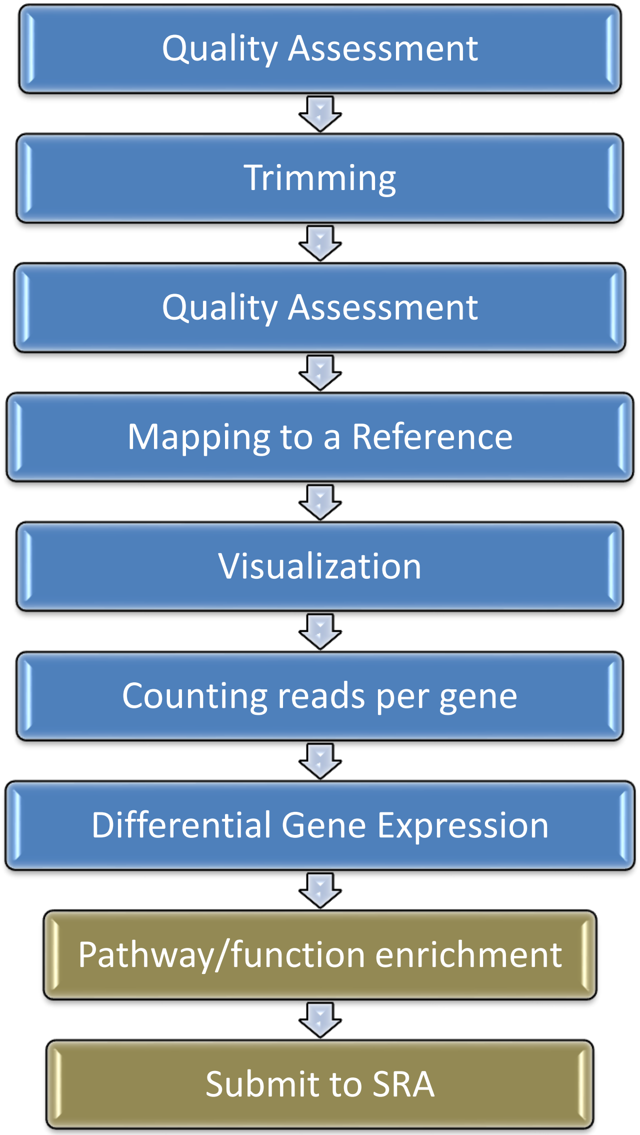
RNASeq Data Analysis Pipeline



Quality Assessment

Trimming

Quality Assessment

Mapping to a Reference

Visualization

Counting reads per gene

Differential Gene Expression

Pathway/function enrichment

Submit to SRA

FastQC

Babraham Institute Babraham Bioinformatics

Skewer

MultiQC

STAR

igv Integrative Genomics Viewer

HTSeq

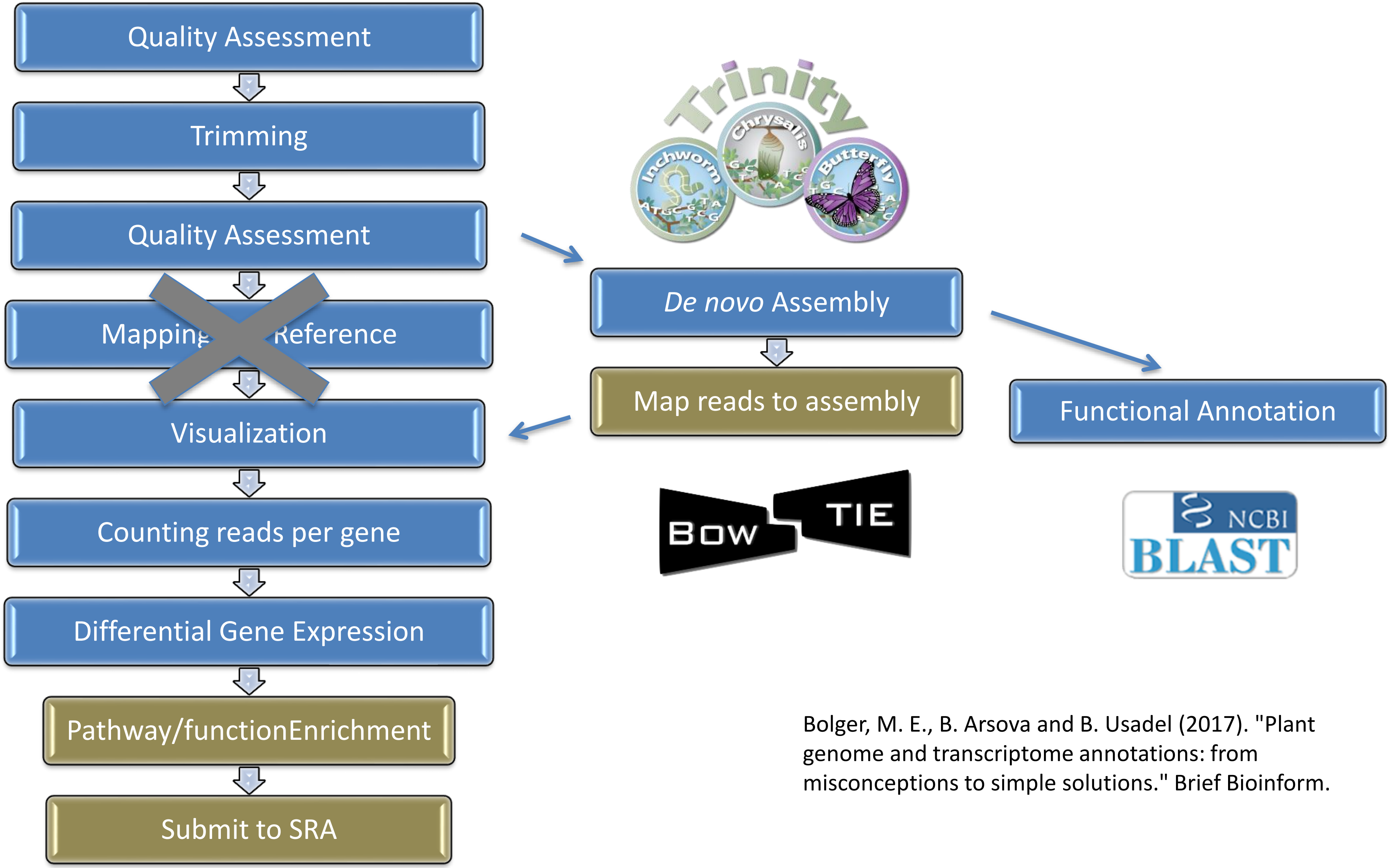
DESeq2

Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

KEGG/GO

NCBI

What if you don’t have a reference?



Quality Assessment

Trimming

Quality Assessment

~~Mapping Reference~~

Visualization

Counting reads per gene

Differential Gene Expression

Pathway/functionEnrichment

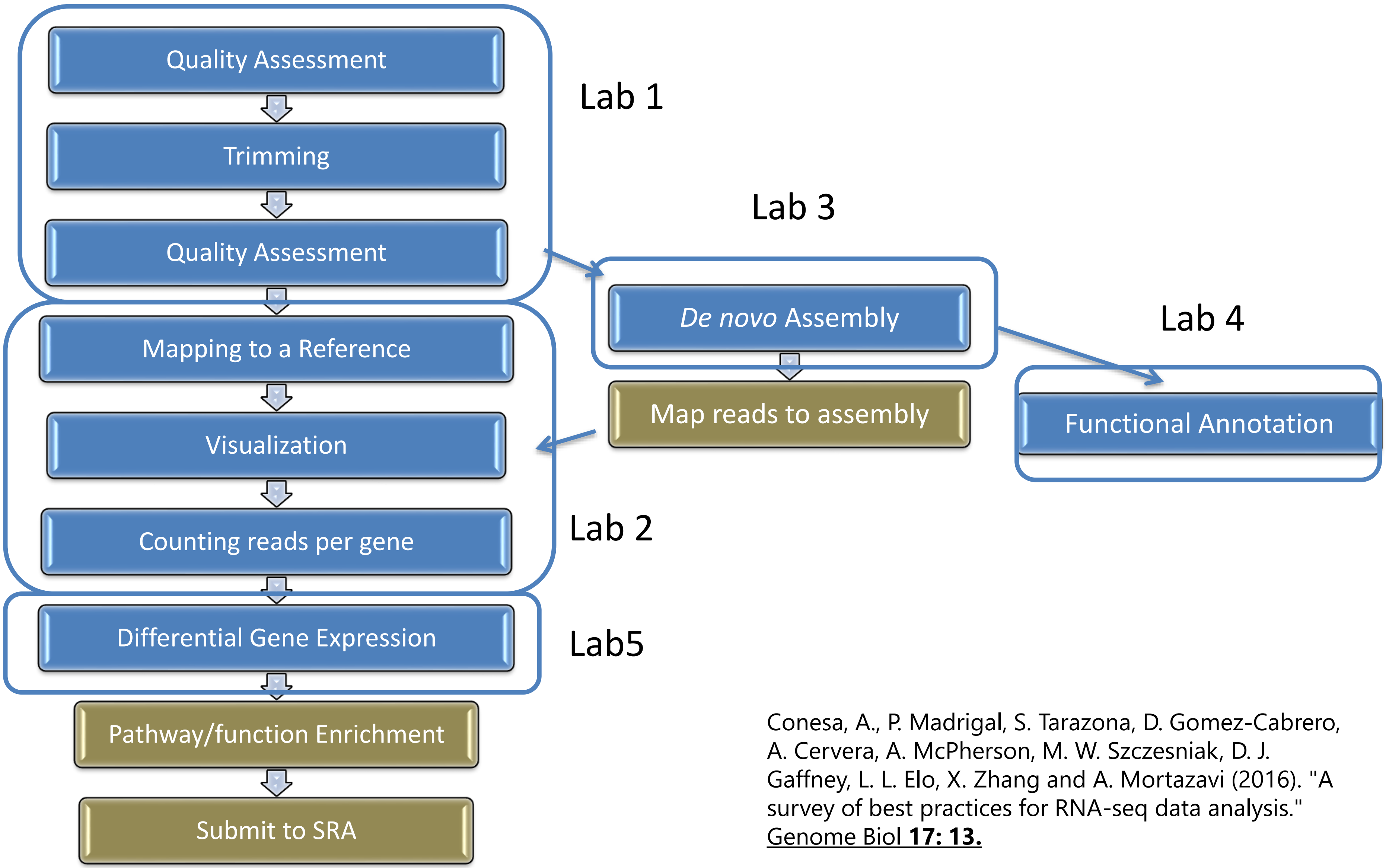
Submit to SRA

De novo Assembly

Map reads to assembly

Functional Annotation

Bolger, M. E., B. Arsova and B. Usadel (2017). "Plant genome and transcriptome annotations: from misconceptions to simple solutions." Brief Bioinform.



Lab 1

Quality Assessment

Trimming

Quality Assessment

Lab 3

De novo Assembly

Map reads to assembly

Lab 4

Functional Annotation

Lab 2

Mapping to a Reference

Visualization

Counting reads per gene

Lab5

Differential Gene Expression

Pathway/function Enrichment

Submit to SRA

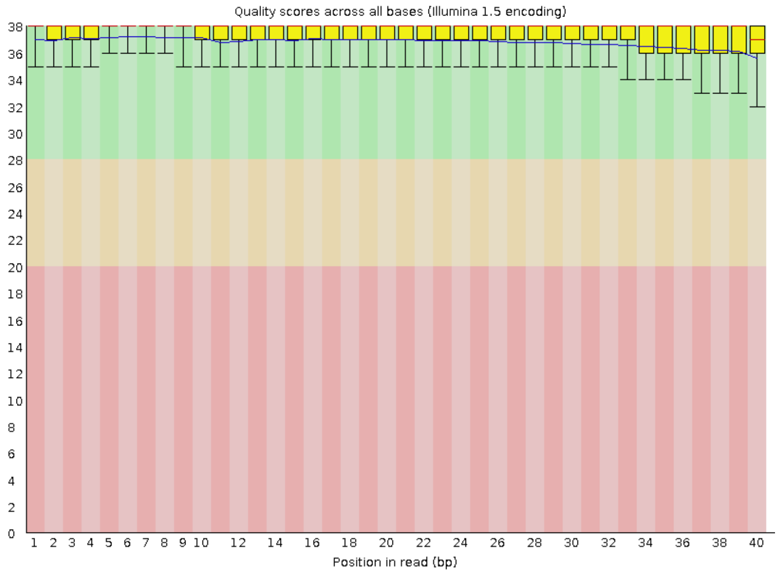
Conesa, A., P. Madrigal, S. Tarazona, D. Gomez-Cabrero, A. Cervera, A. McPherson, M. W. Szczesniak, D. J. Gaffney, L. L. Elo, X. Zhang and A. Mortazavi (2016). "A survey of best practices for RNA-seq data analysis." Genome Biol **17: 13.**

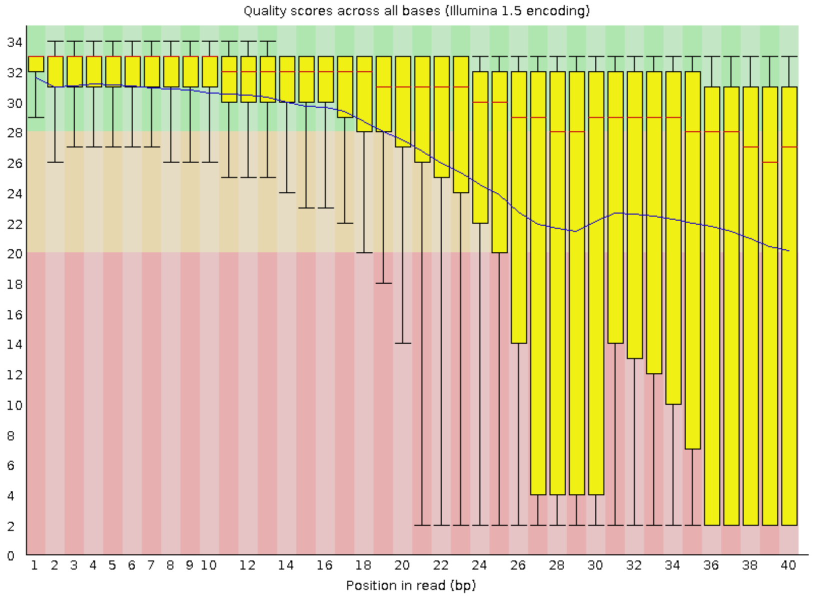
Quality Control

- Is my data of sufficient quality?
- The instrument assigns a confidence value to each base. Are the bases high quality overall?
- Does the complexity look normal?

FastQC

Babraham Bioinformatics





http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Trimming

- Get rid of the bad data, keep the good data
- Adapter trimming
  - Cut adapter and other Illumina-specific sequences from the read
- Quality trimming
  - Trim off low quality bases
  - Drop a read entirely if is too low quality or too short

Skewer

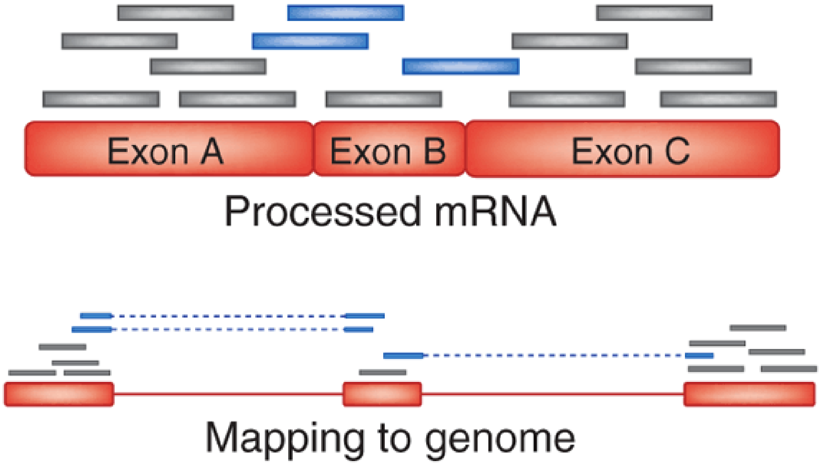
Jiang, H., R. Lei, S. W. Ding and S. Zhu (2014). "Skewer: a fast and accurate adapter trimmer for next-generation sequencing paired-end reads." BMC Bioinformatics **15: 182.**

Newest research:
Gentle trimming is better.

Mapping to the Reference

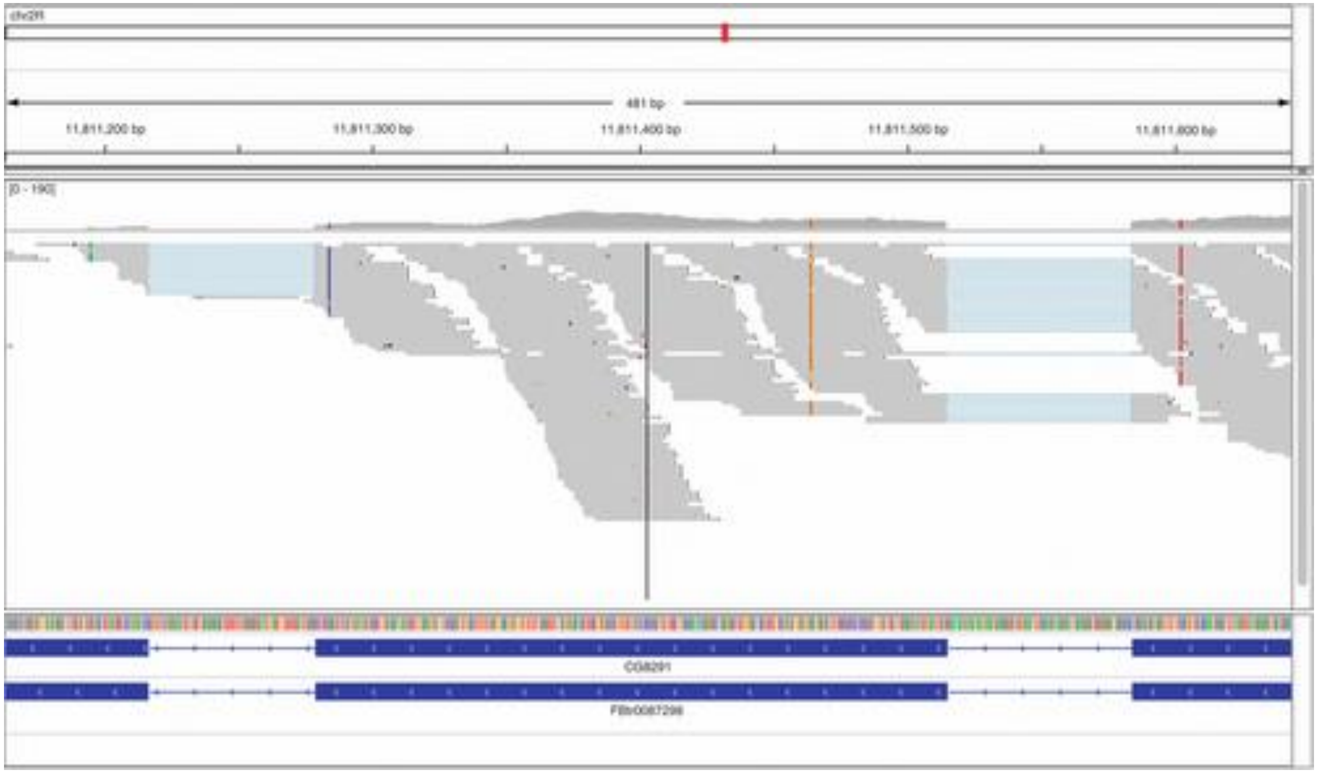
- Mapping RNA to a eukaryotic genome is more complicated than mapping DNA
  - Introns
  - Alternative splicing

- Use a mapping software designed for spliced RNASeq
  - The software will use a file (gff3) to know where the genes are located
  - If this is not available, some mapping software can infer gene structures (cufflinks)



Visualization

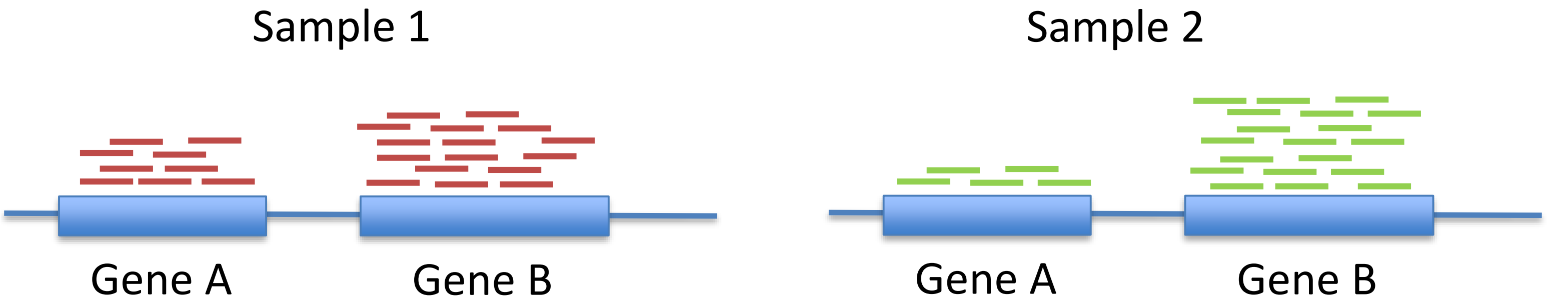
- Look at your data
- The number one most under-appreciate step in data analysis



Integrative Genomics Viewer

Differential expression

- Find genes responding to the conditions
- Replicates give power to your results
  - Biological replicates capture random biological variation
  - Technical replicates measure the random noise of protocols or equipment
- Choose an algorithm that suits the data
  - RNASeq expression levels are discrete counts



Generate sequence counts
for all genes in genome

↓

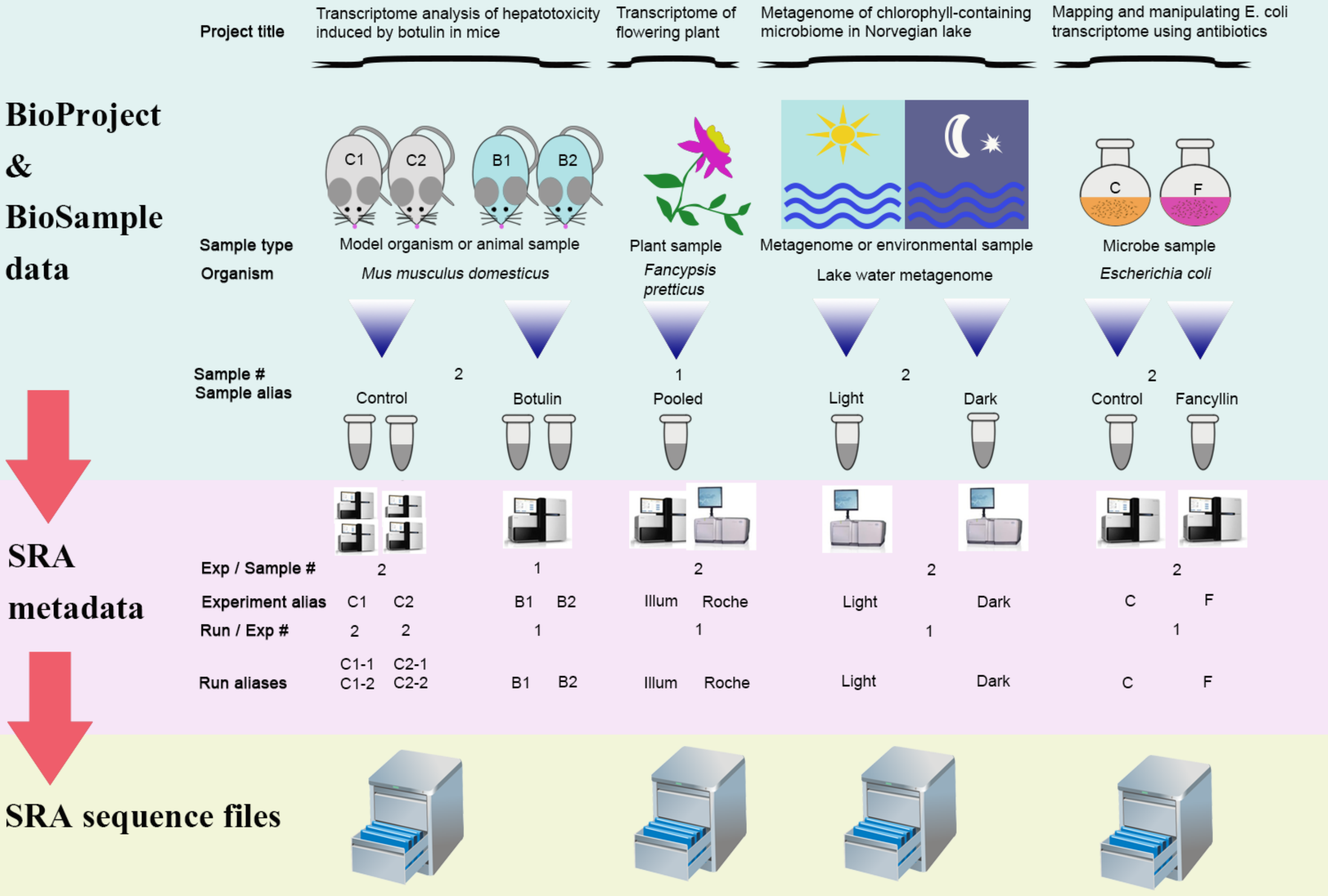
Differential expression

Making data public

- NCBI Short Read Archive (SRA)
  - Stores raw sequence data from "next-generation" sequencing technologies including 454, IonTorrent, Illumina, SOLiD, Helicos and Complete Genomics.
  - SRA also stores alignment information in the form of read placements on a reference sequence.
- Upload to SRA
  - Make a list of all the things you need to know prior to starting the project, and keep it updated.
  - Most journals require an accession number prior to publication
  - Enhances reproducibility and allows for new discovery by comparing data sets.
  - Overview of submission process:

    https://www.ncbi.nlm.nih.gov/sra/docs/submit/

Anatomy of SRA data



BioProject & BioSample data

SRA metadata

SRA sequence files

Project title	Transcriptome analysis of hepatotoxicity induced by botulin in mice		Transcriptome of flowering plant	Metagenome of chlorophyll-containing microbiome in Norvegian lake		Mapping and manipulating E. coli transcriptome using antibiotics	
Sample type	Model organism or animal sample		Plant sample	Metagenome or environmental sample		Microbe sample	
Organism	*Mus musculus domesticus*		*Fancypsis pretticus*	Lake water metagenome		*Escherichia coli*	
Sample #	2		1	2		2	
Sample alias	Control	Botulin	Pooled	Light	Dark	Control	Fancyllin
Exp / Sample #	2	1	2	2		2	
Experiment alias	C1 C2	B1 B2	Illum Roche	Light	Dark	C	F
Run / Exp #	2 2	1	1	1		1	
Run aliases	C1-1 C1-2 C2-1 C2-2	B1 B2	Illum Roche	Light	Dark	C	F

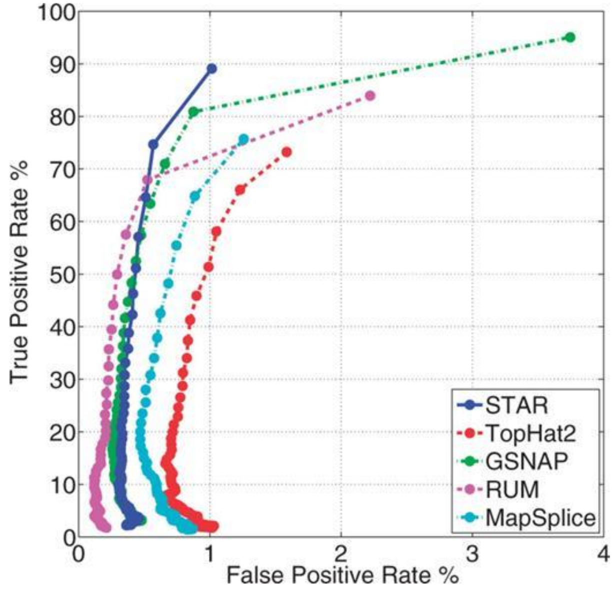
https://www.ncbi.nlm.nih.gov/sra/docs/submitmeta/

Software election

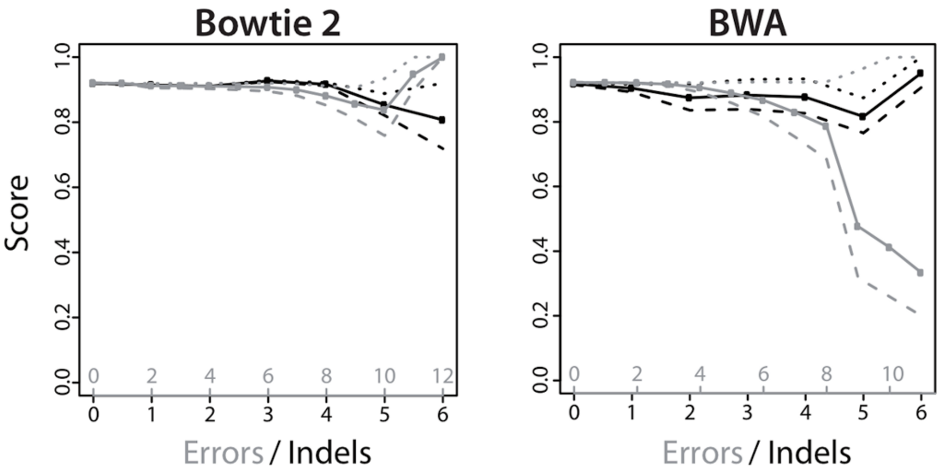
- Multiple software options are available along the analysis
    - Trimming
        - Trimmomatic, skewer
    - Mapping
        - STAR, GSNAP, Stampy, TopHat, HISAT2, bowtie2...
    - Differential expression
        - DESeq2 and edgeR are based on the negative binomial distribution
        - Others are NOISeq, baySeq, SAMseq, limma, cuffdiff...
    - De novo assembly
        - Trinity, SOAP, Trans-ABySS...
- Then... what’s the “best” choice? There are preferred options
- Some articles compared them using real or simulated data

Mapping

- Accuracy, speed, and computational resources
- Mapping of reads to their true location
  - Which fraction of aligned reads is aligned correctly? - precision
  - Which fraction of overall reads were correctly recovered? - recall
  - SNPs and INDELs have a big impact



https://doi.org/10.1093/bioinformatics/bts635

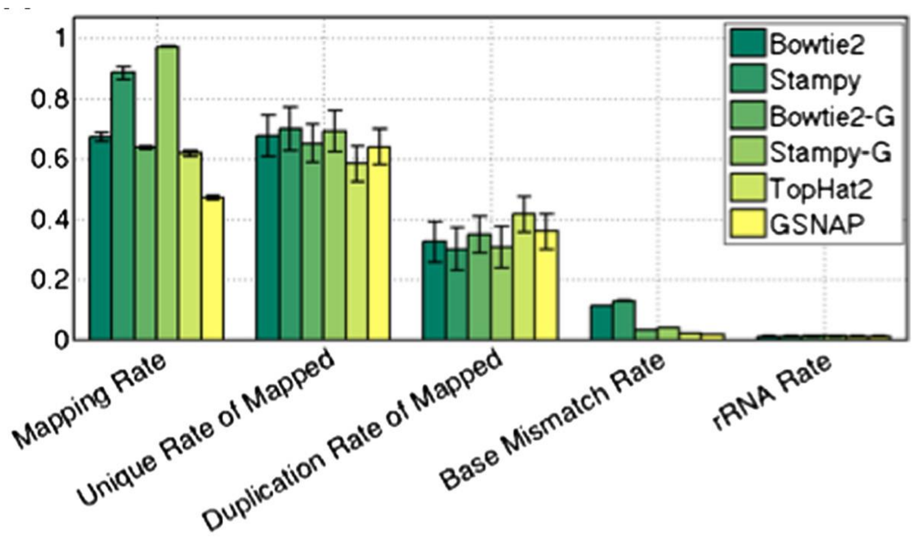


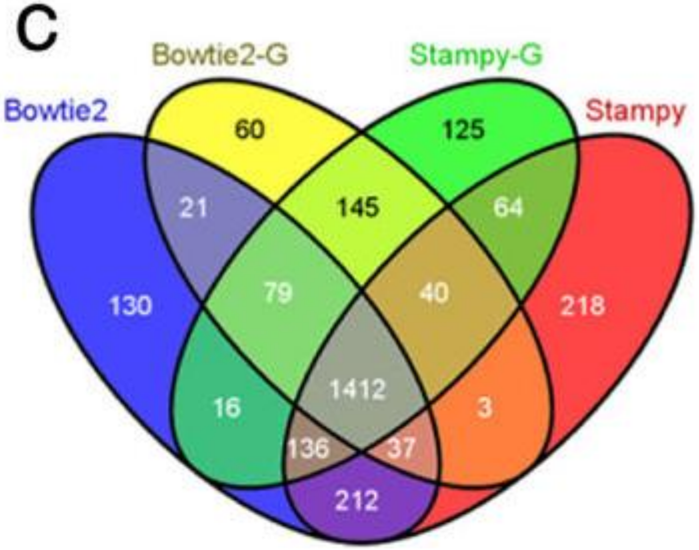
Precision (dotted lines), recall (dashed lines) and F-measure (solid lines –)

https://doi.org/10.1371/journal.pone.0052403.g003

Mapping: genome vs transcriptome

- Is it better to map to a genome or transcriptome?
  - Genome: it provides a predefined annotation that helps comparing results
  - Transcriptome: can be good for identifying novel genes and isoforms



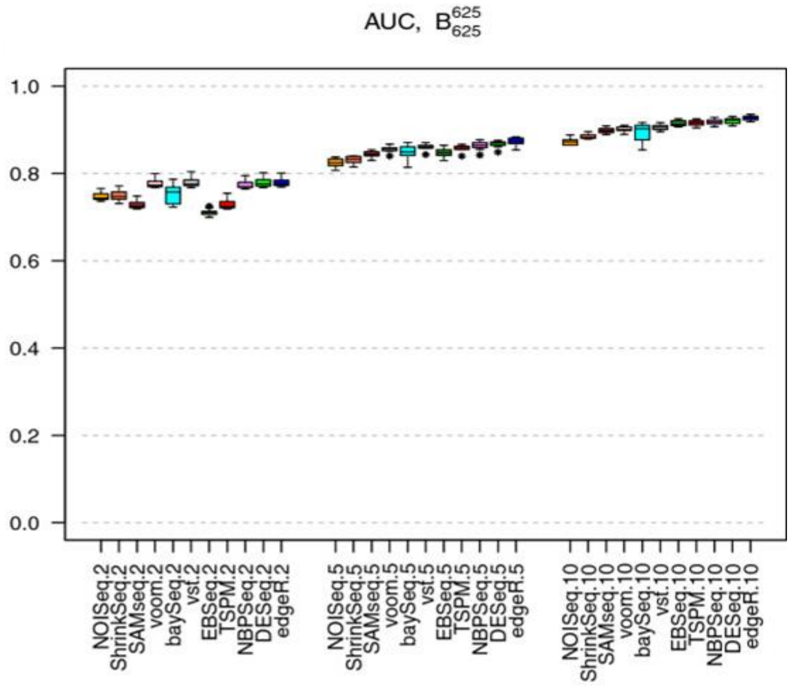


Differential Expression Overlap

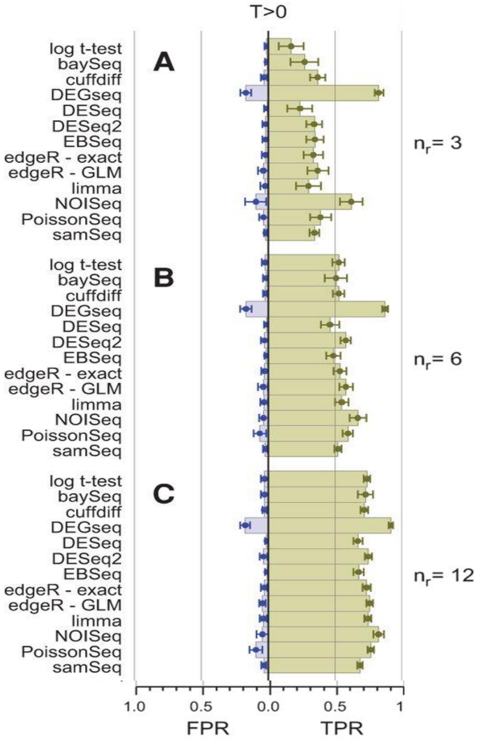
https://doi.org/10.1186/1471-2164-15-570

Differential expression

- High detection of true DE genes (true positive rate) and avoidance of false DEG
- Replicates improve detection of true DEG



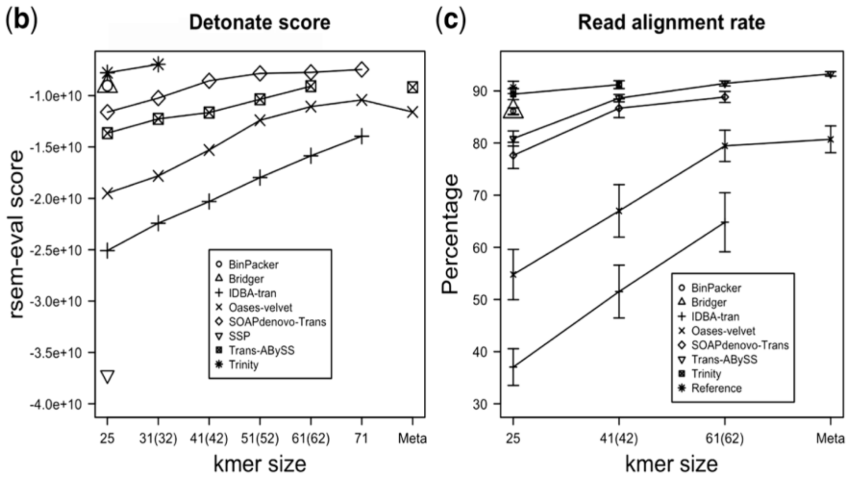
https://doi.org/10.1186/1471-2105-14-91

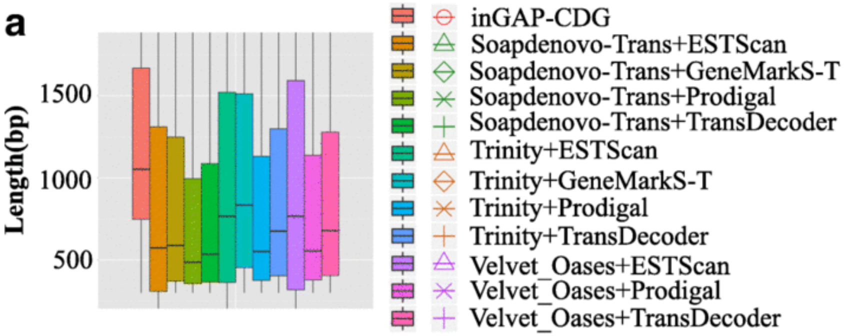


http://rnajournal.cshlp.org/content/22/6/839.full

De novo assembly + annotation

- Assembly: Number of transcripts, transcript length, redundancy, representation of reads
- Annotation: Number of transcripts with ORF, length of encoded ORFs, completeness of essential biological functions in the transcriptome





https://doi.org/10.1186/s13059-016-1094-x

https://doi.org/10.1093/bioinformatics/btw625

Conclusions

- There are decisions to take before starting the analysis
  - What programs to use?
    - Do some tests for optimization?
  - Genome or transcriptome?
  - Algorithm to detect differential expression?
- Results will be affected by these decisions
- Improve robustness by giving support to the software selection made
  - Experimental optimization of parameters
  - References