*INF-BIOx121 2017*

# RNA-seq

## differential expression analysis

Arvind Sundaram
Sep 18-20, 2017

*RNA-seq analysis*

# Transcriptome

Arvind Sundaram

Sep 18, 2017

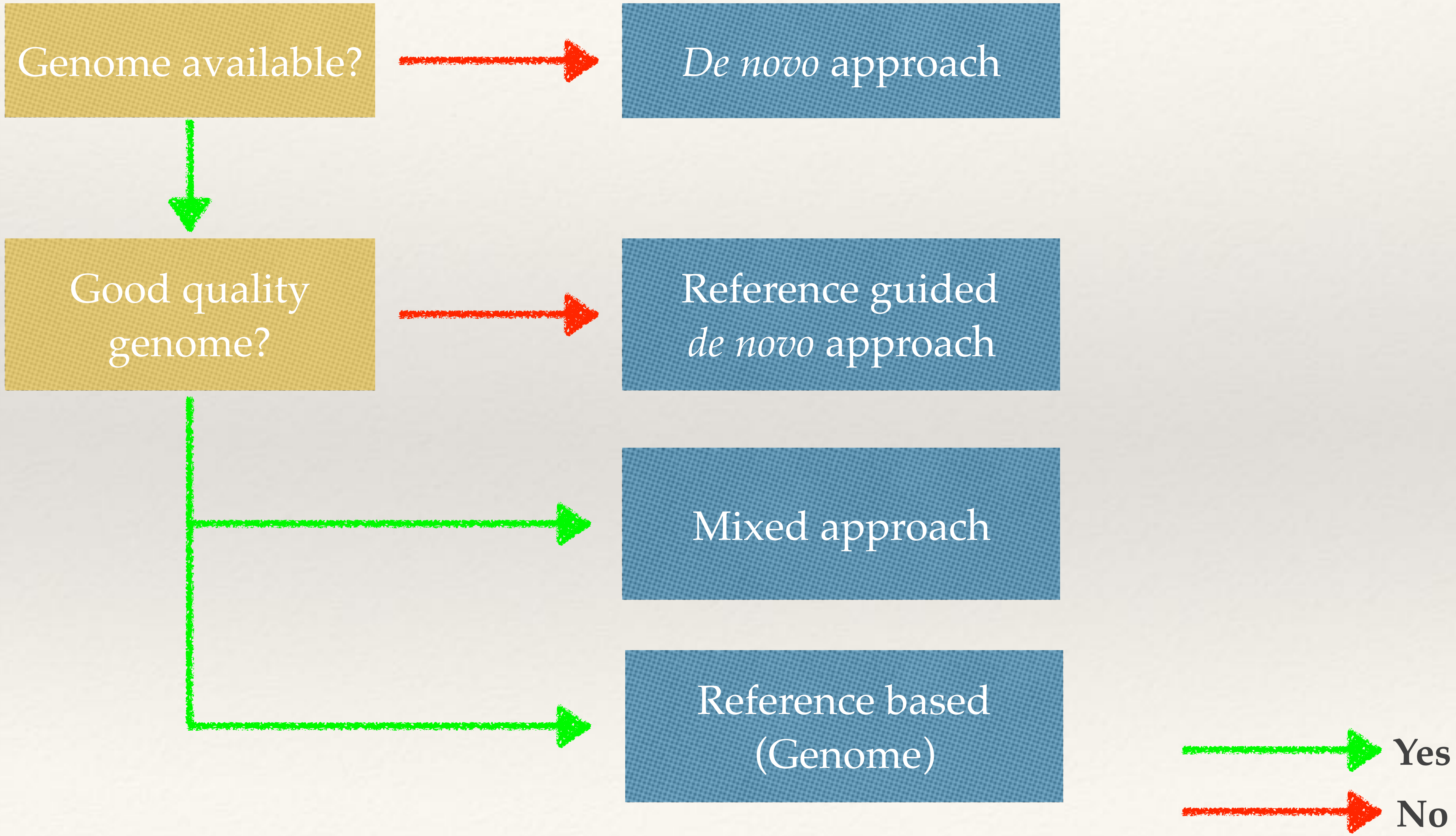
# Transcriptome

- Transcriptome is the total set of transcripts in an organism
- Transcriptome changes across cell types and environmental conditions

So….

- Transcriptome is a set of (all) RNA molecules in one cell or a population of cells in a given moment
- ‘Constructing’ a global transcriptome aims to capture all possible transcripts found across all cell and tissue types.

# Assembling a transcriptome



Genome available?

*De novo* approach

Good quality genome?

Reference guided *de novo* approach

Mixed approach

Reference based (Genome)

Yes

No

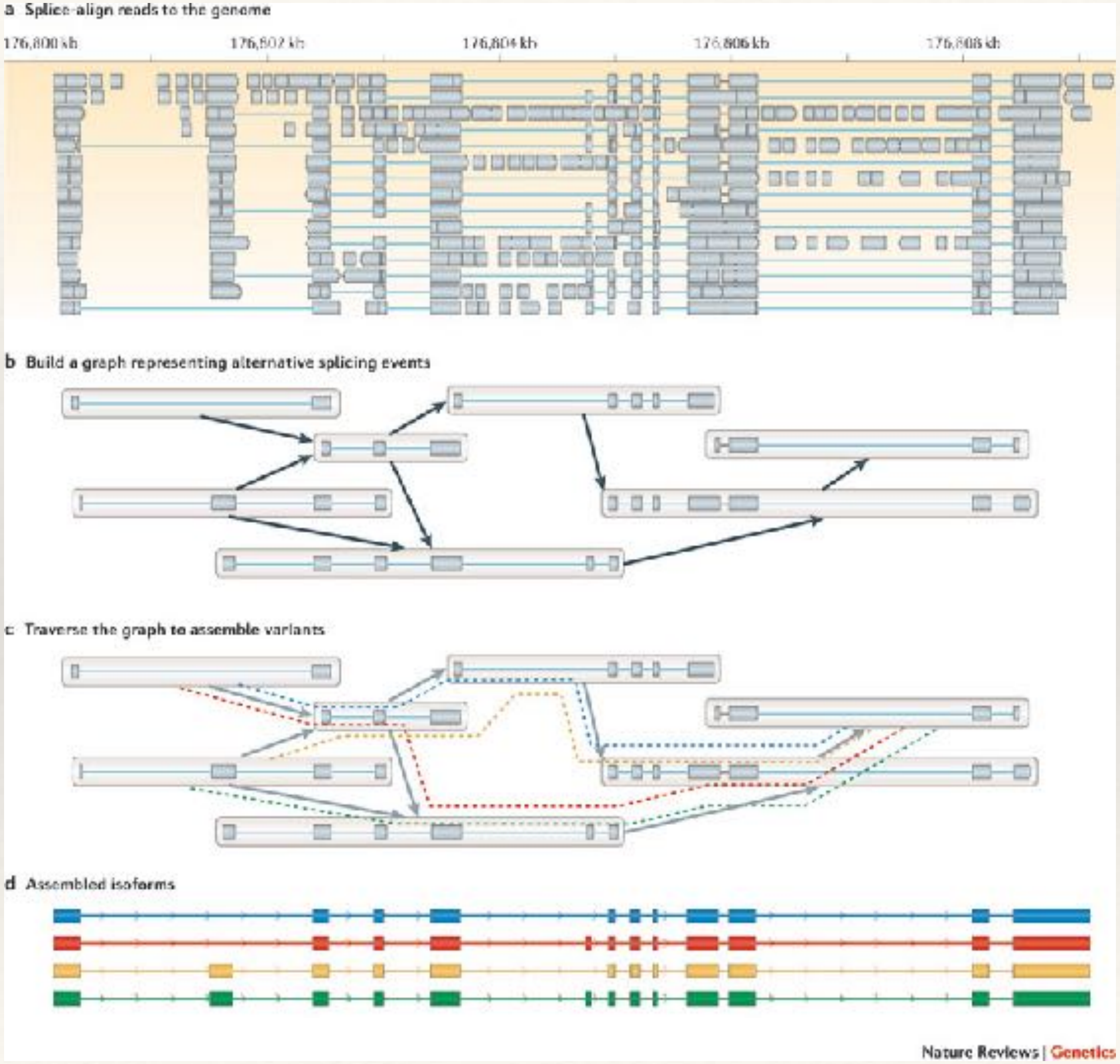
# Reference based transcriptome

- Requires good quality (draft) genome
- Splice-aware aligner
- Improves on existing knowledge

Fast

less CPU

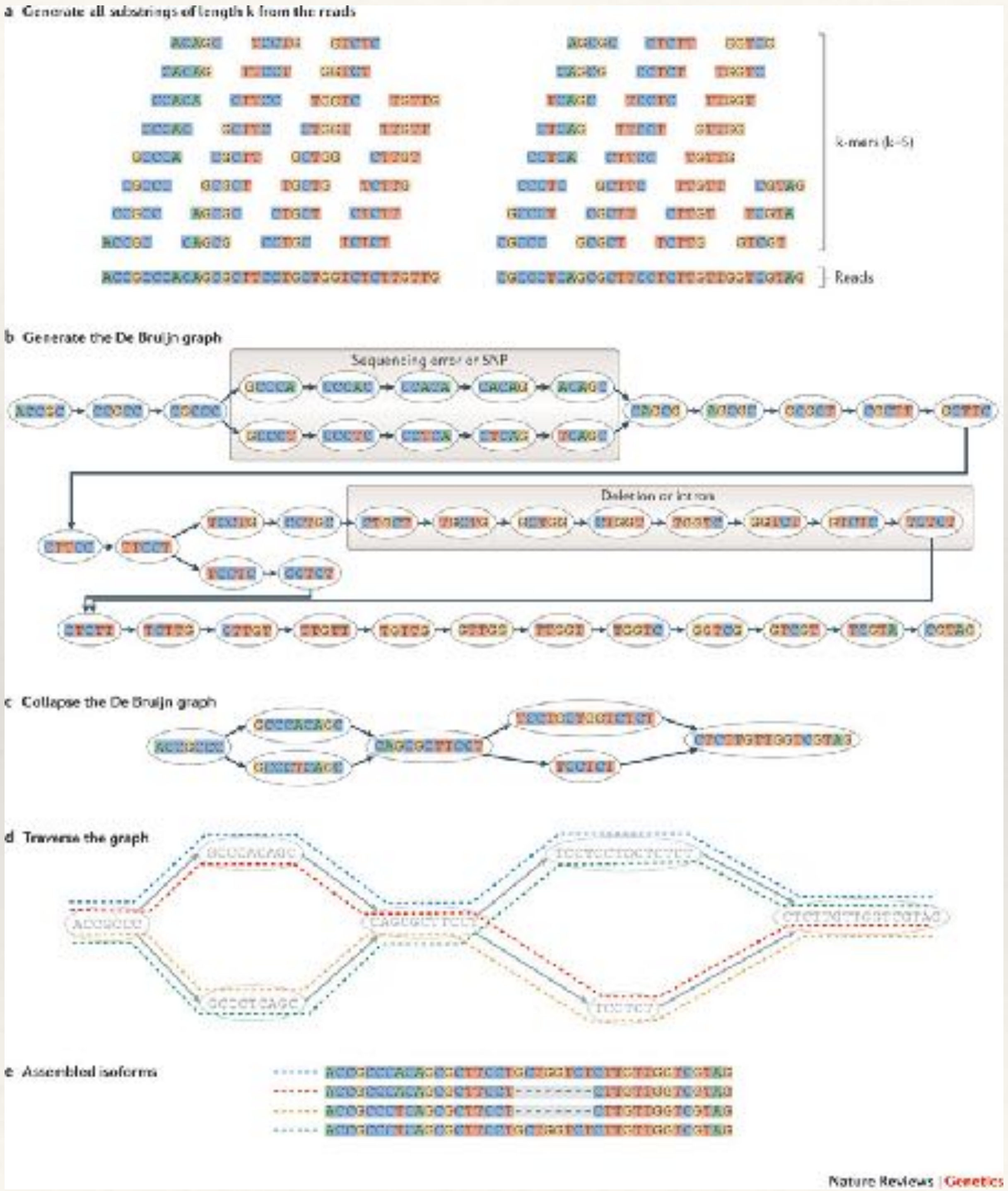
prior knowledge


a Splice-align reads to the genome
176,800 kb
176,802 kb
176,804 kb
176,806 kb
176,808 kb
b Build a graph representing alternative splicing events
c Traverse the graph to assemble variants
d Assembled isoforms
Nature Reviews | Genetics

Next-generation transcriptome assembly
https://doi.org/10.1038/nrg3068

# *de novo* transcriptome

- Requires high coverage sequence data
- Transcript detection based on coverage
- Multiple gene-copies is difficult to resolve


a Generate all substrings of length k from the reads
k-mers (k=5)
Reads
b Generate the De Bruijn graph
Sequencing error or SNP
Deletion or intron
c Collapse the De Bruijn graph
d Traverse the graph
e Assembled isoforms
Nature Reviews | Genetics

Slow

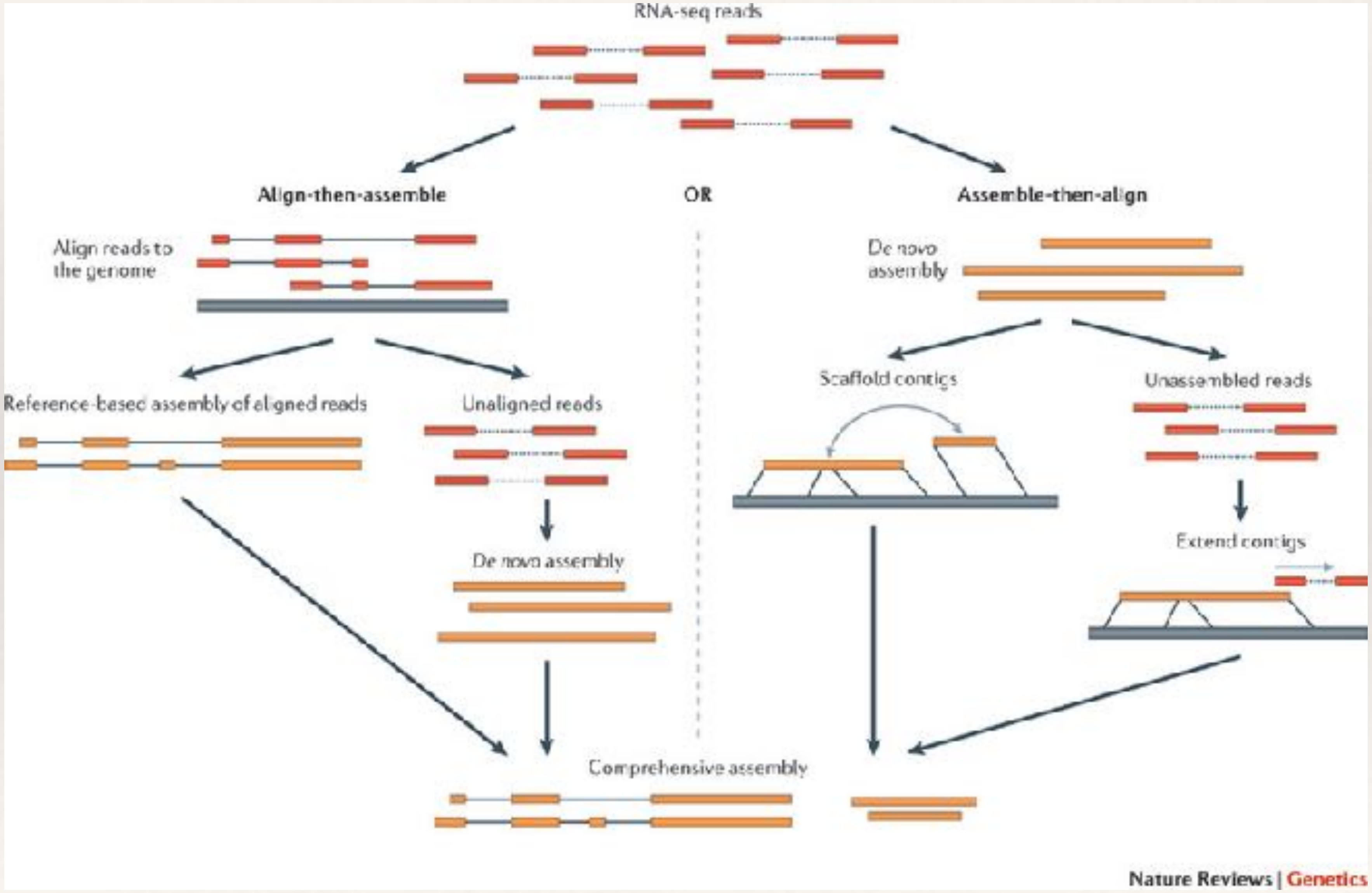
more CPU

No prior knowledge

Next-generation transcriptome assembly
https://doi.org/10.1038/nrg3068

# Mixed approach

- Identifies novel transcripts
- Polyploidy species



Varies

mixed CPU

prior knowledge

Next-generation transcriptome assembly
https://doi.org/10.1038/nrg3068

# *de novo* transcriptome

- One will choose *de novo* / mixed approach due to many reasons
    - Non-model species with less genomic resources
    - Improve gene annotation
    - Genes of interest are not well annotated

https://github.com/trinityrnaseq/trinityrnaseq/wiki

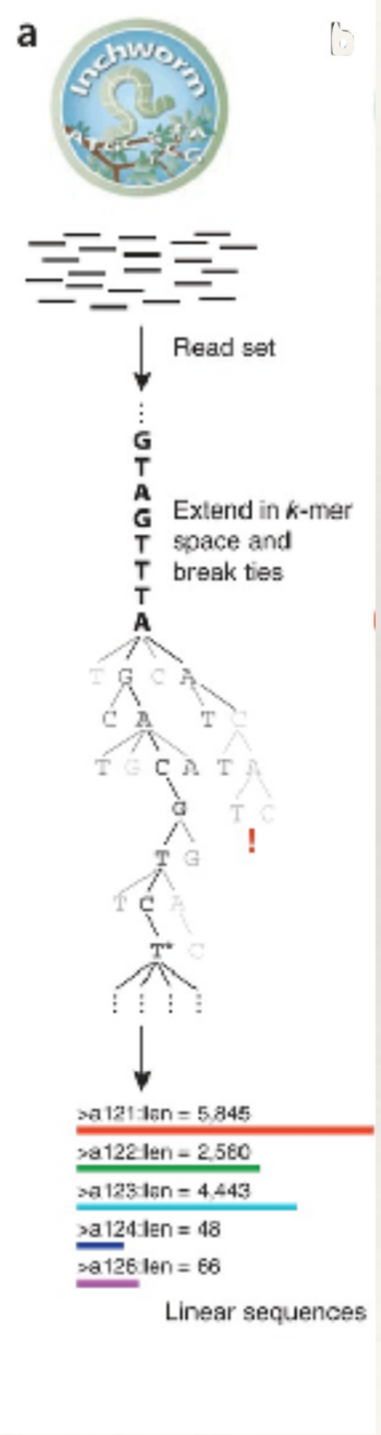
# Trinity assembler

- Trinity is the best single parameter *de novo* RNA assembly pipeline available
- Good on splice variants, full length transcripts and resolution of lowly expressed transcripts
- Contains tools to help with visualisations

https://github.com/trinityrnaseq/trinityrnaseq/wiki

# Trinity pipeline -Inchworm

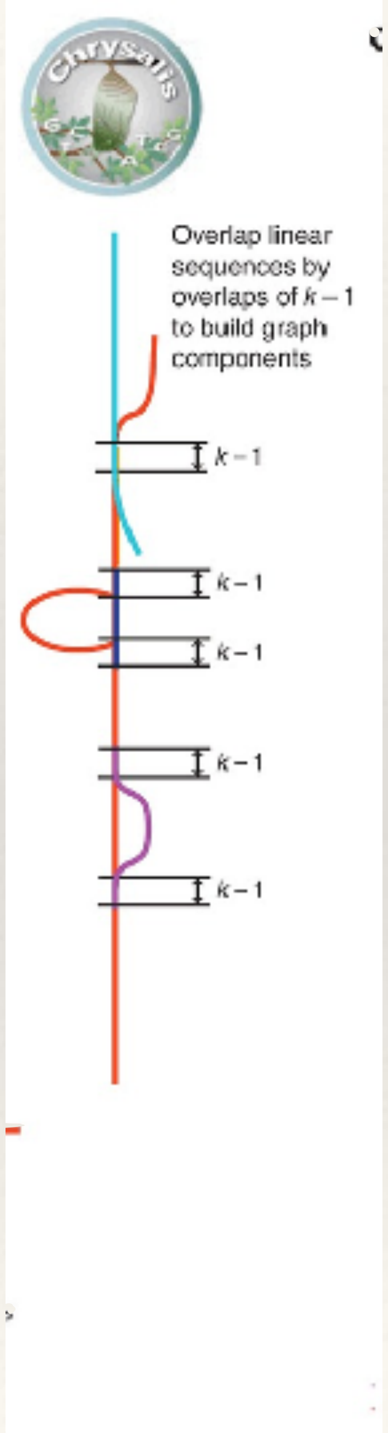
- Inchworm assembles the data by greedily searching for paths in a k-mer graph, resulting in a collection of linear contigs, with each k-mer present only once in the contigs
- Inchworm does not capture the full complexity of the transcriptome; for example, only one alternatively spliced variant can be reported at full length per locus, with partial sequences reported for unique regions of any alternatively spliced transcripts.

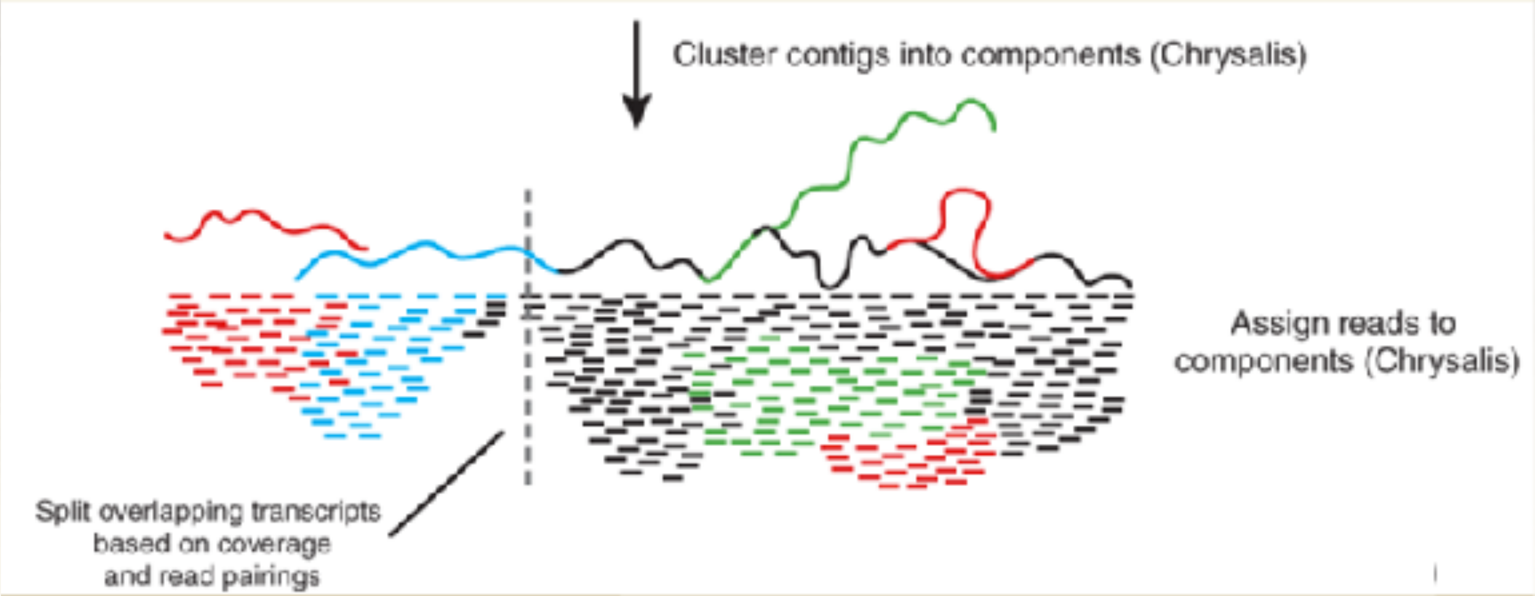


https://github.com/trinityrnaseq/trinityrnaseq/wiki

# Trinity pipeline -Chrysalis

- Chrysalis clusters minimally-overlapping Inchworm contigs into sets of connected components, and constructs complete de Bruijn graphs for each component

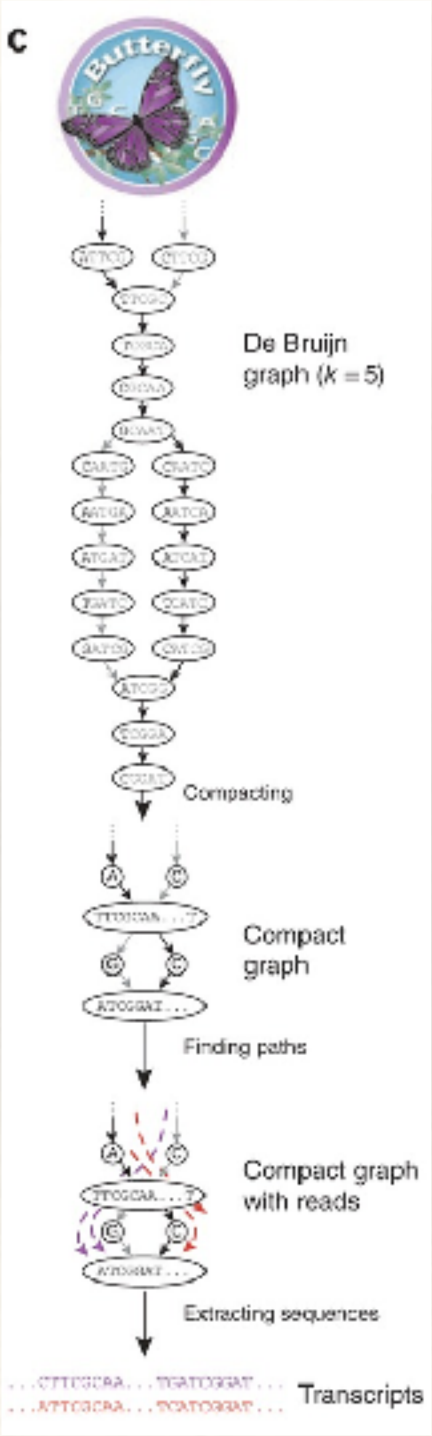

Chrysalis
Overlap linear sequences by overlaps of k − 1 to build graph components
k − 1
k − 1
k − 1
k − 1
k − 1

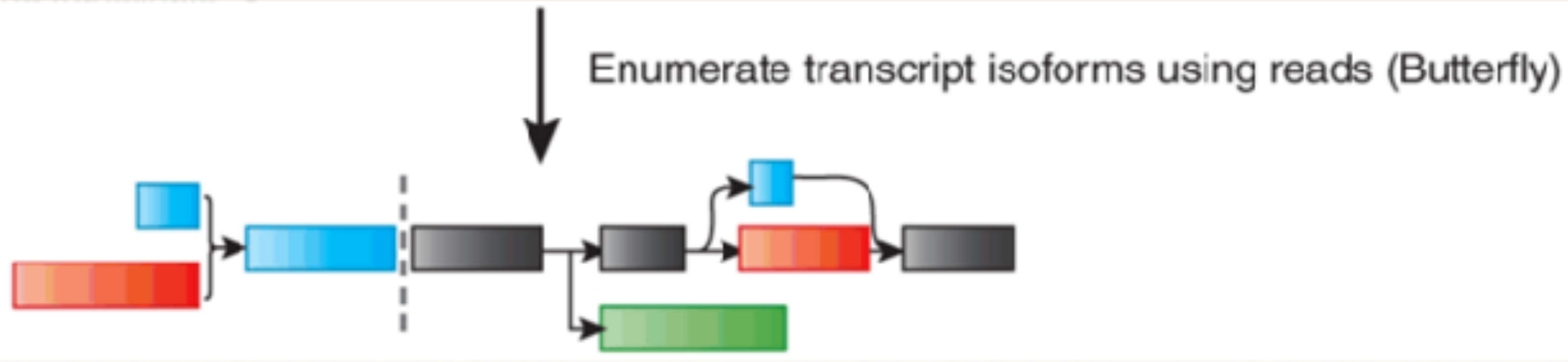

Cluster contigs into components (Chrysalis)
Assign reads to components (Chrysalis)
Split overlapping transcripts based on coverage and read pairings

https://github.com/trinityrnaseq/trinityrnaseq/wiki

# Trinity pipeline - Butterfly

- Butterfly reconstructs plausible full-length, linear transcripts by reconciling the individual de Bruijn graphs generated by Chrysalis with the original reads and paired-ends.

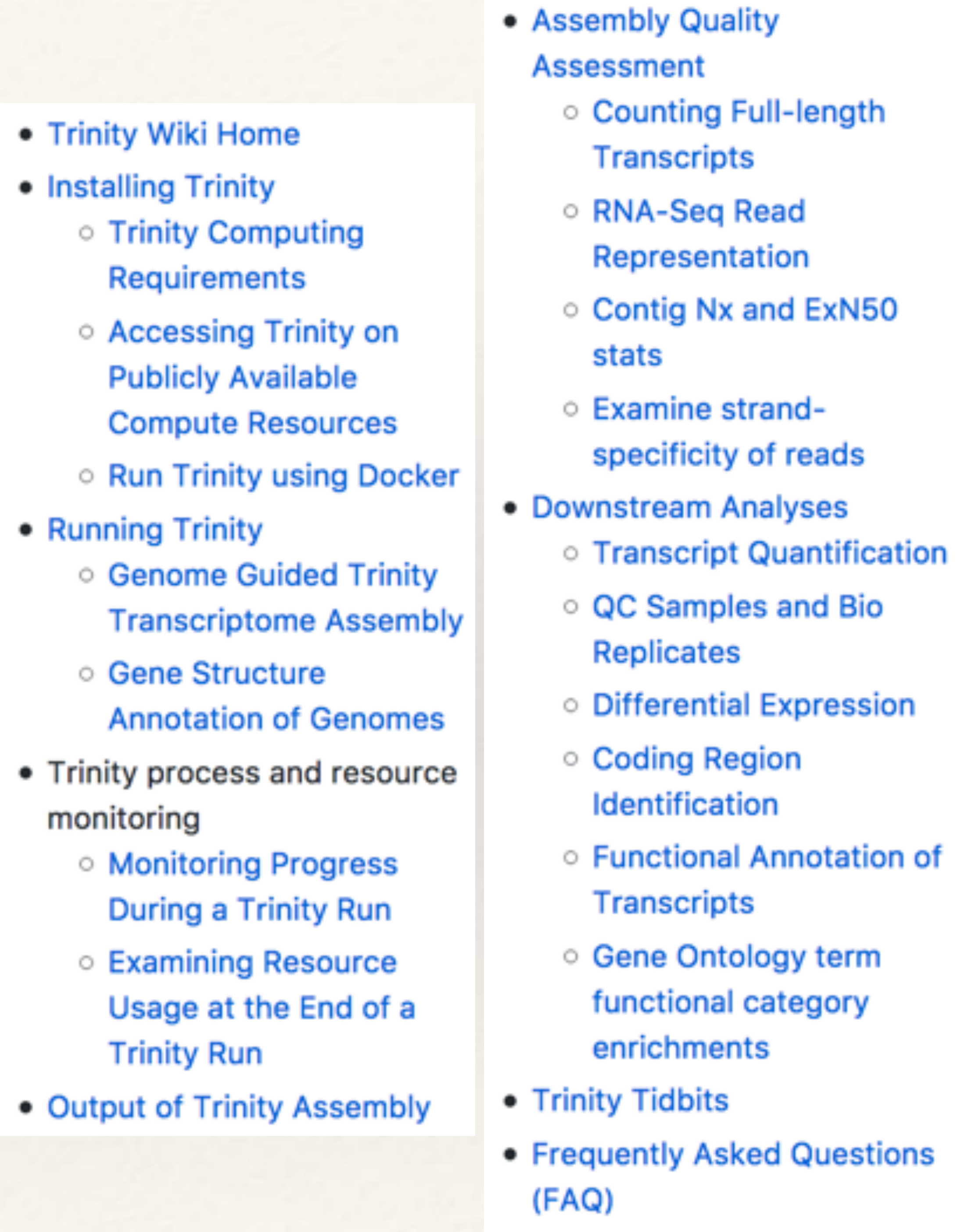




https://github.com/trinityrnaseq/trinityrnaseq/wiki

# Trinity assembler

- Significantly more transcripts than predicted in the same or closely related species!
- Low coverage over splice junctions, sequencing errors and heterozygosity restricts full-length transcript reconstruction



- Trinity Wiki Home
- Installing Trinity
    - Trinity Computing Requirements
    - Accessing Trinity on Publicly Available Compute Resources
    - Run Trinity using Docker
- Running Trinity
    - Genome Guided Trinity Transcriptome Assembly
    - Gene Structure Annotation of Genomes
- Trinity process and resource monitoring
    - Monitoring Progress During a Trinity Run
    - Examining Resource Usage at the End of a Trinity Run
- Output of Trinity Assembly
- Assembly Quality Assessment
    - Counting Full-length Transcripts
    - RNA-Seq Read Representation
    - Contig Nx and ExN50 stats
    - Examine strand-specificity of reads
- Downstream Analyses
    - Transcript Quantification
    - QC Samples and Bio Replicates
    - Differential Expression
    - Coding Region Identification
    - Functional Annotation of Transcripts
    - Gene Ontology term functional category enrichments
- Trinity Tidbits
- Frequently Asked Questions (FAQ)

https://github.com/trinityrnaseq/trinityrnaseq/wiki

# Trinity assembler

- Assembly algorithms require large amounts of memory
- 2/3rds of Trinity is parallelised to save computation time
- Estimate at least 1 week of trial/error/final computation
- Remember to calculate memory/time requirements before starting!
    - – 1Gb RAM / million reads
    - – 30 mins - 1 hour / million reads

https://github.com/trinityrnaseq/trinityrnaseq/wiki

# Transcriptome Assembly Quality Assessment

- Guide from Trinity
  - https://github.com/trinityrnaseq/trinityrnaseq/wiki/Transcriptome-Assembly-Quality-Assessment
- BUSCO
- CEGMA (discontinued)

http://busco.ezlab.org/
http://korflab.ucdavis.edu/datasets/cegma/

# Genome annotation pipeline

- MAKER2
  - Step-by-step annotation guide
  - https://github.com/sujaikumar/assemblage/blob/master/README-annotation.md
- Gene prediction based on ESTs and/or transcriptomes
  - AUGUSTUS, GeneMark-ES, SNAP and more

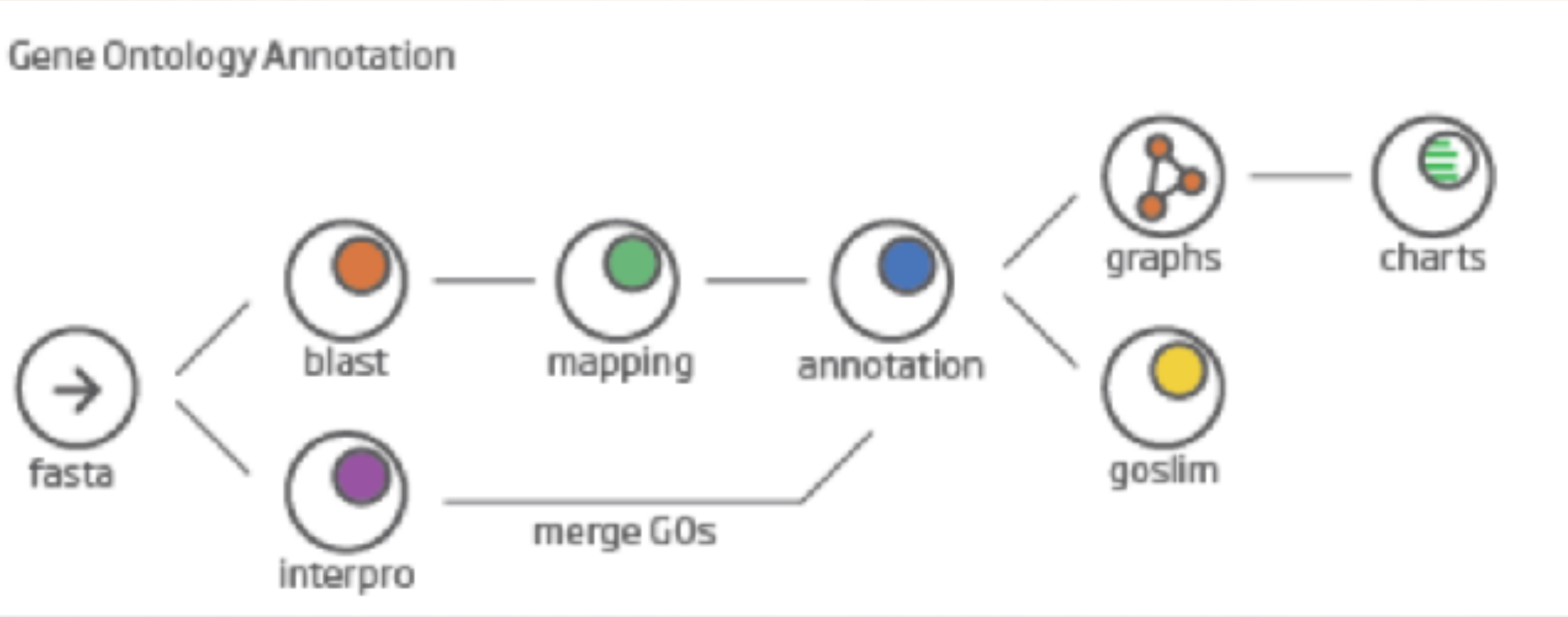
http://www.yandell-lab.org/software/maker.html

# Genome annotation pipeline

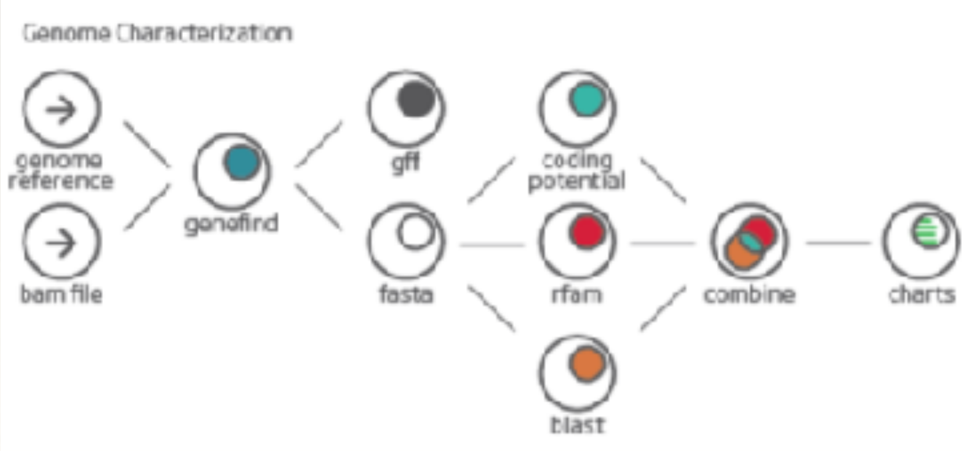
- BLAST against appropriate databases
    - NCBI blast+ toolkit
- Gene ontology
- KEGG pathways
- blast2go

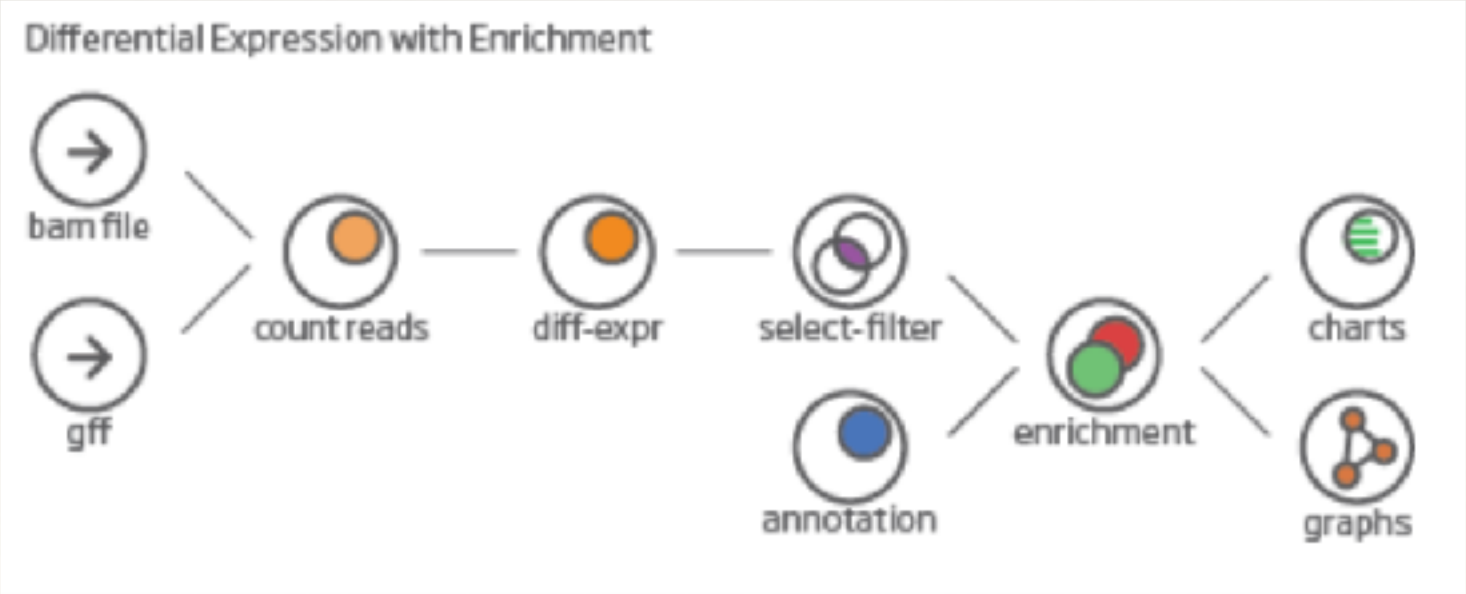
https://www.blast2go.com/

# blast2go


Gene Ontology Annotation

blast2go


Genome Characterization


Differential Expression with Enrichment

https://www.blast2go.com/