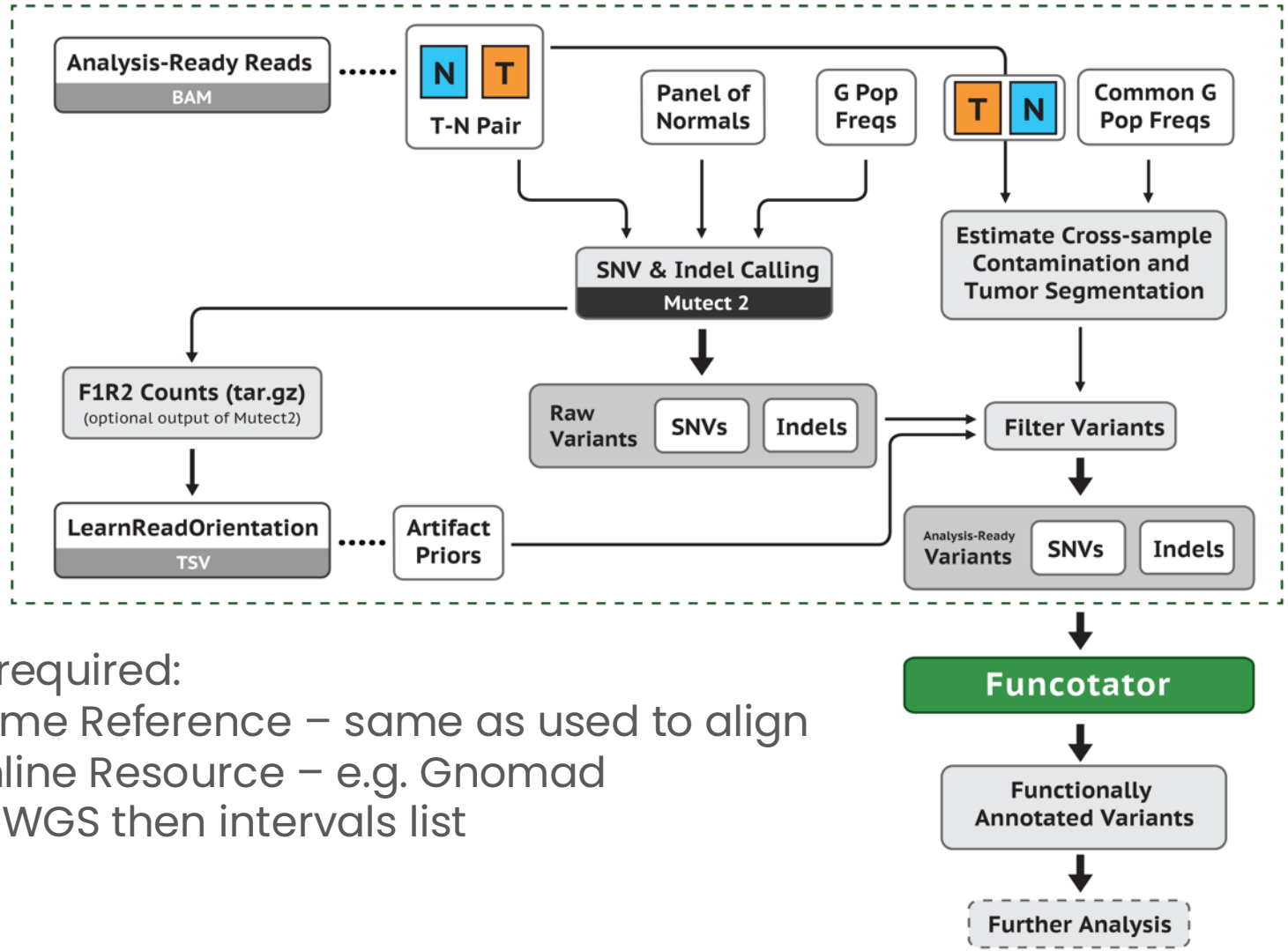
There are a lot of tools for somatic variant calling.

Some of the most popular are:

- Mutect2 (GATK)
- Strelka
- FreeBayes
- VarDict
- VarScan2

They take different approaches, and it is not uncommon to use multiple tools to call variants and then take a consensus

Using Mutect2 from GATK



References required:

- Genome Reference – same as used to align
- Germline Resource – e.g. Gnomad
- If not WGS then intervals list

Variant Call Format (VCF) output

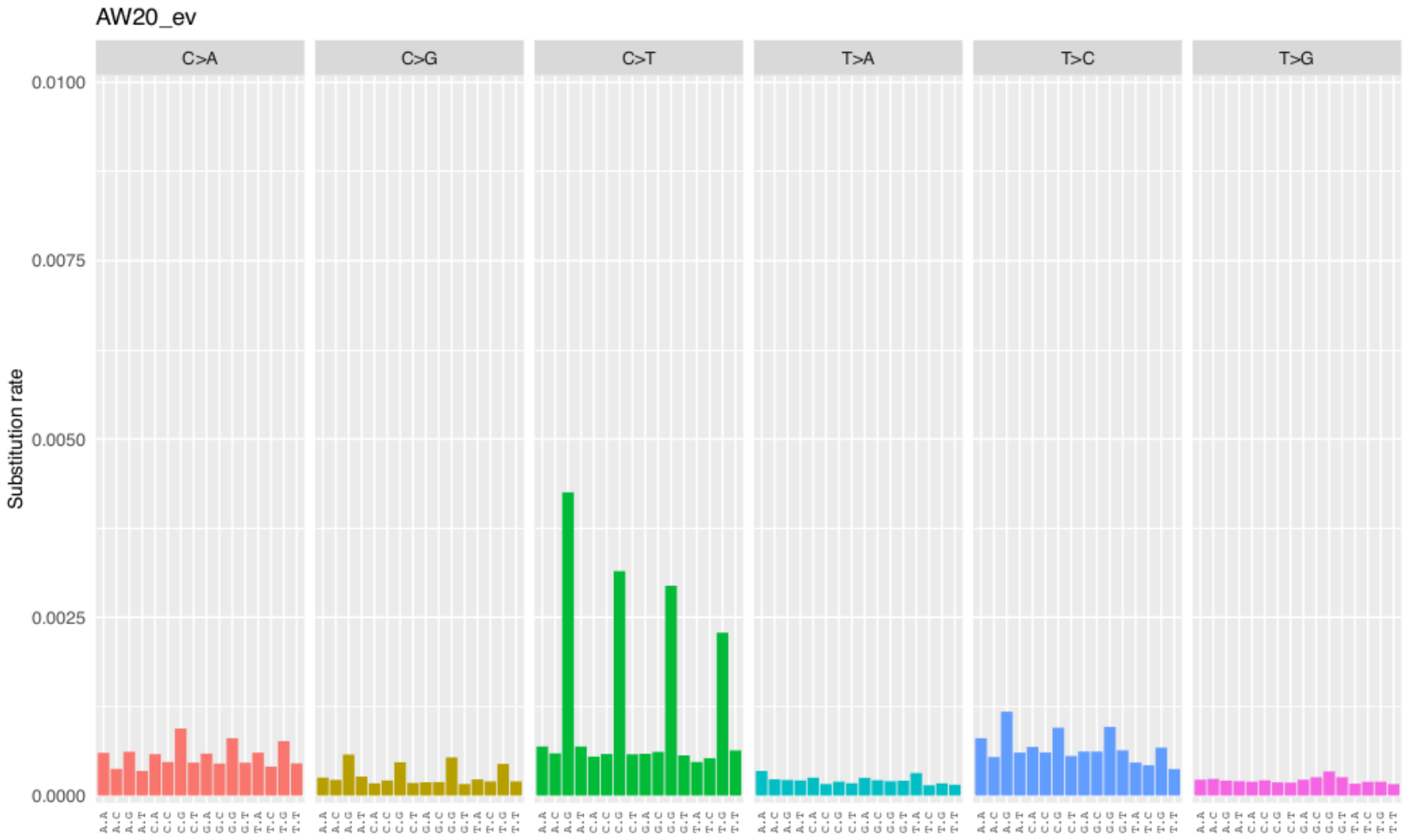
##fileformat=VCFv4.1

##FILTER=<ID=**base_quality**,Description=”Average base quality for variant alleles < 25">
##FILTER=<ID=**germline_risk**,Description="Evidence suggests that the site may be germline, not somatic">

##FORMAT=<ID=**GT**,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read depth at this position in the sample">
##FORMAT=<ID=**AD**,Number=R,Type=Integer,Description="Allelic depths for the ref and alt alleles in the order listed">
##FORMAT=<ID=BQ,Number=.,Type=Integer,Description="Average base quality for reads supporting alleles">

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	NORMAL	TUMOUR
8	142478336	.	A	C		**germline_risk, base_quality**	.	**GT**:DP:**AD**:BQ	**0/0**:24:**23,1**:38,31	**0/1**:41:**35,6**:32,14
8	**142486034**	.	**C**	**T**	.	**PASS**	.	**GT**:DP:**AD**:BQ	**0/0**:43:**43,0**:36,0	**0/1**:52:**44,8**:36,40

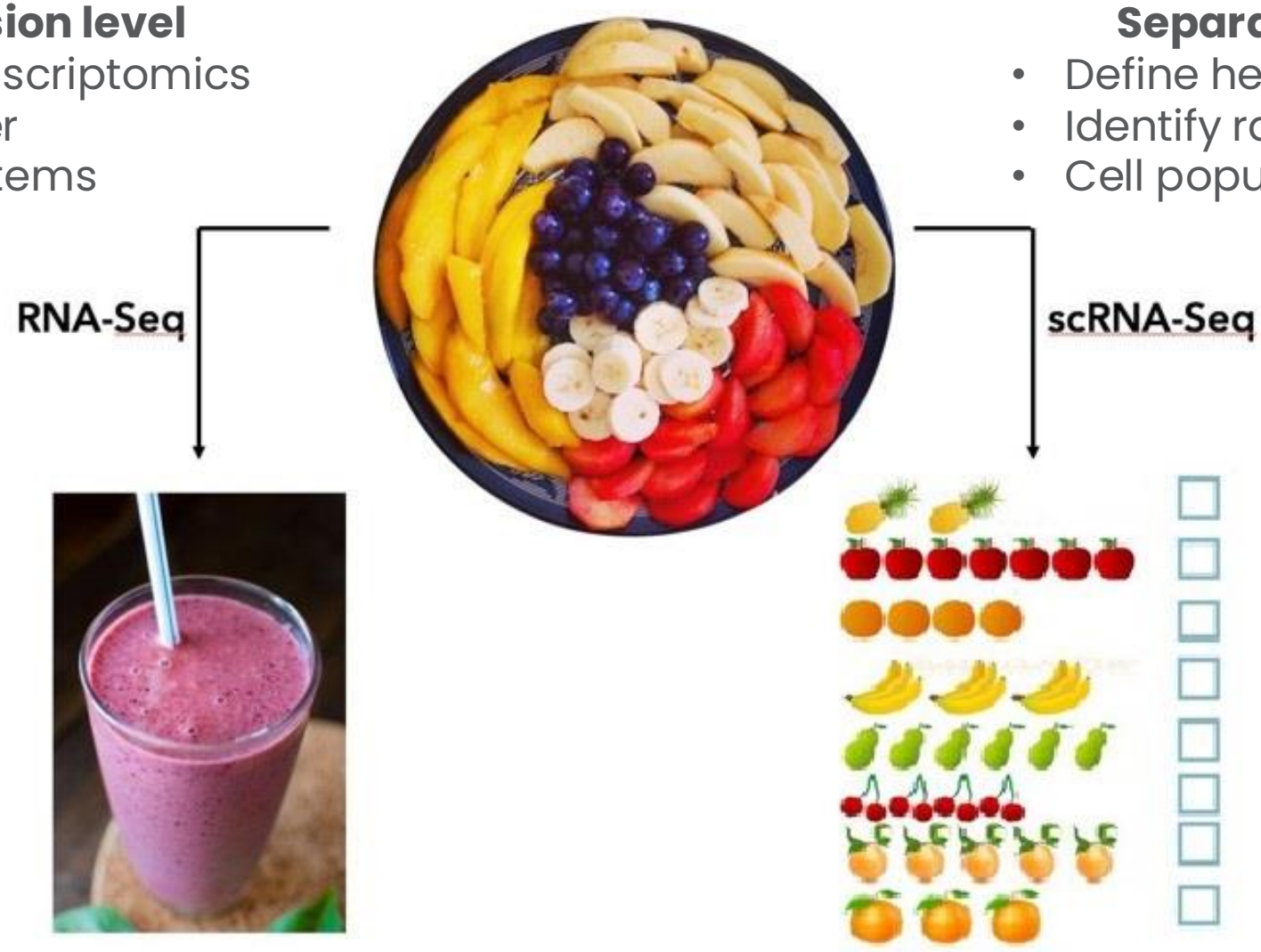
Hypermutation Signatures in Glioblastoma



Single Cell RNAseq

Average expression level

- Comparative transcriptomics
- Disease biomarker
- Homogenous systems



Separate populations

- Define heterogeneity
- Identify rare cell populations
- Cell population dynamics

Single Cell RNAseq

nature medicine

Letter | Published: 08 June 2020

A single-cell atlas of the peripheral immune response in patients with severe COVID-19

Aaron J. Wilk, Arjun Rustagi, Nancy Q. Zhao, Jonasel Roque, Giovanny J. Martínez-Colón, Julia L. McKechnie, Geoffrey T. Ivison, Thanmayi Ranganath, Rosemary Vergara,

LETTER

https://doi.org/10.1038/s41586-018-0394-6

A single-cell atlas of the airway epithelium reveals the CFTR-rich pulmonary ionocyte

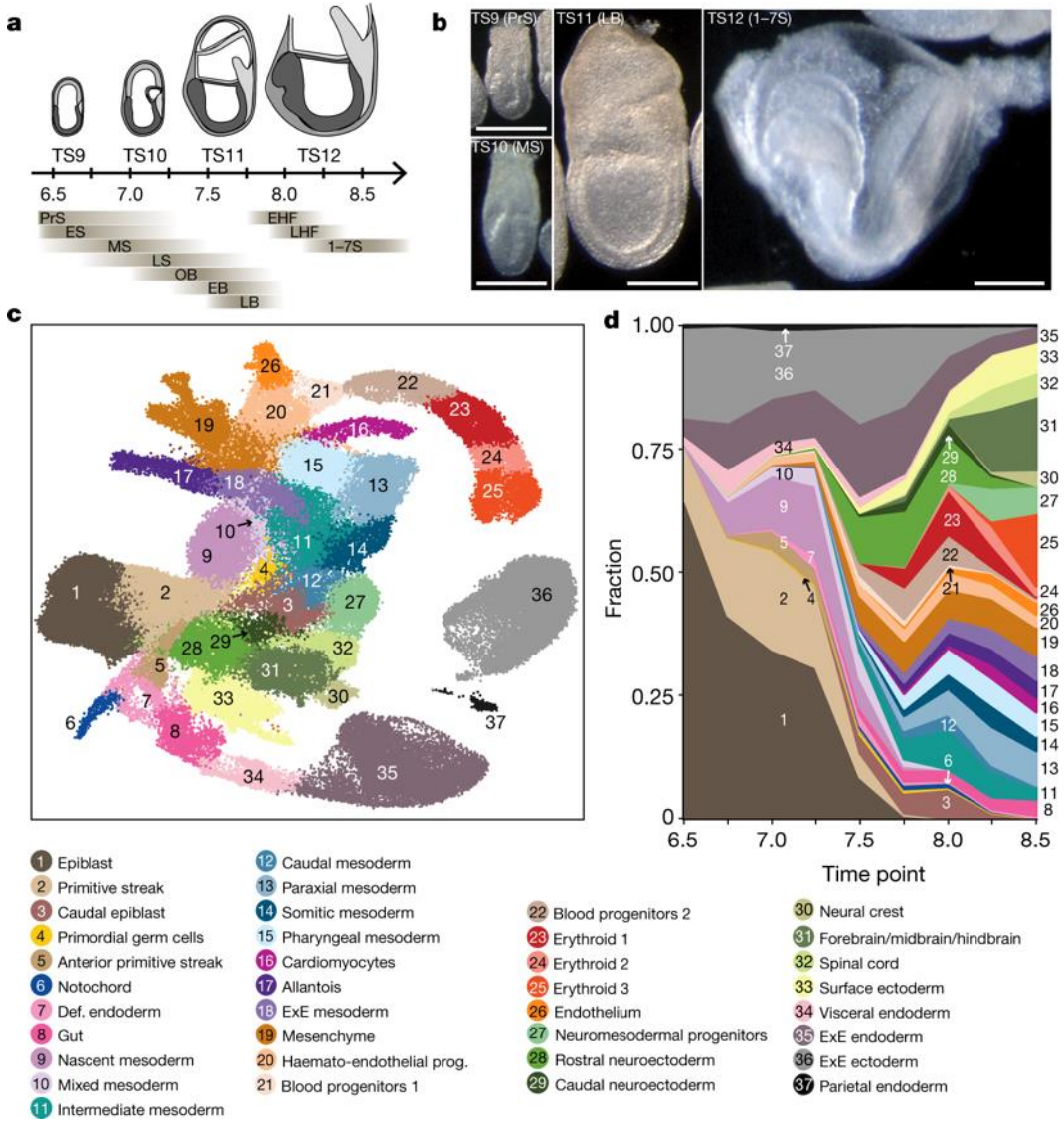
Lindsey W. Plasschaert[1,5,7], Rapolas Žilionis[2,3,7], Rayman Choo-Wing[1,5], Virginia Savova[2,6], Judith Knehr[4], Guglielmo Roma[4], Allon M. Klein[2]* & Aron B. Jaffe[1,5]*

nature

Article | Published: 20 February 2019

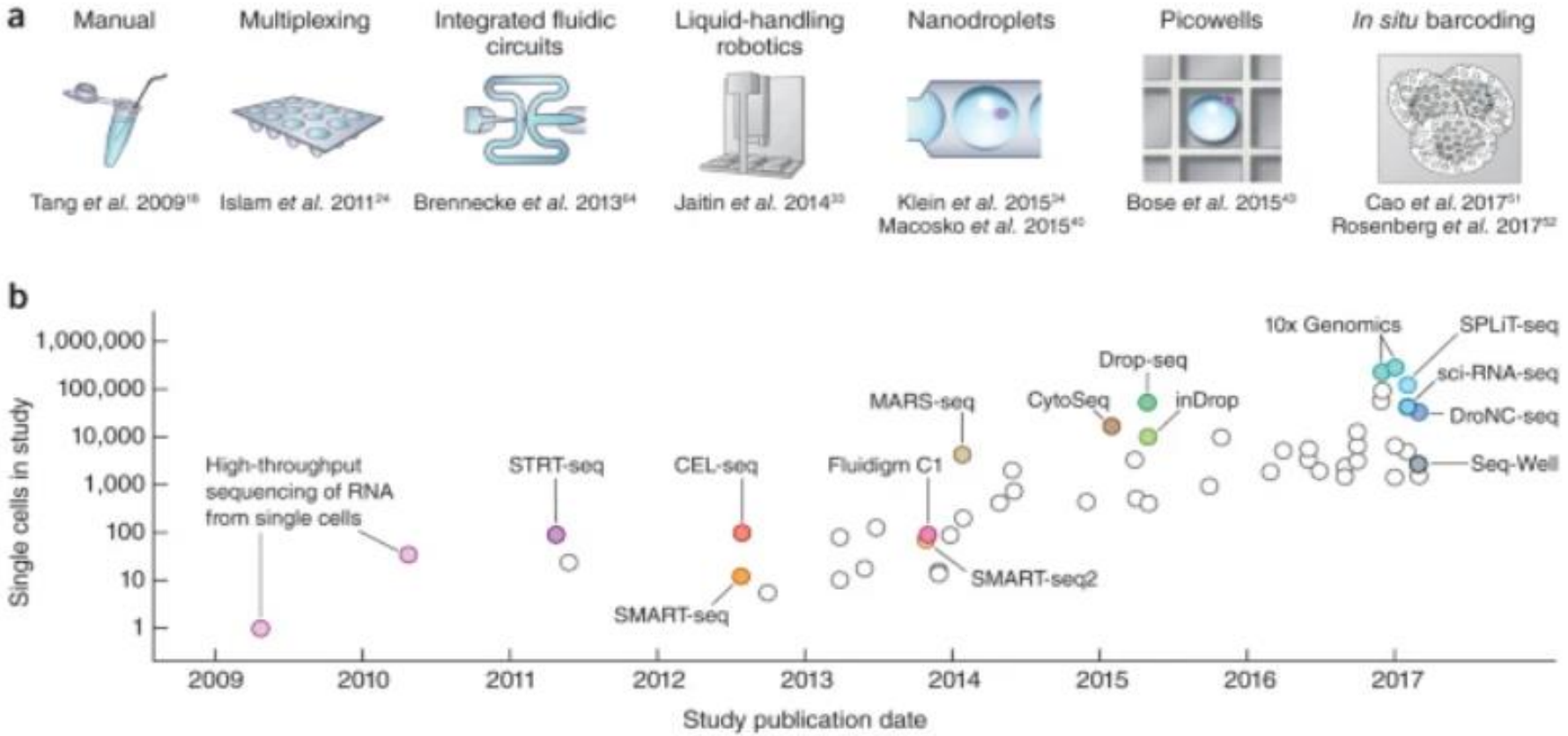
A single-cell molecular map of mouse gastrulation and early organogenesis

Blanca Pijuan-Sala, Jonathan A. Griffiths, Carolina Guibentif, Tom W. Hiscock, Wajid Jawaid, Fernando J. Calero-Nieto, Carla Mulas, Ximena Ibarra-Soria, Richard C. V.



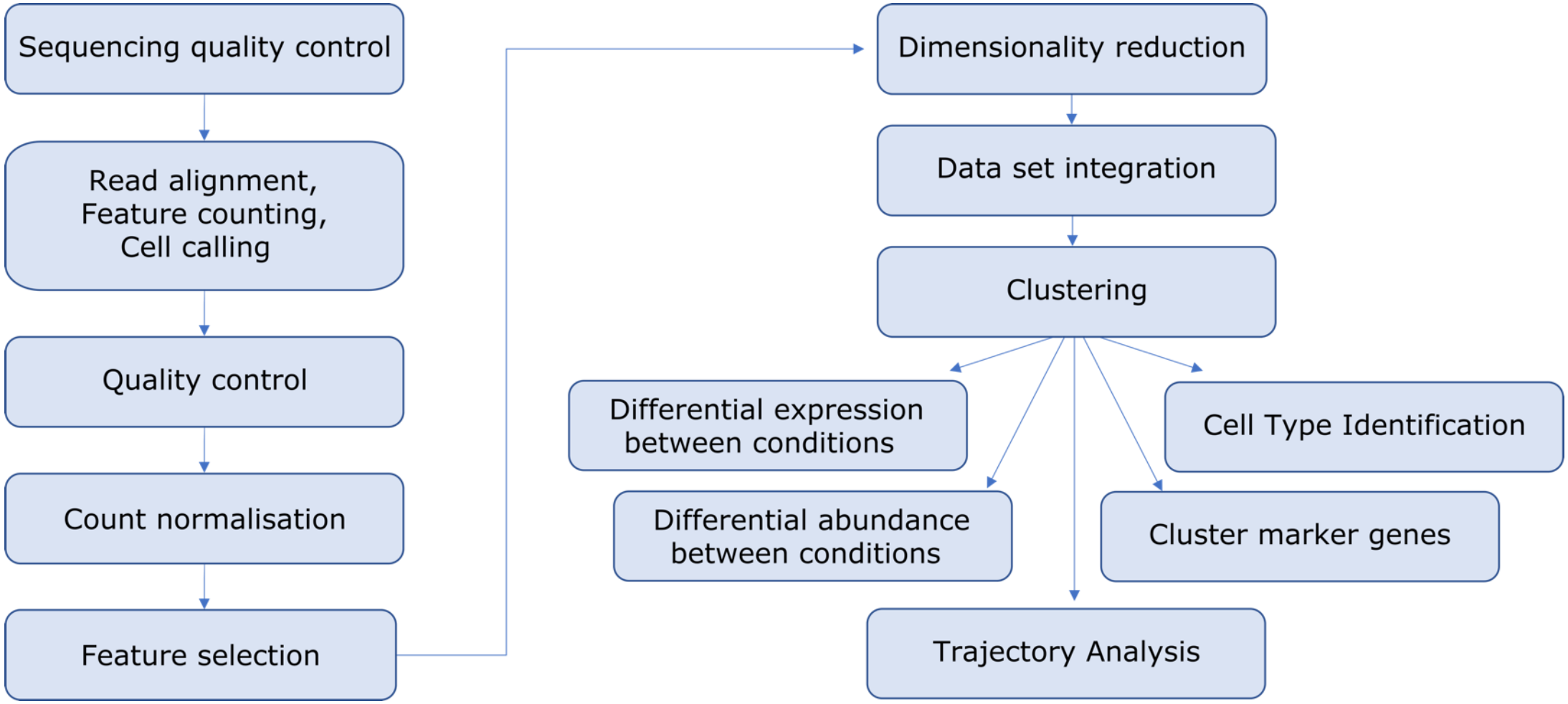
Single Cell RNAseq

Figure 1: Scaling of scRNA-seq experiments.

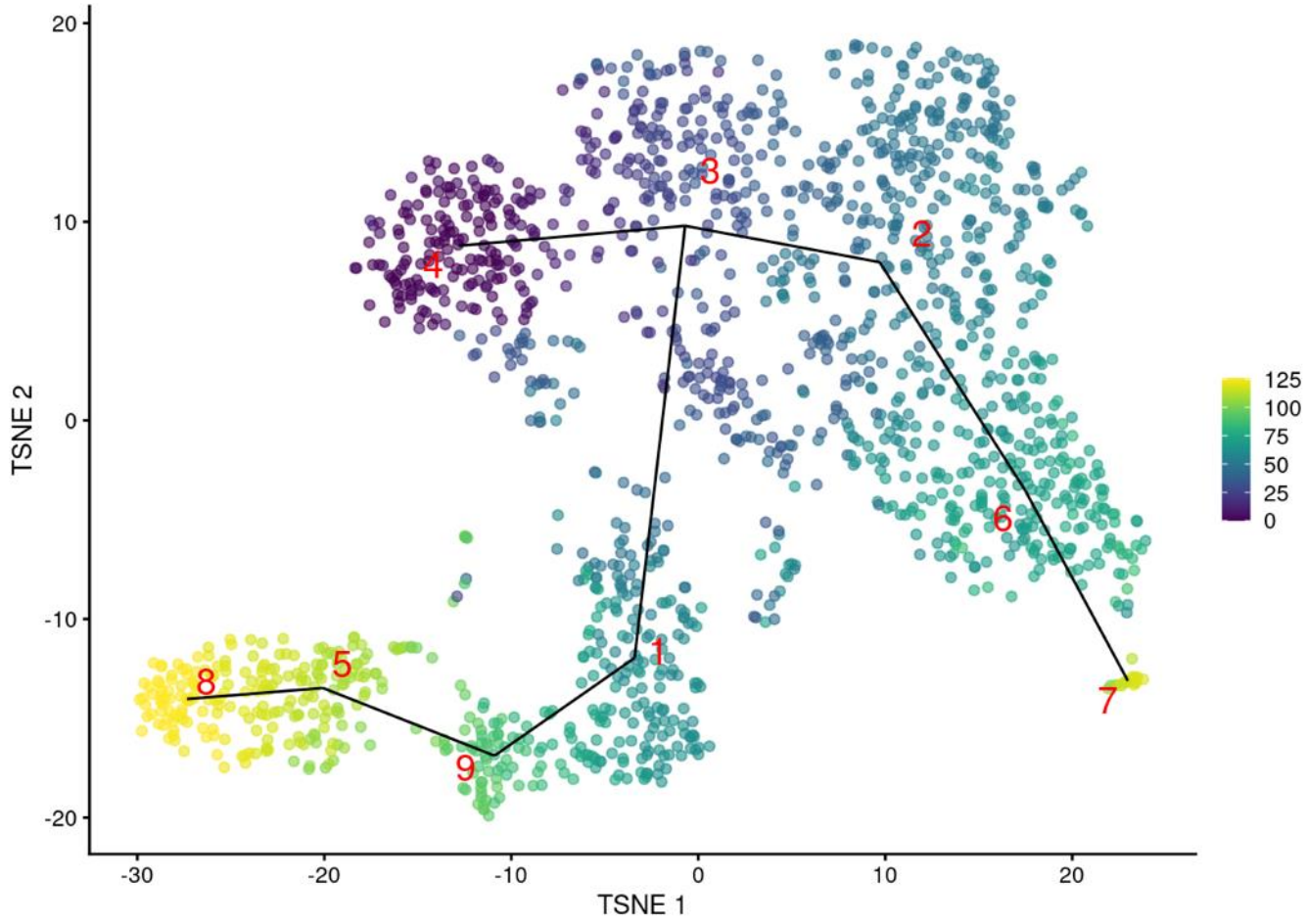


Source: Svensson et al. *Nat Protoc* **13,** 599–604 (2018)

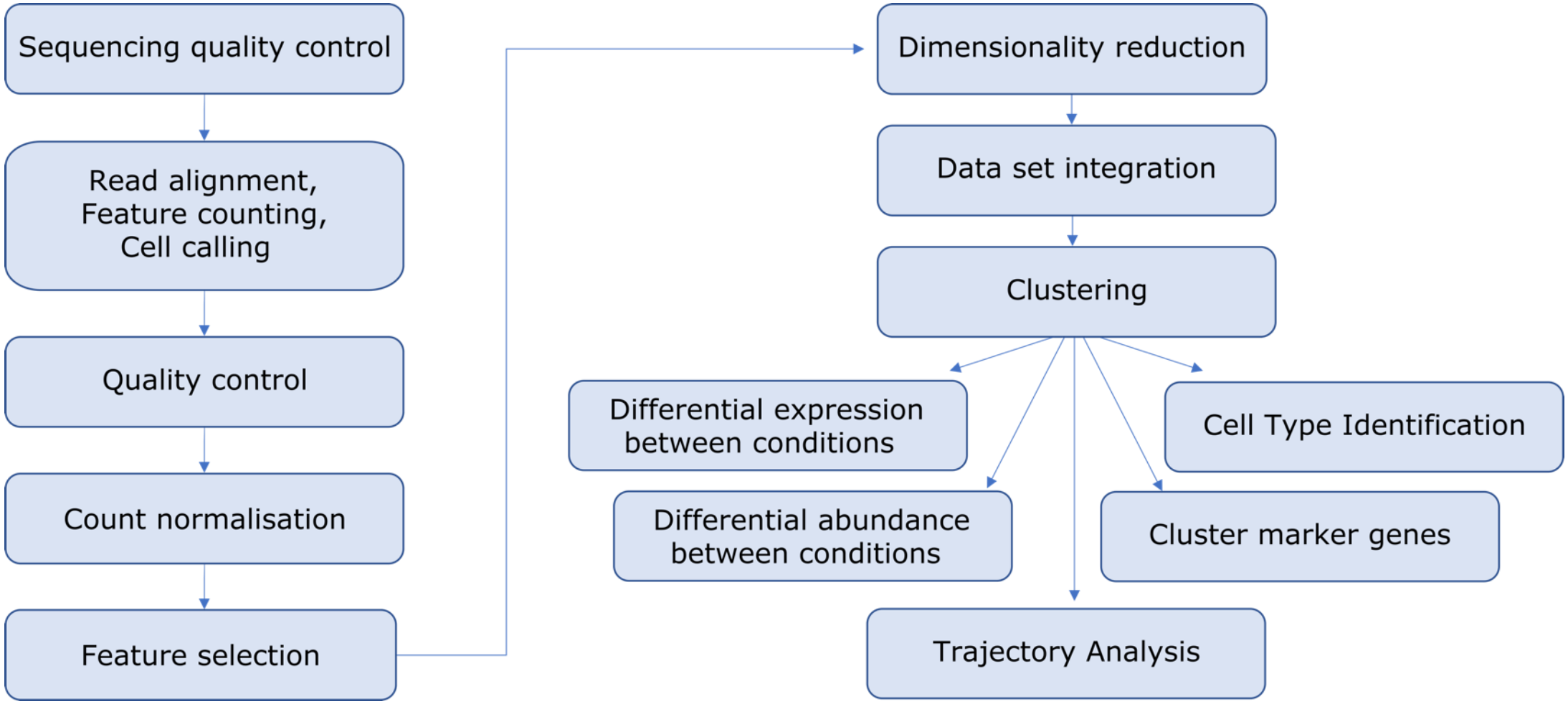
scRNAseq analysis workflow



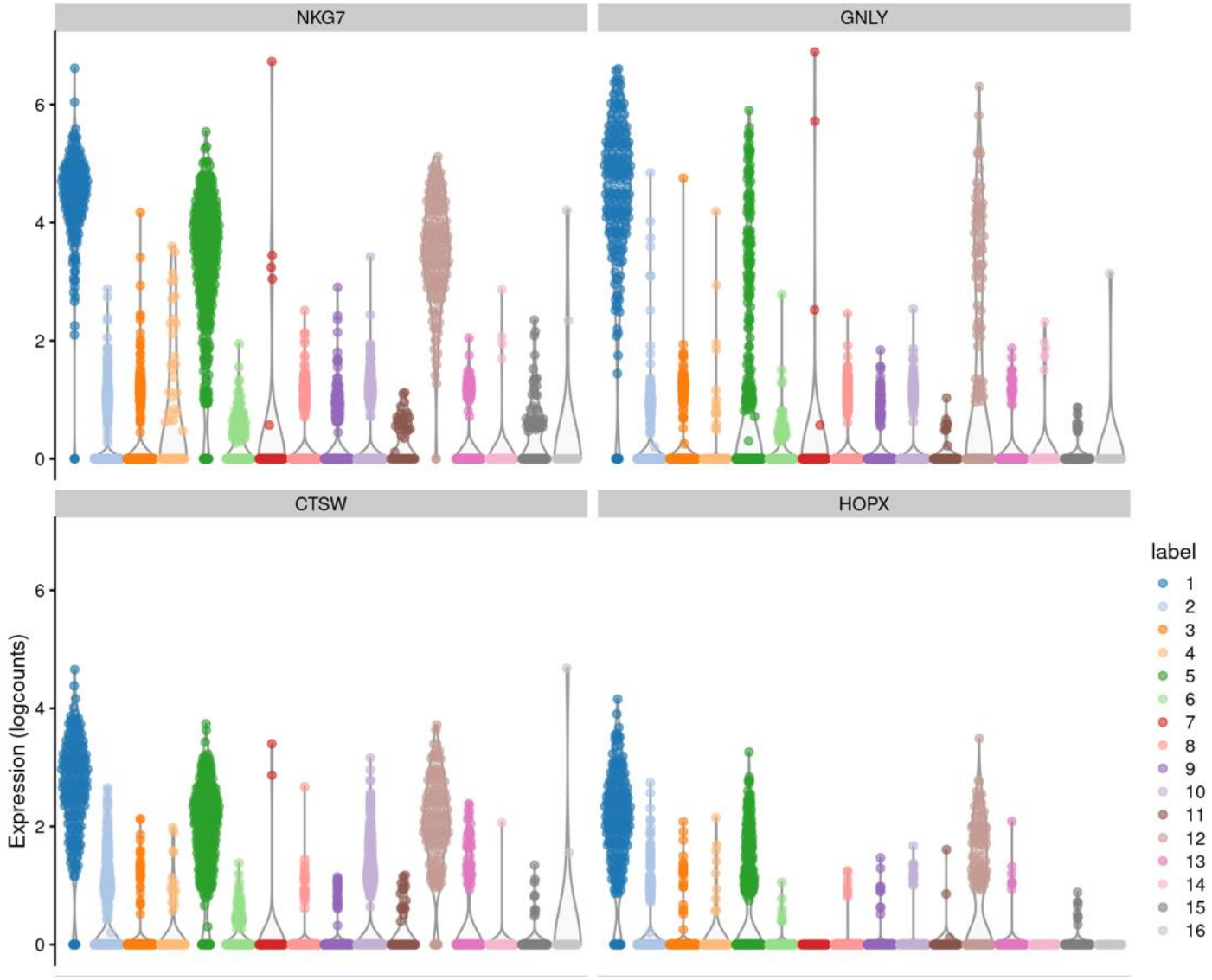
Trajectory Analysis



scRNAseq analysis workflow



Cluster Marker Genes



scRNAseq analysis tools

- Alignment, Gene Expression Quantification, Cell Calling, QC:
  - CellRanger (10X), STARsolo, Alevin
- Data Exploration:
  - Loupe Browser (10X)
- Downstream Analysis:
  - R – Bioconductor packages: scran, scater, bluster, MiloR, SingleR
    - See the OSCA book at https://bioconductor.org/books/release/OSCA/
  - R – Seurat
    - See the Seurat documentation at https://satijalab.org/seurat/
  - Python – Scanpy
    - See the Scanpy documentation at https://scanpy.readthedocs.io/en/stable/

What is Bioinformatics?

In the beginning of the 1970s, Ben Hesper and I started to use the term “bioinformatics” for the research we wanted to do, defining it as “**the study of informatic processes in biotic systems**”.

Paulien Hogeweg, https://doi.org/10.1371/journal.pcbi.1002021

THANK YOU

CANCER RESEARCH UK | CAMBRIDGE INSTITUTE