Syllabus and grading

#	Date	Short title	Lecturer	Other actions	Subject
1	Tuesday, October 10, 2023	introduction	MR		Overview of Bioinformatics, sequence alignment
2	Tuesday, October 17, 2023	Linux/shell/ssh	AD		Introduction to Linux and the command line, bash scripting and ssh
3	Tuesday, October 24, 2023	R (1)	AD		Introduction to the R programming language and Rstudio usage
4	Tuesday, October 31, 2023	QC+RNASeq	MR		Next generation sequencing: introduction, quality control and gene expression analysis for RNAseq
5	Tuesday, November 7, 2023	R (2)	AD		Advances R subjects, introduction to Bioconductor
6	Tuesday, November 14, 2023	bedtools/vcftools/samtools	AD		Command line tool usage: bedtools, vcftools, samtools etc.
7	Tuesday, November 21, 2023	Denovo	MR		NGS for denovo genome and transciptome assembly
8	Tuesday, November 28, 2023	ChipSeq/chirp	MR	assign presentations	NGS analysis for molecular interactions (ChipSeq, (Par-)Clip, structural sequencing, chromosome conformation capture (3C))
9	Tuesday, December 5, 2023	metabolomics	MR		Genome-scale models of metabolism and macromolecular expression, Biological applications of Transformers
10	Tuesday, December 12, 2023	Exome/SNP calling	AD	assign final projects	Pipelines for SNP calling, especially for exome sequencing using the GATK pipeline
11	Tuesday, December 19, 2023	presentations	MR+AD		Paper presentations by students
12	Tuesday, January 9, 2024	presentations	MR+AD		Paper presentations by students
13	Tuesday, January 16, 2024	final projects support	MR+AD		Support for the final project

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

Isoform quantitation tools in the literature (Q3-2020)

- 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)

elixir GREECE

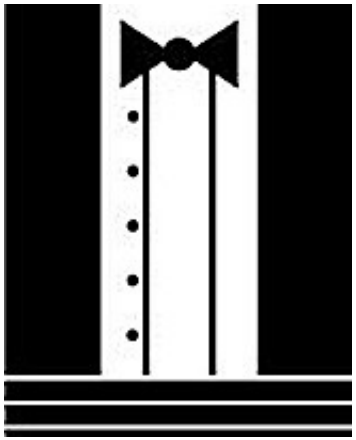
De-novo genome sequence assembly, Genome-Based and Genome-Free Transcript Reconstruction and Analysis Using RNA-Seq Data

based on material from Mathias Haimel, EBI

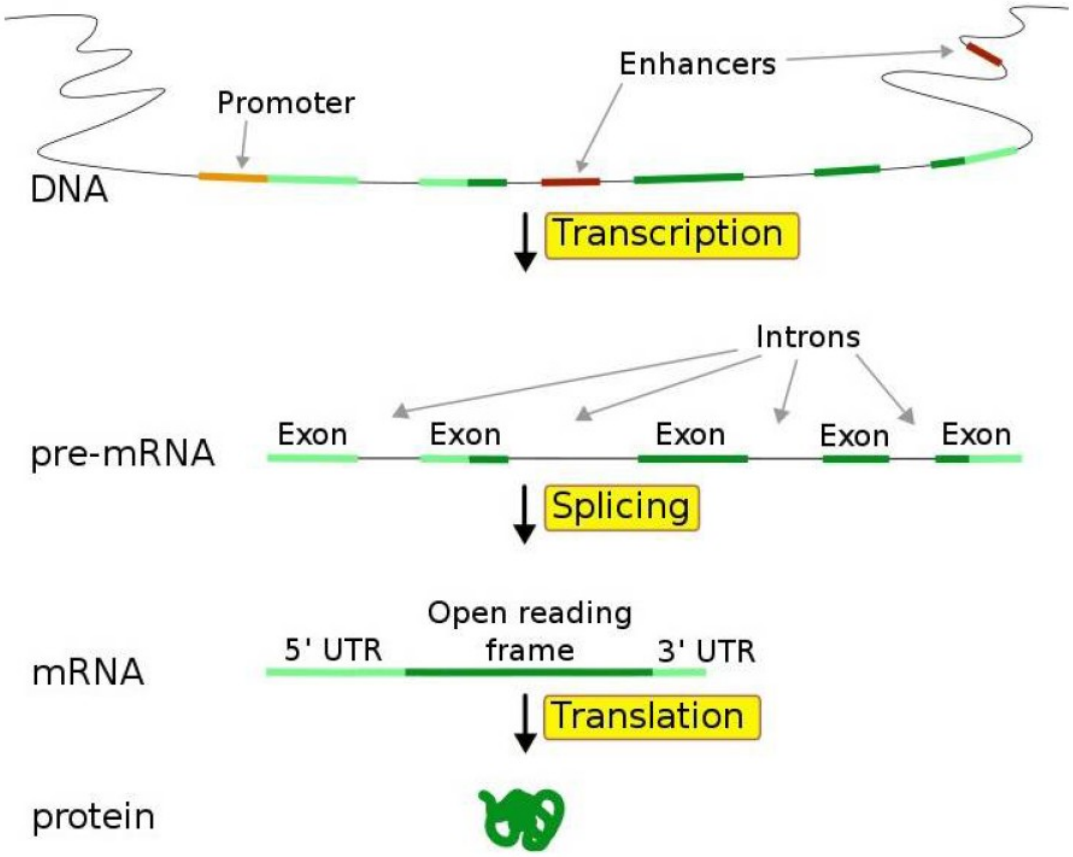
https://www.ebi.ac.uk/training/online/sites/ebi.ac.uk.training.online/files/user/18/private/velvet_1.pdf

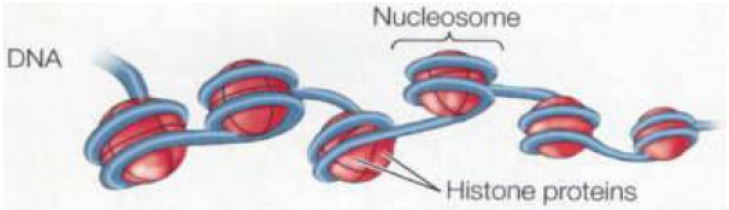
and Brian Haas

Broad Institute, modified by M. Reczko



Next Generation Sequencing



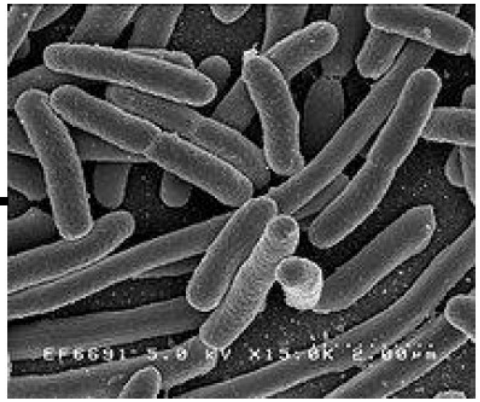


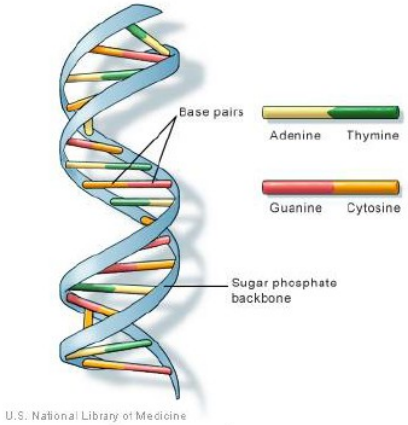
Whole Genome sequencing

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

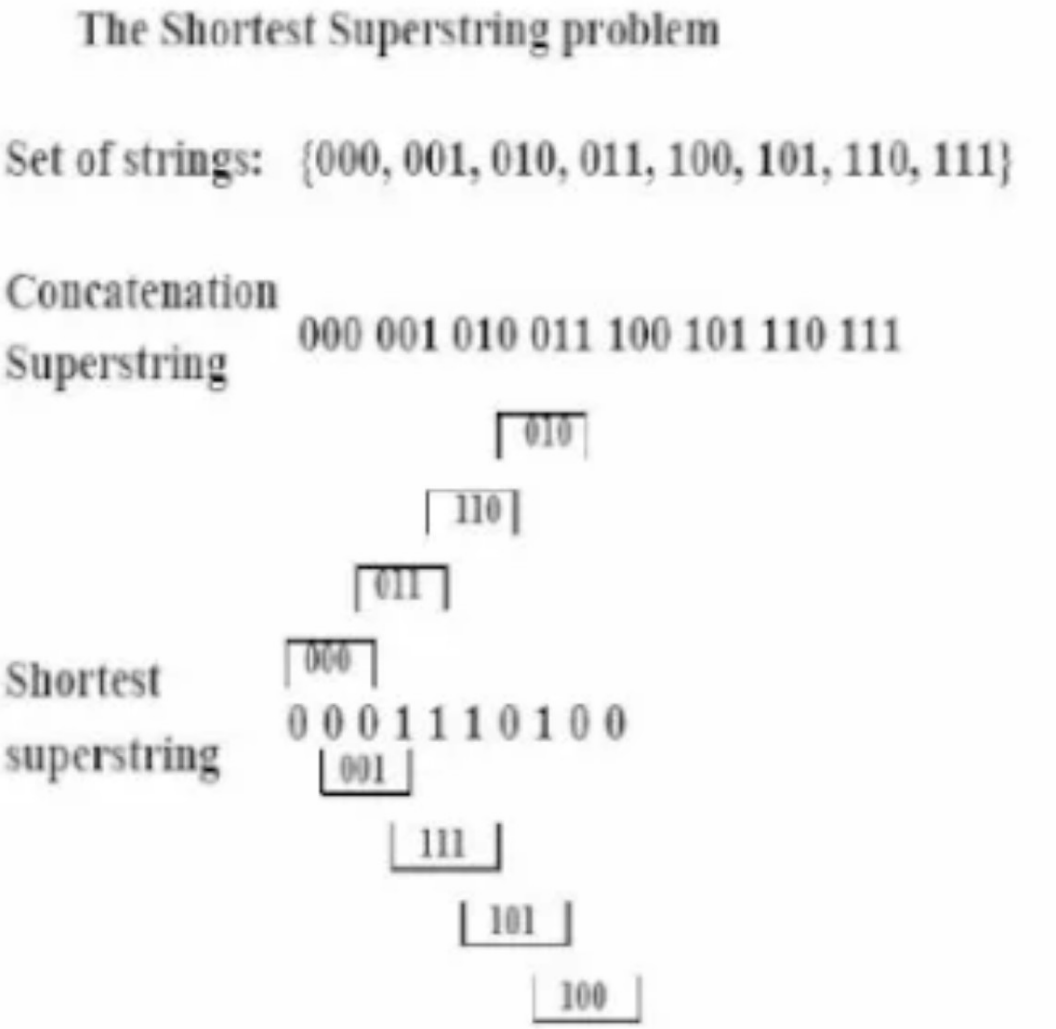
- expressed gene content
- transcript abundance
- differential expression

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



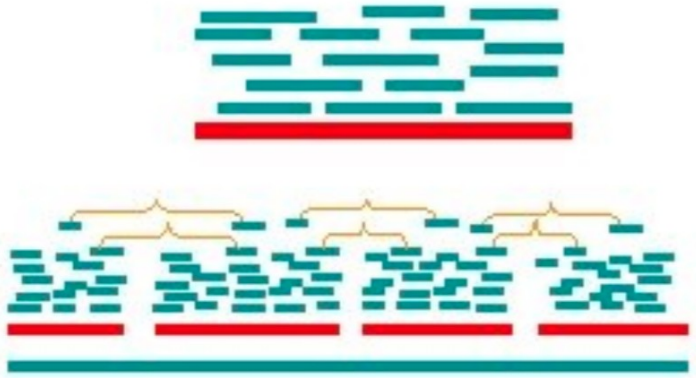
http://feedthedatamonster.com/home/2014/12/24/three-billion-puzzle-pieces-insights-into-the-human-genome-project

Overlap-Layout-Consensus

Assemblers: ARACHNE, PHRAP, CAP, TIGR, CELERA

Overlap: find potentially overlapping reads

Layout: merge reads into contigs and contigs into supercontigs



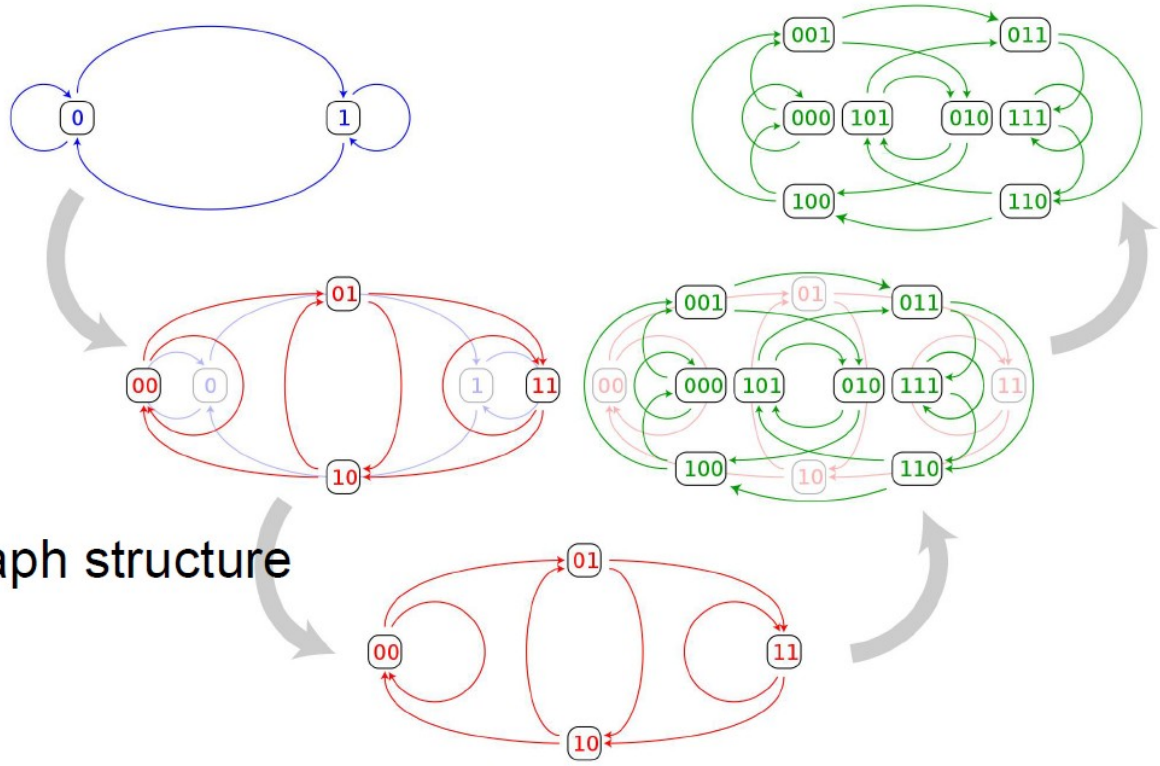
Consensus: derive the DNA sequence and correct read errors

..ACGATTACAATAGGTT..

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

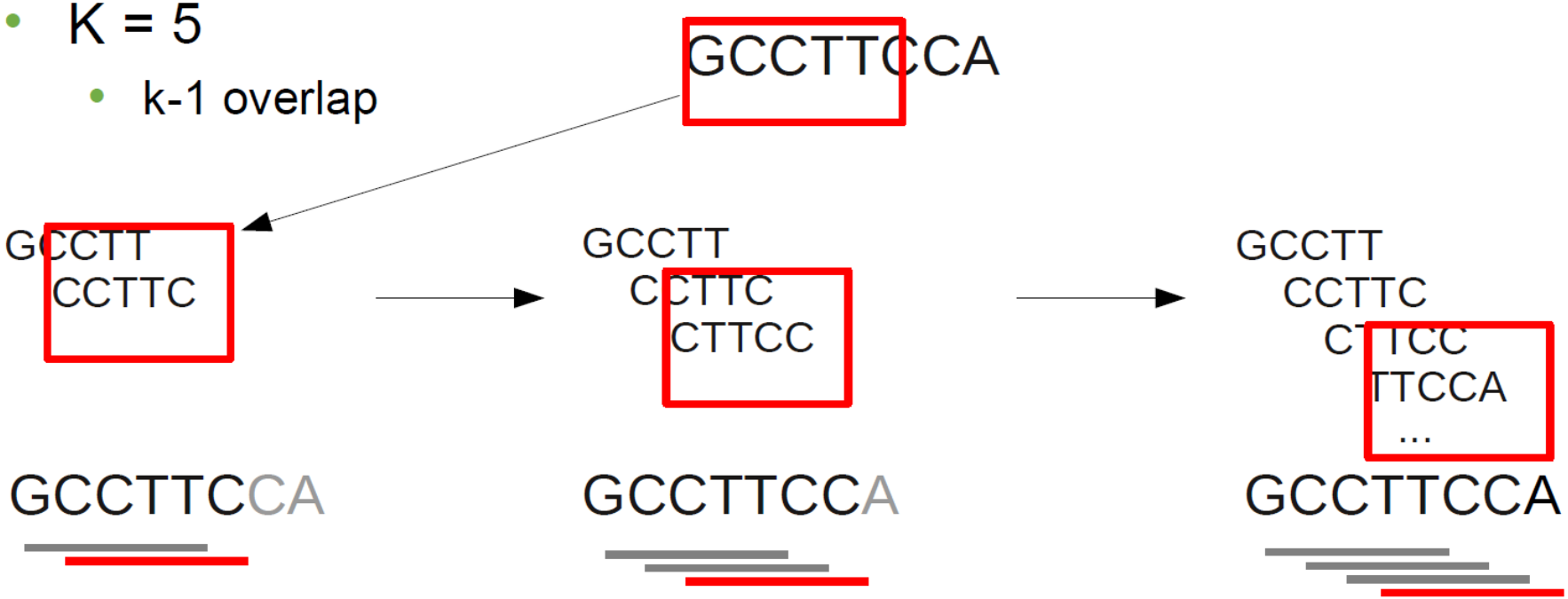
De Bruijn graph

- 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



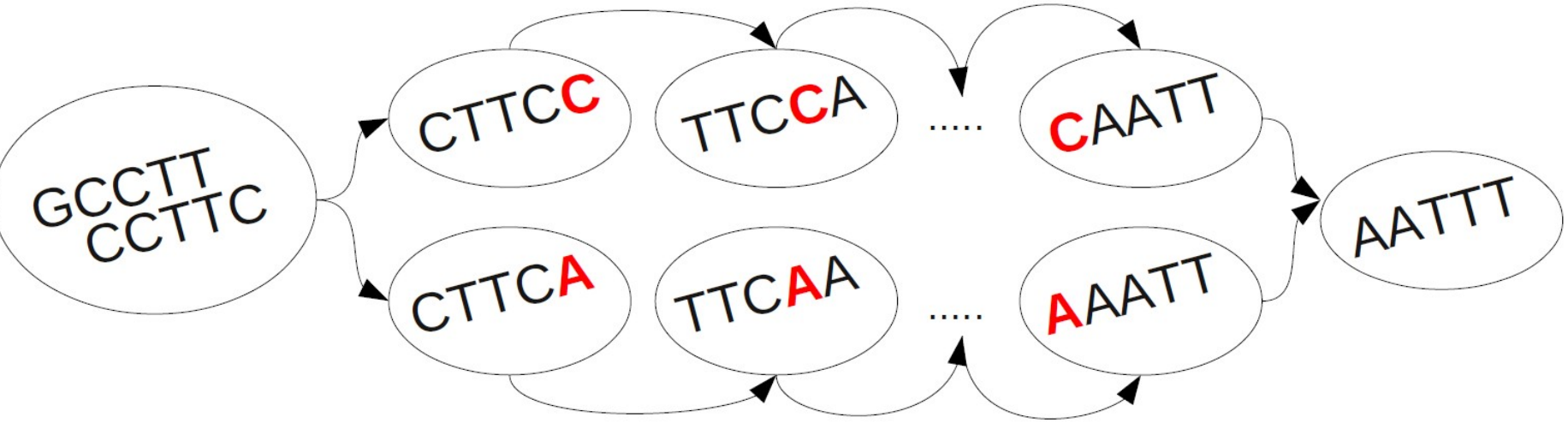
De Bruijn graph (Velvet)

- 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

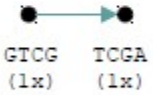
```
 AGTCGAG CTTTAGA  CGATGAG CTTTAGA
  GTCGAGG  TTAGATC  ATGAGGC    GAGACAG
     GAGGCTC   ATCCGAT AGGCTTT GAGACAG
 AGTCGAG    TAGATCC ATGAGGC  TAGAGAA
TAGTCGA  CTTTAGA CCGATGA    TTAGAGA
    CGAGGCT  AGATCCG TGAGGCT  AGAGACA
TAGTCGA GCTTTAG TCCGATG  GCTCTAG
   TCGACGC    GATCCGA GAGGCTT AGAGACA
TAGTCGA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

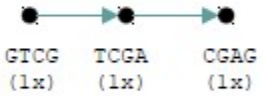
Example

Read: GTCGAGG



Example

Read: GTCGAGG



Example

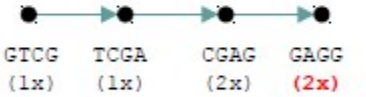
Read: GTCGAGG

Example

New read: CGAGGCT

Example

Read: CGAGGCT



Example

Read: CGAGGCT

Example

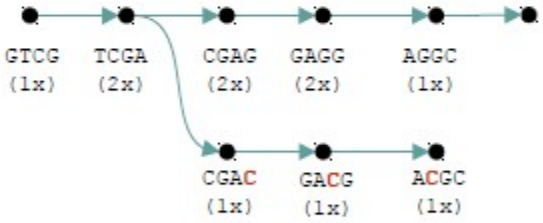
Read: CGAGGCT

Example

New read: TCGACGC

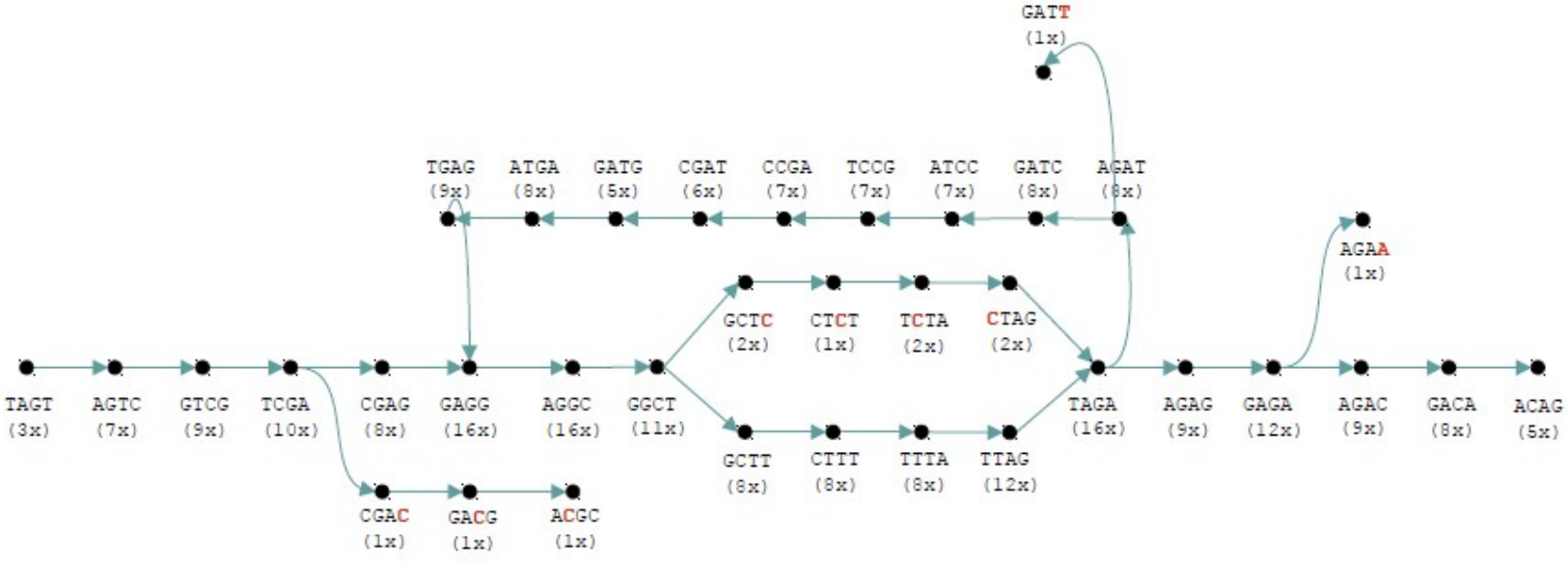
Example

Read: TCGACGC



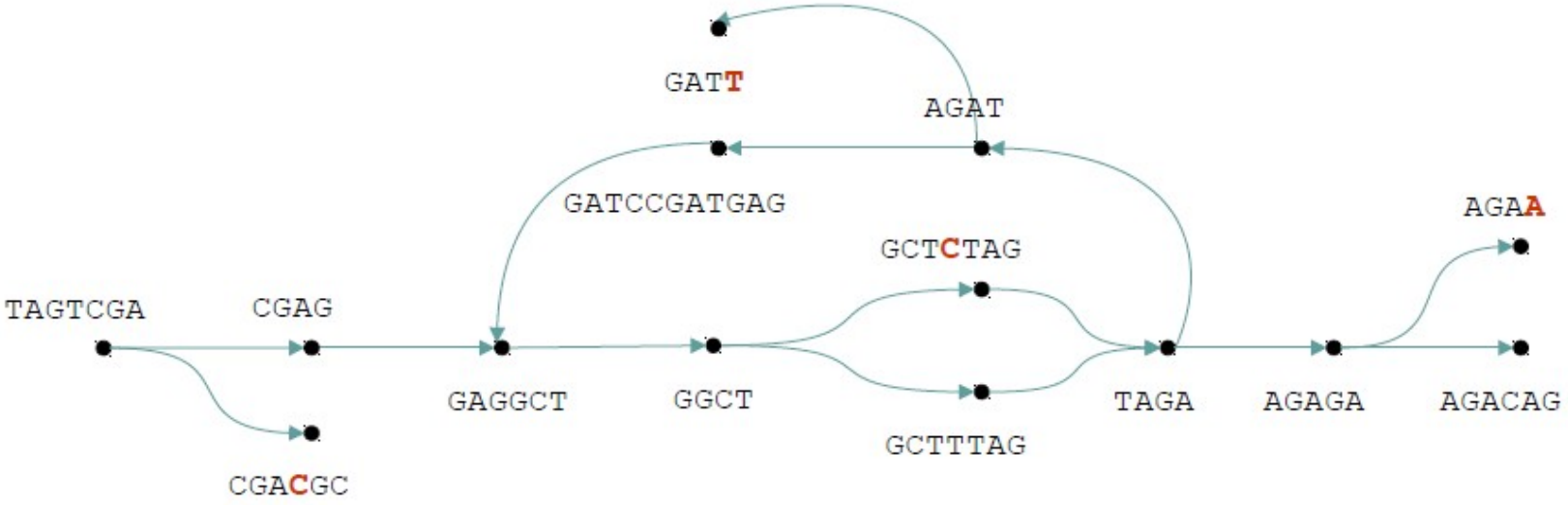
Example

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)

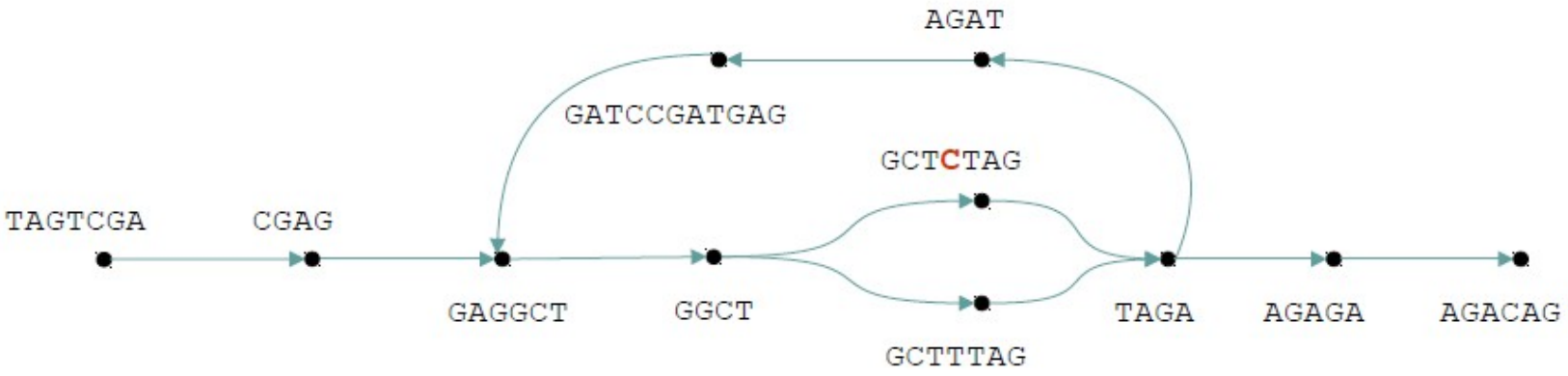
Example

After simplification…



Example

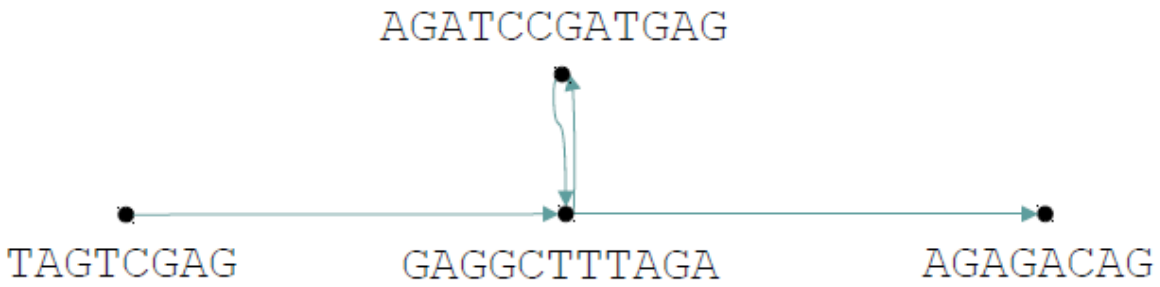
Tips removed…



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**

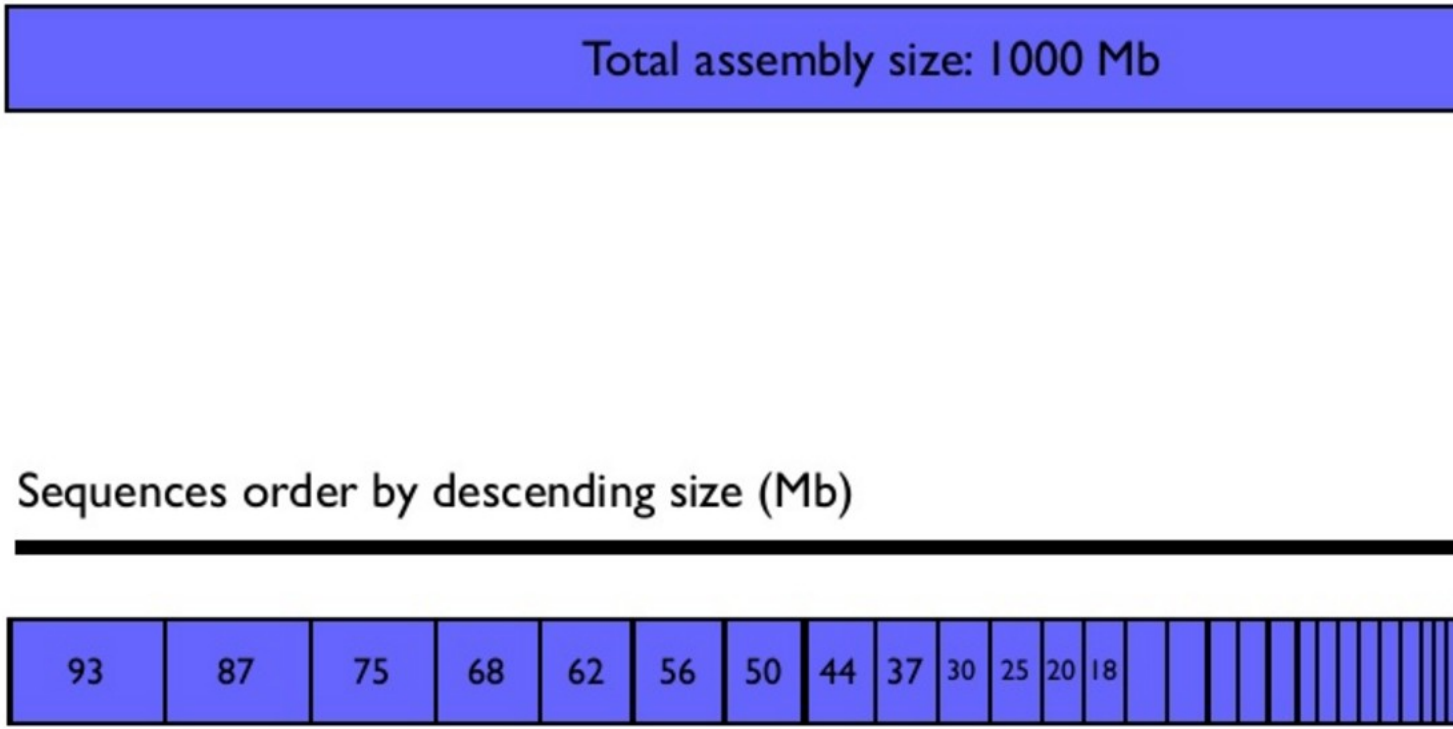
4. **Average scaffold/contig size**

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

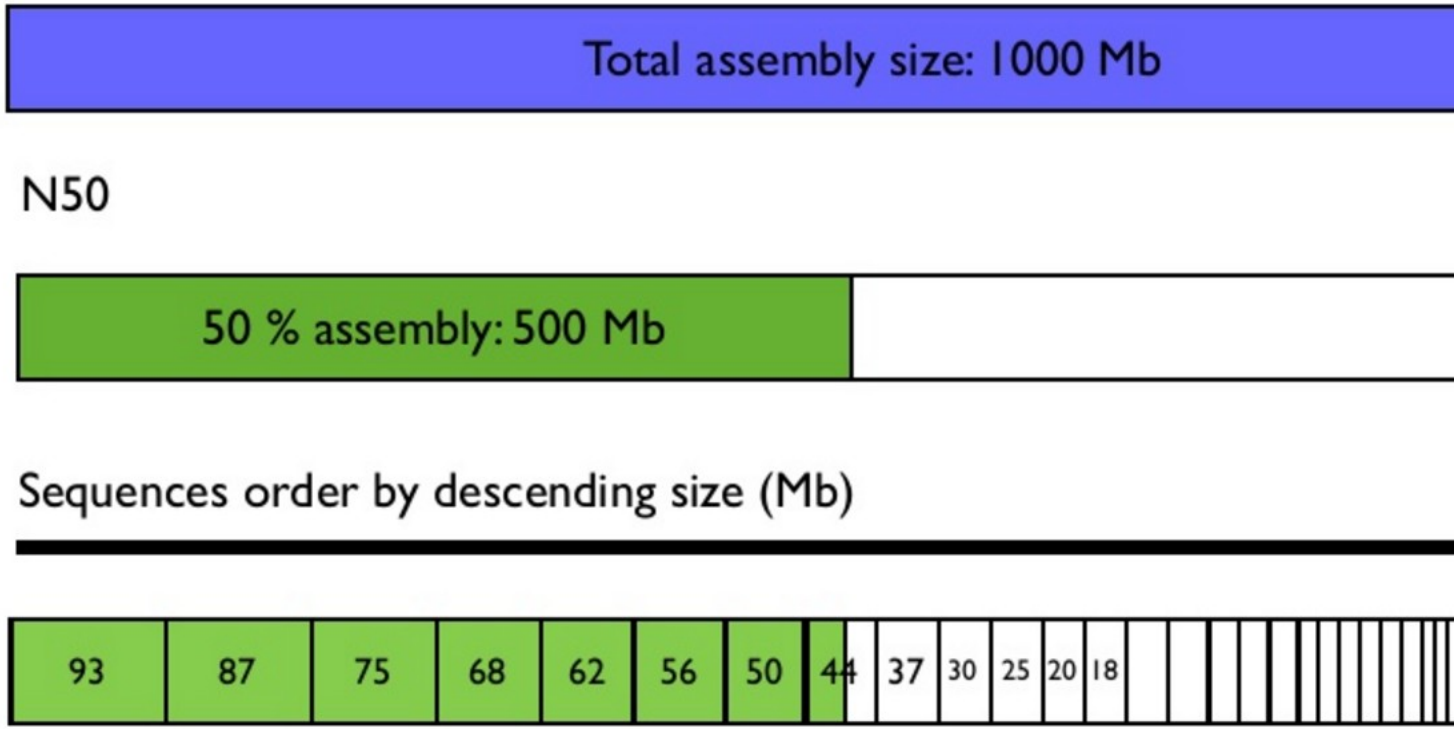
2.6 Assembly evaluation

N50/L50



2.6 Assembly evaluation

N50/L50

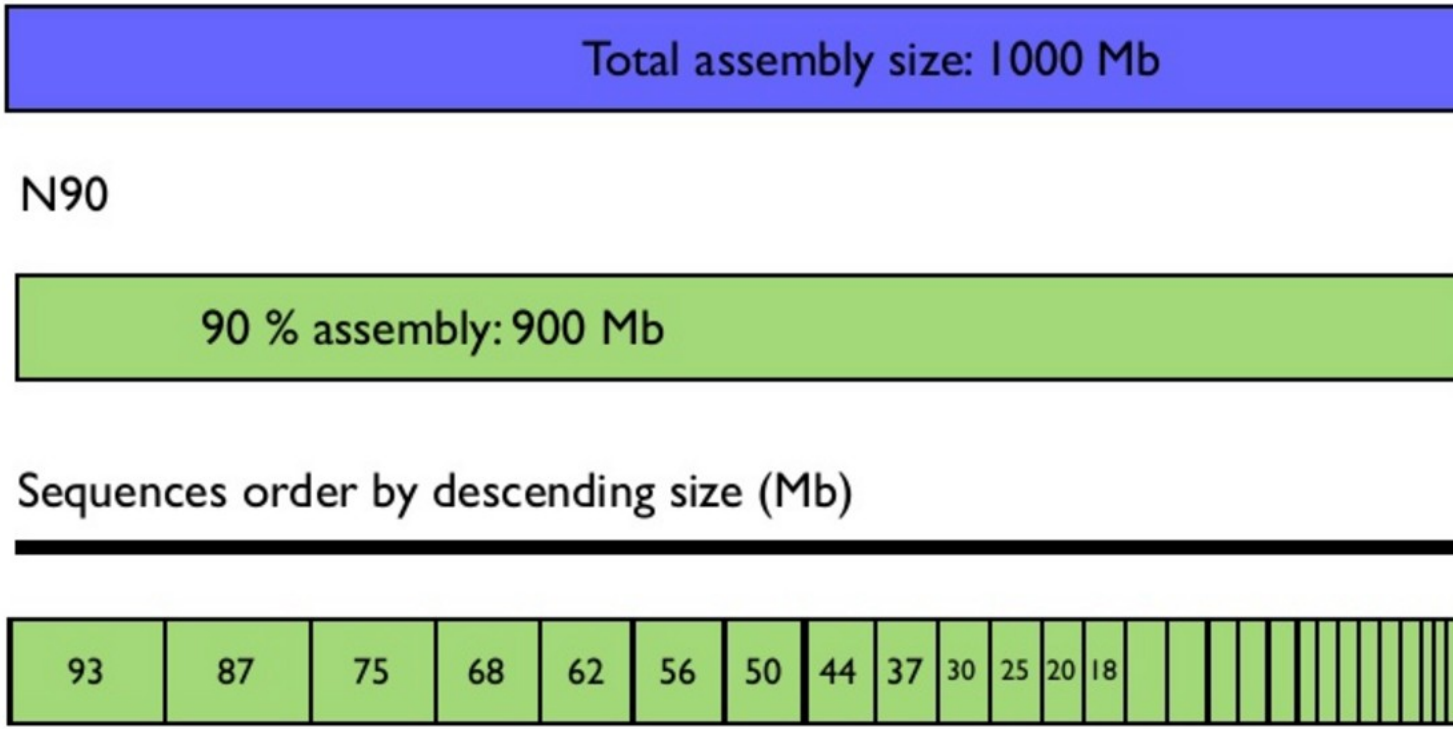


N50 = 7 sequences

L50 = 50 Mb

2.6 Assembly evaluation

N90/L90



N90 = 29 sequences

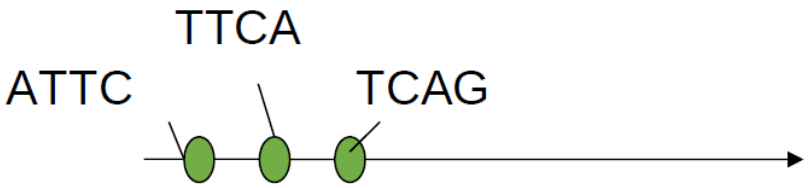
L90 = 12.5 Mb

De Bruijn graph biology extensions (Velvet)

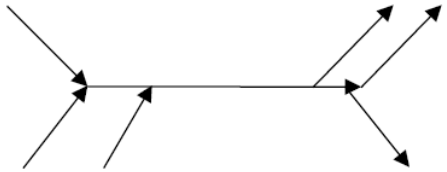
- Handling of reverse strand
  - DNA is read in two directions
  - Paired-end data
- Handling small differences, which are “uninteresting”
  - Errors in sequencing technology
- Memory
  - regularly use 80, 100GB real memory
  - easily get to 1TB real memory requirements

De Bruijn graph representations (Velvet)

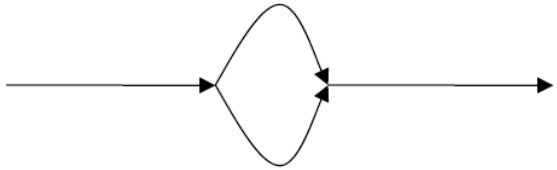
Error free, no repeat,
no polymorphism



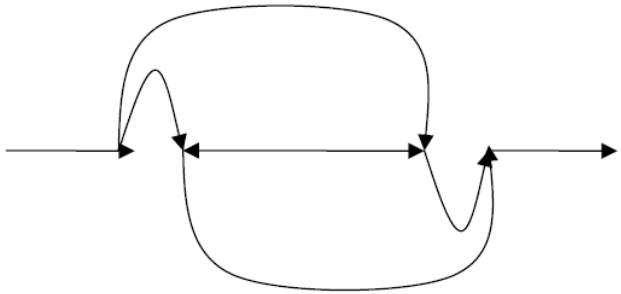
Repeat > kmer length



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

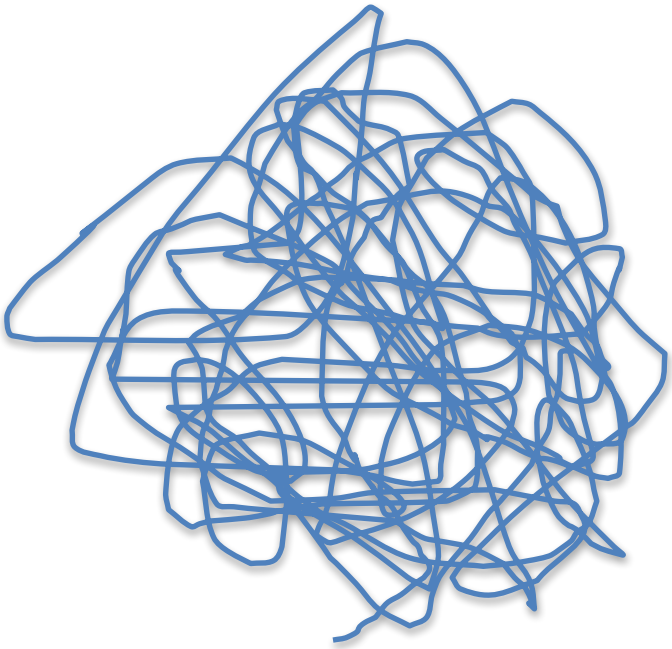
Transcriptome Assembly

- 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

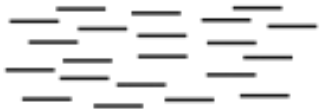
- K. Mathiopoulos, E. Sagri

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...

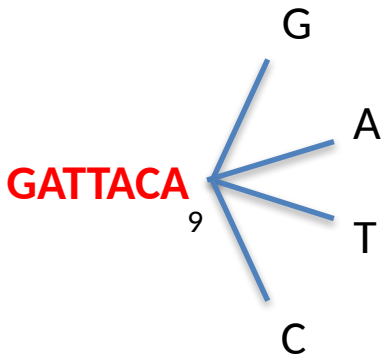
Thousands of disjoint graphs

Inchworm Algorithm

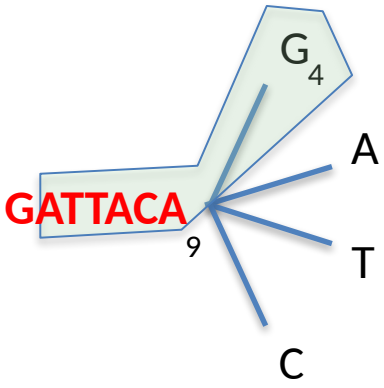
Decompose all reads into overlapping Kmers (25-mers)

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

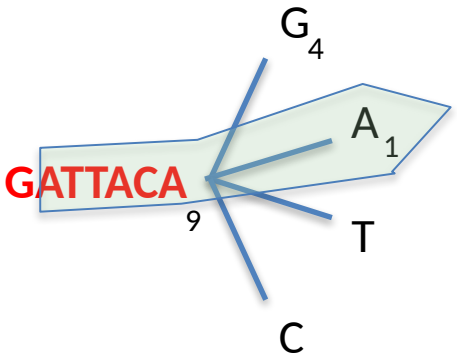
Extend kmer at 3’ end, guided by coverage.



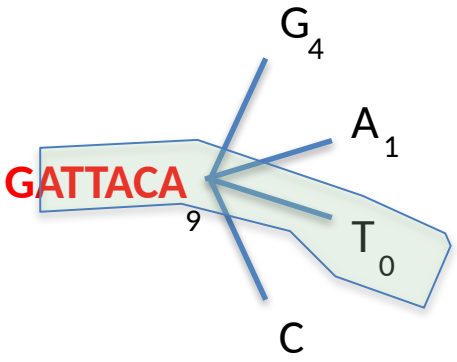
Inchworm Algorithm



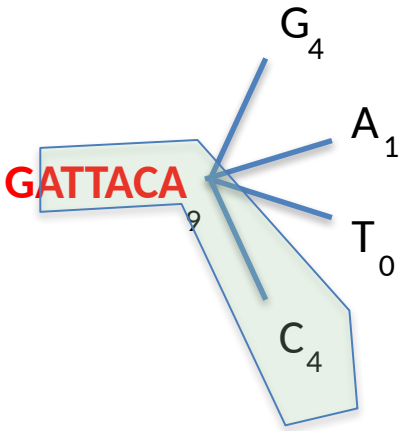
Inchworm Algorithm



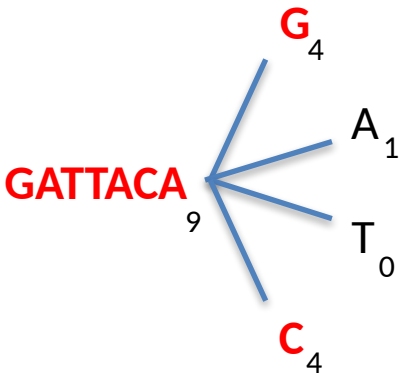
Inchworm Algorithm



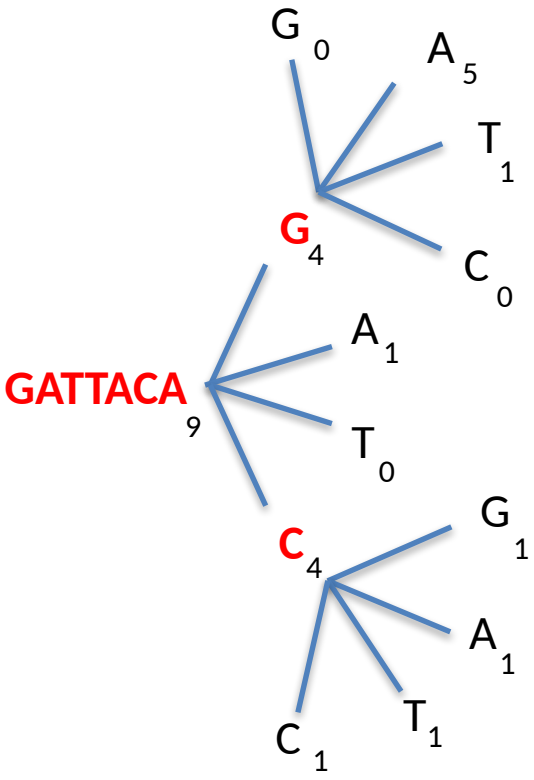
Inchworm Algorithm



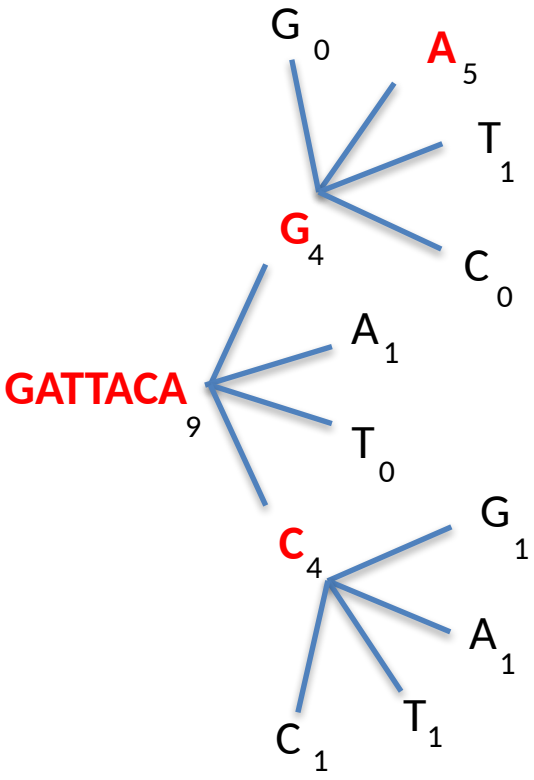
Inchworm Algorithm



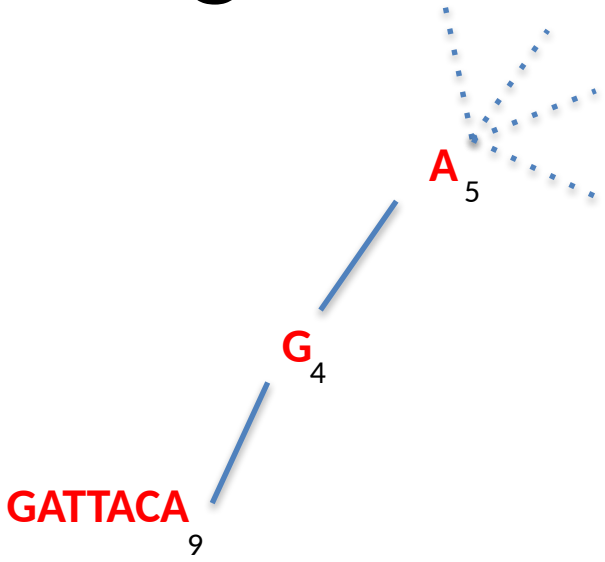
Inchworm Algorithm



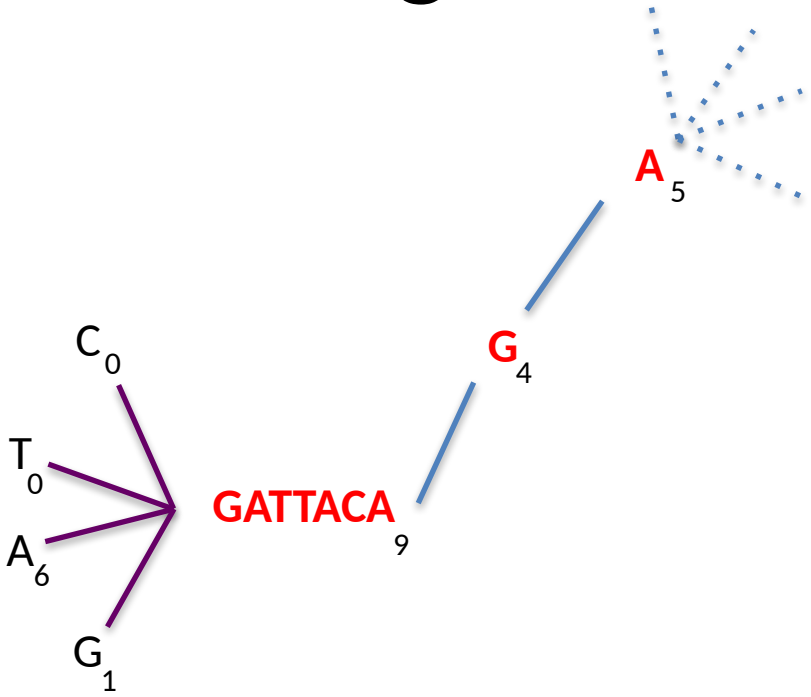
Inchworm Algorithm



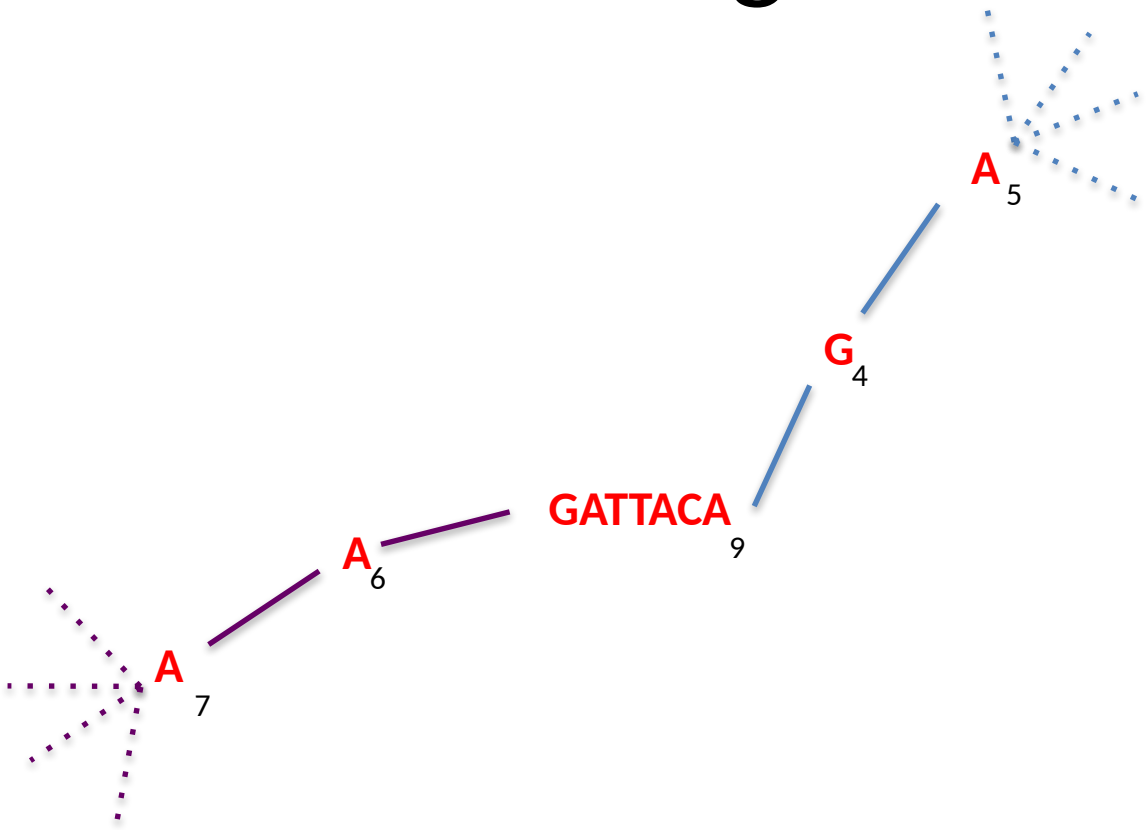
Inchworm Algorithm



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

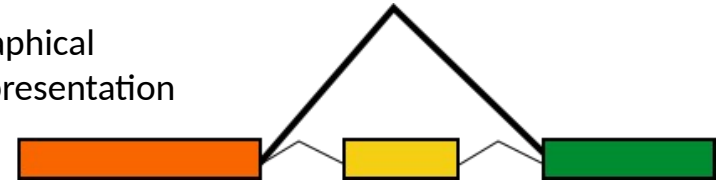
Isoform B

Expression

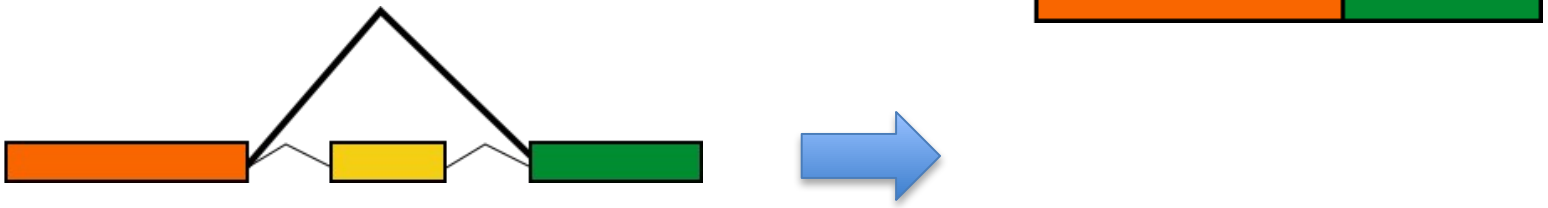
(low)

(high)

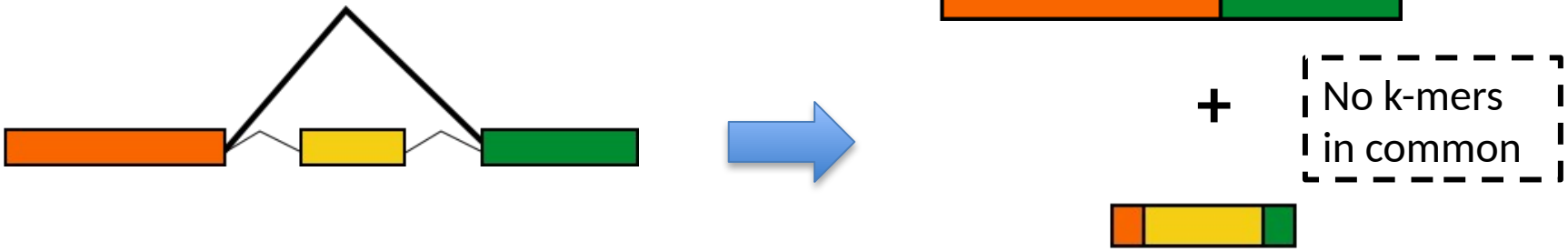
Graphical representation



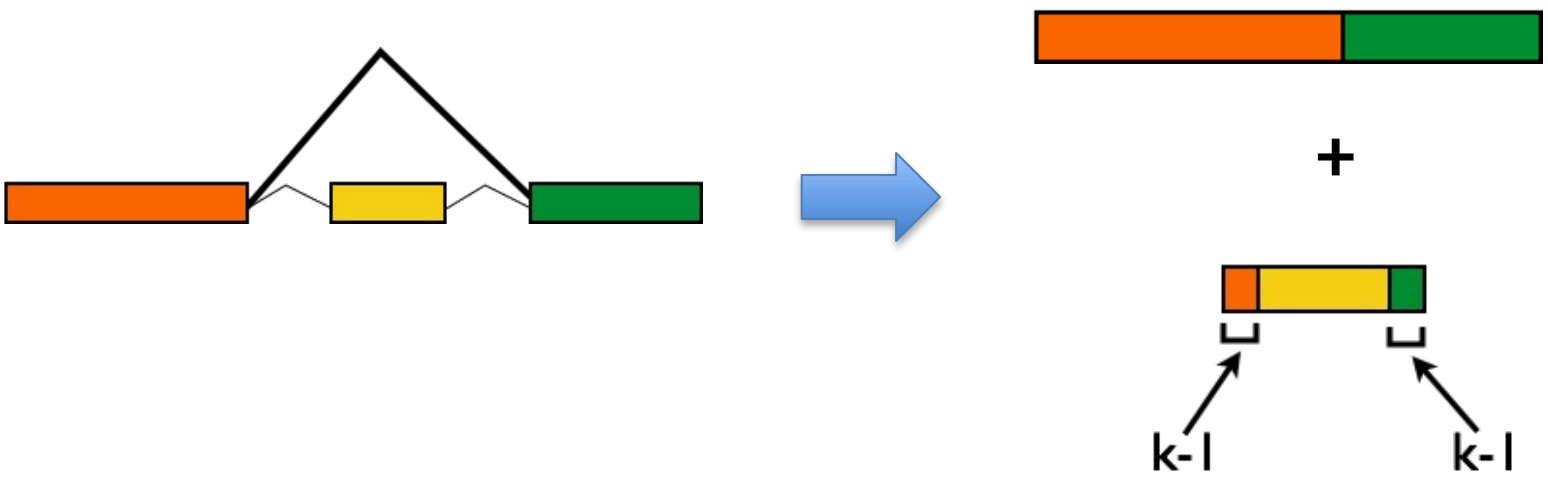
Inchworm Contigs from Alt-Spliced Transcripts



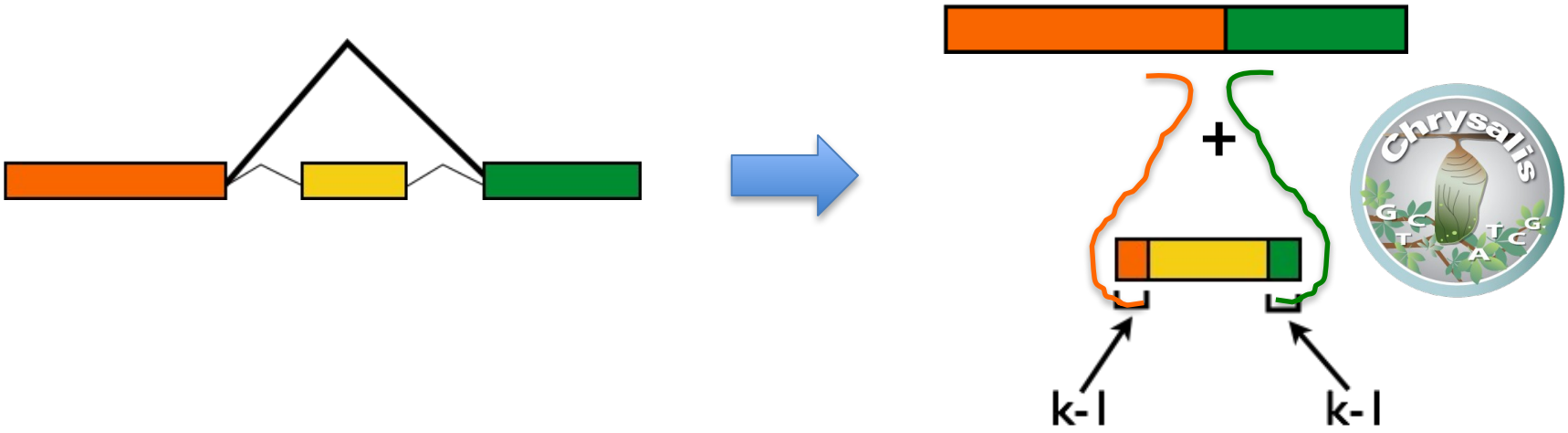
Inchworm Contigs from Alt-Spliced Transcripts



Inchworm Contigs from Alt-Spliced Transcripts



Chrysalis Re-groups Related Inchworm Contigs



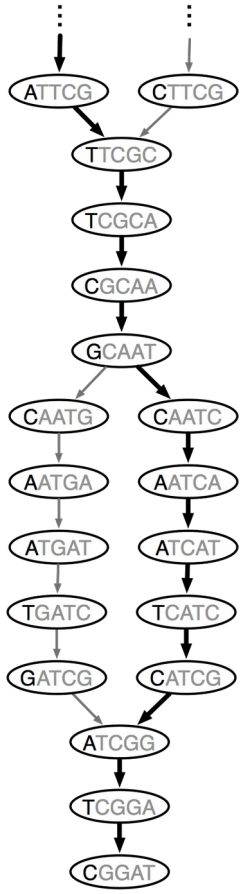
Chrysalis uses (k-1) overlaps and read support to link related Inchworm contigs

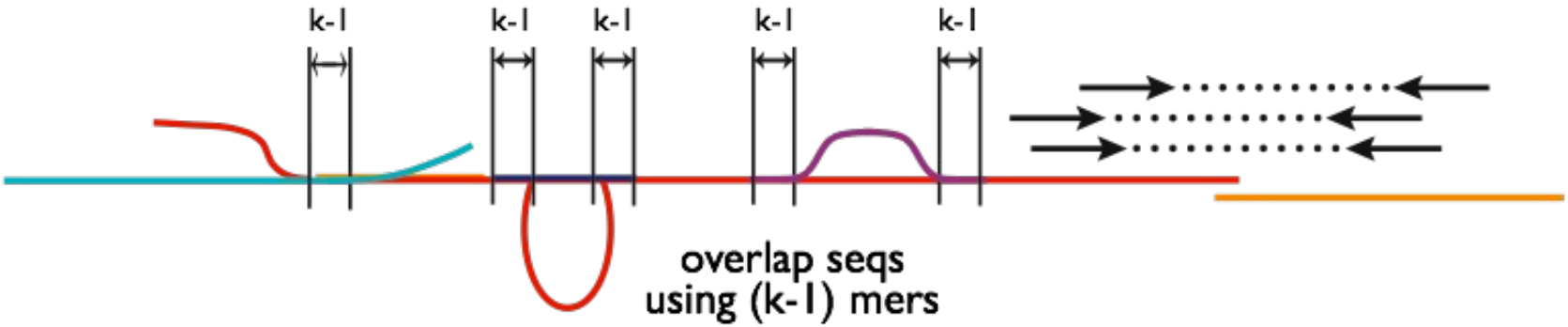
Chrysalis

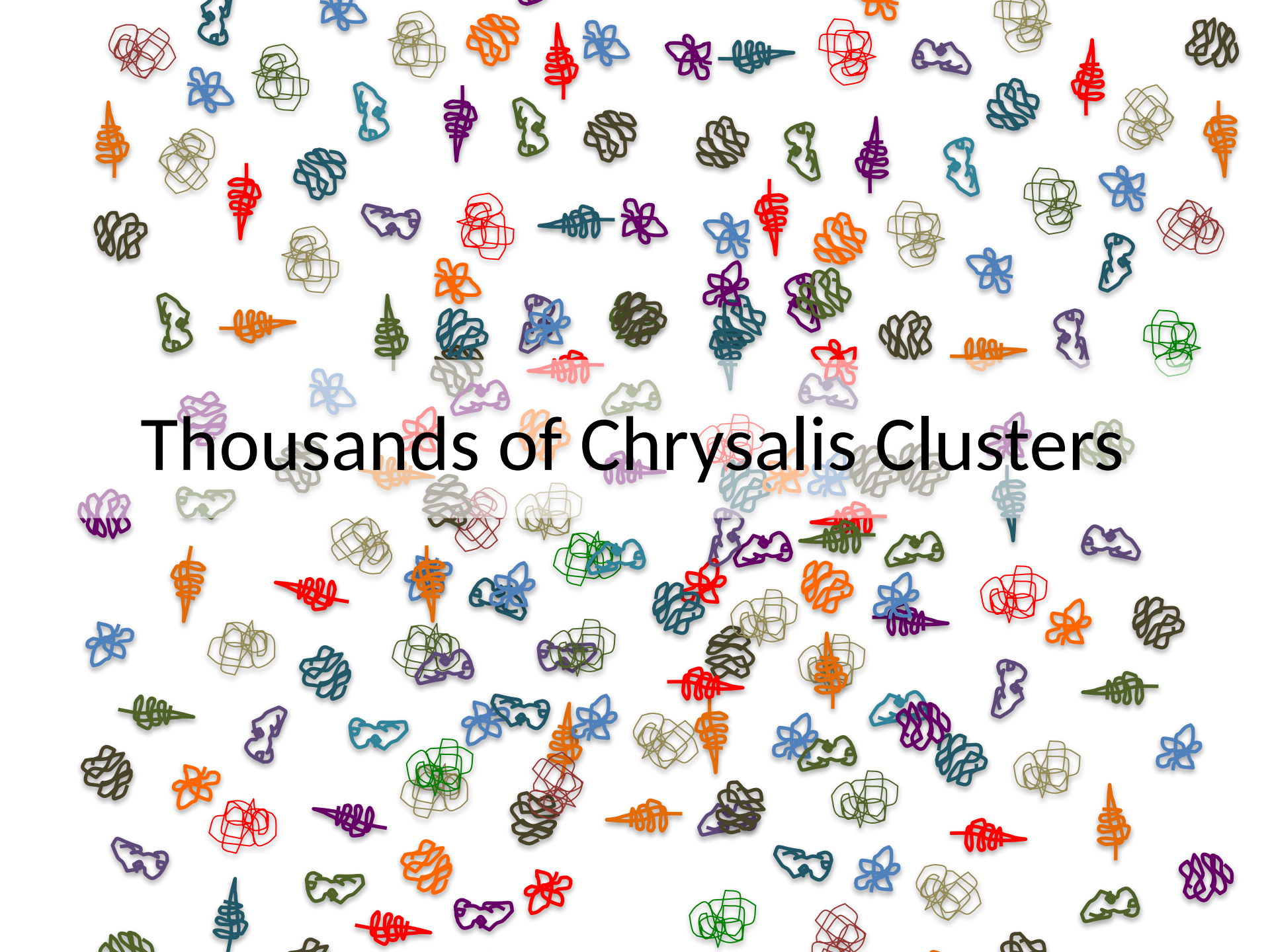
>a121:len=5845
>a122:len=2560
>a123:len=4443
>a124:len=48
>a125:len=8876
>a126:len=68

Integrate isoforms
via k-1 overlaps

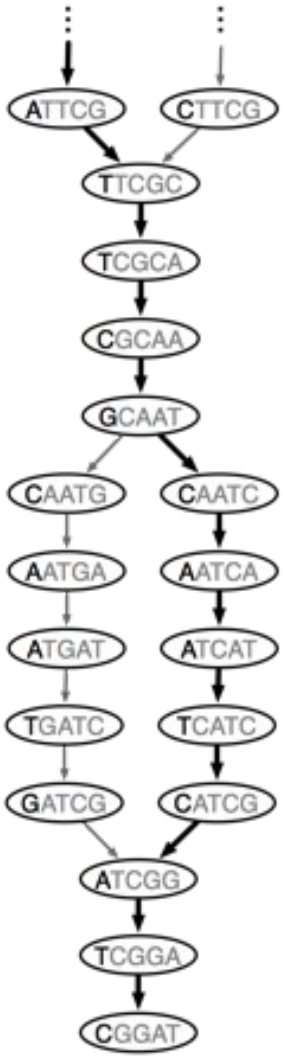
Build de Bruijn Graphs
(ideally, one per gene)





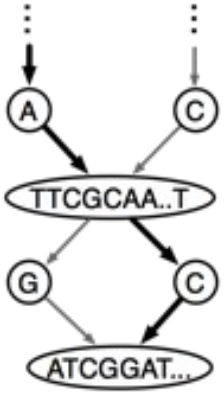

Thousands of Chrysalis Clusters

Butterfly



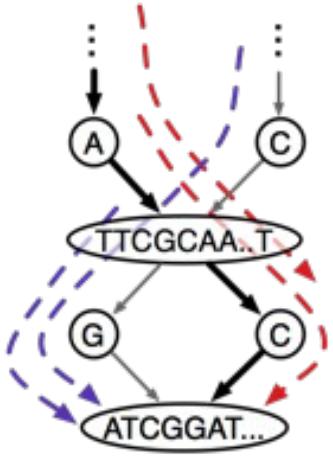
de Bruijn graph

compacting



compact graph

finding paths



compact graph with reads

extracting sequences

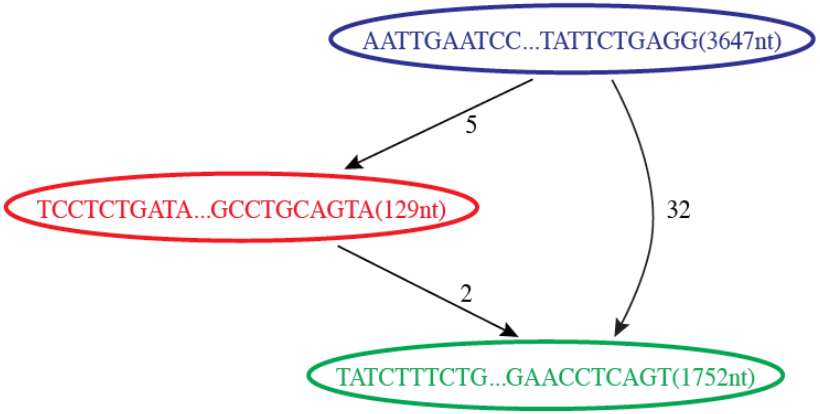
..CTTCGCAA..TGATCGGAT...

..ATTCGCAA..TCATCGGAT...

sequences (isoforms and paralogs)

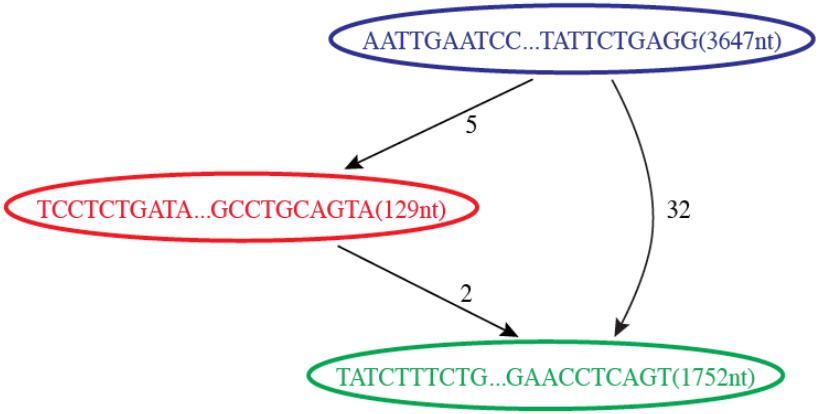
Butterfly Example 1:
Reconstruction of Alternatively Spliced Transcripts

Butterfly's Compacted Sequence Graph



Reconstruction of Alternatively Spliced Transcripts

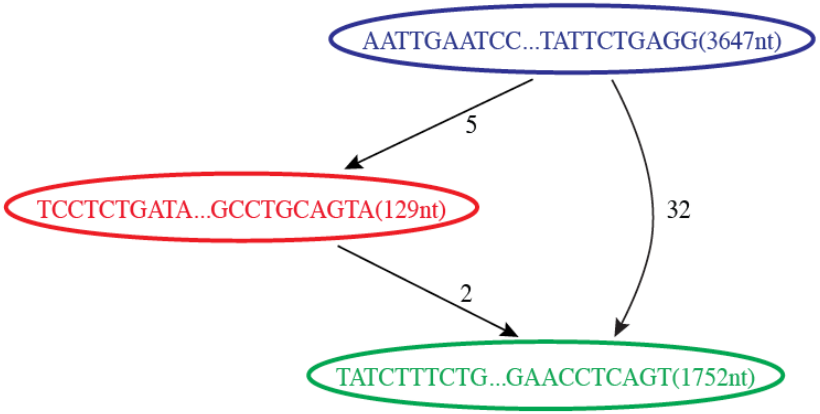
Butterfly's Compacted Sequence Graph



Reconstructed Transcripts

Reconstruction of Alternatively Spliced Transcripts

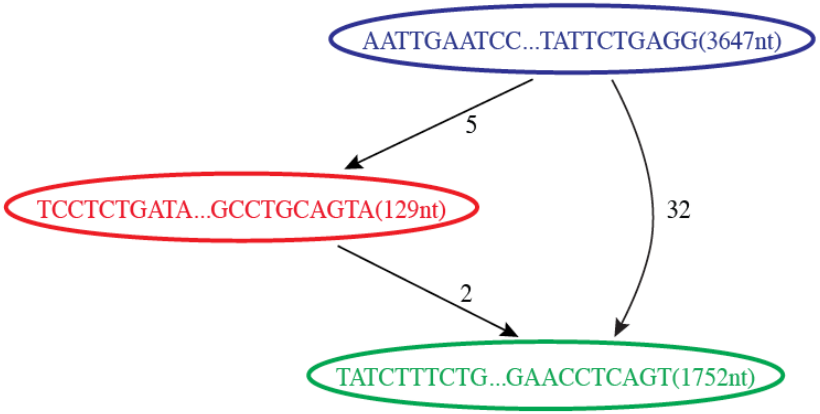
Butterfly's Compacted Sequence Graph



Reconstructed Transcripts

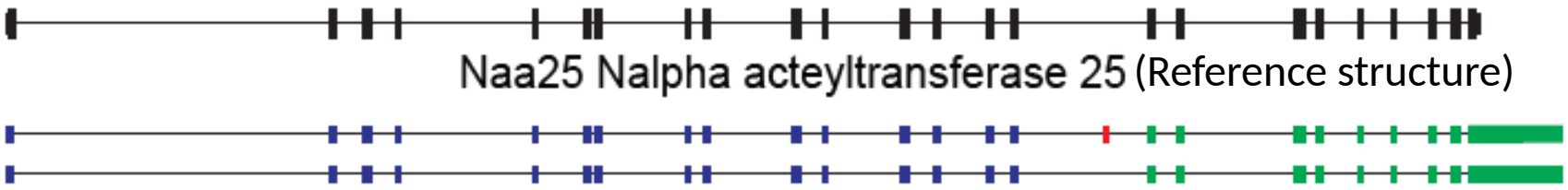
Reconstruction of Alternatively Spliced Transcripts

Butterfly's Compacted Sequence Graph

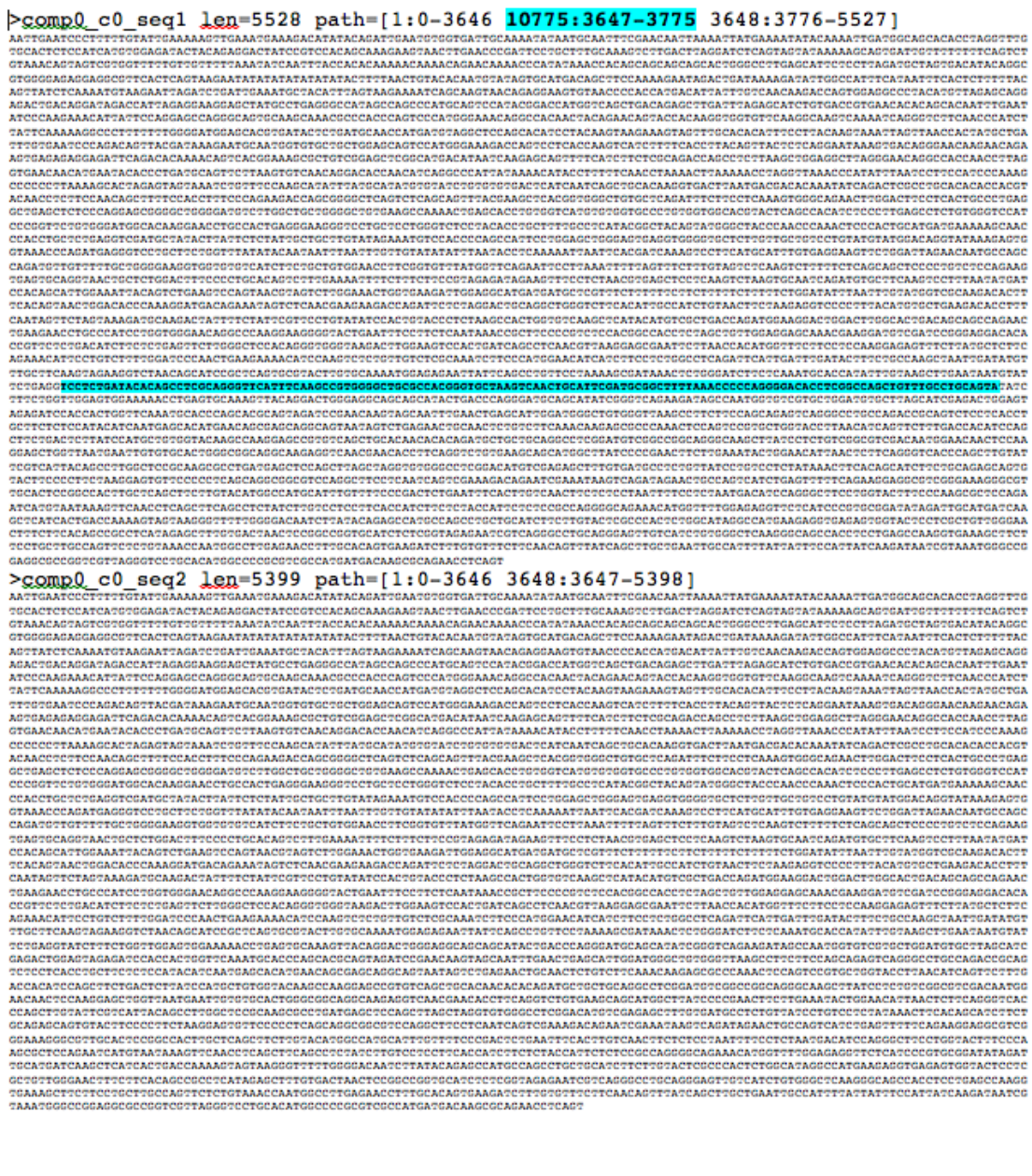


Reconstructed Transcripts

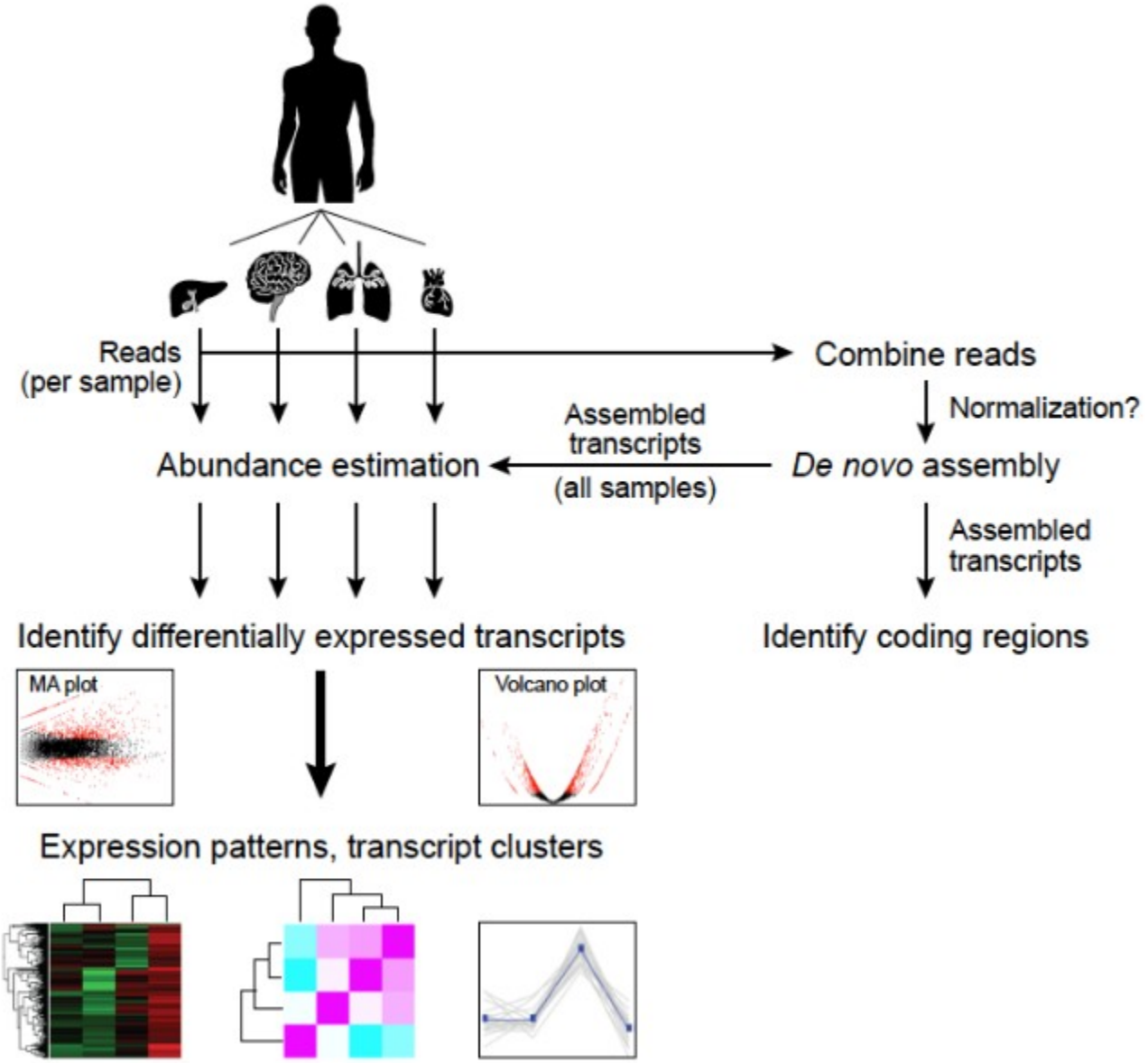
Aligned to Mouse Genome



Trinity output: A multi-fasta file

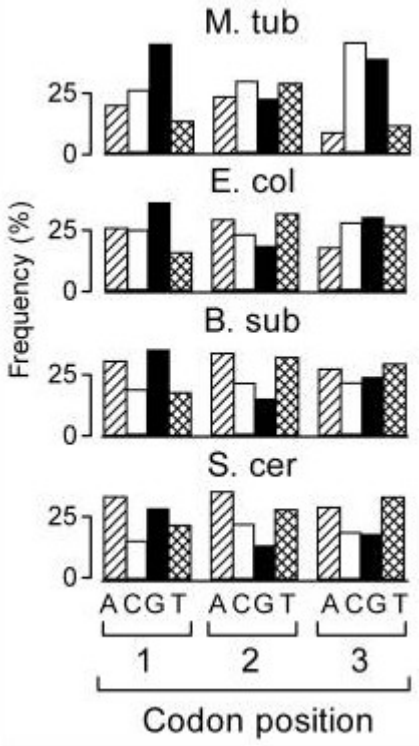


Trinity Demo



Prediction of coding potential

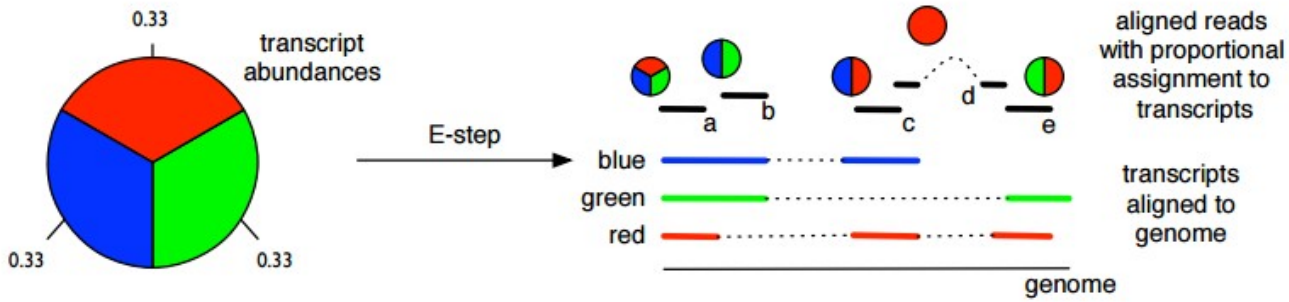
- Periodicity detection
  - Coding sequences have an inherent periodicity of three
  - Especially good on long coding sequences
  - Auto-correlation
    - Seeking the strongest response when shifted sequence is compared with original
    - Michel (1986), *J. Theor. Biol*. **120**, 223-236.
  - Fourier 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.



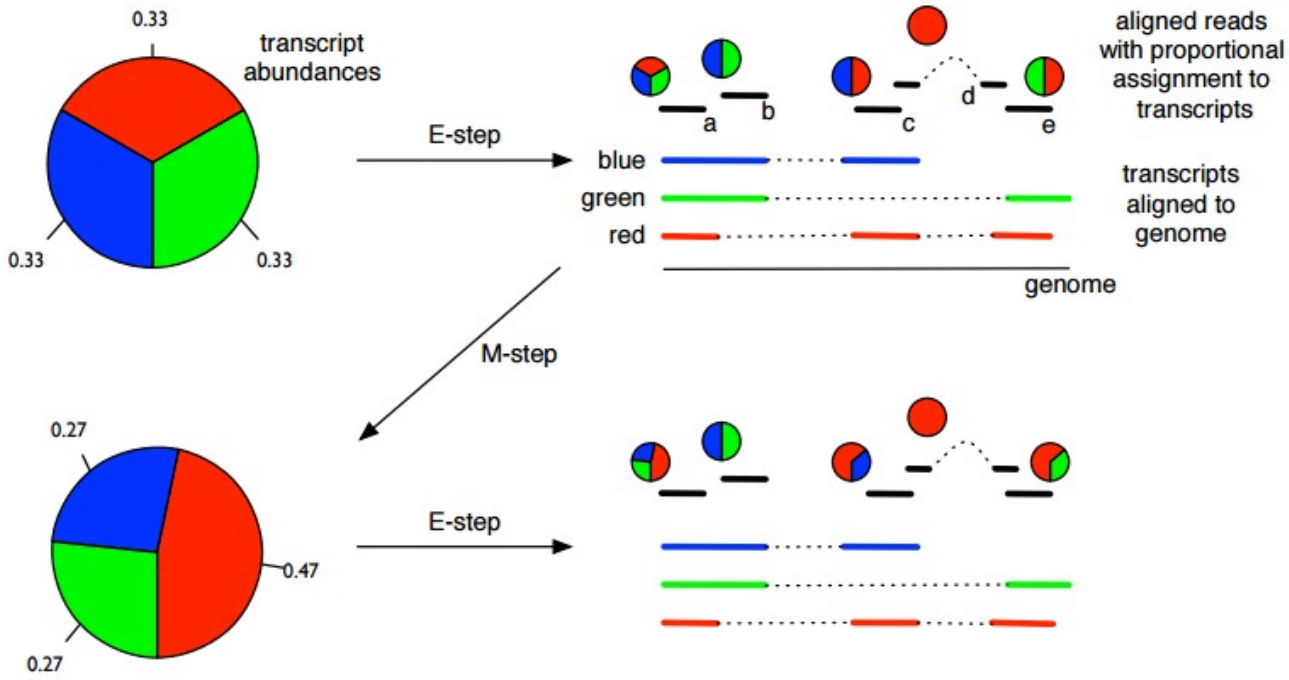
- http://www.fruitfly.org/GASP1/tutorial/presentation/sld082.htm

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

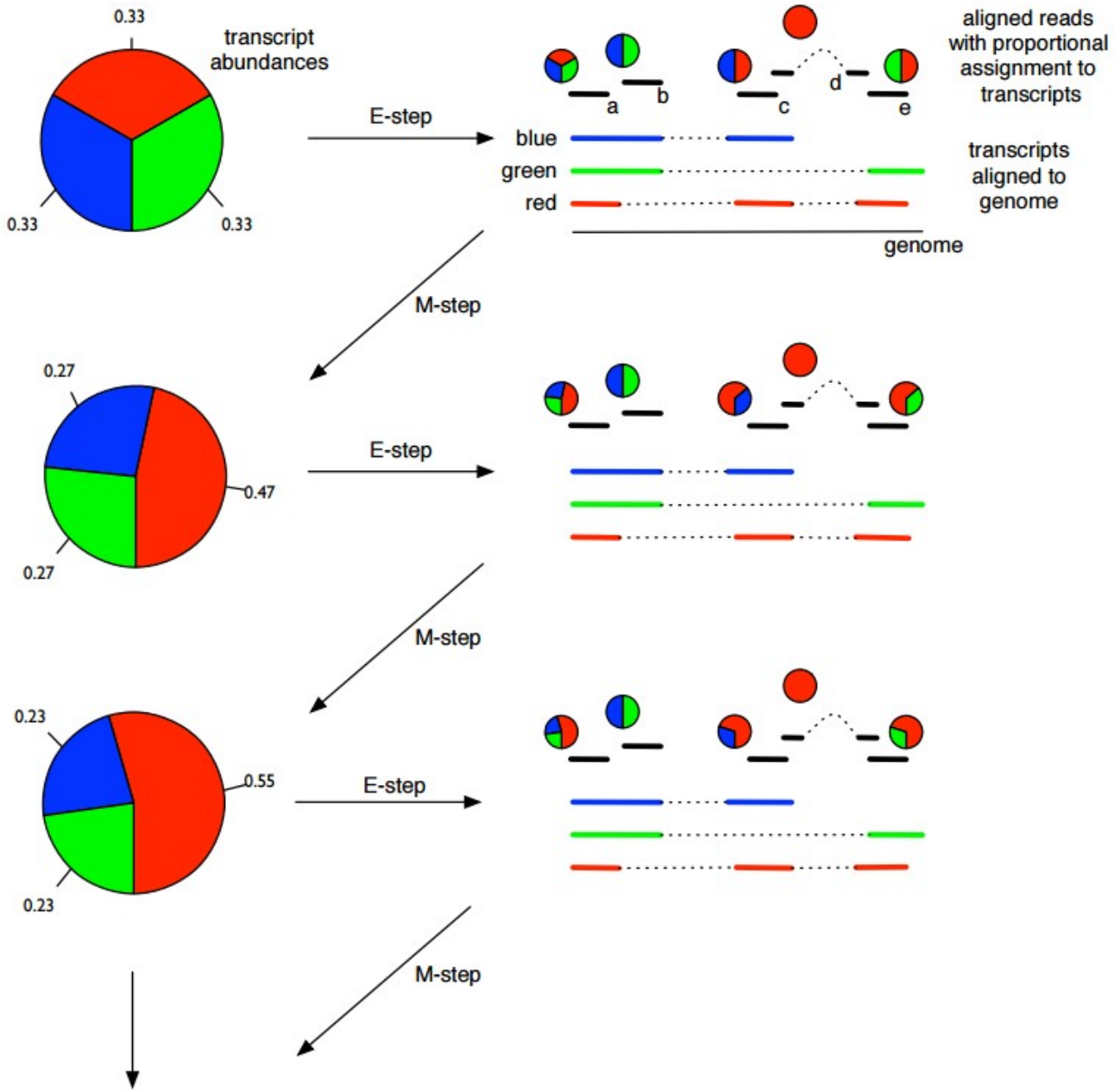
Expectation maximization used in gmap



Expectation maximization used in gmap

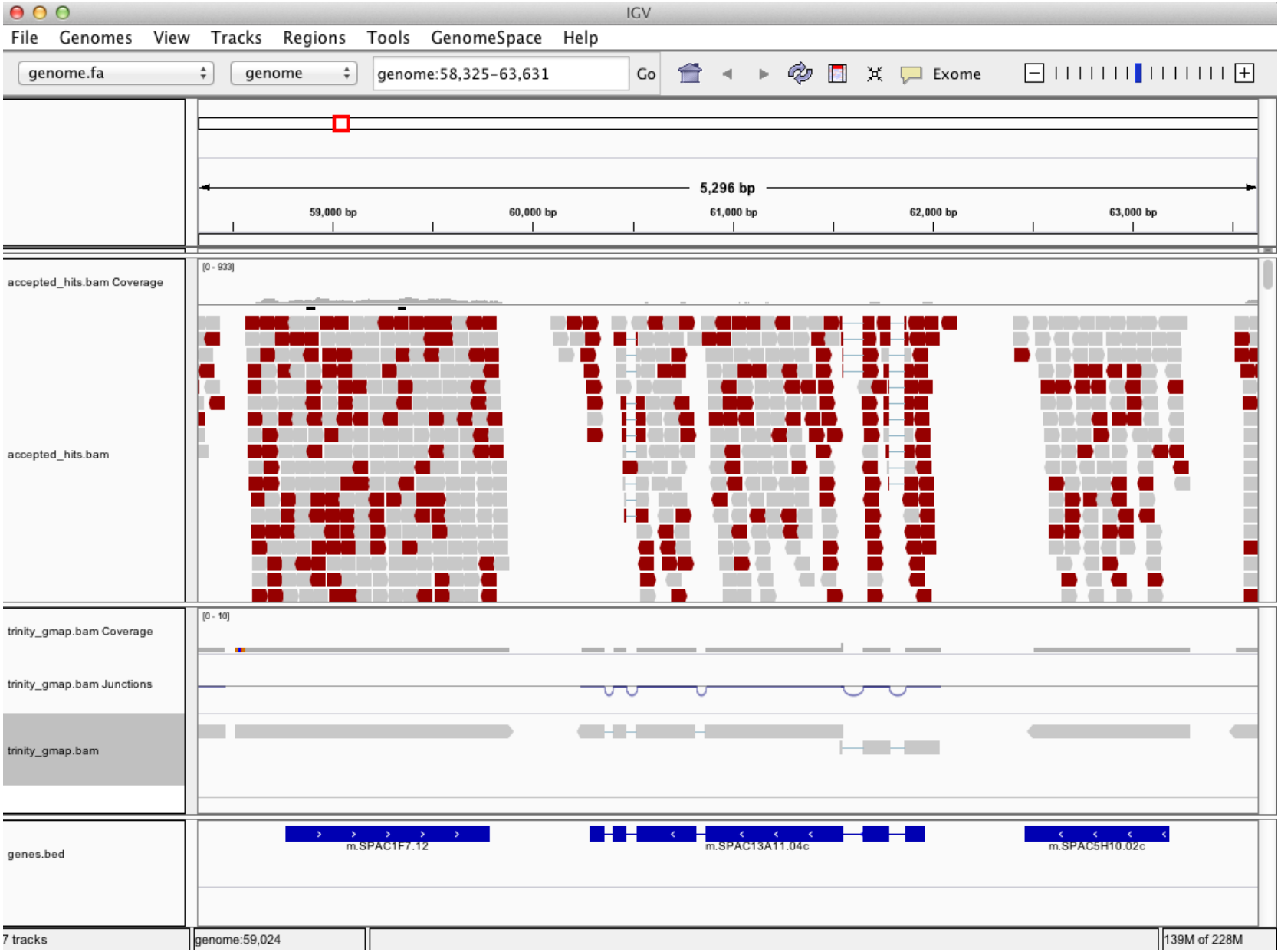


Expectation maximization used in gmap



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

(Trinity transcripts aligned using GMAP)

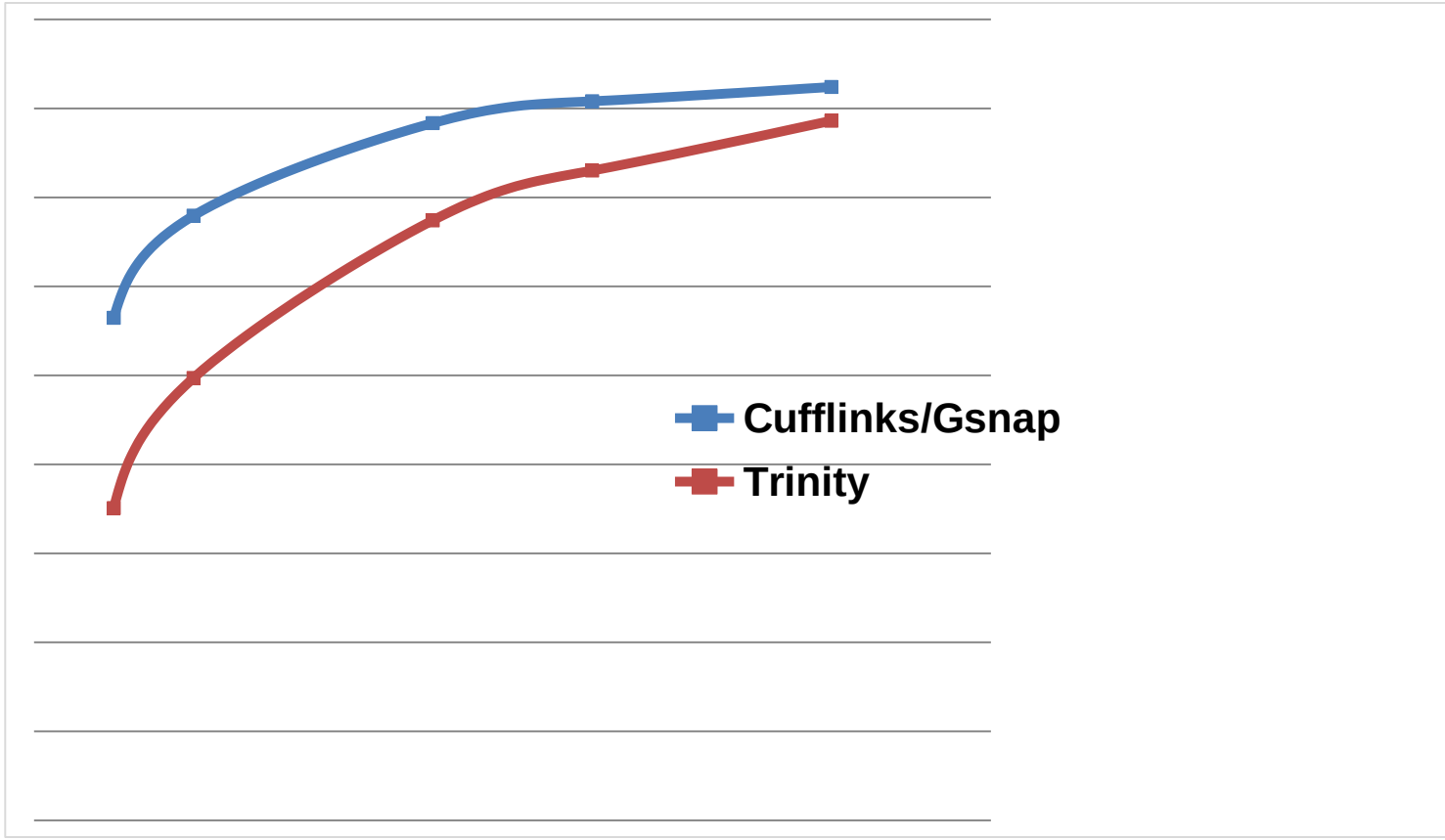


Trinity Demo

- 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

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

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

Single Cell Sequencing

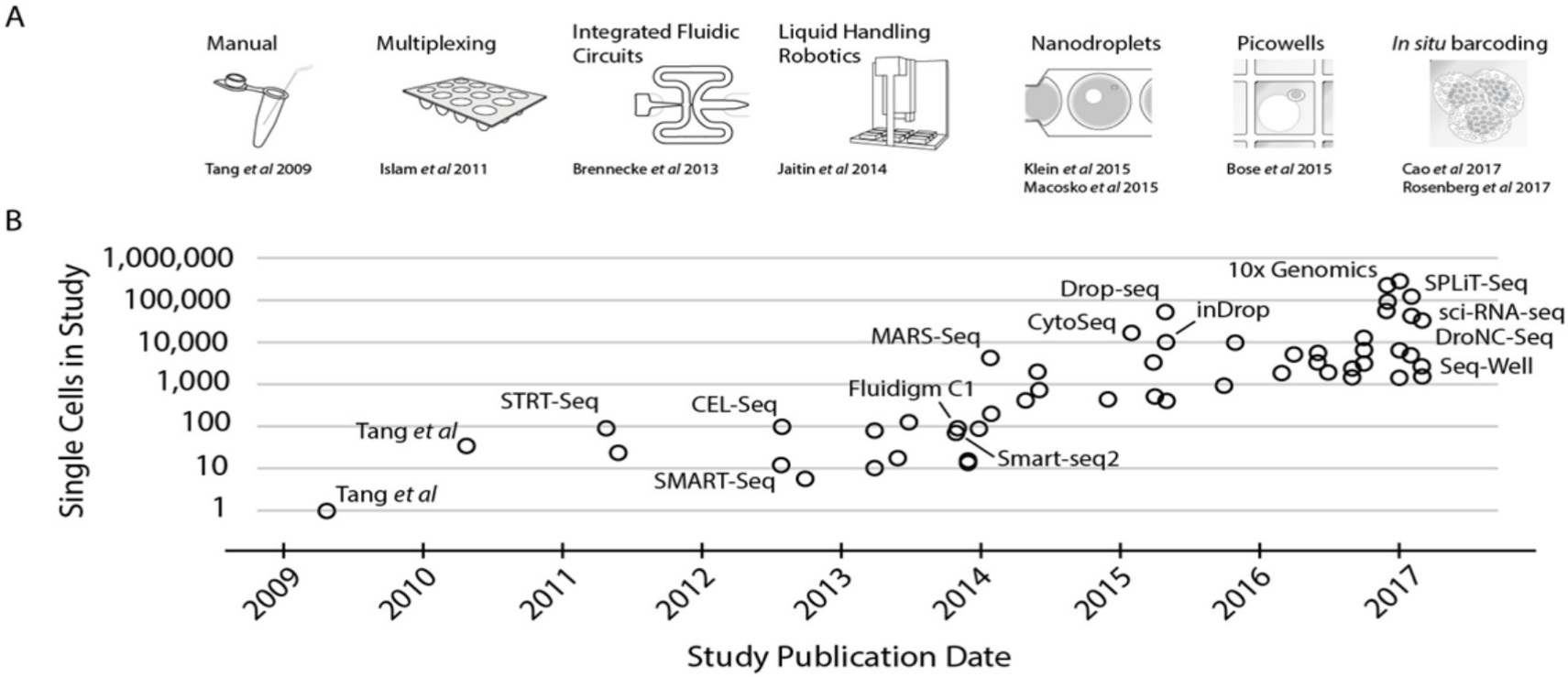
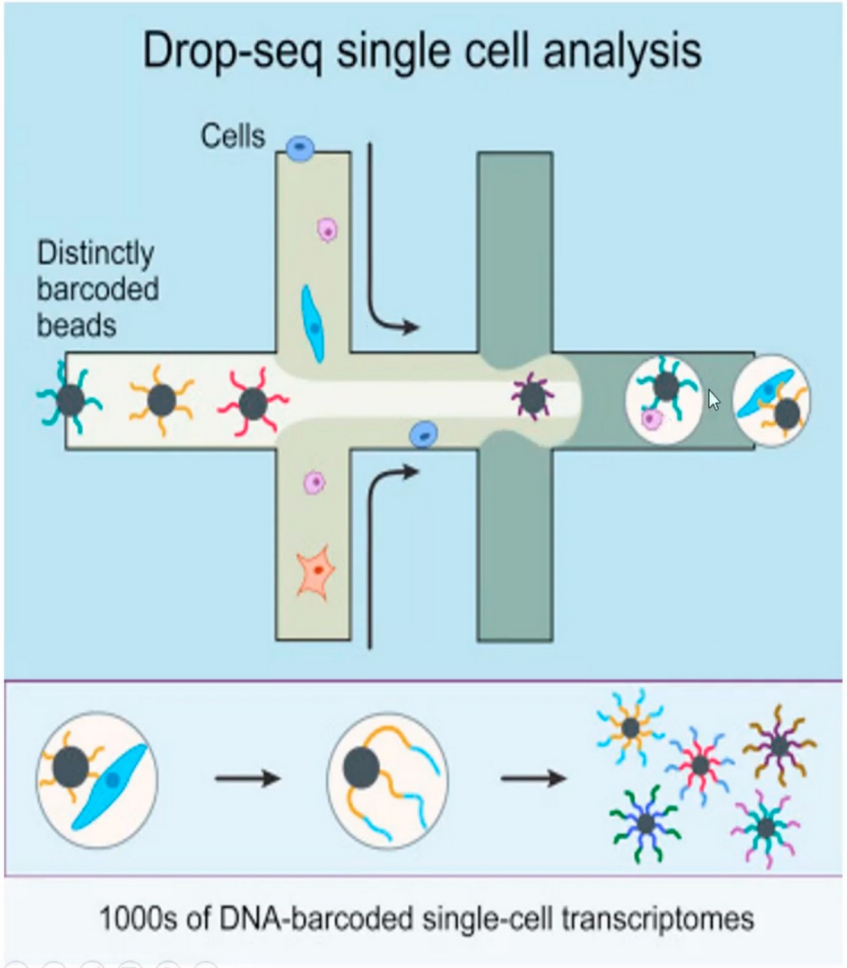


Figure 1: Scaling of scRNA-seq experiments (**A**) Key technologies allowing jumps in experimental scale. A jump to ~100 cells was enabled by sample multiplexing, a jump to ~1,000 cells by large scale studies using integrated fluidic circuits (IFCs), followed by a jump to several thousands using liquid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

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)

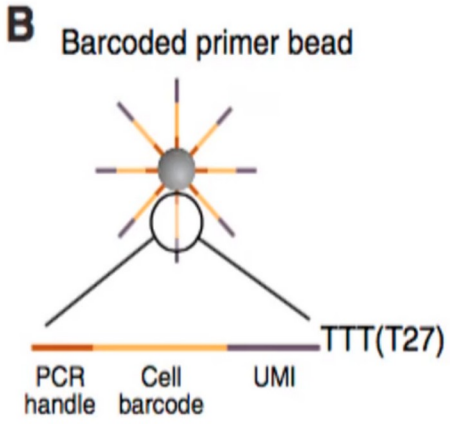


Figure by Macosko et al, Cell, 161:1202-1214, 2015

From reads to digital gene expression matrix (DGE)

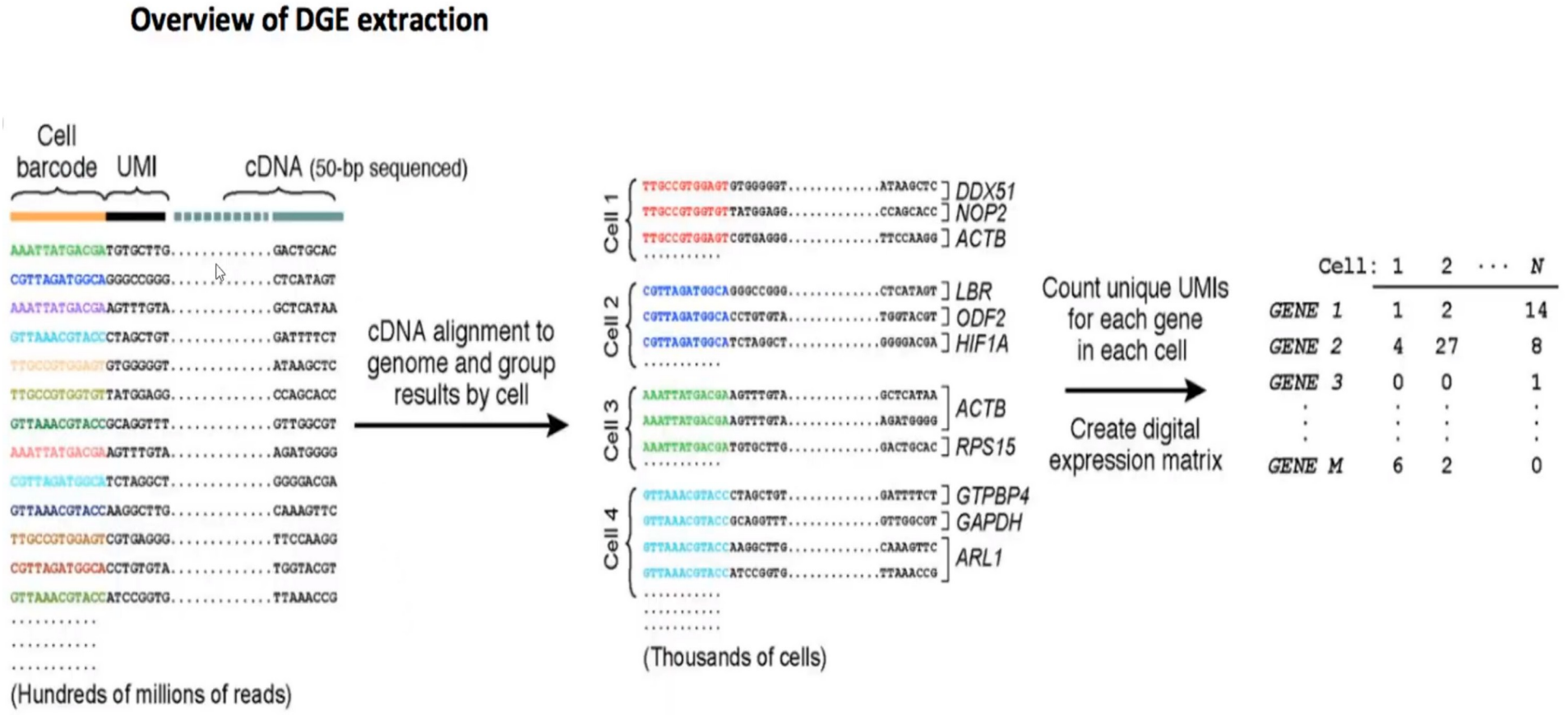
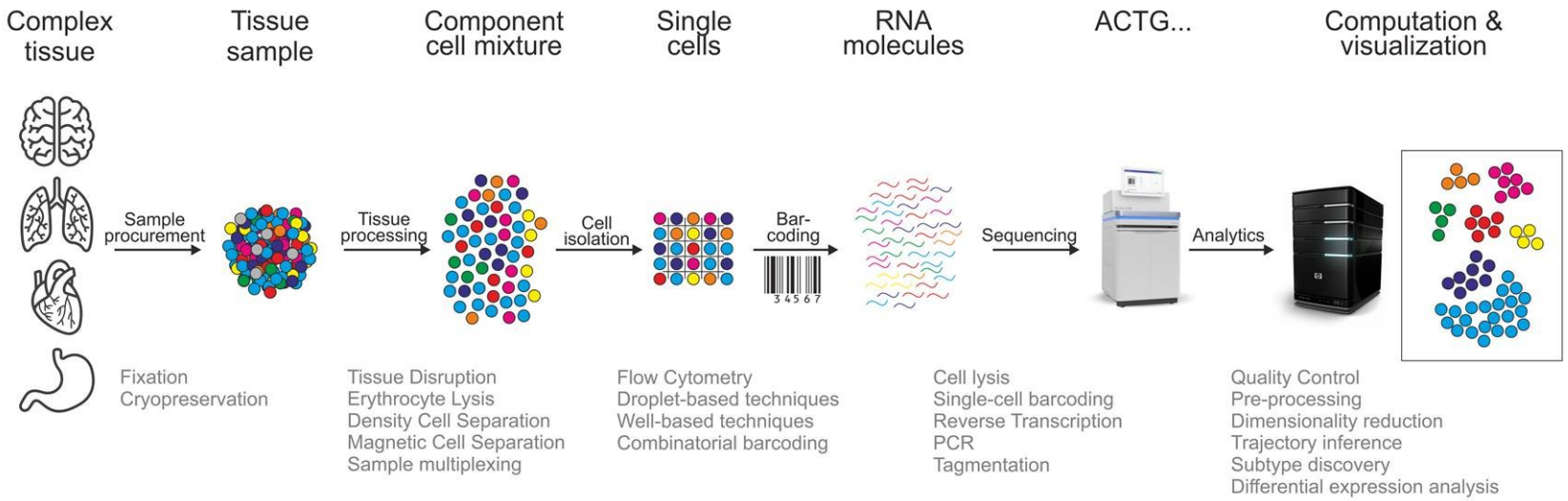


Figure by Macosko et al, Cell, 161:1202-1214, 2015

Single Cell RNA Sequencing and its main applications

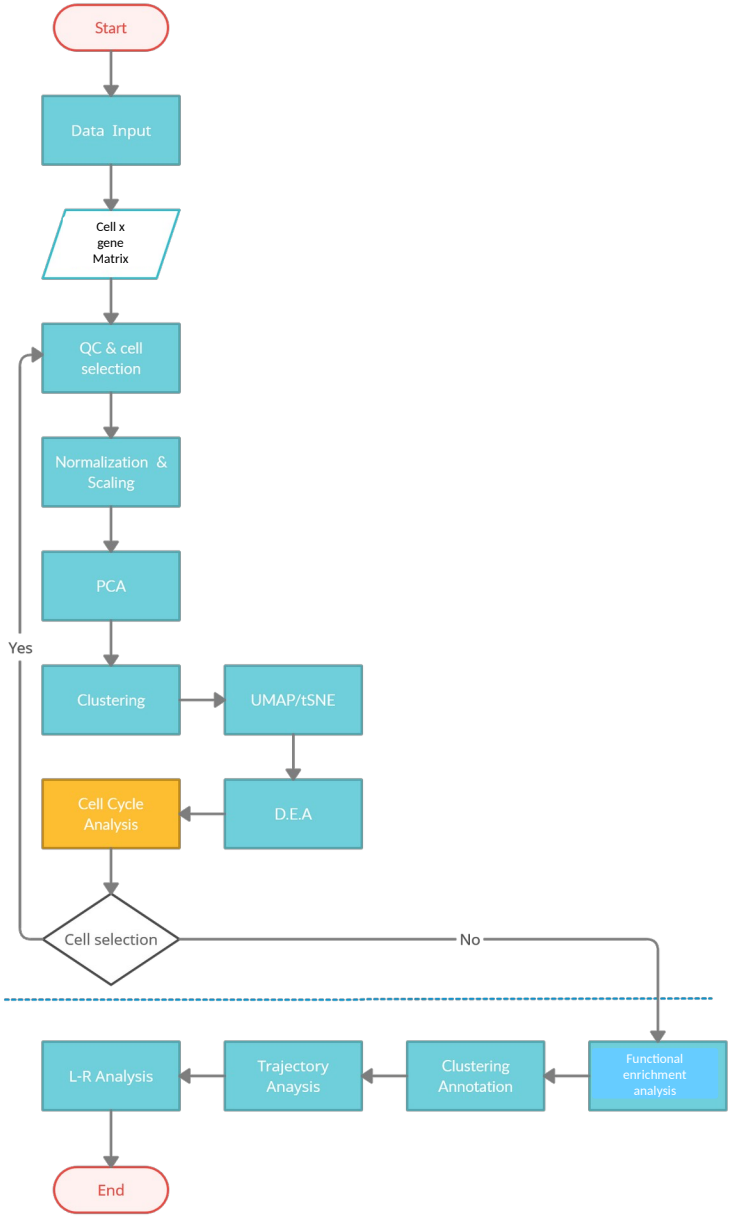
- Identification of new cell populations and subpopulations in complex tissues
- Studying gene dynamics in developmental studies
- Immune cell profiling
- Cancer research
- Personalized medicine



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)

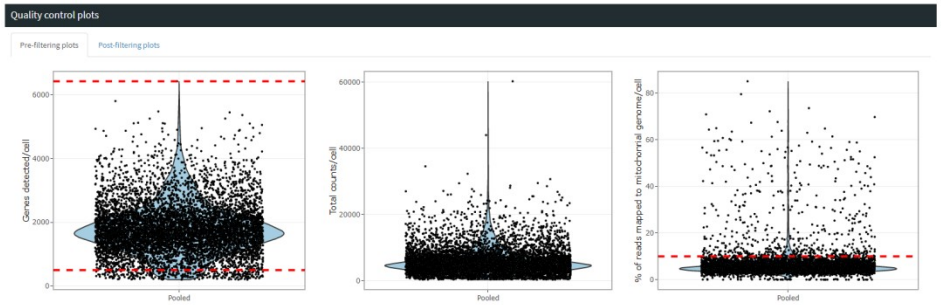
(slides by ITBI student Dimitra Panou)

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

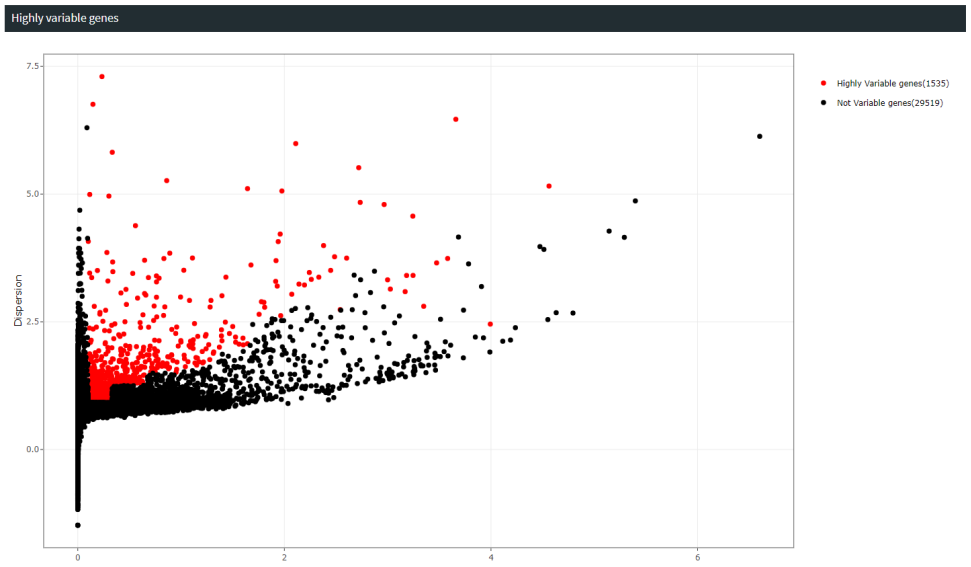


Red lines show the filtering points

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

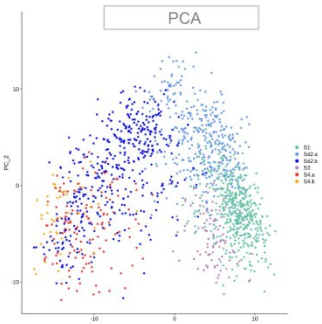


3. Dimensional reduction

PCA analysis

Detection of most informative principal components

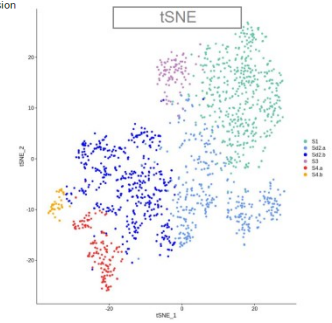
- ❑ Moving to PCA space can help reducing runtime of cell clustering
- ❑ May fail to capture local patterns in scRNA data



Non linear methods

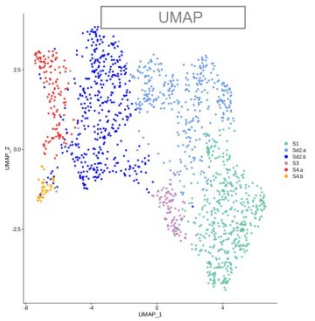
t-Distributed Stochastic Neighbor Embedding (t-SNE)

- ❑ 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

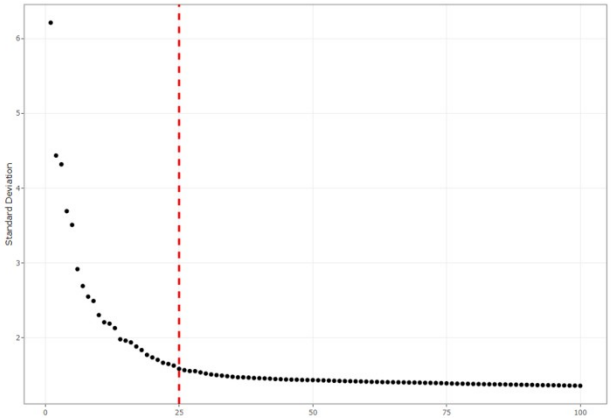


scRNAseq pipeline

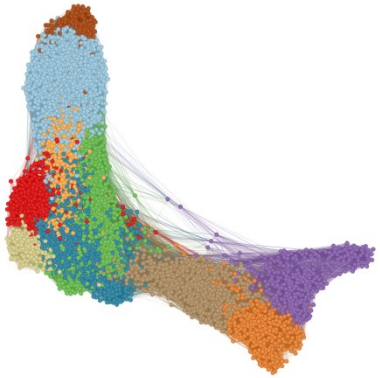
4. <u>Clustering analysis</u>

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

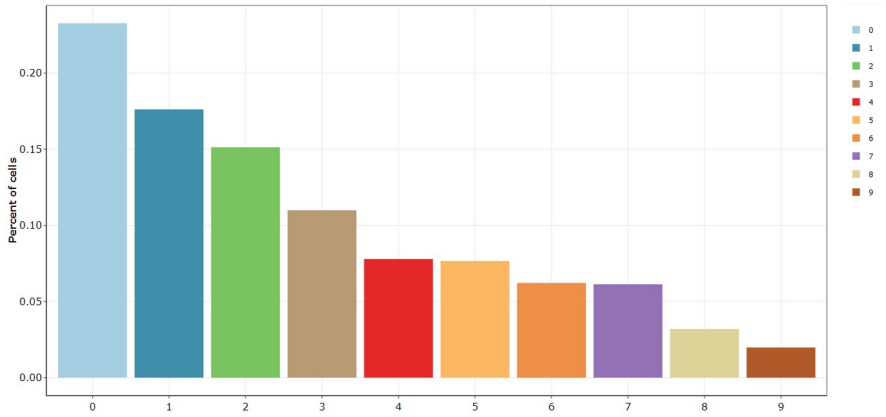
PCA



sNN 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

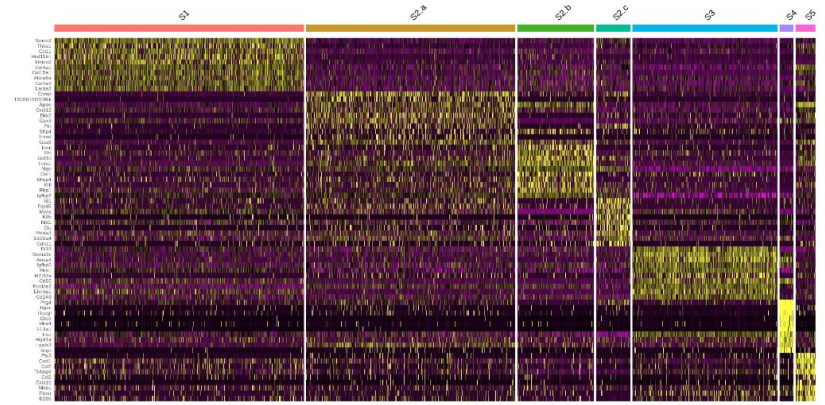
- Identify marker genes per cluster
- Those genes can distinguish one cluster from the rest
- High average expression in cells of the cluster, low in the other cells

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

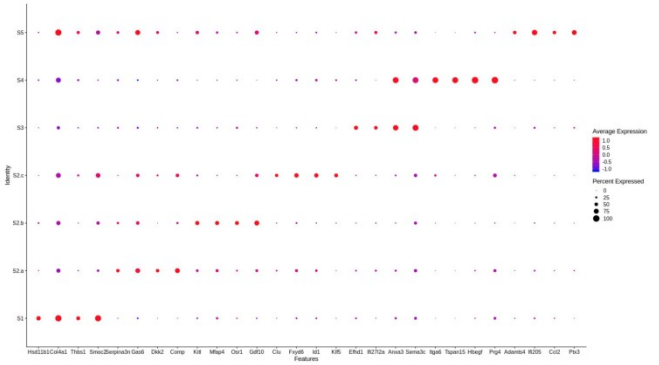
Inspection of top marker genes

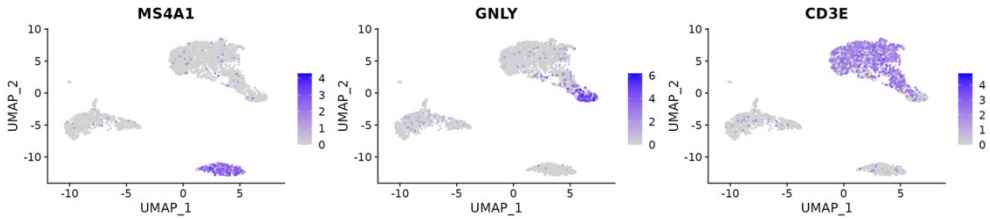
- ❑ Feature plots in UMAP space
- ❑ Color denotes normalized expression

Differential expression analysis



Marker genes for each cluster





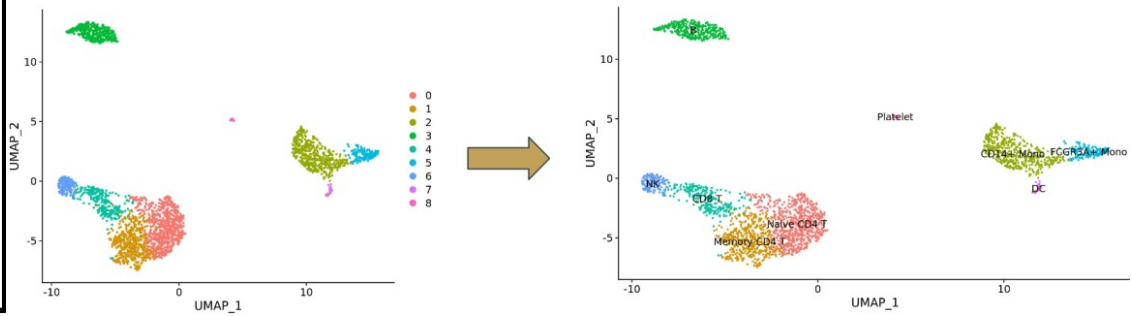
* MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data

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

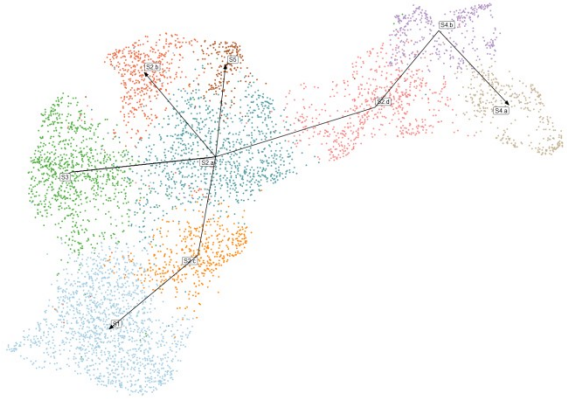


Minimum spanning tree

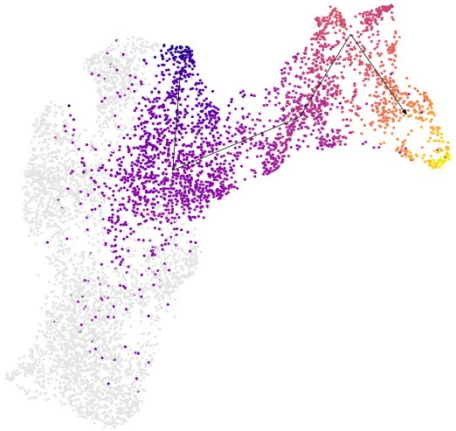
7. Trajectory-Pseudotime analysis

- ❑ 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/