

# Syllabus and grading

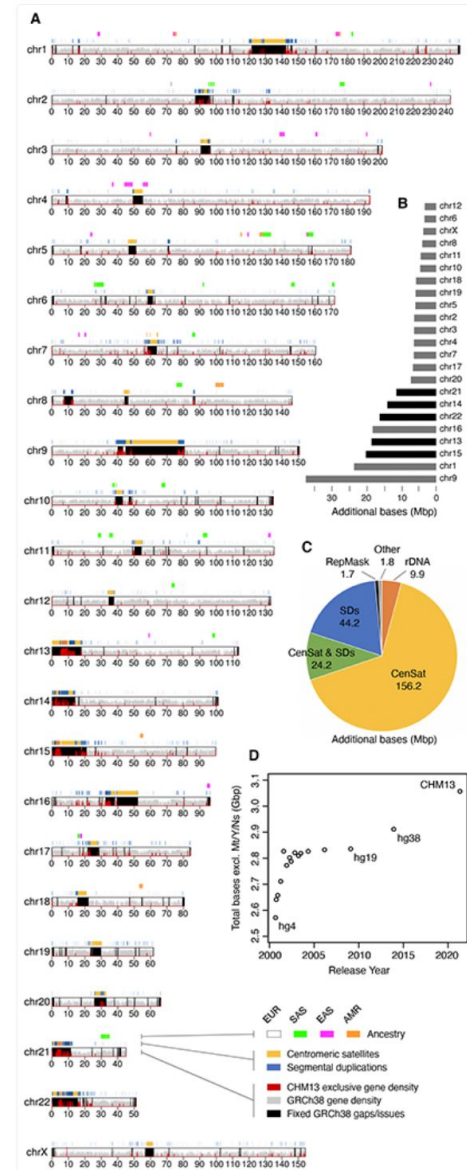
#	Date	Short title	Lecturer	Subject
1	7/10/25	introduction	MR	Overview of Bioinformatics, sequence alignment
2	14/10/25	Linux/shell/ssh	AD	Introduction to Linux and the command line, bash scripting and ssh
3	21/10/25	QC+RNASeq	MR	Next generation sequencing: introduction, quality control and gene expression analysis for RNAseq
4	4/11/25	R (1)	AD	Introduction to the R programming language and Rstudio usage
5	11/11/25	R (2)	AD	Advances R subjects, introduction to Bioconductor
6	18/11/25	bedtools/vcftools/samtools fl	AD	Command line tool usage: bedtools, vcftools, samtools etc.
7	25/11/25	Denovo	MR	NGS for denovo genome and transcriptome assembly
8	2/12/25	Exome/SNP calling	AD	Pipelines for SNP calling, especially for exome sequencing using the GATK pipeline
9	9/12/25	ChipSeq/chirp	MR	NGS analysis for molecular interactions (ChipSeq, (Par-)Clip, structural sequencing, chromosome conformation capture (3C))
10	16/12/25	metabolomics	MR	Pipelines for SNP calling, especially for exome sequencing using the GATK pipeline
11	13/1/26	presentations	MR+AD	Paper presentations by students
12	20/1/26	presentations	MR+AD	Genome-scale models of metabolism and macromolecular expression, Biological applications of Transformers
13	27/1/26	final projects support	MR+AD	Support for the final project

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

**Fig. 1. Summary of the complete T2T-CHM13 human genome assembly.**

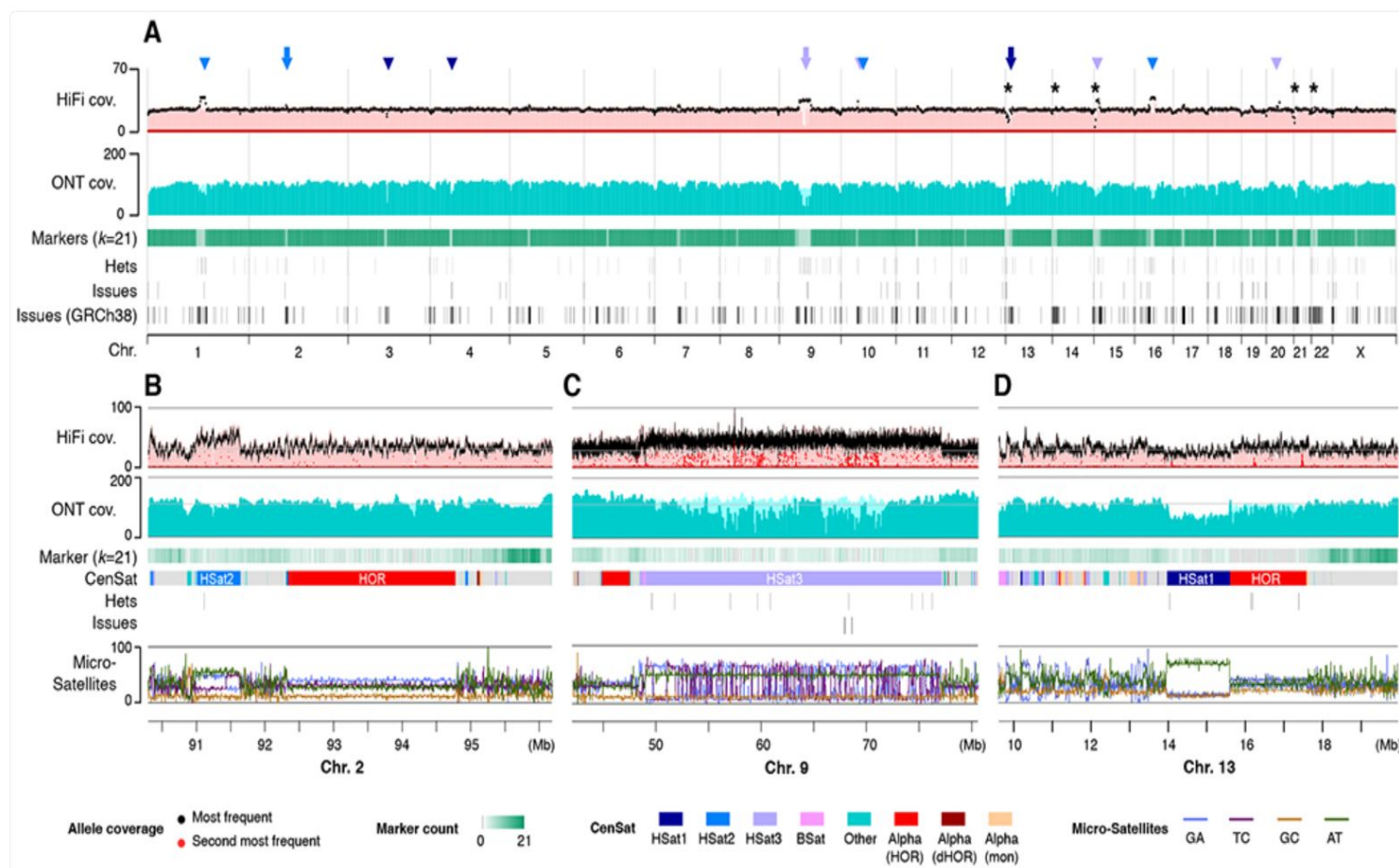
*Science*. 2022 April ; 376(6588): 44–53. doi:10.1126/science.abj6987.

# The complete sequence of a human genome



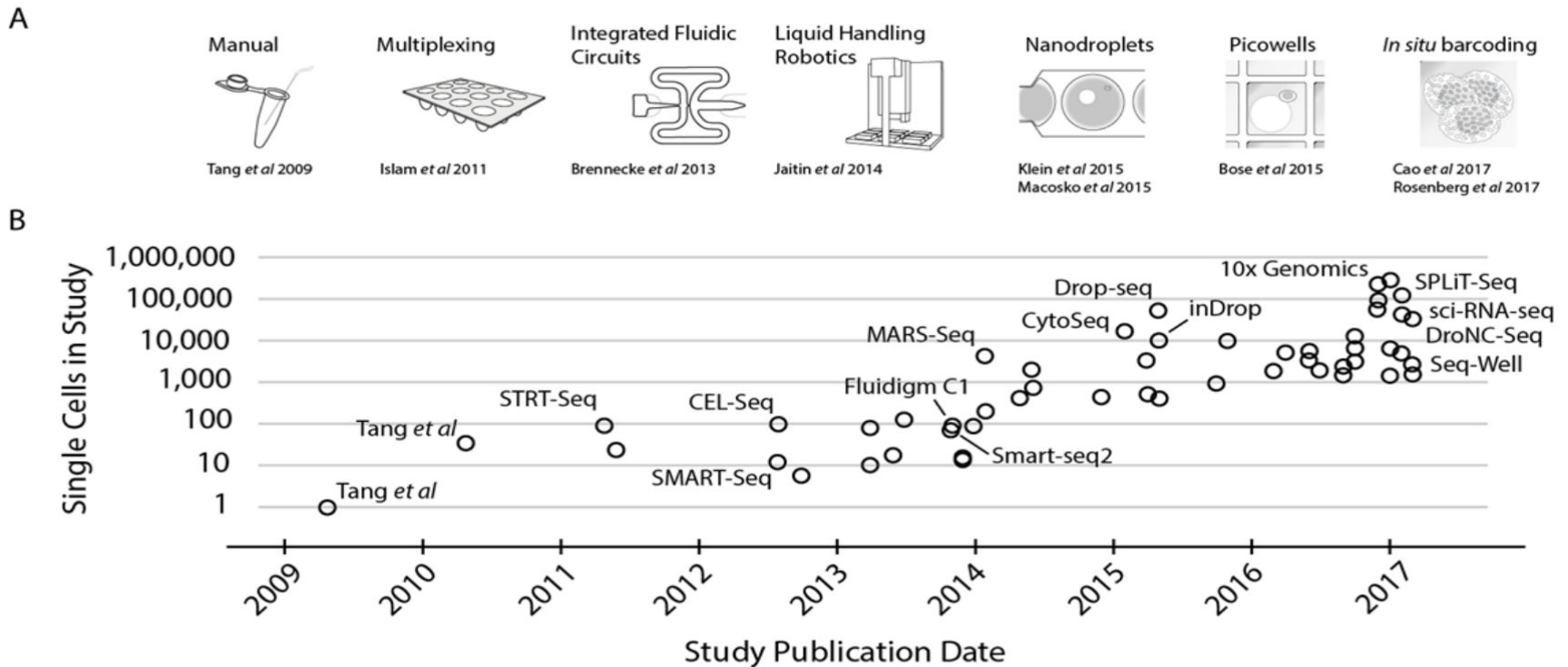
(A) Ideogram of T2T-CHM13v1.1 assembly features. Bottom to top: gaps/issues in GRCh38 fixed by CHM13 overlaid with the density of gene models in red; segmental duplications (SDs) and centromeric satellites (CenSat); and CHM13 ancestry predictions (EUR, European; AS, Asian; EAS, East Asian; AMR, Ad Mixed American). (B) Additional (non-syntenic) bases in the CHM13 assembly relative to GRCh38 per chromosome. The acrocentrics highlighted in black, and (C) by sequence type (note that the CenSat and SD annotations overlap). (D) Total non-gap bases in human genome releases dating back to September 2000 (hg4) and ending with T2T-CHM13 in 2021.

**Fig. 3. Sequencing coverage and assembly validation.**



(A) Uniform whole-genome coverage of mapped HiFi and ONT reads is shown with primary alignments in light shades and marker-assisted alignments overlaid in dark shades. Large HSat arrays (30) are noted by triangles, with inset regions are marked by arrowheads and the location of the rDNA arrays marked with asterisks. Regions with low unique marker frequency (light green) correspond to drops in unique marker density, but are recovered by the lower-confidence primary alignments. Annotated assembly issues are compared for T2T-CHM13 and GRCh38. (B–D) Enlargements corresponding to regions of the genome featured in Fig. 2. Uniform coverage changes within certain satellites are reproducible and likely caused by sequencing bias. Identified heterozygous variants and assembly issues are marked below and typically correspond with low coverage of the primary allele (black) and elevated coverage of the secondary allele (red). % microsatellite repeats for every 128 bp window is shown at the bottom.

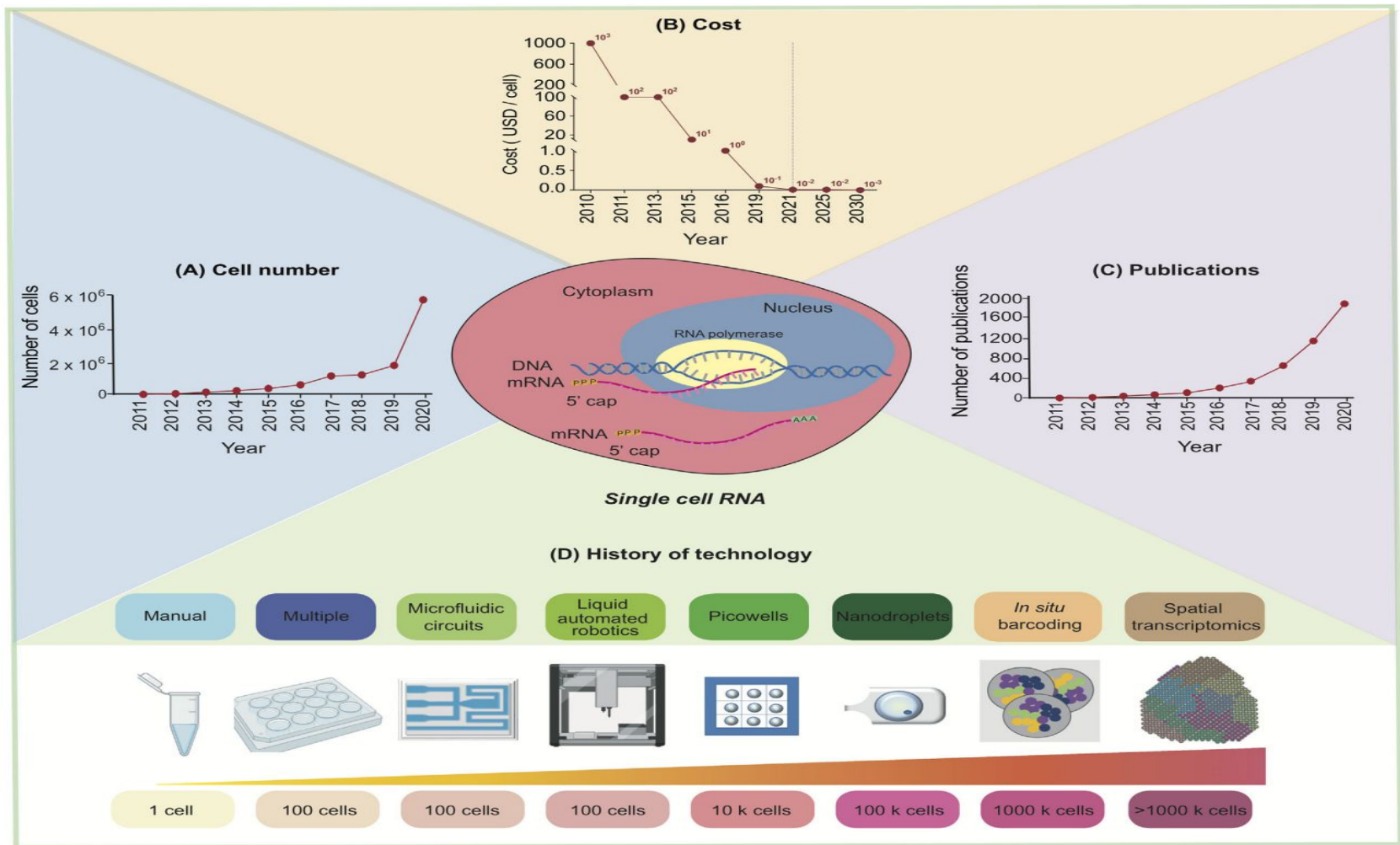
# Single Cell Sequencing



**Figure 1: Scaling of scRNA-seq experiments** (A) Key technologies allowing jumps in experimental scale. A jump to ~100 cells was enabled by sample multiplexing, a jump to ~1,000 cells by large scale studies using integrated fluidic circuits (IFCs), followed by a jump to several thousands using liquid handling robotics. Further order of magnitude jumps were enabled by random capture technologies through nanodroplets and picowell technologies. Recent studies have employed *in situ* barcoding to reach the next order of magnitude. (B) Cell numbers reported in representative publications by publication date. Key technologies and protocols are marked, and a full table with corresponding numbers is available in **Supplementary Table 1**.

Svensson et al. Nature Protocols 2018

# Single Cell Sequencing




**FIGURE 1** Development of single-cell RNA sequencing technology. With the technological advances in single-cell RNA sequencing (scRNA)-seq, (A) the number of analyzed cells increased, (B) the cost (in US dollar) was exponentially reduced, (C) the number of published papers increased and (D) the history of technology evolution in the last decade using more sophisticated, accurate, high throughput analysis was achieved. Part (D) is created with icons from BioRender with license for publication

# Bead production 1

## 1. Synthesis of the Cell Barcode (Fixed Sequence)

The Cell Barcode identifies the specific bead and, thus, the cell. This sequence is **identical** across all millions of oligos attached to one single bead.

- **Method:** This is typically created using **Combinatorial "Split-and-Pool" Synthesis** (often used for creating large, diverse libraries).
  - The entire batch of beads is split into  $N$  different reaction vessels (e.g., four vessels for the four bases: A, T, C, G).
  - A single nucleotide (A, T, C, or G) is added to the oligos on the beads in each vessel.
  - The beads are then pooled back together and mixed. 
  - This process is repeated (e.g., 16 times for a 16 bp barcode).
- **Result:** By the end of the process, each individual bead in the final batch has received a unique, fixed sequence (the **Cell Barcode**), but all the oligos *on that bead* are the same.

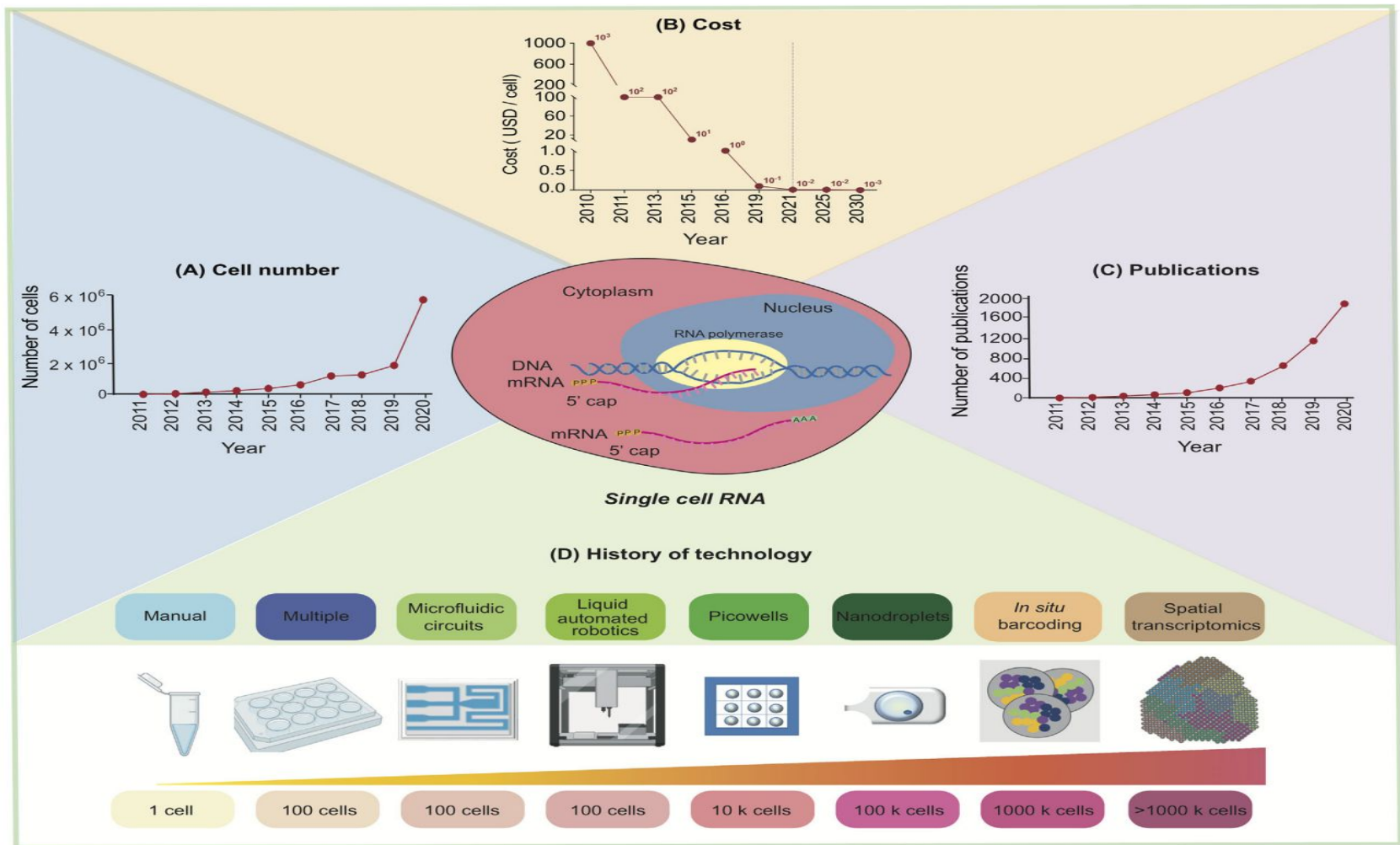
## Bead production 2

### 2. Synthesis of the UMI (Random Sequence)

After the Cell Barcode is synthesized, the UMI segment is added to the oligo. This is where the **randomization** happens.

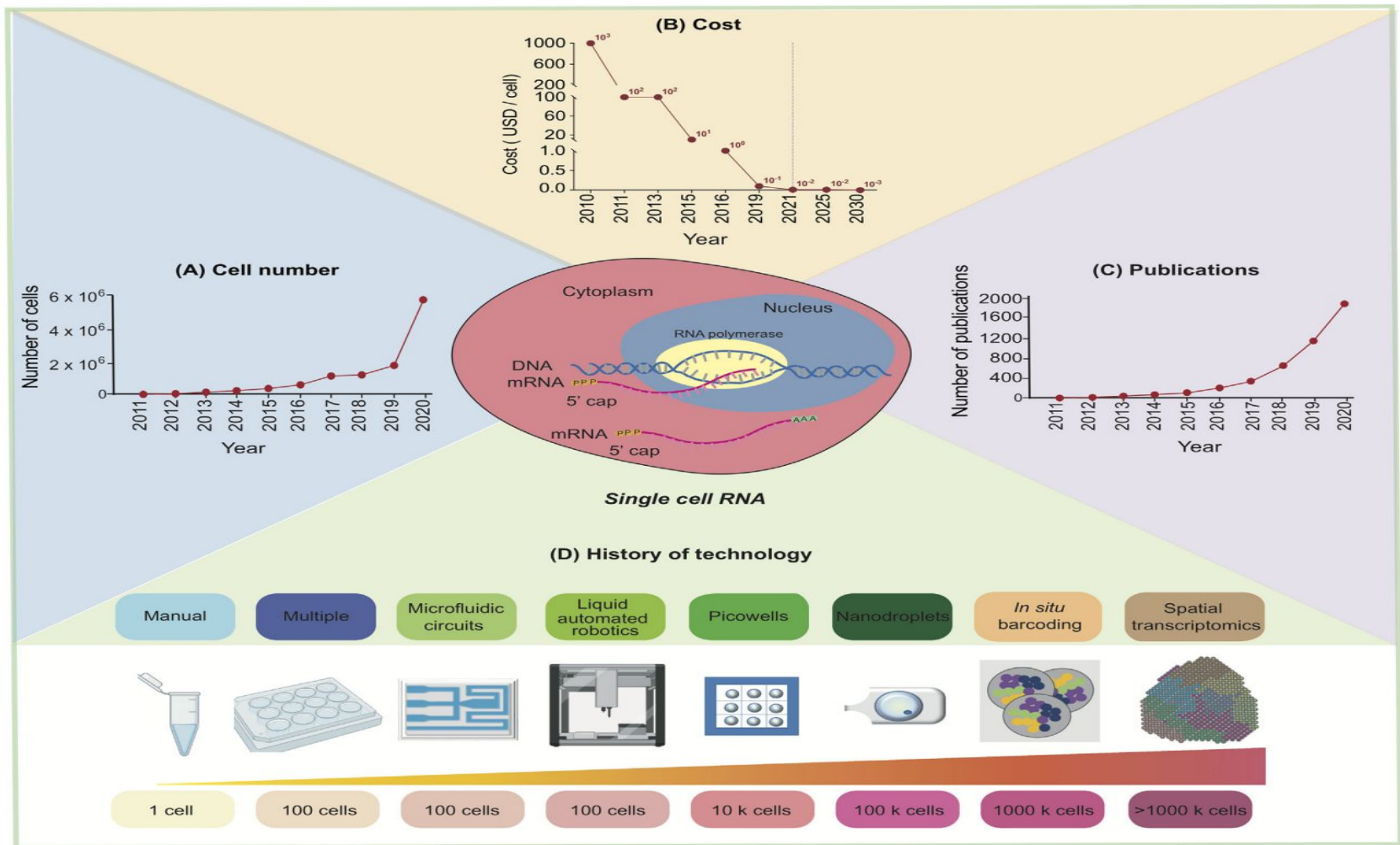
- **Method: Degenerate Synthesis (Random Addition)**
  - The beads are **not split** (or are treated as one large pool).
  - For each position of the UMI (e.g., 10 positions for a 10 bp UMI), the reaction mix is prepared to contain **all four nucleoside phosphoramidites (A, T, C, G) simultaneously**.
  - When the chemical synthesis cycle occurs, the four bases are incorporated randomly onto the growing DNA strand.
- **Result:** Since the incorporation of the bases is a **random chance** event at each position, every individual oligo molecule on that bead (millions of them) receives a different, randomly generated UMI sequence (e.g.,  $4^{10}$  or over a million potential UMI sequences).

# Single Cell Sequencing



**FIGURE 1** Development of single-cell RNA sequencing technology. With the technological advances in single-cell RNA sequencing (scRNA)-seq, (A) the number of analyzed cells increased, (B) the cost (in US dollar) was exponentially reduced, (C) the number of published papers increased and (D) the history of technology evolution in the last decade using more sophisticated, accurate, high throughput analysis was achieved. Part (D) is created with icons from BioRender with license for publication

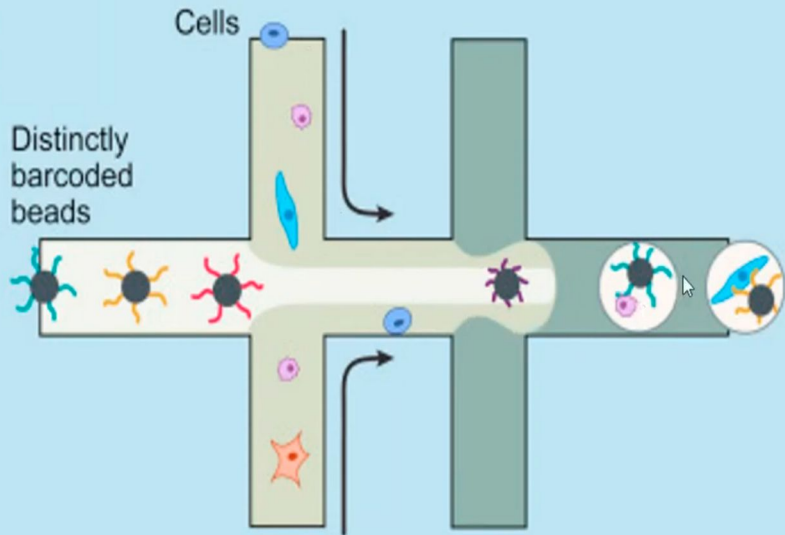
# Single Cell Sequencing



**FIGURE 1** Development of single-cell RNA sequencing technology. With the technological advances in single-cell RNA sequencing (scRNA)-seq, (A) the number of analyzed cells increased, (B) the cost (in US dollar) was exponentially reduced, (C) the number of published papers increased and (D) the history of technology evolution in the last decade using more sophisticated, accurate, high throughput analysis was achieved. Part (D) is created with icons from BioRender with license for publication

# Bead: Cell barcode and unique molecular identifiers (UMIs)

## Drop-seq single cell analysis



1000s of DNA-barcoded single-cell transcriptomes

- Cell barcode: which cell the read comes from
- UMI: which mRNA molecule the read comes from (helps to detect PCR duplicates)

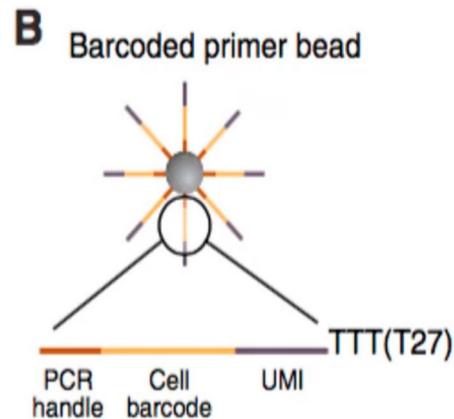


Figure by Macosko et al, *Cell*, 161:1202-1214, 2015

# From reads to digital gene expression matrix (DGE)

## Overview of DGE extraction

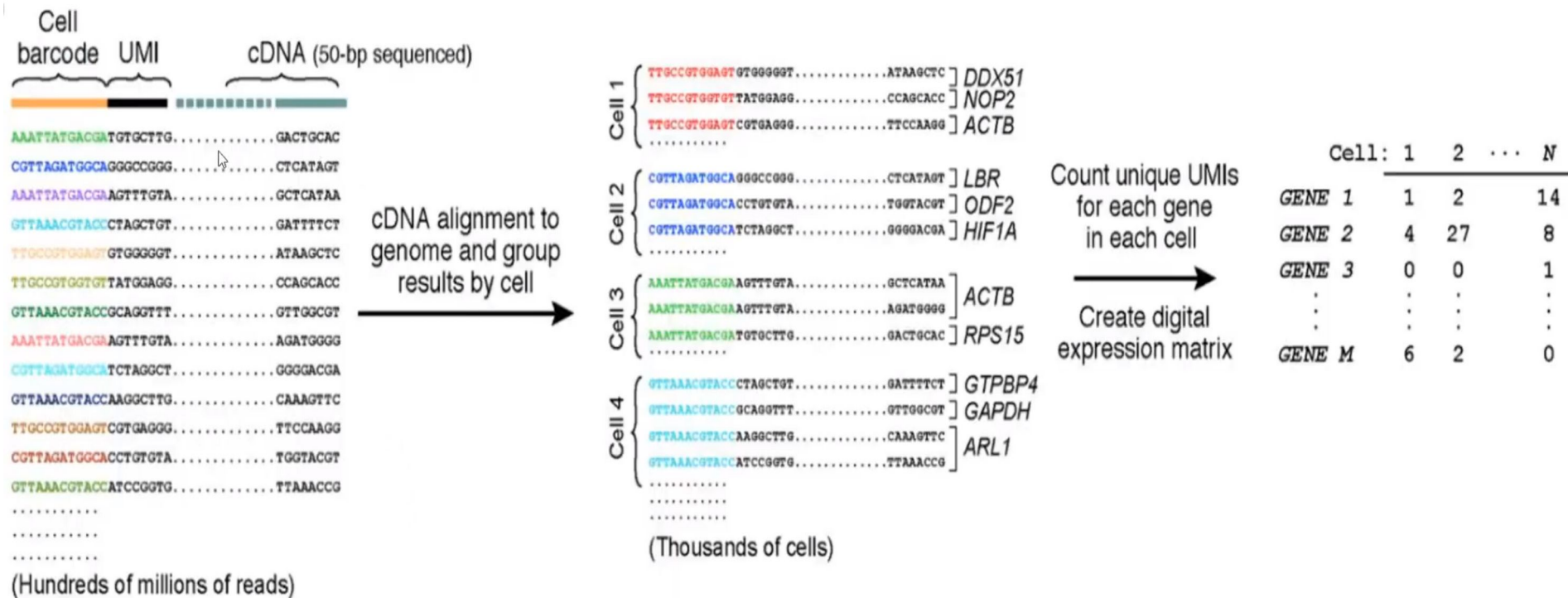
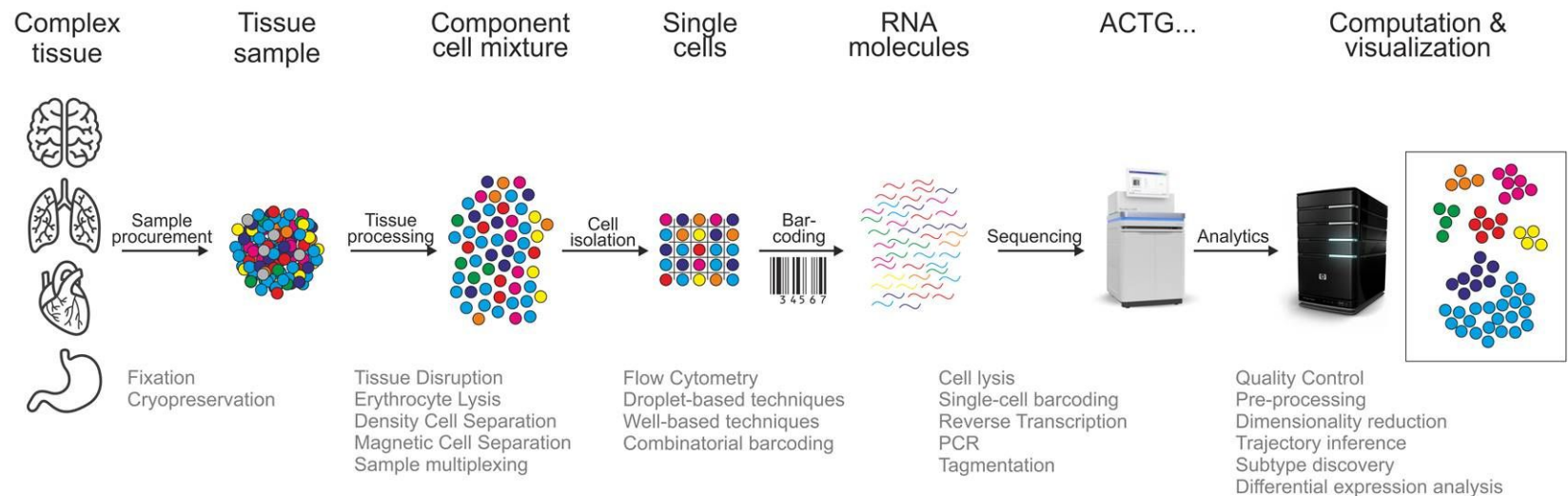


Figure by Macosko et al, Cell, 161:1202-1214, 2015

# Single Cell RNA Sequencing and its main applications

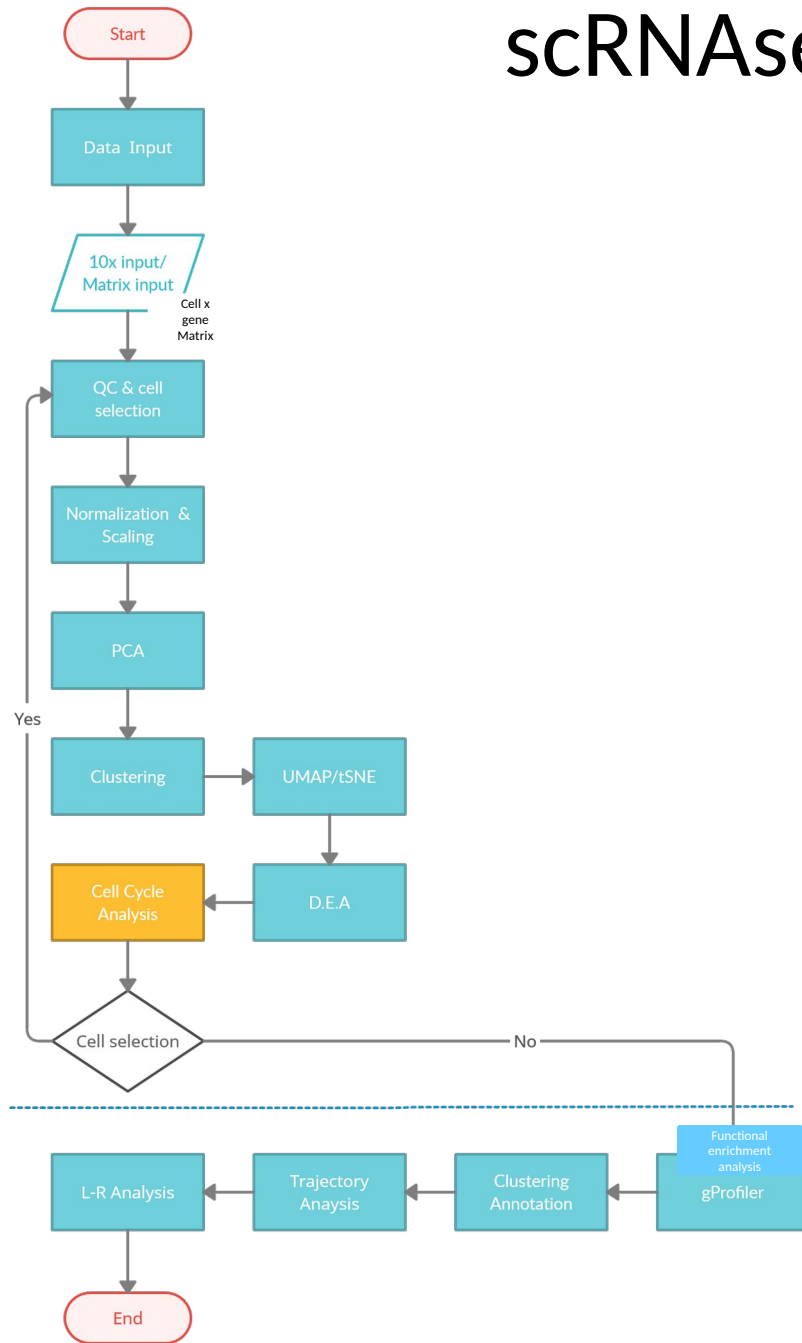
- ❑ Identification of new cell populations and subpopulations in complex tissues
- ❑ Studying gene dynamics in developmental studies
- ❑ Immune cell profiling
- ❑ Cancer research
- ❑ Personalized medicine
- ❑ Cell atlases



Logistics	Cell sources	Miniaturization & optimization	State-of-the-art equipment	Updated algorithms & hardware
Blood Brain BAL Lung Intestine	Fresh cells & tissues Fixed cells & tissues Extracted nuclei	Seq-Well Smart-Seq2 BD Rhapsody Patch-Seq scATAC-Seq	NextSeq NovaSeq	FASTGenomics Auto-encoder Bayesian models Batch correction Sparse2Big (Helmholtz) The Machine (HPE)

(slides by ITBI student Dimitra Panou)

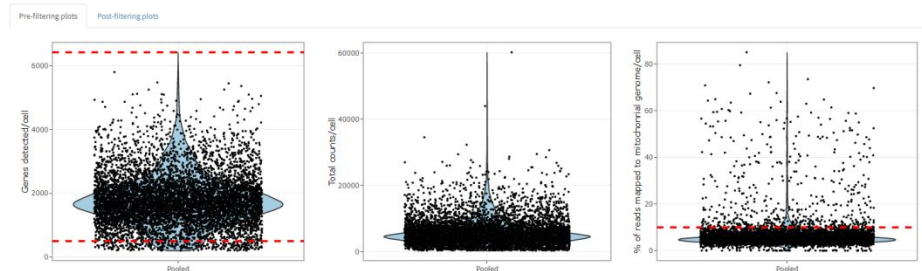
# scRNAseq pipeline



## 1. Quality Control & Cell selection

- ☐ Detect + remove low quality cells from downstream analysis
  - Genes detected/cell
  - Total reads/cell
  - % of reads in mitochondrial genome/cell

Quality control plots



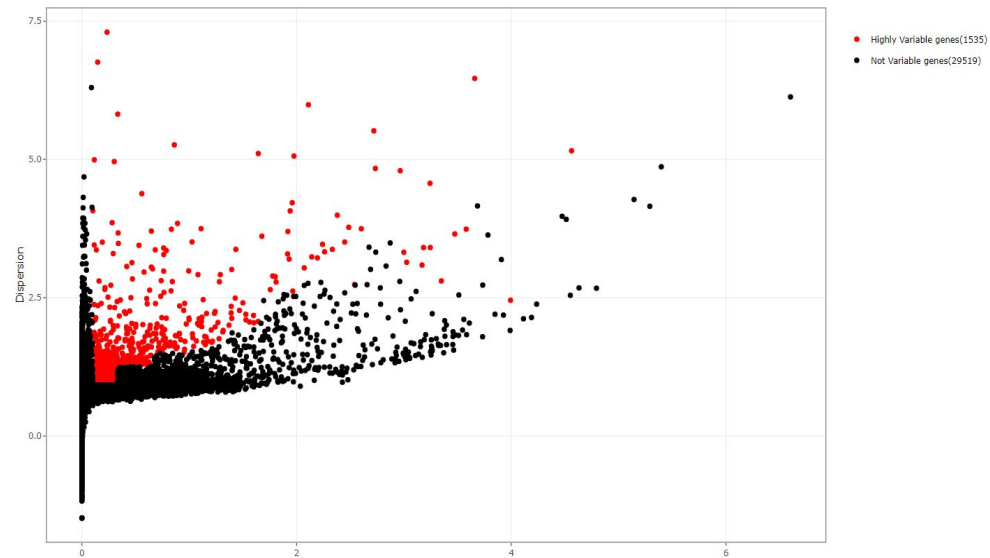
Red lines show the filtering points

# scRNAseq pipeline

## 2. Normalization & Scaling

- **Global normalization**
  - Correcting for sequencing depth differences between cells
  - Log transformation
- **Detection of Highly variable genes**
  - Mean.var.plot method (**mvp**) highly variable genes
  - Scaling transformation
- **Calculation of scaled values for all genes**
  - Scales + centers the genes in dataset

Highly variable genes



## 3. Dimensional reduction

### PCA analysis

Detection of most informative principal components

- ☐ Moving to PCA space can help reducing runtime of cell clustering
- ☐ May fail to capture local patterns in scRNA data

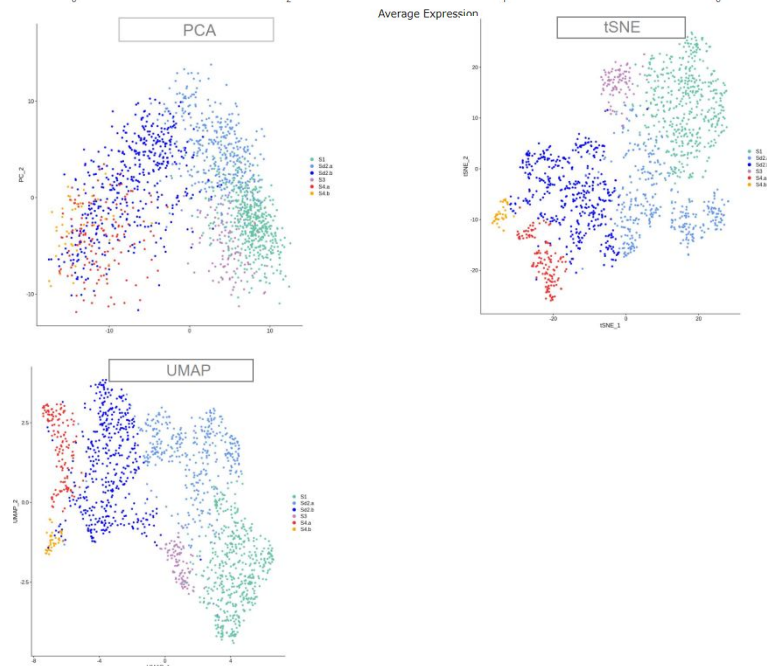
### Non linear methods

t-Distributed Stochastic Neighbor Embedding (t-SNE)

- ☐ Can capture subtle local patterns of expression in the data
- ☐ Places cells with similar local neighborhoods in high dimensional space together in low dimensional space
- ☐ It may fail to give a precise representation of clusters' size and distances

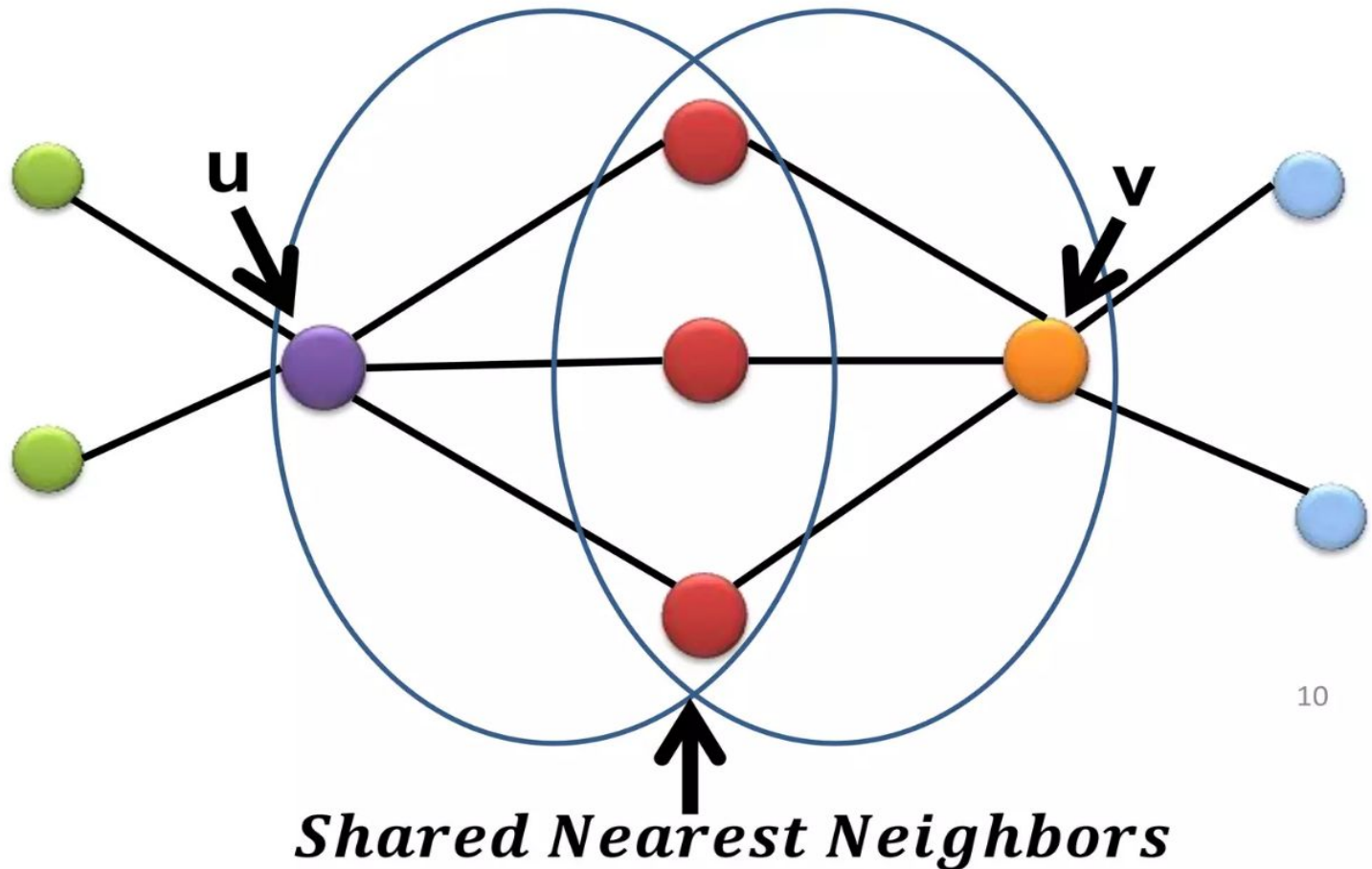
Uniform Manifold Approximation and Projection (UMAP)

- ☐ Preserves better the global structure of the data
- ☐ Faster runtime than tSNE
- ☐ It may fail to illuminate the lineage structure of the data

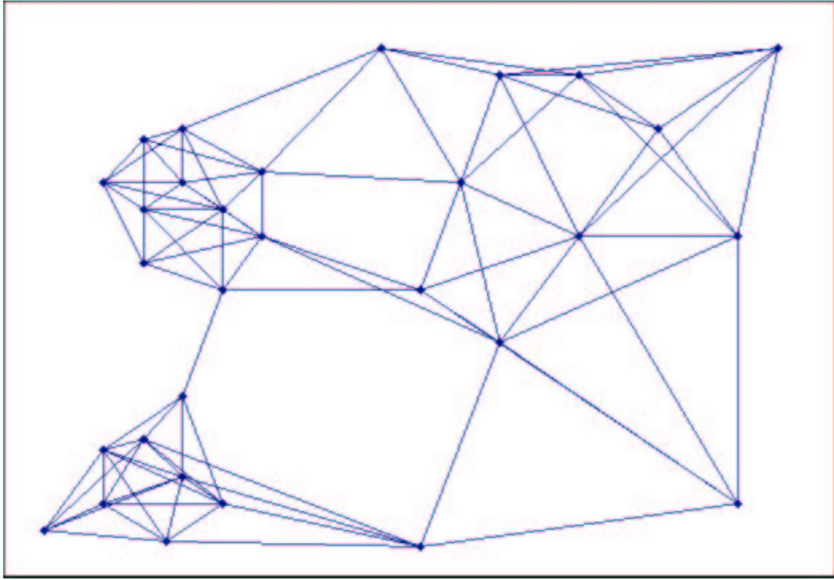


# What is Shared Nearest Neighbor?

**Shared Nearest Neighbor is a proximity measure and denotes the number of neighbor nodes common between any given pair of nodes**

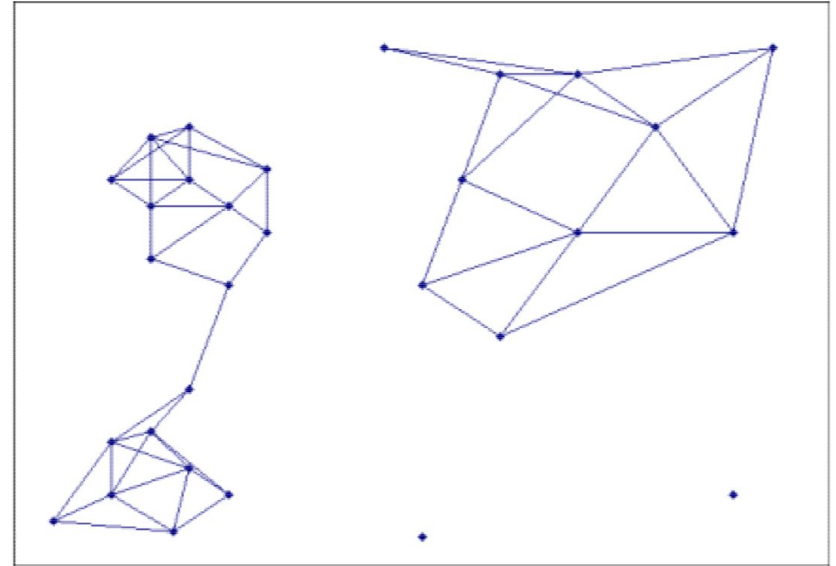


# Shared Nearest Neighbors graph



(a) Near Neighbor Graph.

$k=5$



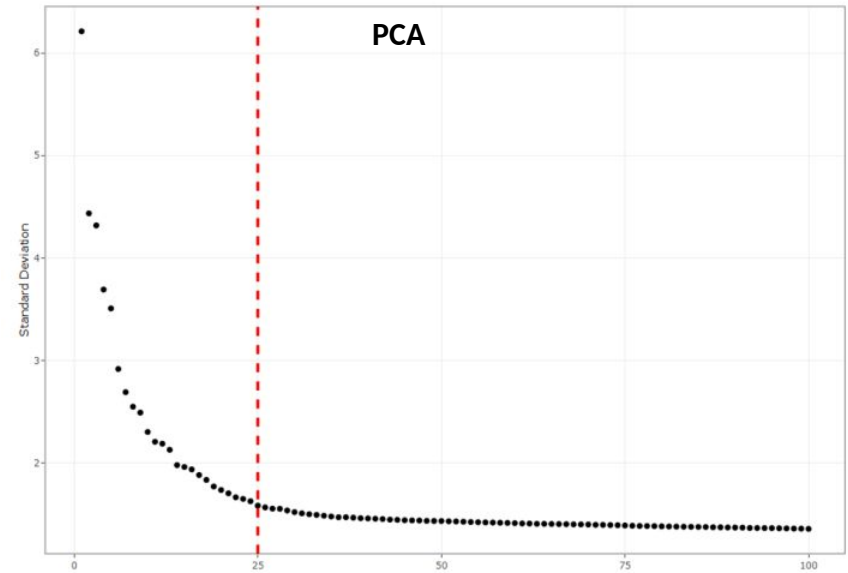
(b) Unweighted Shared Nearest Neighbor.

link if  $p_1$  and  $p_2$  have each other in  
their nearest neighbor lists

# scRNAseq pipeline

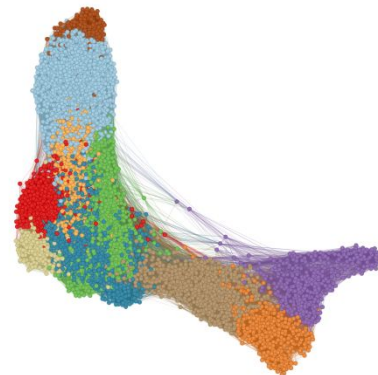
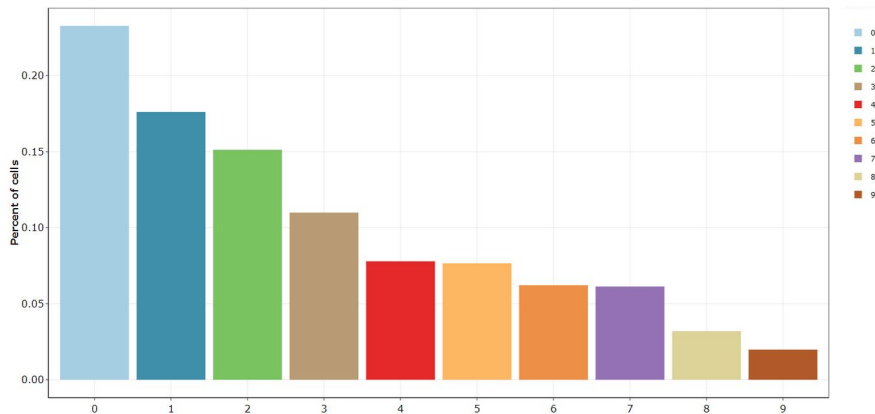
## 4. Clustering analysis

- ❑ Creation of a Shared nearest neighbor (SNN) graph
- ❑ Clusters represent
  - cell population
  - cell sub-population
  - cell state



Shared Nearest Neighbors graph

## Clustering



# scRNAseq pipeline

## 5. Differential Expression Analysis

## Design of the analysis

- Cells belonging to one cluster VS Cells belonging to another
- Cells belonging to one cluster VS Cells belonging to the rest of the clusters

### Selection of D.E.A test

- Wilcoxon test, Student's t-test, Poisson, MAST \*

## Marker gene analysis

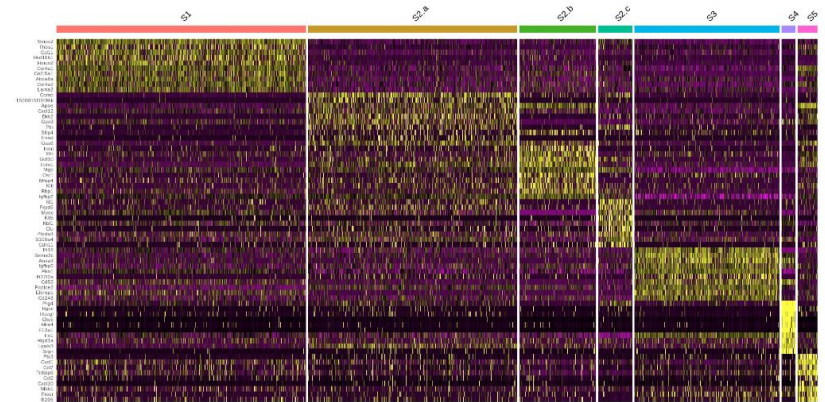
- Identify marker genes per cluster
- Those genes can distinguish one cluster from the rest
- High average expression in cells of the cluster, low in the other cells

- ☐ Wilcoxon test
- ☐  $\log_{2}FC \geq 0.25$
- ☐  $P\text{-value} < 0.01$
- ☐ Percentage of expression  $\geq 25\%$

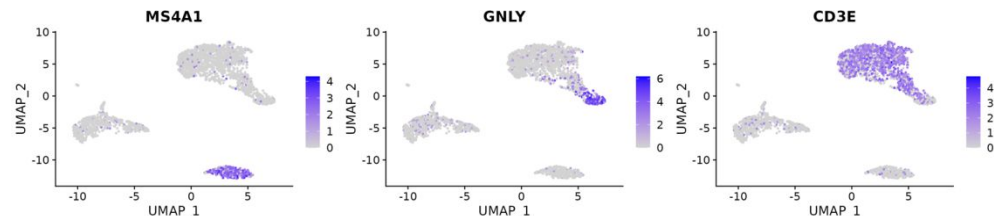
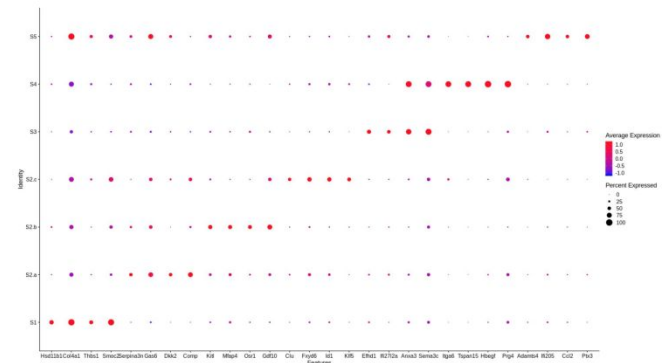
### Inspection of top marker genes

- ❑ Feature plots in UMAP space
- ❑ Color denotes normalized expression

## Differential expression analysis



## Marker genes for each cluster

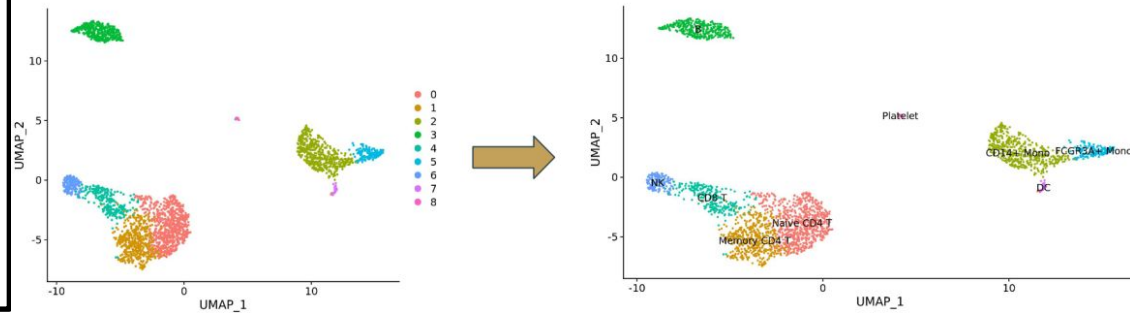


\* MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data

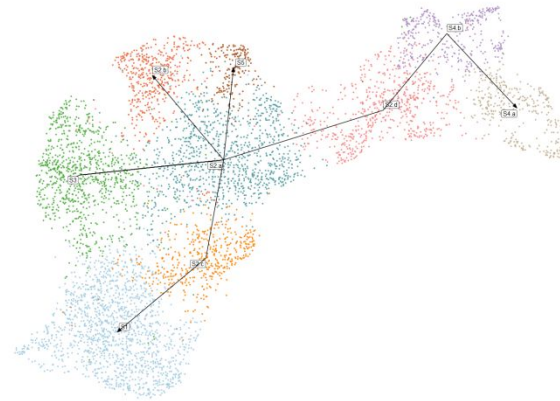
# scRNAseq pipeline

## 6. Cluster annotation

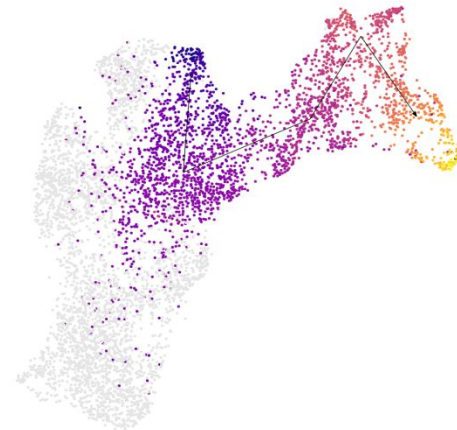
- Compare cluster marker genes to canonical markers for different cell types from the literature
- Using computational methods, match cluster labels from a different dataset (e.g. a cell atlas of the studied organism) to your own clusters



Minimum spanning tree



Pseudotime ordering



## 7. Trajectory-Pseudotime analysis

- ☐ Infer the lineage structure of the dataset
- ☐ Order the cells along the predicted topology
- ☐ PCs as input
- ☐ Output in UMAP plot

- Useful links

- [Seurat - Guided Clustering Tutorial](https://satijalab.org/seurat/articles/pbm3k_tutorial.html)  
[https://satijalab.org/seurat/articles/pbm3k\\_tutorial.html](https://satijalab.org/seurat/articles/pbm3k_tutorial.html)

Online scRNAseq analysis

- <https://singlecell.usegalaxy.eu/>
- <http://scala.fleming.gr/app/scala>
- <https://crescent.cloud/>

# Isoform quantitation tools in the literature

- 26 tools found in literature that support transcript DE
  - 10 still active
  - 6 user friendly enough for being used (!)
    - open-source with source code released under a license

	Name	Since	Citations
1	Tuxedo Suite	2012	5390
2	RSEM	2011	4068
3	New Tuxedo Suite	2016	215
4	sleuth	2017	169
5	BitSeq	2012	164
6	EBSeq	2015	4

(slide by A. Dimopoulos)

# De-novo genome sequence assembly, Genome-Based and Genome-Free Transcript Reconstruction and Analysis Using RNA-Seq Data

based on material from Mathias Haimel, EBI

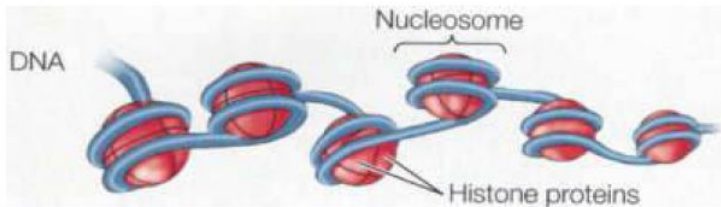
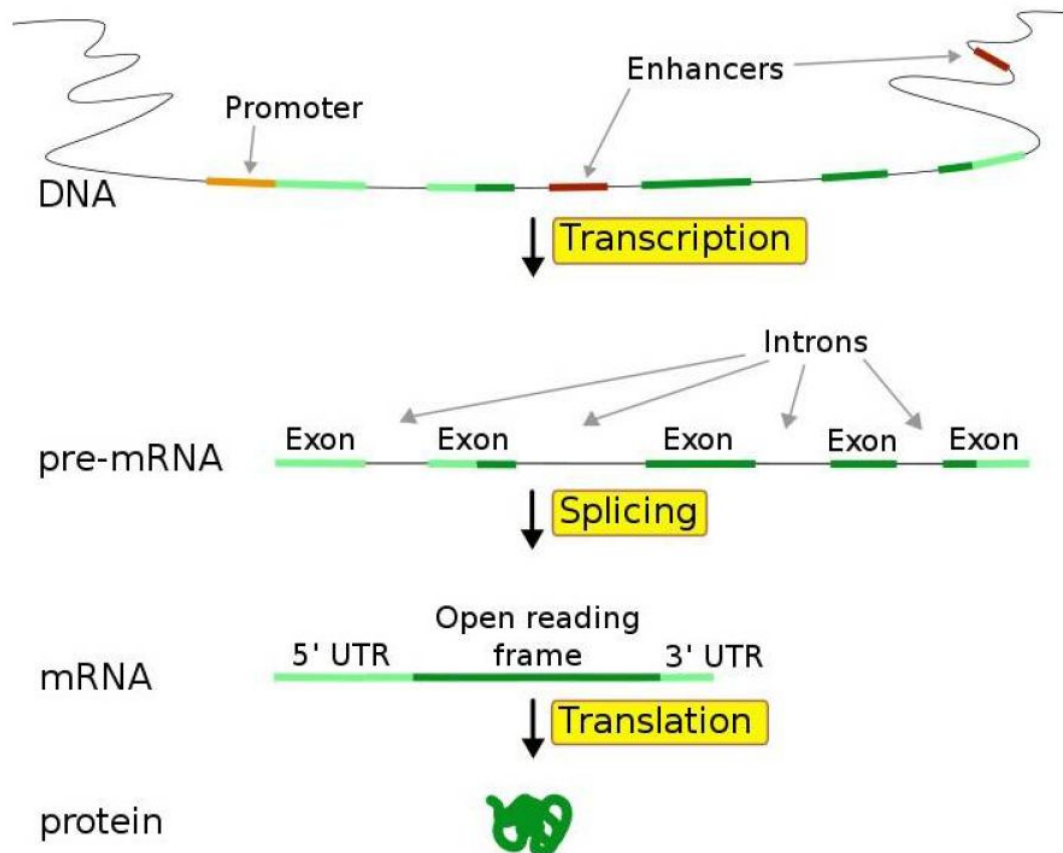
[https://www.ebi.ac.uk/training/online/sites/ebi.ac.uk.training.online/files/user/18/private/velvet\\_1.pdf](https://www.ebi.ac.uk/training/online/sites/ebi.ac.uk.training.online/files/user/18/private/velvet_1.pdf)

and Brian Haas

Broad Institute, modified by M. Reczko



# Next Generation Sequencing

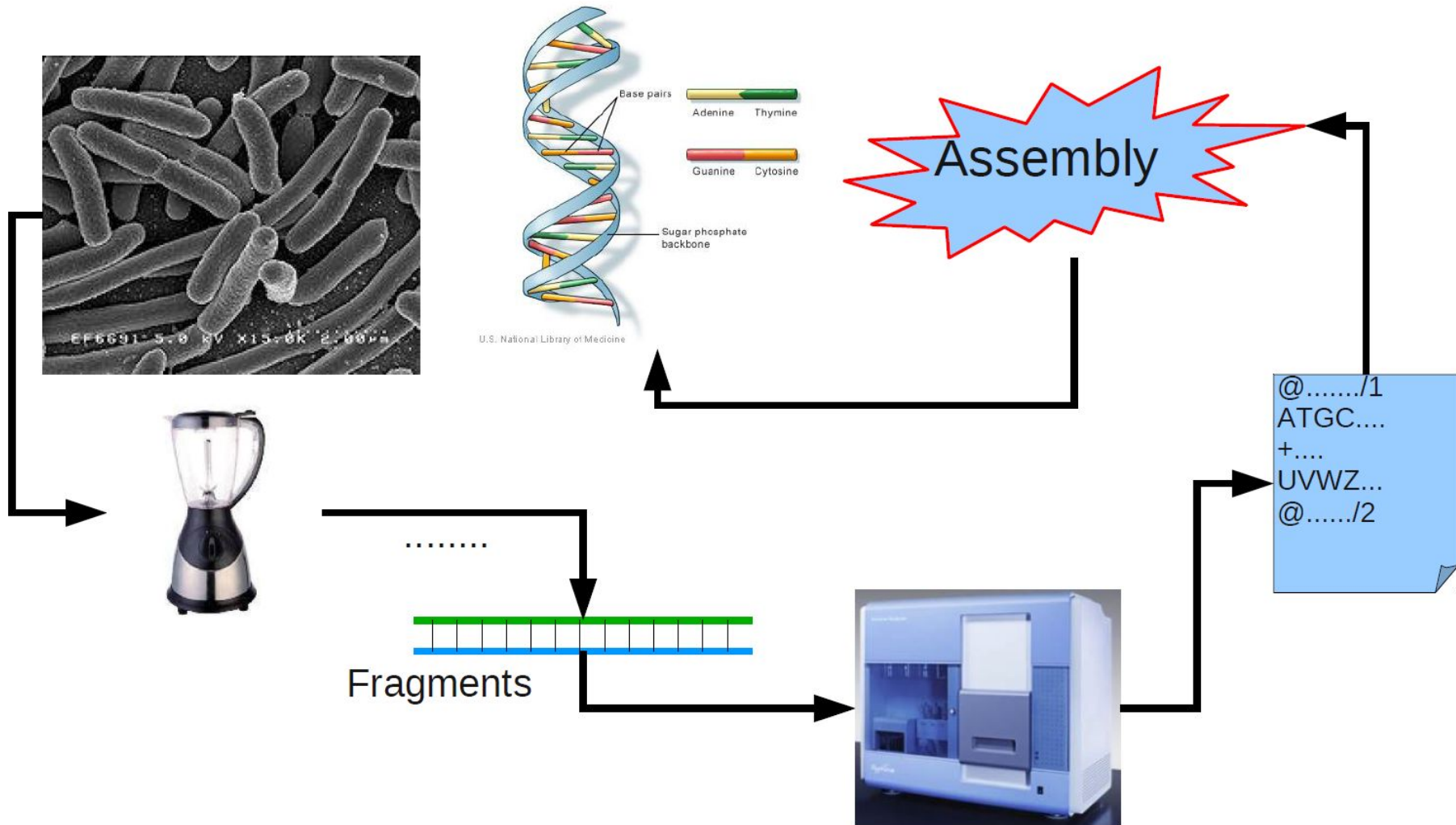


**Whole Genome  
sequencing**

**RNA-Seq**  
Whole Transcriptome  
sequencing

**ChIP-Seq**  
Chromatin Immunoprecipitation  
with DNA sequencing

# Next Generation Sequencing



# *De novo* transcriptome assembly

No genome required

Empower studies of non-model organisms

- Transcript identification
- expressed gene content
- transcript abundance
- differential expression

# Shortest Superstring Problem

- Problem: Given a set of strings, find a shortest string that contains all of them
- Input: Strings  $s_1, s_2, \dots, s_n$
- Output: A string  $s$  that contains all strings  $s_1, s_2, \dots, s_n$  as substrings, such that the length of  $s$  is minimized
- **Complexity**: NP – complete
- **Note**: this formulation does not take into account sequencing errors

# Shortest Superstring Problem: Example

The Shortest Superstring problem

Set of strings: {000, 001, 010, 011, 100, 101, 110, 111}

Concatenation

Superstring

000 001 010 011 100 101 110 111

010

110

011

Shortest

superstring

000

0 0 0 1 1 1 0 1 0 0

001

111

101

100

PHASE : INTERPRETATION  
TWO :



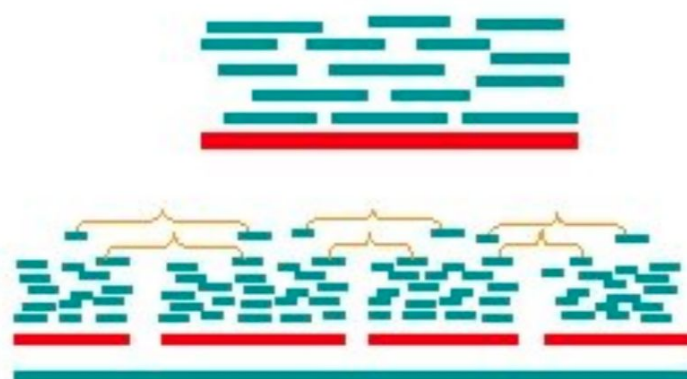
# Overlap-Layout-Consensus

**Assemblers:** ARACHNE, PHRAP, CAP, TIGR, CELERA

**Overlap:** find potentially overlapping reads



**Layout:** merge reads into contigs and contigs into supercontigs



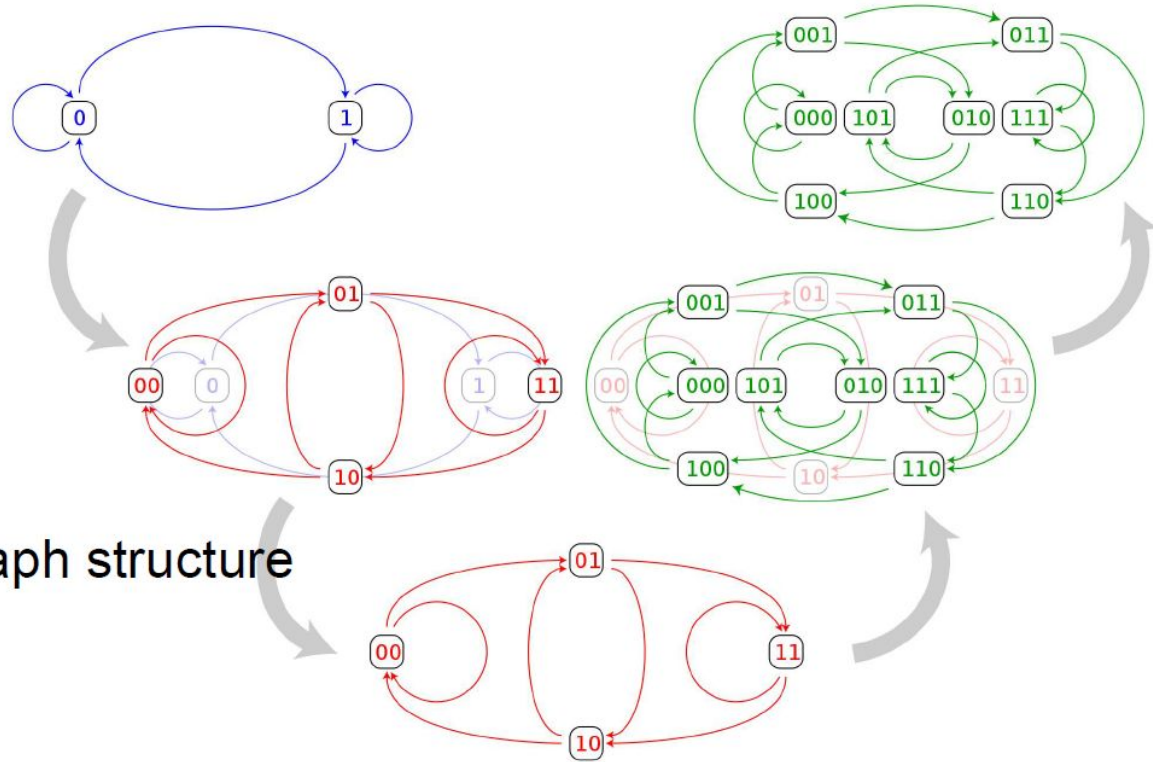
**Consensus:** derive the DNA sequence and correct read errors

..ACGATTACAATAGGTT..

The General Approach to  
*De novo* DNA/RNA-Seq Assembly  
Using De Bruijn Graphs

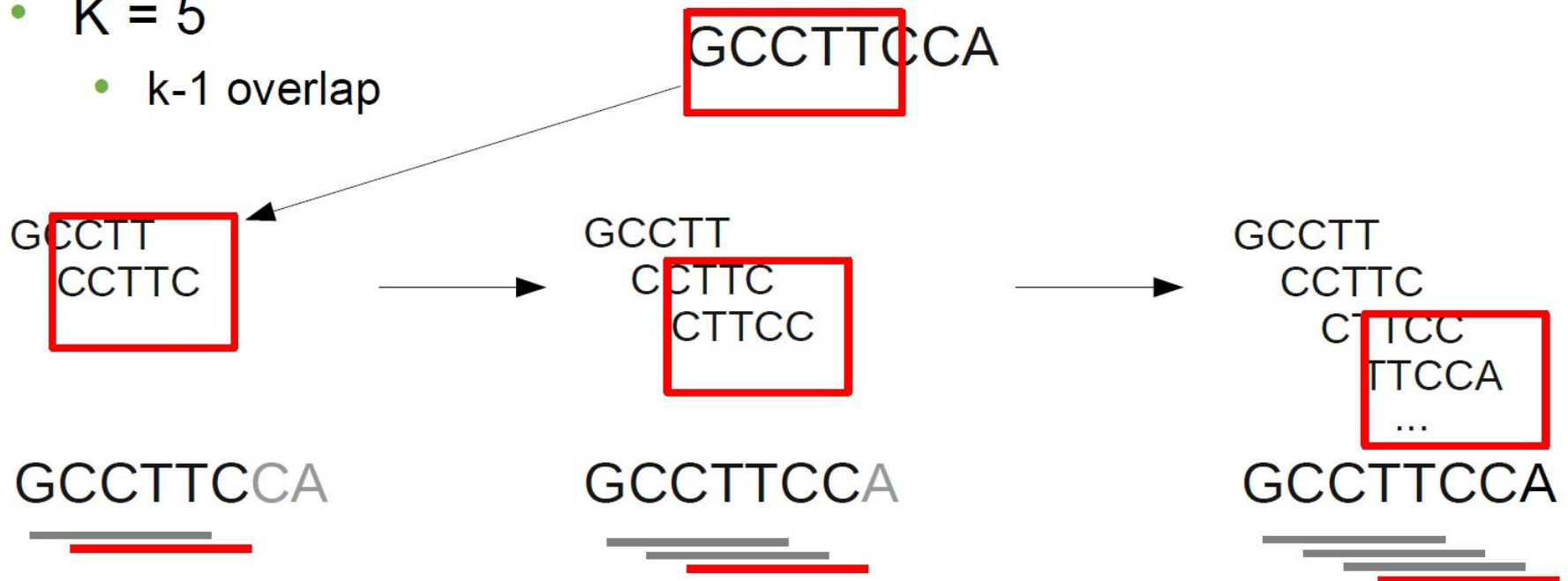
# De Bruijn graph

- A concept in combinatorial mathematics
  - In combinatorics, de bruijn graph is usually fully connected
  - [http://en.wikipedia.org/wiki/De\\_Bruijn\\_graph](http://en.wikipedia.org/wiki/De_Bruijn_graph)
- de bruijn sequence
  - Related concept
  - Path through graph
- Velvet
  - de Bruijn inspired graph structure



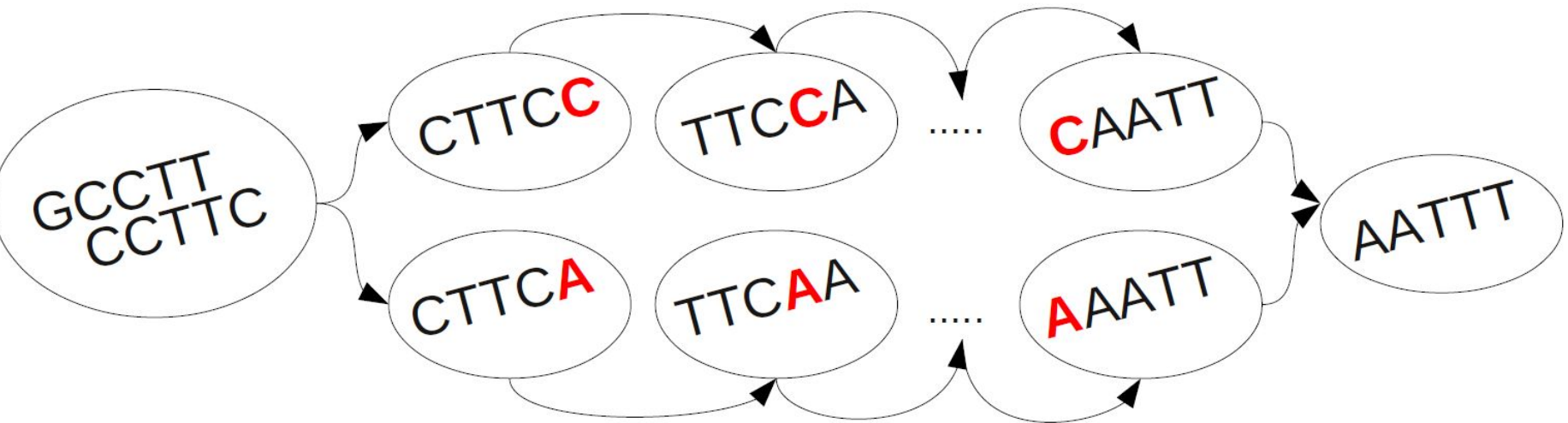
# De Bruijn graph (Velvet)

- Representation of
  - a sequence based on short words (k-mers)
  - overlaps between words
- K-mer: word of length k
- $K = 5$ 
  - k-1 overlap

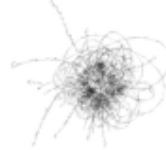


# De Bruijn graph (Velvet)

GCCTTTC**C**AATTT  
GCCTTTC**A**AATTT



# Example



**TAGTCGAGGCTTTAGATCCGATGAGGCTTTAGAGACAG**

```
AGTCGAG CTTTAGA  CGATGAG CTTTAGA
GTCGAGG  TTAGATC  ATGAGGC    GAGACAG
      GAGGCTC    ATCCGAT AGGCTTT GAGACAG
AGTCGAG      TAGATCC ATGAGGC  TAGAGAA
TAGTCGA  CTTTAGA CCGATGA    TTAGAGA
      CGAGGCT  AGATCCG TGAGGCT  AGAGACA
TAGTCGA GCTTTAG TCCGATG  GCTCTAG
      TCGACGC      GATCCGA GAGGCTT AGAGACA
TAGTCGA      TTAGATC GATGAGG TTTAGAG
      GTCGAGG TCTAGAT  ATGAGGC  TAGAGAC
      AGGCTTT  ATCCGAT AGGCTTT GAGACAG
AGTCGAG      TTAGATT  ATGAGGC  AGAGACA
      GGCTTTA  TCCGATG    TTTAGAG
      CGAGGCT TAGATCC  TGAGGCT  GAGACAG
AGTCGAG  TTTAGATC  ATGAGGC  TTAGAGA
      GAGGCTT  GATCCGA GAGGCTT  GAGACAG
```

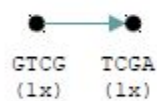
# Example

Read: GTCGAGG

●  
GTCG  
(1x)

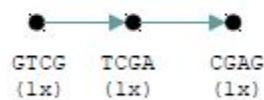
# Example

Read: GTCGAGG



# Example

Read: GTCGAGG



# Example

Read: GTCGAGG



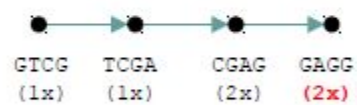
# Example

New read: CGAGGCT



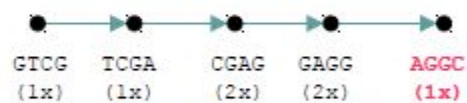
# Example

Read: CGAGGGCT



# Example

Read: CGAGGCT



# Example

Read: CGAGGCT



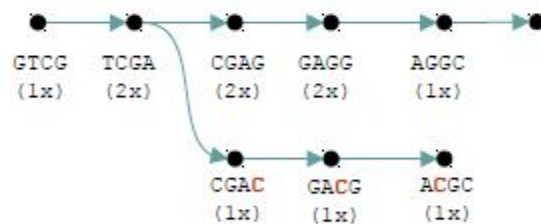
# Example

New read: TCGA**C**GC



# Example

Read: TCGA**CGC**



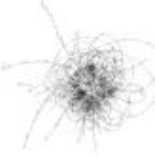
# Example



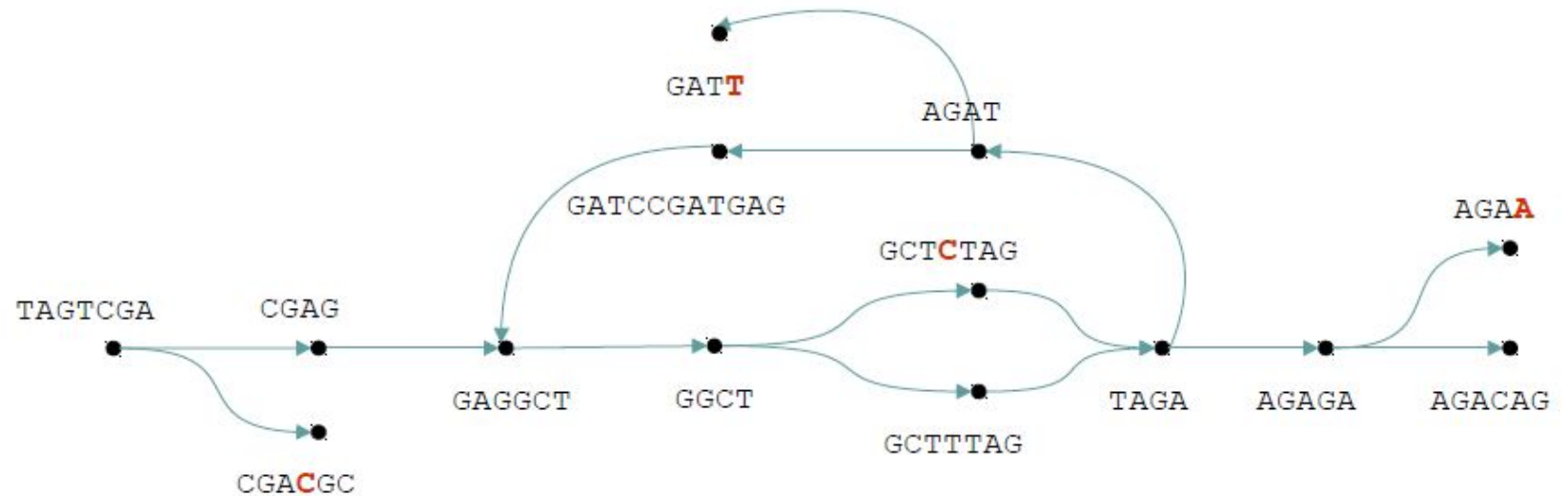
etc...



# Example



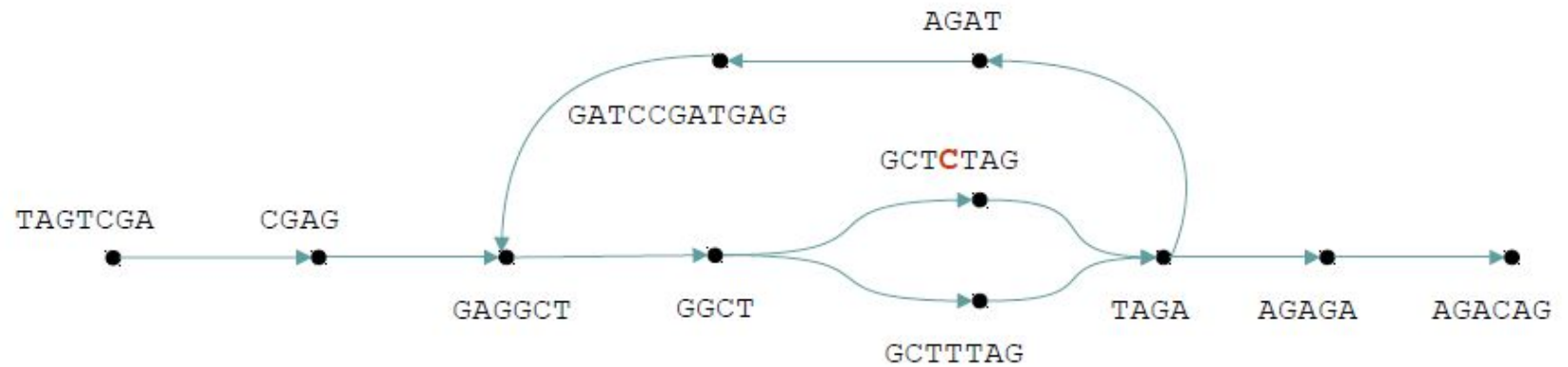
After simplification...



# Example



Tips removed...



# Example

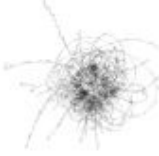
TAGTCGAGGCTTTAGATCCGATGAGGCTTTAGAGACAG

Final simplification...



One possible walk through the graph ...

TAGTCGAG  
    **GAGGCTTTAGA**  
        AGATCCGATGAG  
            **GAGGCTTTAGA**  
                AGAGACAG



2. Sequencing, tools and computers.

### **2.6 Assembly evaluation**

During the assembly optimization will be generated several assemblies. The parameters to evaluate the assembly are:

**1. Total Assembly Size,**

How far is this value from the estimated genome size

**2. Total Number of Sequences (Scaffold/Contigs)**

How far is this value from the number of chromosomes.

**3. Longest scaffold/contig**

**4. Average scaffold/contig size**

**5. N50/L50 (or any other N/L)**

Number sequence (N) and minimum size of them (L) that represents the assembly if the sequences are sorted by size, from bigger to small

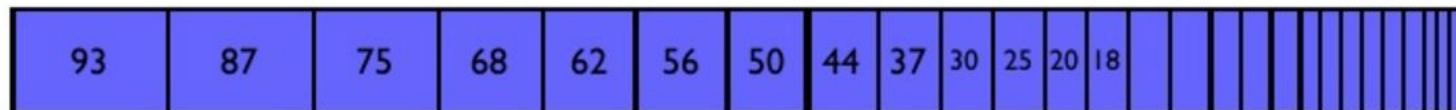
## 2. Sequencing, tools and computers.

### 2.6 Assembly evaluation

**N50/L50**

Total assembly size: 1000 Mb

Sequences order by descending size (Mb)



## 2. Sequencing, tools and computers.

### 2.6 Assembly evaluation

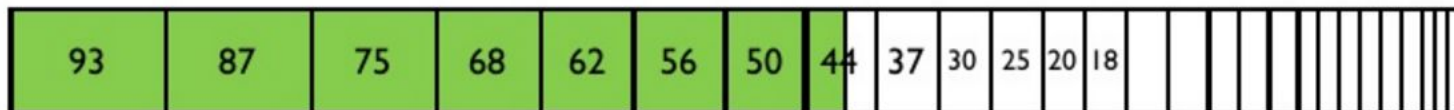
#### N50/L50

Total assembly size: 1000 Mb

N50

50 % assembly: 500 Mb

Sequences order by descending size (Mb)



N50 = 7 sequences

L50 = 50 Mb

## 2. Sequencing, tools and computers.

### 2.6 Assembly evaluation

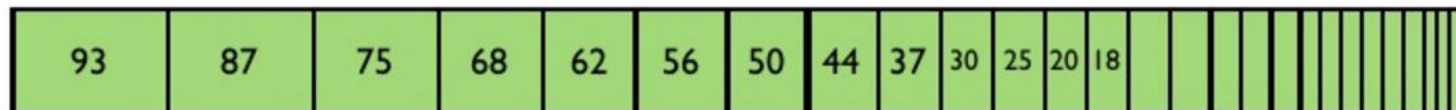
**N90/L90**

Total assembly size: 1000 Mb

N90

90 % assembly: 900 Mb

Sequences order by descending size (Mb)



N90 = 29 sequences

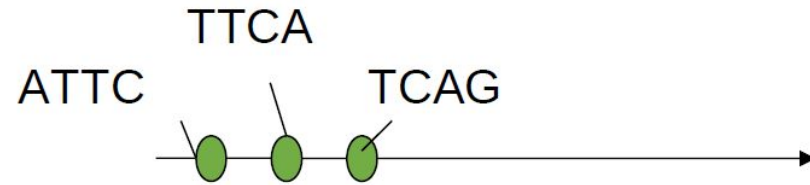
L90 = 12.5 Mb

# De Bruijn graph biology extensions (Velvet)

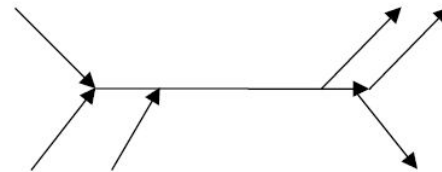
- Handling of reverse strand
  - DNA is read in two directions
  - Paired-end data
- Handling small differences, which are “uninteresting”
  - Errors in sequencing technology
- Memory
  - regularly use 80, 100GB real memory
  - easily get to 1TB real memory requirements

# De Bruijn graph representations (Velvet)

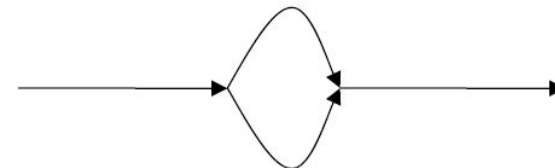
Error free, no repeat,  
no polymorphism



Repeat > kmer length

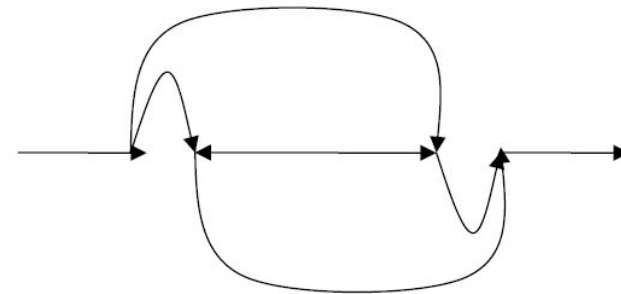


SNP, variant, < kmer length



Structural variant, inversion  
Structural variant, deletion...

...



# Contrasting Genome and Transcriptome Assembly

## Genome Assembly

- Uniform coverage
- Single contig per locus
- Double-stranded

## Transcriptome Assembly

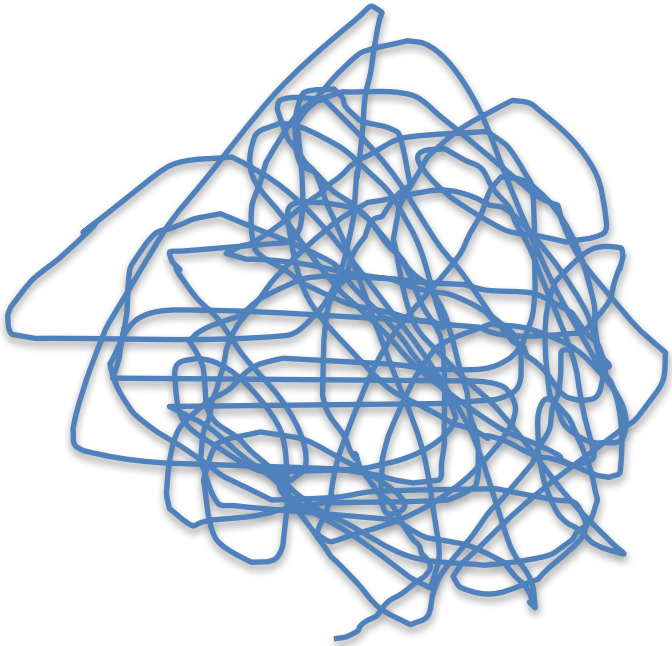
- Exponentially distributed coverage levels
- Multiple contigs per locus (alt splicing)
- Strand-specific



# Trinity Aggregates Isolated Transcript Graphs

## Genome Assembly

Single Massive Graph



Entire chromosomes represented.

## Trinity Transcriptome Assembly

Many Thousands of Small Graphs



Ideally, one graph per expressed gene.

# Applied for: Olive fly *Bactrocera oleae* (dakos)



- Ordo: *Diptera*
- Family: *Tephritidae*
- Genus: *Bactrocera*

- Monophagous
- Production losses > 30% possible
- Affects quantity and quality
- Global economic damage estimated: **800.000.000 \$**



Collaborative effort of

Department of Biochemistry and Biotechnology  
University of Thessaly

Laboratory of Molecular Biology and Genomics

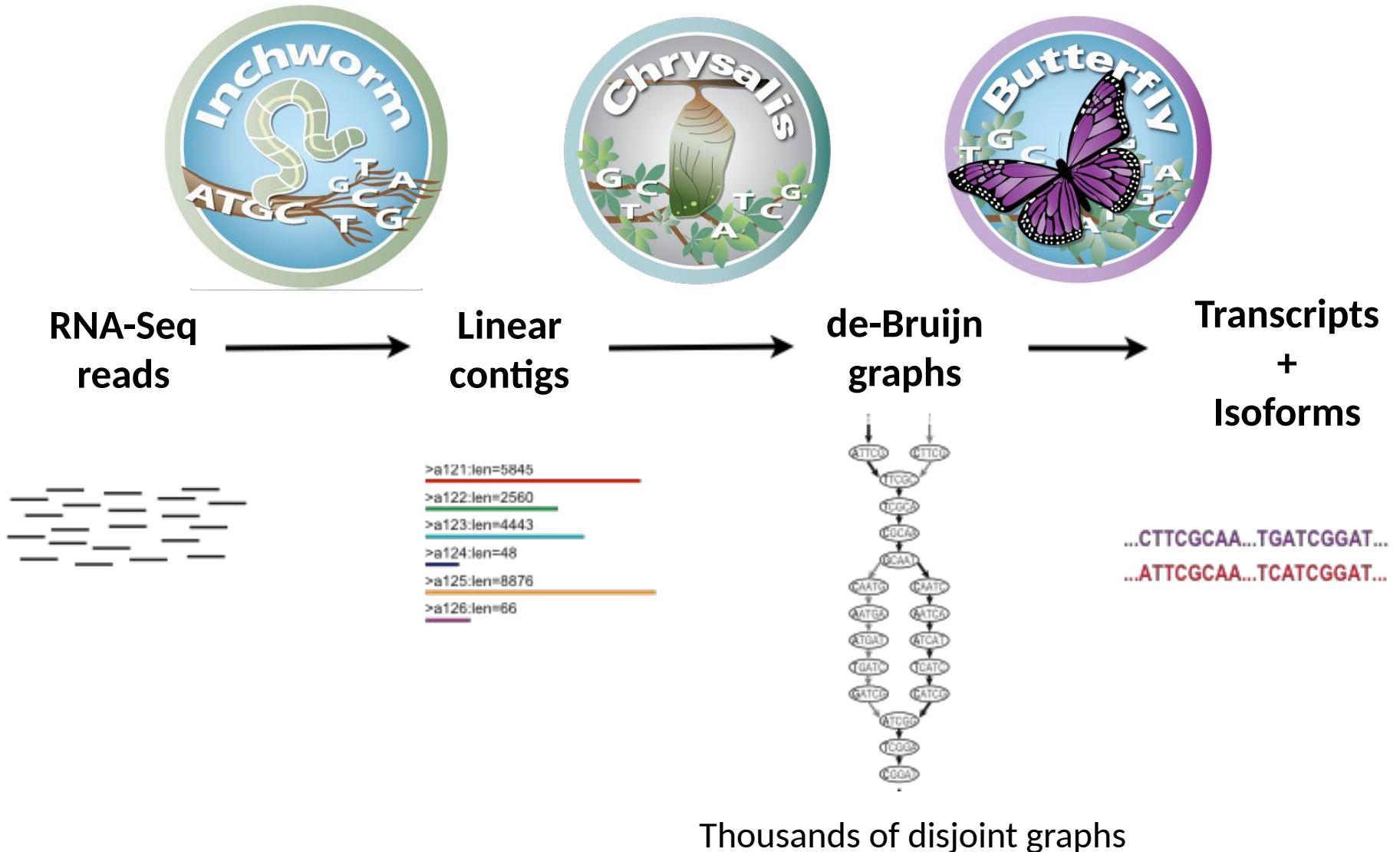
- K. Mathiopoulos, E. Sagri



**ALEXANDER FLEMING**  
Biomedical Sciences Research Center

- J. Ragoussis, M. Reczko, K. Salpea, V. Harokopos, A. Dimopoulos

# Trinity – How it works:



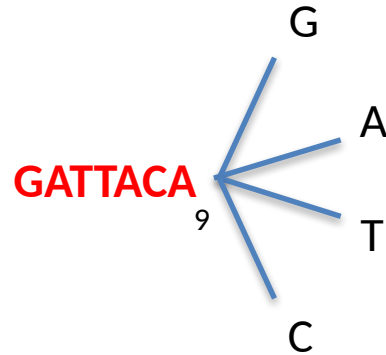


# Inchworm Algorithm

Decompose all reads into overlapping Kmers (25-mers)

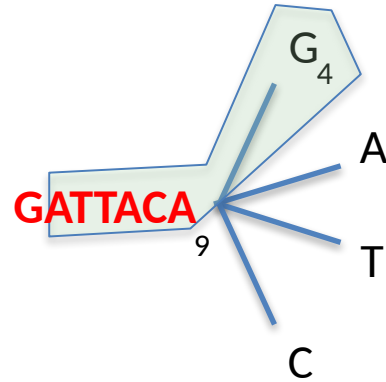
Identify seed kmer as most abundant Kmer, ignoring low-complexity kmers.

Extend kmer at 3' end, guided by coverage.



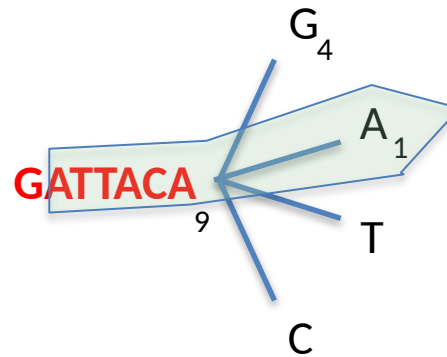


# Inchworm Algorithm



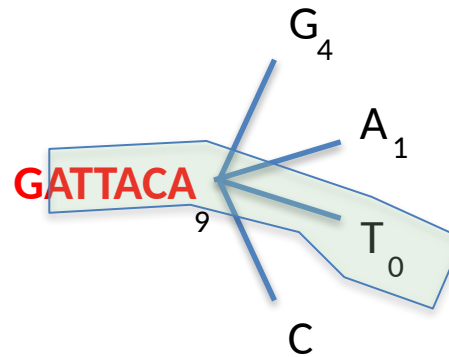


# Inchworm Algorithm



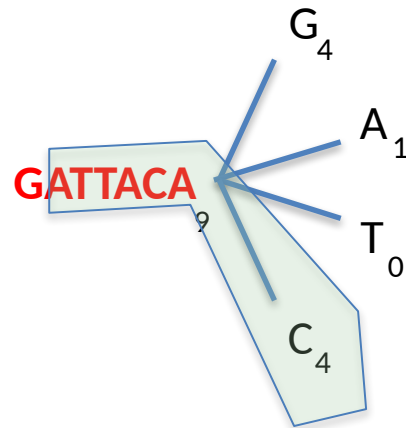


# Inchworm Algorithm



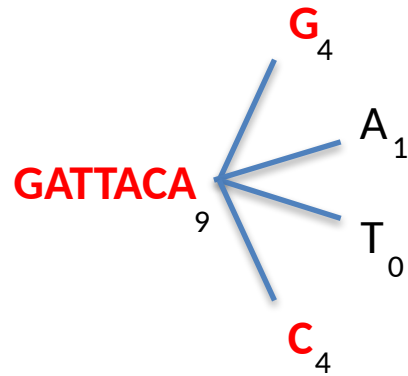


# Inchworm Algorithm



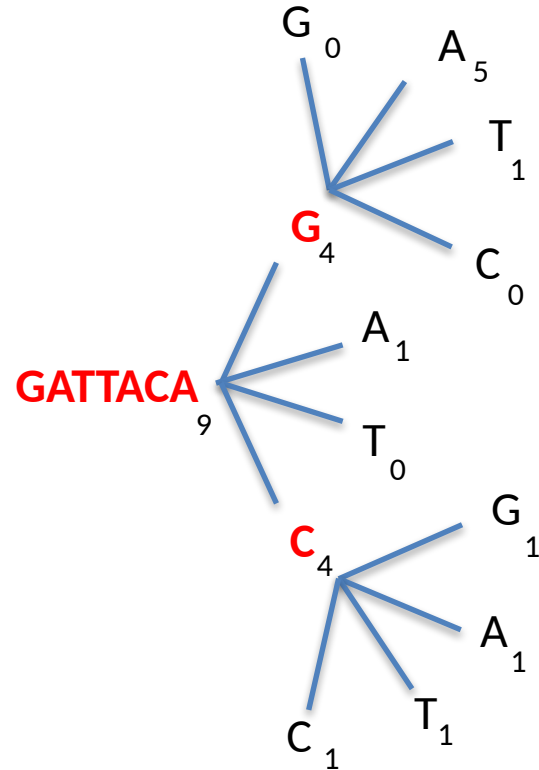


# Inchworm Algorithm



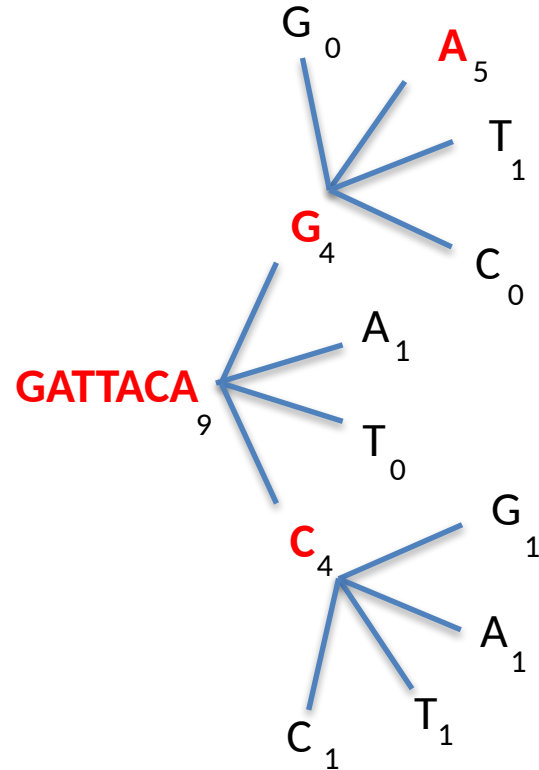


# Inchworm Algorithm



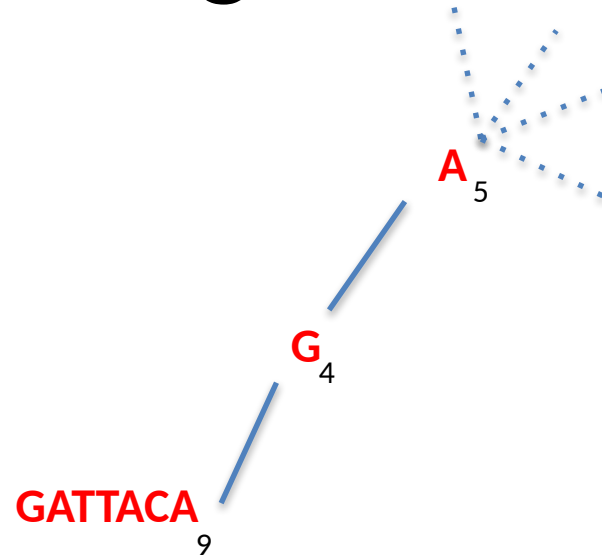


# Inchworm Algorithm



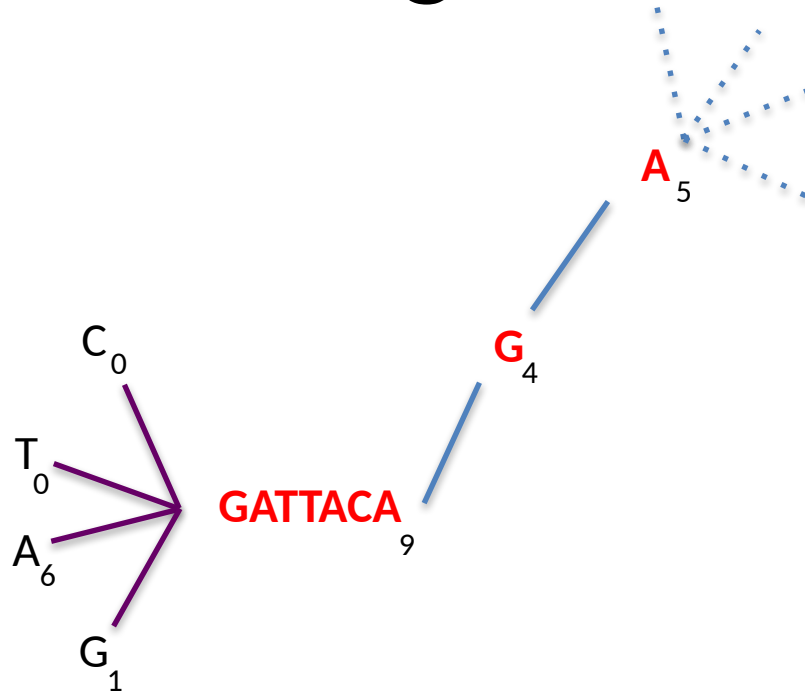


# Inchworm Algorithm



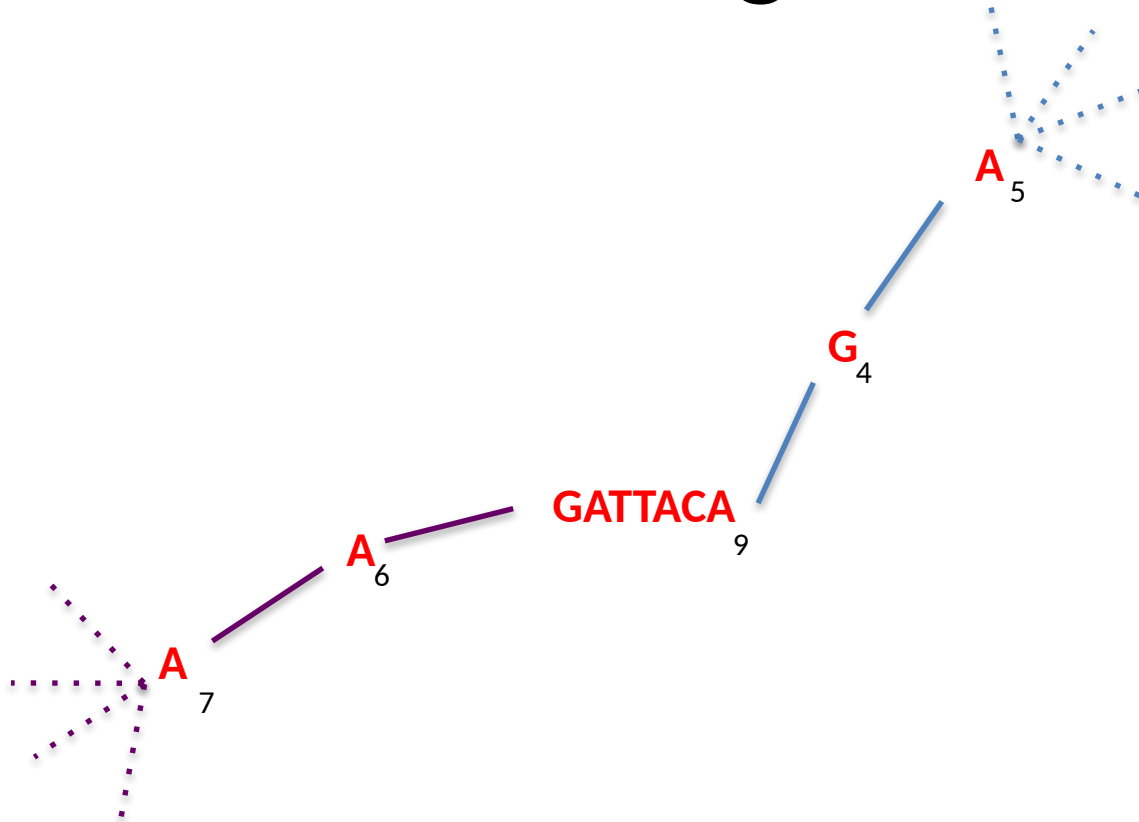


# Inchworm Algorithm





# Inchworm Algorithm



Report contig: ....**AAGATTACAGA**....

Remove assembled kmers from catalog, then repeat the entire process.



# Inchworm Contigs from Alt-Spliced Transcripts

## Expressed isoforms





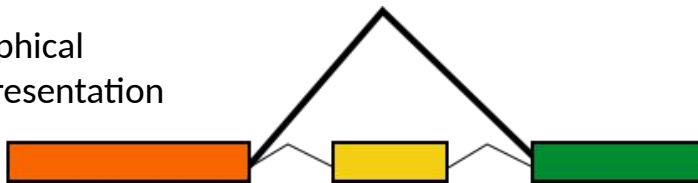
# Inchworm Contigs from Alt-Spliced Transcripts

Expressed isoforms

Expression

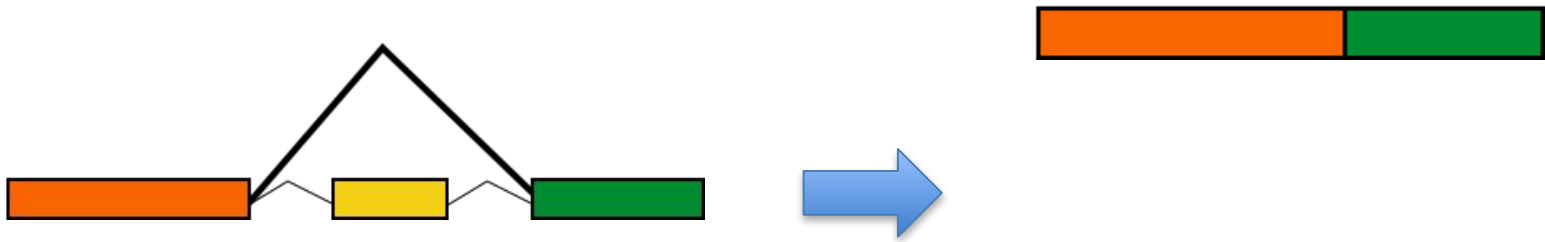


Graphical  
representation



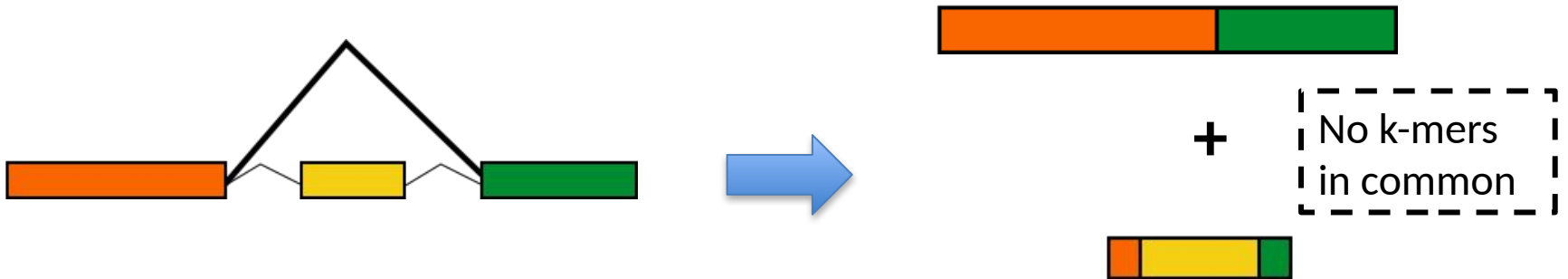


# Inchworm Contigs from Alt-Spliced Transcripts



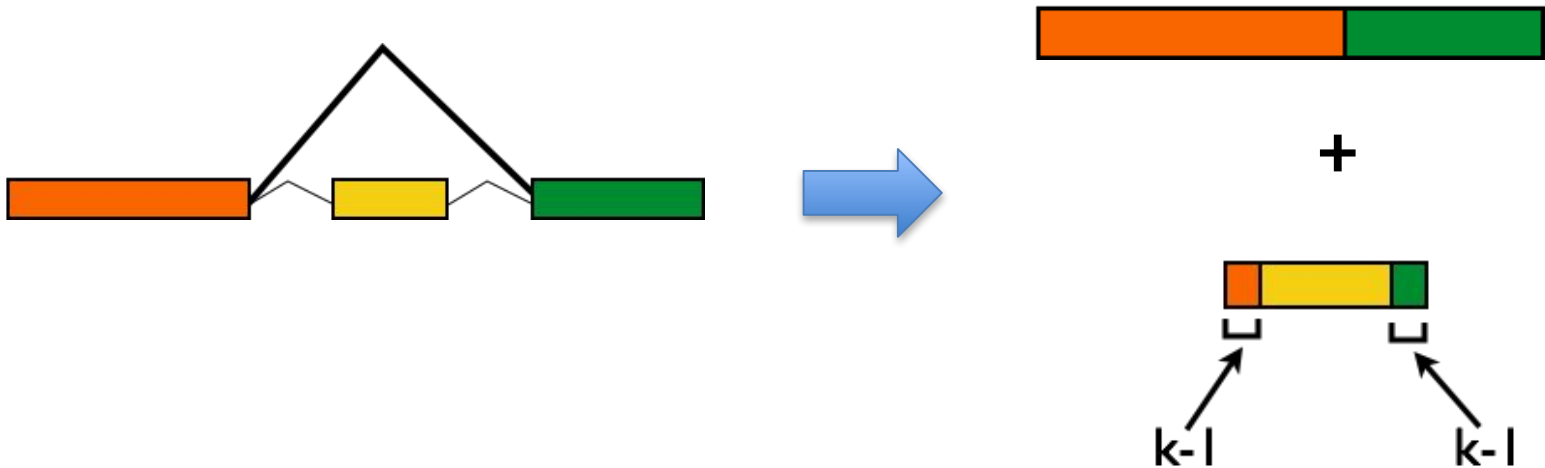


# Inchworm Contigs from Alt-Spliced Transcripts

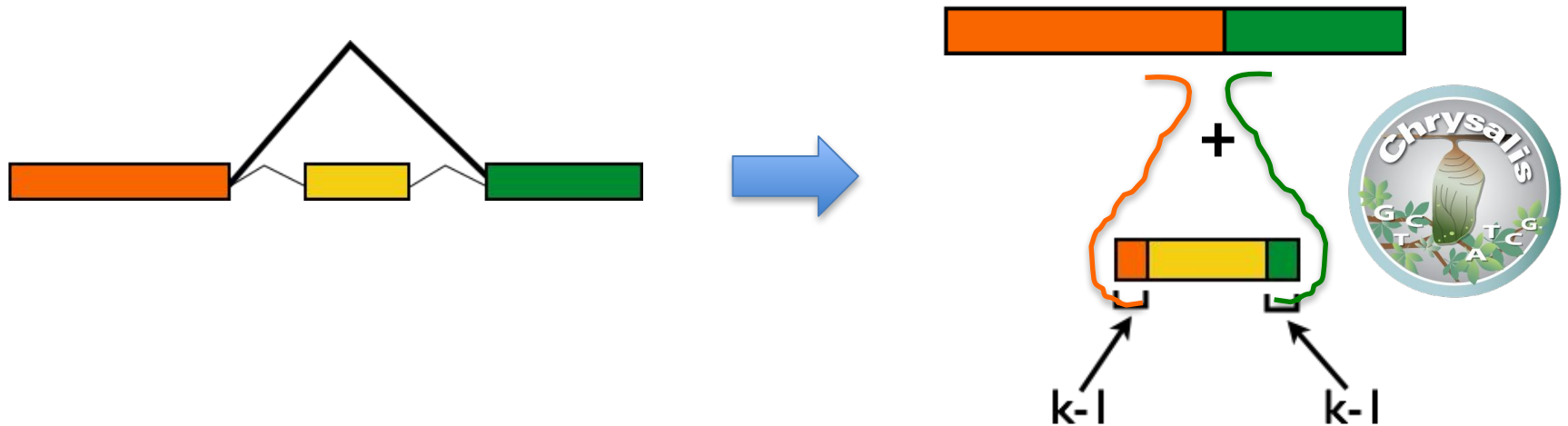




# Inchworm Contigs from Alt-Spliced Transcripts



# Chrysalis Re-groups Related Inchworm Contigs



Chrysalis uses (k-1) overlaps and read support to link related Inchworm contigs

# Chrysalis

>a121:len=5845

>a122:len=2580

>a123:len=4443

>a124:len=48

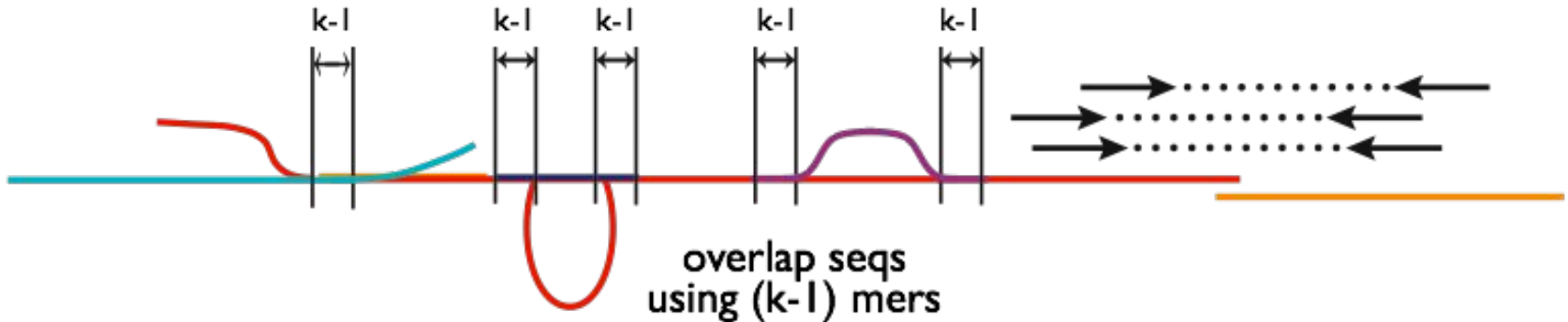
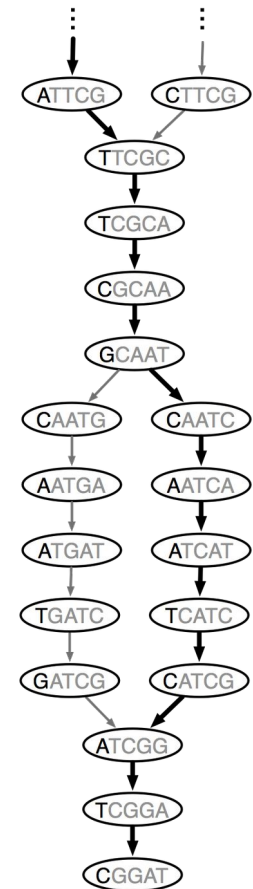
>a125:len=8876

>a126:len=68



Integrate isoforms  
via  $k-1$  overlaps

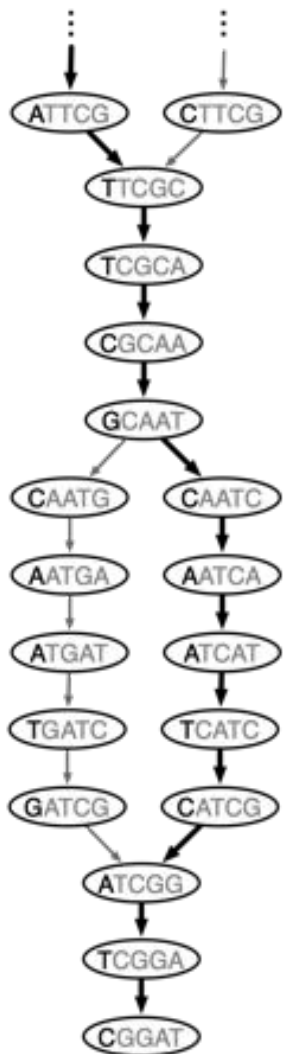
Build de Bruijn Graphs  
(ideally, one per gene)



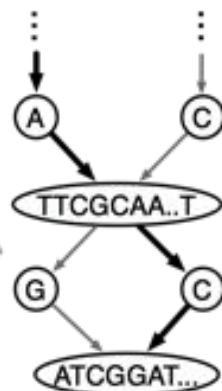


# Thousands of Chrysalis Clusters

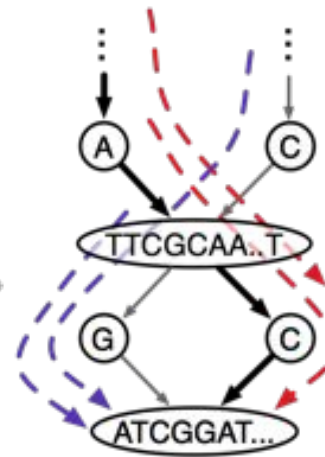
# Butterfly



de Bruijn  
graph



compact  
graph



compact  
graph with  
reads

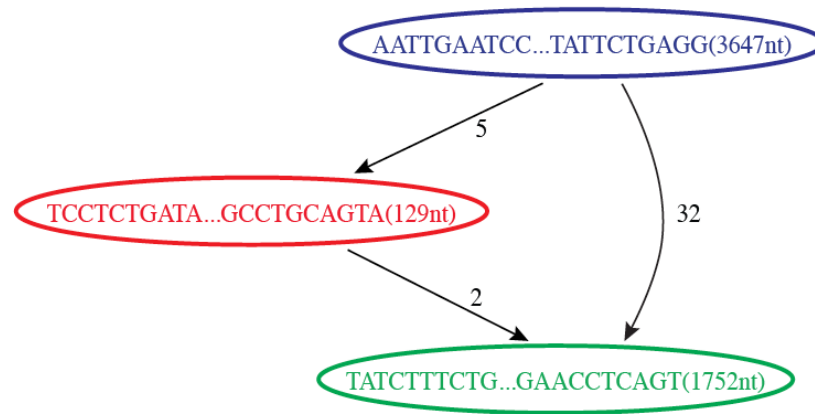


..CTTCGCAA..TGATCGGAT..  
..ATTGCAA..TCATCGGAT..

sequences  
(isoforms and paralogs)

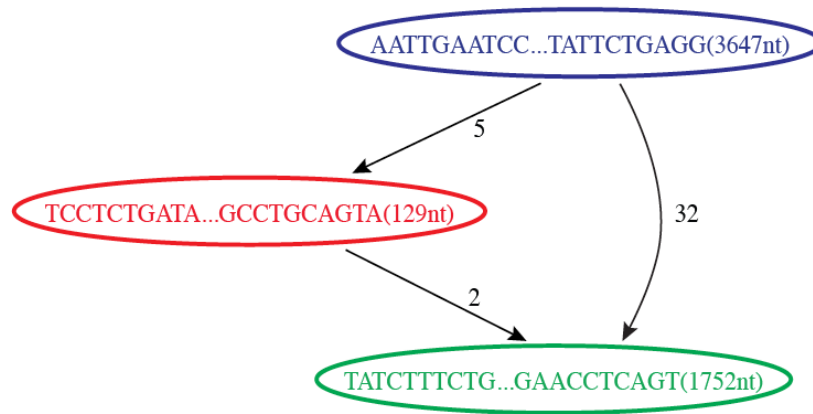
# Butterfly Example 1: Reconstruction of Alternatively Spliced Transcripts

Butterfly's Compacted  
Sequence Graph



# Reconstruction of Alternatively Spliced Transcripts

Butterfly's Compacted  
Sequence Graph

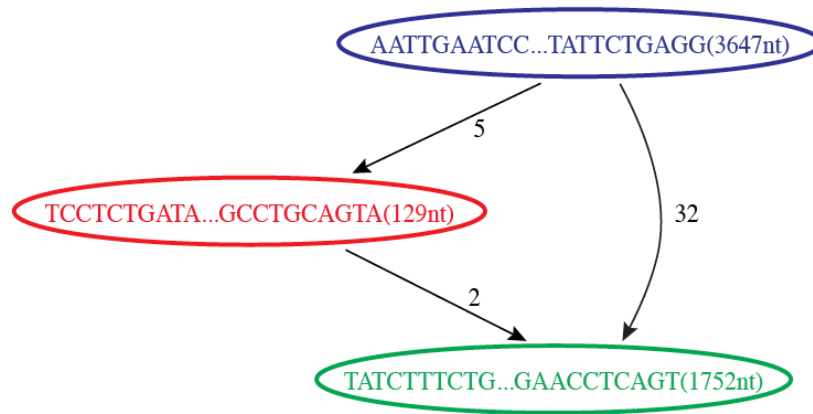


Reconstructed Transcripts



# Reconstruction of Alternatively Spliced Transcripts

Butterfly's Compacted  
Sequence Graph

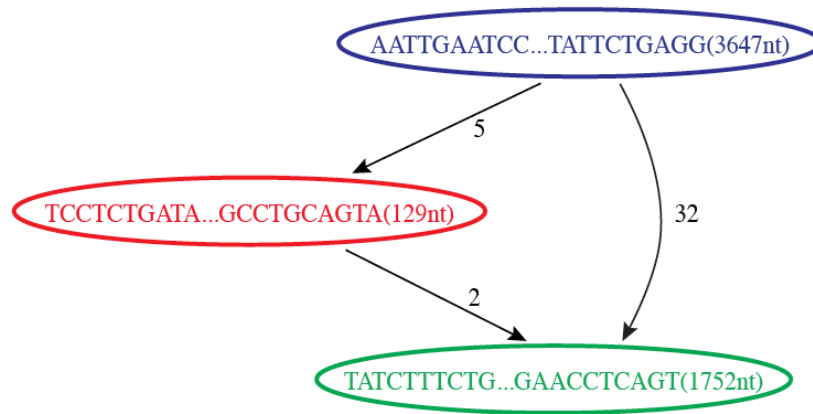


Reconstructed Transcripts



# Reconstruction of Alternatively Spliced Transcripts

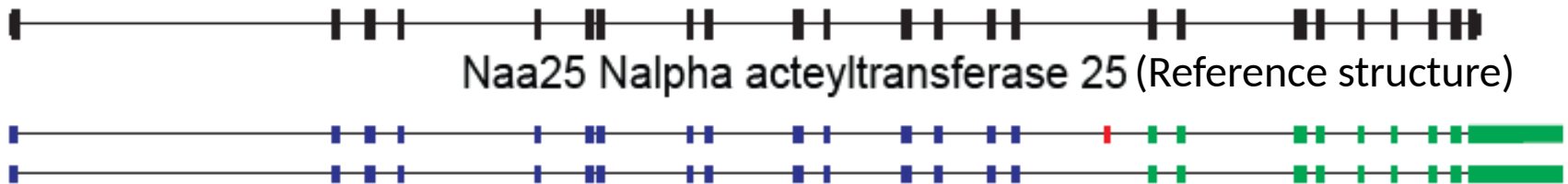
Butterfly's Compacted  
Sequence Graph



Reconstructed Transcripts



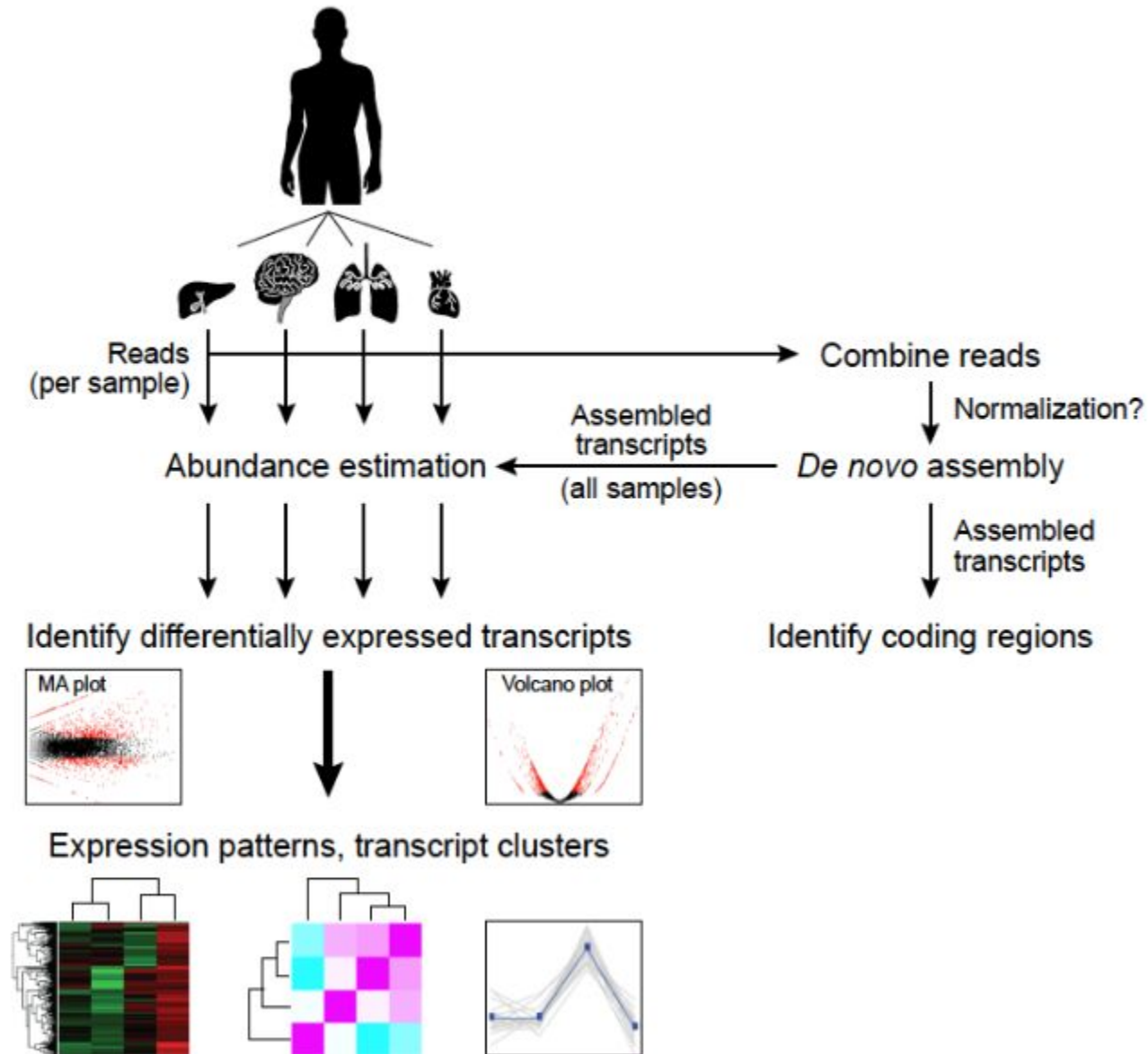
Aligned to Mouse Genome



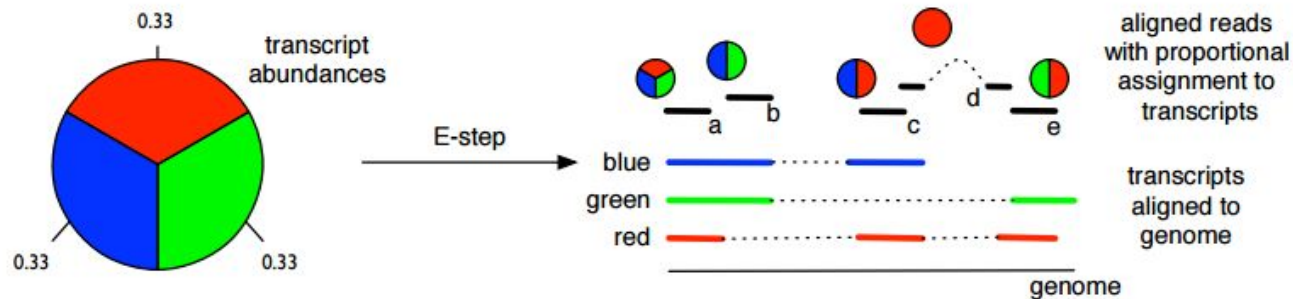
```
>comp0 c0 seq1 len=5528 path=[1:0-3646 10775:3647-3775 3648:3776-5527]
```

>comp1\_c0\_seq2\_len=5399 path=[1:0-3646 3648:3647-5398]

# Trinity Demo



## Expectation maximization used in rsem



### Step E: Expectation (E-Step)

In this step, the algorithm uses the **current best estimate** of transcript expression levels to determine the **probability** that each ambiguous read originated from a specific transcript.

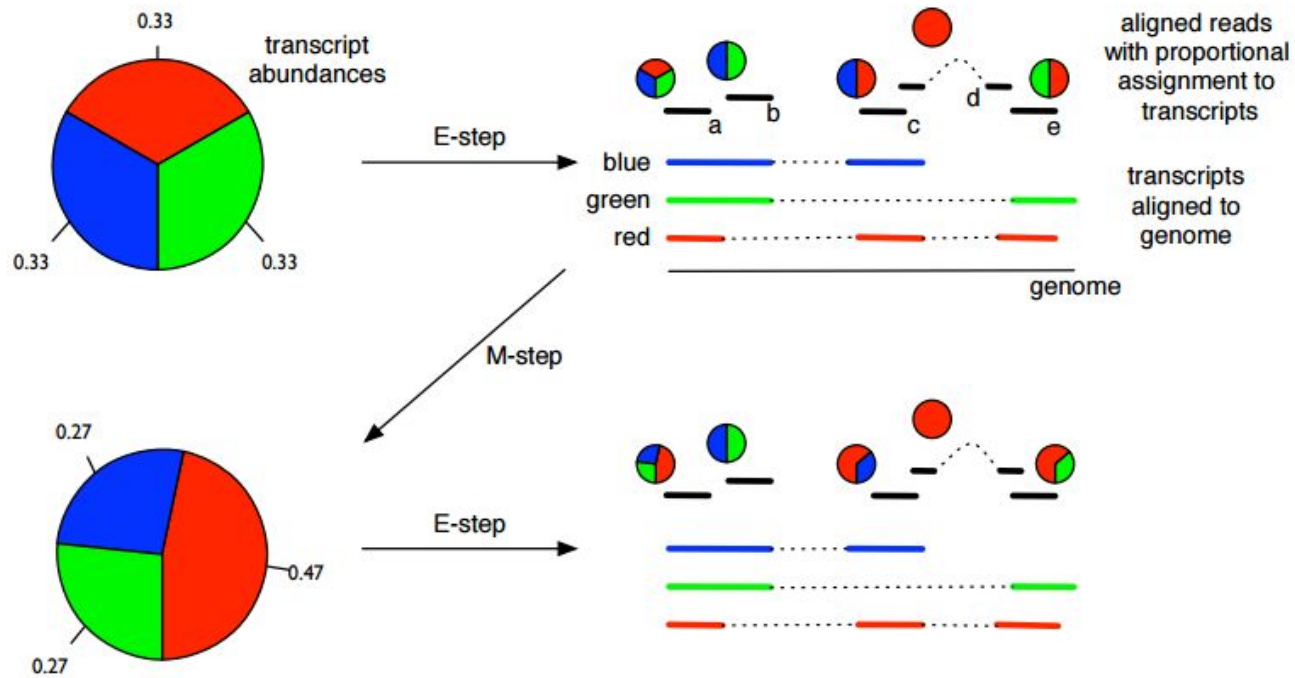
**Calculation:** For every ambiguous read, the probability of it coming from a specific transcript is calculated based on the current relative abundance of that transcript.

### Step M: Maximization (M-Step)

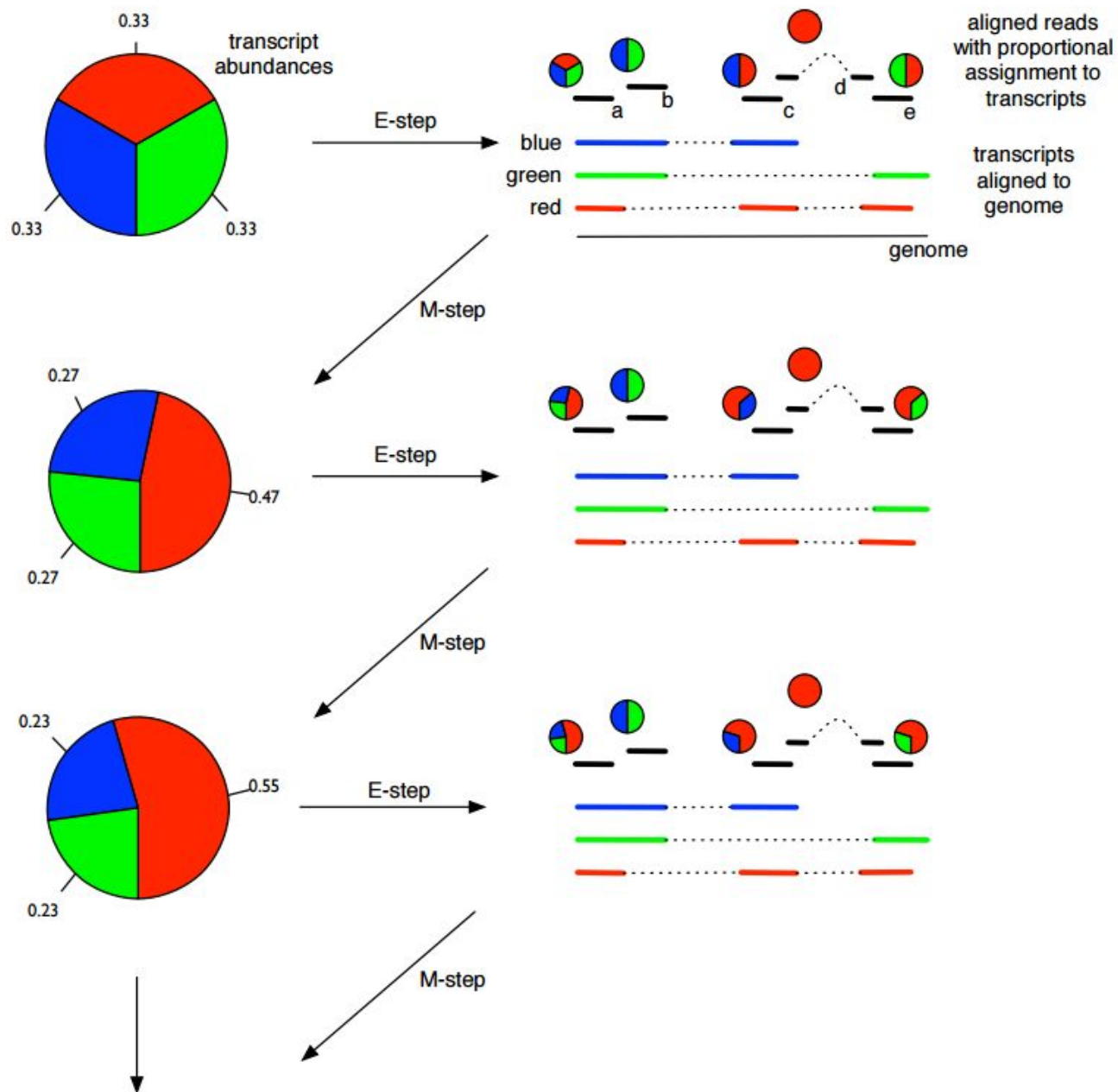
In this step, the algorithm recalculates the **total abundance** of each transcript using the **new, probabilistic counts** determined in the E-step. This maximizes the likelihood of the observed data.

- **Calculation:** The estimated count for each transcript is simply the sum of all probabilities assigned to it from every single read (ambiguous and unambiguous).

# Expectation maximization used in rsem



# Expectation maximization used in rsem



# Trinity Demo

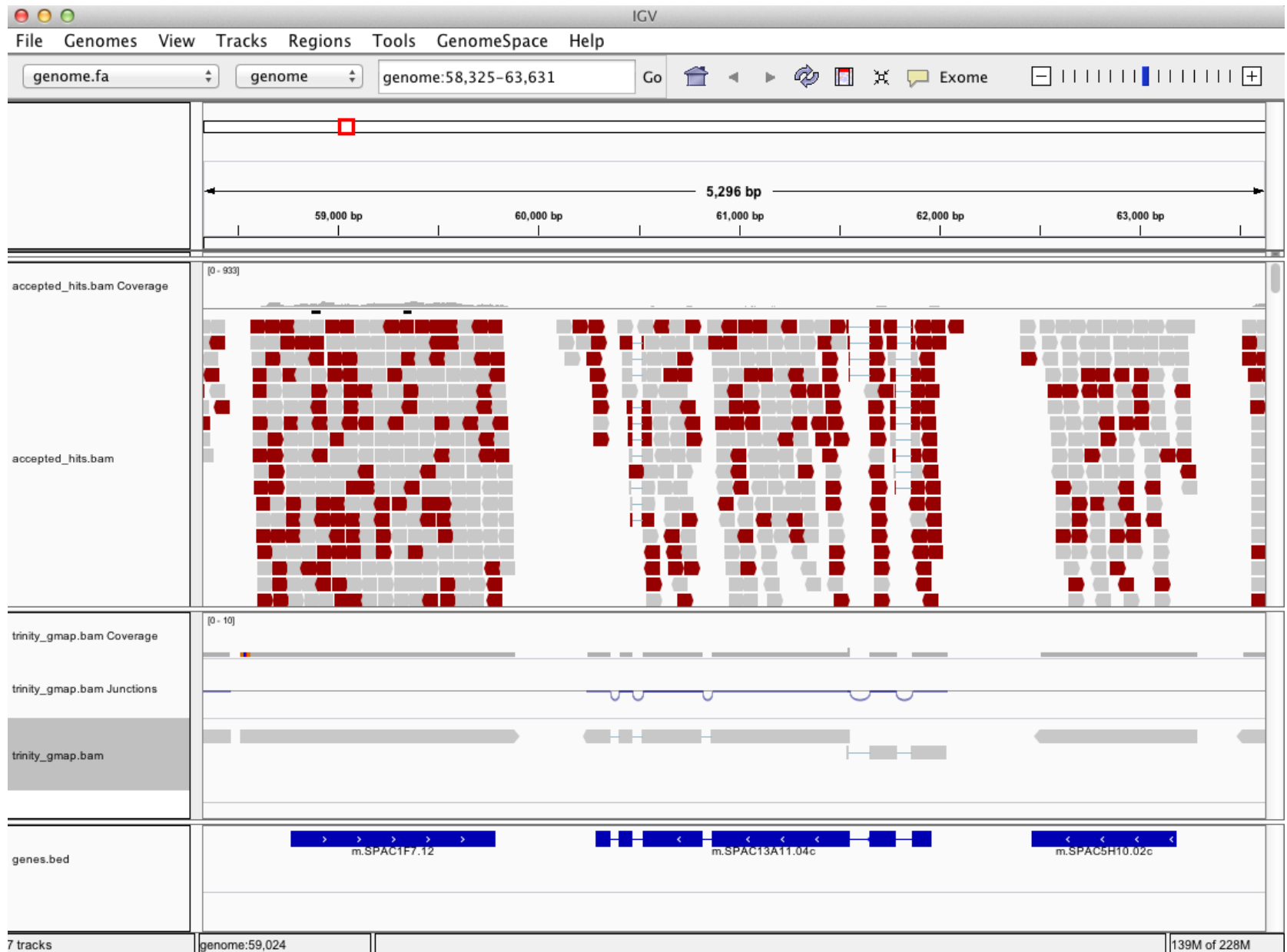
- Assemble RNA-Seq using Trinity
- Examine Trinity in context of a genome:
  - Align Trinity transcripts to the genome using GMAP
  - Align rna-seq reads to genome using Tophat
  - Visualize all alignments using IGV

Try yourself:

```
echo 'export TRINITY_HOME=/home/reczko/tools/trinityrnaseq-v2.15.2' >> ~/.bashrc
source ~/.bashrc
export PATH=$PATH:/usr/local/sbin:/usr/local/bin:/usr/sbin:/usr/bin:/sbin:/bin:/snap/bin
cd ~/rnaseq_workshop
cp /home/reczko/tools/runTrinityDemo.pl .
./runTrinityDemo.pl
```

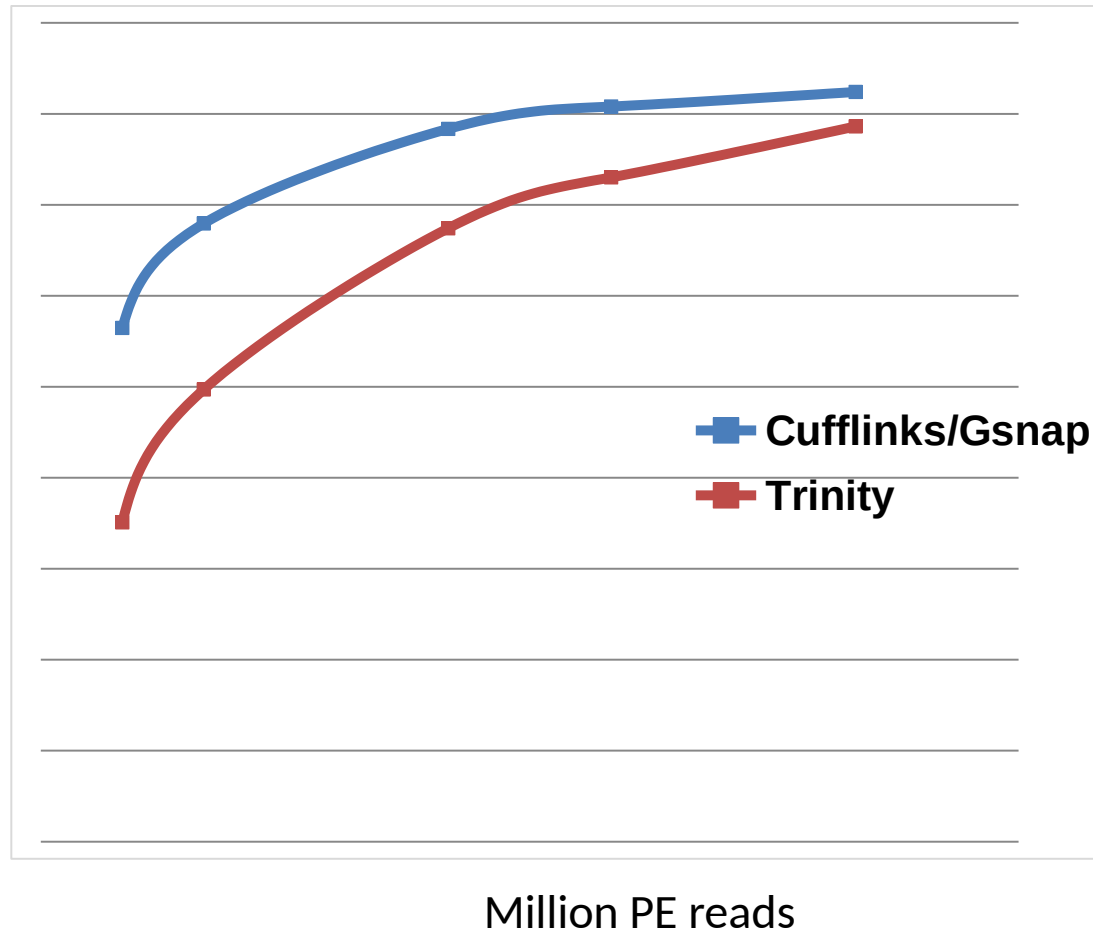
# Trinity transcripts aligned to genome scaffolds to examine intron/exon structures

(Trinity transcripts aligned using GMAP)



# Improved reconstruction with deeper sequencing depth and

Genome-based reconstruction is  
more sensitive than de novo methods



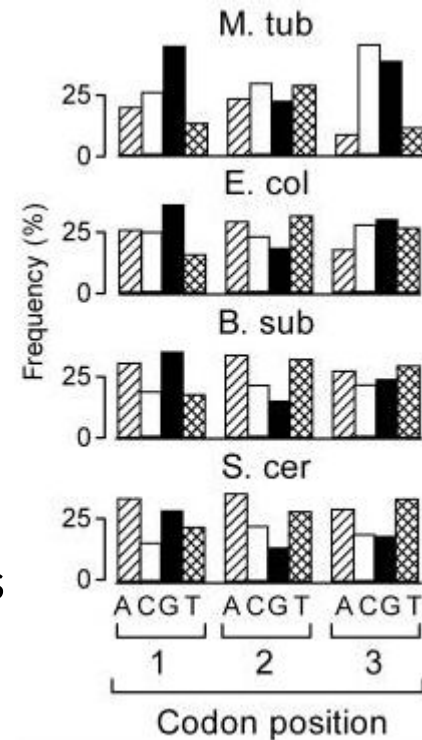
# Genes w/ fully  
reconstructed  
transcripts



Mouse data

# Prediction of coding potential

- Periodicity detection
  - Coding sequences have an inherent periodicity of three
  - Especially good on long coding sequences
  - Auto-correlation
    - Seeking the strongest response when shifted sequence is compared with original
    - Michel (1986), *J. Theor. Biol.* **120**, 223-236.
  - Fourier transformation: Spectral analysis
    - Detection of peak at position corresponding to 1/3 of the frequency
    - Silverman and Linsker (1986), *J. Theor. Biol.* **118**, 295-300.



# 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>