Syllabus and grading

#	Date	Lecturer	Subject
1	6/10/26	MR	Overview of Bioinformatics, sequence alignment
2	13/10/26	AD	Introduction to Linux and the command line, bash scripting and ssh
3	20/10/26	MR	Next generation sequencing: introduction, quality control and gene expression analysis for RNAseq
4	27/10/26	MR	NGS for denovo genome and transciptome assembly
5	3/11/26	AD	Introduction to the R programming language and Rstudio usage
6	10/11/26	AD	Advances R subjects, introduction to Bioconductor
	~~17/11/26~~		
7	24/11/26	MR	NGS analysis for molecular interactions (ChipSeq, (Par-)Clip, structural sequencing, chromosome conformation capture)
8	1/12/26	AD	Command line tool usage: bedtools, vcftools, samtools etc.
9	8/12/26	MR	Genome-scale models of metabolism and macromolecular expression, Biological applications of Transformers
10	15/12/26	AD	Pipelines for SNP calling, especially for exome sequencing using the GATK pipeline
11	22/12/26	MR+AD	Support for the final project (online)
12	12/1/27	MR+AD	Paper presentations by students
13	19/1/27	MR+AD	Paper presentations by students

Grade	**100%**
Presentation	30%
Exercises	20%
Final Project	50%

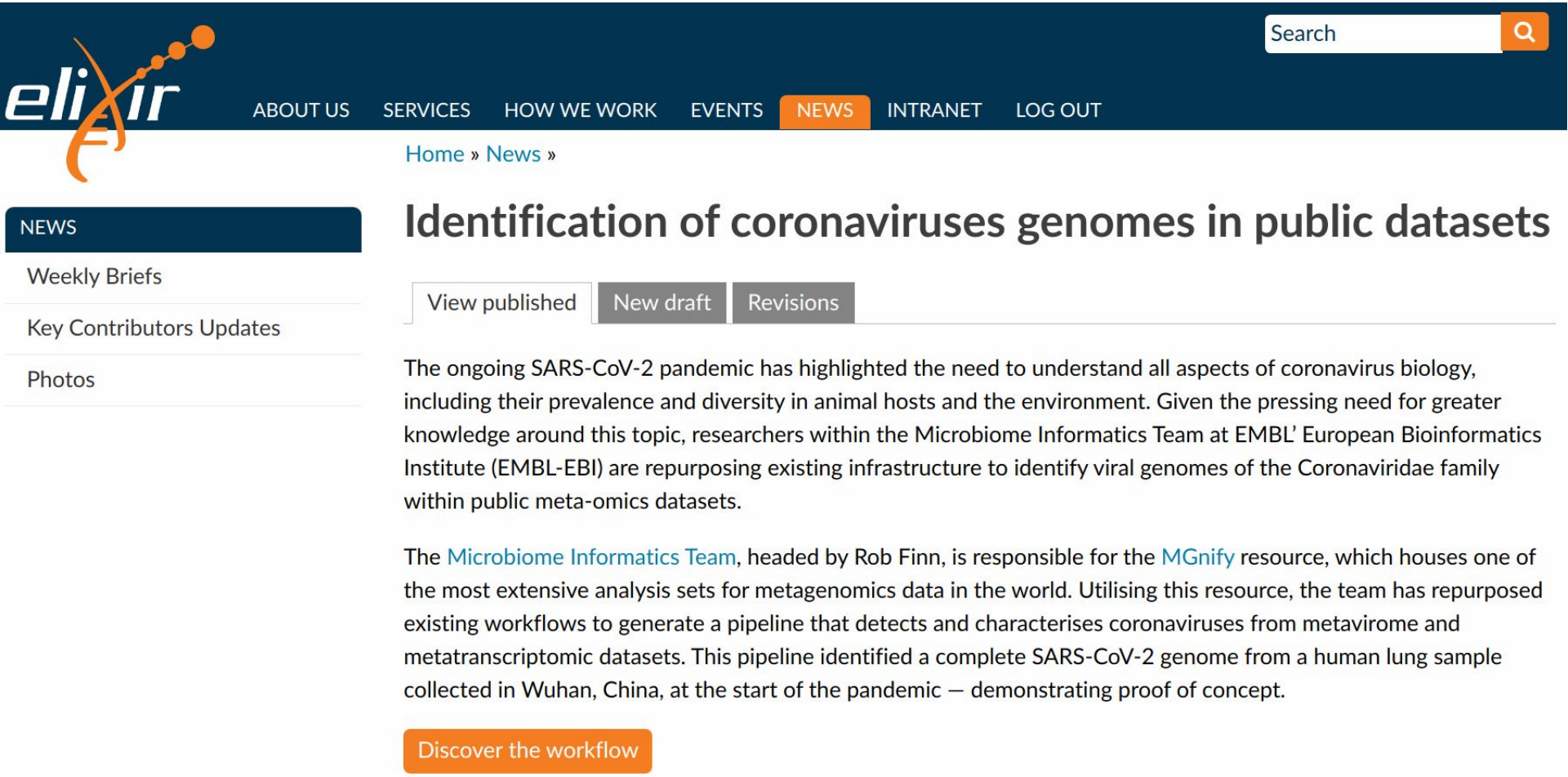
Remarks for exercise 1:

- 13/16 received
- All submissions were correct

- "making of" ? use in e.g.: plagiarism detection

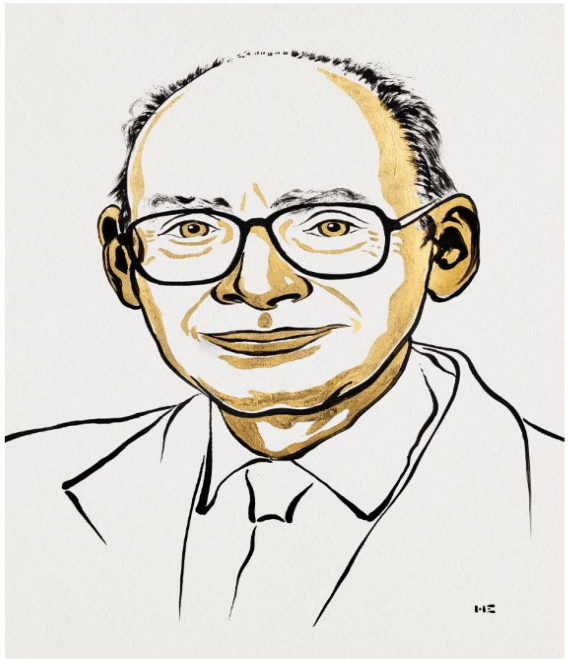
- Reality check .....

Same subject during the pandemic



https://elixir-europe.org/news/identification-coronaviruses-genomes-public-datasets

Nobel Prize in Chemistry 2026



Ill. Niklas Elmehed © Nobel Prize Outreach

Henri B. Kagan

Prize share: 1/2

Ill. Niklas Elmehed © Nobel Prize Outreach

Kenso Soai

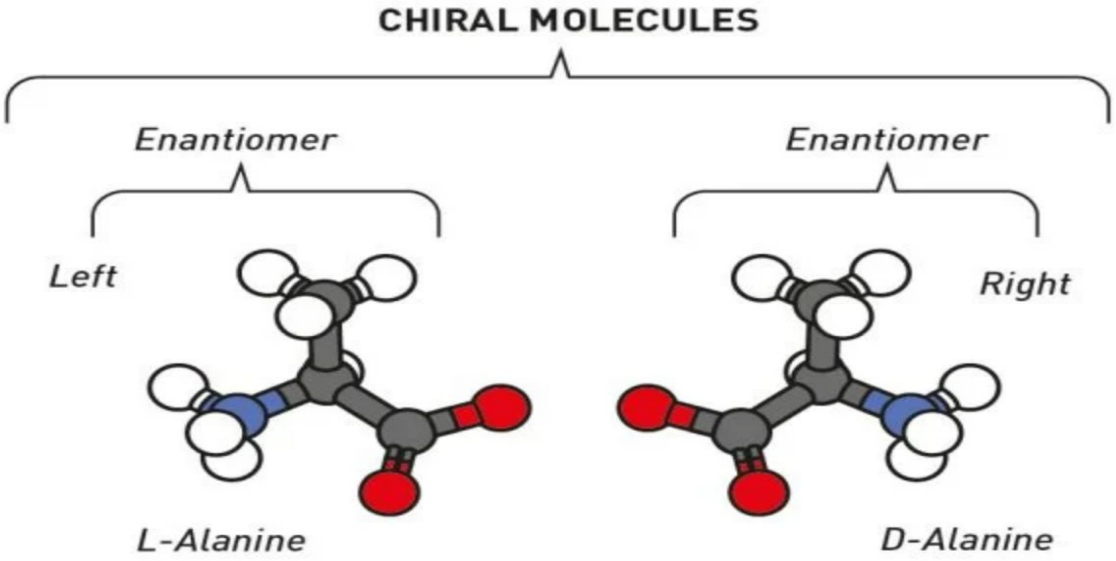
Prize share: 1/2

https://www.nobelprize.org/all-nobel-prizes-2026

Nobel Prize in Chemistry 2026

About chirality

Molecules that exist as two non-identical mirror images are called *chiral* molecules. The mirror images are called *enantiomers* and are defined as either right or left. If a group of molecules – such as amino acids – consists of only right enantiomers or left enantiomers, this is called *homochirality*.



Nobel Prize in Chemistry 2026

A more than 100-year-old mystery

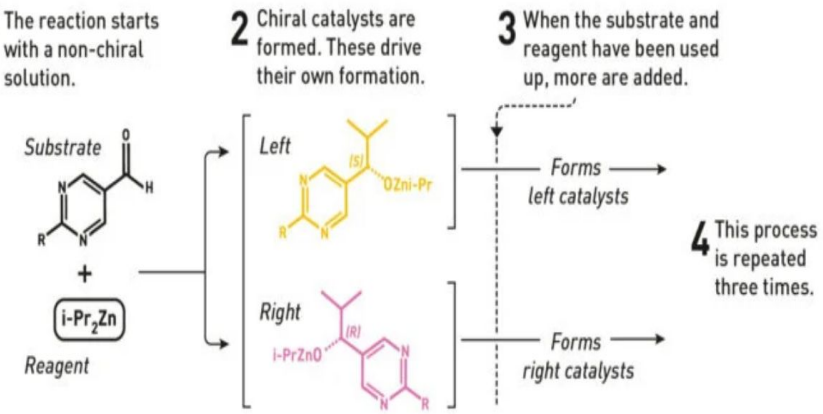
The basis of the chemistry laureates' work was created in the mid-1800s. This is when the chemist Louis Pasteur performed experiments that resulted in scientists realising that molecules may occur in two forms that are mirror images of each other. So, why is only one type found in nature but not the other?

LOUIS PASTEUR IN HIS LABORATORY, PAINTING BY A. EDELFELDT IN 1885, WIKIMEDIA COMMONS

Nobel Prize in Chemistry 2026

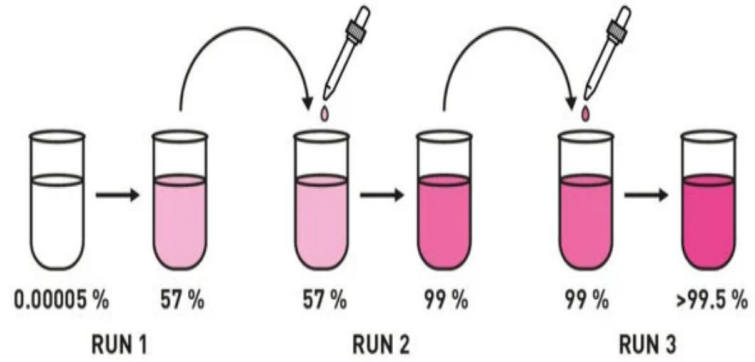
The Soai reaction

Soai designed a reaction that formed a chiral catalyst, which then formed itself. He was thus the first ever chemist to create chirality from a non-chiral combination of molecules.



Self-reinforcing process

In one experiment, Soai started with a small excess of one of the catalyst's enantiomers. This was enough for it to rapidly take over the reaction. After Soai added more substrate and reagent three times, the enantiomer comprised more than 99.5 per cent of the product.



The 2026 Nobel Prize laureates in physiology or medicine

”for their discoveries concerning light-gated ion channels and optogenetics”

“So much mystery now finally seems on the brink of elucidation”

Karl Deisseroth

Karl Deisseroth
Born: 1971, USA

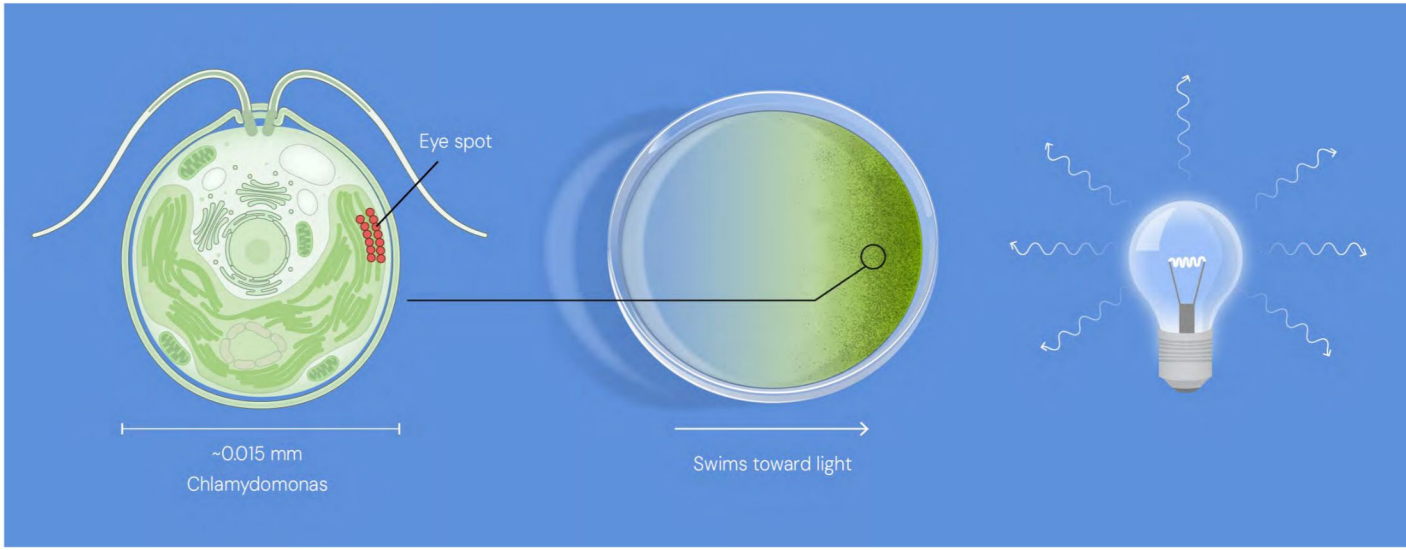
Peter Hegemann
Born: 1954, Germany

Georg Nagel
Born: 1953, Germany

ILL. NIKLAS ELMEHED © NOBEL PRIZE OUTREACH

A light-sensitive alga

The *Chlamydomonas* alga swims towards the light. It reacts very quickly to light. Could this be explained by a light-sensitive protein?



© THE NOBEL COMMITTEE FOR PHYSIOLOGY OR MEDICINE. ILL. MATTIAS KARLÉN

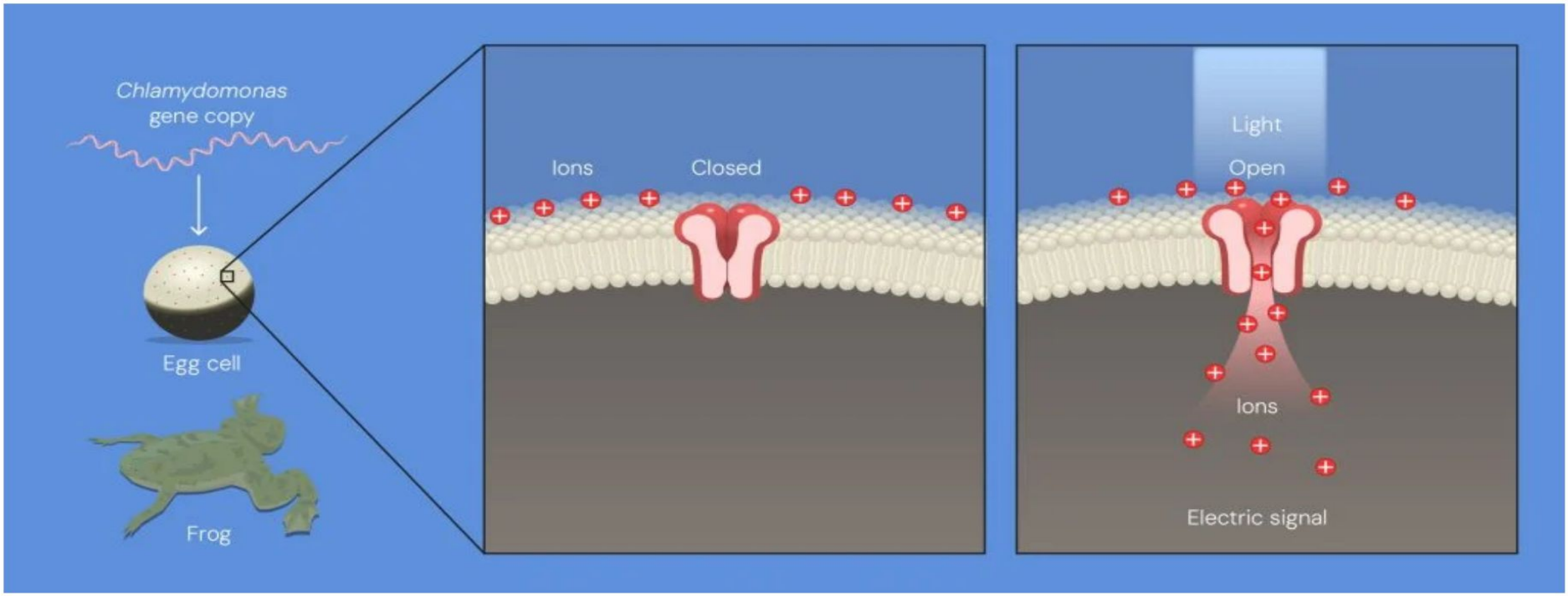
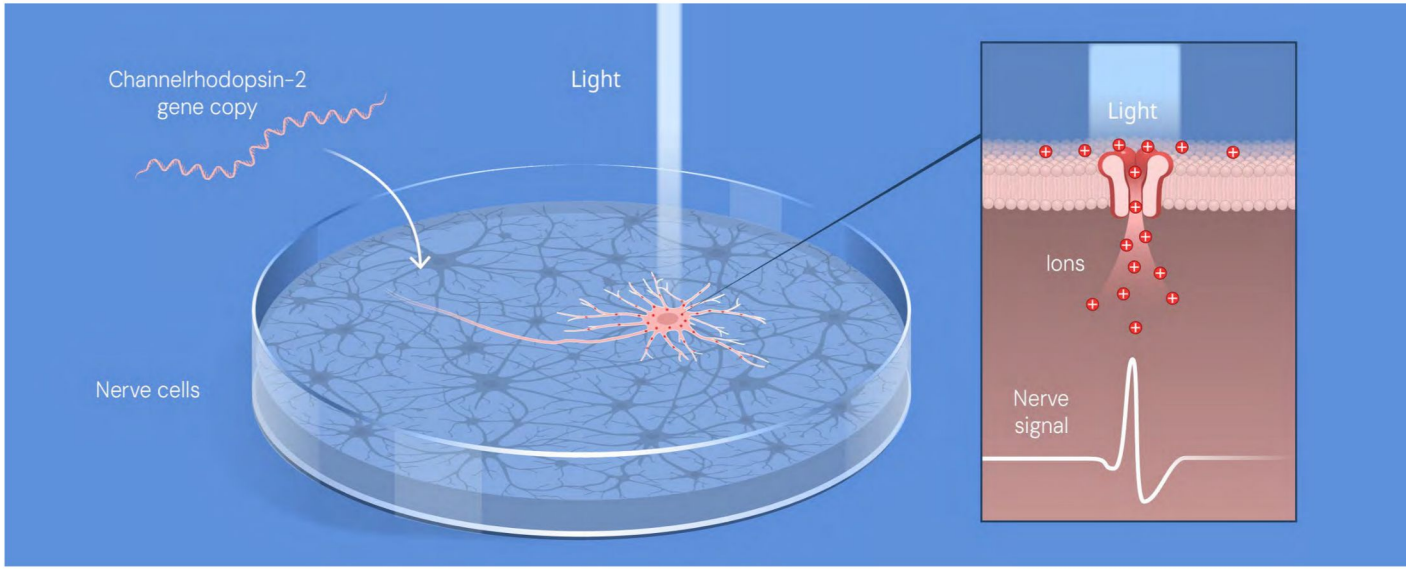


Figure 3. When Nagel and Hegemann introduced the Chlamydomonas genes in frog eggs, they began to produce ion channel proteins that ended up on the surface of the eggs. When the researchers illuminated the ion channels, they opened. Ions then flowed through the channels, creating an electric signal. © The Nobel Committee for Physiology or Medicine. Ill. Mattias Karlén

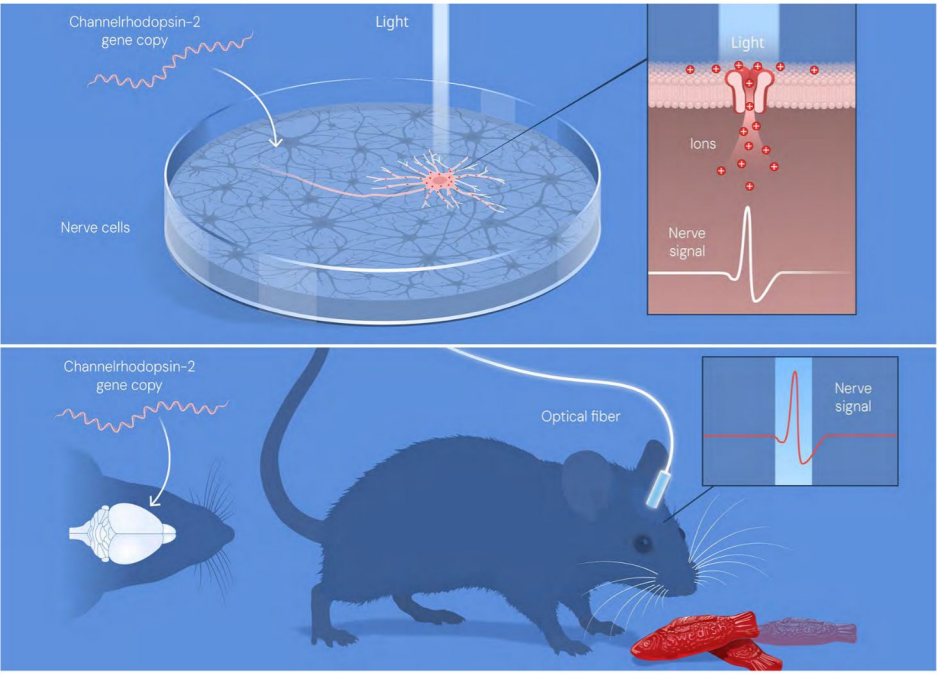
A light-sensitive channel in the cell wall

A protein was found to serve as a channel in the cell wall. Here, ions were able to pass through the wall. The ion channel was closed in darkness and open when exposed to light.



© THE NOBEL COMMITTEE FOR PHYSIOLOGY OR MEDICINE. ILL. MATTIAS KARLE

Ion channels in nerve cells



© THE NOBEL COMMITTEE FOR PHYSIOLOGY OR MEDICINE. ILL. MATTIAS KARLÉN

The gene for the light-sensitive ion channel could be introduced into nerve cells. This enabled researchers to control their activity by using light. This method, called optogenetics, can be used to study which nerve cells control different feelings and behaviours.

For the greatest benefit to humankind

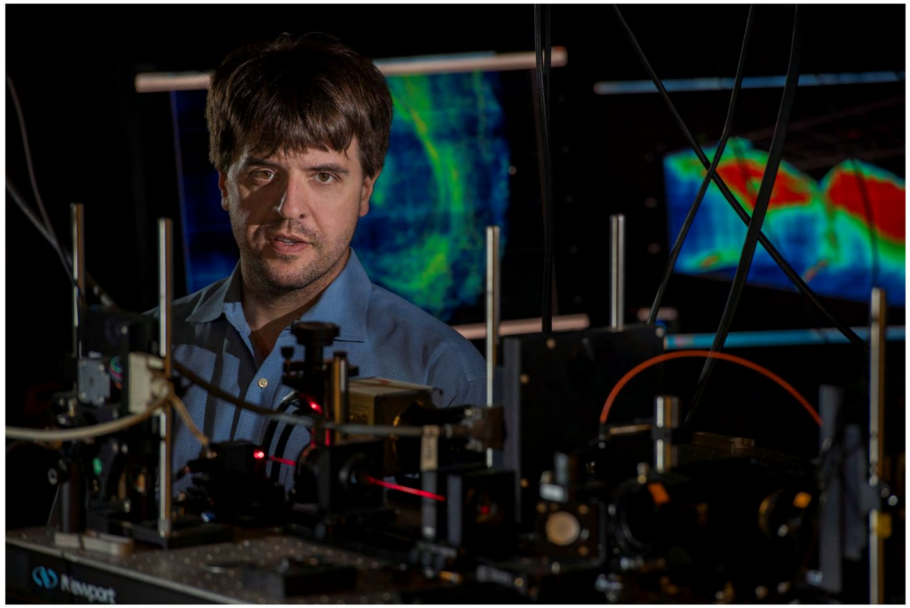


PHOTO: COURTESY STANFORD MEDICINE

Possible new treatment methods for psychiatric disorders and neurological diseases. The ability to restore sight for people who have become blind due to a disease. Designing improved hearing implants.

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

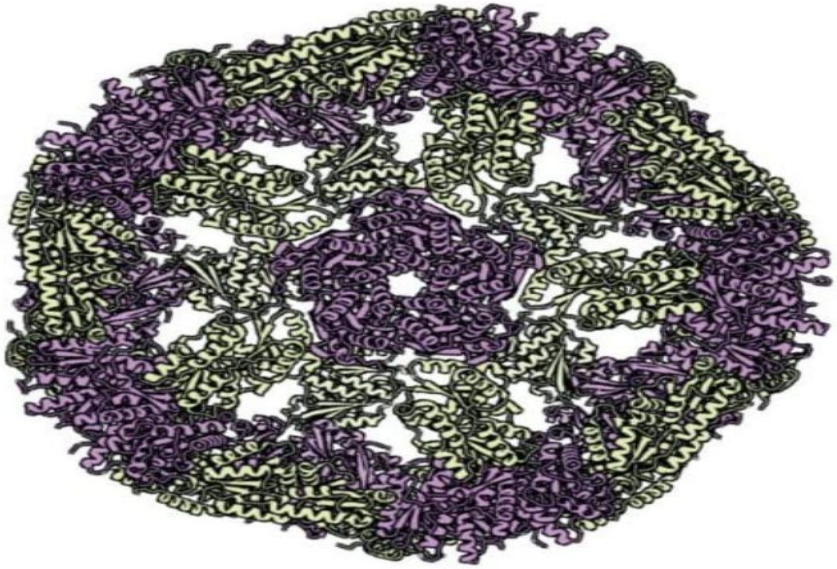
John Jumper

"for protein structure prediction"

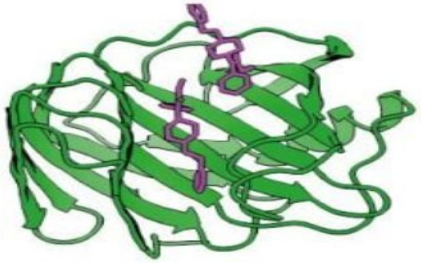
John Jumper. Ill. Niklas Elmehed © Nobel Prize Outreach

https://www.nobelprize.org/all-nobel-prizes-2024/

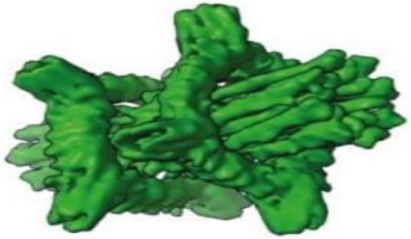
David Baker's designed proteins



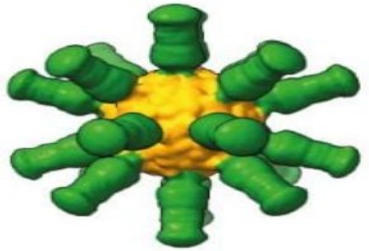
2016: New nanomaterials where up to 120 proteins spontaneously link together.



2017: Proteins that bind to an opioid called fentanyl (purple). These could be used to detect fentanyl in the environment.



2022: Proteins that function as a type of molecular rotor.



2021: Nanoparticles (yellow) with proteins imitating influenza virus on the surface (green) that can be used as a vaccine for influenza. Successful in animal models.

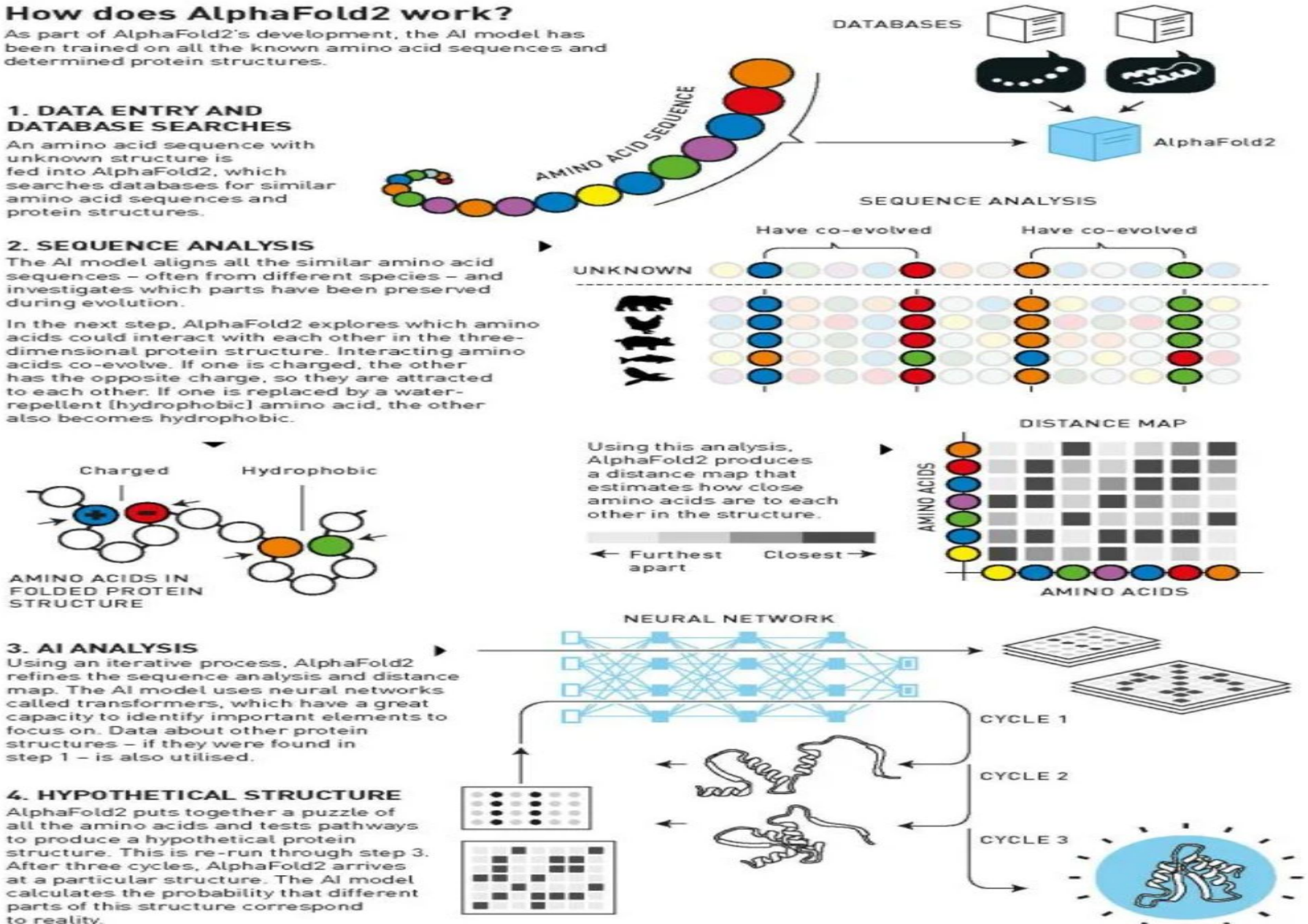
2024: Geometrically shaped proteins that can change their shape due to external influences. Could be used for producing tiny sensors.

Figure 4. Proteins developed using Baker's program Rosetta.

©Terezia Kovalova/The Royal Swedish Academy of Sciences

https://www.nobelprize.org/prizes/chemistry/2024/popular-information/

Finding the ‘Holy Grail’ of Bioinformatics



https://www.nobelprize.org/prizes/chemistry/2024/popular-information/

The Nobel Prize in Physiology or Medicine 2024

Victor Ambros

"for the discovery of microRNA and its role in post-transcriptional gene regulation"

Victor Ambros. Ill. Niklas Elmehed © Nobel Prize Outreach

Gary Ruvkun

"for the discovery of microRNA and its role in post-transcriptional gene regulation"

Gary Ruvkun. Ill. Niklas Elmehed © Nobel Prize Outreach

https://www.nobelprize.org/all-nobel-prizes-2024/

New type of gene regulation

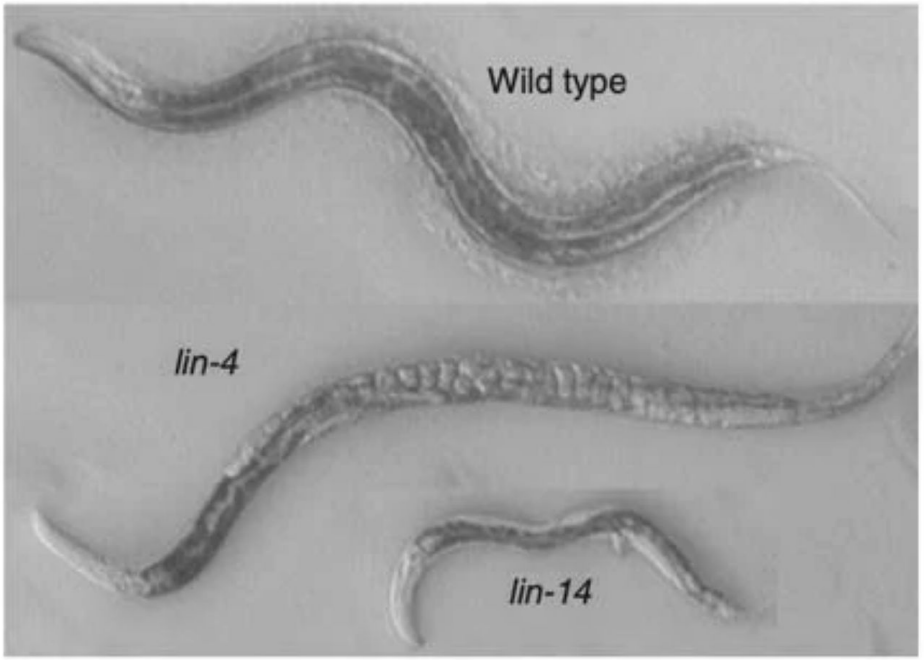


Figure 2. Heterochronic worm mutants with developmental defects. Nematode *lin-4* and *lin-14* mutants with disrupted animal development. Mutant *lin-4* worms reiterate developmental programs for cell lineages to accumulate internal eggs without forming a vulva, while *lin-14* mutants are small and lack larval development.

Worms adapted from (Ambros, 2008)

https://www.nobelprize.org/prizes/medicine/2024/advanced-information/

New type of gene regulation

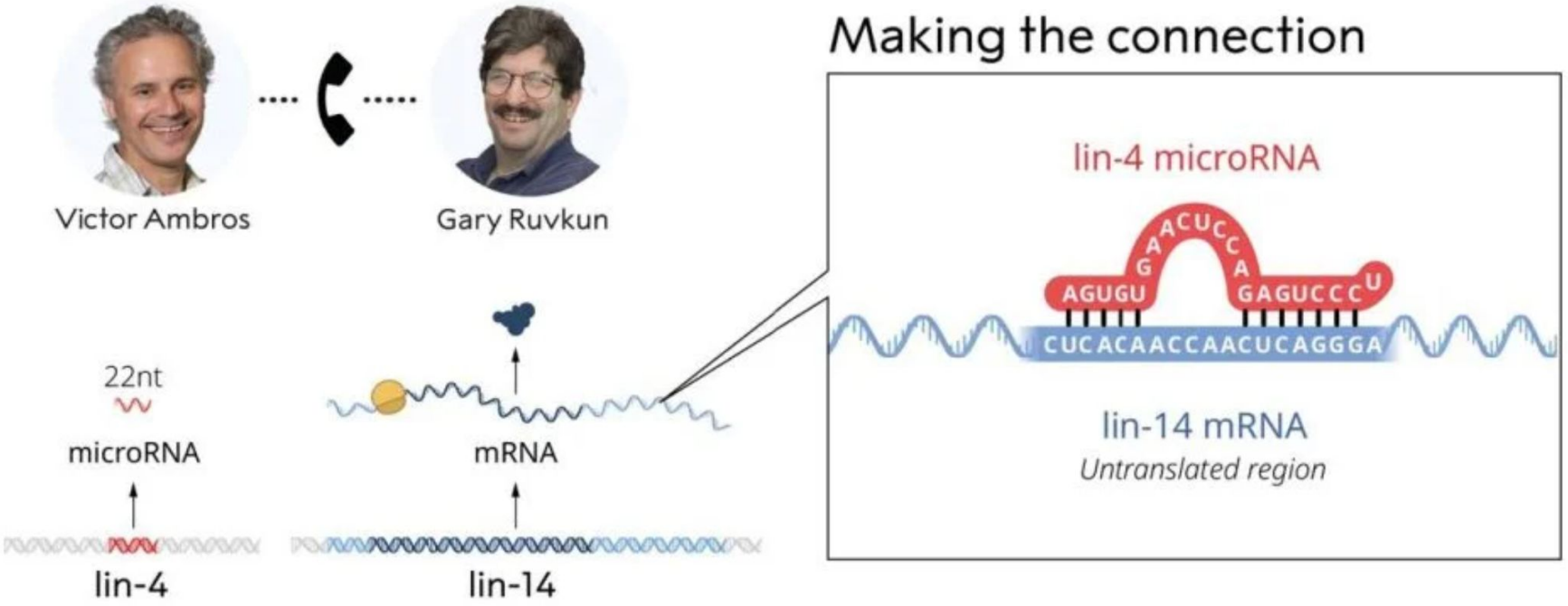


Figure 4. Complementary sequence elements in *lin-4* and *lin-14* RNA. Upon comparing cloned sequences for *lin-4* and *lin-14*, it was revealed that the short 22 nt *lin-4* RNA had partial complementarity to repeated elements in the *lin-14* 3'UTR. © The Nobel Committee for Physiology or Medicine. Ill. Mattias Karlén

https://www.nobelprize.org/prizes/medicine/2024/advanced-information/

The Nobel Prize in Physics 2024

John Hopfield

"for foundational discoveries and inventions that enable machine learning with artificial neural networks"

John Hopfield. Ill. Niklas Elmehed © Nobel Prize Outreach

Geoffrey Hinton

"for foundational discoveries and inventions that enable machine learning with artificial neural networks"

Geoffrey Hinton. Ill. Niklas Elmehed © Nobel Prize Outreach

https://www.nobelprize.org/all-nobel-prizes-2024

Nobel Prize in Physiology or Medicine 2025

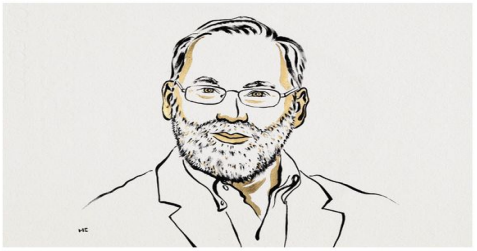
Mary E. Brunkow

"for their discoveries concerning peripheral immune tolerance"

Ill. Niklas Elmehed © Nobel Prize Outreach

Fred Ramsdell

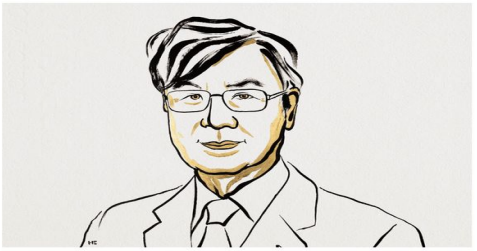
"for their discoveries concerning peripheral immune tolerance"



Ill. Niklas Elmehed © Nobel Prize Outreach

Shimon Sakaguchi

"for their discoveries concerning peripheral immune tolerance"

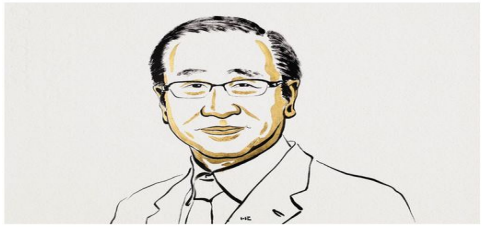


Ill. Niklas Elmehed © Nobel Prize Outreach

Nobel Prize in Chemistry 2025

Susumu Kitagawa

"for the development of metal-organic frameworks"



Ill. Niklas Elmehed © Nobel Prize Outreach

Richard Robson

"for the development of metal-organic frameworks"

Ill. Niklas Elmehed © Nobel Prize Outreach

Omar M. Yaghi

"for the development of metal-organic frameworks"

Ill. Niklas Elmehed © Nobel Prize Outreach

Nobel Prize in Physics 2025

John Clarke

"for the discovery of macroscopic quantum mechanical tunnelling and energy quantisation in an electric circuit"

Ill. Niklas Elmehed © Nobel Prize Outreach

Michel H. Devoret

"for the discovery of macroscopic quantum mechanical tunnelling and energy quantisation in an electric circuit"

Ill. Niklas Elmehed © Nobel Prize Outreach

John M. Martinis

"for the discovery of macroscopic quantum mechanical tunnelling and energy quantisation in an electric circuit"

Ill. Niklas Elmehed © Nobel Prize Outreach

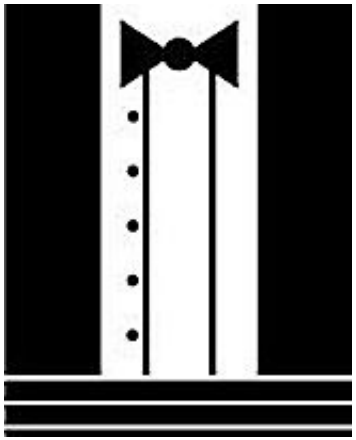
https://www.nobelprize.org/all-nobel-prizes-2025

NGS intro + Genome-Based Transcript Reconstruction and Analysis Using RNA-Seq Data

Based on material from: Brian Haas

Broad Institute

Martin Reczko



Source: https://trinityrnaseq.github.io/workshop/rnaseq_workshop.html

Overview

- Next generation sequencing (NGS) introduction
- Quality control
- Genome-based and genome-free (de-novo) transcript reconstruction from RNA-Seq
- Running the Tuxedo and Trinity software and visualizing the results.
- Principles of transcript abundance estimation
- Principles of differential expression analysis
- Single cell RNA-Seq basics

A quick history of sequencing

1995 – First bacterial genome – *H. influenzae* (1.8 Mb)

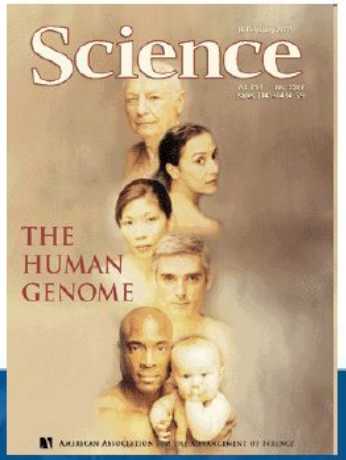
1998 – First animal genome – *C. elegans* (97 Mb)

2003 – Completion of Human Genome Project (3 Gb)

– 13 years, $2.7 bn

2005 – First “next-generation” sequencing instrument

2013– >10,000 genome sequences in NCBI database



Sanger sequencing: chain termination method

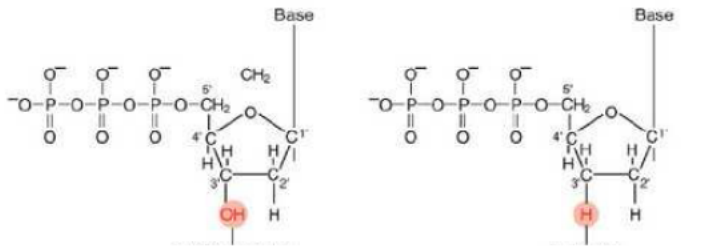
TCTGATGCAT*

TCTGATGCATGAACT*

TCTGATGCATGAACTGCT*

TCTGATGCATGAACTGCTCAT*

AGACTACGTACTTGACGAGTAC.....

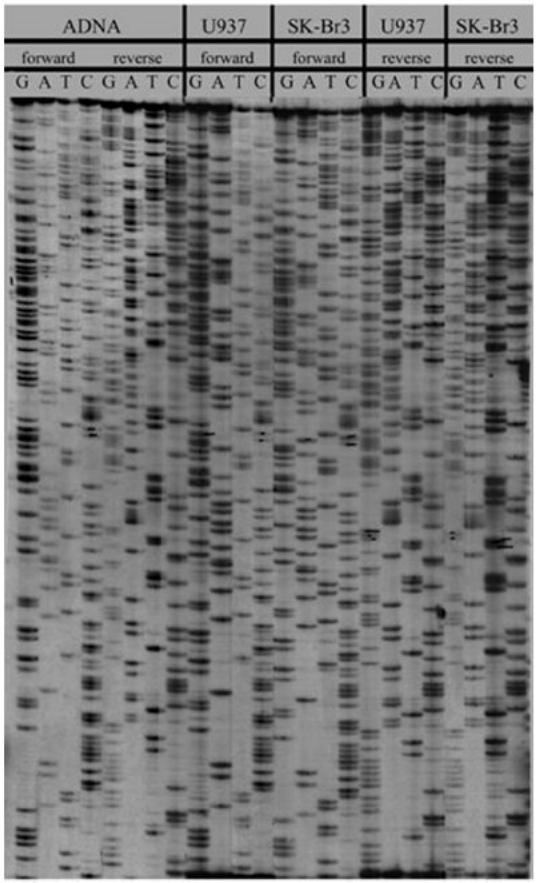


deoxynucleotide

dideoxynucleotide

Sanger sequencing: chain termination method

Separation of fragments by gel electrophoresis



Next-gen sequencing technologies

- Six main technologies
- All massively parallel sequencing
  - Sequencing by synthesis
  - Sequencing by ligation
- Mostly produce short reads- from <400bp
- Read numbers vary from ~ 1 million to ~ 1 billion per run

Next-gen sequencing technologies

Roche GS-FLX

Life Technologies SOLiD

Illumina HiSeq

Life Technologies Ion Torrent/Proton

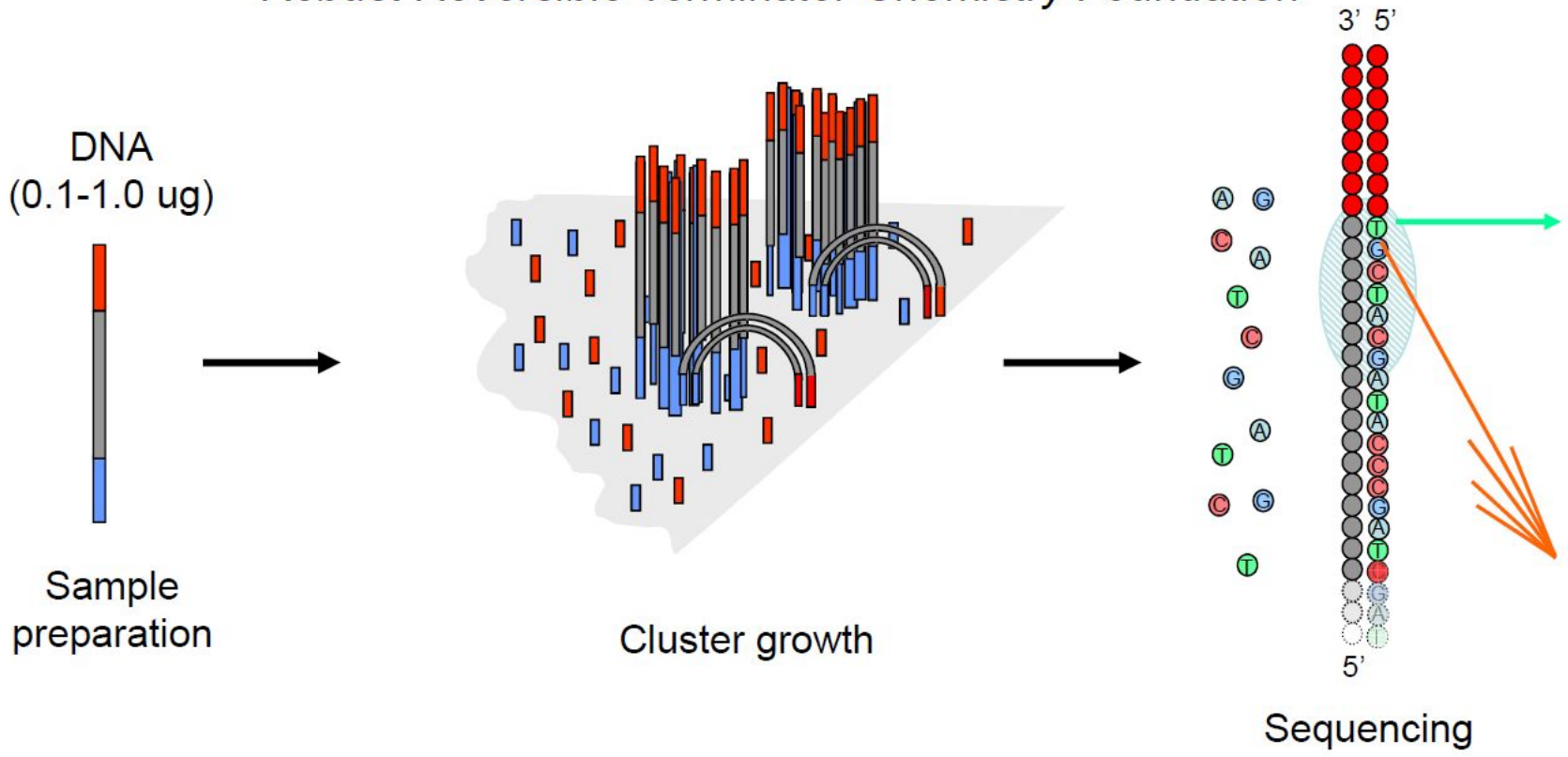
Oxford NANOPORE Technologies

PACBIO®

Illumina HiSeq

Illumina Sequencing Technology

Robust Reversible Terminator Chemistry Foundation



Platform Updates

Solexa 1G	• 18bp reads, ~1Gbp / run
Illumina GA	• 36bp reads ~3Gbp / run
Illumina GAII	• 75bp paired ends ~10Gbp / run (8 days)
Illumina GAIIx	• 75bp paired end reads ~40Gbp / run (8 days)
Illumina HiSeq 2000	• 100 bp paired end reads ~200 Gbp/ run (10 days)
Illumina HiSeq, v3 SBS	• 100bp paired end reads ~600Gbp / run (12 days)
Illumina HiSeq 2500 (Rapid)	• 150 bp paired end reads ~ 180 Gbp/ run (2 days)
MiSeq	• 250 bp paired end reads ~8 Gb/run (2 days)

Maximum yield / day 50,Gbp
~16x the human genome

Recent Platform Updates

NextSeq 550 Series

NextSeq 2000

NovaSeq 6000

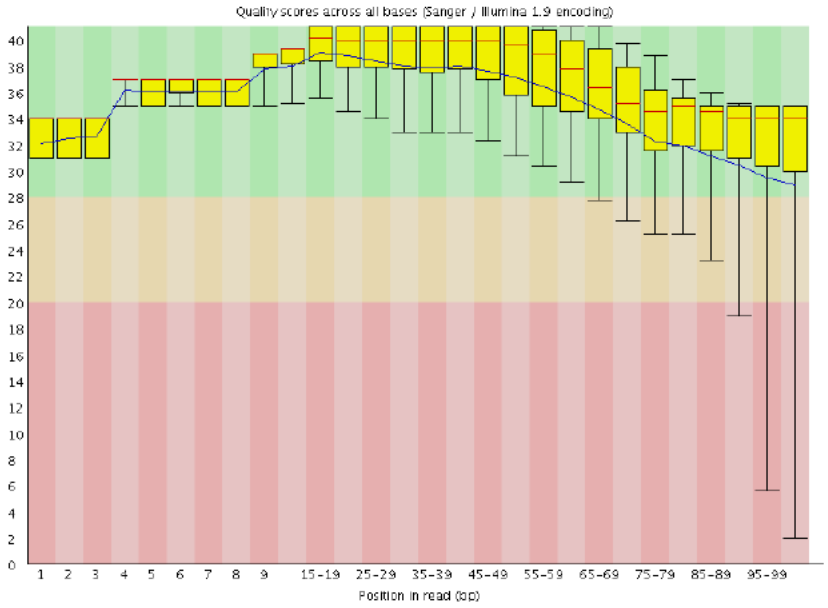
	NextSeq 550 Series	NextSeq 2000	NovaSeq 6000
Run Time	12–30 hours	24-48 hours	~13 - 38 hours (dual SP flow cells) ~13–25 hours (dual S1 flow cells) ~16–36 hours (dual S2 flow cells) ~44 hours (dual S4 flow cells)
Maximum Output	120 Gb	300 Gb*	6000 Gb
Maximum Reads Per Run	400 million	1 billion*	20 billion
Maximum Read Length	2 x 150 bp	2 x 150 bp	2 x 250**

MiSeq i100 Plus	NextSeq 2000 (P4)	NovaSeq X Plus
4-15 h	8-44 h	14-48 h
30 Gb (2×150)	540 Gb (2×150)	16-21 Tb (dual 25B flow cells, 2×150)
200 M PE	3.6 B PE	52-70 B PE
2×500 bp	2×300 bp	2×300 bp

Max. yield per day: 6000Gb = 200 human genomes

Illumina Sequencing Output

- *.fastq *(sequence and corresponding quality score encoded with an ASCII character, phred-like quality score + 33)*



Illumina fastq

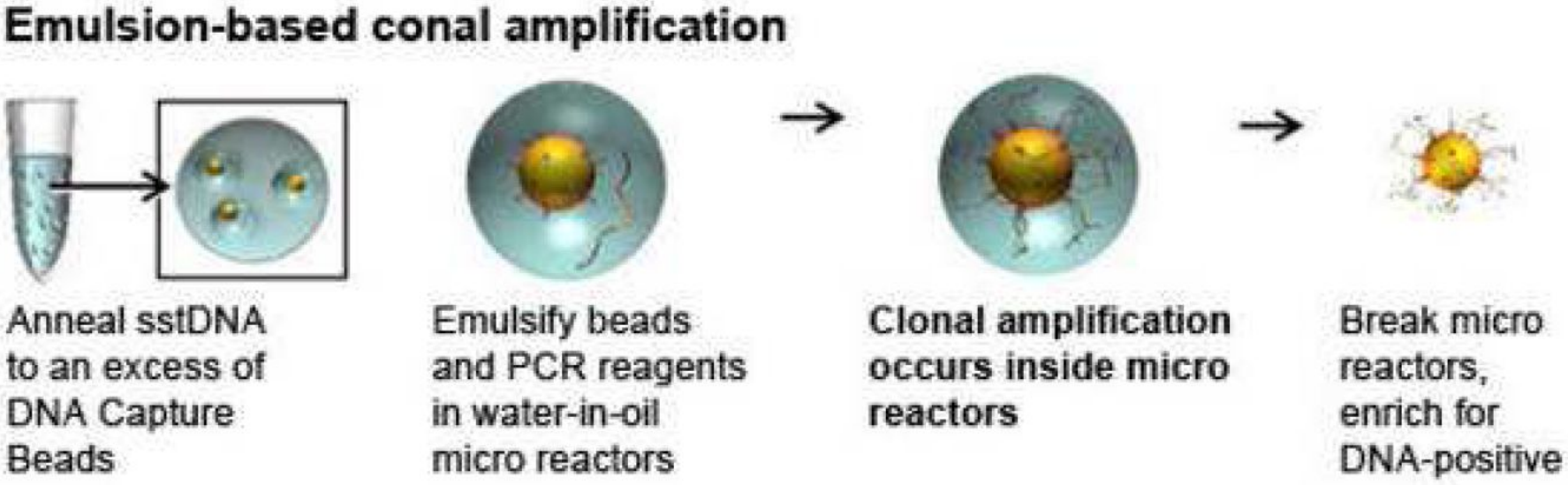
```
        1             2        3     4     5    6 7      8
@HWI-ST226:253:D14WFACXX:2:1101:2743:29814 1:N:0:ATCACG
TGCGGAAGGATCATTGTGGAATTCTCGGGTGCCAAGGAACTCCAGTCACATCACGATCTCGTATGCCGTCTTCTGCTT
GAAAAAAAAAAAAAAAAAATTA
+
B@CFFFFFHHFFHJIIGHIHIJJIJIIJJGDCHIIIJJJJJJGJGIHHEH@)=F@EIGHHEHFFFFDCBBD:@CC@C
:<CDDDD50559<B########
```

1. unique instrument ID and run ID
2. Flow cell ID and lane
3. tile number within the flow cell lane
4. 'x'-coordinate of the cluster within the tile
5. 'y'-coordinate of the cluster within the tile
6. the member of a pair, /1 or /2 *(paired-end or mate-pair reads only)*
7. N if the read passes filter, Y if read fails filter otherwise
8. Index sequence

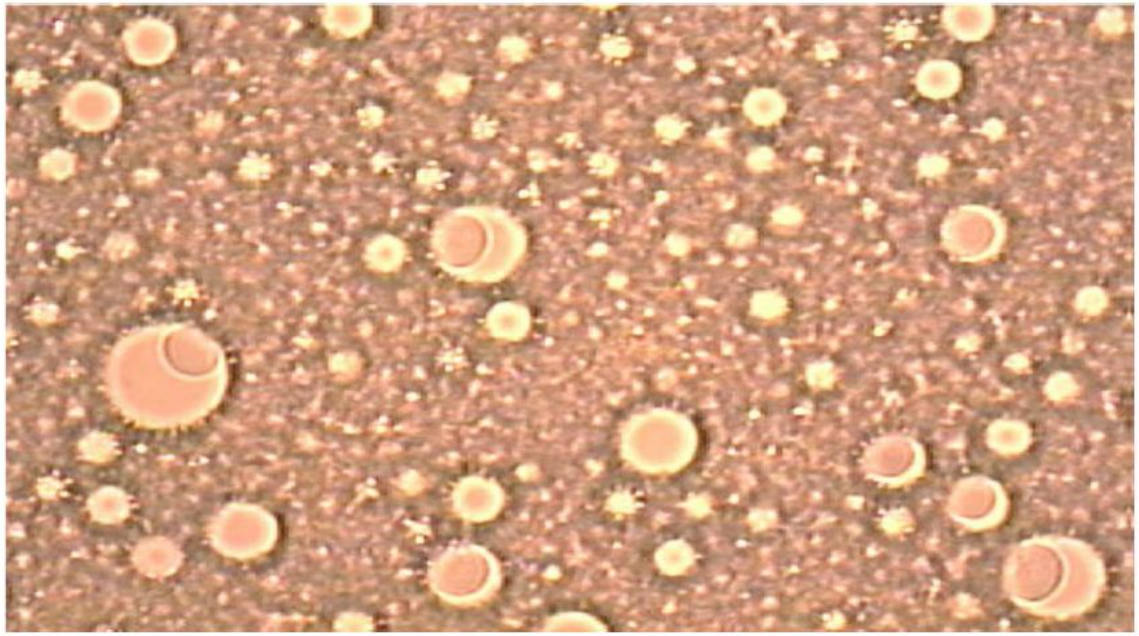
Applied Biosystems SOLiD

emPCR

Emulsion PCR is a method of clonal amplification which allows for millions of unique PCRs to be performed at once through the generation of micro-reactors.

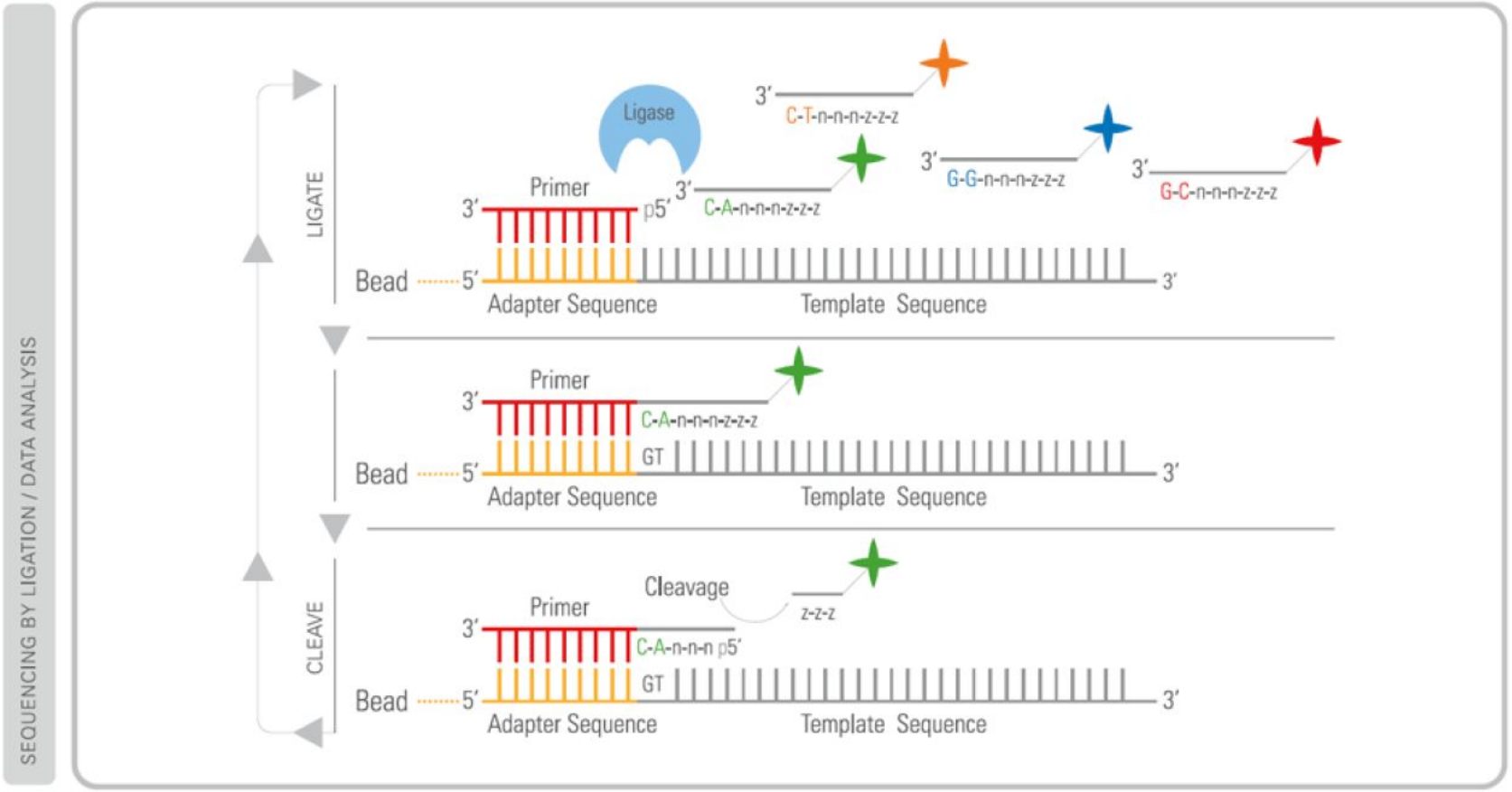


emPCR



The Water-in-Oil-Emulsion

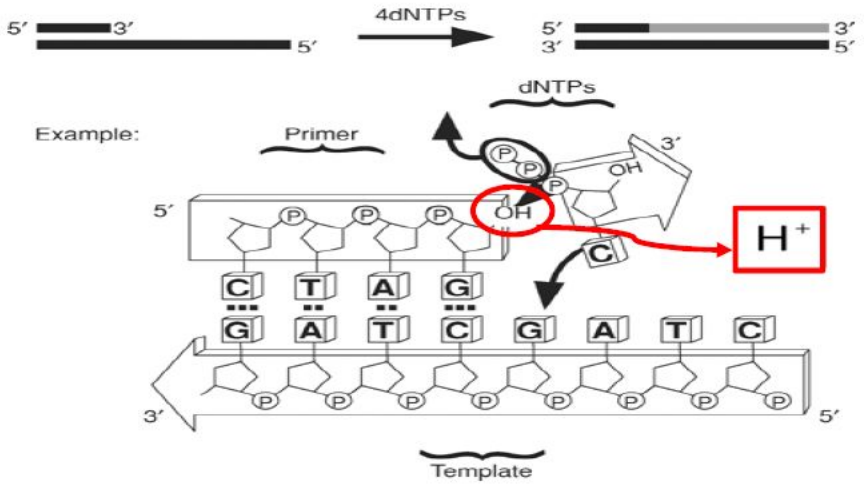
Sequencing by Ligation



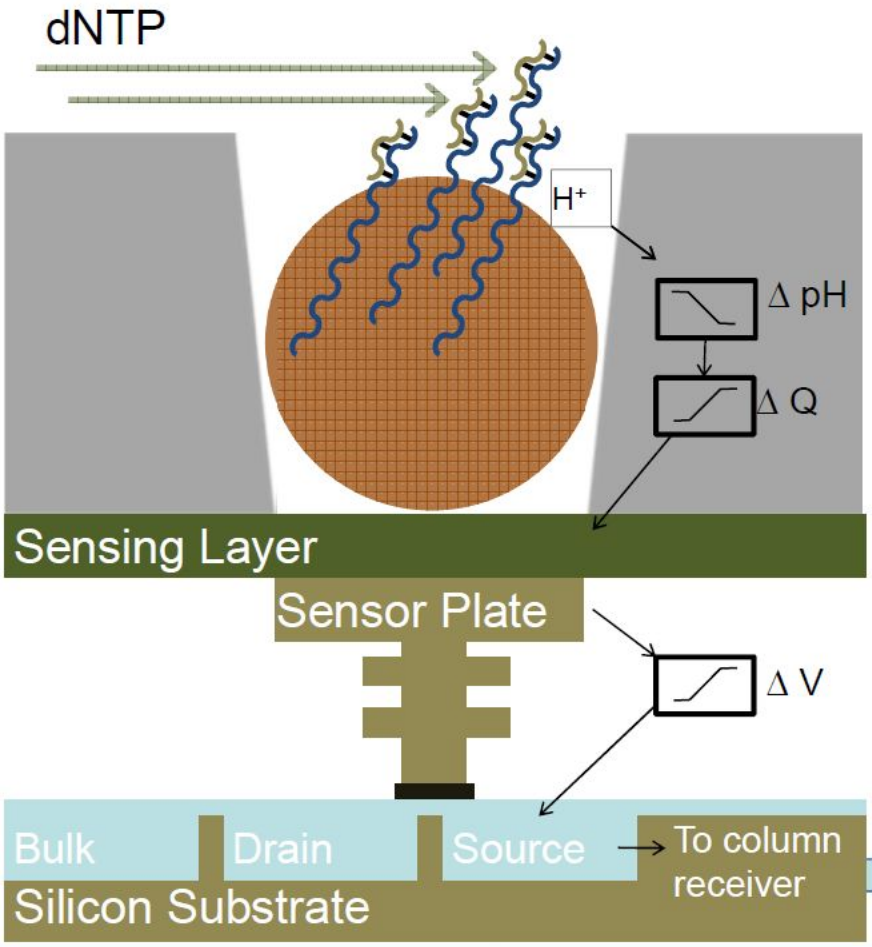
Applied Biosystems: Ion Torrent PGM

Ion Torrent

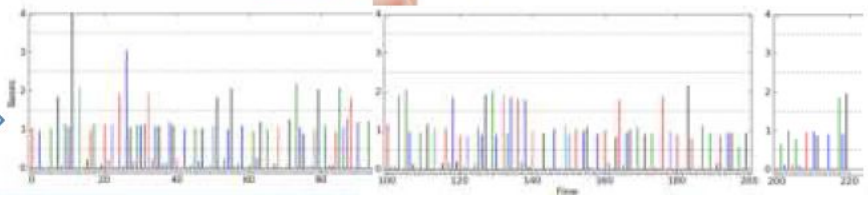
- Ion Semiconductor Sequencing
- Detection of hydrogen ions during the polymerization DNA
- Sequencing occurs in microwells with ion sensors
- No modified nucleotides
- No optics



Ion Torrent



- DNA → Ions → Sequence
  - Nucleotides flow sequentially over Ion semiconductor chip
  - One sensor per well per sequencing reaction
  - Direct detection of natural DNA extension
  - Millions of sequencing reactions per chip
  - Fast cycle time, real time detection



Ion Torrent: System Updates

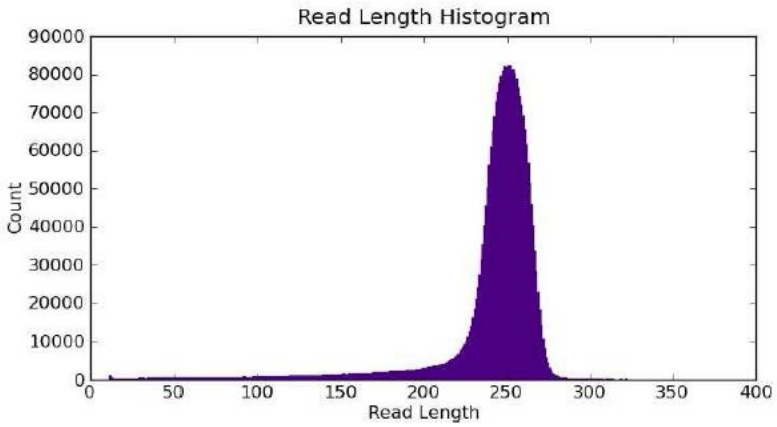
314 Chip	•100bp reads ~10 Mb/run (1.5 hrs)
316 Chip	•100 bp reads ~100 Mbp / run (2 hrs) •200 bp reads ~200 Mbp/run (3 hrs)
318 Chip	•200 bp reads ~1 Gbp / run (4.5 hrs)
P1 Chip	•100 bp reads ~8 Gbp/run
Ion GeneStudio S5 Ion 540 Chip	• 15 Gbp/run

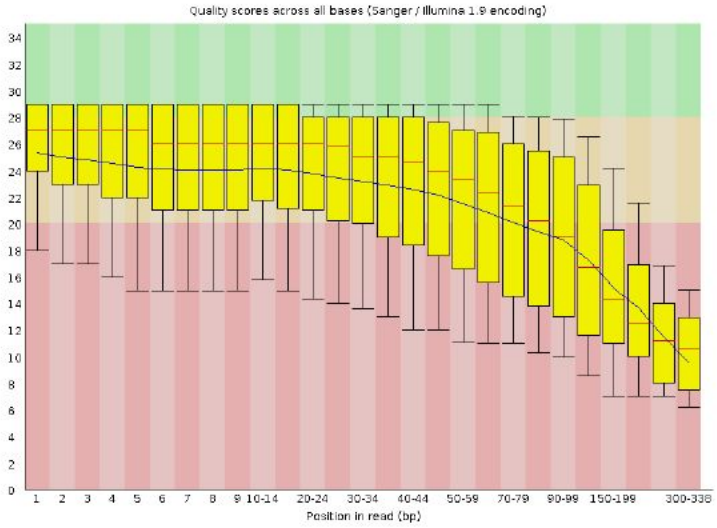
Ion Torrent: System Updates

Platform	Chip	Reads / Run	Output / Run	Read Length	Turnaround
Ion GeneStudio S5	510	2-3 M	0.3-1 Gb	200-400 bp	3-5 h
	520	4-6 M	0.6-2 Gb	200-400 bp	3-6 h
	530	15-20 M	3-8 Gb	200-400 bp	4-8 h
	540	60-80 M	10-15 Gb	200 bp	6.5-10 h
Ion GeneStudio S5 Prime	550	100-130 M	20-25 Gb	200 bp	8.5 h
Ion GeneStudio S5 Prime / Plus	560	240-280 M	45-50 Gb	200 bp	8.5-16 h
Genexus Integrated Sequencer	GX5	48-60 M	~6 Gb	200-400 bp	<24 h specimen-to-report

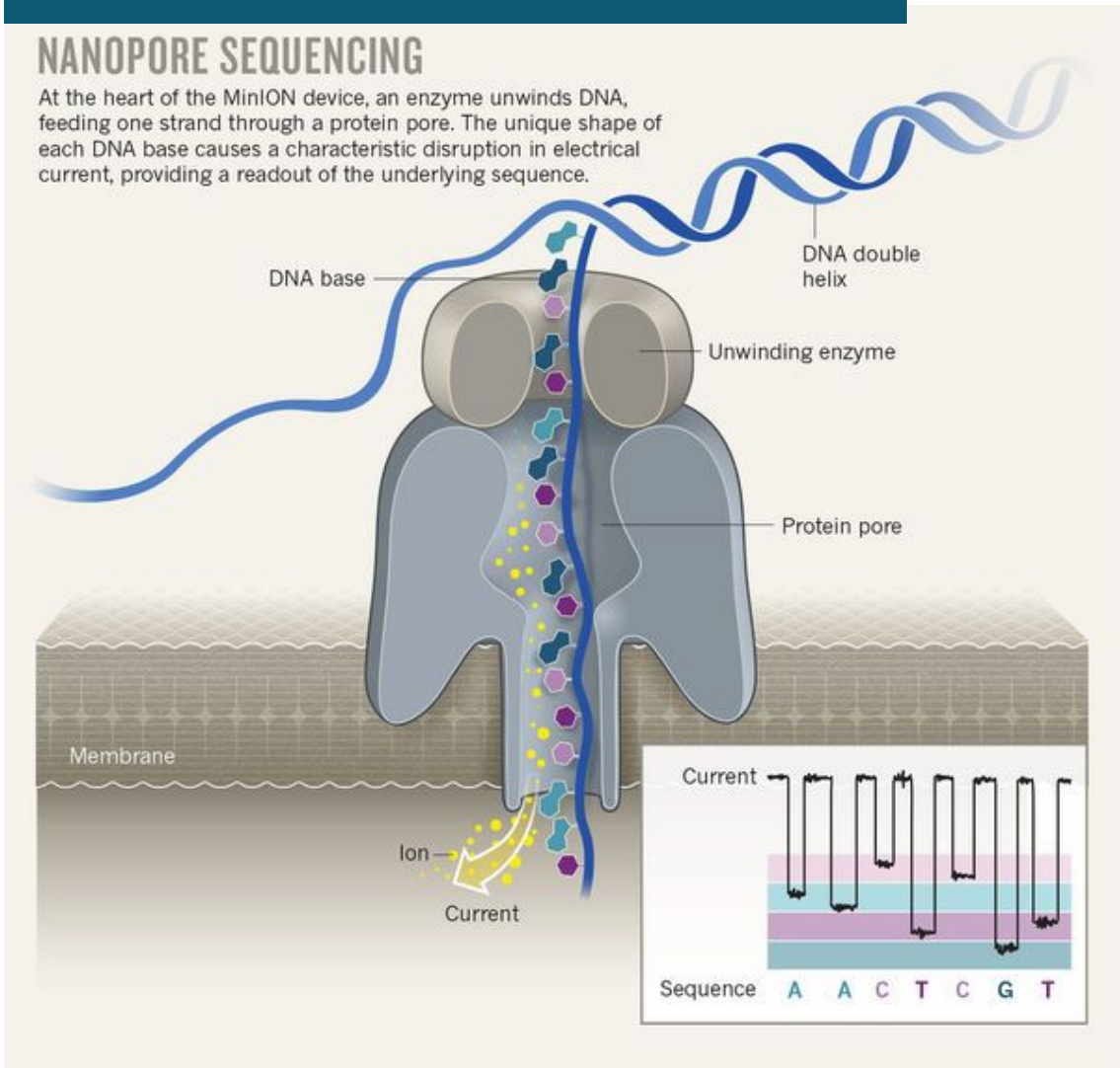
Ion Torrent Reads

- *.sff *(standard flowgram format)*
- *.fastq (*sequence and corresponding quality score encoded with an ASCII character, phred-like quality score + 33)*





Oxford NANOPORE Technologies



http://blogs.nature.com/blog/tag/oxford-nanopore-technologies/

‘Handheld’ sequencing

Figure 2: In-field surveillance of Zika virus
The advent of highly portable sequencing devices has enabled low-cost disease surveillance and characterisation at point of infection, providing faster access to informative results. Device shown: MinION™ Mk1B from Oxford Nanopore Technologies. Image courtesy of Professor Nuno Faria, University of Oxford, UK.

Efficient viral monitoring

Sequence SARS-CoV-2 genomes rapidly:
From RNA to answer in as little as

7 h 15 m

Scale to your needs: from MinION to PromethION, from

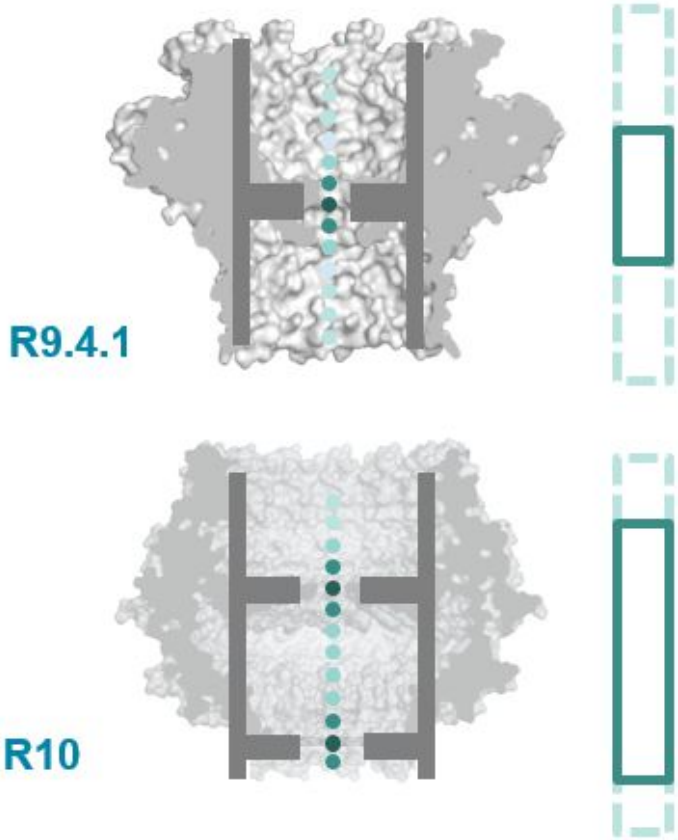
12 to 1,000+ samples

in a single sequencing run, with costs as low as

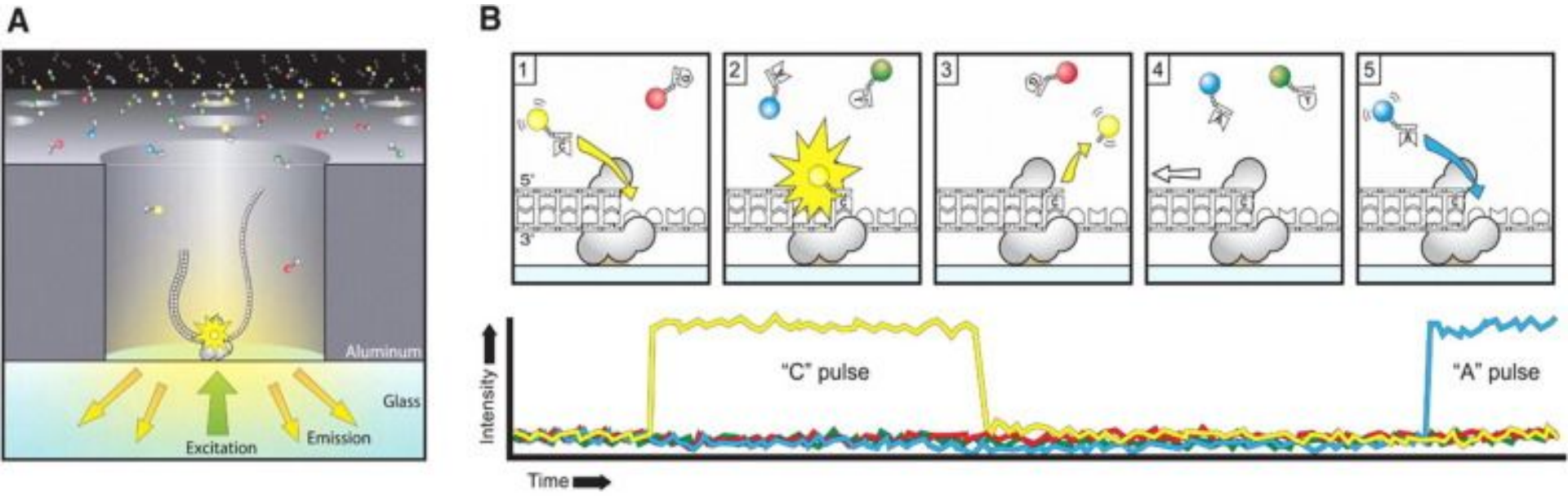
$9.55

per sample.

Accuracy with dual heads



PacBio sequencing



A. A SMRTbell (gray) diffuses into a Zero-Mode Waveguide (ZMW), and the adaptor binds to a polymerase immobilized at the bottom. **B.** Each of the four nucleotides is labeled with a different fluorescent dye (indicated in red, yellow, green, and blue, respectively for G, C, T, and A) so that they have distinct emission spectrums. As a nucleotide is held in the detection volume by the polymerase, a light pulse is produced that identifies the base. (1) A fluorescently-labeled nucleotide associates with the template in the active site of the polymerase. (2) The fluorescence output of the color corresponding to the incorporated base (yellow for base C as an example here) is elevated. (3) The dye-linker-pyrophosphate product is cleaved from the nucleotide and diffuses out of the ZMW, ending the fluorescence pulse. (4) The polymerase translocates to the next position. (5) The next nucleotide associates with the template in the active site of the polymerase, initiating the next fluorescence pulse, which corresponds to base A here.

https://www.sciencedirect.com/science/article/pii/S1672022915001345#f0015

DNA nanoball sequencing (DNBSEQ)

Comparison of various BGI NGS instruments [38].

Methods/applications	DNBSEQ-T7	DNBSEQ-G400 FAST	DNBSEQ-G400	DNBSEQ-G50
Major applications	WGS, DES, EGS, TS	WGS, WES, TS, MGS, RNA-seq	WGS, WES	Targeted sequencing (DNA & RNA), pathogen identification, and SPS
Max. run time (hours)	30	13	37	40
Maximum output	6 Tb	330 Gb	1440 Gb	150 Gb
Maximum reads per run	5000 million	550 million	1800 million	770 million
Maximum read length	150 PE	150 PE	200 PE/400 SE	150 PE
Data quality	> 85% > Q30	> 85% > Q30	> 85% > Q30	> 85% > Q30

https://www.hindawi.com/journals/bmri/2022/3457806/tab4/

Tutorial at https://www.youtube.com/watch?v=CAZwdtORXMw

DNA nanoball sequencing (DNBSEQ)

	DNBSEQ-G99	DNBSEQ-G400	DNBSEQ-T7+	DNBSEQ-T20×2
Segment	Benchtop	Core facility	Ultra-high-throughput	Population genomics factory
Applications	Targeted seq., RNA-seq, clinical research	WGS, WES, RNA-seq, single-cell	Large-scale multiomics, population genomics	National-scale genomics
Max output/run	8-96 Gb	55-1440 Gb	>14 Tb/day	48-72 Tb/run
Reads/run	up to 80 M/flowcell	up to 1.8 B/flowcell	up to 48 B	40 B per slide × 6 slides
Read length	PE150	PE150 / PE200	PE150	PE100 / PE150
Run time	hours	up to ~2 days	~24 h	60-80 h
Data quality	>85% Q30	>85% Q30	Q40-class performance	>85% Q30

Comparison of sequencing technologies

Method	Generation	Read length (bp)	Single pass error rate (%)	No. of reads per run	Time per run	Cost per million bases (USD)	Refs.
Sanger ABI 3730×l	1st	600–1000	0.001	96	0.5–3 h	500	[14], [18], [19], [20], [21]
Ion Torrent	2nd	200	1	8.2×10^7	2–4 h	0.1	[15], [25]
454 (Roche) GS FLX+	2nd	700	1	1×10^6	23 h	8.57	[14], [17], [27]
Illumina HiSeq 2500 (High Output)	2nd	2 × 125	0.1	8×10^9 (paired)	7–60 h	0.03	[9], [16], [26]
Illumina HiSeq 2500 (Rapid Run)	2nd	2 × 250	0.1	1.2×10^9 (paired)	1–6 days	0.04	[9], [16], [26]
SOLiD 5500×l	2nd	2 × 60	5	8×10^8	6 days	0.11	[14], [24]
PacBio RS II: P6-C4	3rd	$1.0–1.5 \times 10^4$ on average	13	$3.5–7.5 \times 10^4$	0.5–4 h	0.40–0.80	[5], [12], [15]
Oxford Nanopore MinION	3rd	$2–5 \times 10^3$ on average	38	$1.1–4.7 \times 10^4$	50 h	6.44–17.90	[22], [23]

Comparison of sequencing technologies

Technology	Representative Platform	Typical Read Length	Accuracy	Output / Run	Approx. Cost / Million Bases (USD)
Sanger	ABI 3730xl	700-1000 bp	>99.99%	96 reactions	**$100-500**
Ion Torrent	Genexus / S5 Prime	200-400 bp	~99%	up to 50 Gb	**$0.05-0.20**
Illumina SBS	NovaSeq X Plus	2×150 bp	>99.9%	up to 21 Tb	**$0.0005-0.005**
MGI DNBSEQ	DNBSEQ-T20×2	PE150	>99.9%	up to 72 Tb	**$0.0003-0.004**
PacBio HiFi	Revio	15-25 kb	99.8-99.9%	>100 Gb / SMRT Cell	**$0.02-0.10**
Oxford Nanopore	PromethION 48	10 kb-1 Mb+	>99% duplex	multi-Tb	**$0.005-0.05**

Comparison of sequencing technologies

Comparison of various high-performing sequencing instruments*.

Manufacturer	Read length	Data output	Max. run time (hours)	Chemistry	Key applications**
Illumina (NovaSeq 6000)	300 PE	6 Tb (6000 Gb)	44	Sequencing by synthesis	SS-WGS and TGS, TGEP, 16sMGS, WES, SCP, LS-WGS, CA, MS, MGP, CFS, LBA
Thermo Fisher Scientific Ion Torrent (Ion GeneStudio S5 Prime)	600 SE	50 Gb	12	Sequencing by synthesis	WGS, WES, TGS
GenapSys (16 chips)	150 SE	2 Gb	24	Sequencing by synthesis	TS, SS-WGS, GEV, 16S rRNA sequencing, sRNA sequencing, TSCAS
QIAGEN (GeneReader)	100 SE	Not available	Not available	Sequencing by synthesis	Cancer research and identifying mutations
BGI/Complete Genomics	400 SE	6 Tb (6000 Gb)	40	DNA nanoball	Small and large WGS, WES and TGS
PacBio (HiFi Reads)	25 Kb	66.5 Gb	30	Real-time sequencing	DN sequencing, FT, identifying ASI, mutations, and EPM
Nanopore (PromethION)	4 Mb	14 Tb (14000 Gb)	72	Real-time sequencing	SV, GS, phasing, DNA and RNA base modifications, FT, and isoform detection

*Performance comparison is given as per manufacturer's description. **Applications by all sequencers of the respective manufacturer are listed. **Full names are given in Abbreviations.

https://www.hindawi.com/journals/bmri/2022/3457806/tab7/

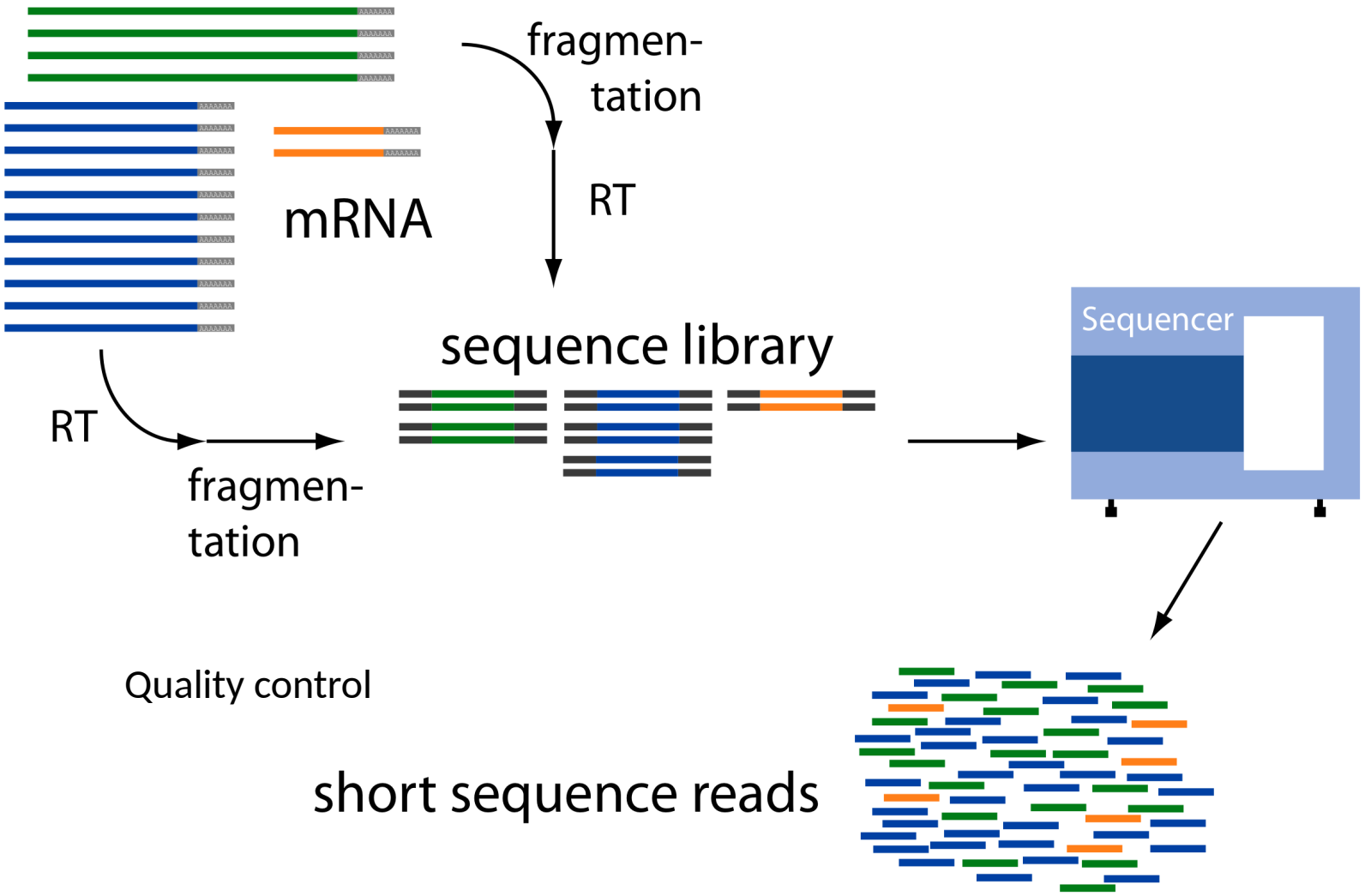
Comparison of sequencing generations

Advantages and disadvantages of sequencing generations.

Sequencing generation	Advantages	Disadvantages
First generation	High accuracy Helps in validating findings of NGS	High cost Low throughput
Second generation	High throughput Low cost Have clinical applications Short run time	Short read length Difficult sample preparation PCR amplification Long run time
Third generation	No PCR amplification Require less starting material Longer read lengths Very low cost Low error rate during library preparation	High sequencing error rate 10–15% in the PacBio and 5–20% in the ONT Fresh DNA requires for ensuring quality of ultralong reads
Fourth generation	Advantages of 3rdGS+ Ultrafast: scan of whole genome in 15 minutes Spatial distribution of the sequencing reads over the sample can be seen	Database systems and algorithms/tools are rare for analyzing 3rd and 4th GS data

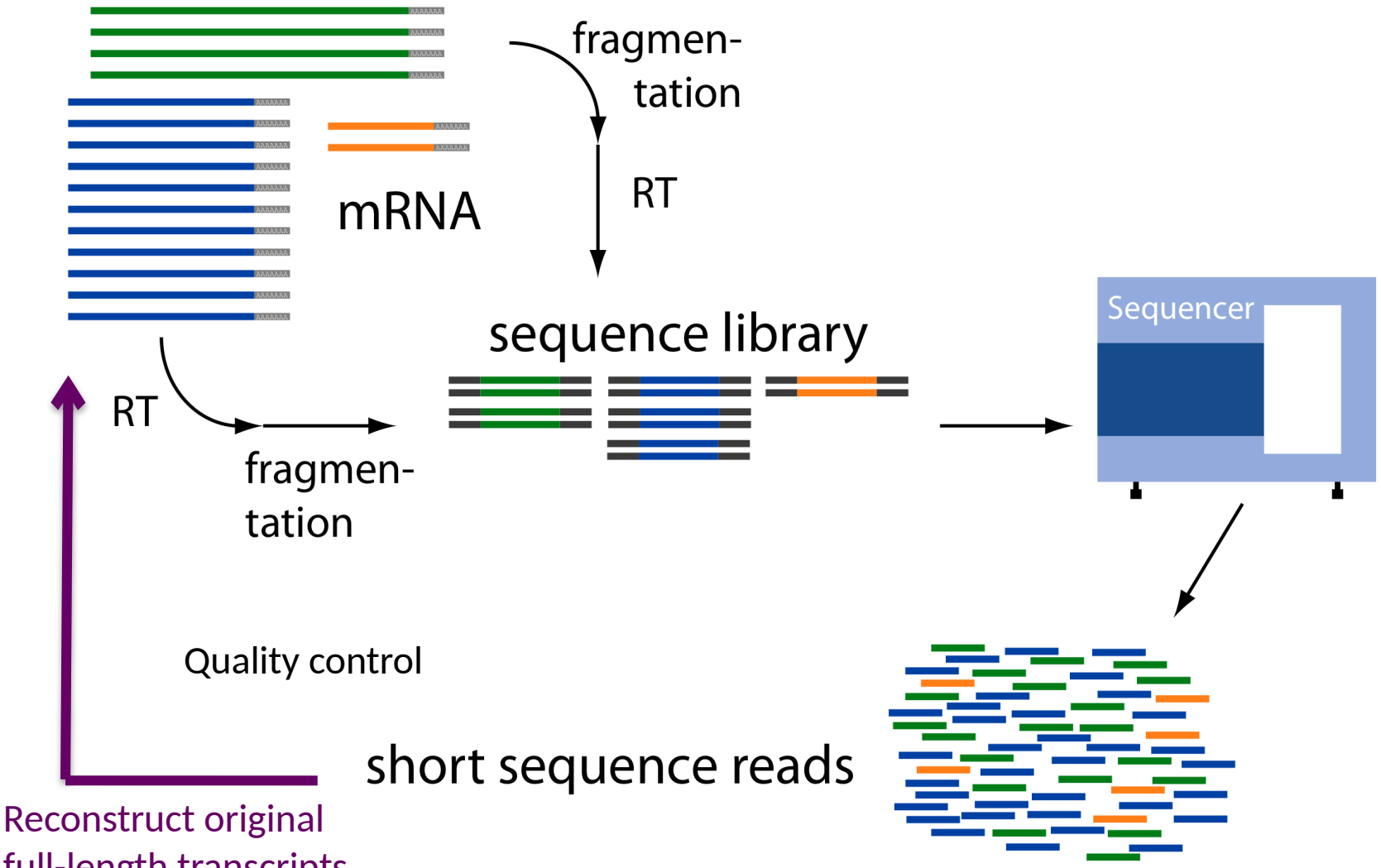
https://www.hindawi.com/journals/bmri/2022/3457806/tab8/

Overview of RNA-Seq



From: http://www2.fml.tuebingen.mpg.de/raetsch/members/research/transcriptomics.html

Overview of RNA-Seq



From: http://www2.fml.tuebingen.mpg.de/raetsch/members/research/transcriptomics.html

Common Data Formats for RNA-Seq

FASTA format:

```
>61DFRAAXX100204:1:100:10494:3070/1
AAACAACAGGGCACATTGTCACTCTTGTATTTGAAAAACACTTTCCGGCCAT
```

FASTQ format:

```
@61DFRAAXX100204:1:100:10494:3070/1
AAACAACAGGGCACATTGTCACTCTTGTATTTGAAAAACACTTTCCGGCCAT
+
ACCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCBC?CCCCCCCCC@@CACCCCCA
```

Common Data Formats for RNA-Seq

FASTA format:

```
>61DFRAAXX100204:1:100:10494:3070/1
AAACAACAGGGCACATTGTCACTCTTGTATTTGAAAAACACTTTCCGGCCAT
```

FASTQ format:

```
@61DFRAAXX100204:1:100:10494:3070/1
AAACAACAGGGCACATTGTCACTCTTGTATTTGAAAAACACTTTCCGGCCAT
+
ACCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCBC?CCCCCCCCC@@CACCCCCA
```

Read

Quality values

Common Data Formats for RNA-Seq

FASTA format:

```
>61DFRAAXX100204:1:100:10494:3070/1
AAACAACAGGGCACATTGTCACTCTTGTATTTGAAAAACACTTTCCGGCCAT
```

FASTQ format:

```
@61DFRAAXX100204:1:100:10494:3070/1
AAACAACAGGGCACATTGTCACTCTTGTATTTGAAAAACACTTTCCGGCCAT
+
ACCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCBC?CCCCCCCCC@@CACCCCCA
```

Read

Quality values

AsciiEncodedQual(x) = -10 * log10(Pwrong(x)) + 33

Common Data Formats for RNA-Seq

FASTA format:

```
>61DFRAAXX100204:1:100:10494:3070/1
AAACAACAGGGCACATTGTCACTCTTGTATTTGAAAAACACTTTCCGGCCAT
```

FASTQ format:

```
@61DFRAAXX100204:1:100:10494:3070/1
AAACAACAGGGCACATTGTCACTCTTGTATTTGAAAAACACTTTCCGGCCAT
+
ACCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCBC?CCCCCCCCC@@CACCCCCA
```

Read

Quality values

AsciiEncodedQual(x) = -10 * log10(Pwrong(x)) + 33

AsciiEncodedQual (‘C’) = **64**

Common Data Formats for RNA-Seq

FASTA format:

```
>61DFRAAXX100204:1:100:10494:3070/1
AAACAACAGGGCACATTGTCACTCTTGTATTTGAAAAACACTTTCCGGCCAT
```

FASTQ format:

```
@61DFRAAXX100204:1:100:10494:3070/1
AAACAACAGGGCACATTGTCACTCTTGTATTTGAAAAACACTTTCCGGCCAT
+
ACCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCBC?CCCCCCCCC@@CACCCCCA
```

Read

Quality values

AsciiEncodedQual(x) = -10 * log10(Pwrong(x)) + 33

AsciiEncodedQual (‘C’) = **64**

So, Pwrong(‘C’) = 10^((**64**-33/ (-10)) = 10^-3.4 = **0.0004**

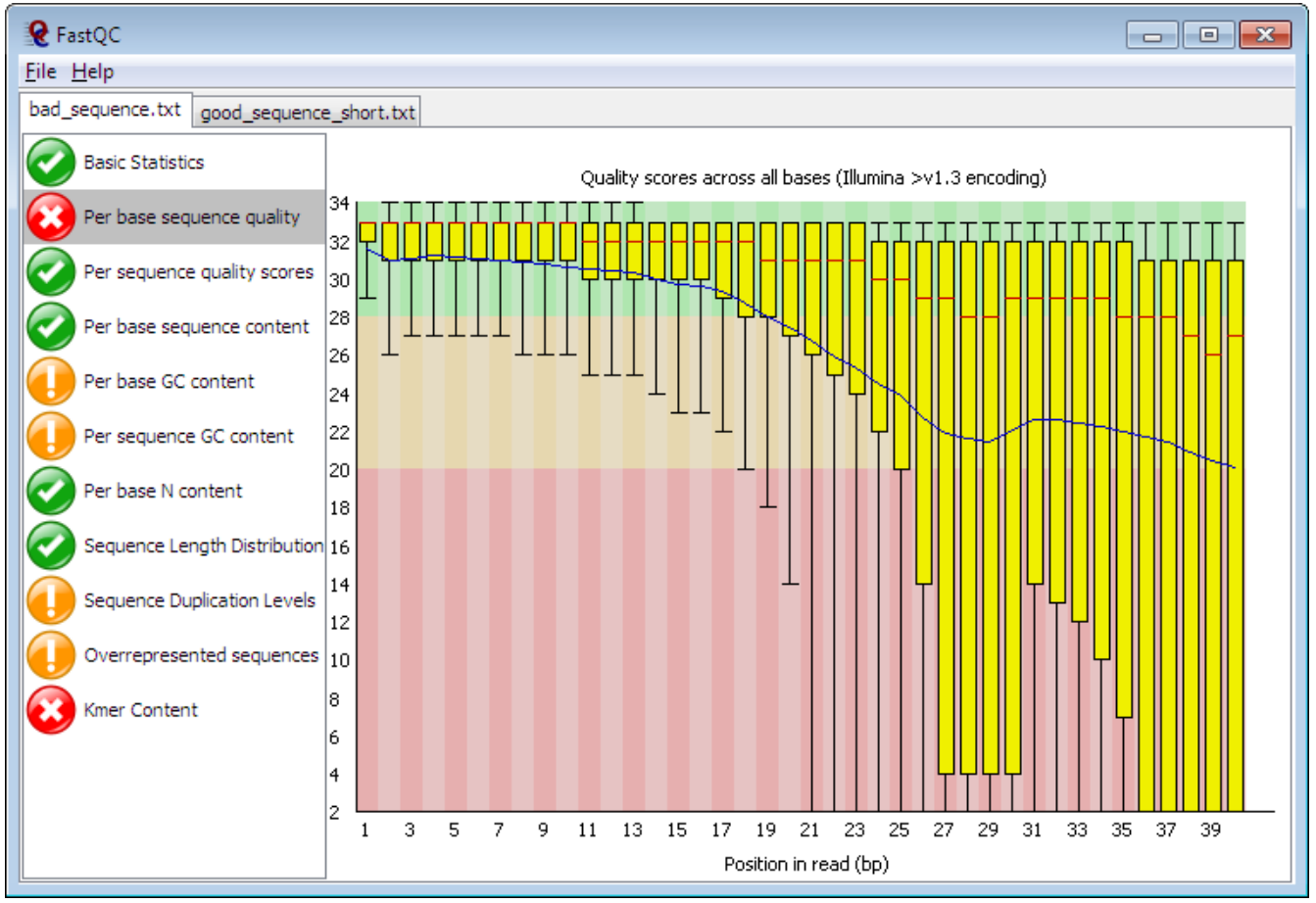
Paired-end Sequences

Two FastQ files, read name indicates left (/1) or right (/2) read of paired-end

```
@61DFRAAXX100204:1:100:10494:3070/1
AAACAACAGGGCACATTGTCACTCTTGTATTTGAAAAACACTTTCCGGCCAT
+
ACCCCCCCCCCCCCCCCCCCCCCCCCCCCCBC?CCCCCCCCC@@CACCCCCA
```

```
@61DFRAAXX100204:1:100:10494:3070/2
CTCAAATGGTTAATTCTCAGGCTGCAAATATTCGTTCAGGATGGAAGAACA
+
C<CCCCCCCACCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCBCCCC
```

Good QC
Summarise, Visualise and Flag



FastQC Report

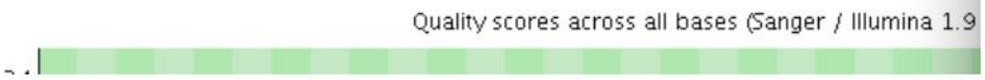
Summary

- Basic Statistics
- Per base sequence quality
- Per sequence quality scores
- Per base sequence content
- Per sequence GC content
- Per base N content
- Sequence Length Distribution
- Sequence Duplication Levels
- Overrepresented sequences
- Adapter Content
- Kmer Content

Basic Statistics

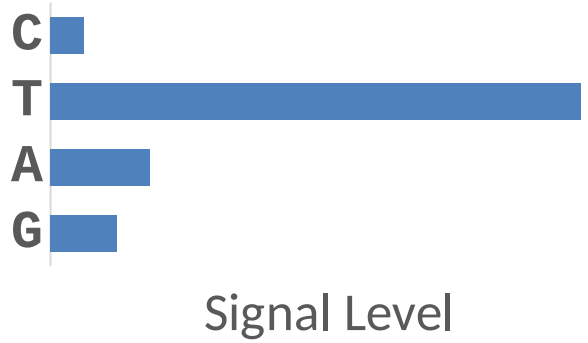
Measure	Value
Filename	read2.fq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	75000
Sequences flagged as poor quality	0
Sequence length	35
%GC	33

Per base sequence quality



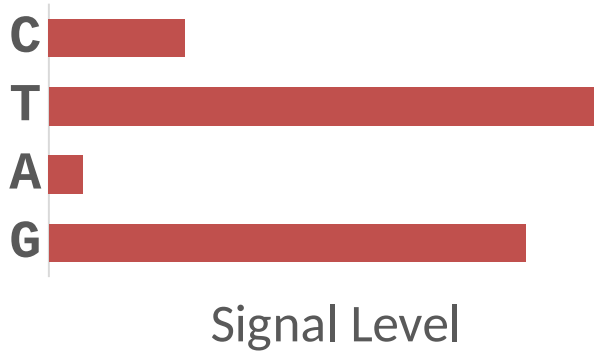
Technical

Technical Failures



Call = T

Confidence = High



Call = T

Confidence = Low

Call = T

Confidence = Low

Technical

Phred Scores

$$\text{Phred} = -10 \log_{10} p$$

p = Probability call is incorrect

10% error	Phred10
1% error	Phred20
0.1% error	Phred30

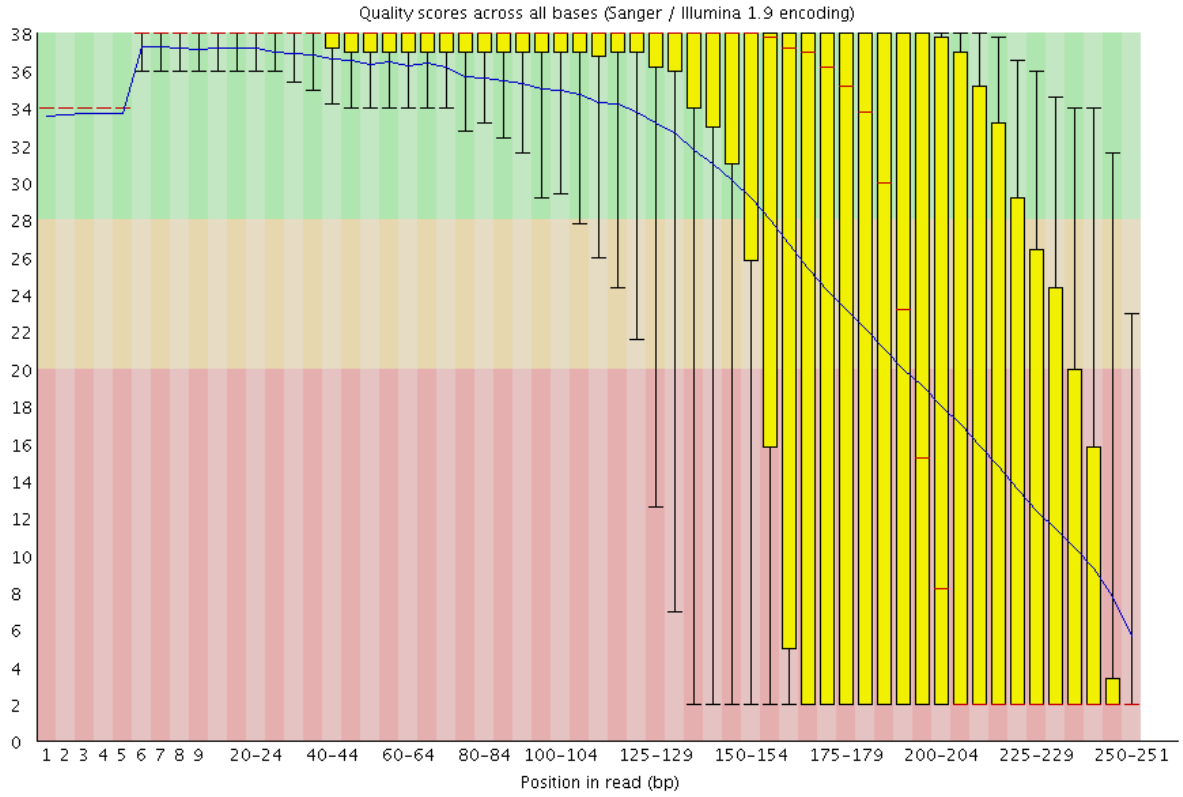
Technical

Phred Scores

$$\text{Phred} = -10 \log_{10} p$$

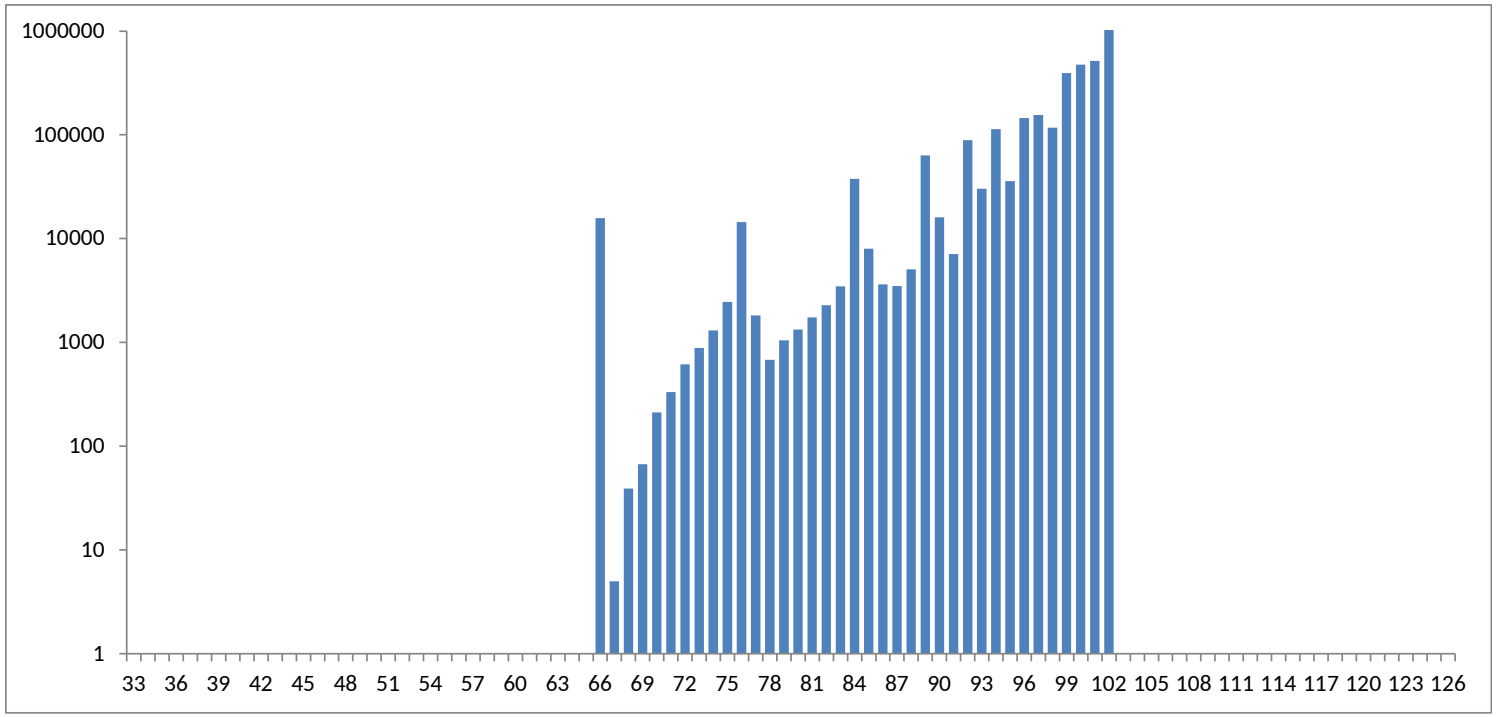
p = Probability call is incorrect

10% error	Phred10
1% error	Phred20
0.1% error	Phred30



Technical

Incorrect Encoding



Phred64 (Illumina)

Phred33 (Sanger)

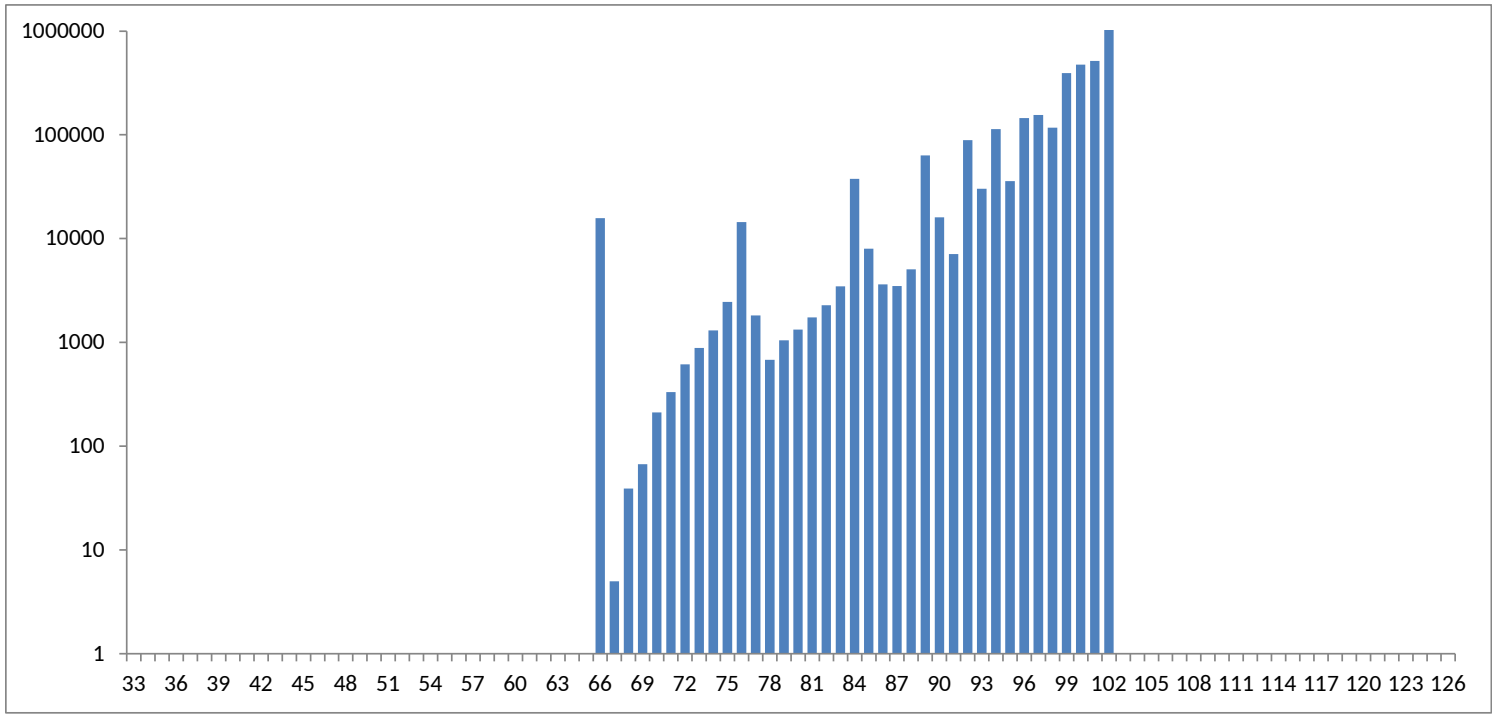
Technical

Incorrect Encoding

Phred64

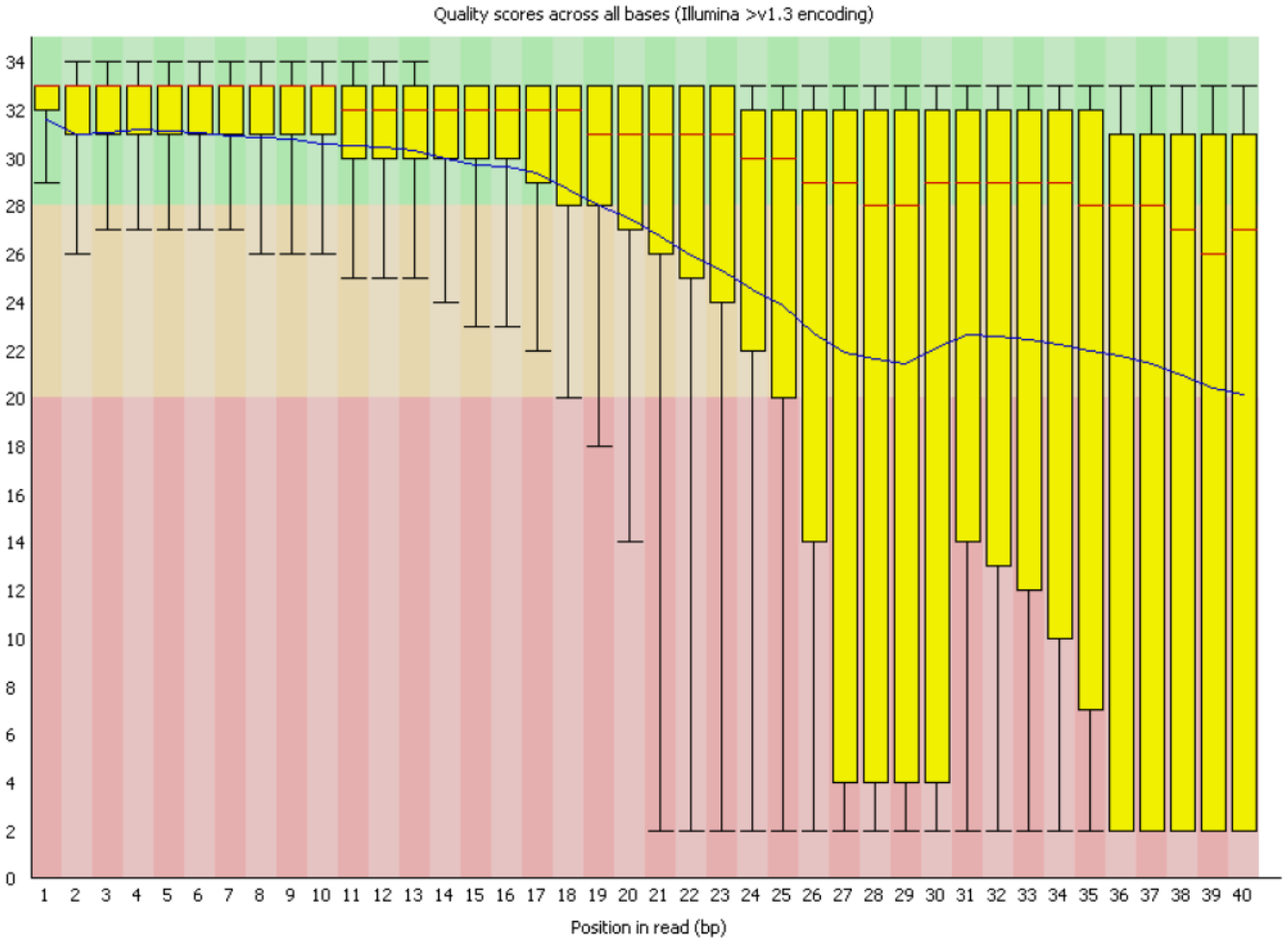
```
!"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefgh
```

Phred33



Phred64 (Illumina)

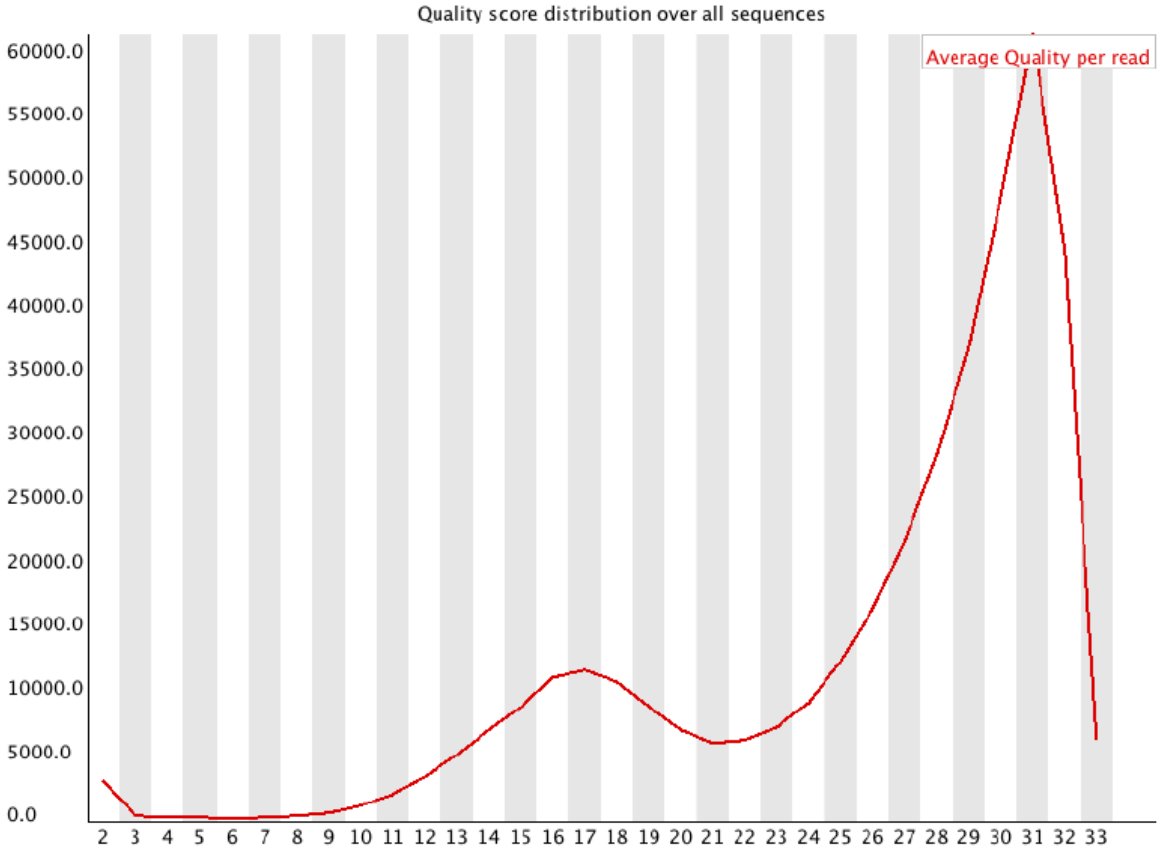
Phred33 (Sanger)



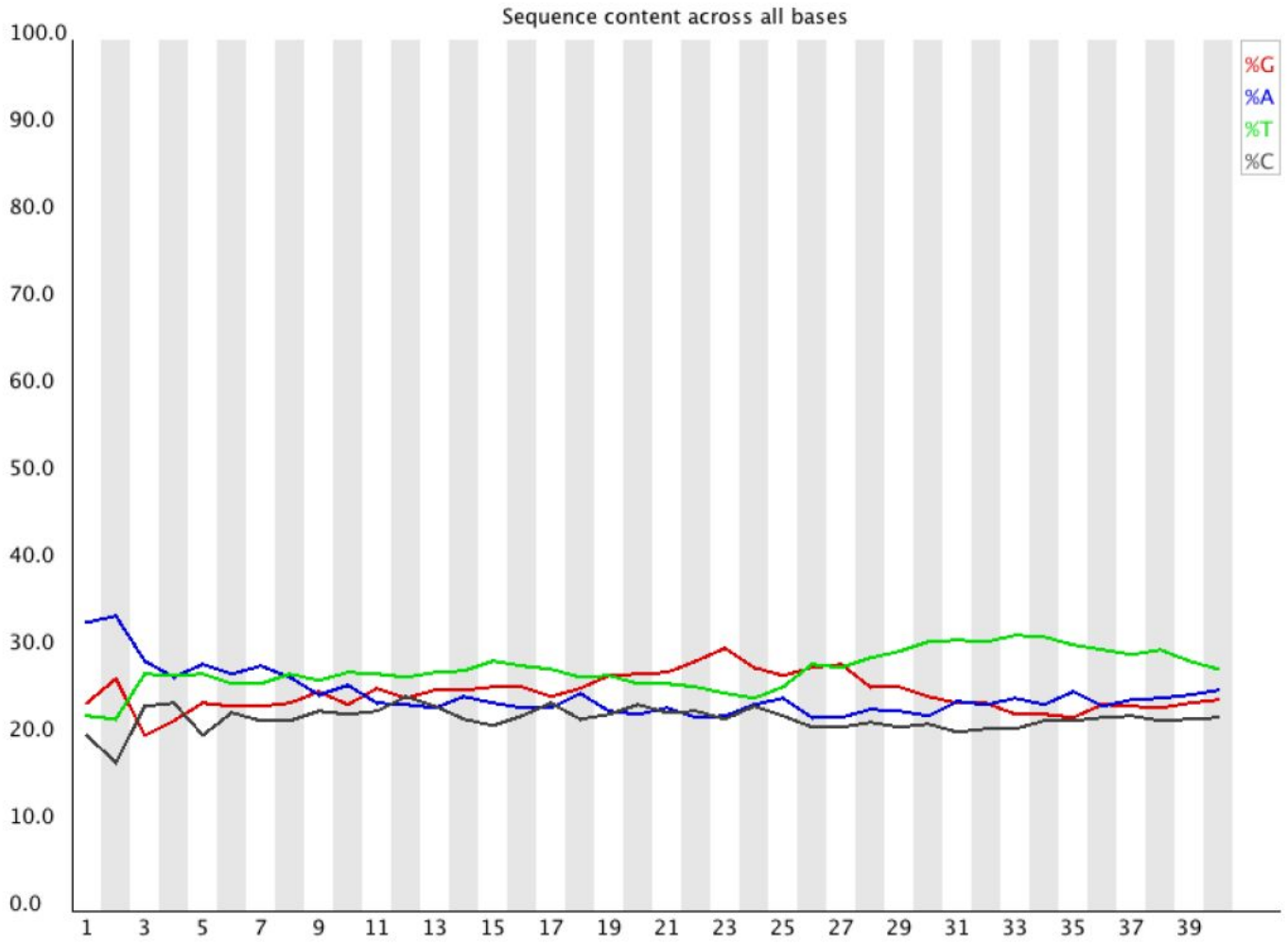
For each position a BoxWhisker type plot is drawn. The elements of the plot are as follows:

- The central red line is the median value
- The yellow box represents the inter-quartile range (25-75%)
- The upper and lower whiskers represent the 10% and 90% points
- The blue line represents the mean quality

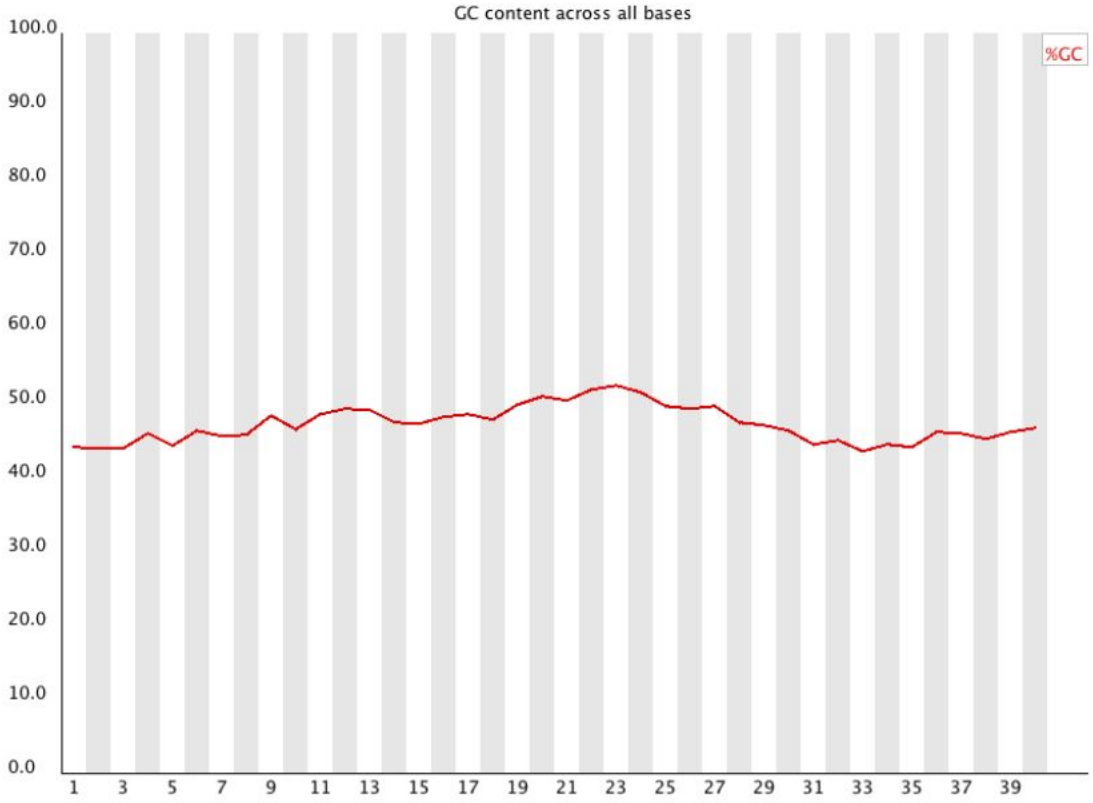
The per sequence quality score report allows you to see if a subset of your sequences have universally low quality values. It is often the case that a subset of sequences will have universally poor quality, often because they are poorly imaged (on the edge of the field of view etc), however these should represent only a small percentage of the total sequences.

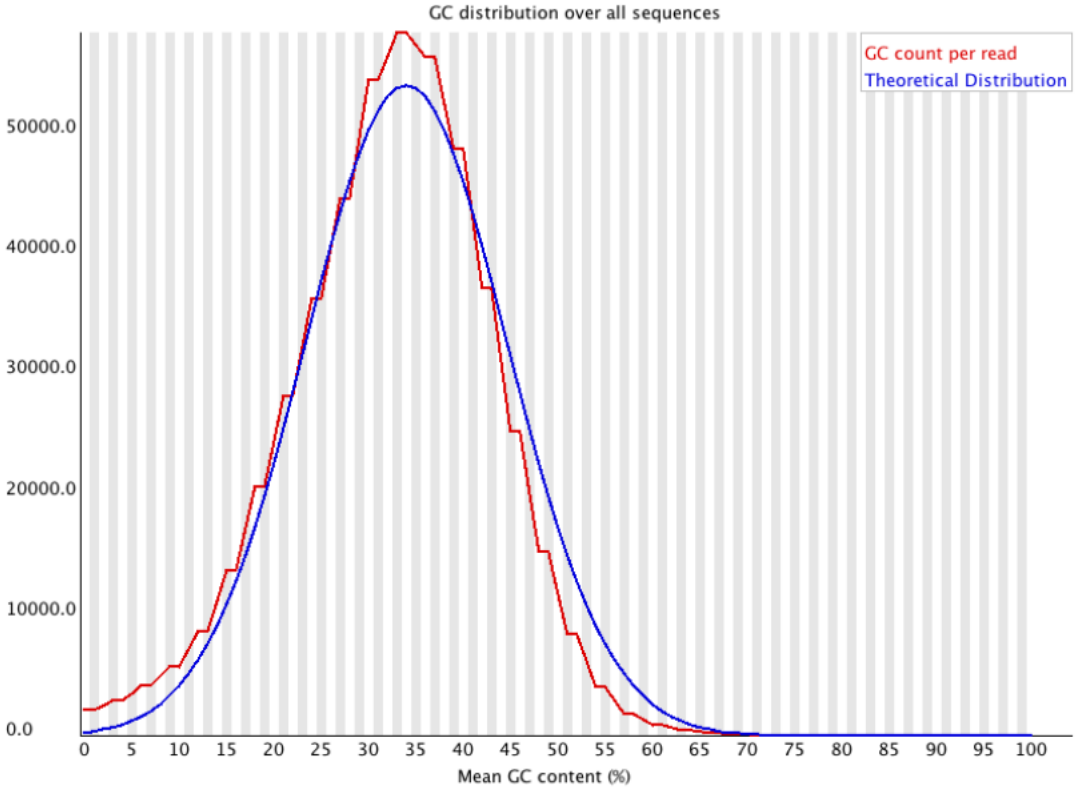


Per Base Sequence Content plots out the proportion of each base position in a file for which each of the four normal DNA bases has been called.



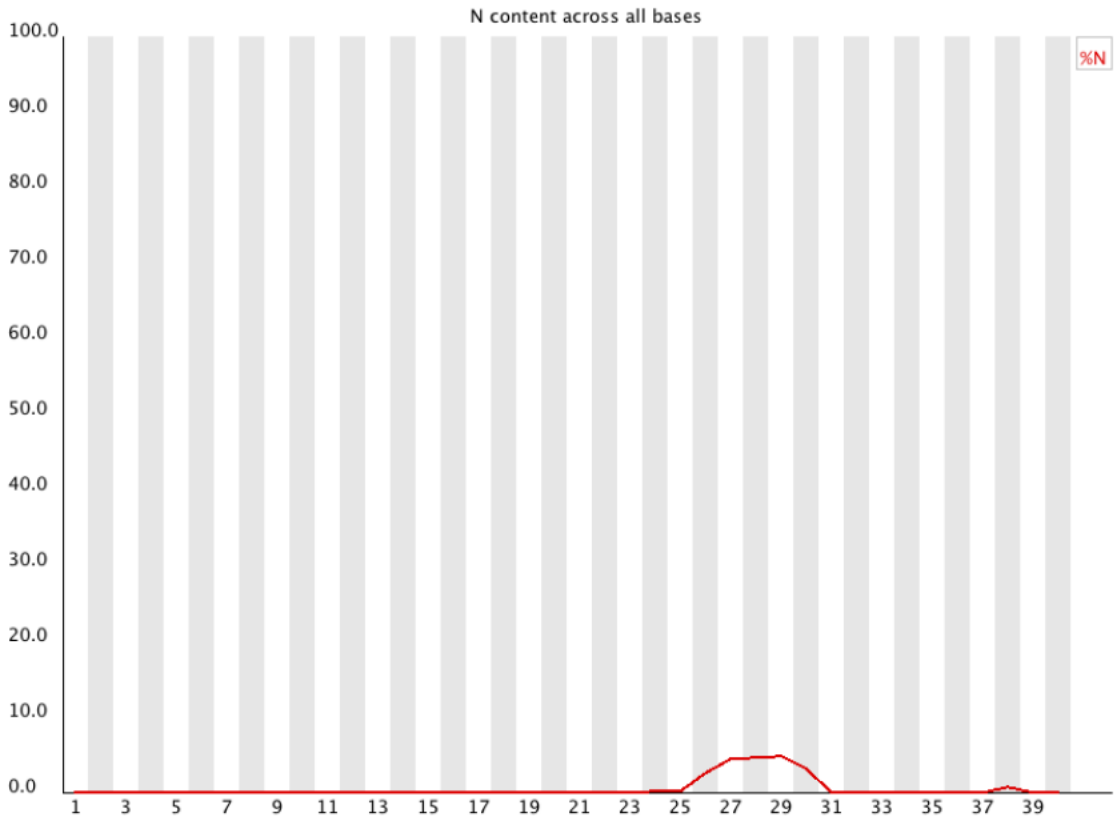
Per Base GC Content plots out the GC content of each base position in a file.





In a normal random library you would expect to see a roughly normal distribution of GC content where the central peak corresponds to the overall GC content of the underlying genome. Since we don't know the the GC content of the genome the modal GC content is calculated from the observed data and used to build a reference distribution.

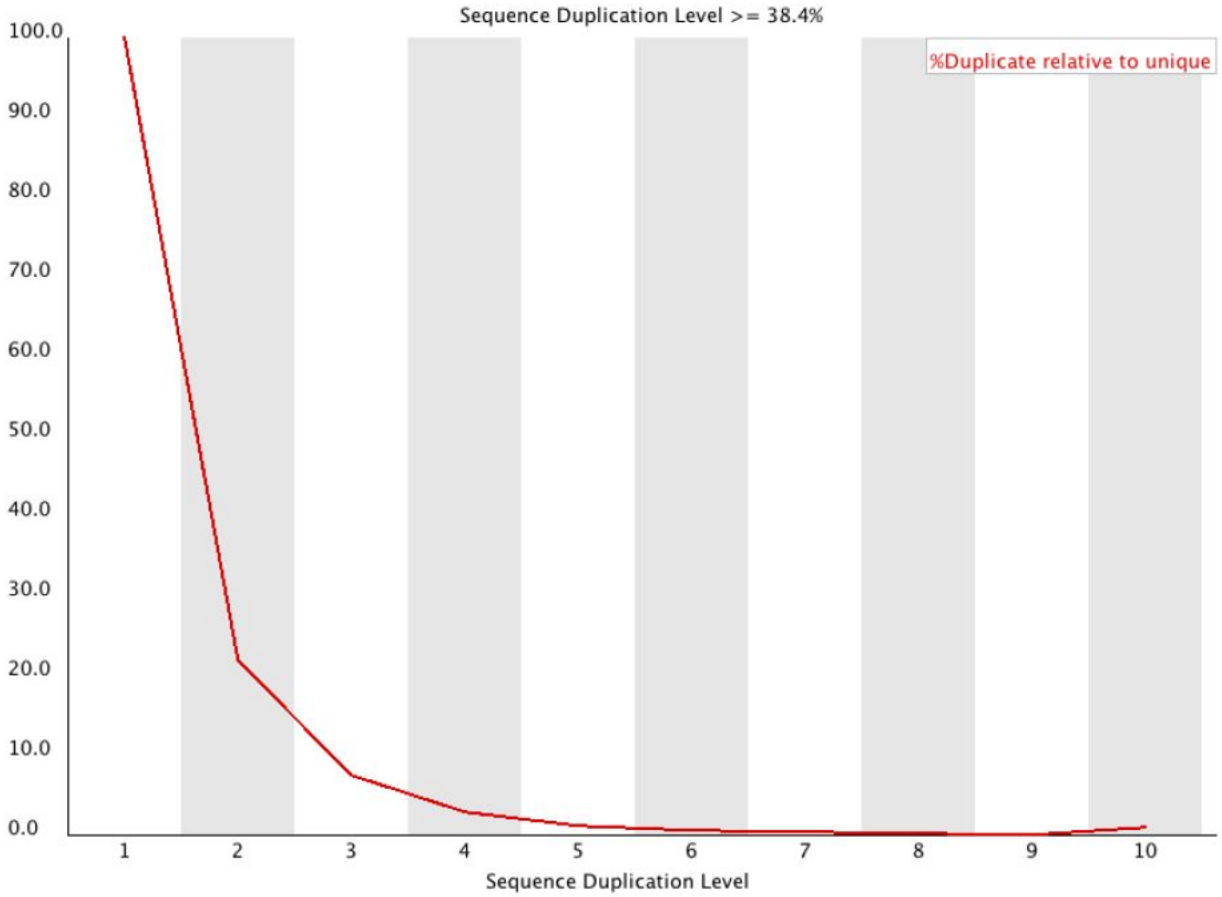
An unusually shaped distribution could indicate a contaminated library or some other kinds of biased subset. A normal distribution which is shifted indicates some systematic bias which is independent of base position. If there is a systematic bias which creates a shifted normal distribution then this won't be flagged as an error by the module since it doesn't know what your genome's GC content should be.



It's not unusual to see a very low proportion of Ns appearing in a sequence, especially nearer the end of a sequence. However, if this proportion rises above a few percent it suggests that the analysis pipeline was unable to interpret the data well enough to make valid base calls.

In a diverse library most sequences will occur only once in the final set. A low level of duplication may indicate a very high level of coverage of the target sequence, but a high level of duplication is more likely to indicate some kind of enrichment bias (eg PCR over amplification).

This module counts the degree of duplication for every sequence in the set and creates a plot showing the relative number of sequences with different degrees of duplication.



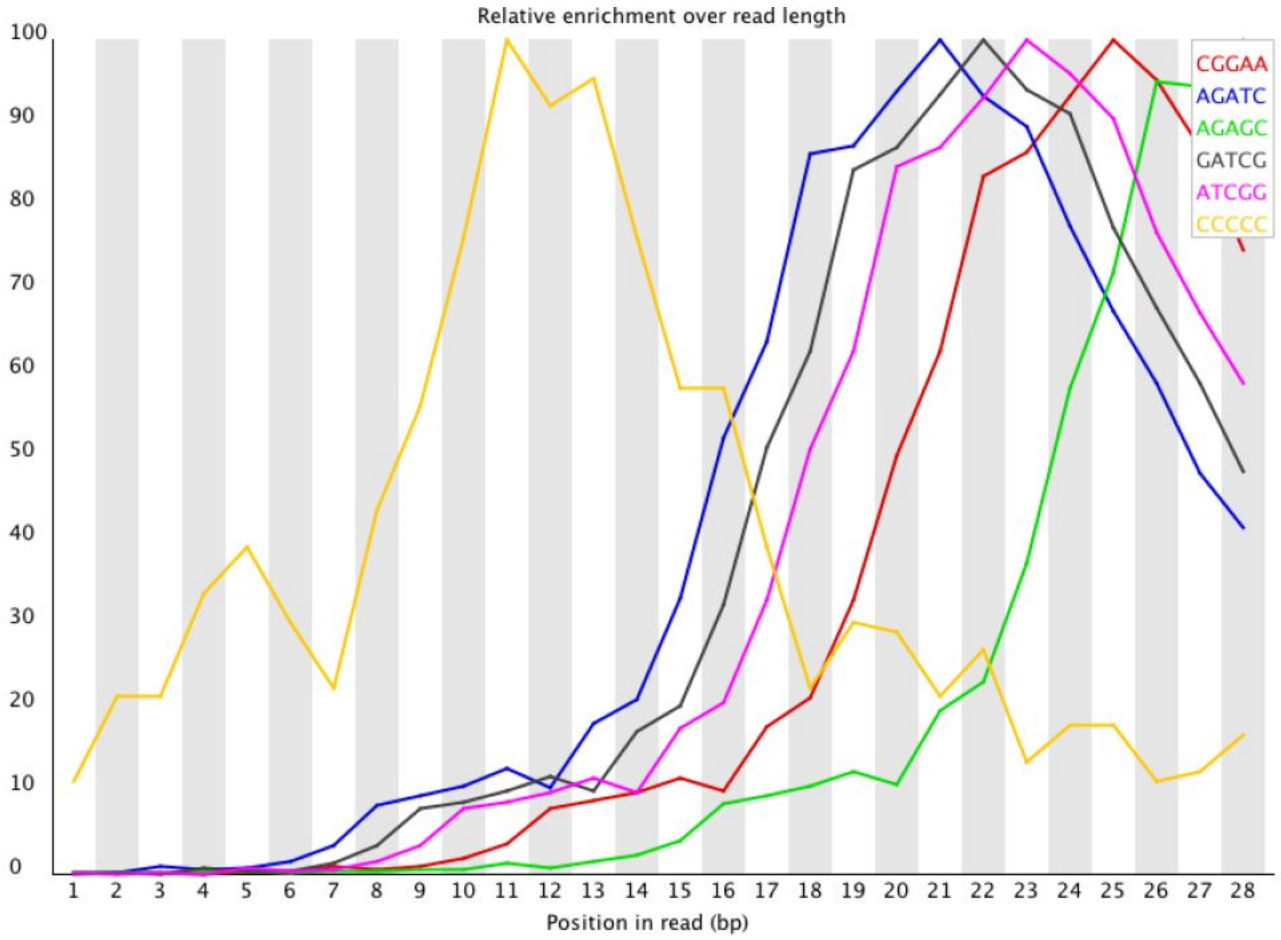
3.10 Overrepresented Sequences

Summary

A normal high-throughput library will contain a diverse set of sequences, with no individual sequence making up a tiny fraction of the whole. Finding that a single sequence is very overrepresented in the set either means that it is highly biologically significant, or indicates that the library is contaminated, or not as diverse as you expected.

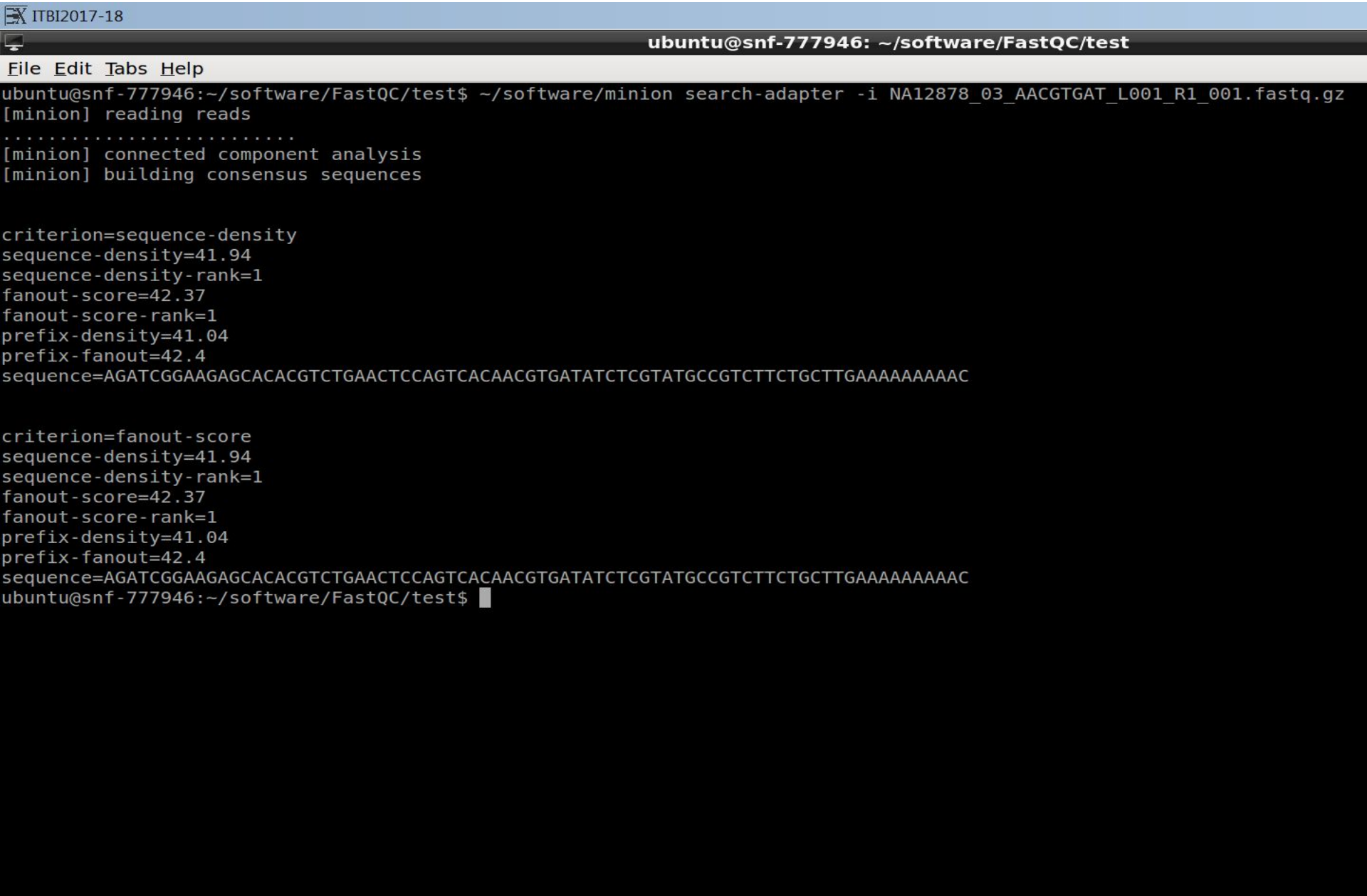
This module lists all of the sequence which make up more than 0.1% of the total. To conserve memory only sequences which appear in the first 200,000 sequences are tracked to the end of the file. It is therefore possible that a sequence which is overrepresented but doesn't appear at the start of the file for some reason could be missed by this module.

For each overrepresented sequence the program will look for matches in a database of common contaminants and will report the best hit it finds. Hits must be at least 20bp in length and have no more than 1 mismatch. Finding a hit doesn't necessarily mean that this is the source of the contamination, but may point you in the right direction. It's also worth pointing out that many adapter sequences are very similar to each other so you may get a hit reported which isn't technically correct, but which has very similar sequence to the actual match.



Any k-mer showing more than a 3 fold overall enrichment or a 5 fold enrichment at any given base position will be reported by this module.

Sequencing adapter identification



Sequencing adapter removal

```
150     27      0.0     3       25 2
153     1       0.0     3       1
155     1       0.0     3       1
156     1       0.0     3       0 0 0 1
158     11369   0.0     3       10957 350 27 35

ubuntu@snf-777946:~/software/FastQC/test$   cd ..
ubuntu@snf-777946:~/software/FastQC$ ./fastqc
ubuntu@snf-777946:~/software/FastQC$ cd test/
ubuntu@snf-777946:~/software/FastQC/test$ ~/.local/bin/cutadapt -a AGATCGGAAGAGC -o NA12878_03_AACGTGAT_L001_R1_001.fastq.trm.gz NA12878_03_AACGTGAT_L001_R1_00
1.fastq.gz | tee trm.log2
This is cutadapt 1.14 with Python 2.7.12
Command line parameters: -a AGATCGGAAGAGC -o NA12878_03_AACGTGAT_L001_R1_001.fastq.trm.gz NA12878_03_AACGTGAT_L001_R1_001.fastq.gz
Trimming 1 adapter with at most 10.0% errors in single-end mode ...
```

Transcript Reconstruction from RNA-Seq Reads

Advancing RNA-Seq analysis

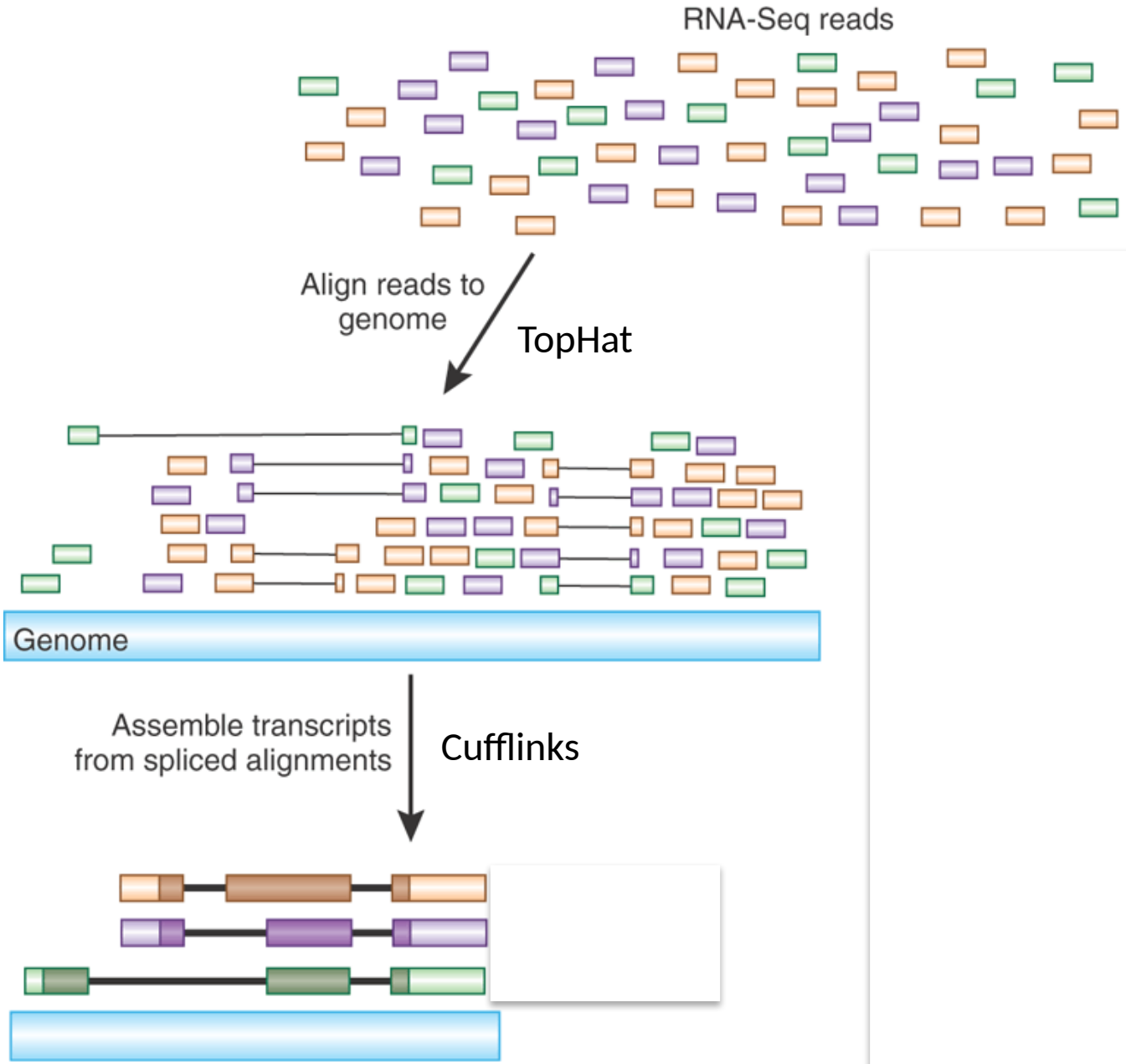
Brian J Haas & Michael C Zody

Nature Biotech, 2010

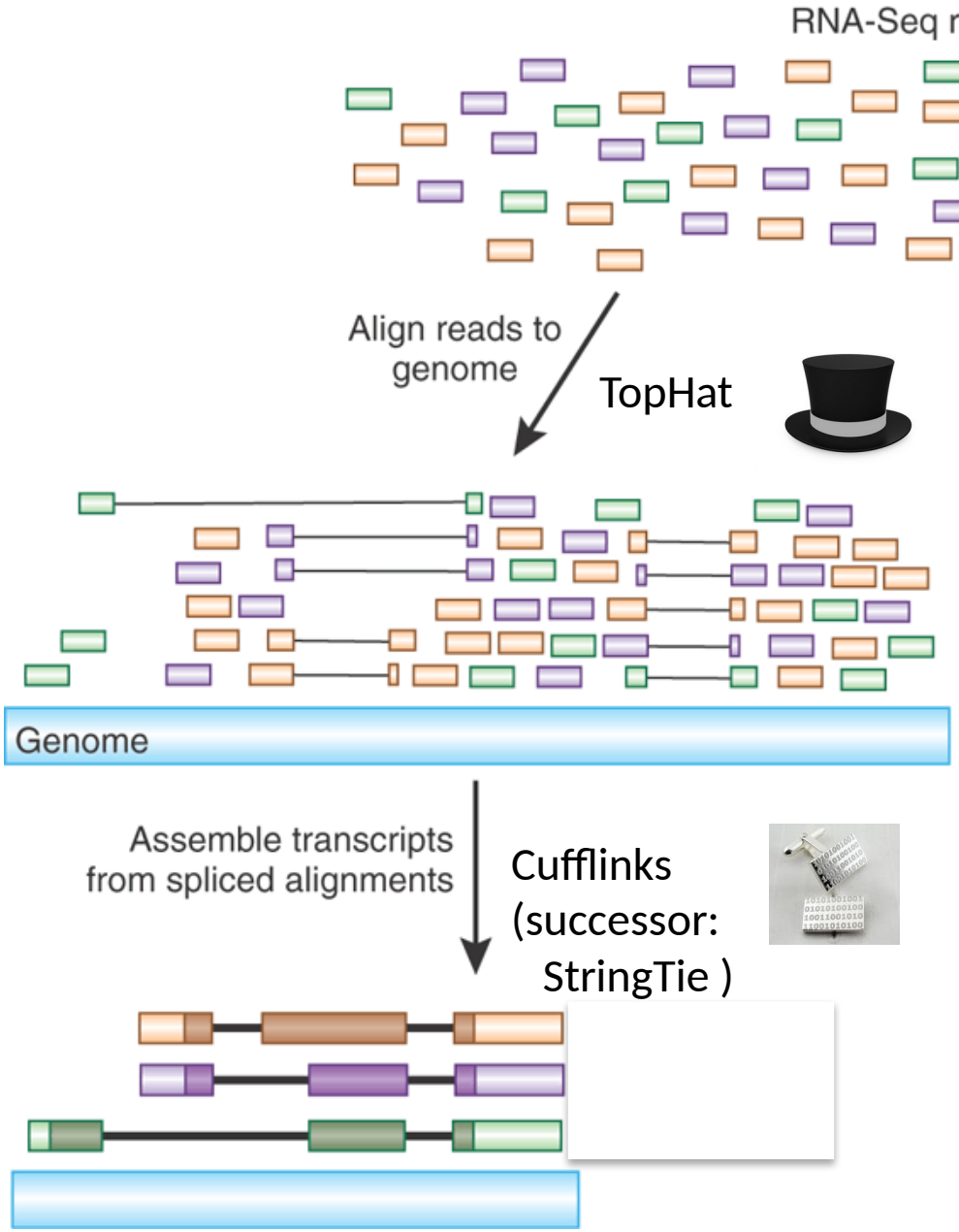
New methods for analyzing RNA-Seq data enable *de novo* reconstruction of the transcriptome.

Transcript Reconstruction from RNA-Seq Reads

Transcript Reconstruction from RNA-Seq Reads



Transcript Reconstruction from RNA-Seq Reads



The Tuxedo Suite:

End-to-end **Genome**-based RNA-Seq Analysis Software Package

NATURE PROTOCOLS | PROTOCOL

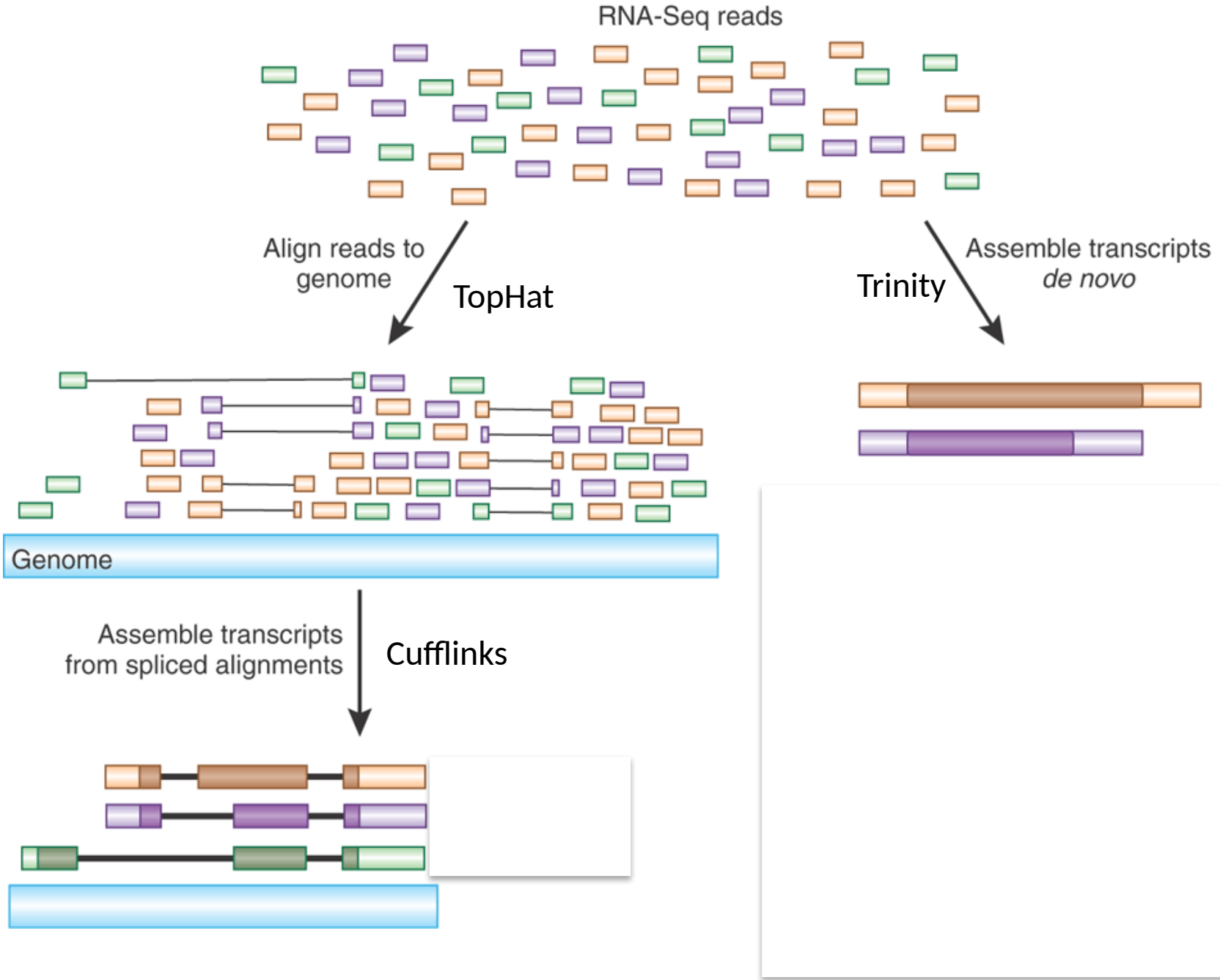
Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks

Cole Trapnell, Adam Roberts, Loyal Goff, Geo Pertea, Daehwan Kim, David R Kelley, Harold Pimentel, Steven L Salzberg, John L Rinn & Lior Pachter

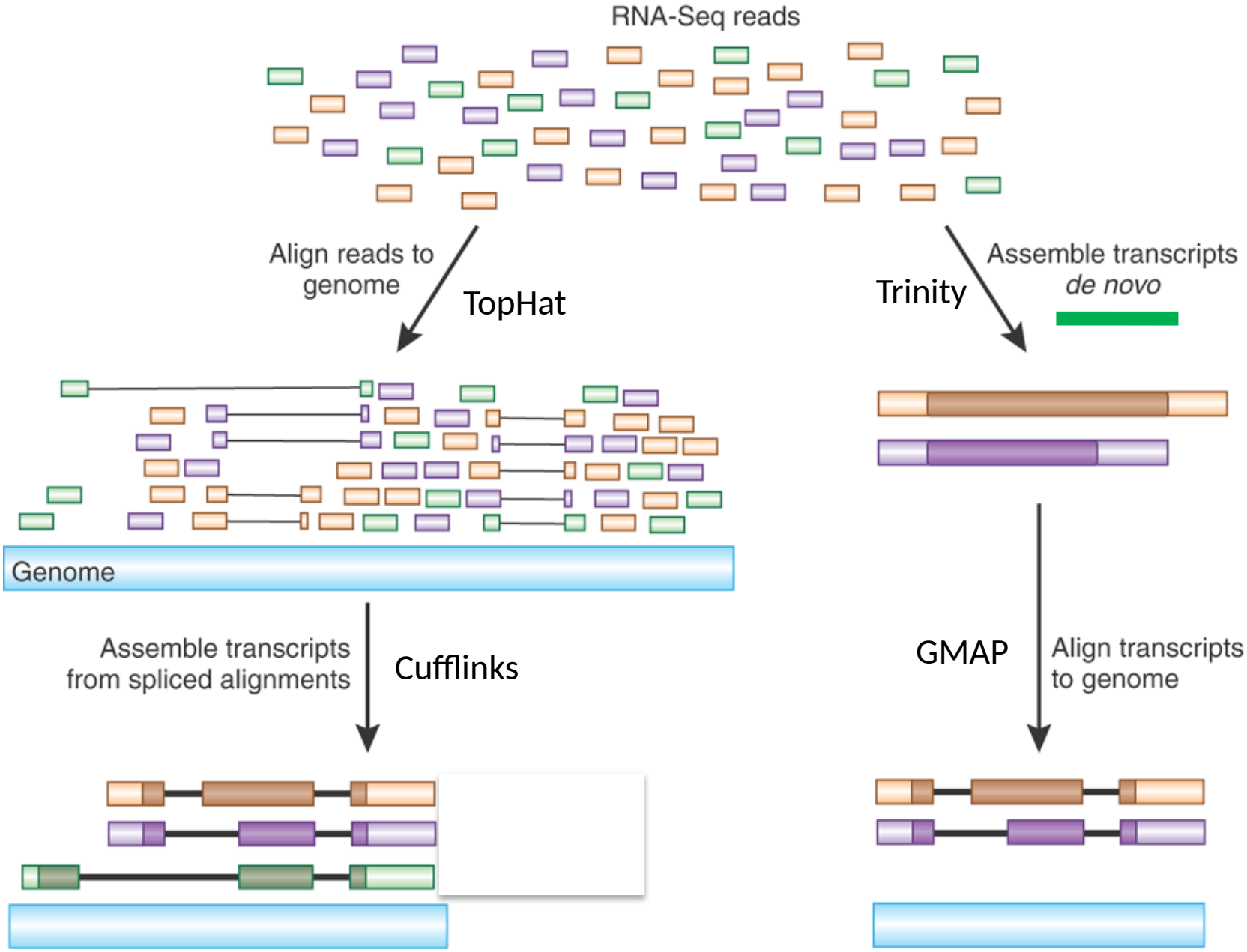
Affiliations | Contributions | Corresponding author

Nature Protocols **7**, 562–578 (2012) | doi:10.1038/nprot.2012.016
Published online 01 March 2012

Transcript Reconstruction from RNA-Seq Reads



Transcript Reconstruction from RNA-Seq Reads



Transcript Reconstruction from RNA-Seq Reads

End-to-end **Transcriptome**-based
RNA-Seq Analysis
Software Package

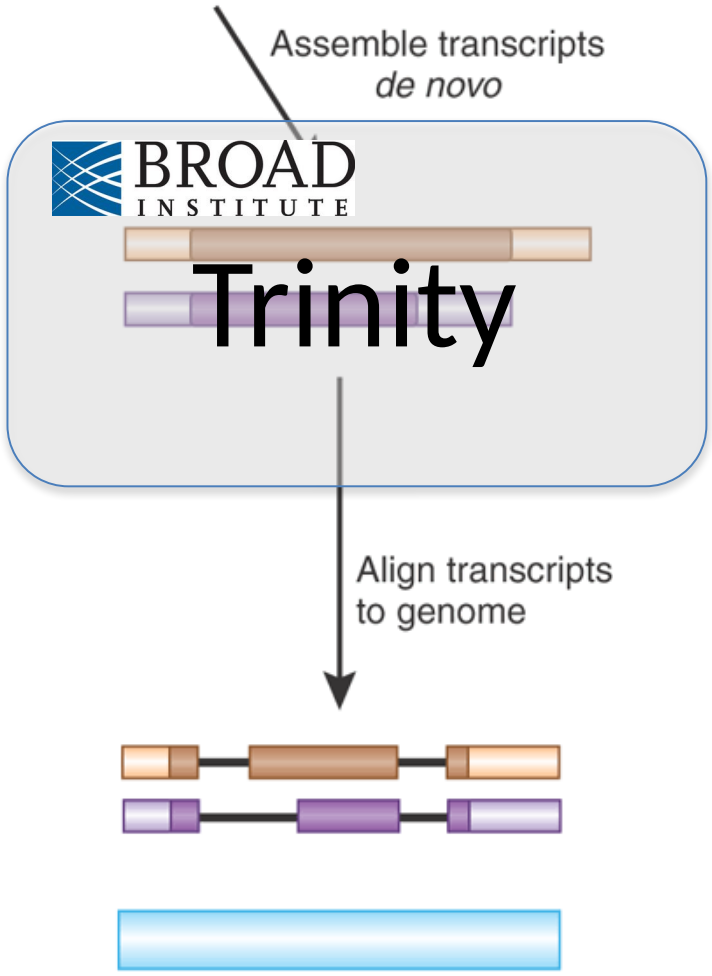
NATURE PROTOCOLS | PROTOCOL

De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis

Brian J Haas, Alexie Papanicolaou, Moran Yassour, Manfred Grabherr, Philip D Blood, Joshua Bowden, Matthew Brian Couger, David Eccles, Bo Li, Matthias Lieber, Matthew D MacManes, Michael Ott, Joshua Orvis, Nathalie Pochet, Francesco Strozzi, Nathan Weeks, Rick Westerman, Thomas William, Colin N Dewey, Robert Henschel, Richard D LeDuc, Nir Friedman & Aviv Regev

Affiliations | Contributions | Corresponding authors

Nature Protocols **8**, 1494–1512 (2013) | doi:10.1038/nprot.2013.084
Published online 11 July 2013



Overview of the Tuxedo Software Suite

Bowtie (fast short-read alignment)

TopHat (spliced short-read alignment)

Cufflinks (transcript reconstruction from alignments)
(now: StringTie)

Cuffdiff (differential expression analysis)
(now: BallGroom)

CummeRbund (visualization & analysis)
(now: BallGroom)

Slide courtesy of Cole Trapnell

Tuxedo development team

Daehwan Kim
Johns Hopkins

Ben Langmead
Johns Hopkins

Ali Mortazavi
Caltech

Barbara Wold
Caltech

Geo Pertea
Johns Hopkins

Steven Salzberg
Johns Hopkins

Lior Pachter
UC Berkeley

Cole Trapnell
Harvard

Brian Williams
Caltech

Mihai Pop
U. Maryland

Jeltje van Baren
MBARI

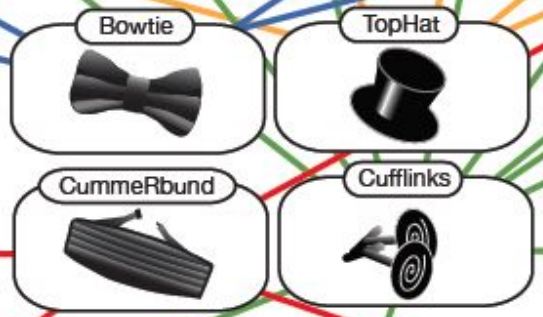
Loyal Goff
Harvard

Adam Roberts
UC Berkeley

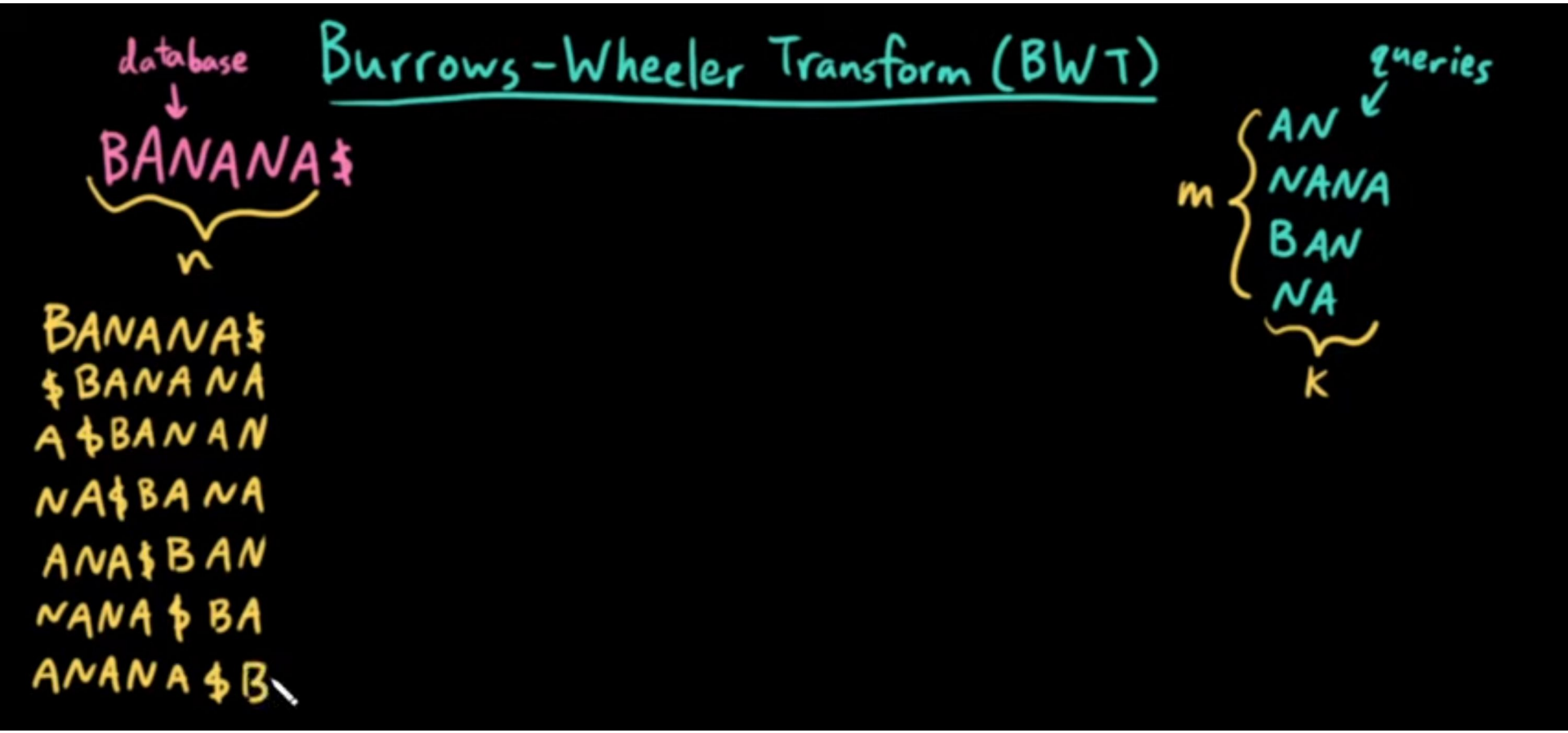
John Rinn
Harvard

Dave
Hendrickson
Harvard

Manolis Kellis
MIT



Bowtie, BWA, HiSat are based on the BWT (linear time matching):



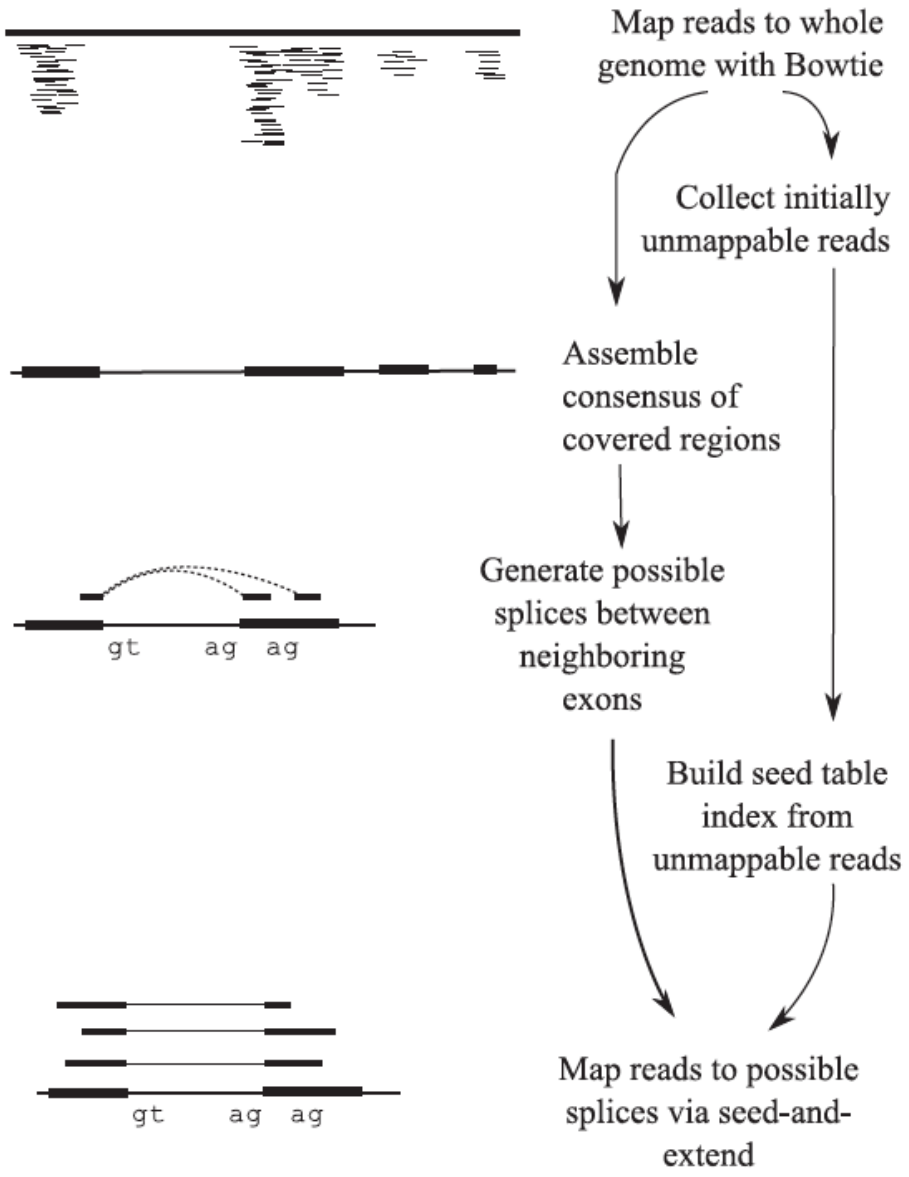
Niema Moshiri (UCSC) https://niema.net

https://www.youtube.com/watch?v=Lc-ACiJIrnM #BWT intro

https://www.youtube.com/watch?v=ni_w-rdItG8 #BWT inversion

https://www.youtube.com/watch?v=uKreghMwLLE #BWT matching

The TopHat Pipeline



IUM reads

From Trapnell, Pachter, & Salzberg. Bioinformatics. 2009

‘seed and extend’

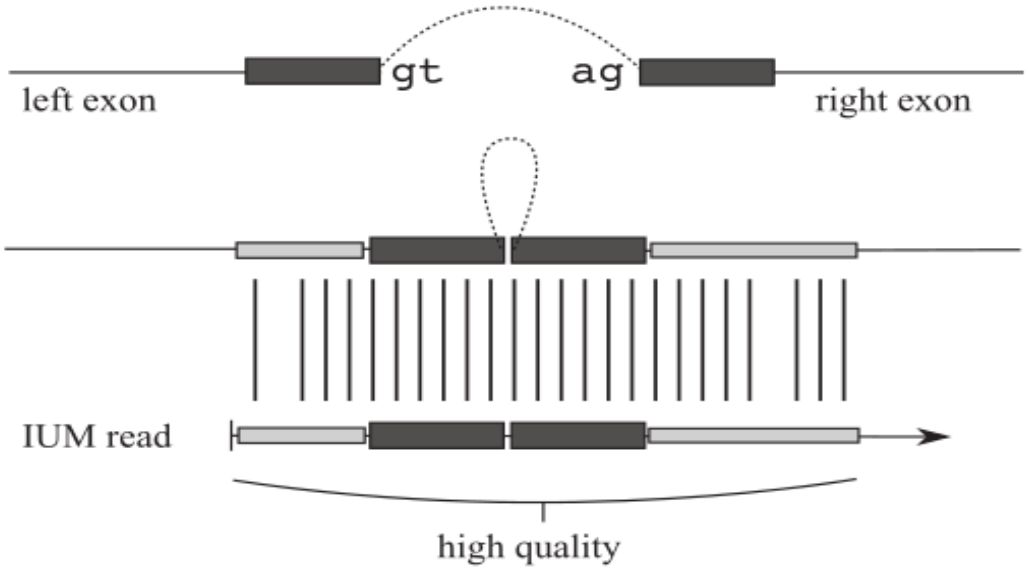


Fig. 3. The seed and extend alignment used to match reads to possible splice sites. For each possible splice site, a seed is formed by combining a small amount of sequence upstream of the donor and downstream of the acceptor. This seed, shown in dark gray, is used to query the index of reads that were not initially mapped by Bowtie. Any read containing the seed is checked for a complete alignment to the exons on either side of the possible splice. In the light gray portion of the alignment, TopHat allows a user-specified number of mismatches. Because reads typically contain low-quality base calls on their 3′ ends, TopHat only examines the first 28 bp on the 5′ end of each read by default.

Alignments are reported in a compact representation: SAM format

```
0       61G9EAAXX100520:5:100:10095:16477
1       83
2       chr1
3       51986
4       38
5       46M
6       =
7       51789
8       -264
9       CCCAAACAAGCCGAACTAGCTGATTTGGCTCGTAAAGACCCGGAAA
10      ###CB?=ADDBCBCDEEFFDEFFFDEFFGDBEFGEDGCFGFGGGGG
11      MD:Z:67
12      NH:i:1
13      HI:i:1
14      NM:i:0
15      SM:i:38
16      XQ:i:40
17      X2:i:0
```

SAM format specification: http://samtools.sourceforge.net/SAM1.pdf

Alignments are reported in a compact representation: SAM format

```
0     61G9EAAXX100520:5:100:10095:16477
1     83
2     chr1
3     51986
4     38
5     46M
6     =
7     51789
8     -264
9     CCCAAACAAGCCGAACTAGCTGATTTGGCTCGTAAAGACCCGGAAA
10    ###CB?=ADDBCBCDEEFFDEFFFDEFFGDBEFGEDGCFGFGGGGG
11    MD:Z:67
12    NH:i:1
13    HI:i:1
14    NM:i:0
15    SM:i:38
16    XQ:i:40
17    X2:i:0
```

(read name)

(FLAGS stored as bit fields; 83 = 00001010011)

(alignment target)

(position alignment starts)

(Compact description of the alignment in CIGAR format)

(read sequence, oriented according to the forward alignment)

(base quality values)

(Metadata)

SAM format specification: http://samtools.sourceforge.net/SAM1.pdf

Alignments are reported in a compact representation: SAM format

```
0       61G9EAAXX100520:5:100:10095:16477 (read name)
1       83 (FLAGS stored as bit fields; 83 = 00001010011 )
2       chr1 (alignment target)
```

Still not compact enough…
Millions to billions of reads takes up a lot of space!!

Convert SAM to binary – BAM format.

```
15      SM:i:38
16      XQ:i:40
17      X2:i:0
```

SAM format specification: http://samtools.sourceforge.net/SAM1.pdf

Samtools

- Tools for
  - converting SAM <-> BAM
  - Viewing BAM files (eg. samtools view file.bam | less)
  - Sorting BAM files, and lots more:

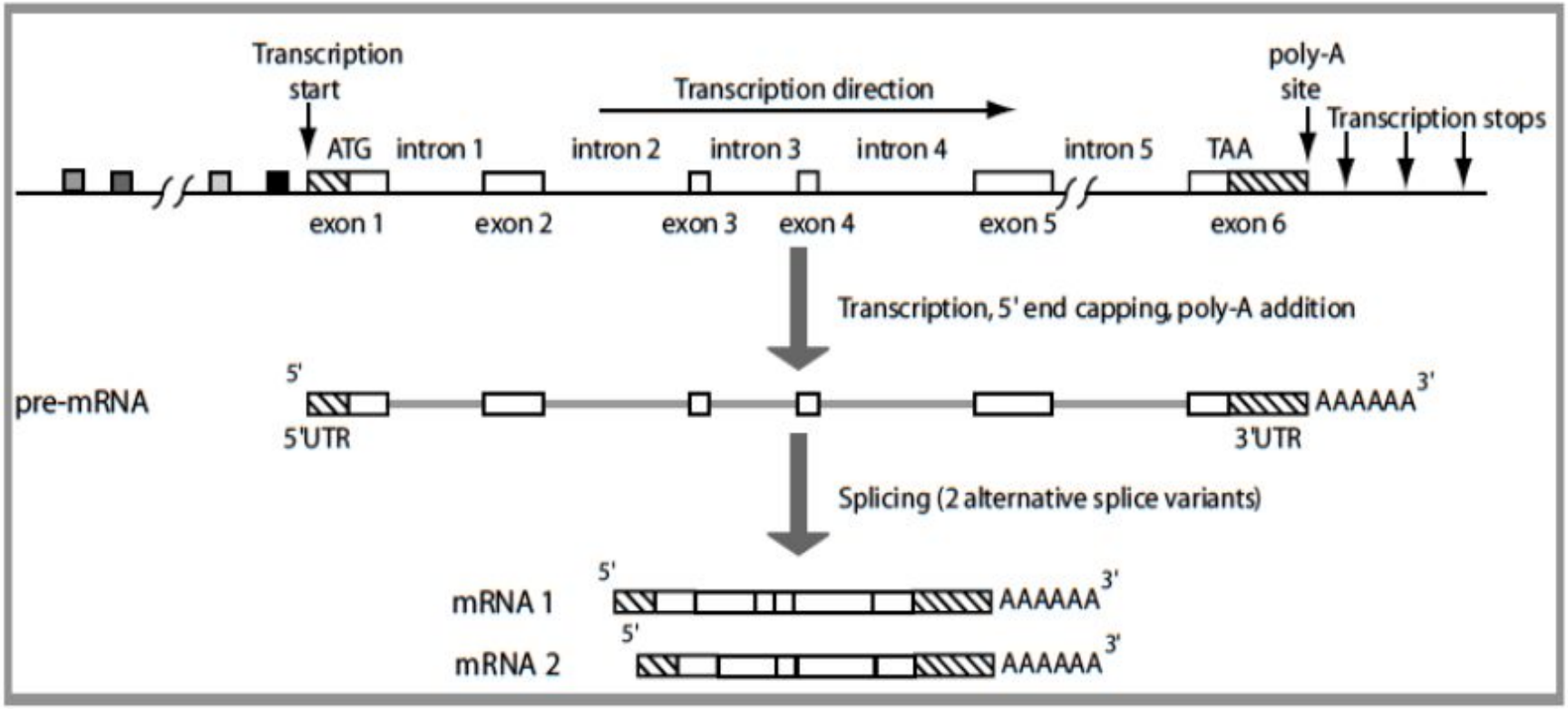
```
Program: samtools (Tools for alignments in the SAM format)
Version: 0.1.18 (r982:295)

Usage:   samtools <command> [options]

Command: view        SAM<->BAM conversion
         sort        sort alignment file
         mpileup     multi-way pileup
         depth       compute the depth
         faidx       index/extract FASTA
         tview       text alignment viewer
         index       index alignment
         idxstats    BAM index stats (r595 or later)
         fixmate     fix mate information
         flagstat    simple stats
         calmd       recalculate MD/NM tags and '=' bases
         merge       merge sorted alignments
         rmdup       remove PCR duplicates
         reheader    replace BAM header
         cat         concatenate BAMs
         targetcut   cut fosmid regions (for fosmid pool only)
         phase       phase heterozygotes
```

Cufflinks

Recall:

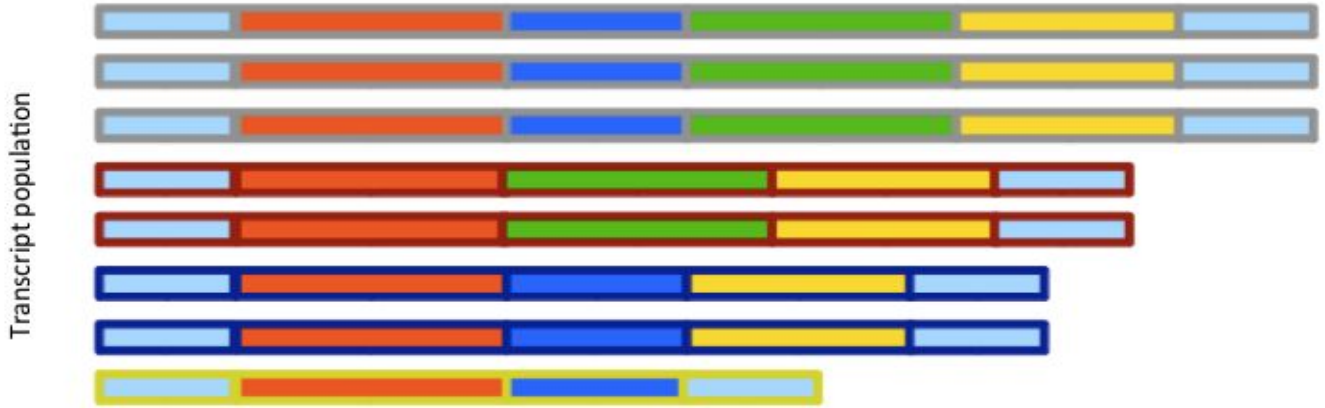


Source: Computational Genome Analysis

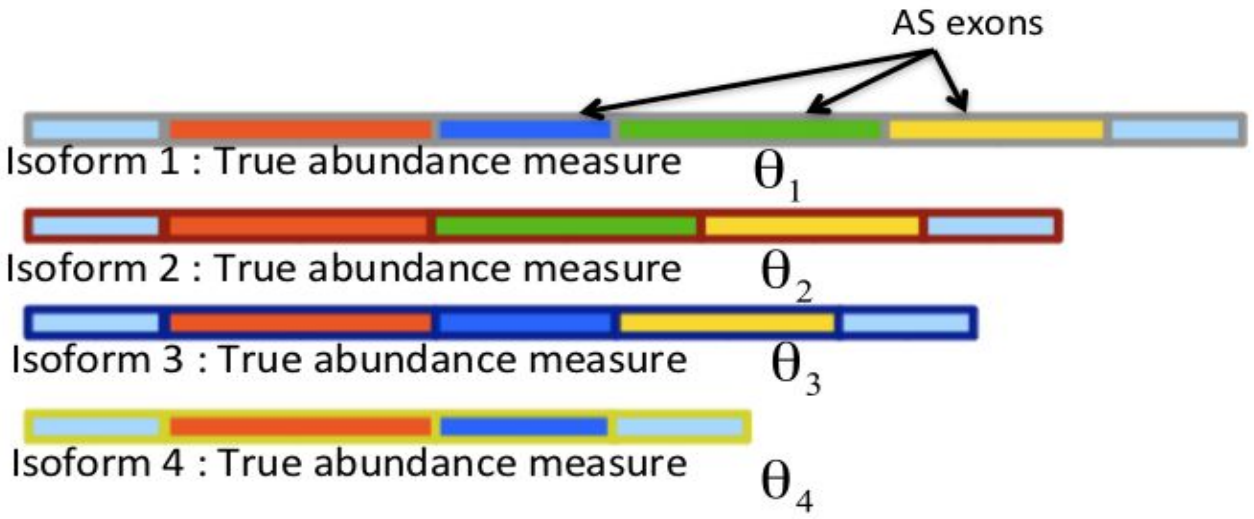
Goal: Develop and analyze a statistical model for measuring differential expression of **Isoforms** of the same gene using Rna-Seq.

The assembly algorithm is designed to aim for a parsimonious explanation of the fragments from the RNA-seq experiment, i.e.:

1. Every fragment is consistent with at least one assembled transcript.
2. Every transcript is tiled by reads.
3. The number of transcripts is the smallest required to satisfy requirement (1)
4. The resulting RNA-Seq models display some desirable qualities

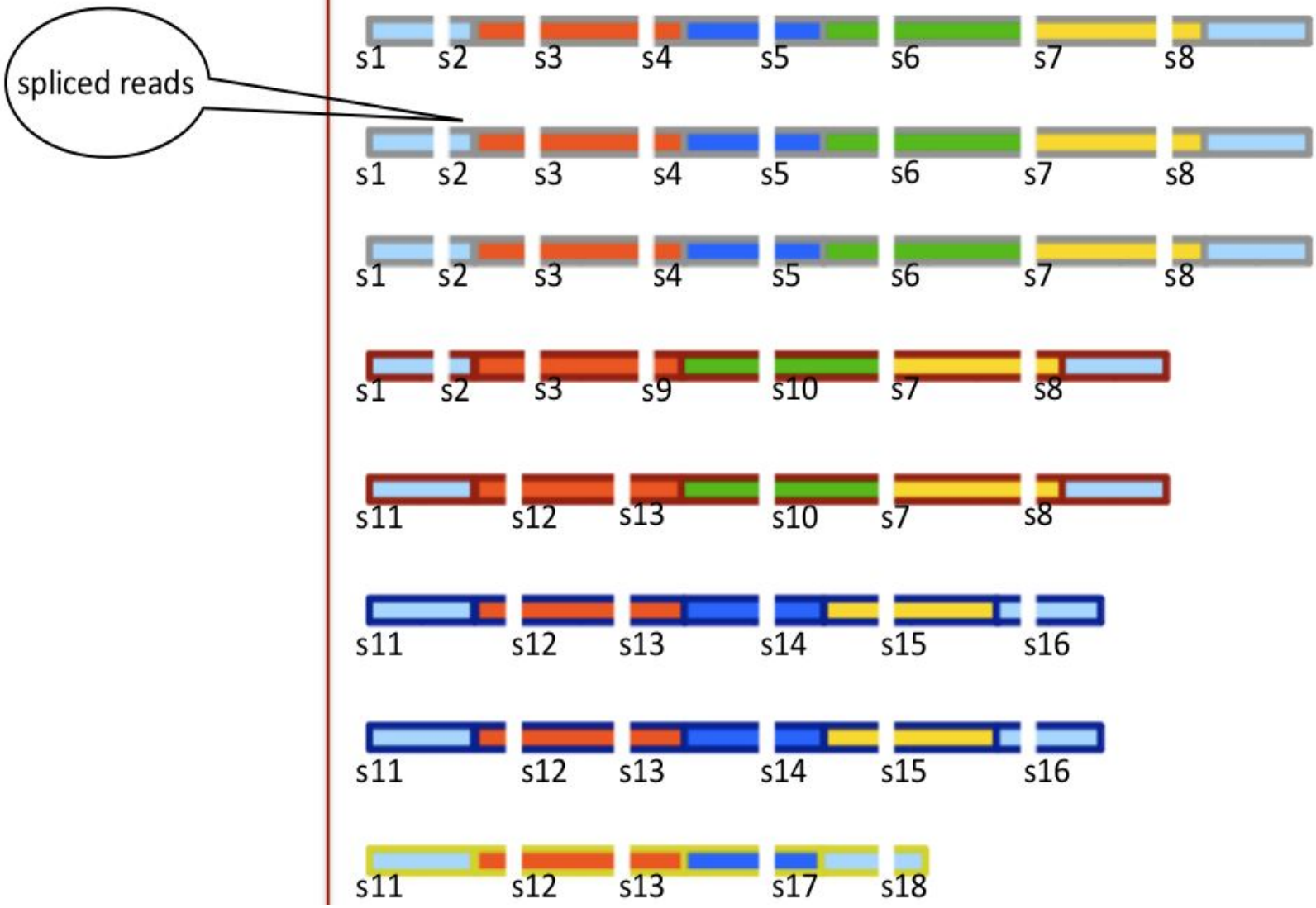


Suppose we have a gene with 4 isoforms and 3 alternatively spliced (AS) exons as shown above.



The goal is to estimate the true abundance measure of the 4 isoforms.

Fragmented mRNAs: 54 total reads with 18 unique types.



Reads vs. transcripts

n_{ij} matrix = the number of reads type s_j generated by transcript θ_i.

	s1	s2	s3	s4	s5	s6	s7	s8	s9	s10	s11	s12	s13	s14	s15	s16	s17	s18	
θ1	3	3	3	3	3	3	3	3	0	0	0	0	0	0	0	0	0	0	24
θ2	1	1	1	0	0	0	2	2	1	2	1	1	1	0	0	0	0	0	13
θ3	0	0	0	0	0	0	0	0	0	0	2	2	2	2	2	2	0	0	12
θ4	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	1	1	5
n_j	4	4	4	3	3	3	5	5	1	2	4	4	4	2	2	2	1	1	54

For each read type, we only observe n_j.

We want to estimate last column (transcript abundance).

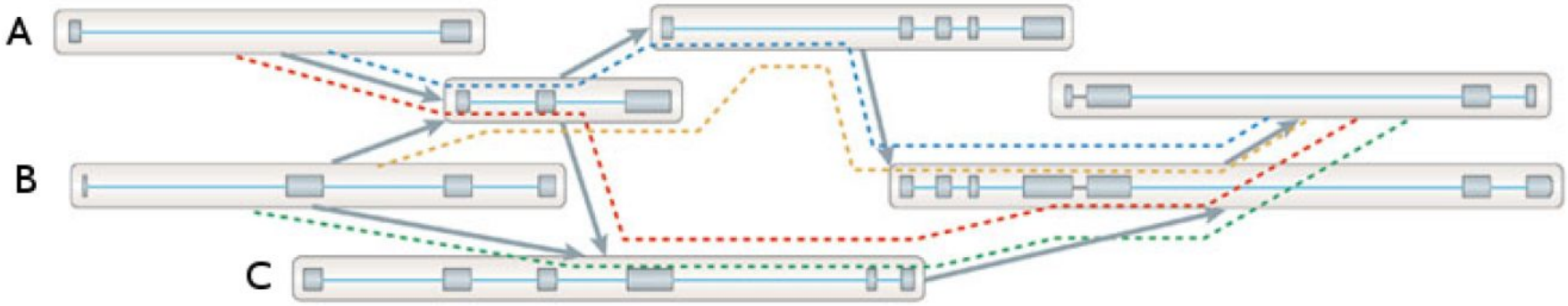
n_{ij} matrix = the number of reads type s_j generated by transcript θ_i.

	s1	s2	s3	s4	s5	s6	s7	s8	s9	s10	s11	s12	s13	s14	s15	s16	s17	s18	
θ1	3	3	3	3	3	3	3	3	0	0	0	0	0	0	0	0	0	0	24
θ2	1	1	1	0	0	0	2	2	1	2	1	1	1	0	0	0	0	0	13
θ3	0	0	0	0	0	0	0	0	0	0	2	2	2	2	2	2	0	0	12
θ4	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	1	1	5
n_j	4	4	4	3	3	3	5	5	1	2	4	4	4	2	2	2	1	1	54

- In reality we only observe $n_j = \sum_{i=1}^{I} n_{ij}$.
- $n_j \sim Poisson(\sum_{i=1}^{I} \theta_i a_{ij} = \theta^T a_j)$, where $\theta = \begin{bmatrix} \theta_1 \\ \dots \\ \theta_I \end{bmatrix}$, $a_j = \begin{bmatrix} a_{1j} \\ \dots \\ a_{Ij} \end{bmatrix}$.
- Likelihood: $f_\theta(n_1, n_2, \dots, n_J) = \prod_{j=1}^{J} \frac{(\theta^T a_j)^{n_j} e^{-\theta^T a_j}}{n_j!}$.

Compatible reads/fragments

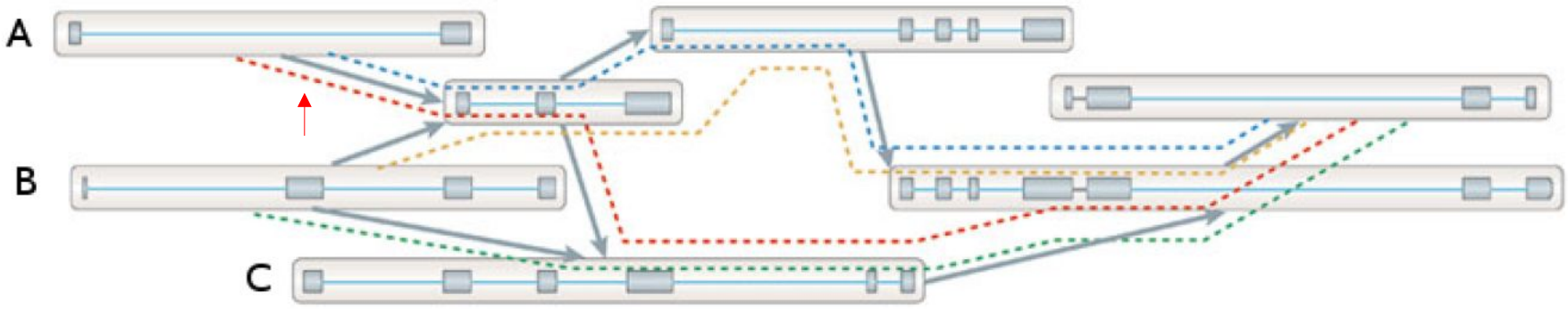
Two reads are **compatible** if their overlap contains the exact same implied introns (or none). If two reads are not compatible they are **incompatible**.



- Read A is incompatible with reads B and C
- Read B is compatible with read C

Compatible reads/fragments

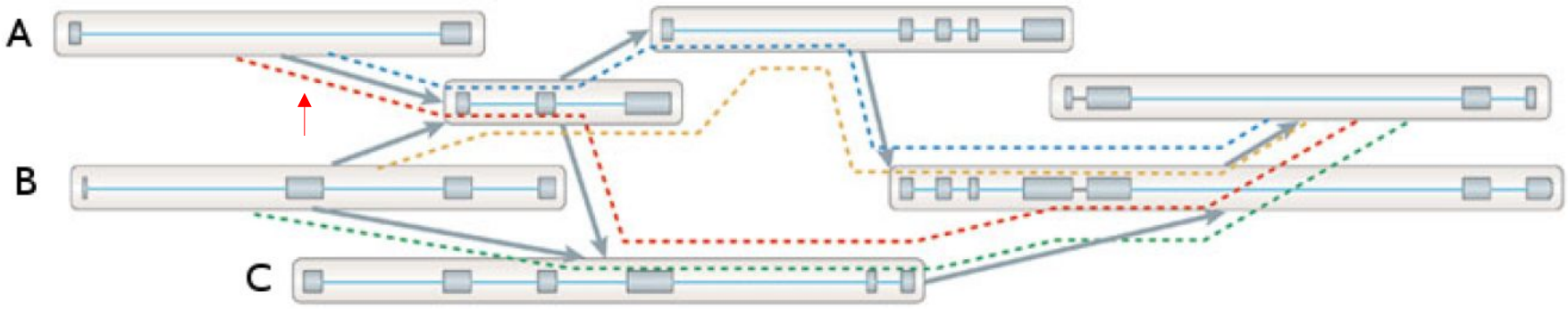
Two reads are **compatible** if their overlap contains the exact same implied introns (or none). If two reads are not compatible they are **incompatible**.



- Read A is incompatible with reads B and C
- Read B is compatible with read C

Compatible reads/fragments

Two reads are **compatible** if their overlap contains the exact same implied introns (or none). If two reads are not compatible they are **incompatible**.



- Read A is incompatible with reads B and C
- Read B is compatible with read C

Alternative transcript events	Total events ($\times 10^3$)	Number detected ($\times 10^3$)	Both isoforms detected	Number tissue-regulated	% Tissue-regulated (observed)	% Tissue-regulated (estimated)
Skipped exon	37	35	10,436	6,822	65	72
Retained intron	1	1	167	96	57	71
Alternative 5′ splice site (A5SS)	15	15	2,168	1,386	64	72
Alternative 3′ splice site (A3SS)	17	16	4,181	2,655	64	74
Mutually exclusive exon (MXE)	4	4	167	95	57	66
Alternative first exon (AFE)	14	13	10,281	5,311	52	63
Alternative last exon (ALE)	9	8	5,246	2,491	47	52
Tandem 3′ UTRs	7	7	5,136	3,801	74	80
Total	105	100	37,782	22,657	60	68

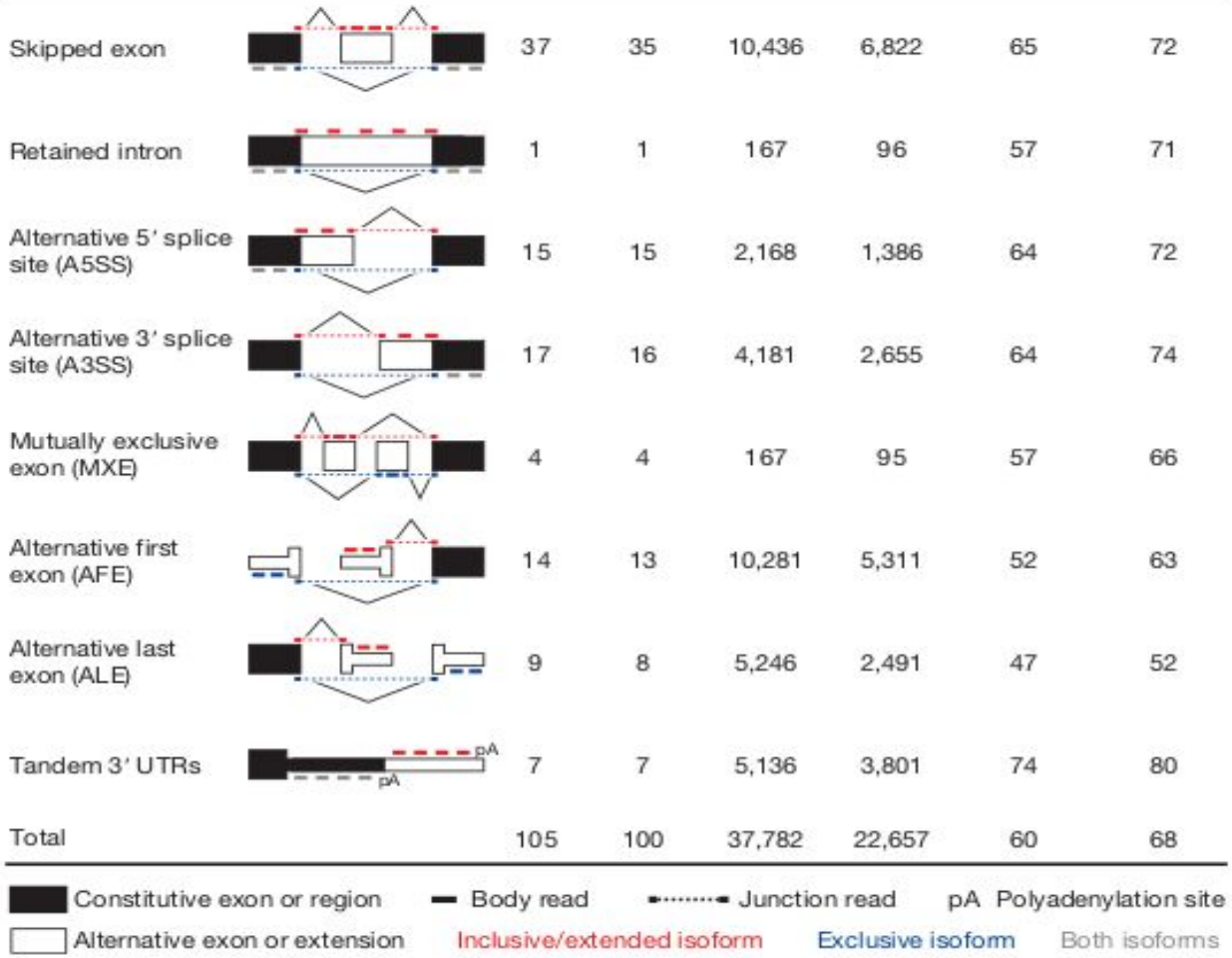
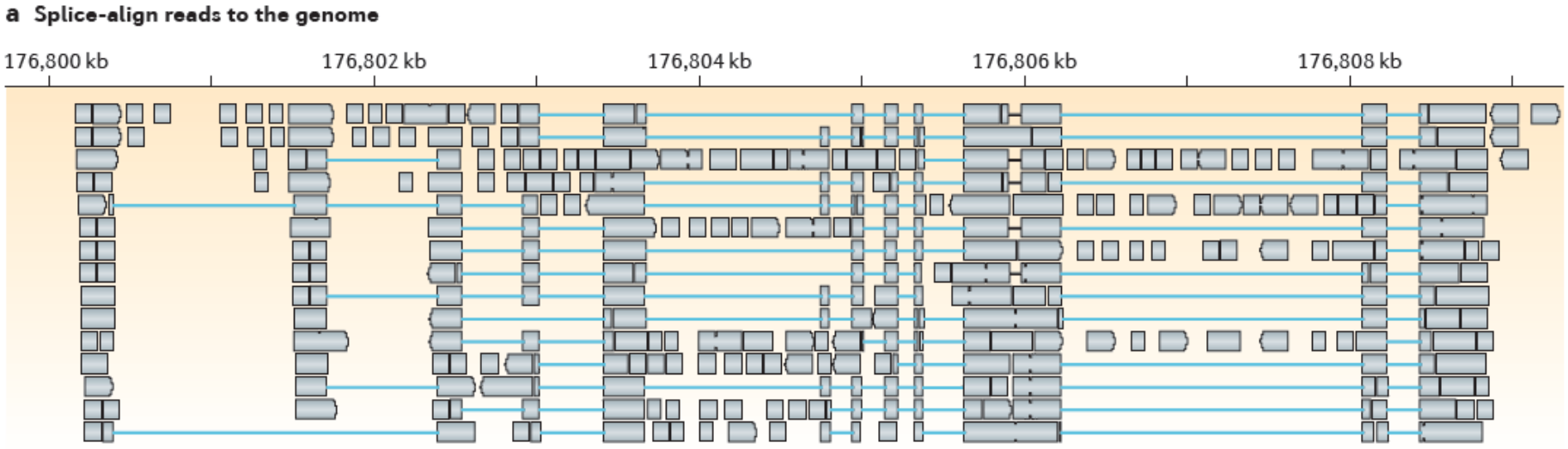


Figure 2 | Pervasive tissue-specific regulation of alternative mRNA isoforms. Rows represent the eight different alternative transcript event types diagrammed. Mapped reads supporting expression of upper isoform, lower isoform or both isoforms are shown in blue, red and grey, respectively. Columns 1–4 show the numbers of events of each type: (1) supported by cDNA and/or EST data; (2) with ≥1 isoform supported by mRNA-Seq reads; (3) with both isoforms supported by reads; and (4) events detected as tissue-regulated (Fisher's exact test) at an FDR of 5% (assuming negligible technical variation[10]). Columns 5 and 6 show: (5) the observed percentage of events with both isoforms detected that were observed to be tissue-regulated; and (6) the estimated true percentage of tissue-regulated isoforms after correction for power to detect tissue bias (Supplementary Fig. 6) and for the FDR. For some event types, 'common reads' (grey bars) were used in lieu of (for tandem 3′ UTR events) or in addition to 'exclusion' reads for detection of changes in isoform levels between tissues.

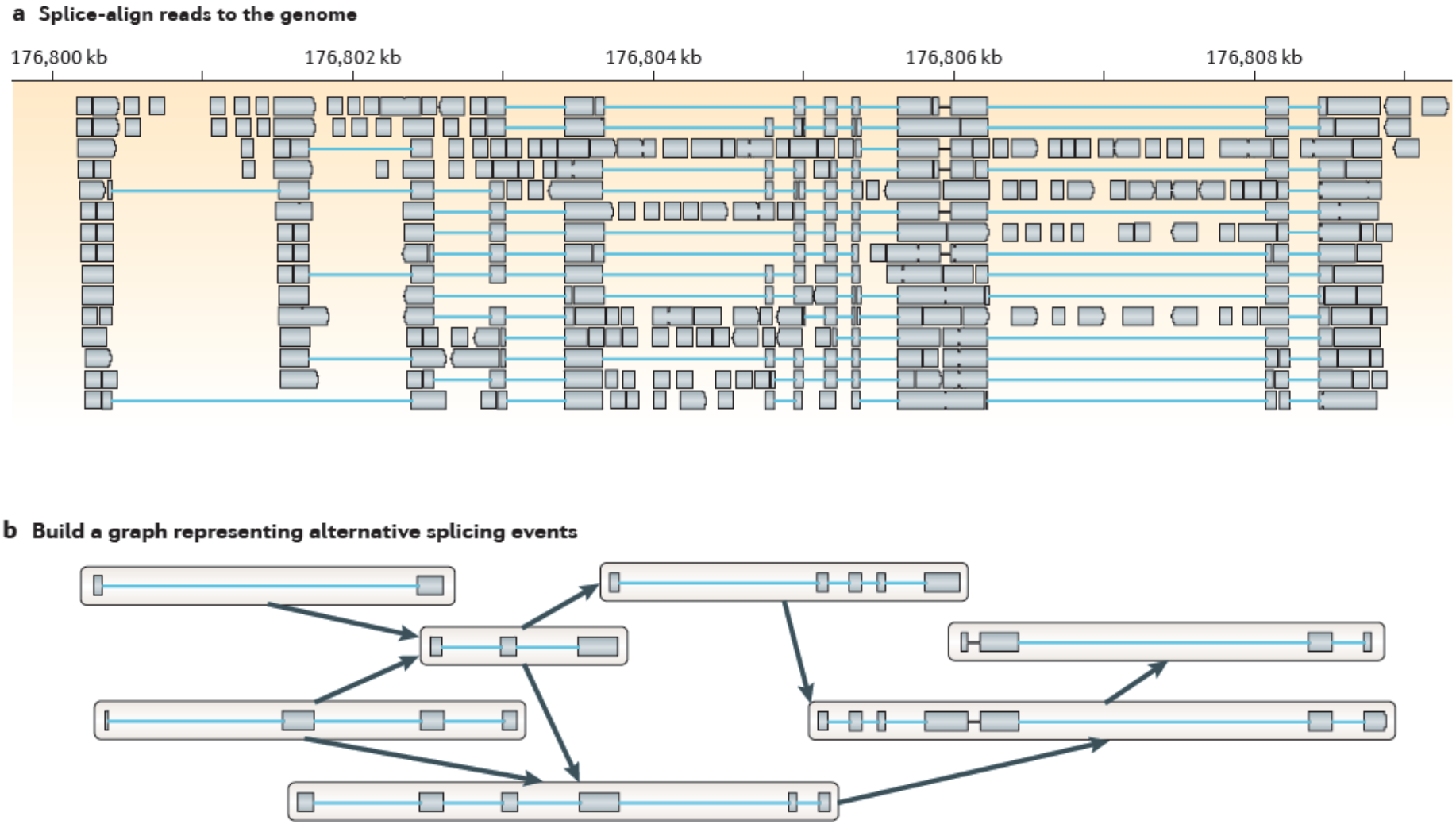
Wang et al., Alternative isoform regulation in human tissue transcriptomes, Nature (2008)

Transcript Reconstruction Using Cufflinks



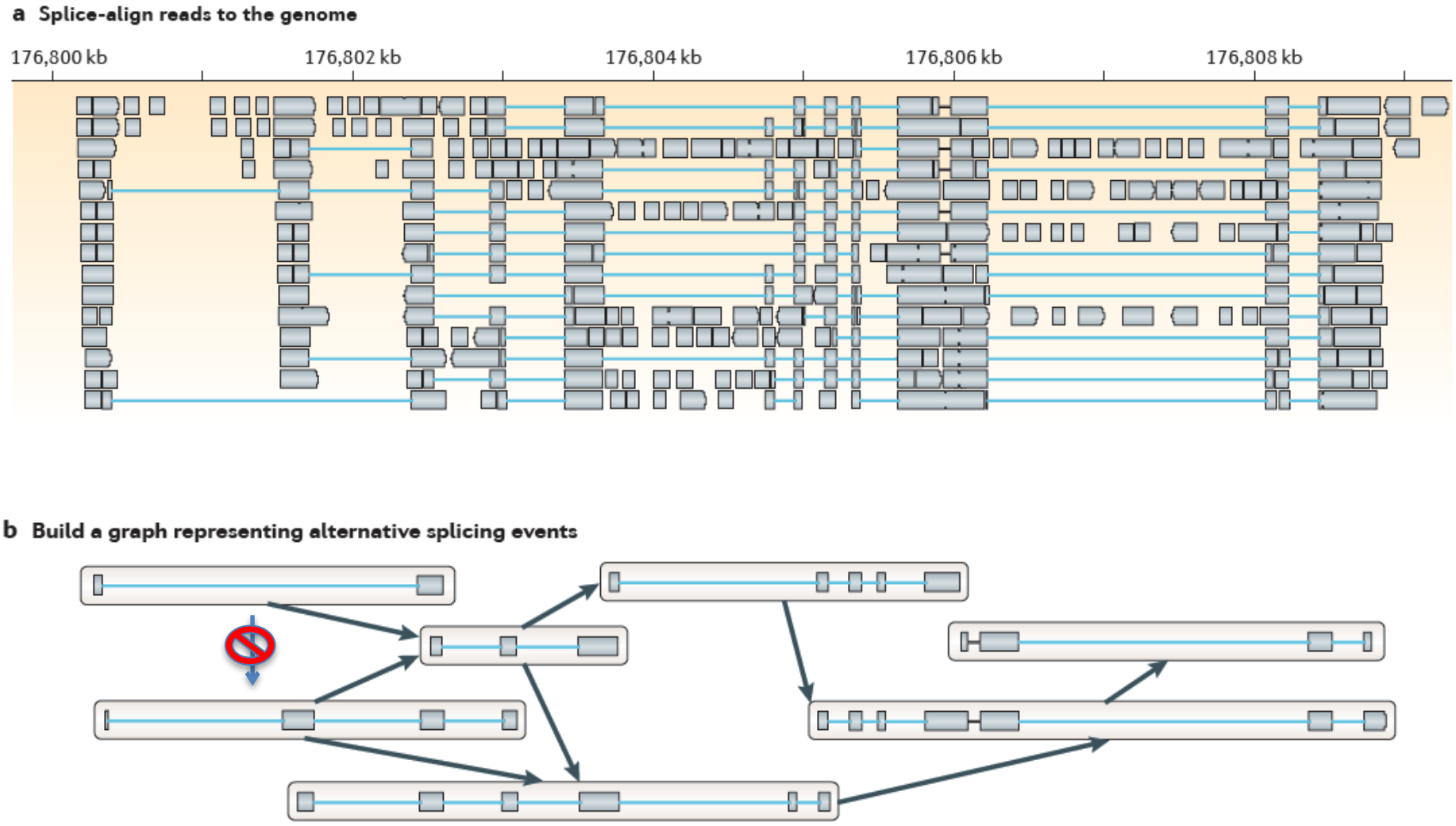
From Martin & Wang. Nature Reviews in Genetics. 2011

Transcript Reconstruction Using Cufflinks



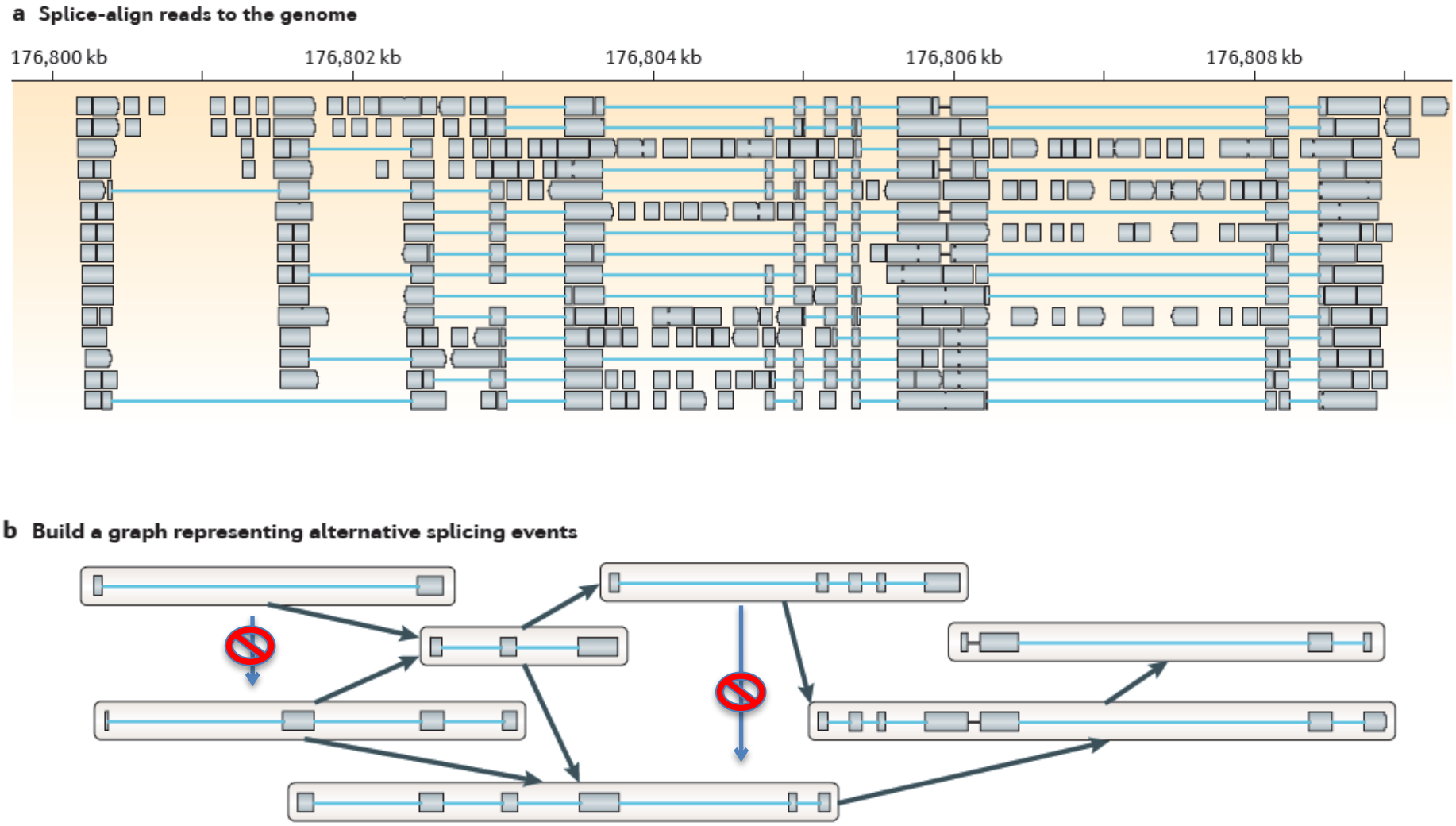
From Martin & Wang. Nature Reviews in Genetics. 2011

Transcript Reconstruction Using Cufflinks



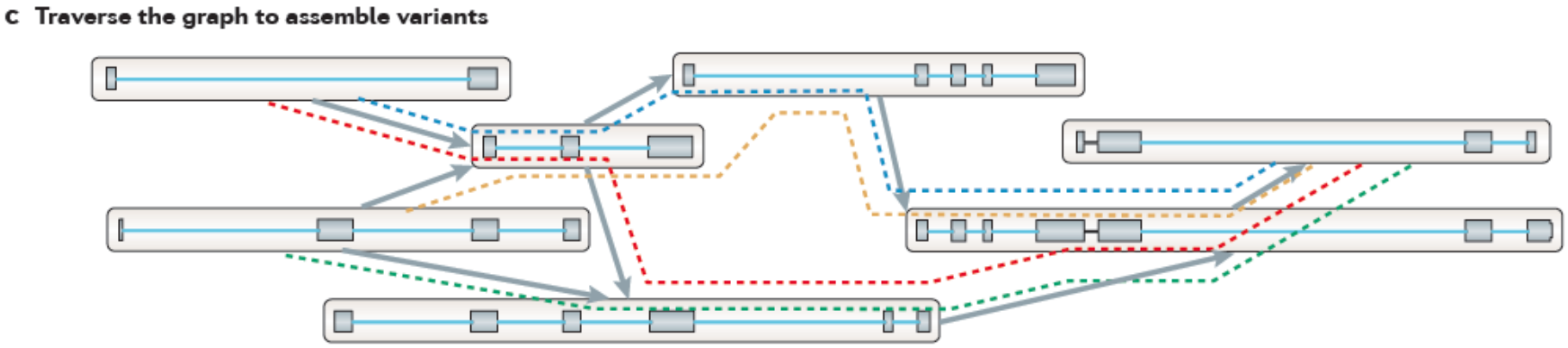
From Martin & Wang. Nature Reviews in Genetics. 2011

Transcript Reconstruction Using Cufflinks



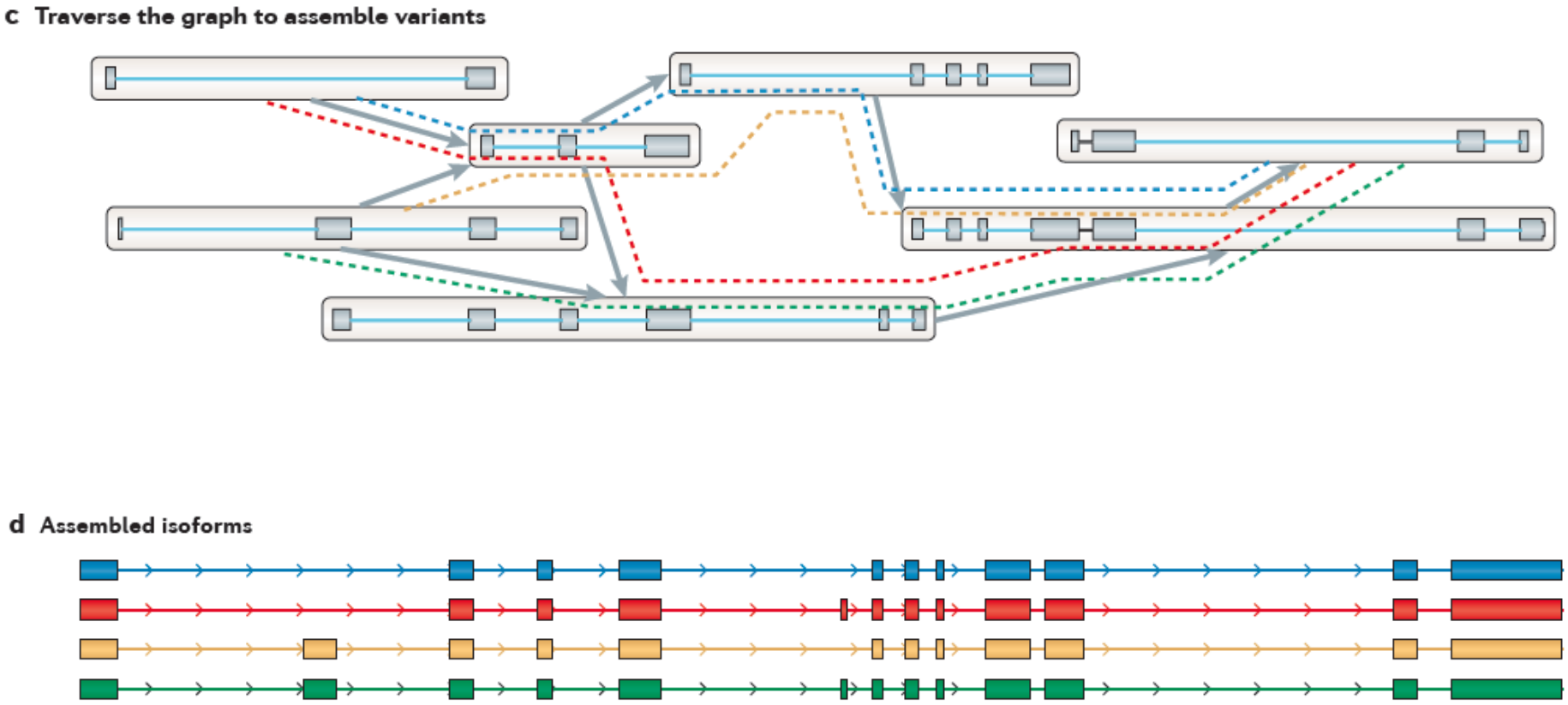
From Martin & Wang. Nature Reviews in Genetics. 2011

Transcript Reconstruction Using Cufflinks



From Martin & Wang. Nature Reviews in Genetics. 2011

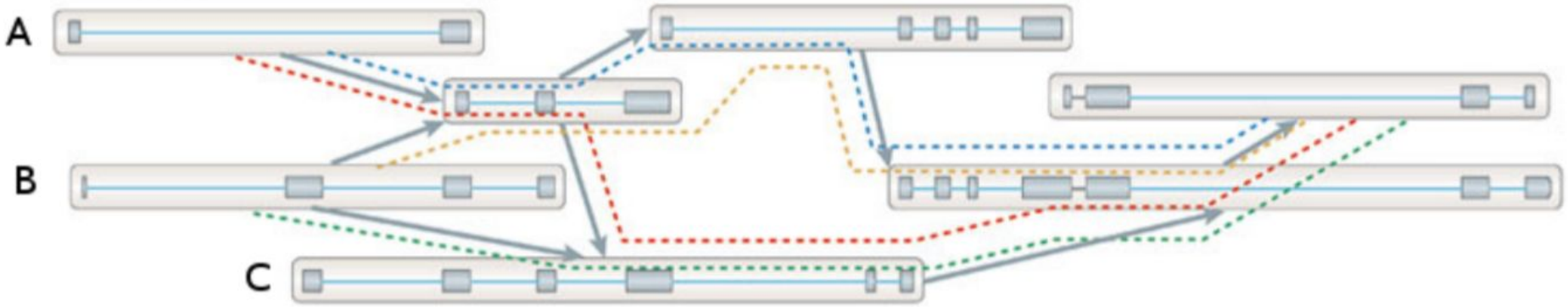
Transcript Reconstruction Using Cufflinks



From Martin & Wang. Nature Reviews in Genetics. 2011

Dilworth's theorem

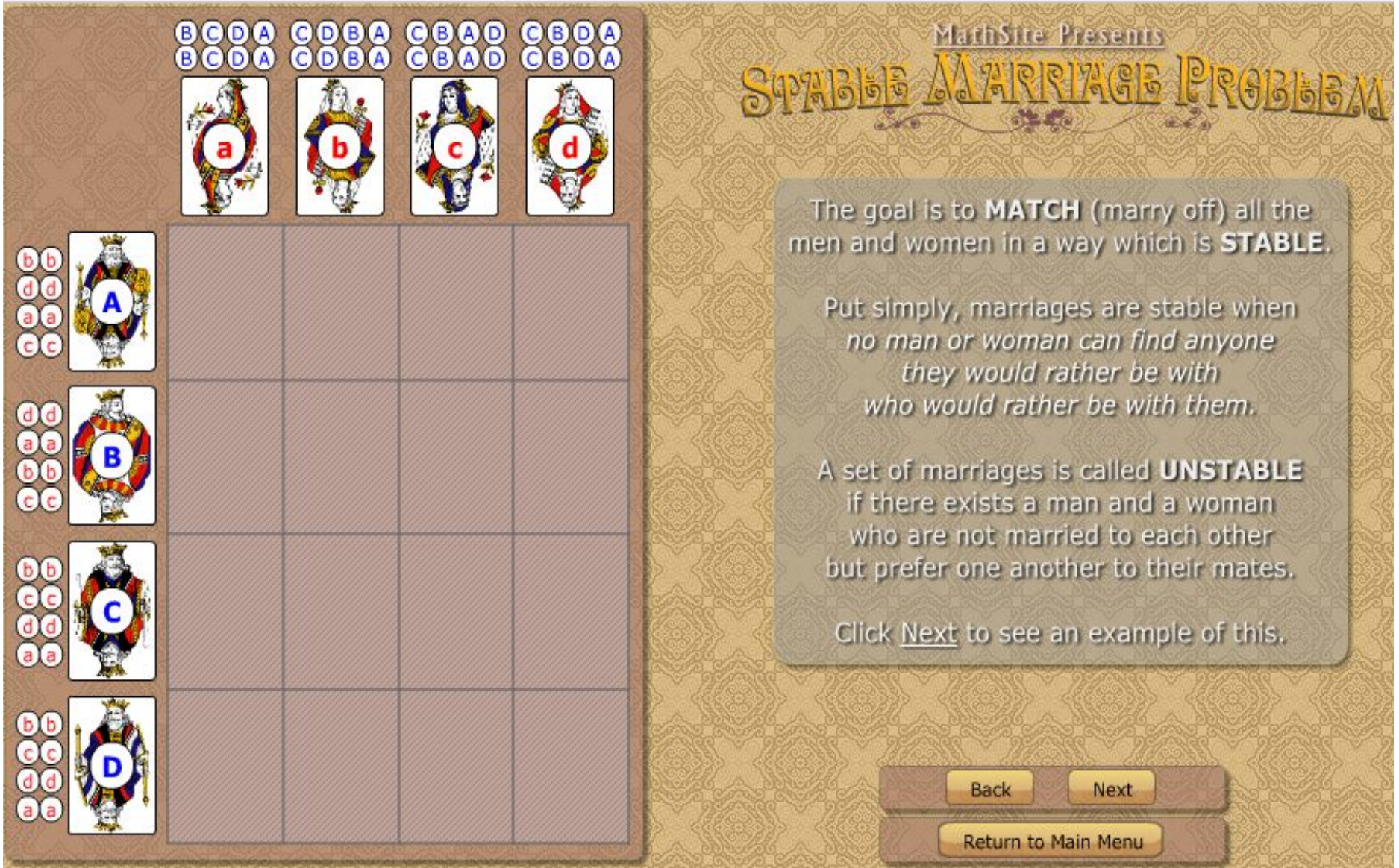
- In the setting of RNA-seq, this essentially means that the maximum cardinality of a set of fragments that are pairwise incompatible is the same as the minimum number of isoforms needed to explain the reads.

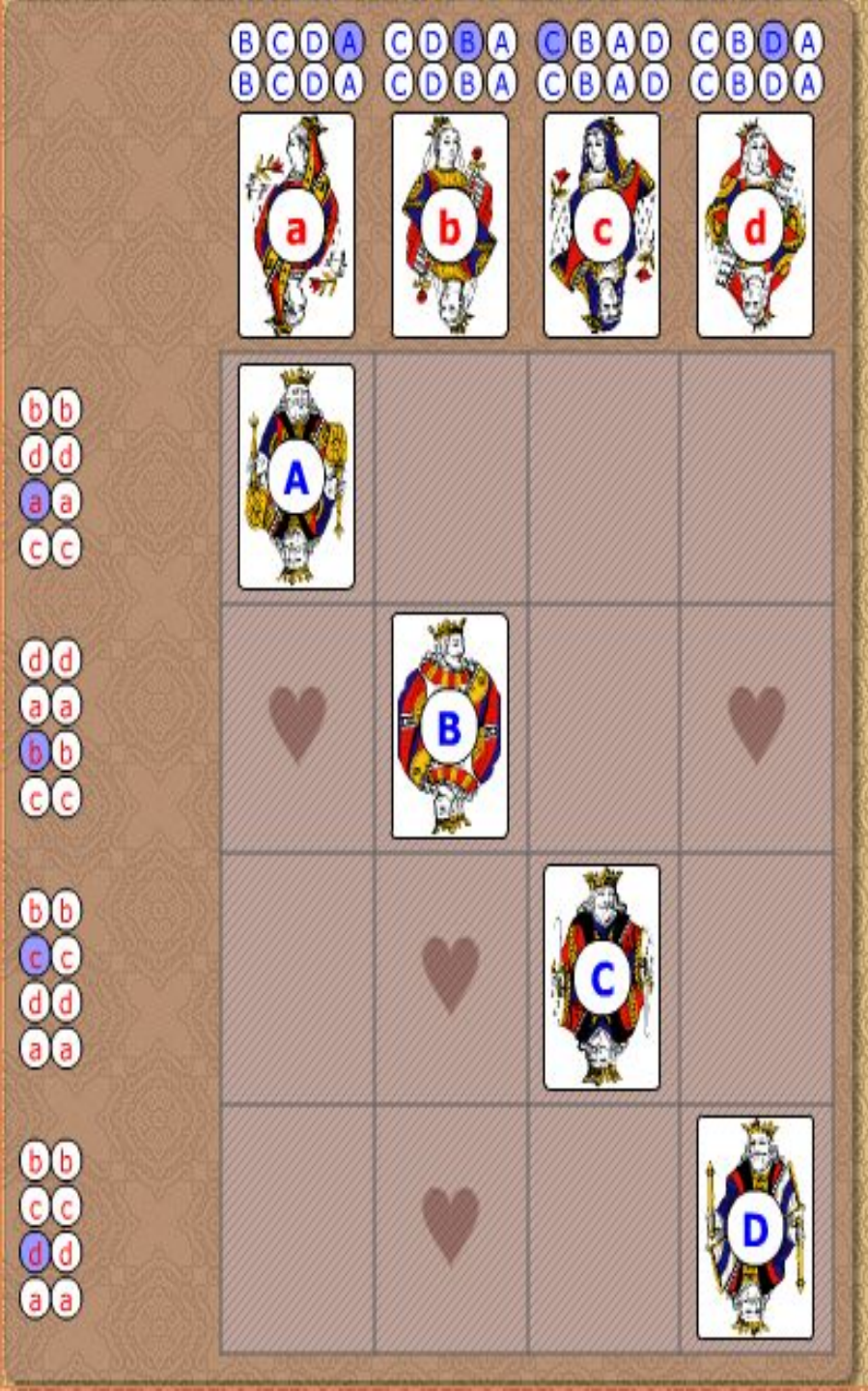


More details at:
http://www.mi.fu-berlin.de/wiki/pub/ABI/GenomicsLecture12Materials/rnaseq1.pdf

'Marrying' reads with transcripts

http://mathsite.math.berkeley.edu/smp/smp.html





MathSite Presents

Stable Marriage Problem

This arrangement is **UNSTABLE**.
The hearts show potential matchings that **DESTABILIZE** the current marriages.
They are called **blocking pairs**.

For example, look at man **C** and woman **b**.
They both prefer each other to their spouses.
Currently, man **C** gets his 2nd choice
and woman **b** gets her 3rd choice.
But if they were married to each other,
they would *both* get their 1st choice.

Do you see how the other three blocking pairs destabilize the current set of marriages?

Back Next

Return to Main Menu

MathSite Presents

Stable Marriage Problem

The Gale-Shapley algorithm works like this:

Each man proposes to his highest ranked woman.
If a woman is not engaged she automatically accepts.
If she *is* engaged she picks the more preferred man.
The rejected man moves on to his next desired woman.
When each man is engaged the problem is solved.
(The order in which men propose is not important.)

This will be clearer in the next example.

MathSite Presents

Stable Marriage Problem

Let's start with the first steps:

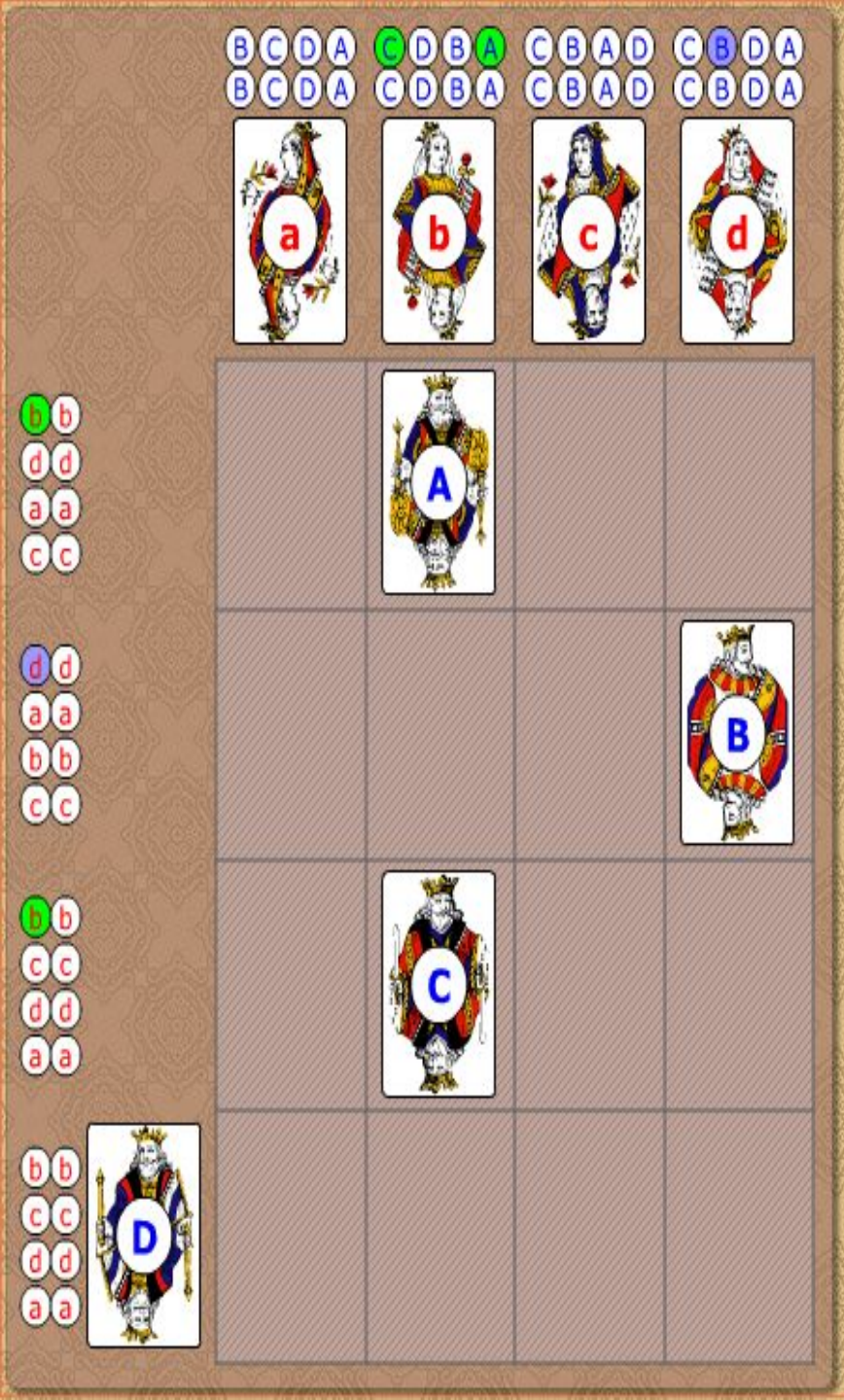
Each man proposes to his highest ranked woman.
If a woman is not engaged she automatically accepts.

We go down the list of men, starting with A,
and propose to the most preferred women.
A proposes to b and B proposes to d.

Click Next to see the other men propose.

Back Next

Return to Main Menu



MathSite Presents

Stable Marriage Problem

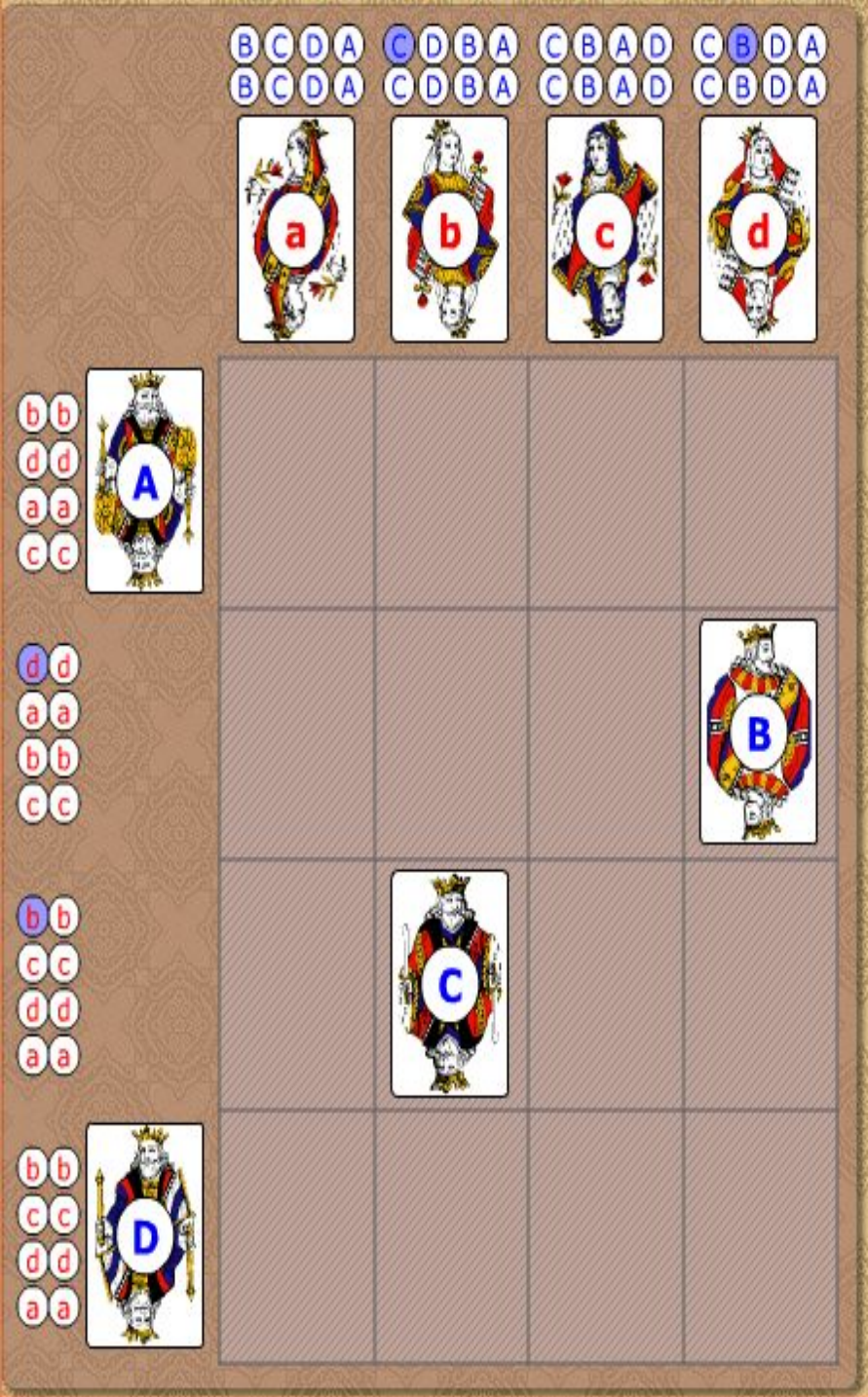
Man C prefers b the most, but she is engaged.

If a woman is engaged she picks the more preferred man.

Woman b must decide between men A and C.
Who will she pick?

Back Next

Return to Main Menu



MathSite Presents

Stable Marriage Problem

Woman b prefers man C over man A.
She rejects A and accepts C's proposal.

Back Next

Return to Main Menu

MathSite Presents

Stable Marriage Problem

The rejected man moves on to his next desired woman.

Man A prefers woman d next.
Now woman d must decide between A and B.

Who will she pick?

Back Next

Return to Main Menu

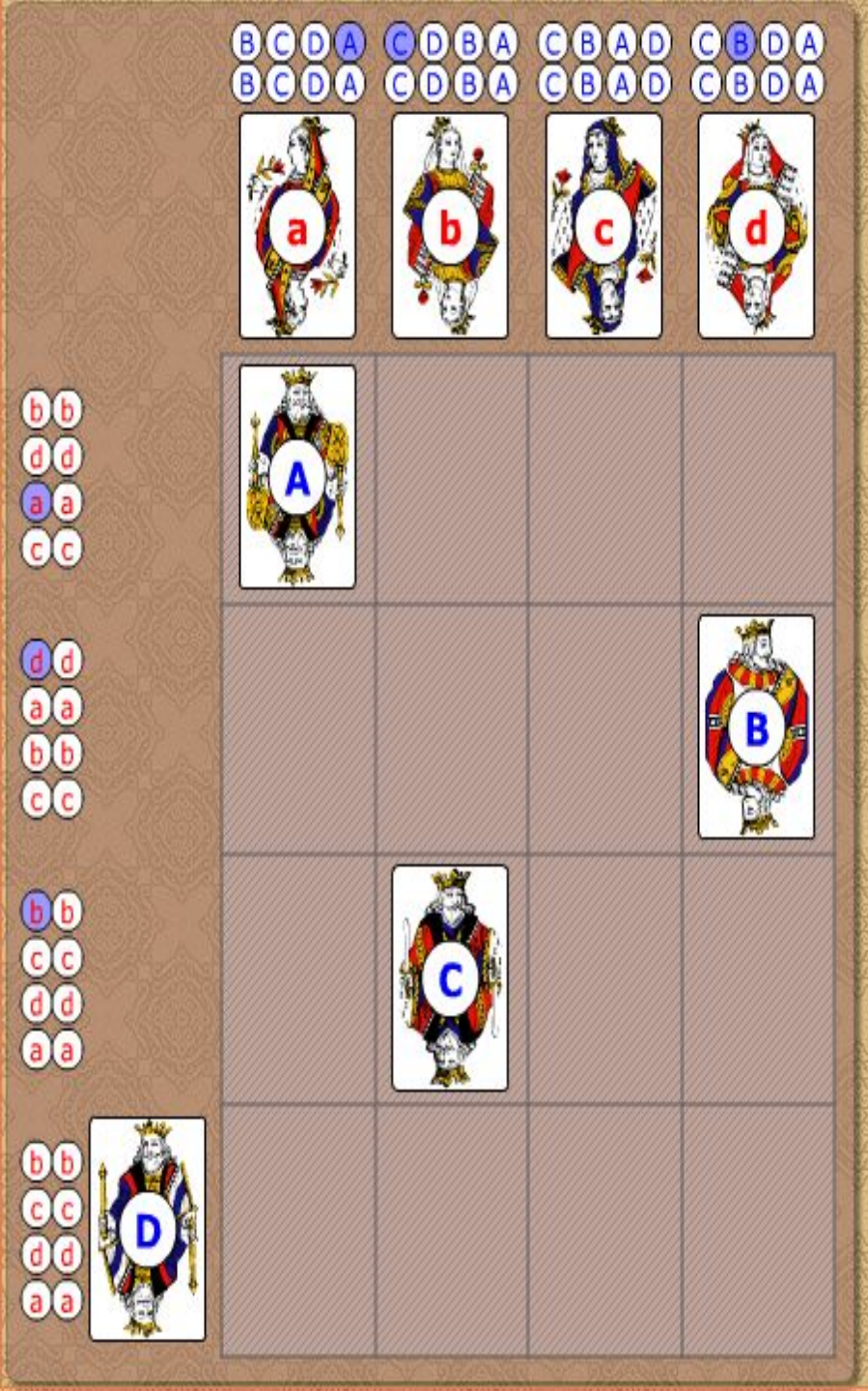
MathSite Presents

Stable Marriage Problem

Woman d stays with man B.

Back Next

Return to Main Menu



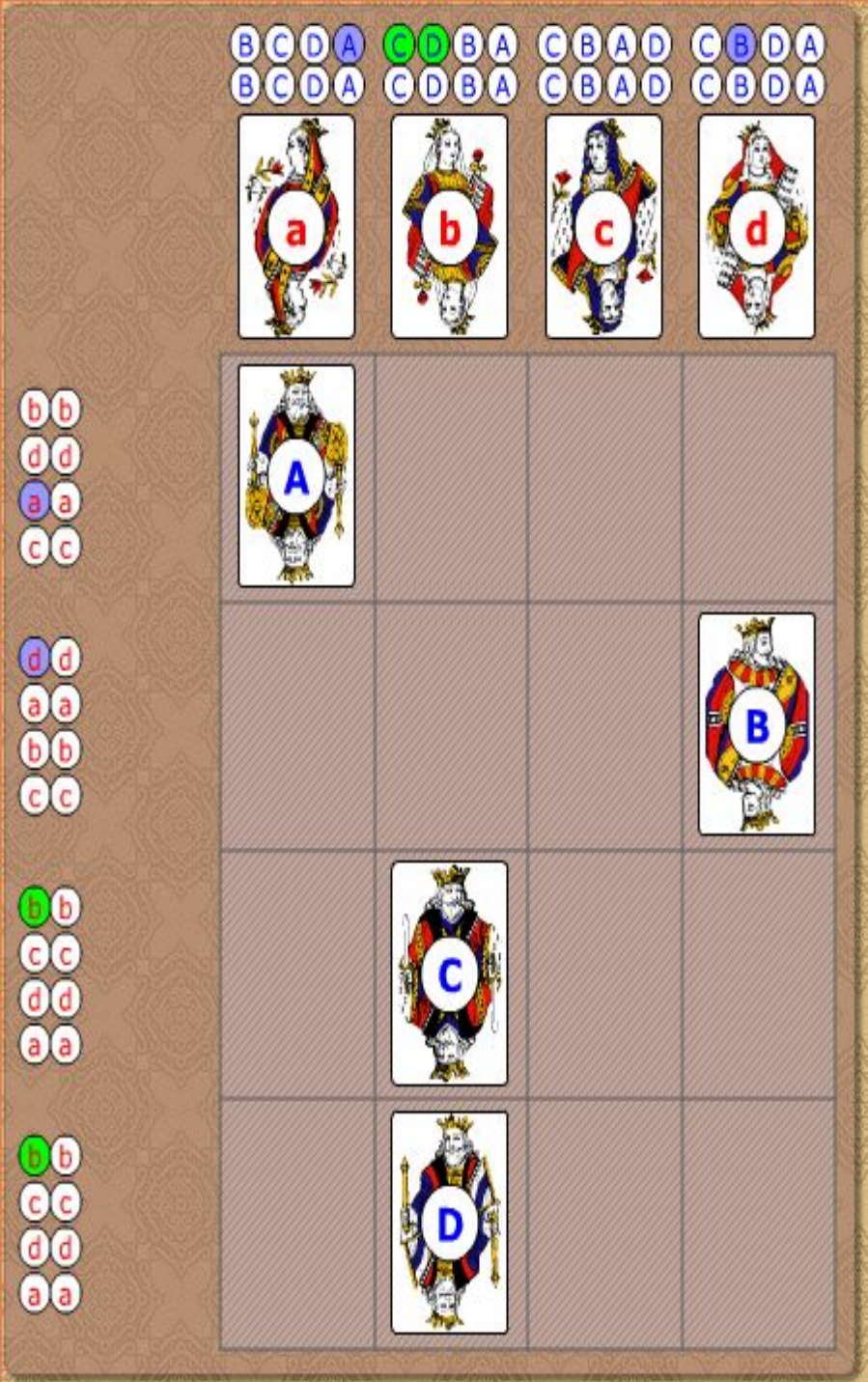
MathSite Presents

Stable Marriage Problem

Man A now goes on to his 3rd choice, woman a.
Woman a is not engaged, so she accepts.

Back Next

Return to Main Menu



MathSite Presents

Stable Marriage Problem

Next, man D seeks his first choice, b.
Woman b is already engaged.
Who will she pick?

Back Next

Return to Main Menu

MathSite Presents
STABLE MARRIAGE PROBLEM
B C D A
B C D A
C D B A
C D B A
C B A D
C B A D
C B D A
C B D A
a
b
c
d
A
B
C
D
b b
d d
a a
c c
d d
a a
b b
c c
b b
c c
d d
a a
b b
c c
d d
a a
Woman b rejects D and stays with C.
Back
Next
Return to Main Menu

MathSite Presents

Stable Marriage Problem

Man D moves on to his next choice, c.

When each man is engaged the problem is solved.

The marriages are now stable:
No person can break their engagement for a more desired partner who would be willing to do the same.
The algorithm has generated a stable matching.

Back Next

Return to Main Menu

MathSite Presents

Stable Marriage Problem

The Stable Marriage Problem has many real-world applications.

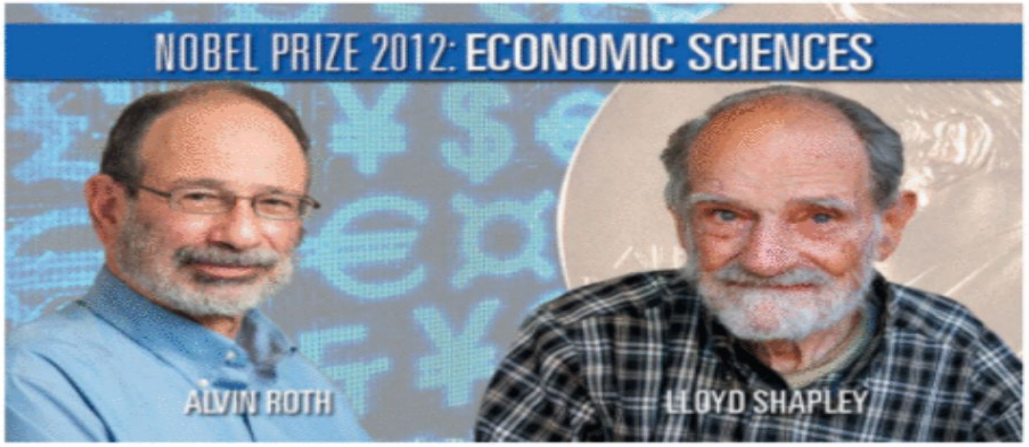
Every year in the US, some thirty thousand graduating medical school students are matched with hospitals. The hospitals and students rank each other and the *National Residents Matching Program* then arranges a stable matching between them. The NRMP used to use a hospital-optimal (hospitals proposing) matching, but has recently changed to a student-optimal arrangement.

Although the NRMP uses a more complicated algorithm than the one we have seen, this should nonetheless give you an idea of how matching is applied in the real world.

Economics Nobel Honors Matchmaking Finesse

How do you make a perfect match—between doctors and hospitals, between schools and students, or even between kidneys and patients? Lloyd Shapley and Alvin Roth have won the 2012 Nobel Prize in economics for helping to answer that question. Shapley, 89, is a professor emeritus at the University of California, Los Angeles, and Roth, 60, is a professor at Harvard University and Harvard Business School in Boston.

The Nobel committee awarded the prize to the two economists "for the theory of stable allocations and the practice of market design." Shapley pioneered theoretical concepts to understand and solve the matching problem; Roth clarified those ideas and applied them to engineer algorithms that are now widely used in the real world.



Working in the 1950s and 1960s, Shapley used cooperative game theory to explore matching. Trade between a group of rational actors, Shapley reasoned, would reach a stable state if no individual had anything further to gain by making a new trade. Along with the late economist and mathematician David Gale, Shapley applied this idea of stability to the case of marriages, setting up an example of pairing 10 women and 10 men into opposite-sex couples.

The two came up with a matchmaking method called the "deferred acceptance" algorithm, which could proceed in either of two ways. In one scenario, the men propose to the women, and each woman rejects the men she finds unsuitable while holding on to—but not yet accepting—the proposal she likes best. In the next round, the rejected men propose to their second-best choice, and the women again retain or choose what they view as the best offer while rejecting the rest. Over several rounds, the pairings become stable.

In the other scenario, the women propose to the men. The result is a different set of pairings, also stable. Although stable matches emerge no matter who does the choosing, Gale and Shapley showed that women end up with a more favorable outcome when they propose to men, and vice versa.

In the 1980s, Roth explored matching between medical interns and hospitals and showed that the algorithm then in use by the National Resident Matching Program (NRMP)—a clearinghouse that matched residency applicants to hospitals—was successful because it followed principles similar to those encoded by the Gale-Shapley algorithm. In the mid-1990s, Roth was asked to improve on the NRMP algorithm to make matching more efficient and better at accommodating special needs such as placing dual-doctor couples in the same hospital. Roth delivered an improved algorithm, working with Elliott Peranson, the founder of the Toronto, Canada–based company National Matching Services Inc.

Roth went on to apply the Gale-Shapley algorithm to redesign the admission process used by public high schools in New York City. The new process has led to better matching between schools and students, reducing by 90% the number of students assigned to schools they didn't list as preferences.

The Gale-Shapley algorithm has also been applied, with important modifications, to matching kidneys with patients. New versions of the algorithm are also being used to conduct Internet auctions in which search companies sell advertising space. The Nobel committee's announcement of the prize this morning describes the work being honored as "an outstanding example of economic engineering."

Roth's keenness to seek out real-world problems to solve makes him stand out among academic economists, says Peranson, who has been a personal friend of Roth for years and has watched him speak at university symposiums. "The Ph.D. students are very enamored with Roth because when they hear him talk, they realize, 'Not only am I studying something that's interesting, but it could have real impact.'"

Roth, who is moving to Stanford University next year, is the author of a popular blog, "Market Design" (marketdesigner.blogspot.com). "Blog may be delayed today," he posted on Monday morning, shortly after getting that life-changing phone call from Stockholm. "Count me as surprised."

–YUDHIJIT BHATTACHARJEE

CREDITS (LEFT TO RIGHT): LINDA A. CICERO/STANFORD UNIVERSITY; JONATHAN ALCORN/REUTERS/LANDOV; (BACKGROUND) ISTOCKPHOTO.COM; NOBEL FOUNDATION

Novel isoform discovery

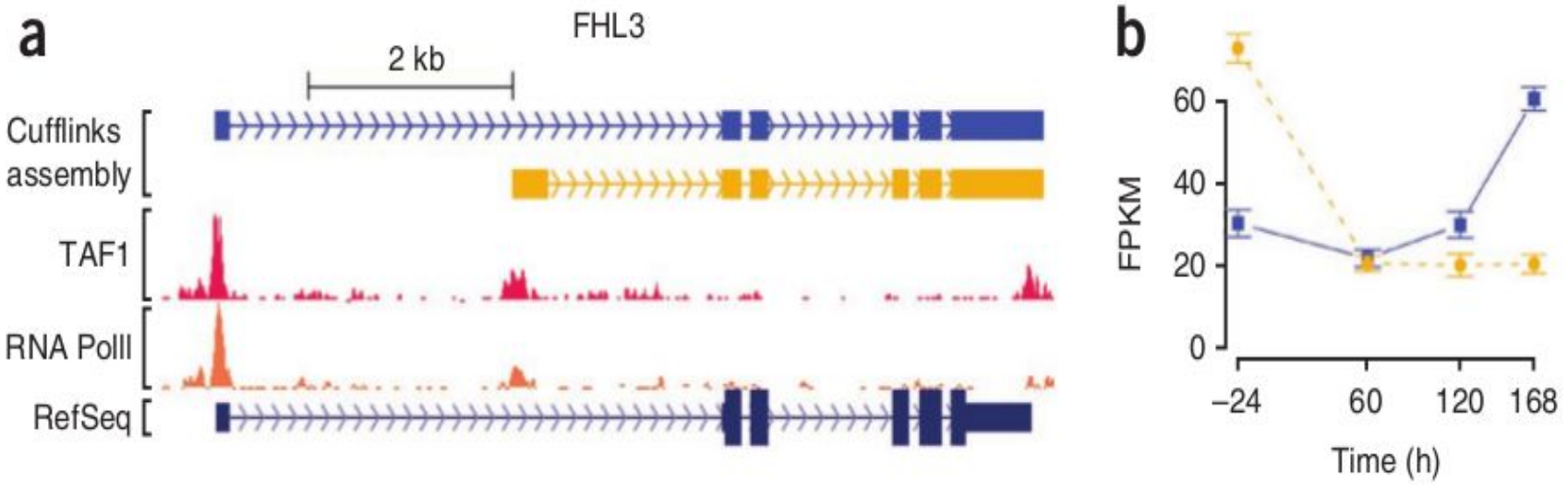
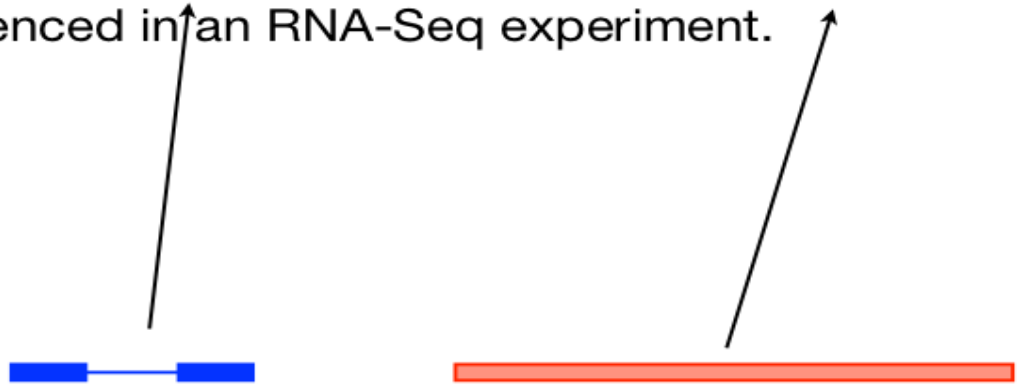


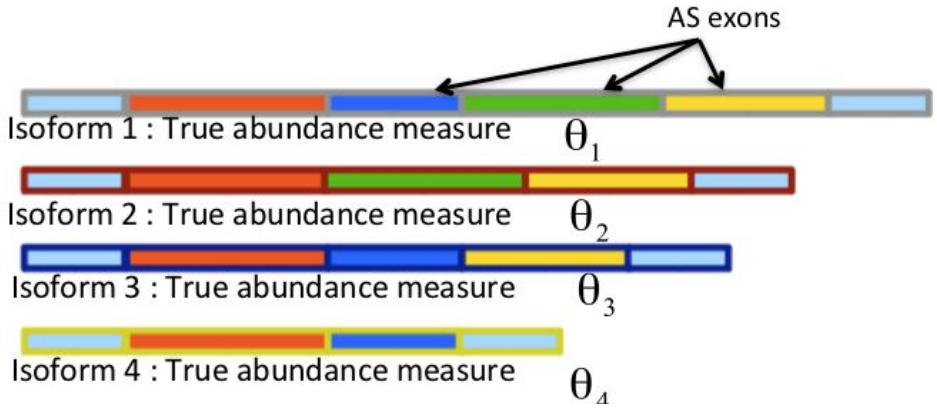
Figure 3 Excluding isoforms discovered by Cufflinks from the transcript abundance estimation affects the abundance estimates of known isoforms, in some cases by orders of magnitude. FHL3 inhibits myogenesis by binding MyoD and attenuating its transcriptional activity. (**a**) The C2C12 transcriptome contains a novel isoform that is dominant during proliferation. The new TSS for FHL3 is supported by proximal TAF1 and RNA polymerase II ChIP-Seq peaks. (**b**) The known isoform (solid line) is preferred at time points following differentiation.

FPKM

- Expected number of **F**ragments **P**er **K**ilobase (of transcript) per **M**illion fragments sequenced in an RNA-Seq experiment.



- These units are proportional to the θ_i.



FPKM (Fragments Per Kilobase per Million) vs TPM (Transcipts per Million)

Suppose we have RNA-seq read counts for three genes in one sample:

Gene	Length (kb)	Raw read counts
Gene A	2 kb	500
Gene B	1 kb	1000
Gene C	0.5 kb	500

We first compute **Reads Per Kilobase (RPK)**:

$$RPK = \frac{\text{Raw counts}}{\text{Gene length (kb)}}$$

Gene	Raw counts	Length (kb)	RPK
A	500	2	250
B	1000	1	1000
C	500	0.5	1000

🔹 Step 2a: FPKM

- Compute scaling factor = total mapped reads (in millions).
  Suppose total mapped reads = 2,000,000 → 2 million.
- FPKM = (RPK ÷ total mapped reads in millions).

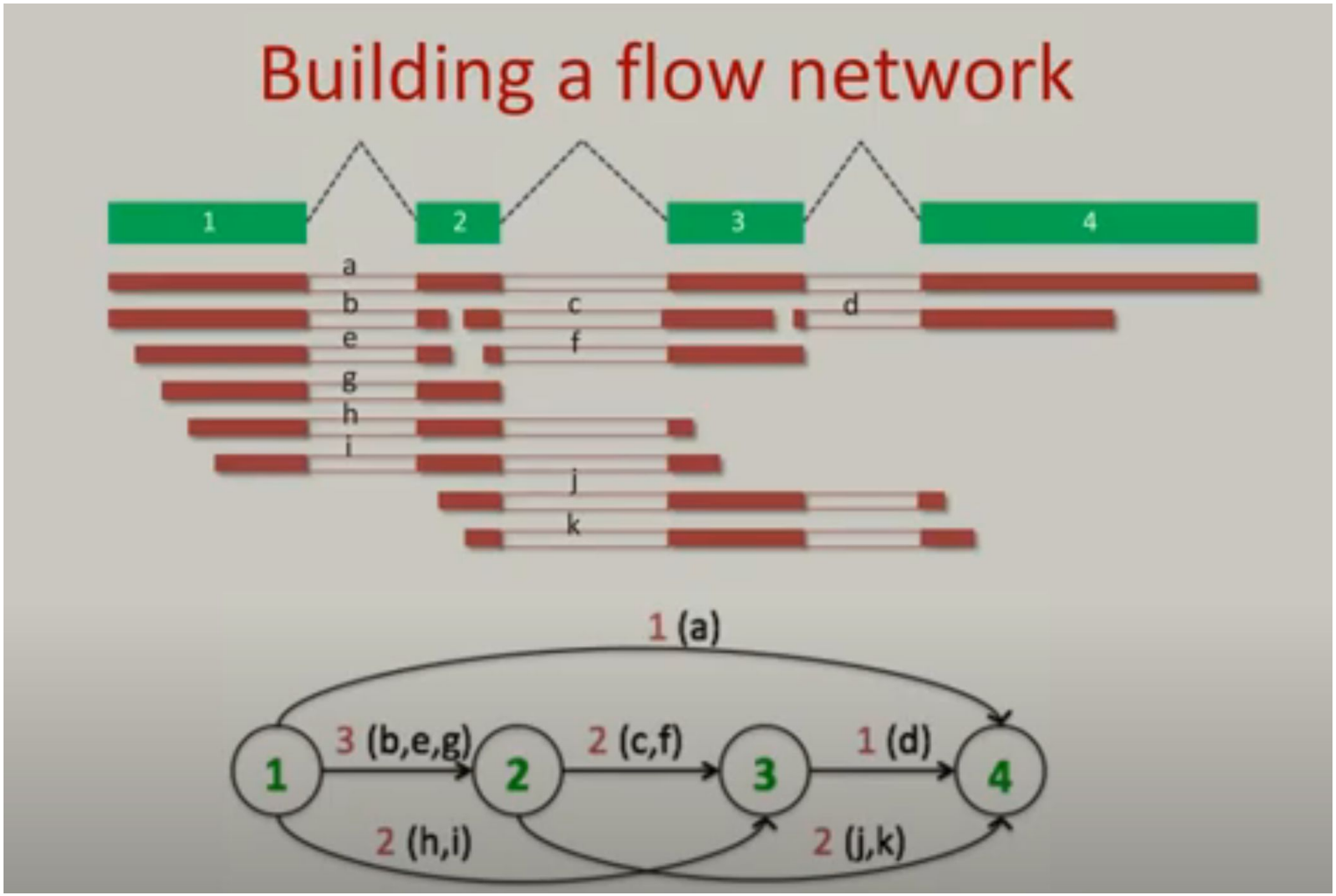
Gene	RPK	FPKM
A	250	125
B	1000	500
C	1000	500

🔹 Step 2b: TPM

- First sum all RPK values: 250 + 1000 + 1000 = 2250.
- Divide each gene's RPK by the sum, then multiply by 1,000,000.

Gene	RPK	TPM
A	250	111,111
B	1000	444,444
C	1000	444,444

STRINGTIE (better than Cufflinks, see www.youtube.com/watch?v=2qGiw4MRK3c)



Network flow algorithm, see https://www.youtube.com/watch?v=Tl90tNtKvxs&t=66s

FPKM Tracking Files

Column number	Column name	Example	Description
1	tracking_id	TCONS_00000001	A unique identifier describing the object (gene, transcript, CDS, primary transcript)
2	class_code	=	The class_code attribute for the object, or "-" if not a transcript, or if class_code isn't present
3	nearest_ref_id	NM_008866.1	The reference transcript to which the class code refers, if any
4	gene_id	NM_008866	The gene_id(s) associated with the object
5	gene_short_name	Lypla1	The gene_short_name(s) associated with the object
6	tss_id	TSS1	The tss_id associated with the object, or "-" if not a transcript/primary transcript, or if tss_id isn't present
7	locus	chr1:4797771-4835363	Genomic coordinates for easy browsing to the object
8	length	2447	The number of base pairs in the transcript, or '-' if not a transcript/primary transcript
9	coverage	43.4279	Estimate for the absolute depth of read coverage across the object
10	q0_FPKM	8.01089	FPKM of the object in sample 0
11	q0_FPKM_lo	7.03583	the lower bound of the 95% confidence interval on the FPKM of the object in sample 0
12	q0_FPKM_hi	8.98595	the upper bound of the 95% confidence interval on the FPKM of the object in sample 0
13	q0_status	OK	Quantification status for the object in sample 0. Can be one of OK (deconvolution successful), LOWDATA (too complex or shallowly sequenced), HIDATA (too many fragments in locus), or FAIL, when an ill-conditioned covariance matrix or other numerical exception prevents deconvolution.

Class Codes

Priority	Code	Description
1	=	Complete match of intron chain
2	c	Contained
3	j	Potentially novel isoform (fragment): at least one splice junction is shared with a reference transcript
4	e	Single exon transfrag overlapping a reference exon and at least 10 bp of a reference intron, indicating a possible pre-mRNA fragment.
5	i	A transfrag falling entirely within a reference intron
6	o	Generic exonic overlap with a reference transcript
7	p	Possible polymerase run-on fragment (within 2Kbases of a reference transcript)
8	r	Repeat. Currently determined by looking at the soft-masked reference sequence and applied to transcripts where at least 50% of the bases are lower case
9	u	Unknown, intergenic transcript
10	x	Exonic overlap with reference on the opposite strand
11	s	An intron of the transfrag overlaps a reference intron on the opposite strand (likely due to read mapping errors)
12	.	(.tracking file only, indicates multiple classifications)

Cufflinks pipelines

Discovering novel genes and transcripts

1 Map the reads for each tissue to the reference genome

2 Run Cufflinks on each mapping file

3 Merge the resulting assemblies

4 (optional) Compare the merged assembly with known or annotated genes

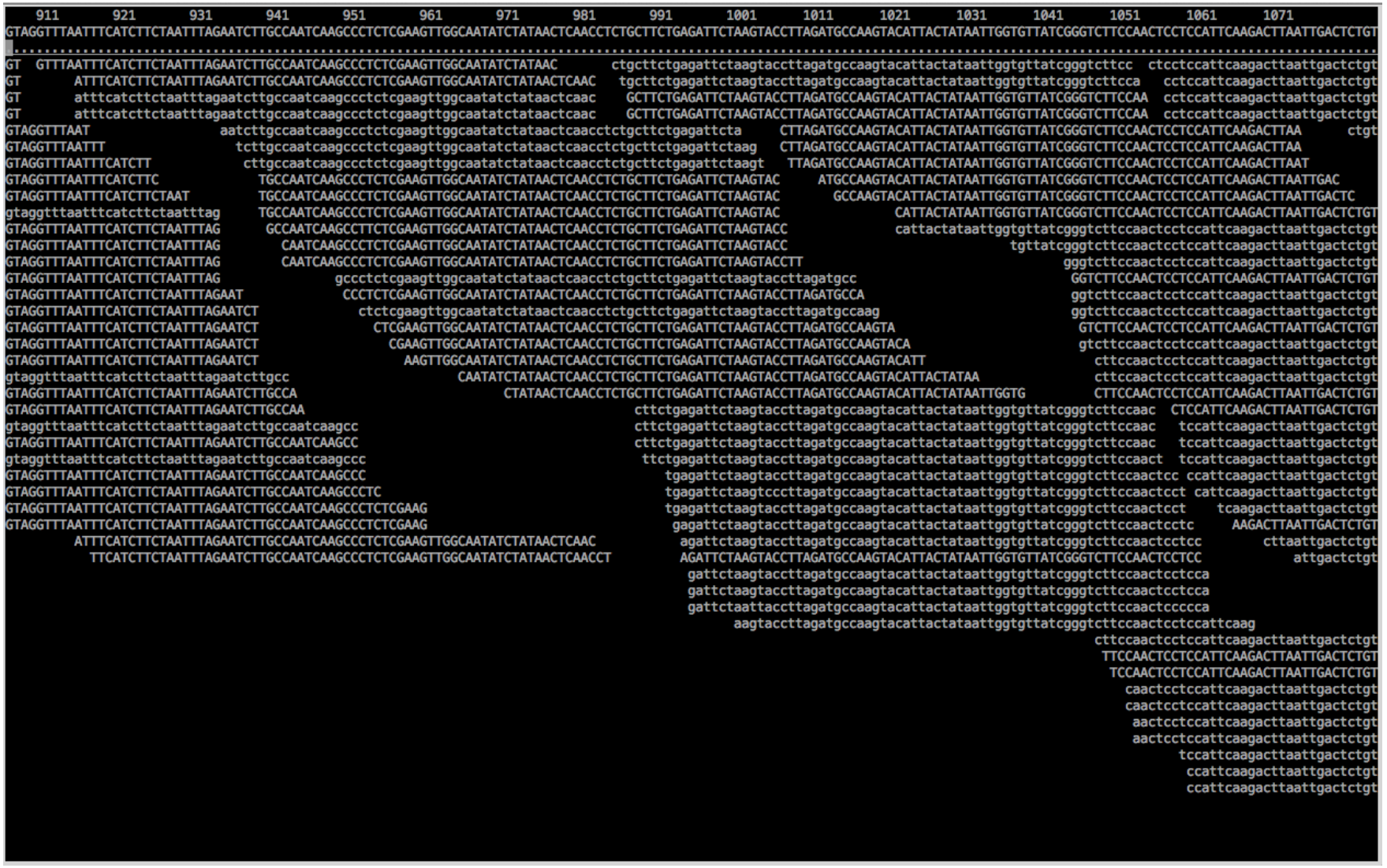
Differential expression

5 Run cuffdiff

Visualizing Alignments of RNA-Seq reads

Text-based Alignment Viewer

% samtools tview alignments.bam target.fasta



IGV

Integrative Genomics Viewer

- Home
- Downloads
- Documents
  - IGV User Guide
  - Tutorial Videos
  - File Formats
  - Hosted Genomes
  - FAQ
  - Release Notes
  - Credits
- Contact

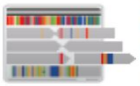
Search website

search

© 2013-2021
Broad Institute
and the Regents of the
University of California

Home

Overview



The **Integrative Genomics Viewer (IGV)** is a high-performance, easy-to-use, interactive tool for the visual exploration of genomic data. It supports flexible integration of all the common types of genomic data and metadata, investigator-generated or publicly available, loaded from local or cloud sources.

IGV is available in multiple forms, including:

- the original **IGV** - a Java desktop application,
- **IGV-Web** - a web application,
- **igv.js** - a JavaScript component that can be embedded in web pages *(for developers)*

This site is focused on the IGV desktop application. See https://igv.org for links to all forms of IGV.

Download IGV

Download the IGV desktop application and igvtools.

Citing IGV

To cite your use of IGV in your publication, please reference one or more of:

James T. Robinson, Helga Thorvaldsdóttir, Wendy Winckler, Mitchell Guttman, Eric S. Lander, Gad Getz, Jill P. Mesirov. **Integrative Genomics Viewer**. Nature Biotechnology 29, 24–26 (2011). (Free PMC article here).

Helga Thorvaldsdóttir, James T. Robinson, Jill P. Mesirov. **Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration**. Briefings in Bioinformatics 14, 178-192 (2013).

James T. Robinson, Helga Thorvaldsdóttir, Aaron M. Wenger, Ahmet Zehir, Jill P. Mesirov. **Variant Review with the Integrative Genomics Viewer (IGV).** Cancer Research 77(21) 31-34 (2017).

James T. Robinson, Helga Thorvaldsdóttir, Douglass Turner, Jill P. Mesirov. **igv.js: an embeddable JavaScript implementation of the Integrative Genomics Viewer (IGV)**. bioRxiv 2020.05.03075499.

IGV: Viewing Tophat Alignments

Transcript Structures in **GTF** Format

(tab-delimited fields per line shown transposed to a column format here)

0	7000000090838467
1	Cufflinks
2	transcript
3	101
4	5716
5	1000
6	+
7	.
8	gene_id "CUFF.1"; transcript_id "CUFF.1.1"; FPKM "378.0239937260“

0	7000000090838467
1	Cufflinks
2	exon
3	101
4	5716
5	1000
6	+
7	.
8	gene_id "CUFF.1"; transcript_id "CUFF.1.1"; exon_number "1"; FPKM "378.0239937260“

Transcript Structures in **GTF** Format

(tab-delimited fields per line shown transposed to a column format here)

0	7000000090838467	(genomic contig identifier)
1	Cufflinks	
2	transcript	
3	101	(left coordinate)
4	5716	(right coordinate)
5	1000	
6	+	(strand)
7	.	
8	gene_id "CUFF.1"; transcript_id "CUFF.1.1"; FPKM "378.0239937260“	(annotations)

0	7000000090838467
1	Cufflinks
2	exon
3	101
4	5716
5	1000
6	+
7	.
8	gene_id "CUFF.1"; transcript_id "CUFF.1.1"; exon_number "1"; FPKM "378.0239937260“

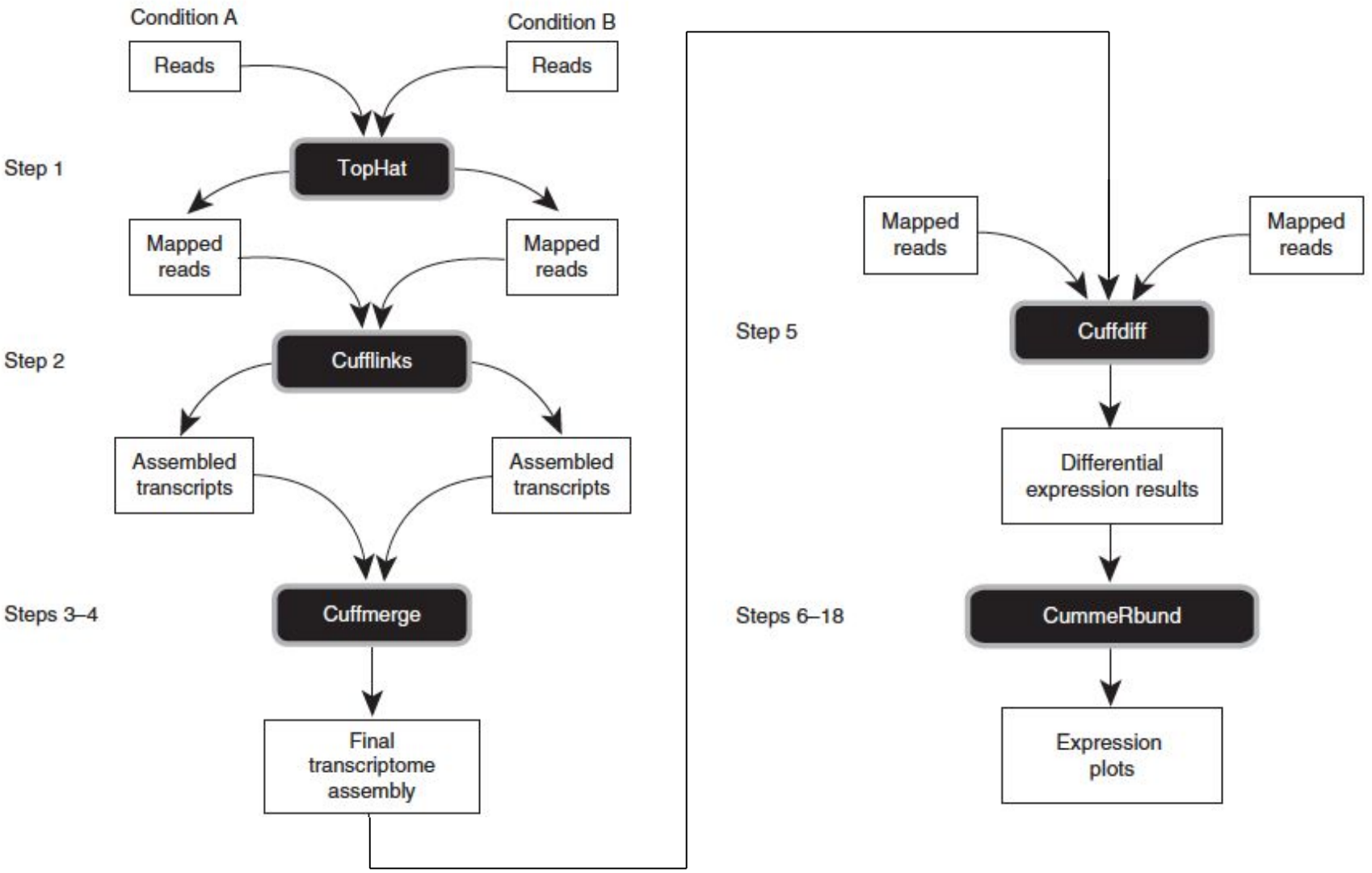
Demo: Tuxedo and IGV

- Run Tophat to align reads to the genome
- Reconstruct transcripts using cufflinks
- View genome-aligned reads and reconstructed transcripts using IGV

Full Tuxedo Framework Demo

- See: Tuxedo_workshop_activities.pdf

Tuxedo Framework for Transcriptome Analysis



Derived from: Nat Protoc. 2012 Mar 1;7(3):562-78. doi: 10.1038/nprot.2012.016.

Differential Expression Analysis Using RNA-Seq

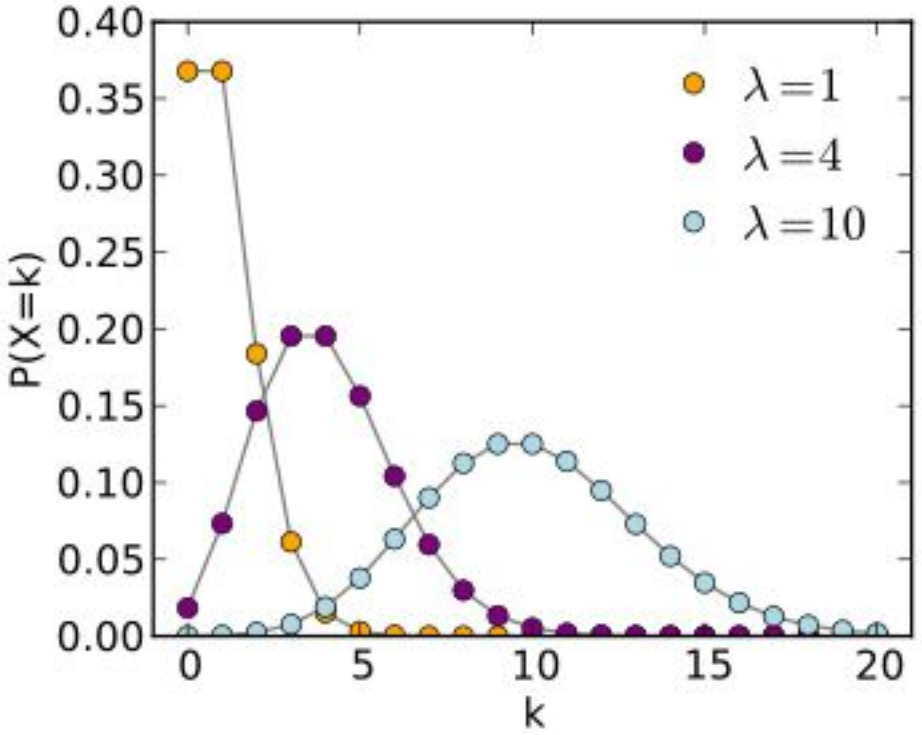
Diff. Expression Analysis Involves

- Counting reads
- Statistical significance testing

	Sample_A	Sample_B	Fold_Change	Significant?
Gene A	1	2	2-fold	No
Gene B	100	200	2-fold	Yes

Observed RNA-Seq Counts Result from Random Sampling of the Population of Reads

Technical variation in RNA-Seq counts per feature is well modeled by the Poisson distribution



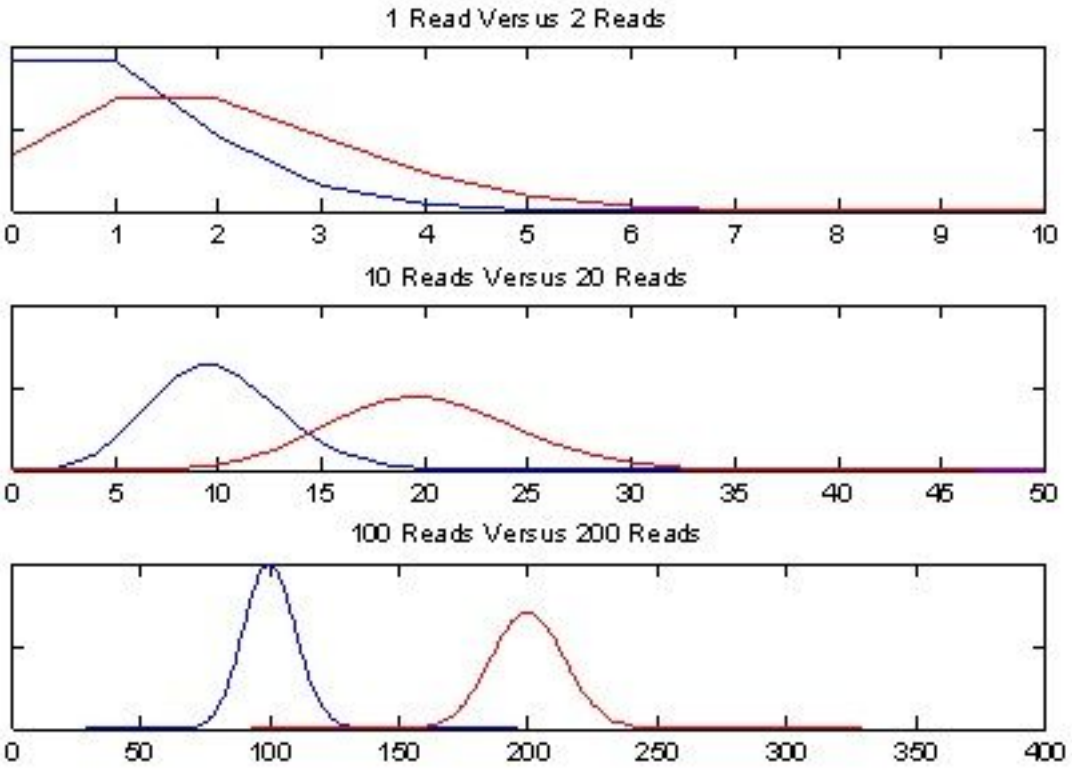
Mean # fragments

(observed read counts)

See: http://en.wikipedia.org/wiki/Poisson_distribution

Beware of concluding fold change from small numbers of counts

Poisson distributions for counts based on **2-fold** expression differences



No confidence in 2-fold difference. Likely observed by chance.

High confidence in 2-fold difference. Unlikely observed by chance.

From: http://gkno2.tumblr.com/post/24629975632/thinking-about-rna-seq-experimental-design-for

More Counts = More Statistical Power

Example: 5000 total reads per sample.

Observed 2-fold differences in read counts.

	SampleA	Sample B	Fisher's Exact Test (P-value)
geneA	1	2	1.00
geneB	10	20	0.098
geneC	100	200	< 0.001

Tools for DE analysis with RNA-Seq

ShrinkSeq
NoiSeq
baySeq
Vsf
Voom
SAMseq
TSPM
DESeq
EBSeq
NBPSeq
edgeR (metaSeqR)

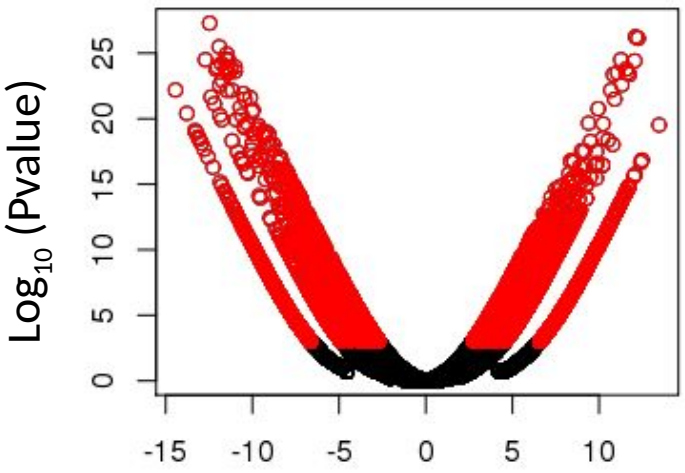
+ other (not-R)
including CuffDiff

See: http://www.biomedcentral.com/1471-2105/14/91

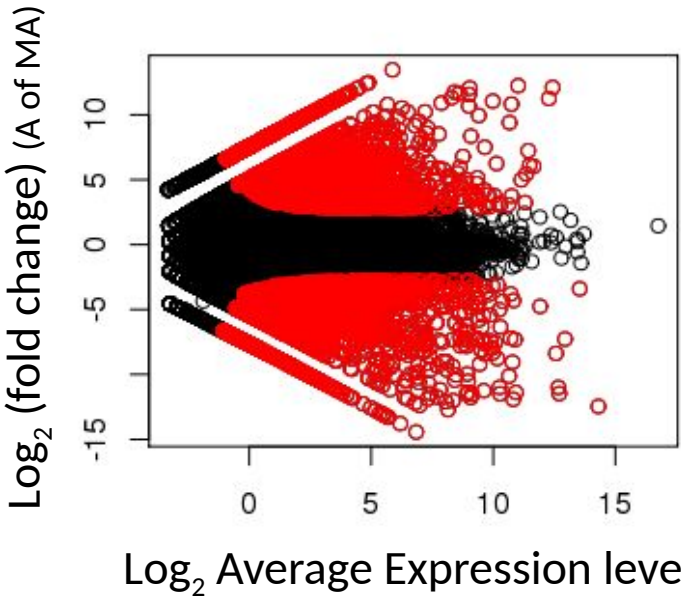
Visualization of DE results and Expression Profiling

Plotting Pairwise Differential Expression Data

Volcano plot
(fold change vs. significance)

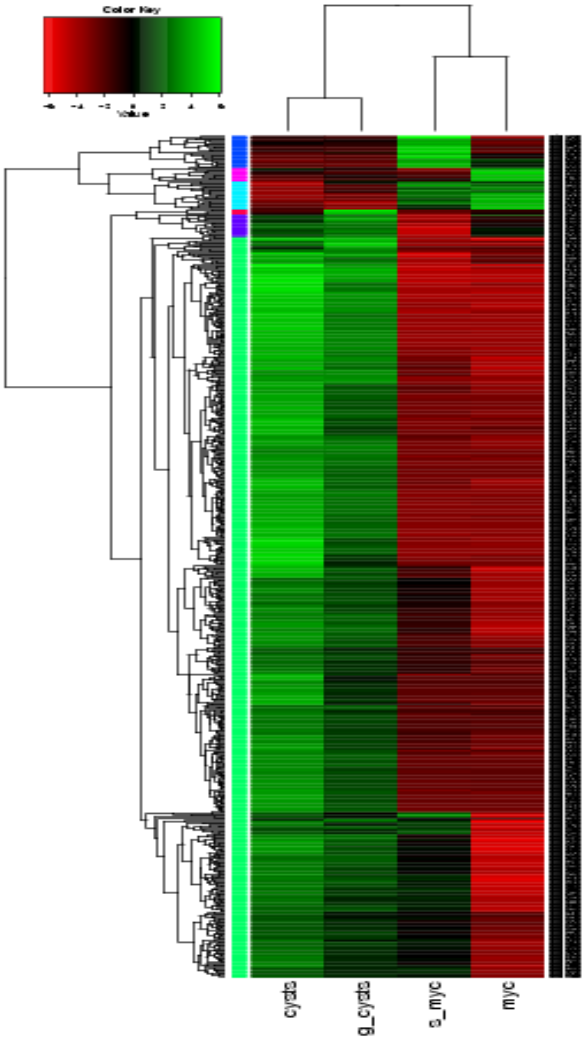


MA plot
(abundance vs. fold change)



Significantly differently expressed transcripts have FDR <= 0.001
(shown in red)

Comparing Multiple Samples



Heatmaps provide an effective tool for navigating differential expression across multiple samples.

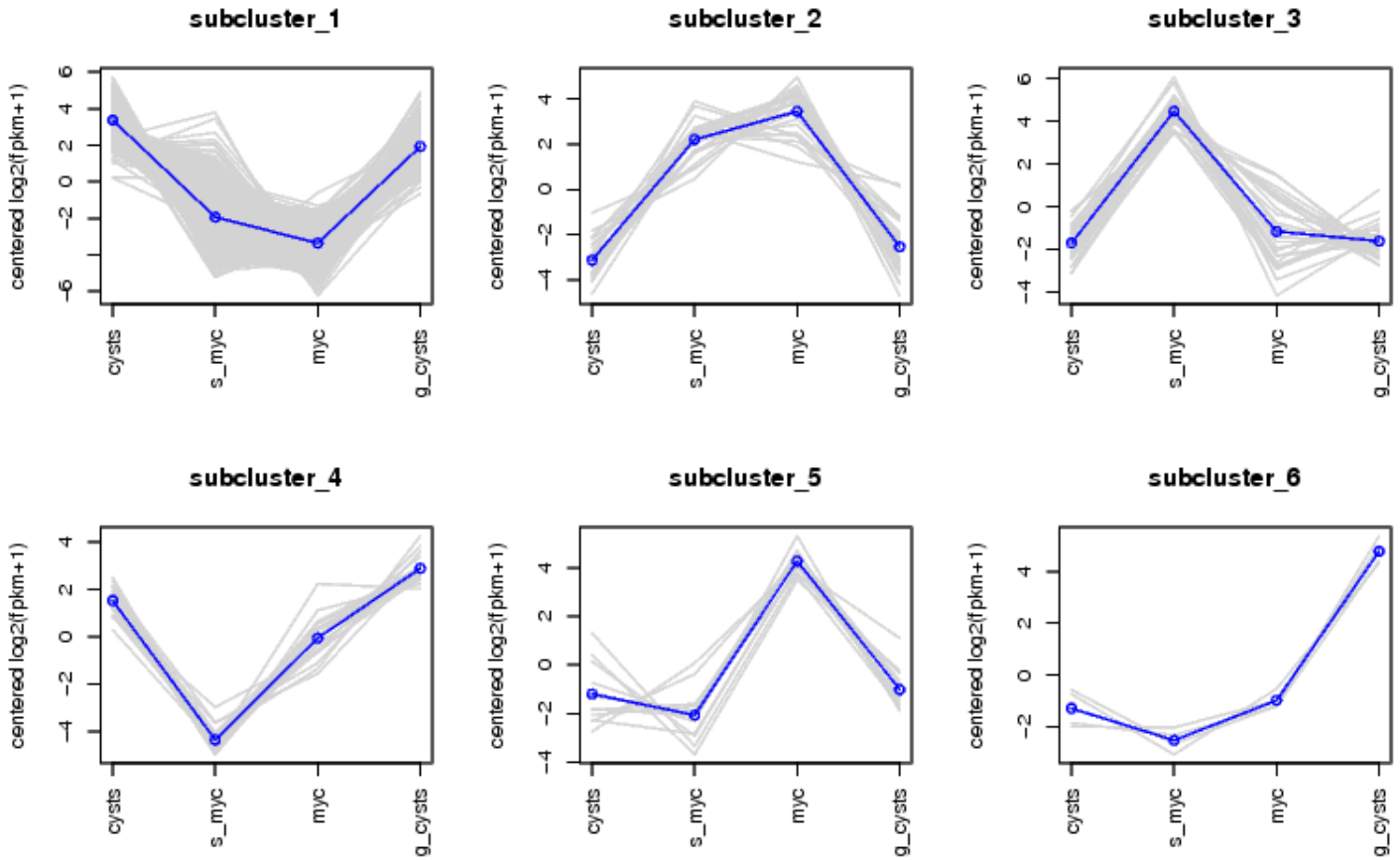
Clustering can be performed across both axes:

-cluster transcripts with similar expression patters.

-cluster samples according to similar expression values among transcripts.

Examining Patterns of Expression Across Samples

Can extract clusters of transcripts and examine them separately.



Summary of Key Points

- RNA-Seq is a versatile method for transcriptome analysis enabling quantification and novel transcript discovery.
- Genome-based and genome-free methods exist for transcript reconstruction
- Expression quantification is based on sampling and counting reads derived from transcripts
- Fold changes based on few read counts lack statistical significance.
- Multiple analysis frameworks are available – alternative and often complementary approaches to support biological investigations.

Software Links

- Tuxedo
  - Bowtie: http://bowtie-bio.sourceforge.net/index.shtml
  - Tophat: http://tophat.cbcb.umd.edu/
  - Cufflinks: http://cufflinks.cbcb.umd.edu/
- Trinity

  http://trinityrnaseq.sourceforge.net/
- IGV for Visualization

  http://www.broadinstitute.org/igv/
- GMAP

  http://research-pub.gene.com/gmap/
- Samtools

  http://samtools.sourceforge.net/

Papers of Interest

- Next generation transcriptome assembly
  - http://www.nature.com/nrg/journal/v12/n10/full/nrg3068.html
- Tuxedo protocol
  - http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3334321/
- Trinity
  - http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3571712/
  - http://www.nature.com/nprot/journal/v8/n8/full/nprot.2013.084.html