Syllabus and grading

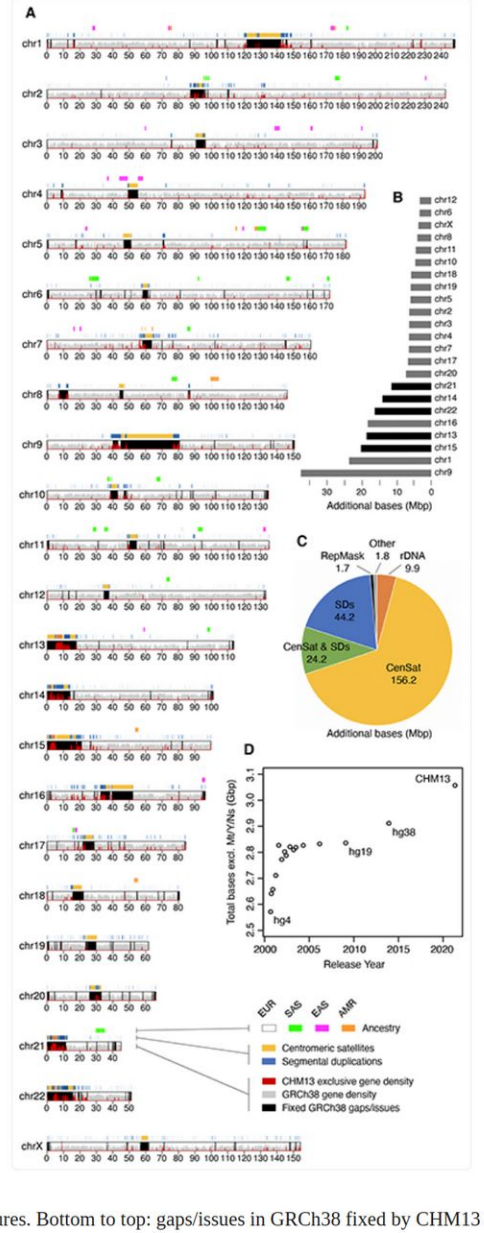
#	Date	Lecturer	Subject
1	6/10/26	MR	Overview of Bioinformatics, sequence alignment
2	13/10/26	AD	Introduction to Linux and the command line, bash scripting and ssh
3	20/10/26	MR	Next generation sequencing: introduction, quality control and gene expression analysis for RNAseq
4	27/10/26	MR	NGS for denovo 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%

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

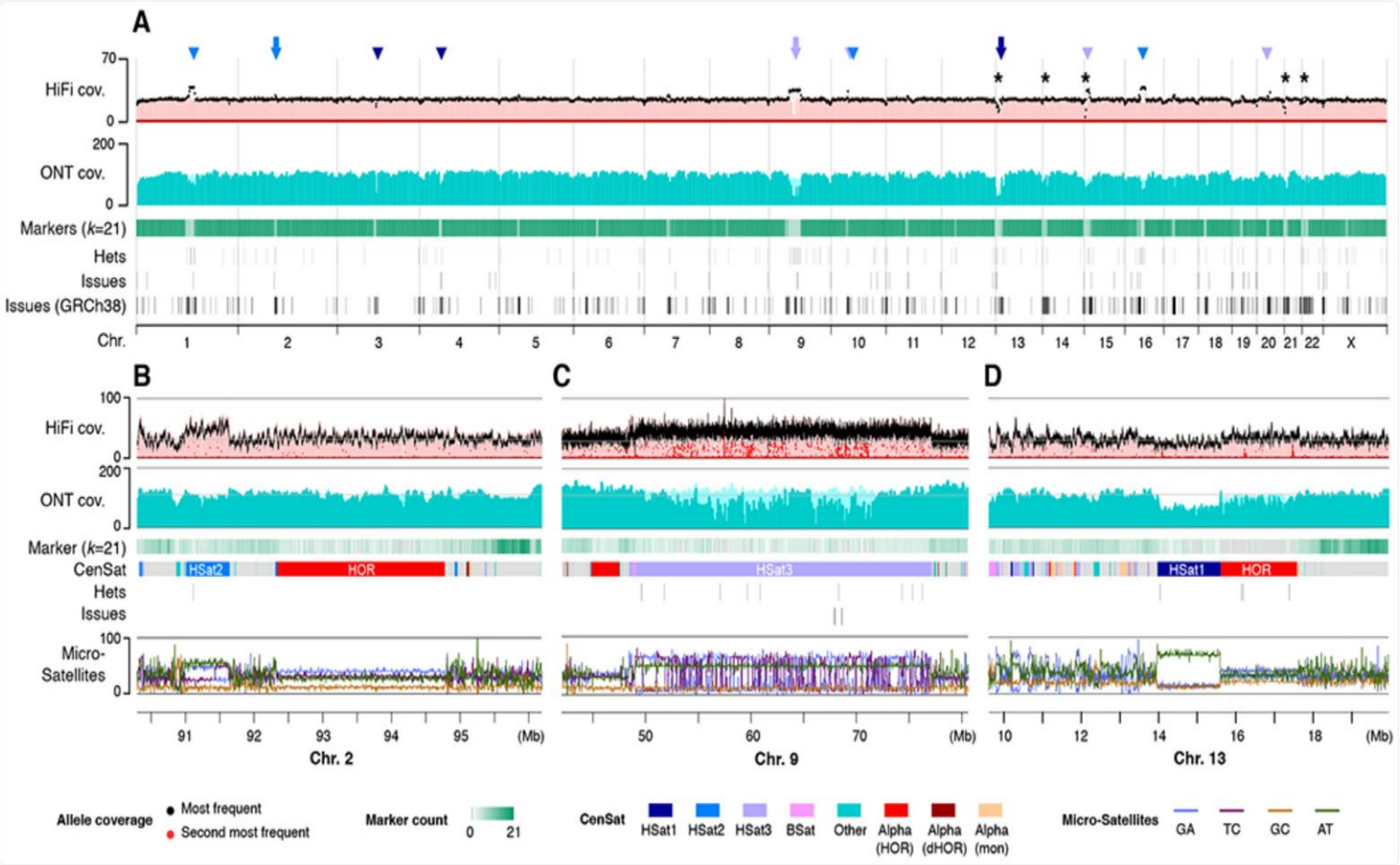
The complete sequence of a human genome

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



(**A**) Ideogram of T2T-CHM13v1.1 assembly features. Bottom to top: gaps/issues in GRCh38 fixed by CHM13 overlaid with the density of g CHM13 in red; segmental duplications (SDs) (42) and centromeric satellites (CenSat) (30); and CHM13 ancestry predictions (EUR, Europea Asian; EAS, East Asian; AMR, Ad Mixed American). (**B**) Additional (non-syntenic) bases in the CHM13 assembly relative to GRCh38 per c the acrocentrics highlighted in black, and (**C**) by sequence type (note that the CenSat and SD annotations overlap). (**D**) Total non-gap bases i 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 Genomics: From Hundreds of Cells to Spatial Multiomic Atlases (2009-2026)

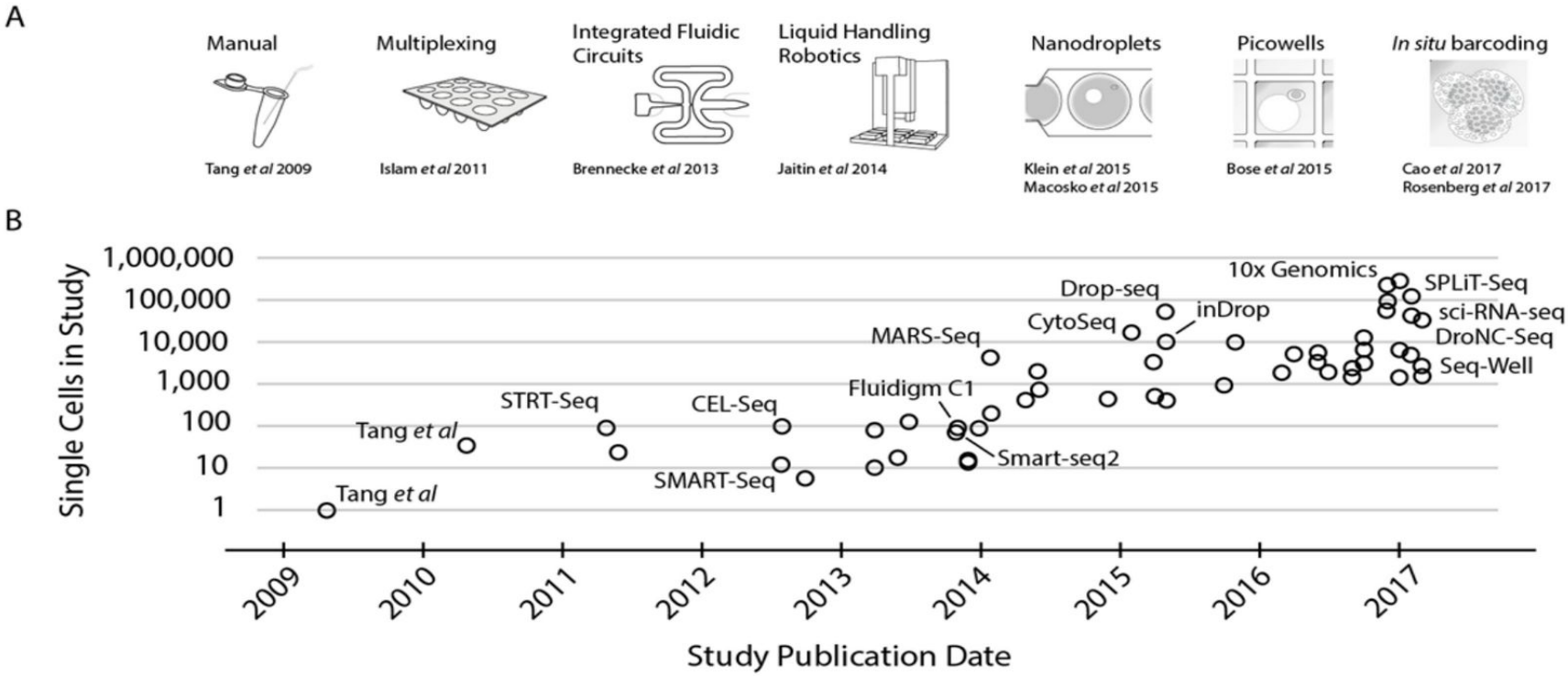


Figure 1: Scaling of scRNA-seq experiments (**A**) Key technologies allowing jumps in experimental scale. A jump to ~100 cells was enabled by sample multiplexing, a jump to ~1,000 cells by large scale studies using integrated fluidic circuits (IFCs), followed by a jump to several thousands using liquid 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 - Eric Chow (UCSF)
https://www.youtube.com/watch?v=k9VFNLLQP8c

Single-Cell Genomics: From Hundreds of Cells to Giga-Scale Atlases

Historical Foundation (2009–2017 Breakthroughs)

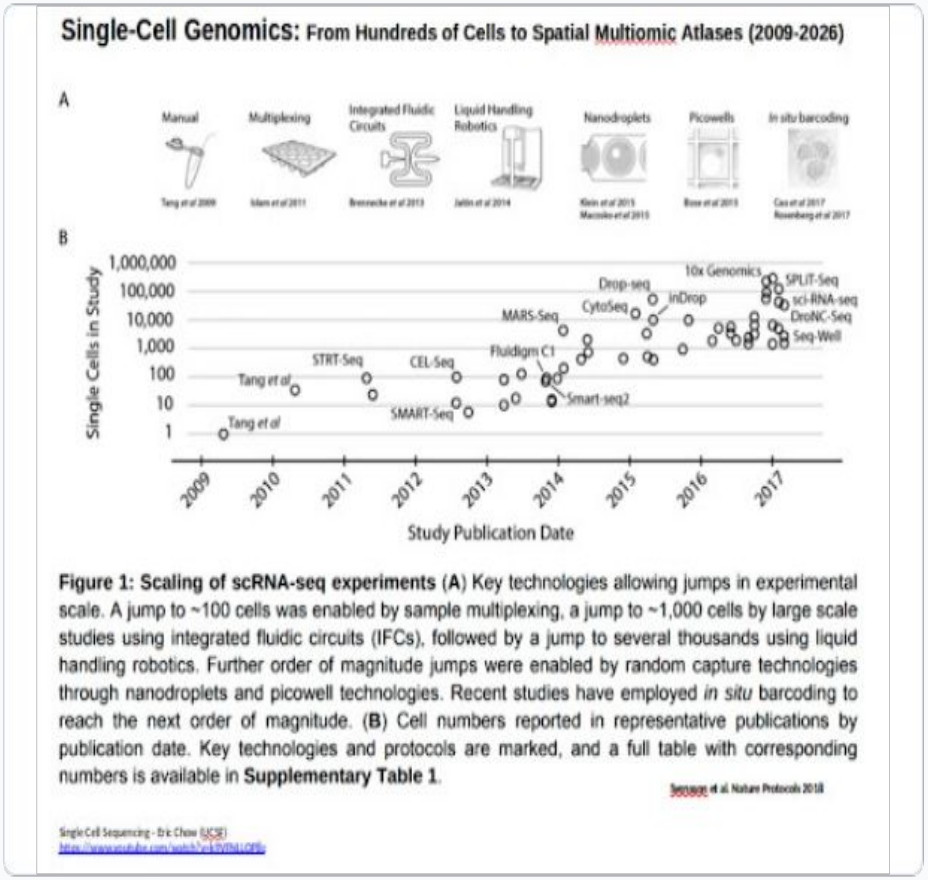


Figure 1: Scaling of scRNA-seq experiments (**A**) Key technologies allowing jumps in experimental scale. A jump to ~100 cells was enabled by sample multiplexing, a jump to ~1,000 cells by large scale studies using integrated fluidic circuits (IFCs), followed by a jump to several thousands using liquid 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 - Eric Chow (UCSF)
https://www.youtube.com/watch?v=kVVNLLOPBs

Modern SOTA (2018–2026)

100M+ Cells & Multimodal AI

2018 – 2022 • SUB-CELLULAR SPATIAL RESOLUTION

Spatial Multiomics & In Situ Sequencing

Integration of spatial transcriptomics and epigenomics. Direct *in situ* molecular profiling preserving tissue architecture.

2023 – 2025 • GIGA-SCALE THROUGHPUT

100M+ Cell Atlas Datasets

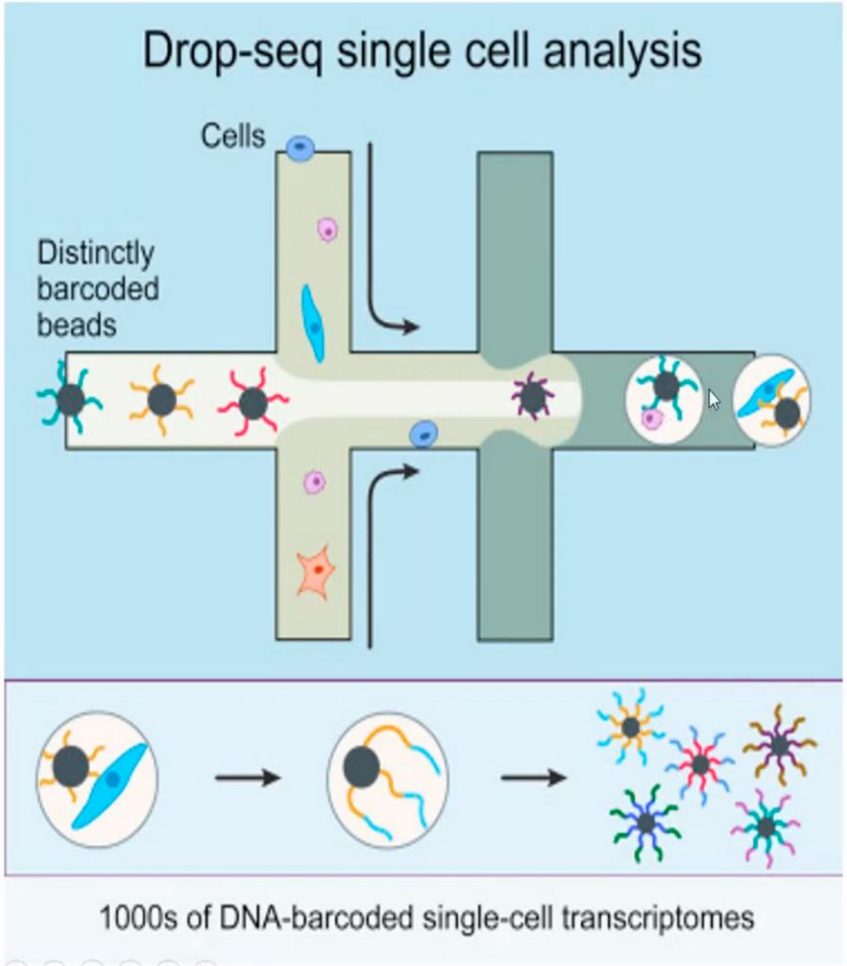
Droplet-free combinatorial indexing enables ultra-scale screening (e.g. Tahoe-100M, Perturb-seq) across diverse cellular contexts.

2025 – 2026 SOTA FRONTIER

Single-Cell AI Foundation Models

Billion-parameter transformers (e.g. Tahoe-x1, Virtual Cell Challenge) driving accurate zero-shot biological perturbation predictions.

Bead: Cell barcode and unique molecular identifiers (UMIs)



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

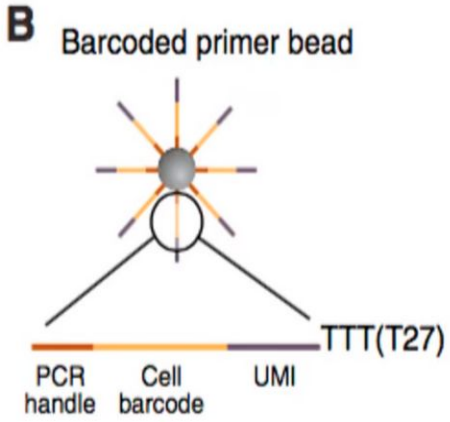


Figure by Macosko et al, Cell, 161:1202-1214, 2015

From reads to digital gene expression matrix (DGE)

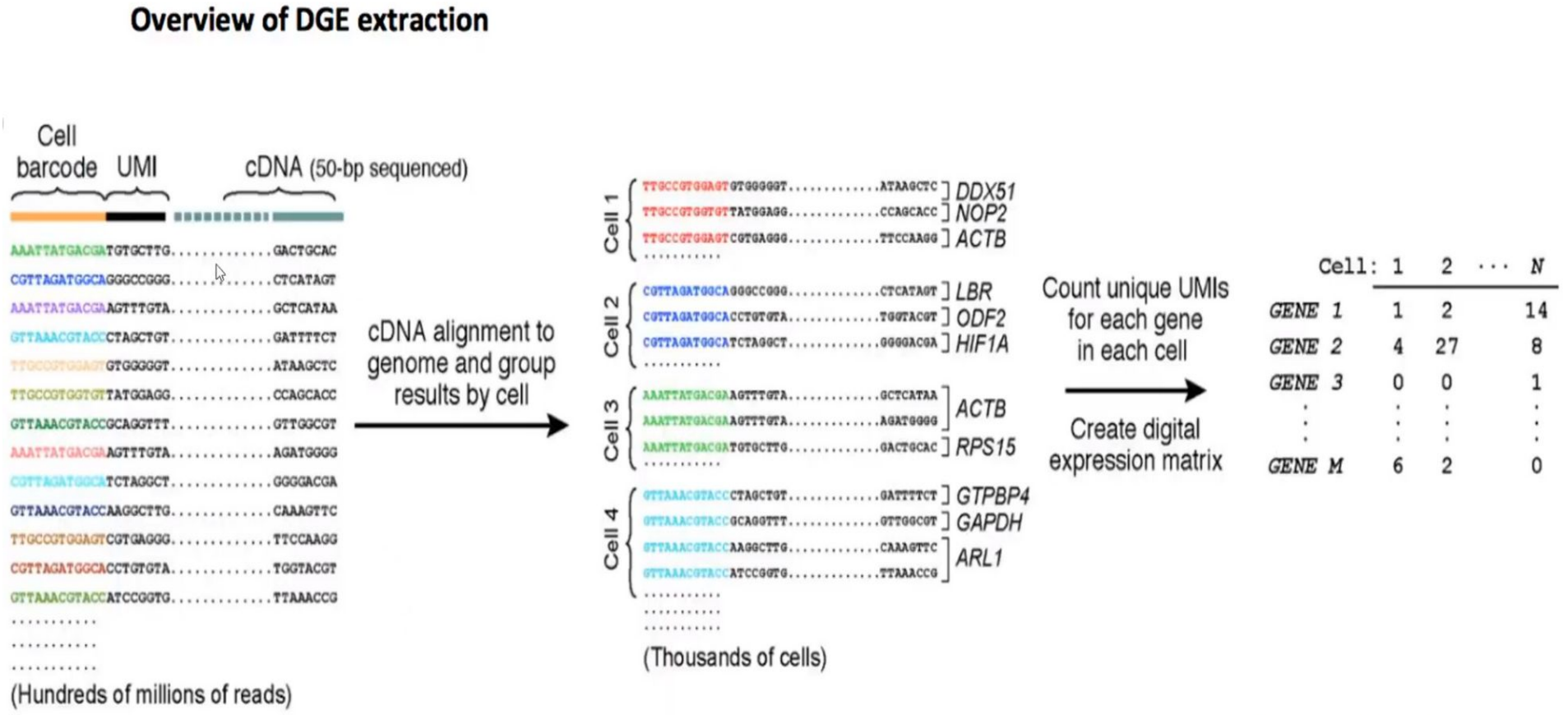


Figure by Macosko et al, Cell, 161:1202-1214, 2015

What is *single-nucleus RNA-seq*?

snRNA-seq measures which genes are turned on inside the nucleus of one cell at a time. Scientists isolate the nucleus (the control center of the cell), capture its RNA messages, and use barcodes so every message can be traced back to its original nucleus.

01

Frozen & archived tissue

Many valuable samples (biobanks, hospital blocks) are frozen or fixed. Whole cells often die or break; nuclei survive much better.

02

Fragile or sticky cells

Brain neurons, fat cells, heart muscle cells, and some immune cells are hard to keep intact. Isolating nuclei avoids the damage of enzymatic digestion.

03

Less cell-type bias

When you try to make a single-cell suspension, some cell types are lost or stressed. Nuclei give a more complete picture of the tissue.

From tissue to data — *step by step*

1. **Isolate nuclei**
   Tissue is gently broken open (often by douncing or mild detergent). Nuclei are cleaned by filtering and centrifugation. Quality is checked under a microscope.

2. **Load into droplets**
   Nuclei + barcoded beads + enzymes are mixed in a microfluidic chip. Each droplet ideally holds one nucleus + one bead.

3. **RT + Barcoding**
   Inside the droplet the bead’s poly-T oligo captures mRNA. Reverse transcriptase copies RNA → cDNA while adding a cell barcode and a unique molecular identifier (UMI).

4. **Library prep**
   Droplets are broken, cDNA is amplified, and sequencing adapters are added. The final library carries both the gene sequence and the barcode.

5. **Sequence & analyze**
   Millions of reads are sequenced. Software groups reads by barcode (same nucleus) and uses UMIs to count unique molecules of each gene.

Reverse Transcription *Barcoding* — the critical step

WHAT HAPPENS INSIDE THE DROPLET

Each gel bead carries thousands of identical DNA oligos that contain three important parts:

Cell Barcode

A unique 16-base sequence that identifies which nucleus the RNA came from.

UMI

A short random sequence that tags each individual mRNA molecule so we can count true molecules instead of PCR duplicates.

poly-T

Captures the poly-A tail of messenger RNA so reverse transcriptase can start copying.

Current platforms

10x Genomics Chromium

Most common droplet system. Excellent for large-scale atlases.

snRandom-seq

Uses random primers instead of poly-T. Works on old FFPE clinical samples.

Smart-seq2 / full-length

Higher sensitivity per nucleus; better for rare transcripts or isoforms.

snMultiome

Measures RNA + chromatin accessibility (ATAC) from the same nucleus.

PIP-seq & newer kits

Improved chemistry for better gene detection from nuclei.

What scientists can *discover* with snRNA-seq

01 **Brain cell atlases**

Map every type of neuron and support cell without the damage caused by trying to keep whole neurons alive.

02 **Disease vs healthy**

Compare frozen patient samples (Alzheimer’s, cancer, heart disease) with healthy tissue at single-nucleus resolution.

03 **Drug & treatment response**

See which exact cell types change after a drug or therapy — critical for precision medicine.

04 **Development & aging**

Track how gene activity shifts as tissues grow, mature, or age — one nucleus at a time.

PRACTICAL REALITY CHECK

Nuclei contain less RNA than whole cells, so gene detection is usually a bit lower. Modern analysis pipelines therefore count both exonic and intronic reads. Quality control still checks for low-quality nuclei, doublets, and ambient RNA before biological conclusions are drawn.

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).

Why location *matters*.

Single-cell and single-nucleus RNA-seq destroy tissue structure. You learn what each cell is expressing, but you lose where those cells were sitting relative to one another. Spatial transcriptomics solves this by keeping the address of every measurement.

01

Cell neighborhoods

A neuron next to a blood vessel behaves differently from the same type of neuron deep in the tissue. Position shapes identity and function.

02

Communication

Signaling molecules, immune infiltration, and developmental gradients all depend on distance and contact between cells.

03

Disease architecture

Tumors, plaques, and inflammatory foci have spatial organization that bulk or dissociated methods completely miss.

Two main *approaches*.

IMAGING-BASED

In situ hybridization & sequencing

Fluorescent probes or sequencing reactions are performed directly on the tissue section. High-resolution microscopes then read the signals at each location.

- MERFISH / seqFISH+ — highly multiplexed FISH
- Xenium (10x) — targeted in situ sequencing
- CosMx / Molecular Cartography
- Typically higher resolution, gene-panel limited

ARRAY / SEQUENCING-BASED

Barcoded capture surfaces

Tissue is placed on a slide covered with spatially barcoded spots or beads. RNA is captured, reverse-transcribed, and sequenced. Each barcode reports a location.

- Visium (10x) — widely used, ~55 µm spots
- Slide-seq / Curio — bead-based, higher res
- Stereo-seq — nanometer-scale DNA nanoballs
- Near-genome-wide coverage, lower resolution

How spatial data are *generated*.

1. **Section the tissue**
   A thin slice (usually 5–10 µm) is cut from fresh-frozen or FFPE tissue and placed on the capture slide or imaging surface.

2. **Capture or probe**
   For array methods: RNA is released and hybridizes to spatially barcoded oligos. For imaging methods: fluorescent probes or padlock probes bind target transcripts.

3. **Record location**
   Either the barcode on the capture surface encodes XY coordinates, or the microscope records the exact position of each fluorescent signal.

4. **Sequence or decode**
   Libraries are sequenced (array methods) or sequential imaging rounds decode the identity of each transcript (imaging methods).

5. **Map & analyze**
   Gene counts are aligned back to tissue coordinates. Downstream tools cluster spots/cells, find spatially variable genes, and reconstruct cell–cell interactions.

APPLICATIONS

What spatial transcriptomics *enables*.

Developmental biology

Map how cell identities and signaling gradients change across space and time in embryos and organoids.

Cancer microenvironment

Resolve tumor–immune–stromal architecture, identify niches that promote invasion or therapy resistance.

Neuroscience

Link molecular cell types to precise brain regions, circuits, and pathological lesions (e.g., plaques).

Pathology & diagnostics

Move toward spatially informed molecular diagnostics on routine clinical tissue sections.

Single Cell Sequencing

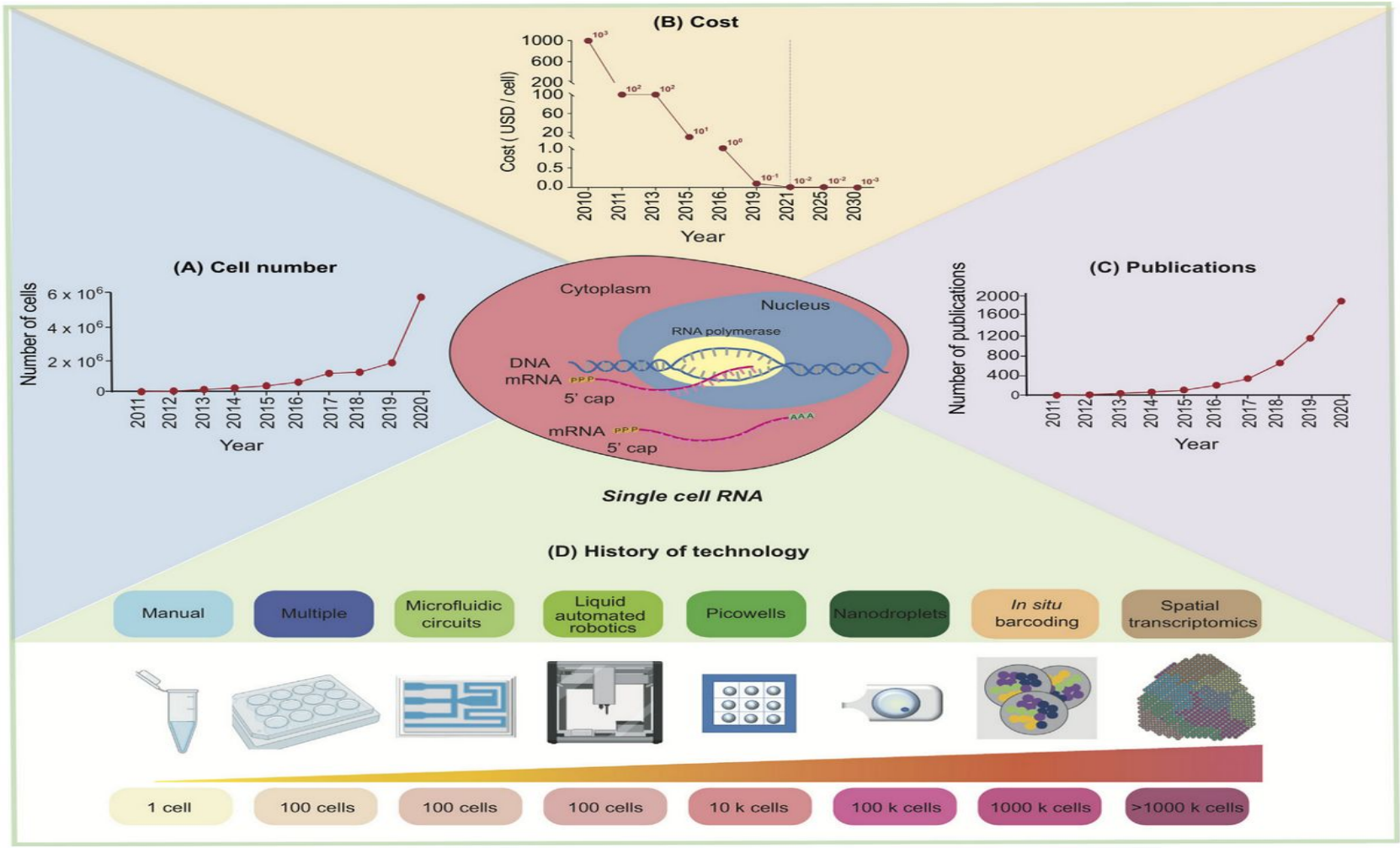
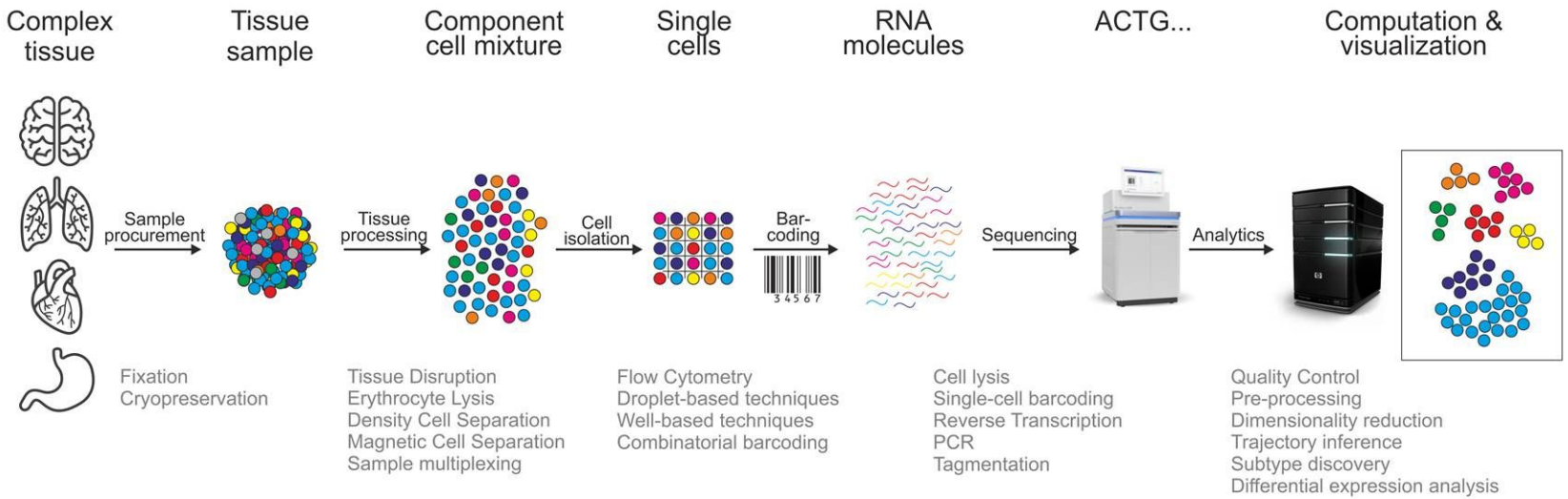


FIGURE 1 Development of single-cell RNA sequencing technology. With the technological advances in single-cell RNA sequencing (scRNA)-seq, (A) the number of analyzed cells increased, (B) the cost (in US dollar) was exponentially reduced, (C) the number of published papers increased and (D) the history 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

Jovic et al. Clin Transl Med. 2022;12:e694.

Single Cell RNA Sequencing and its main applications

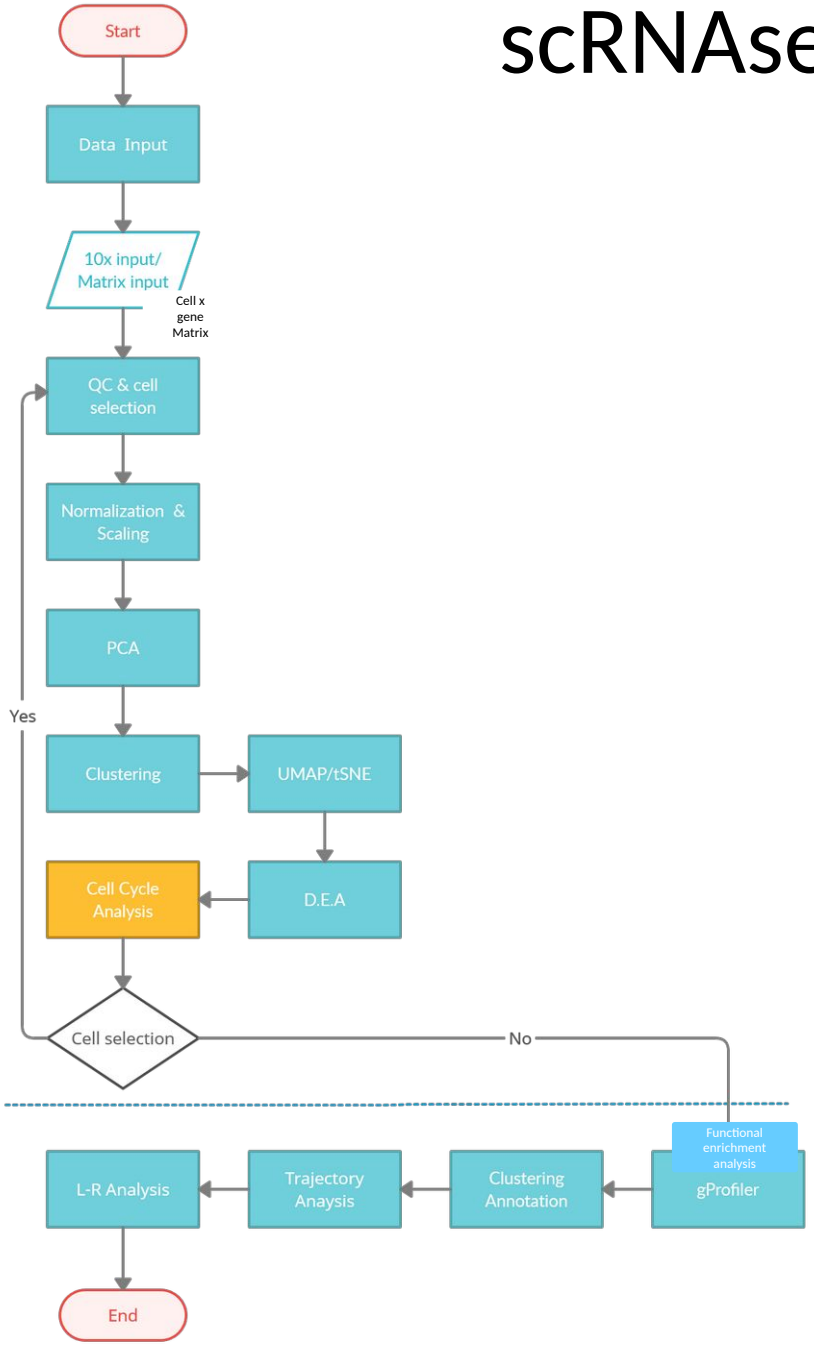
- Identification of new cell populations and subpopulations in complex tissues
- Studying gene dynamics in developmental studies
- Immune cell profiling
- Cancer research
- Personalized medicine
- Cell atlases



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

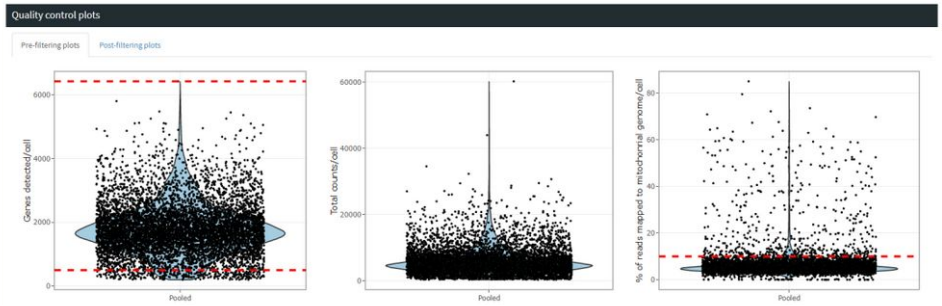
(slides by ITBI student Dimitra Panou)

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

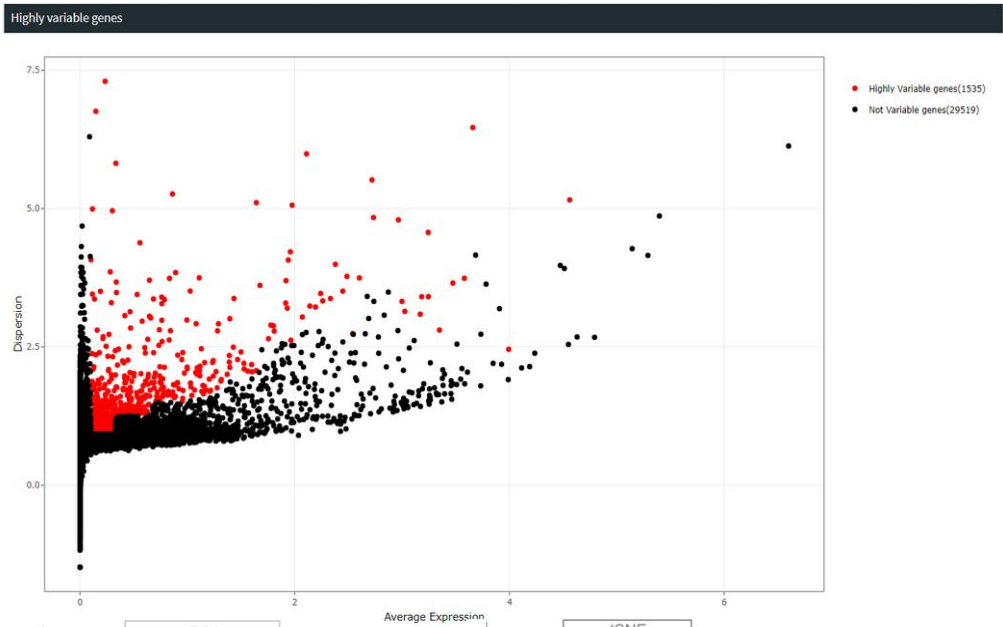


Red lines show the filtering points

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



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

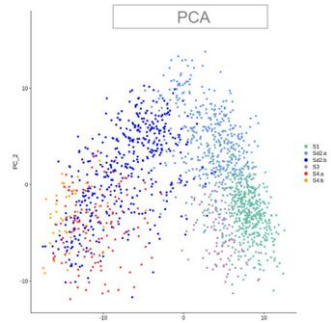
Non linear methods

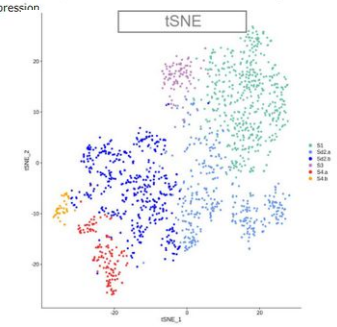
t-Distributed Stochastic Neighbor Embedding (t-SNE)

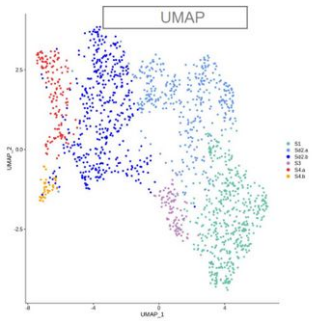
- ❑ Can capture subtle local patterns of expression in the data
- ❑ Places cells with similar local neighborhoods in high dimensional space together in low dimensional space
- ❑ It may fail to give a precise representation of clusters' size and distances

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

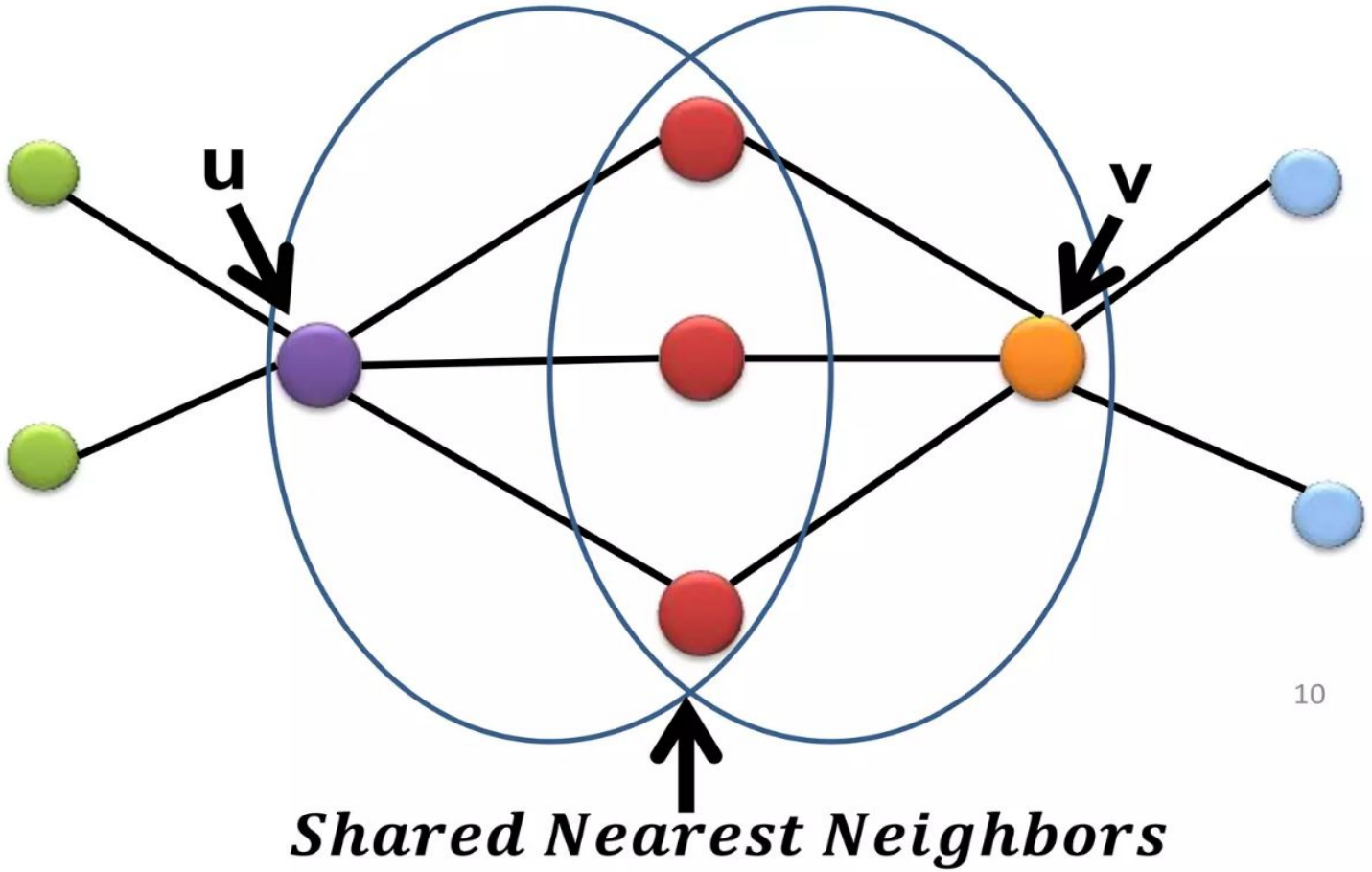






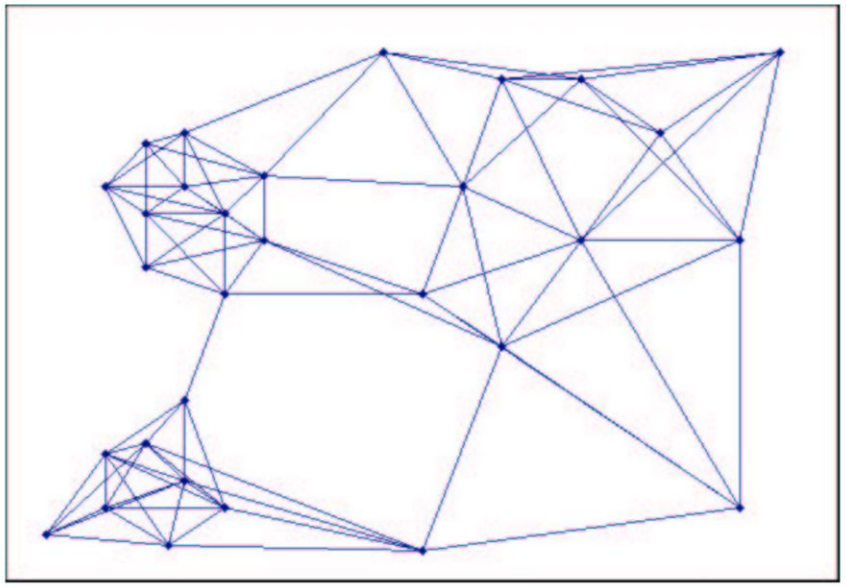
What is Shared Nearest Neighbor?

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



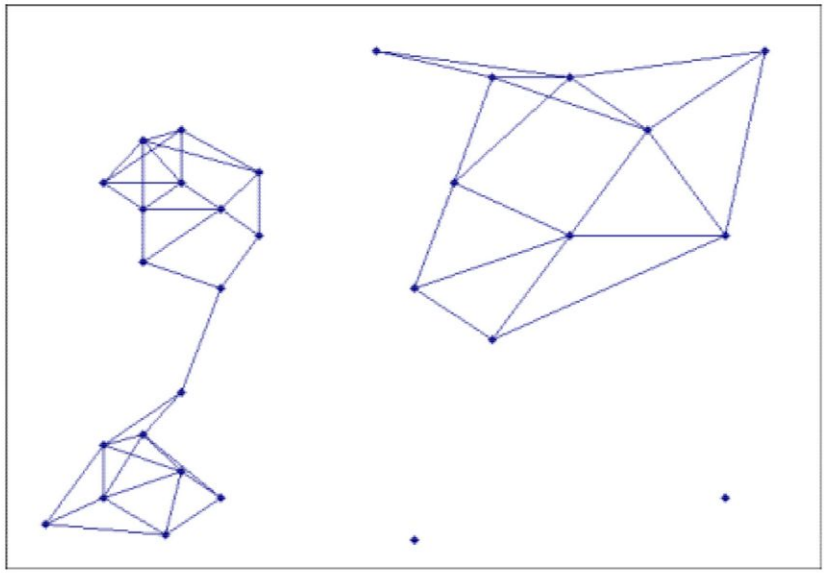
https://www.slideshare.net/slideshow/graph-clustering/250914505#7

Shared Nearest Neighbors graph



(a) Near Neighbor Graph.

k=5



(b) Unweighted Shared Nearest Neighbor.

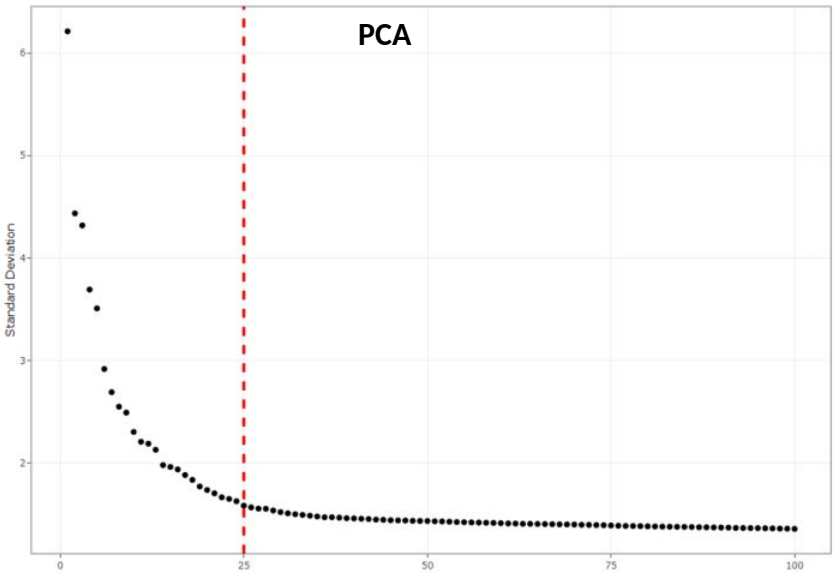
link if p1 and p2 have each other in their nearest neighbor lists

Ertöz et al. DOI: 10.1137/1.9781611972733.5

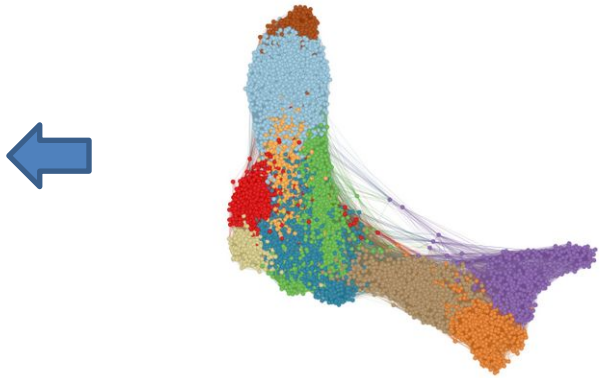
scRNAseq pipeline

4. <u>Clustering analysis</u>

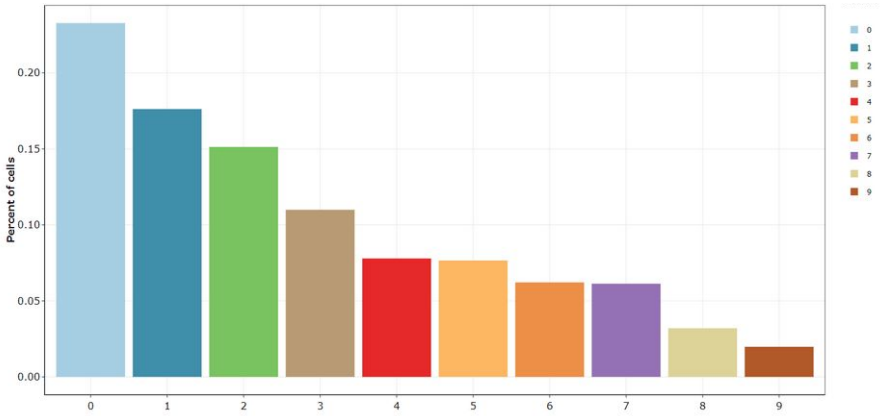
- ❑ Creation of a Shared nearest neighbor (SNN) graph
- ❑ Clusters represent
  - cell population
  - cell sub-population
  - cell state



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

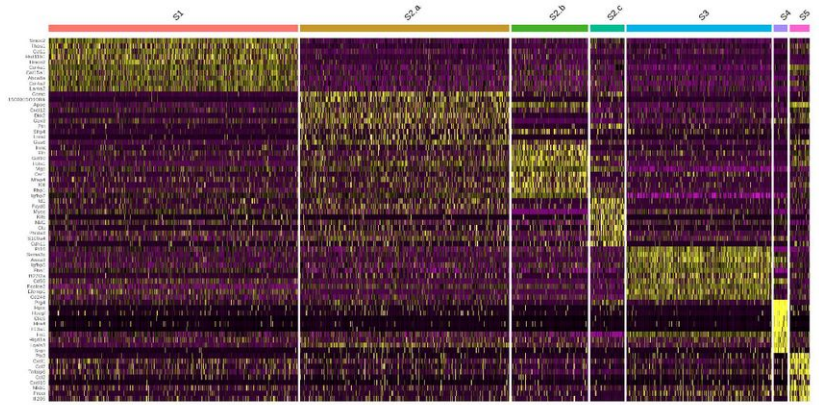
- Identify marker genes per cluster
- Those genes can distinguish one cluster from the rest
- High average expression in cells of the cluster, low in the other cells

- ❑ Wilcoxon test
- ❑ logFC >= 0.25
- ❑ Pval < 0.01
- ❑ Percentage of expression >= 25%

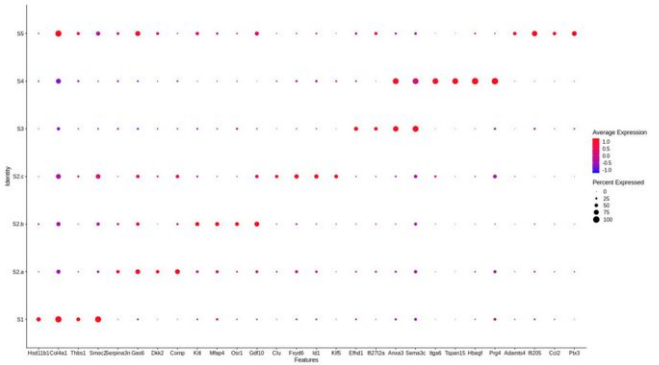
Inspection of top marker genes

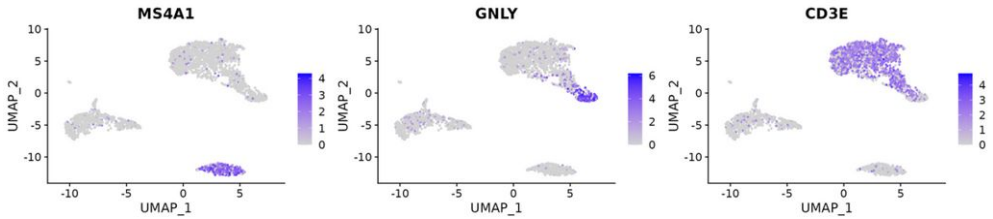
- ❑ Feature plots in UMAP space
- ❑ Color denotes normalized expression

Differential expression analysis



Marker genes for each cluster





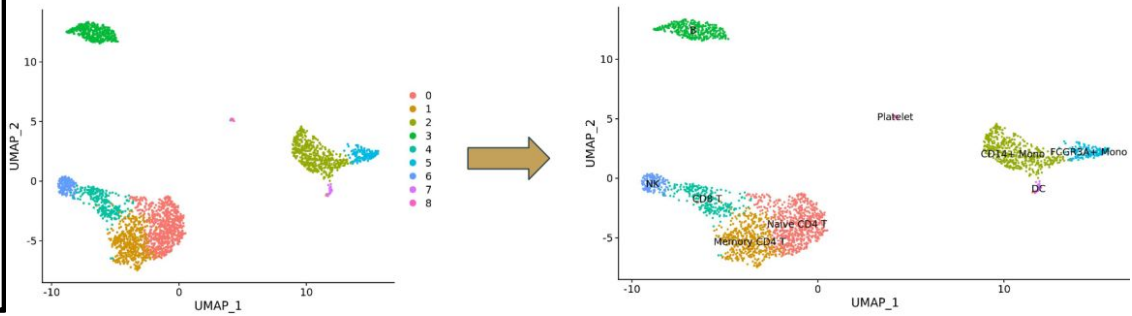
* MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data

G Finak, A McDavid, M Yajima, J Deng, V Gersuk, AK Shalek, CK Slichter et al Genome biology 16 (1), 278

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

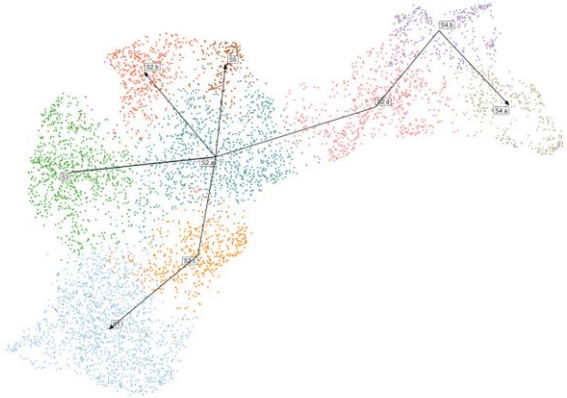


Minimum spanning tree

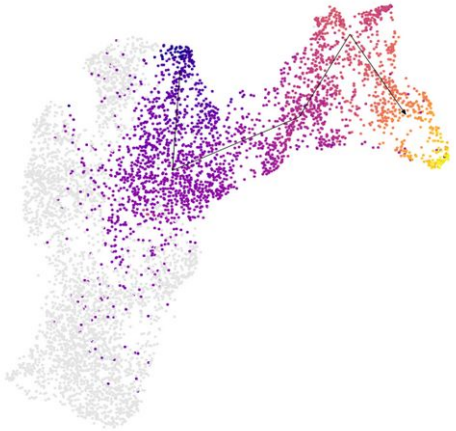
7. Trajectory-Pseudotime analysis

- ❑ Infer the lineage structure of the dataset
- ❑ Order the cells along the predicted topology

- ❑ PCs as input
- ❑ Output in UMAP plot



Pseudotime ordering



- Useful links
  - Seurat - Guided Clustering Tutorial https://satijalab.org/seurat/articles/pbmc3k_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
  - Widely used open-source tools (with source under a license)

	Name	Since	Citations*
1	RSEM	2011	~22,900
2	Tuxedo Suite	2012	~14,200
3	Salmon	2017	~13,000
4	kallisto	2016	~11,800
5	New Tuxedo Suite	2016	~6,600
6	sleuth	2017	~1,750
7	EBSeq	2015	~1,250
8	BitSeq	2012	~275

** Approximate Google Scholar / publisher citation counts as of late 2026*

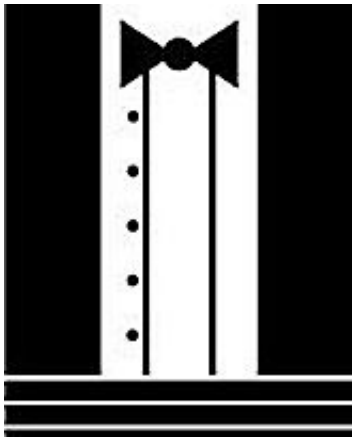
De-novo genome sequence assembly, Genome-Based and Genome-Free Transcript Reconstruction and Analysis Using RNA-Seq Data

based on material from Mathias Haimel, EBI

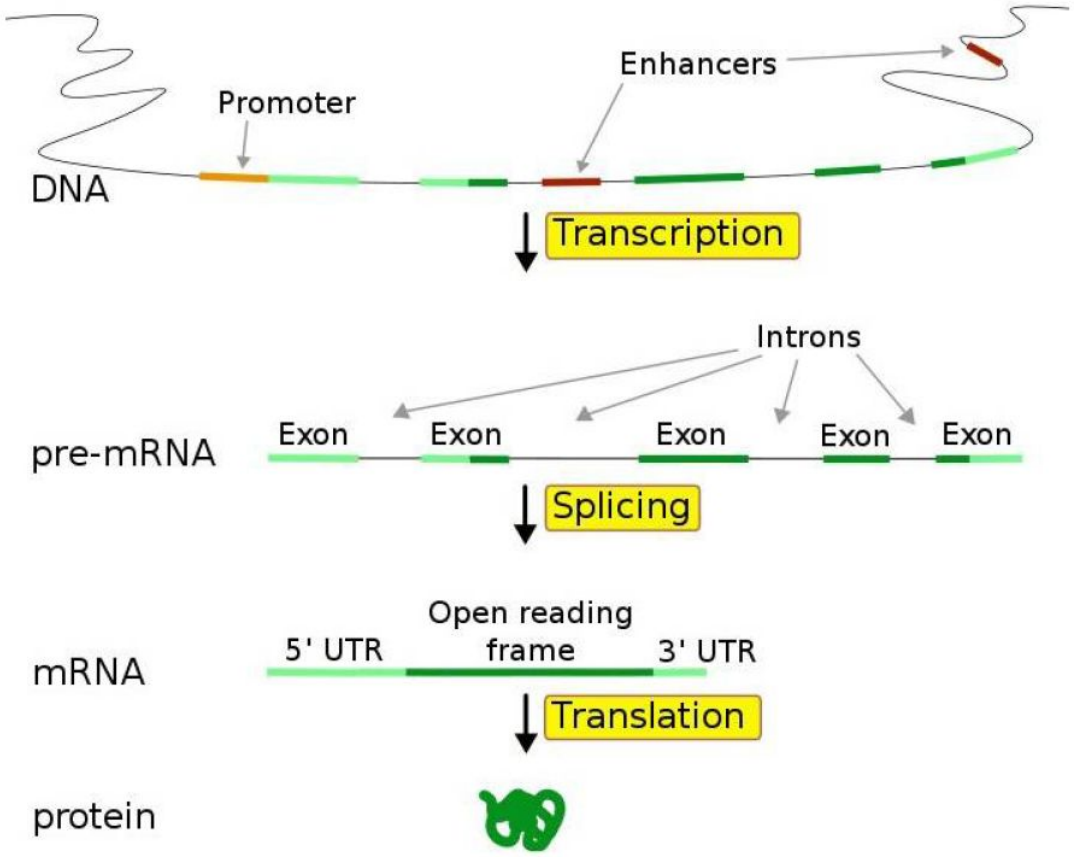
https://www.ebi.ac.uk/training/online/sites/ebi.ac.uk.training.online/files/user/18/private/velvet_1.pdf

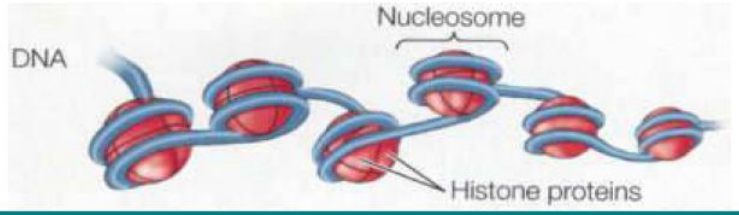
and Brian Haas

Broad Institute, modified by M. Reczko



Next Generation Sequencing



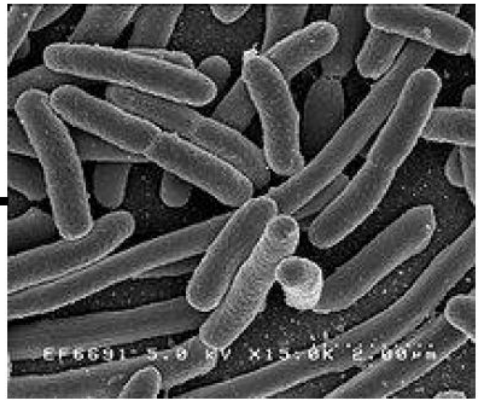


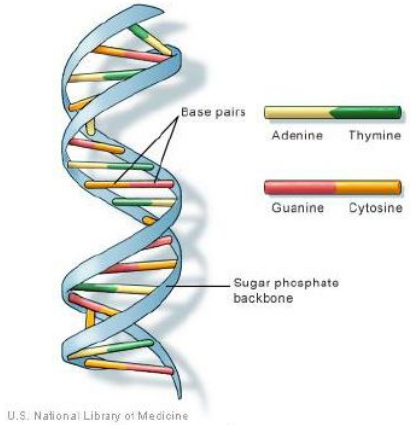
Whole Genome sequencing

RNA-Seq
Whole Transcriptome sequencing

ChIP-Seq
Chromatin Immunoprecipitation with DNA sequencing

Next Generation Sequencing

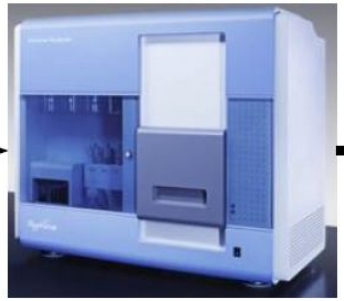




Assembly

.....

Fragments



@...../1
ATGC....
+....
UVWZ...
@...../2

De novo transcriptome assembly

No genome required

Empower studies of non-model organisms

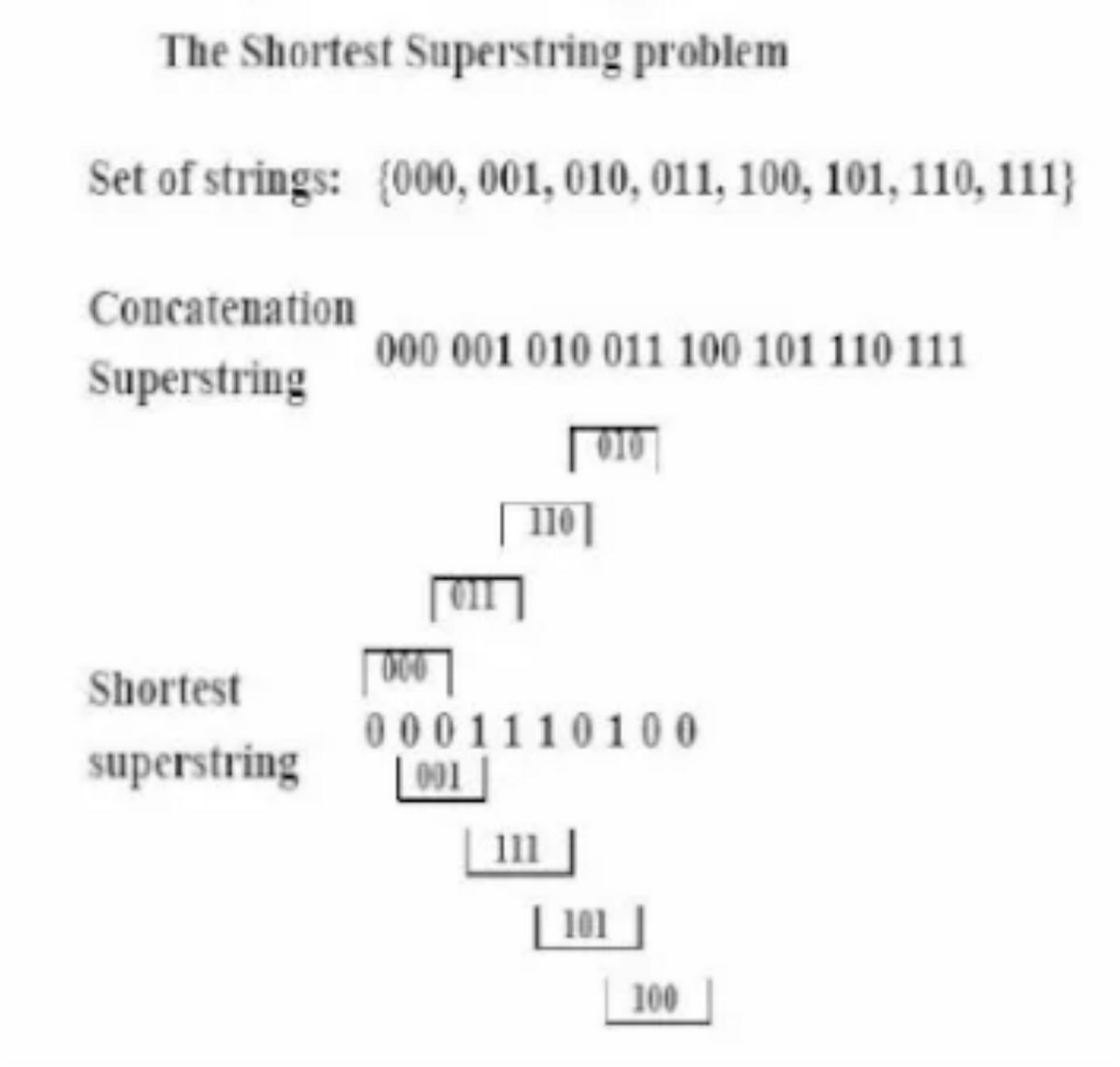
- Transcript identification
- expressed gene content
- transcript abundance
- differential expression

Shortest Superstring Problem

- Problem: Given a set of strings, find a shortest string that contains all of them
- Input: Strings $s_1, s_2, \ldots, s_n$
- Output: A string s that contains all strings $s_1, s_2, \ldots, s_n$ 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



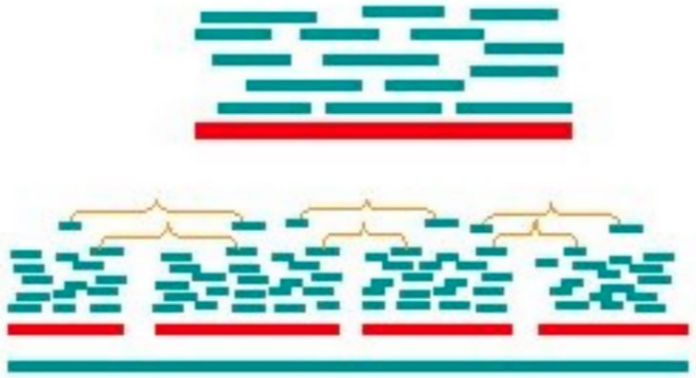
http://feedthedatamonster.com/home/2014/12/24/three-billion-puzzle-pieces-insights-into-the-human-genome-project

Overlap-Layout-Consensus

Assemblers: ARACHNE, PHRAP, CAP, TIGR, CELERA

Overlap: find potentially overlapping reads

Layout: merge reads into contigs and contigs into supercontigs



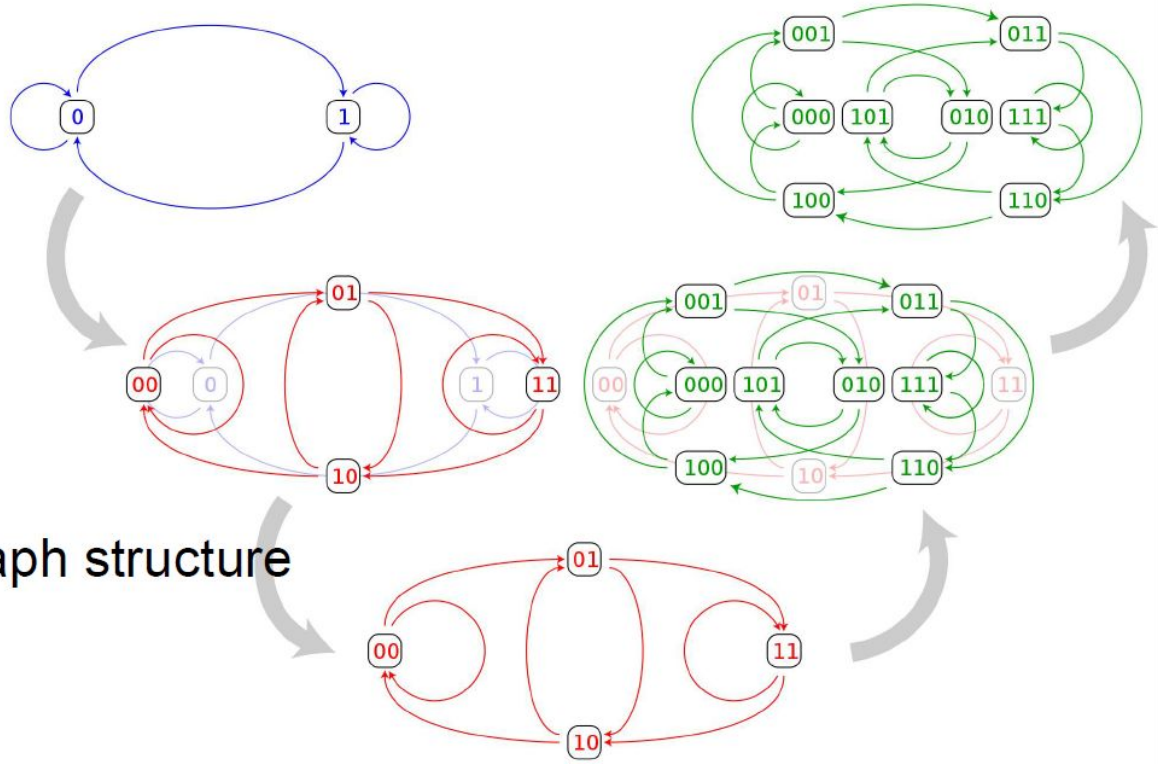
Consensus: derive the DNA sequence and correct read errors

..ACGATTACAATAGGTT..

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

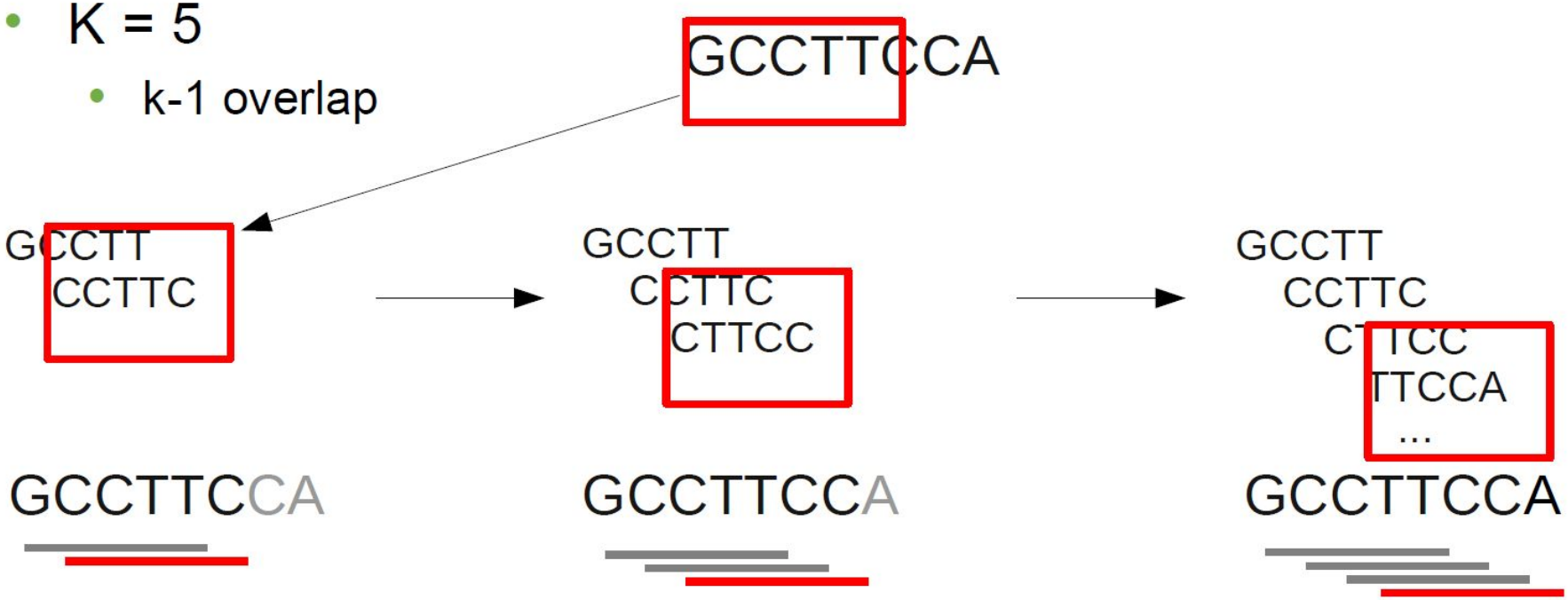
De Bruijn graph

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



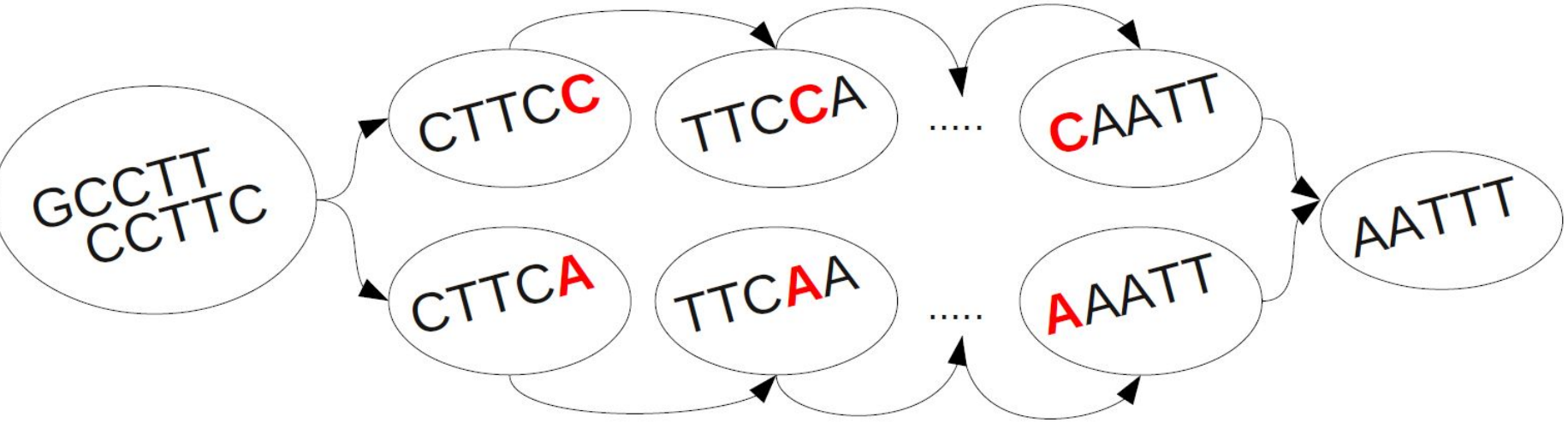
De Bruijn graph (Velvet)

- 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)

GCCTTCCAATTT
GCCTTCAAATTT



Example

TAGTCGAGGCTTTAGATCCGATGAGGCTTTAGAGACAG

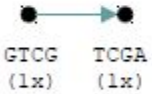
```
 AGTCGAG CTTTAGA  CGATGAG CTTTAGA
  GTCGAGG  TTAGATC  ATGAGGC    GAGACAG
     GAGGCTC   ATCCGAT AGGCTTT GAGACAG
 AGTCGAG    TAGATCC ATGAGGC  TAGAGAA
TAGTCGA  CTTTAGA CCGATGA    TTAGAGA
    CGAGGCT  AGATCCG TGAGGCT  AGAGACA
TAGTCGA GCTTTAG TCCGATG  GCTCTAG
   TCGACGC    GATCCGA GAGGCTT AGAGACA
TAGTCGA    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

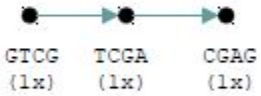
Example

Read: GTCGAGG



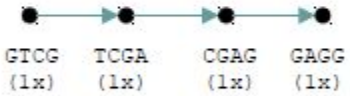
Example

Read: GTCGAGG



Example

Read: GTCGAGG



Example

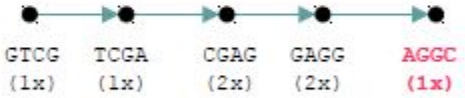
New read: CGAGGCT

Example

Read: CGAGGCT

Example

Read: CGAGGCT



Example

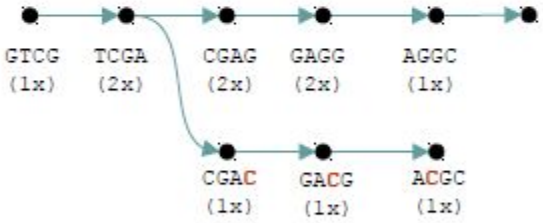
Read: CGAGGCT

Example

New read: TCGACGC

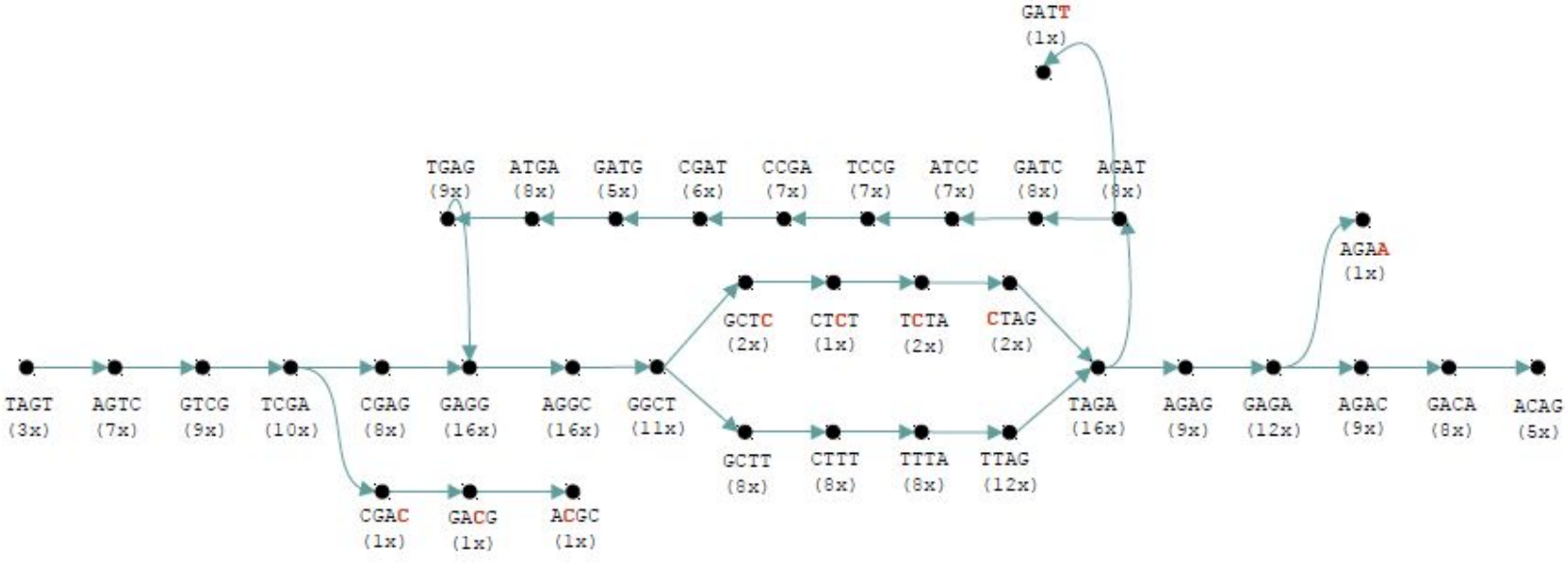
Example

Read: TCGACGC



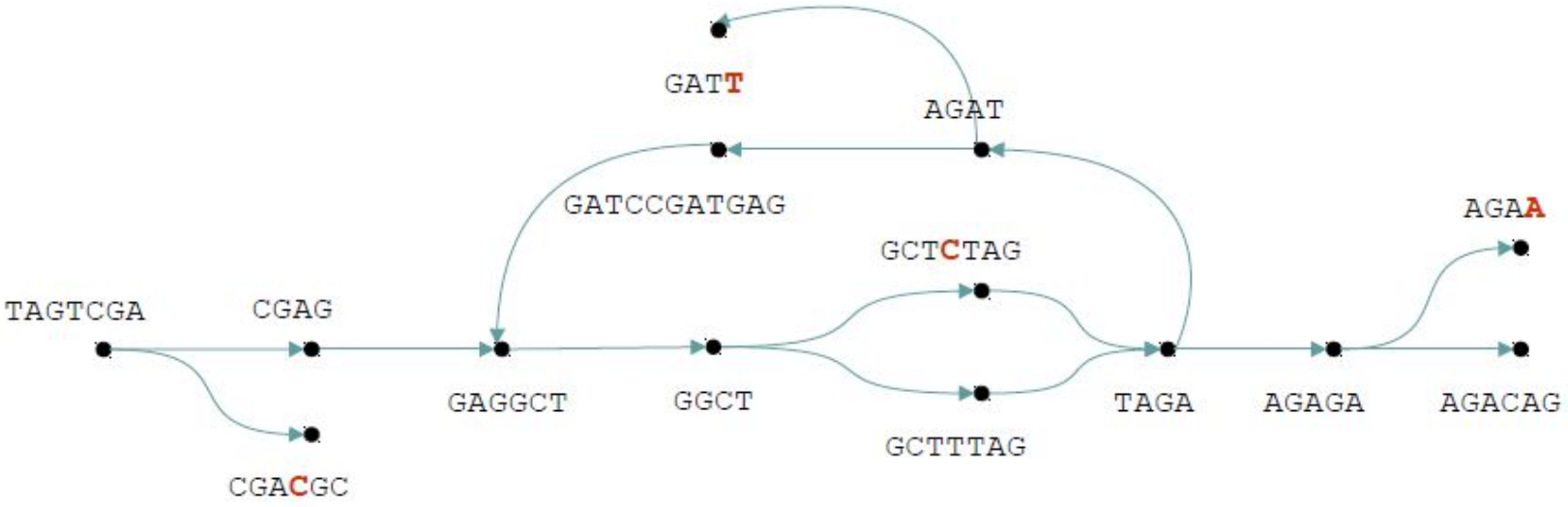
Example

etc…


GATT
(1x)
TGAG
(9x)
ATGA
(8x)
GATG
(5x)
CGAT
(6x)
CCGA
(7x)
TCCG
(7x)
ATCC
(7x)
GATC
(8x)
AGAT
(8x)
AGAA
(1x)
GCTC
(2x)
CTCT
(1x)
TCTA
(2x)
CTAG
(2x)
TAGT
(3x)
AGTC
(7x)
GTCG
(9x)
TCGA
(10x)
CGAG
(8x)
GAGG
(16x)
AGGC
(16x)
GGCT
(11x)
TAGA
(16x)
AGAG
(9x)
GAGA
(12x)
AGAC
(9x)
GACA
(8x)
ACAG
(5x)
GCTT
(8x)
CTTT
(8x)
TTTA
(8x)
TTAG
(12x)
CGAC
(1x)
GACG
(1x)
ACGC
(1x)

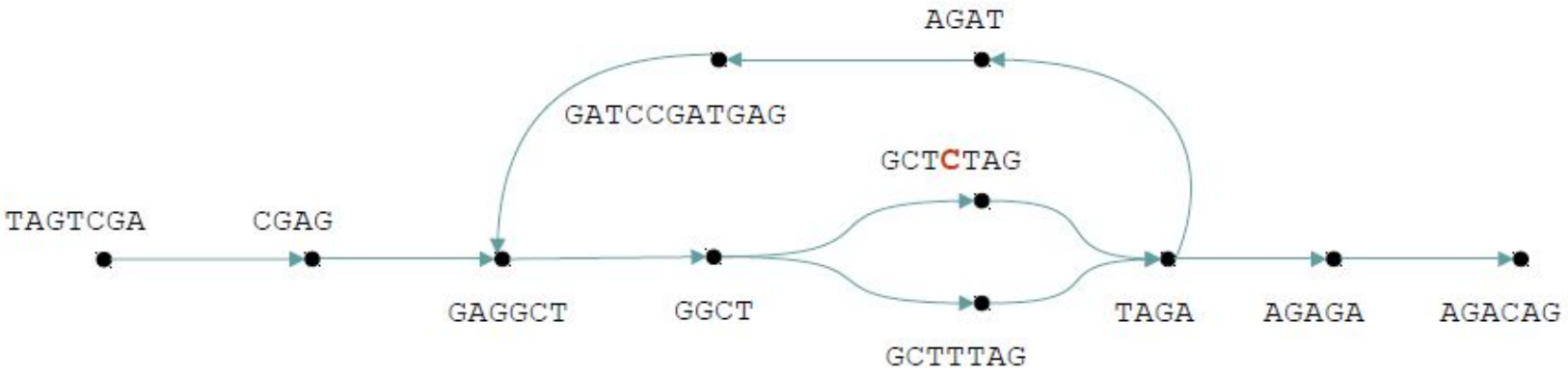
Example

After simplification…



Example

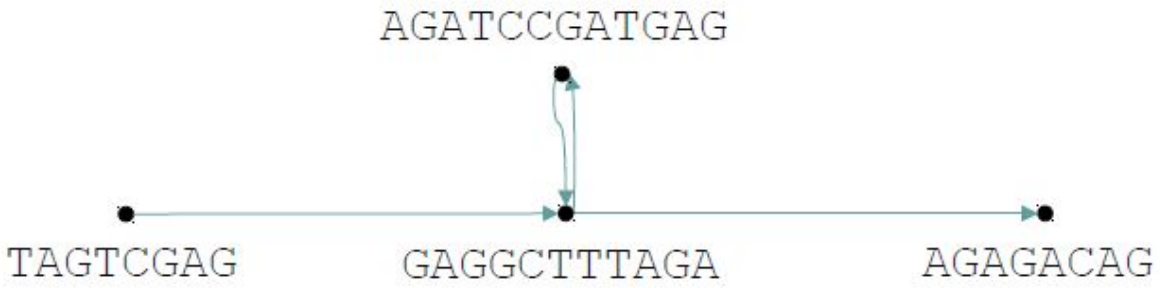
Tips removed…



Example

TAGTCGAGGCTTTAGATCCGATGAGGCTTTAGAGACAG

Final simplification…



One possible walk through the graph ...

TAGTCGAG
GAGGCTTTAGA
AGATCCGATGAG
GAGGCTTTAGA
AGAGACAG

2.6 Assembly evaluation

During the assembly optimization will be generated several assemblies. The m
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**

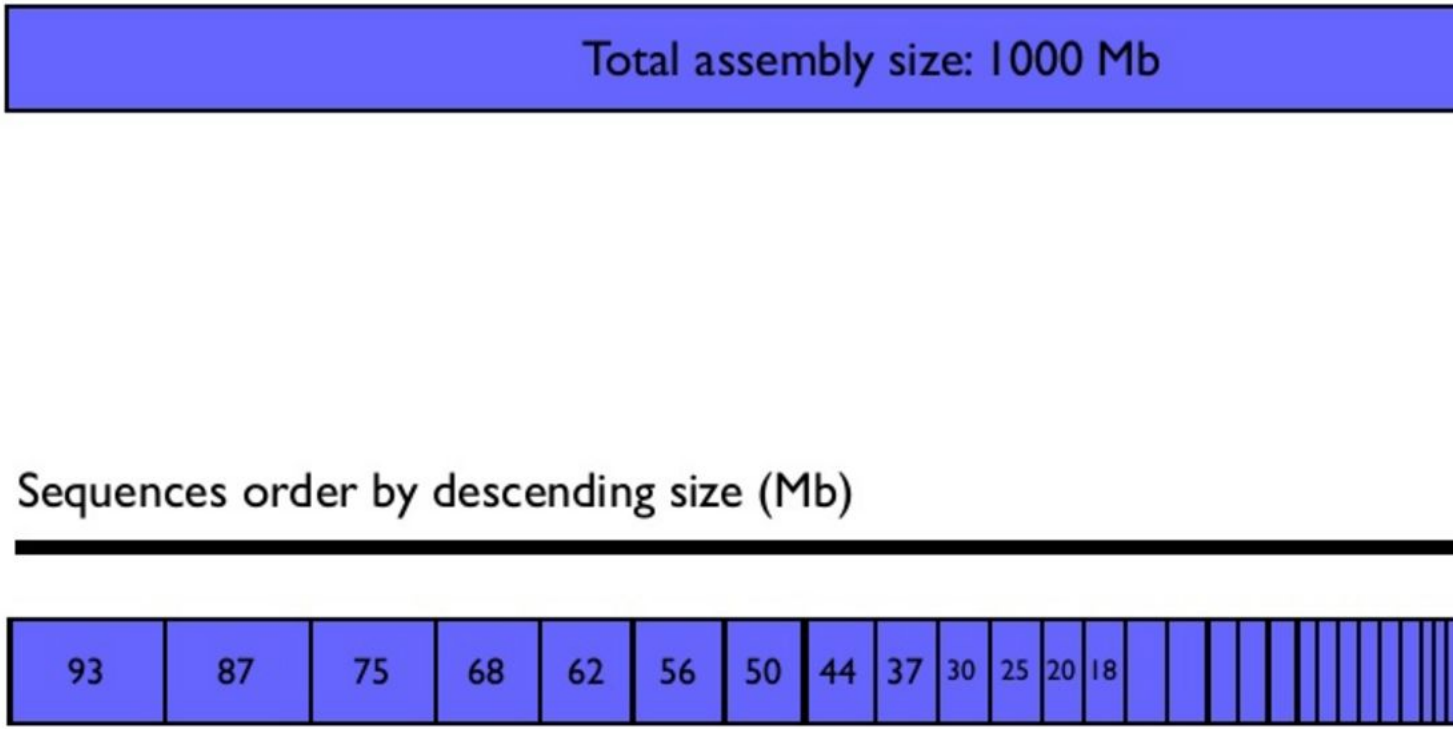
4. **Average scaffold/contig size**

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 smal

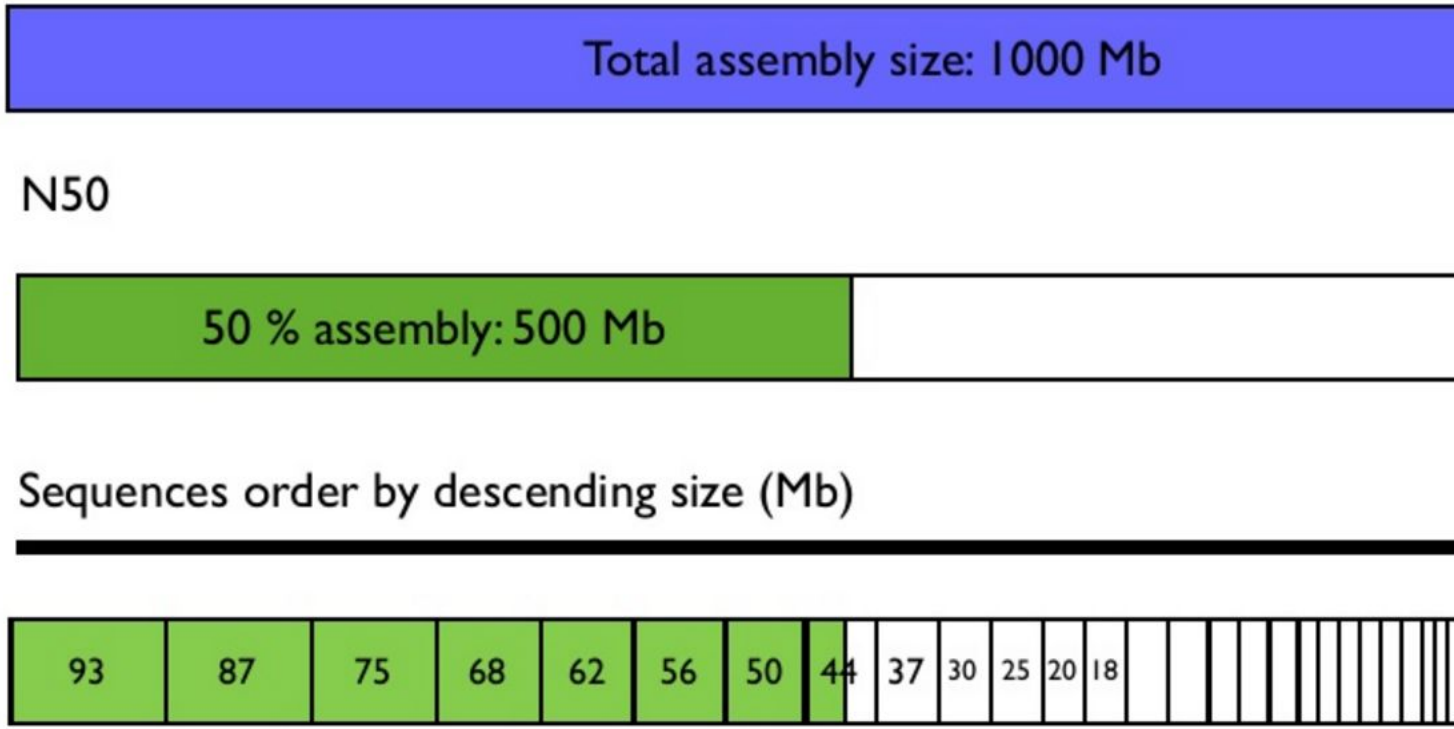
2.6 Assembly evaluation

N50/L50



2.6 Assembly evaluation

N50/L50

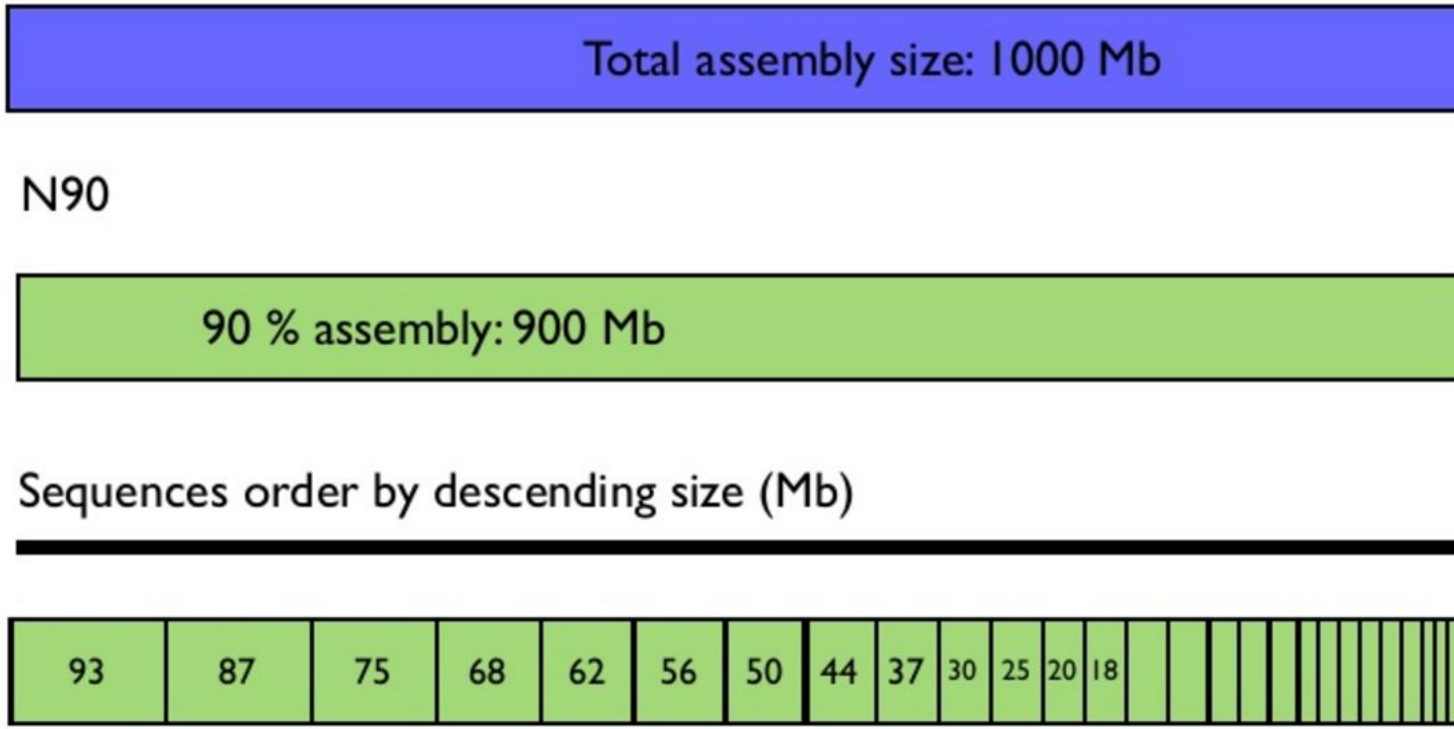


N50 = 7 sequences

L50 = 50 Mb

2.6 Assembly evaluation

N90/L90



N90 = 29 sequences

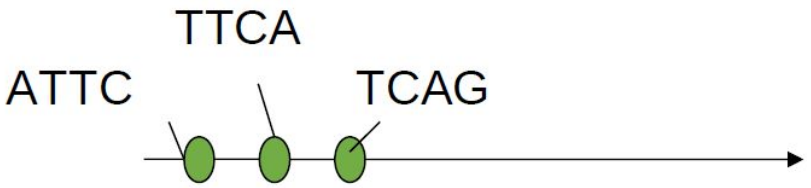
L90 = 12.5 Mb

De Bruijn graph biology extensions (Velvet)

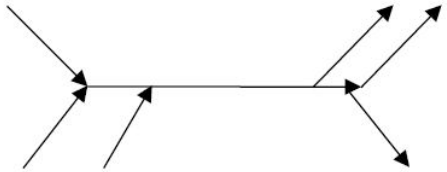
- Handling of reverse strand
  - DNA is read in two directions
  - Paired-end data
- Handling small differences, which are “uninteresting”
  - Errors in sequencing technology
- Memory
  - regularly use 80, 100GB real memory
  - easily get to 1TB real memory requirements

De Bruijn graph representations (Velvet)

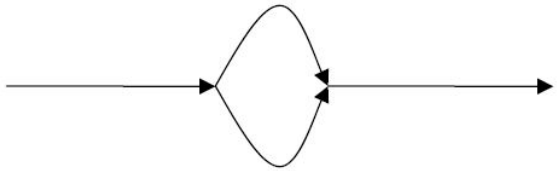
Error free, no repeat,
no polymorphism



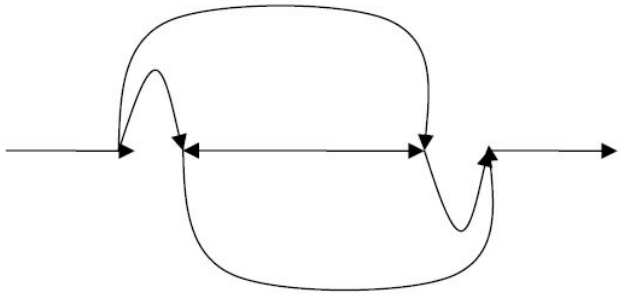
Repeat > kmer length



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

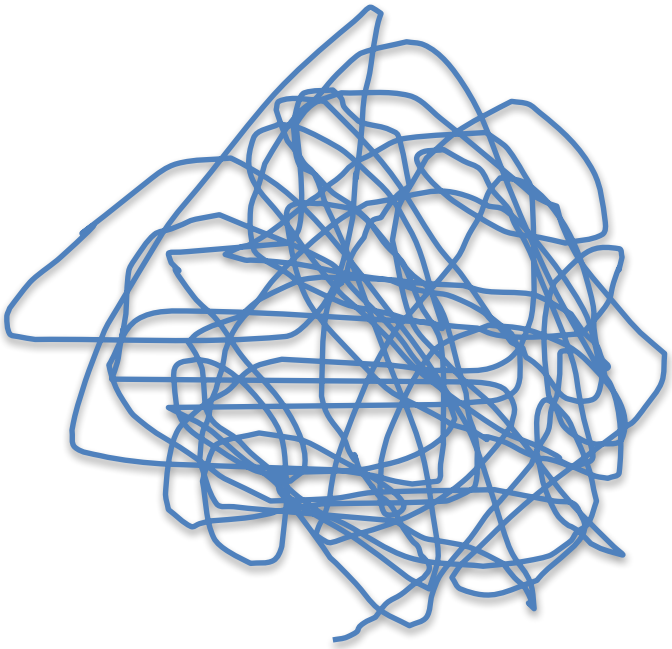
Transcriptome Assembly

- 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 Mlecular Biology and Genomics

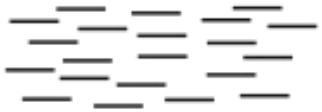
- K. Mathiopoulos, E. Sagri

ALEXANDER FLEMING
Biomedical Sciences Research Center

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

Trinity – How it works:

RNA-Seq reads → **Linear contigs** → **de-Bruijn graphs** → **Transcripts + Isoforms**



>a121:len=5845
>a122:len=2560
>a123:len=4443
>a124:len=48
>a125:len=8876
>a126:len=66

...CTTCGCAA...TGATCGGAT...
...ATTCGCAA...TCATCGGAT...

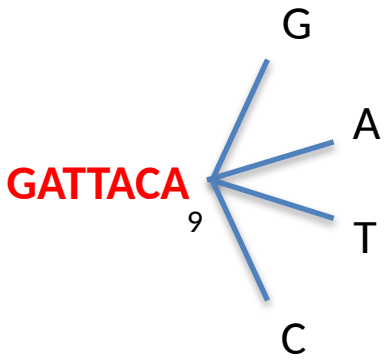
Thousands of disjoint graphs

Inchworm Algorithm

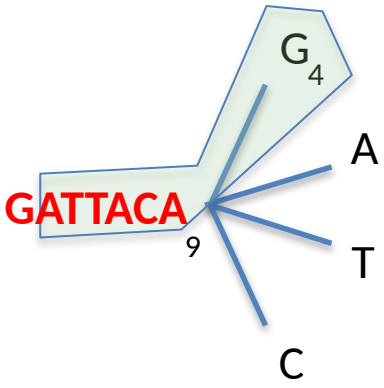
Decompose all reads into overlapping Kmers (25-mers)

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

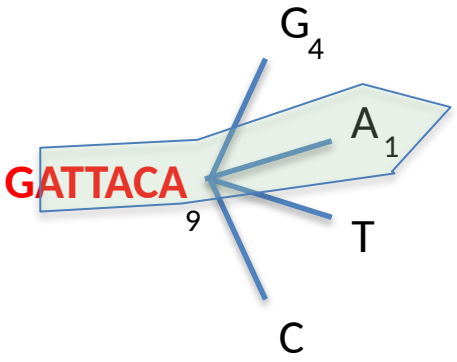
Extend kmer at 3’ end, guided by coverage.



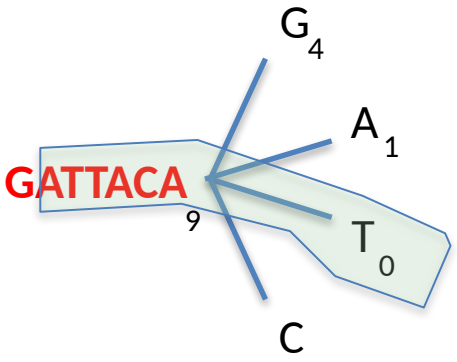
Inchworm Algorithm



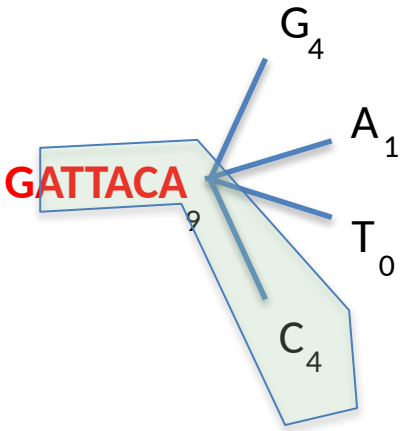
Inchworm Algorithm



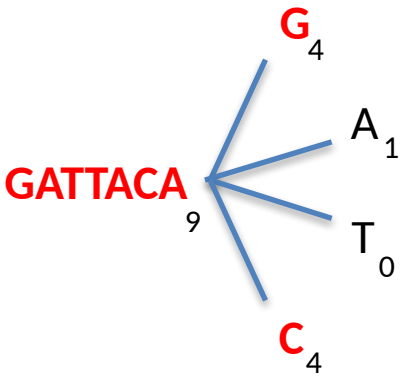
Inchworm Algorithm



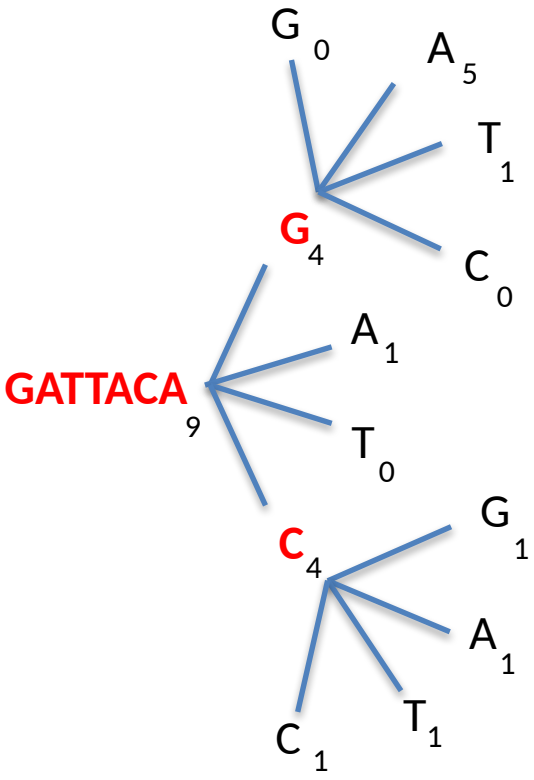
Inchworm Algorithm



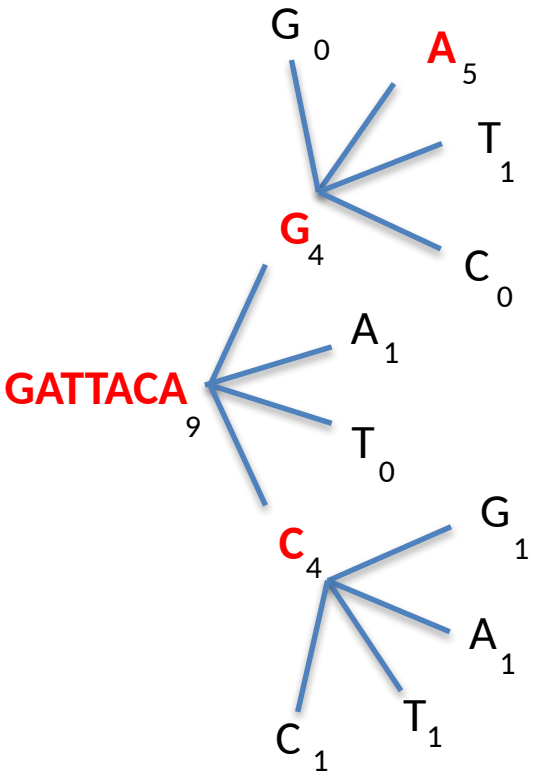
Inchworm Algorithm



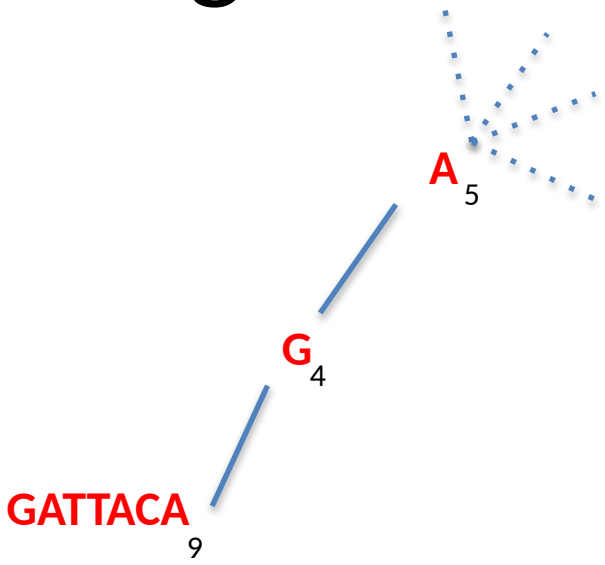
Inchworm Algorithm



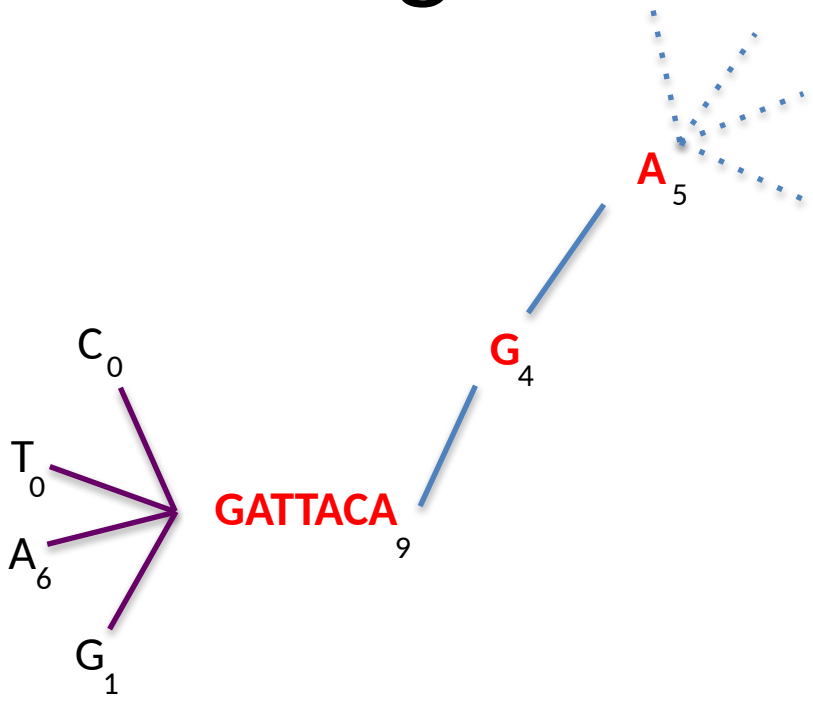
Inchworm Algorithm



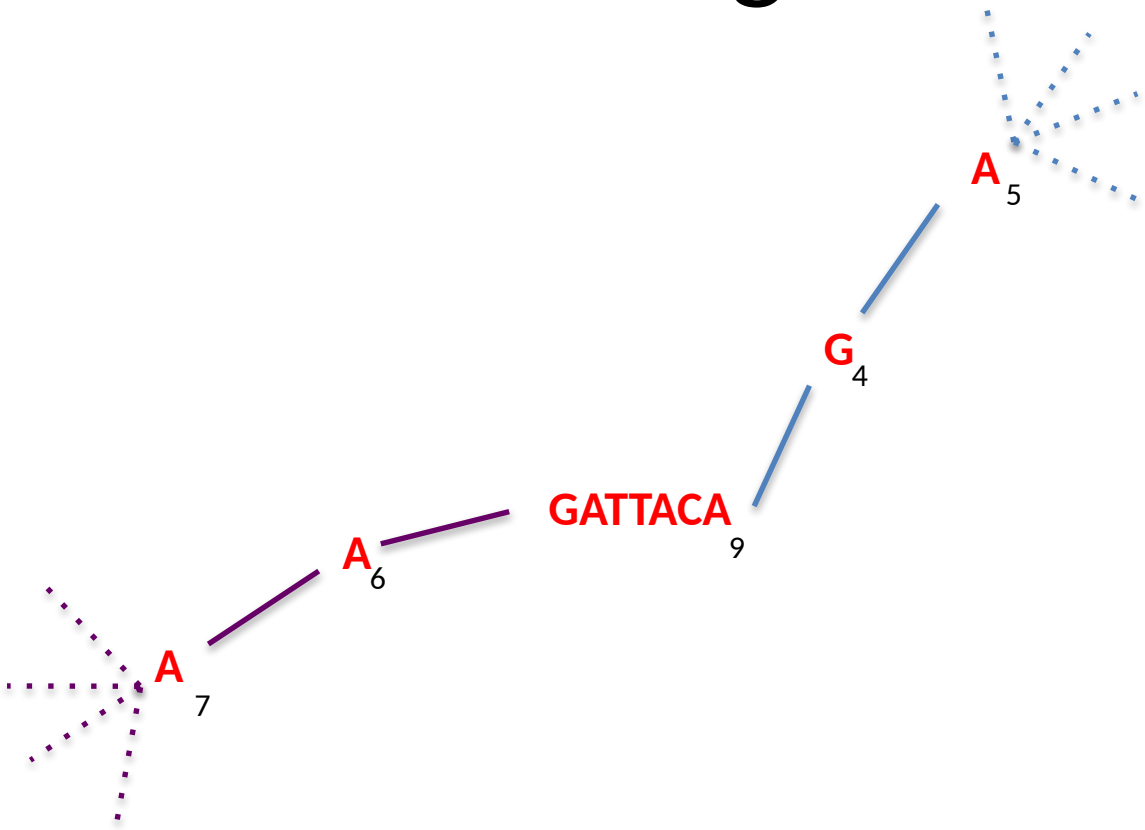
Inchworm Algorithm



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

Isoform A

Isoform B

Inchworm Contigs from Alt-Spliced Transcripts

Expressed isoforms

Isoform A

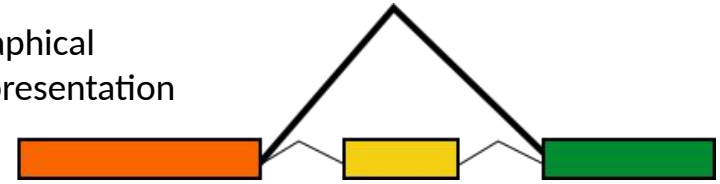
Isoform B

Expression

(low)

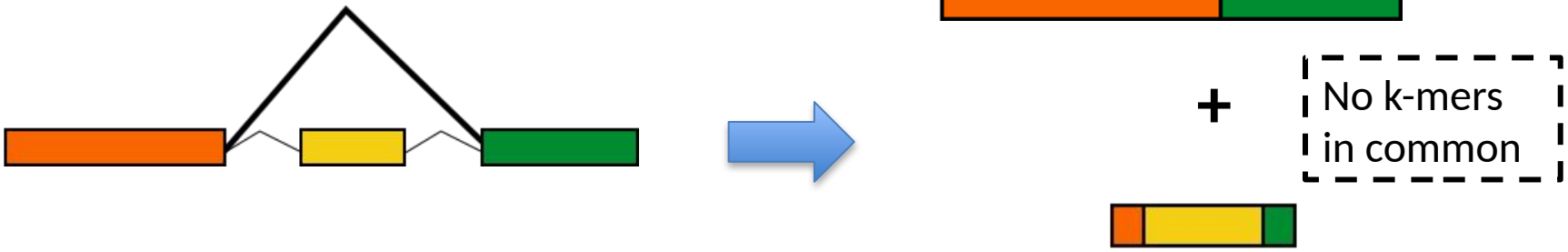
(high)

Graphical representation

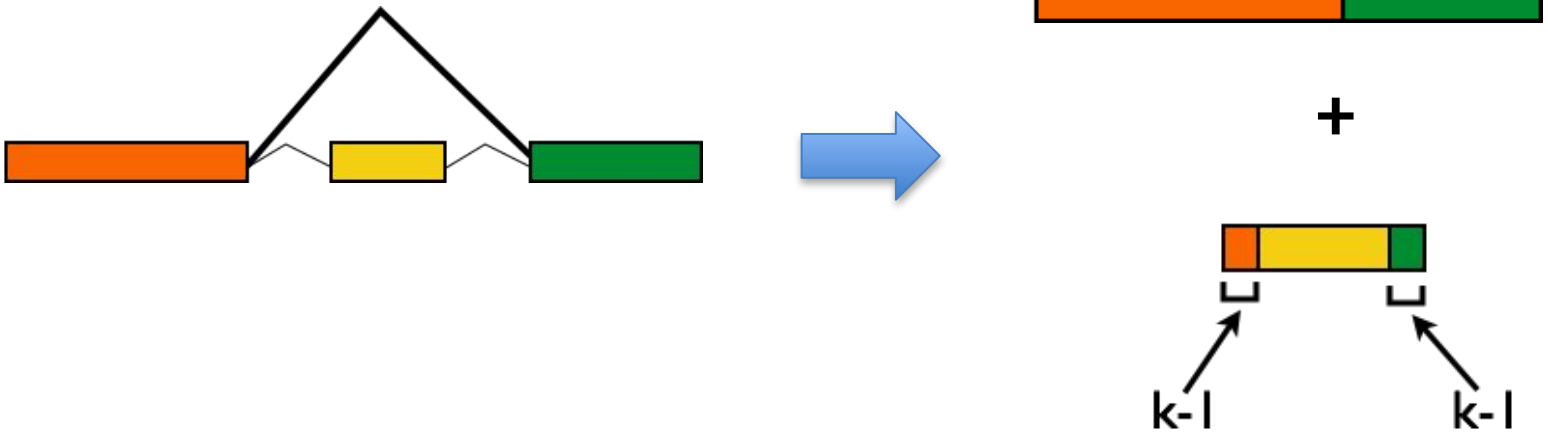


Inchworm Contigs from Alt-Spliced Transcripts

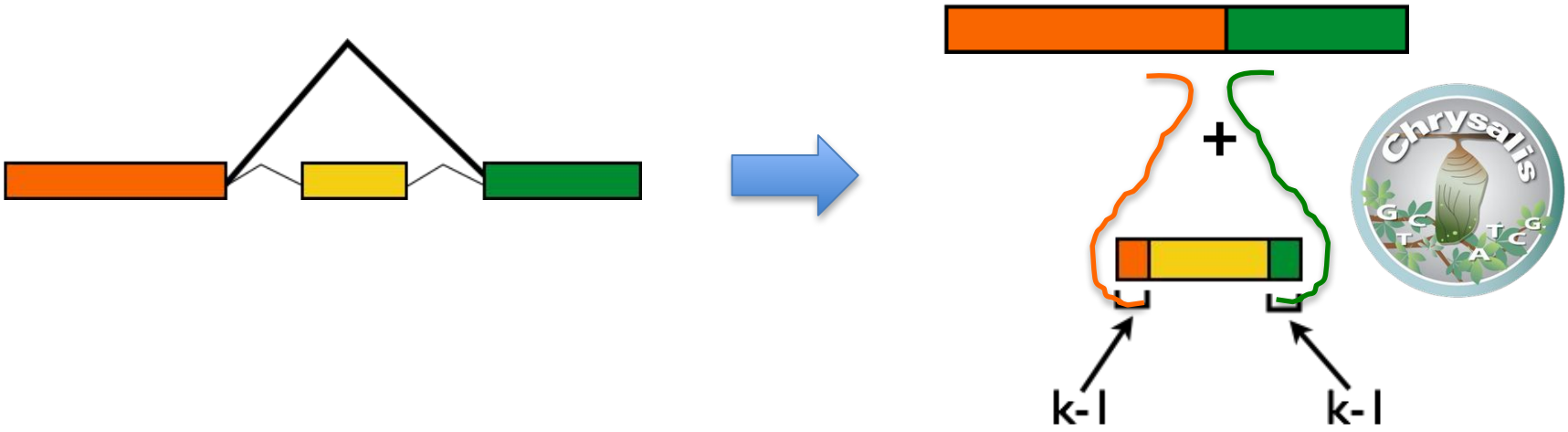
Inchworm Contigs from Alt-Spliced Transcripts



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=2560

>a123:len=4443

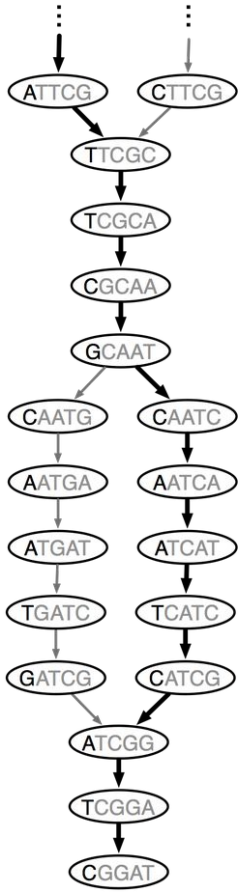
>a124:len=48

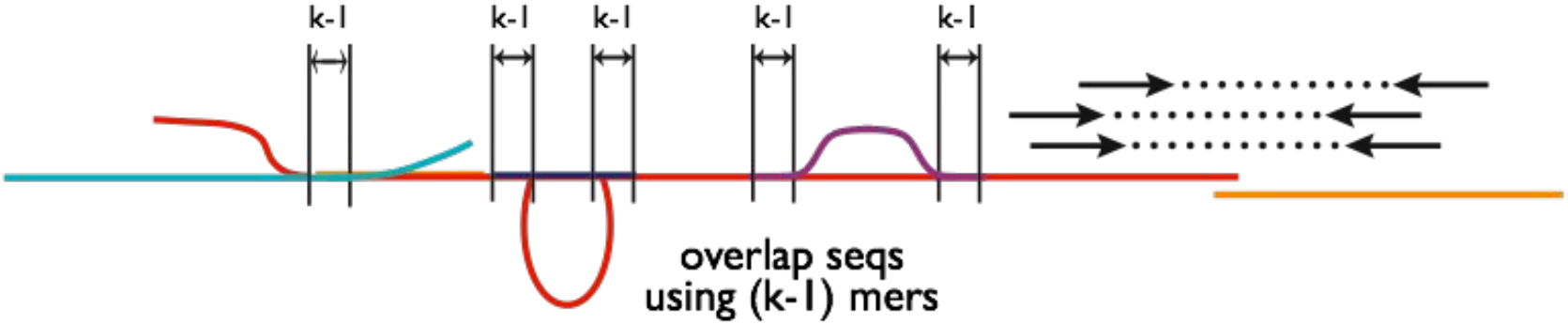
>a125:len=8876

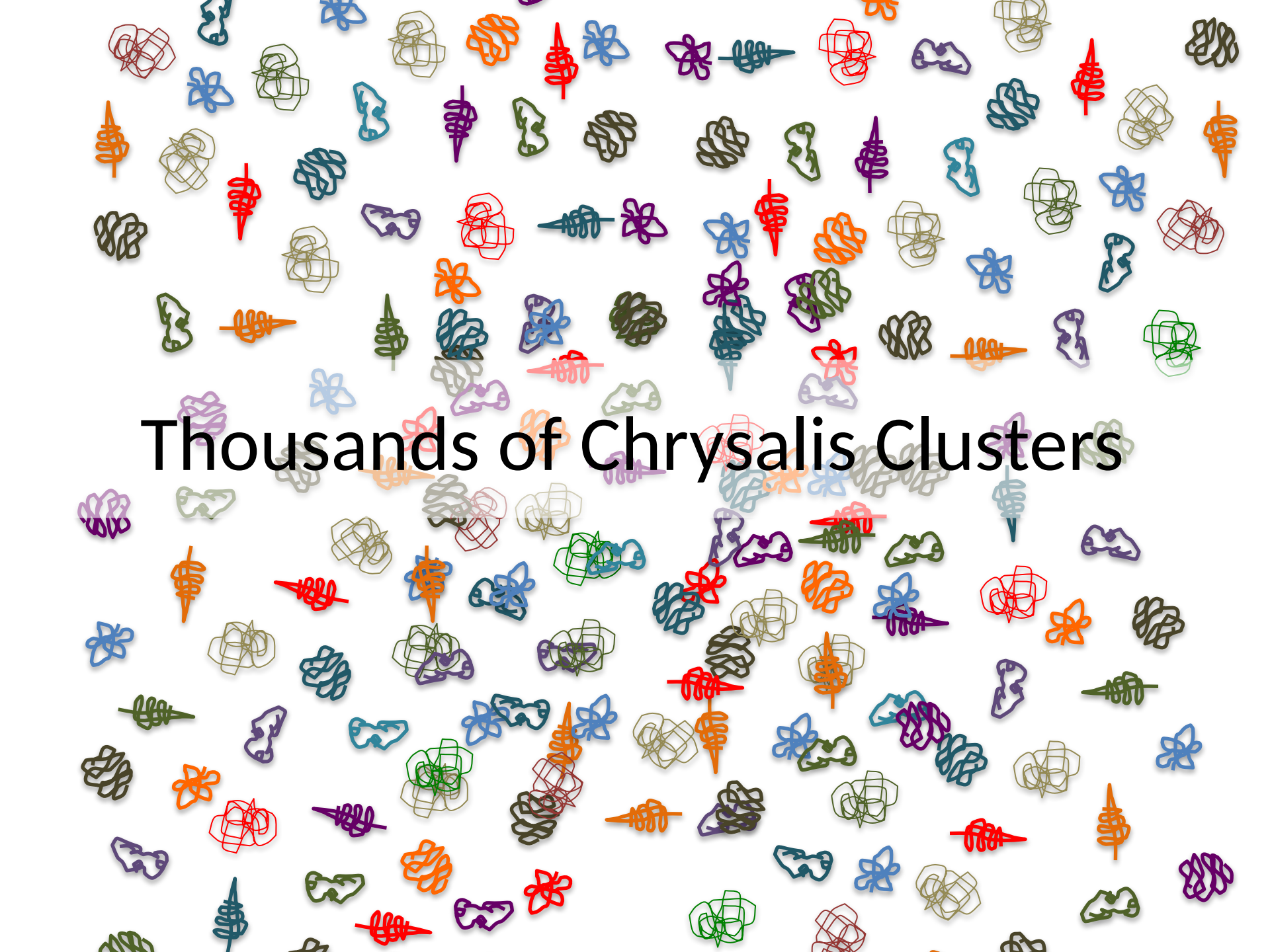
>a126:len=68

Integrate isoforms via k-1 overlaps

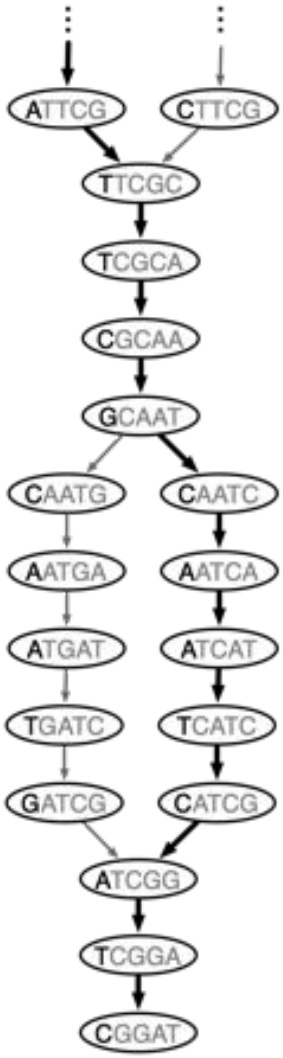
Build de Bruijn Graphs (ideally, one per gene)





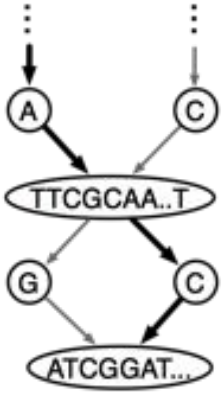

Thousands of Chrysalis Clusters

Butterfly



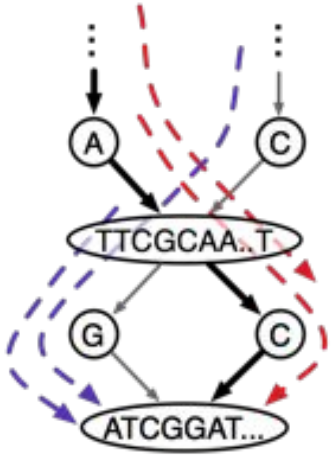
de Bruijn graph

compacting



compact graph

finding paths



compact graph with reads

extracting sequences

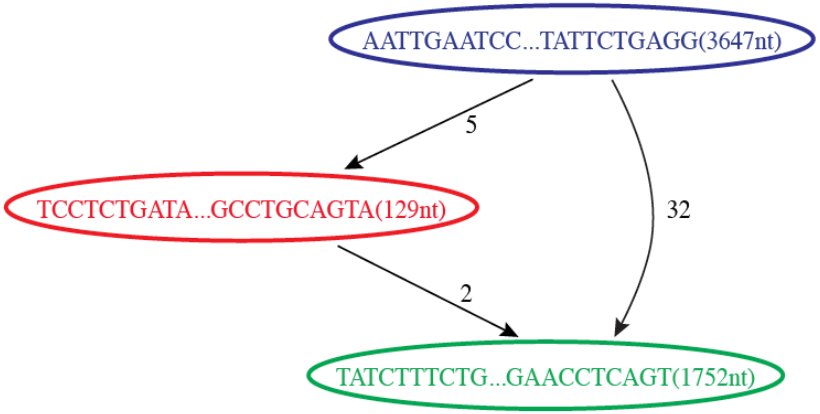
..CTTCGCAA..TGATCGGAT...

..ATTCGCAA..TCATCGGAT...

sequences (isoforms and paralogs)

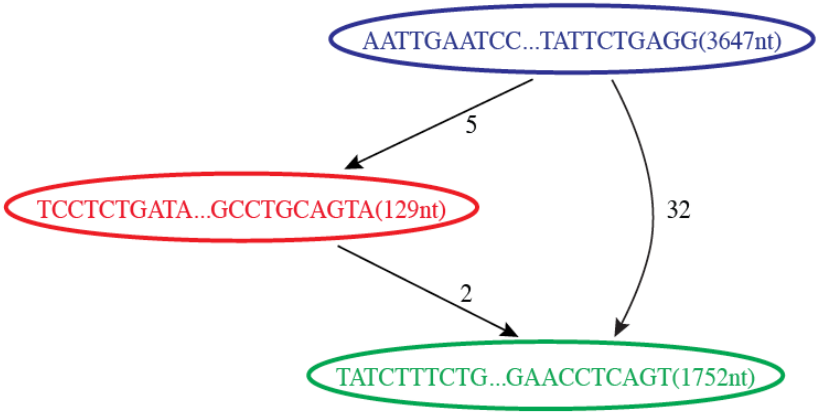
Butterfly Example 1:
Reconstruction of Alternatively Spliced Transcripts

Butterfly's Compacted Sequence Graph



Reconstruction of Alternatively Spliced Transcripts

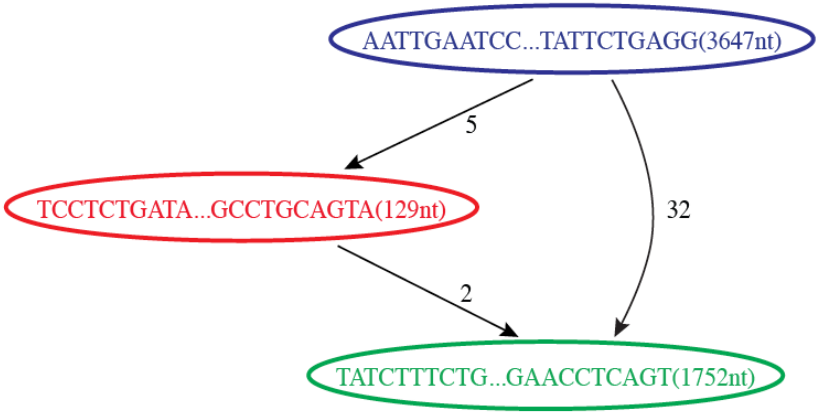
Butterfly's Compacted Sequence Graph



Reconstructed Transcripts

Reconstruction of Alternatively Spliced Transcripts

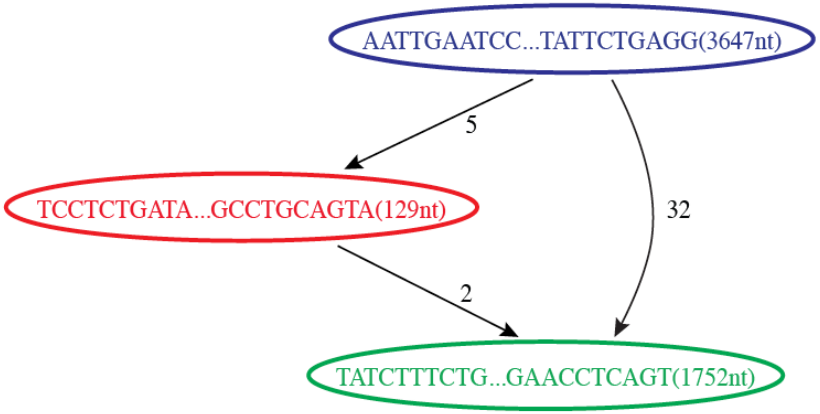
Butterfly's Compacted Sequence Graph



Reconstructed Transcripts

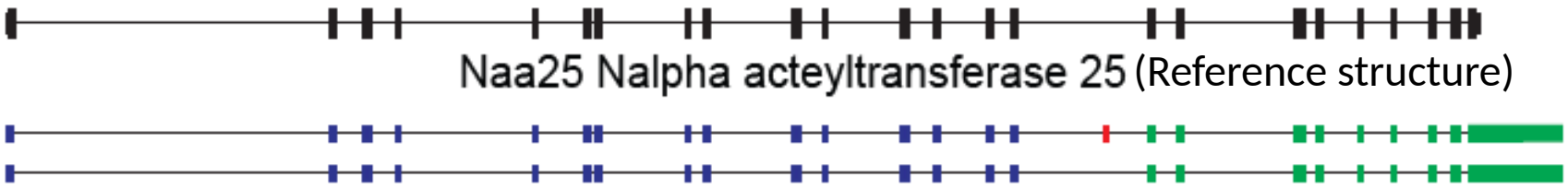
Reconstruction of Alternatively Spliced Transcripts

Butterfly's Compacted Sequence Graph

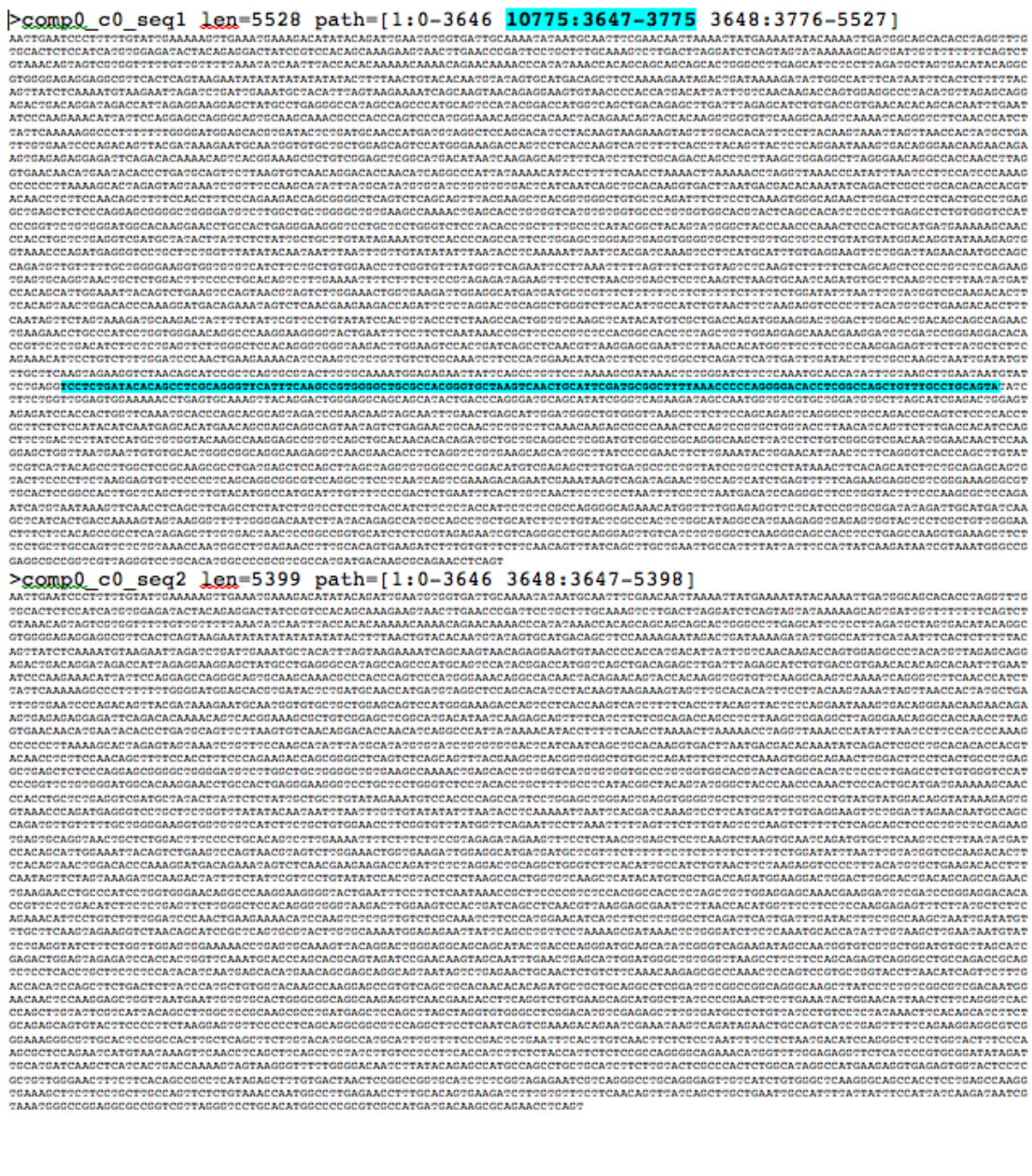


Reconstructed Transcripts

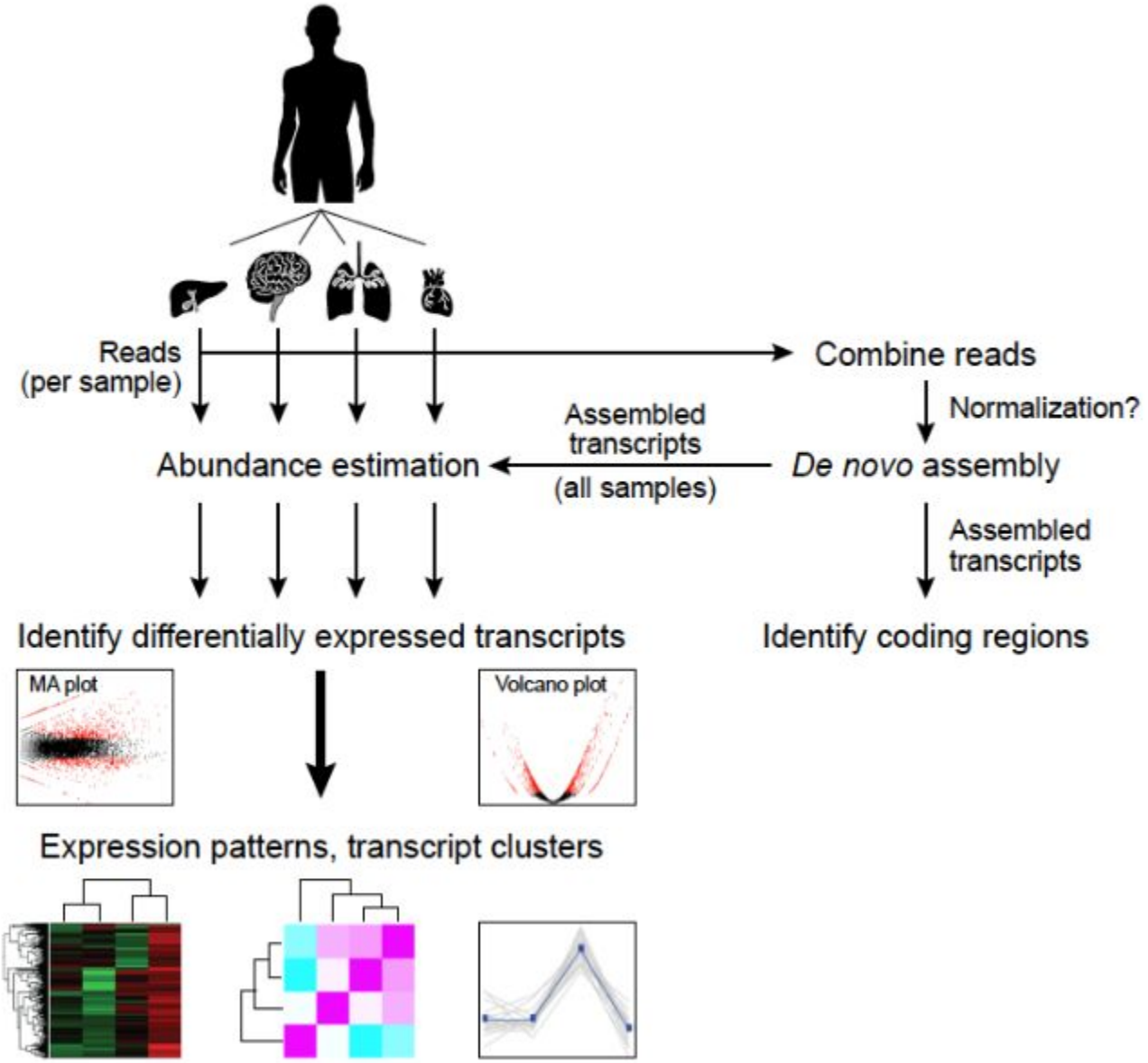
Aligned to Mouse Genome



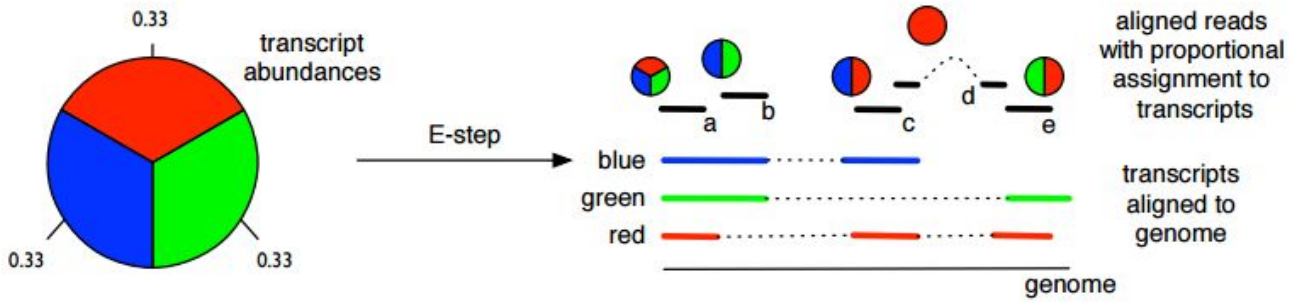
Trinity output: A multi-fasta file



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.

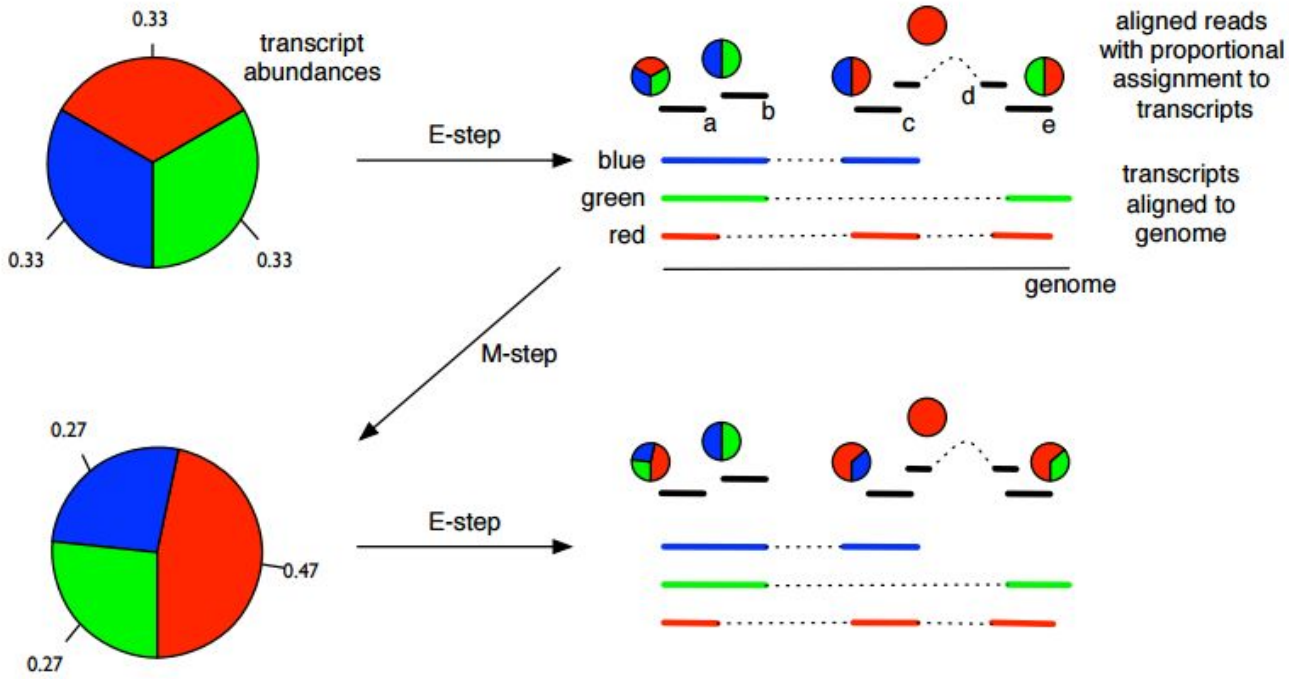
Calculation: For every ambiguous read, the probability of it coming from a specific transcript is calculated based on the current relative abundance of that transcript.

Step M: Maximization (M-Step)

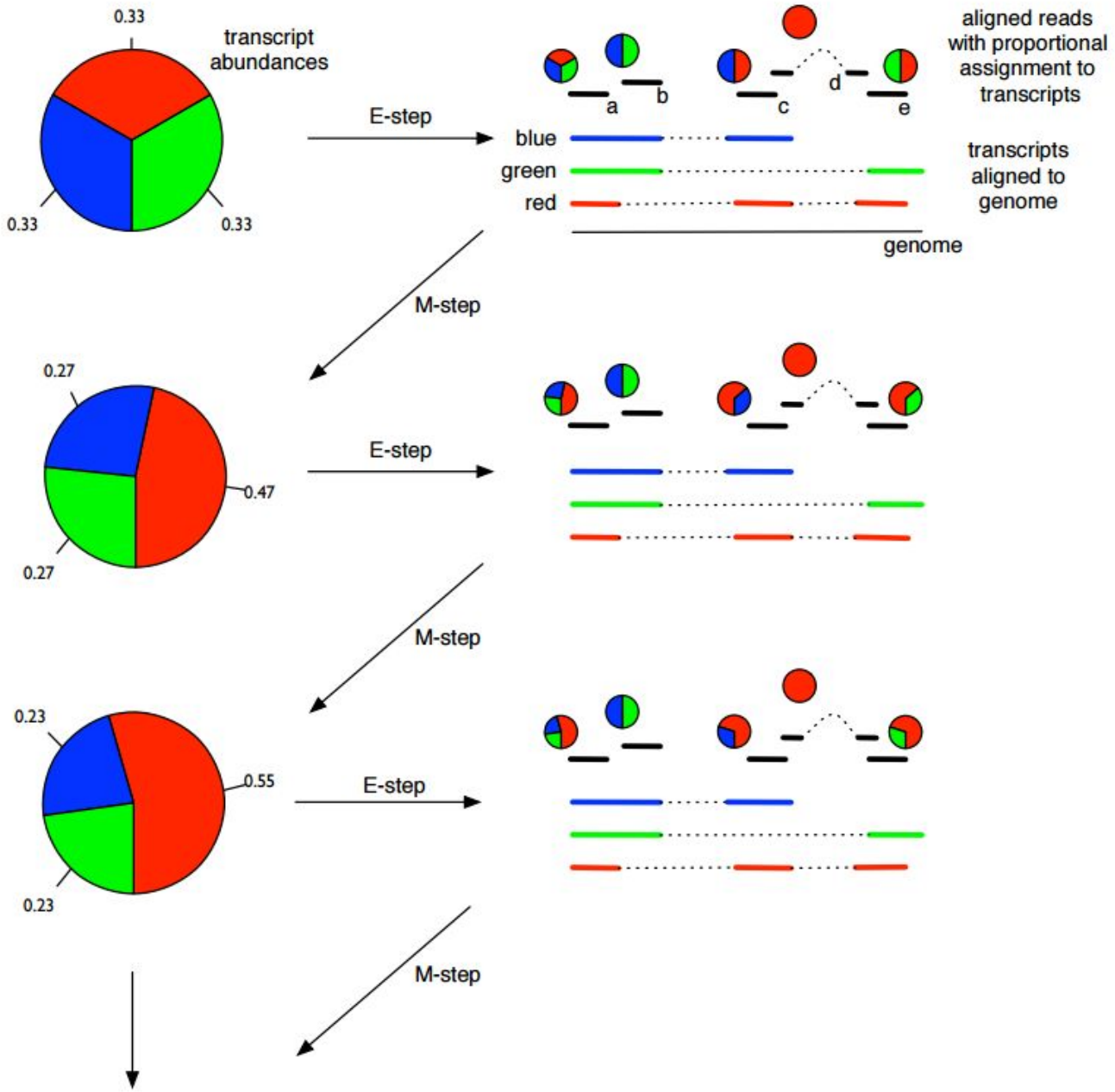
In this step, the algorithm recalculates the **total abundance** of each transcript using the **new, probabilistic counts** determined in the E-step. This maximizes the likelihood of the observed data.

- **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

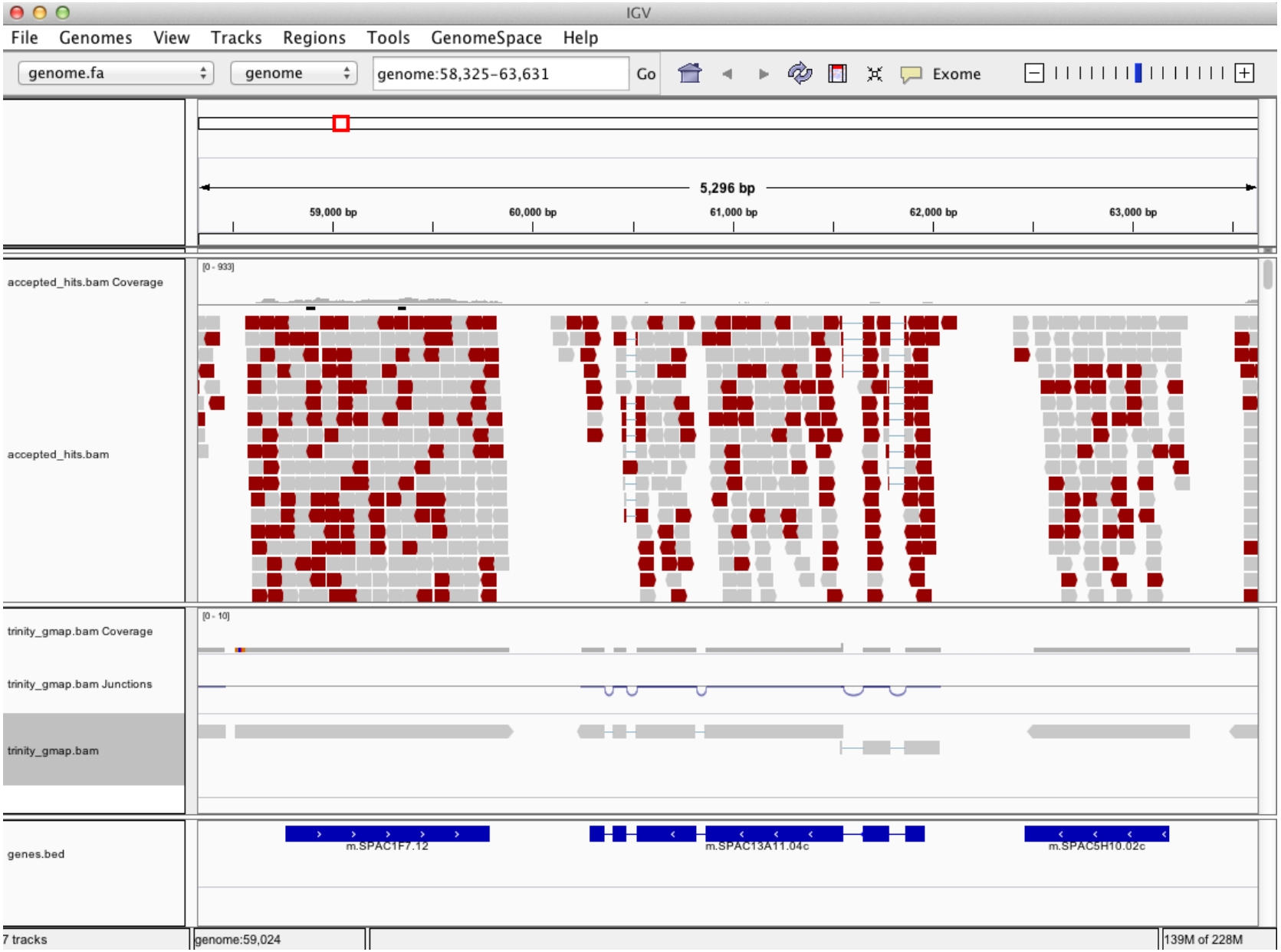
- Assemble RNA-Seq using Trinity
- Examine Trinity in context of a genome:
  - Align Trinity transcripts to the genome using GMAP
  - Align rna-seq reads to genome using Tophat
  - Visualize all alignments using IGV

Try yourself:

```
echo 'export TRINITY_HOME=/home/reczko/tools/trinityrnaseq-v2.15.2' >> ~/.bashrc
source ~/.bashrc
export PATH=/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
```

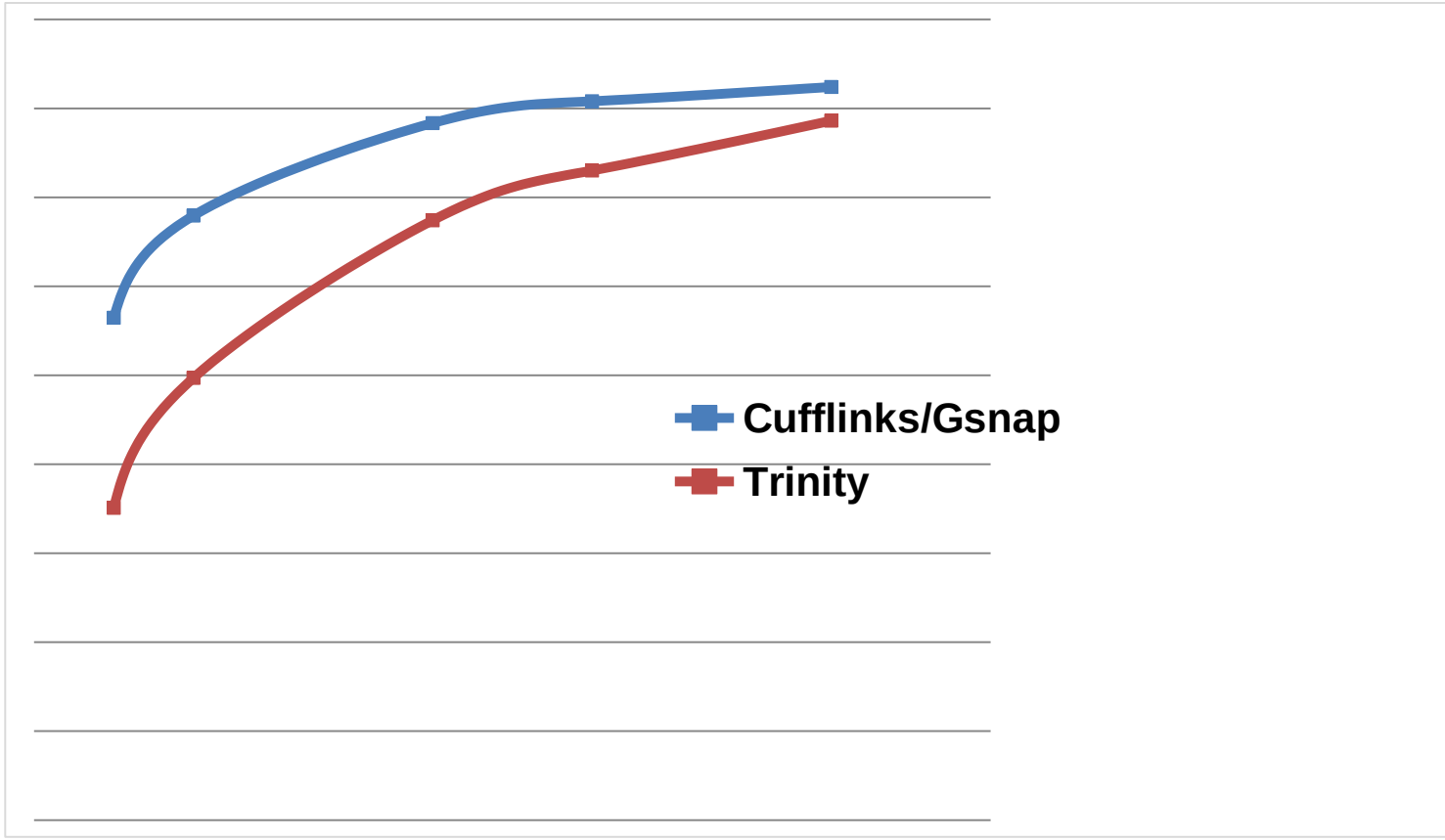
Trinity transcripts aligned to genome scaffolds to examine intron/exon structures

(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

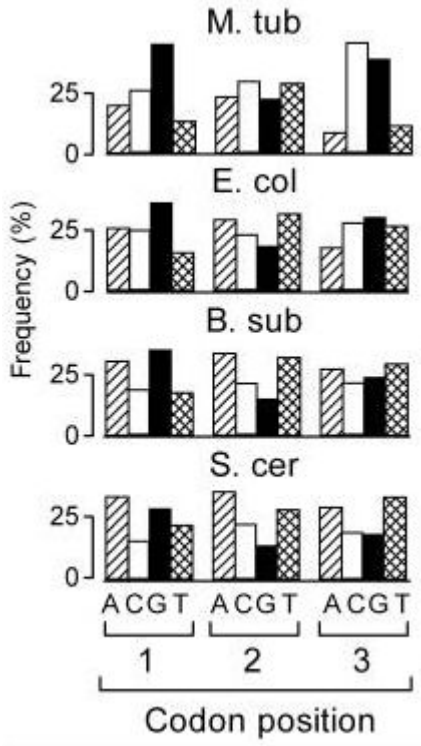


Million PE reads

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.



- http://www.fruitfly.org/GASP1/tutorial/presentation/sld082.htm

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3125259/

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