EMBL-EBI

Cinvestav

De novo approach to RNA-seq

Introduction to RNA-seq and functional interpretation
EMBL-EBI Training

February 21st, 2023

Selene L. Fernández-Valverde
regRNAlab.github.io
@SelFdz

Learning objectives

In this session we learn:

- What is a *de novo* transcriptome assembly and why it can be useful or just necessary
- A general overview on how Trinity works
- To use Trinity to assemble *de novo* high-throughput data

What is a *de novo* transcriptome assembly?

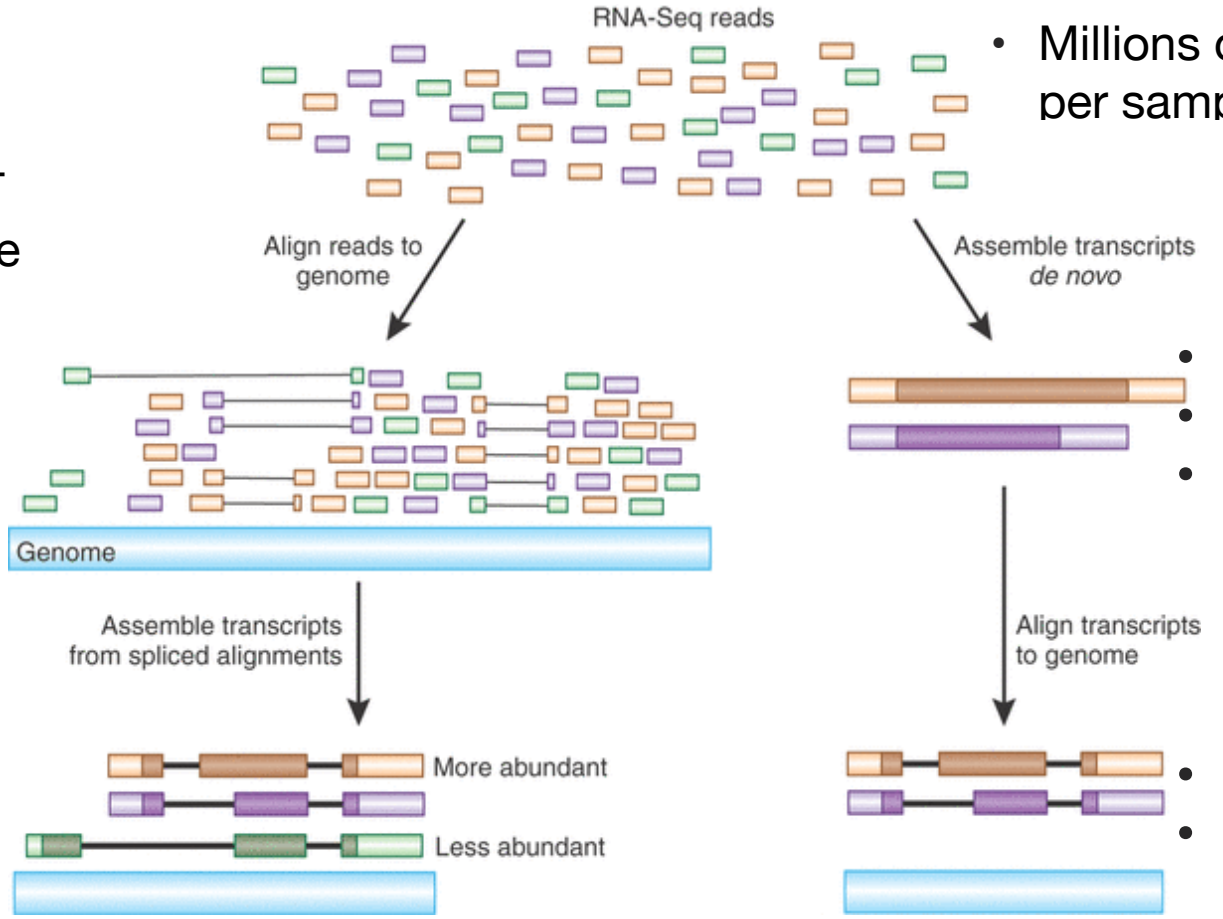
A *de novo* transcriptome assembly takes RNA-Seq reads and assembles them into complete transcripts (including isoforms) **without the help of a genome as a reference**.

Useful for organisms that do not have a high quality genome and/or are too gene dense, also for discovering transcripts that might not be in the references (e.g. chimeric transcripts, viruses and otehr pathogens). Also useful for complex samples with rearranged genomes (e.g. cancer)

How do we assemble a transcriptome?

- Millions of reads per sample !

- HISAT
- Bowtie
- BWA



- Trinity
- Oases de Novo
- trans-ABySS

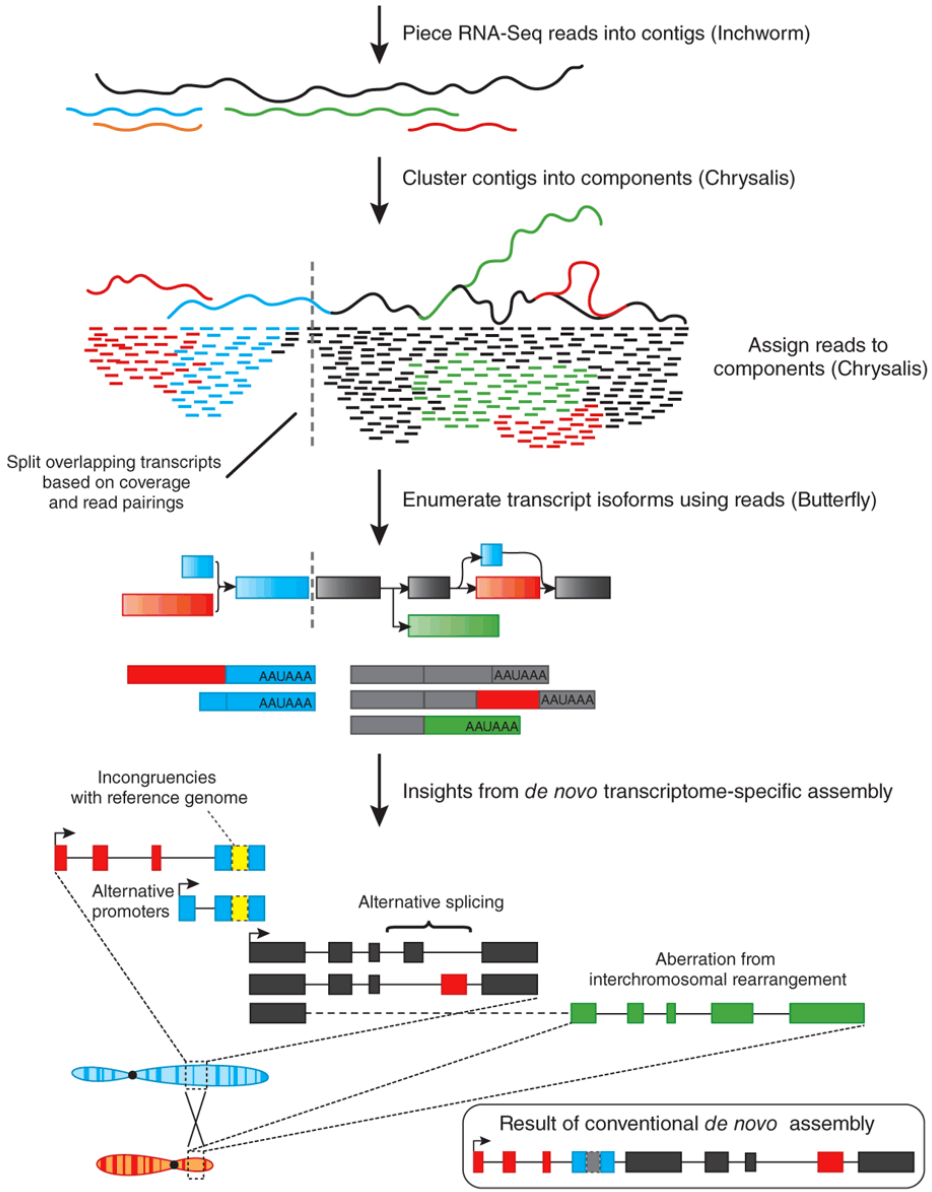
- StringTie

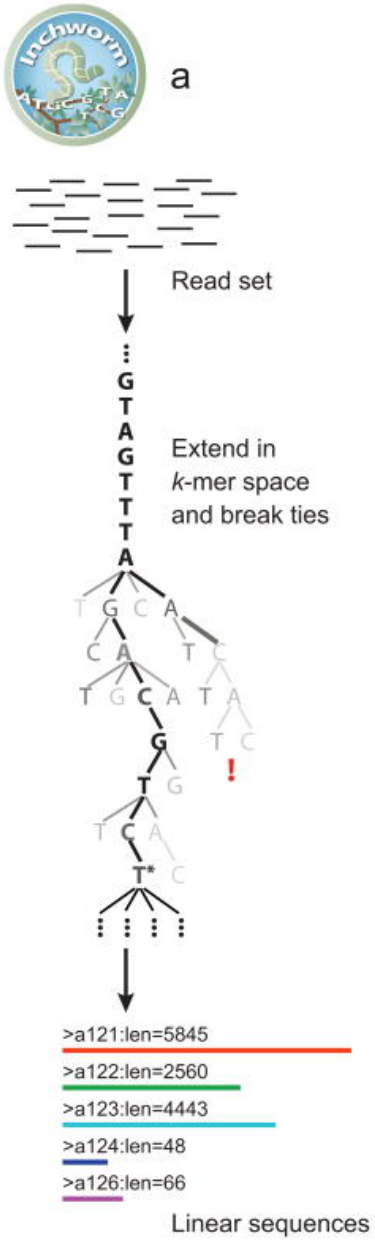
- Blat
- Exonerate

Programs to assemble transcriptomes *de novo*

- **Trinity** (https://github.com/trinityrnaseq/trinityrnaseq)
- Trans-ABySS (https://github.com/bcgsc/transabyss)
- SOAPdenovo-Trans (http://soap.genomics.org.cn/SOAPdenovo-Trans.html)
- Velvet/Oases (https://www.ebi.ac.uk/~zerbino/oases/)

It is a method for the reconstruction of transcriptomes based on RNA-Seq data. Trinity combines 3 main modules



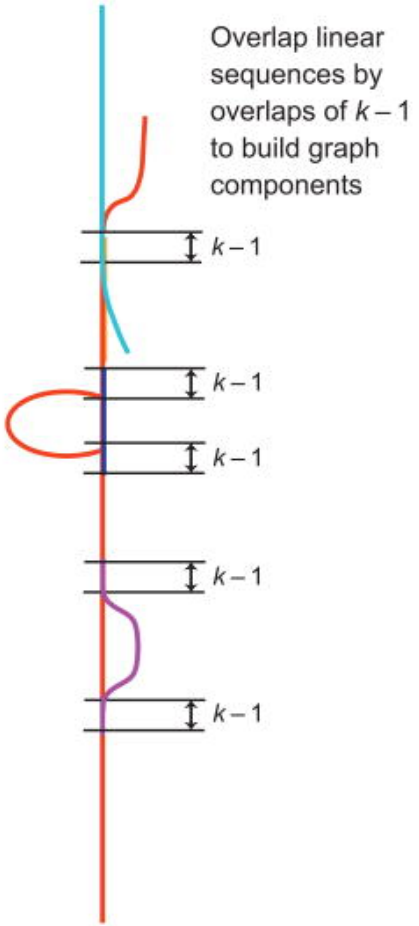


Inchworm

- Create a catalog of overlapping kmers (25-mers)
- Save these kmers and their sequence (do not create graph yet)
- Take the most abundant kmer and use it as a seed
- Extends the 3 'end guided by coverage
- If there is a tie, recursively look for options to identify numbers that provide the most cumulative coverage
- Extensions are made until there are no more compatible numbers
- The 5 'end is then extended
- Reports the longest contig and removes used catalogs from the catalog
- Restart this process with a new seed
- Stop when there are no more kmers in the catalog

Grabherr MG, et al. Nature Biotechnology, 2011

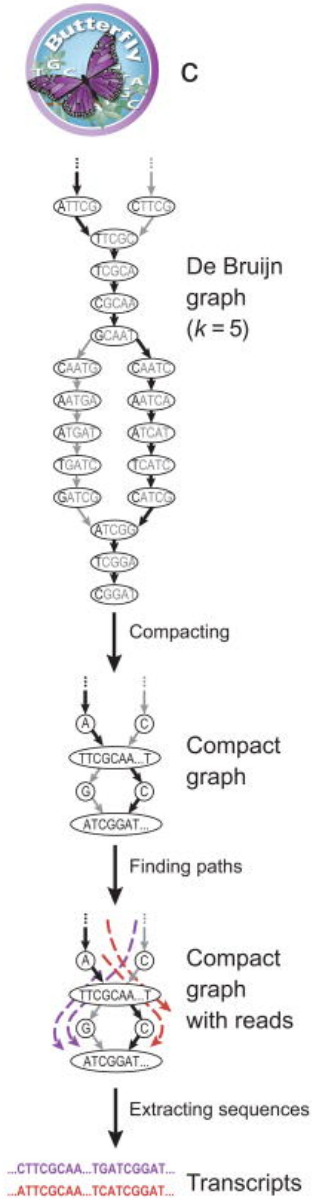
Chrysalis



- Inchworm cannot rebuild isoforms
- Take the contigs generated by inchworm that do not have full compatible numbers at their ends
- Explore partial kmers (k-1) to regroup related contigs
- If there is support Chrysalis unites contigs generated by Inchworm
- Finally build a graph of Bruijn for each group
- The graph branches into variation sites
- This results in one graph per gene

Grabherr MG, et al. Nature Biotechnology, 2011

Butterfly



- Works on each independent graph in parallel (processing many thousands of small graphs)
- It collapses unbranched structures of the graph
- Embed the original readings within the graph by tracking the path of each reading and verifying the congruence of readings in pairs
- Issues complete assembled transcripts including isoforms and paralogs

Grabherr MG, et al. Nature Biotechnology, 2011

General recommendations

- Absolutely essential to have **stranded** RNA-seq to avoid transcript chimeras
- Need a reasonable amount of memory (RAM)
- Samples should be pooled within the software to assemble a single *de novo* reference transcriptome that can be used for downstream analyses
- Transcripts can be mapped back to the reference genome if it's quality is subpart, they can even be used to stitch genomic contigs together!

Practical - assembling a transcriptome using Trinity

https://liz-fernandez.github.io/EBI-Intro-RNA-seq-2023/01-assembly_denovo.html