Tissue collection, read processing, assembly, and translation

Ya Yang

yangya@umn.edu

Transcriptome workshop, Botany 2018

Herbarium and Department of Plant Biology

University of Minnesota-Twin Cities

The most important slide

- Transcriptomics is an extremely active research area
  - The principles are general, the specifics change every couple of months

The most important slide

- Transcriptomics is an extremely active research area
- “Best practice” differs by plant group and data set
- Invest time to get familiar with command line and regular expression. Unit test on subset of data

The most important slide

- Transcriptomics is an extremely active research area
- “Best practice” differs by plant group and data set
- Invest time to get familiar with command line and regular expression. Unit test on subset of data

Sources of data

- NCBI Sequence Read Archive (SRA)
  - Access from command line directly using SRA Toolkit
  - Be aware of submission that does not properly metadata such as sample source, library type, and sequencing info; many SRA data set does not have associated publication information!

Sources of data

- NCBI Sequence Read Archive (SRA)
  - Access from command line directly using SRA Toolkit
  - Be aware of submission that does not properly metadata such as sample source, library type, and sequencing info; many SRA data set does not have associated publication information!

- The 1000 plants (1KP) initiative. Now all data available through NCBI SRA https://sites.google.com/a/ualberta.ca/onekp/

Sources of data

- NCBI Sequence Read Archive (SRA)
  - Access from command line directly using SRA Toolkit
  - Be aware of submission that does not properly metadata such as sample source, library type, and sequencing info; many SRA data set does not have associated publication information!

- The 1000 plants (1KP) initiative. Now all data available through NCBI SRA https://sites.google.com/a/ualberta.ca/onekp/

- Genomes – ingroups and outgroups https://phytozome.jgi.doe.gov

Sources of data

- NCBI Sequence Read Archive (SRA)
  - Access from command line directly using SRA Toolkit
  - Be aware of submission that does not properly metadata such as sample source, library type, and sequencing info; many SRA data set does not have associated publication information!

- The 1000 plants (1KP) initiative. Now all data available through NCBI SRA https://sites.google.com/a/ualberta.ca/onekp/

- Genomes – ingroups and outgroups https://phytozome.jgi.doe.gov

- Generate your own data from fresh tissue collected from green house, botanical gardens, or the field

Tissue collection:

liquid N_2 vs. RNA*later*

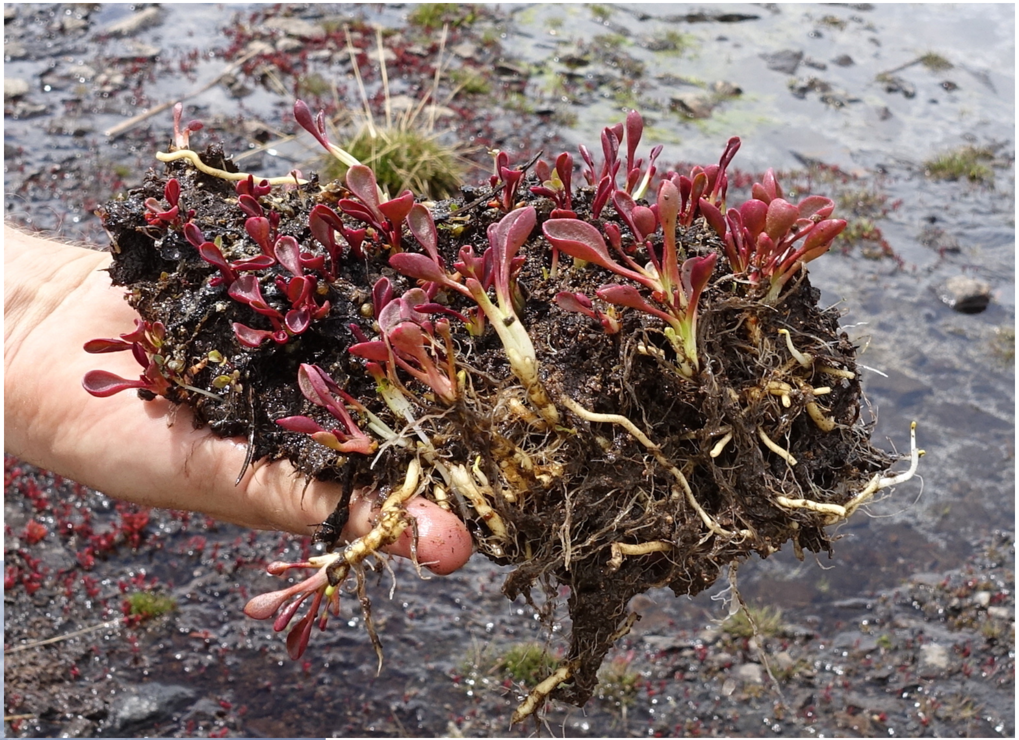
Photo by Mike Moore

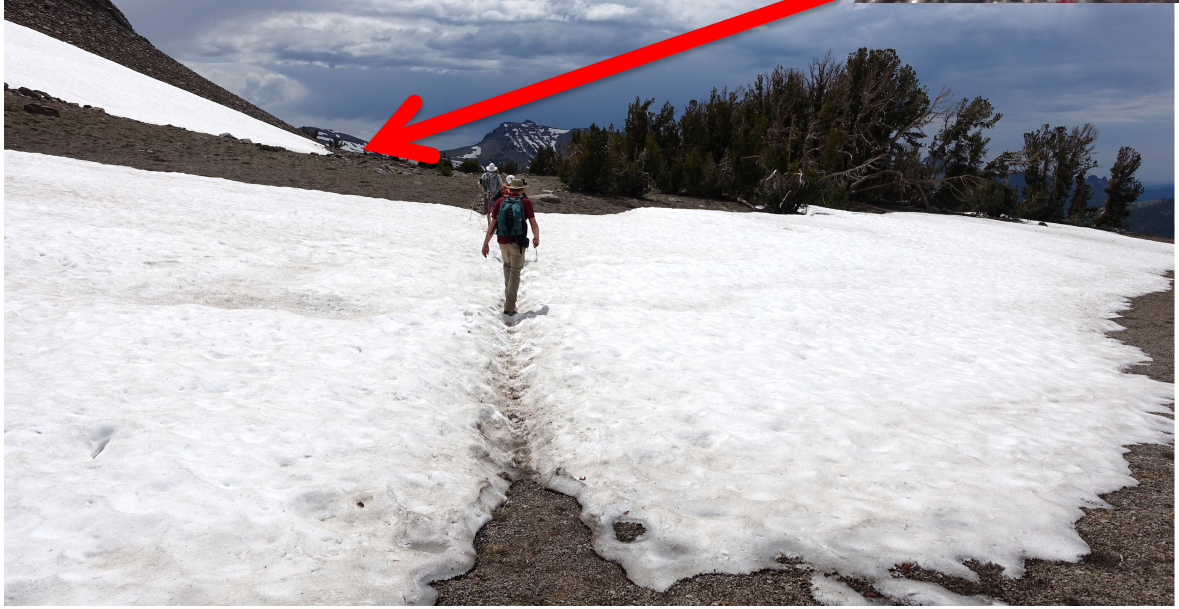
http://onsnetwork.org/blog/tag/rnalater/

- Flash freeze in liquid N_2. Small dry shippers can fit into a backpack
- Store in -80°C or in liquid N_2 vapor freezers
- Preserves DNA, RNA, secondary metabolites

- 1 week at RT, 4°C for a month, -20°C forever
- Preserves DNA and RNA
- May not work in certain groups. Test before using for trips

Sierra Nevada, California, United States

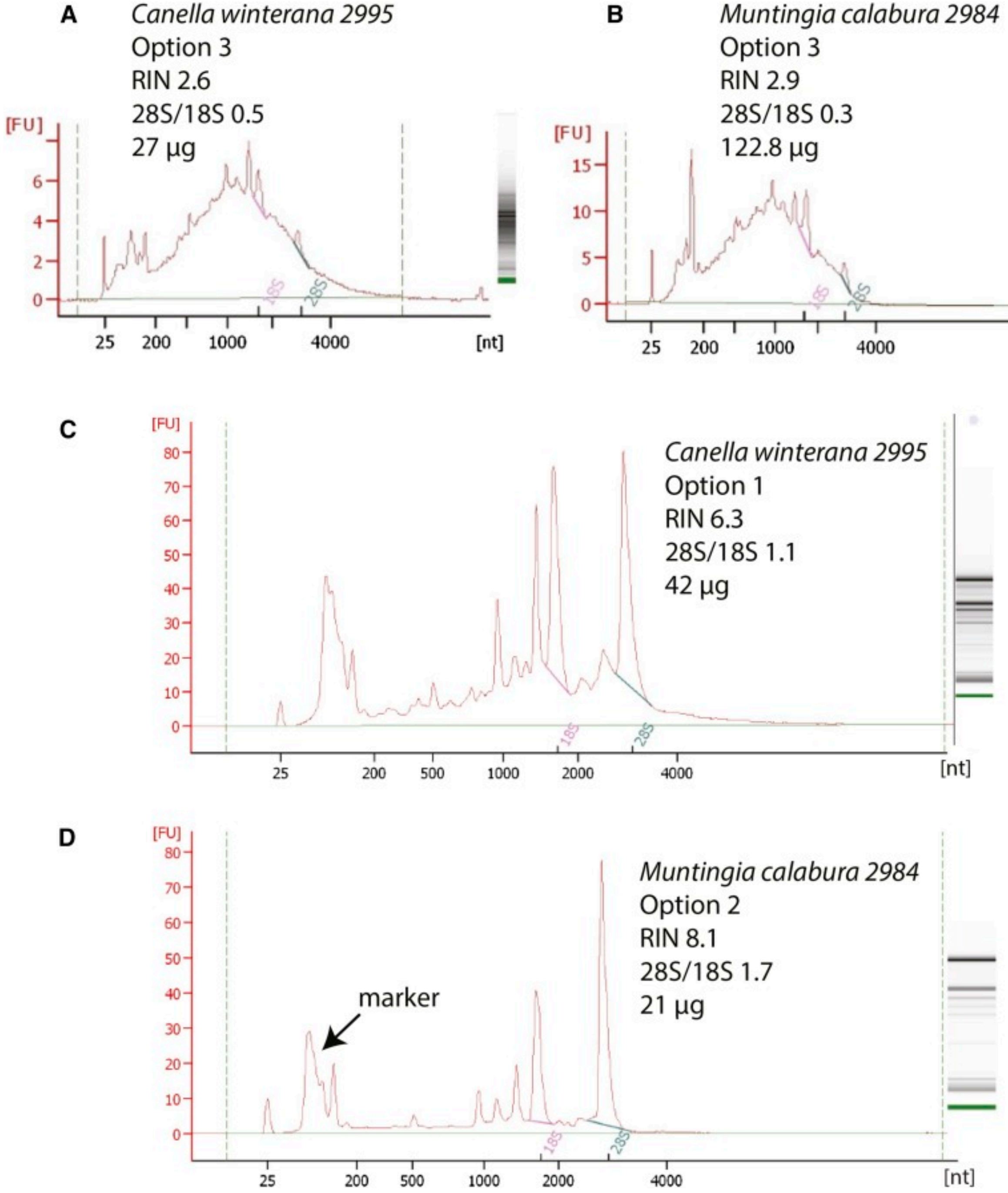




Claytonia nevadensis
Montiaceae

What tissue type?

- Young leaves are better than mature leaves
- Flower buds: easier to extract RNA and add flower-specific genes.
  - Avoid open flower to avoid additional alleles
- Mix tissue types to increase number of genes recovered



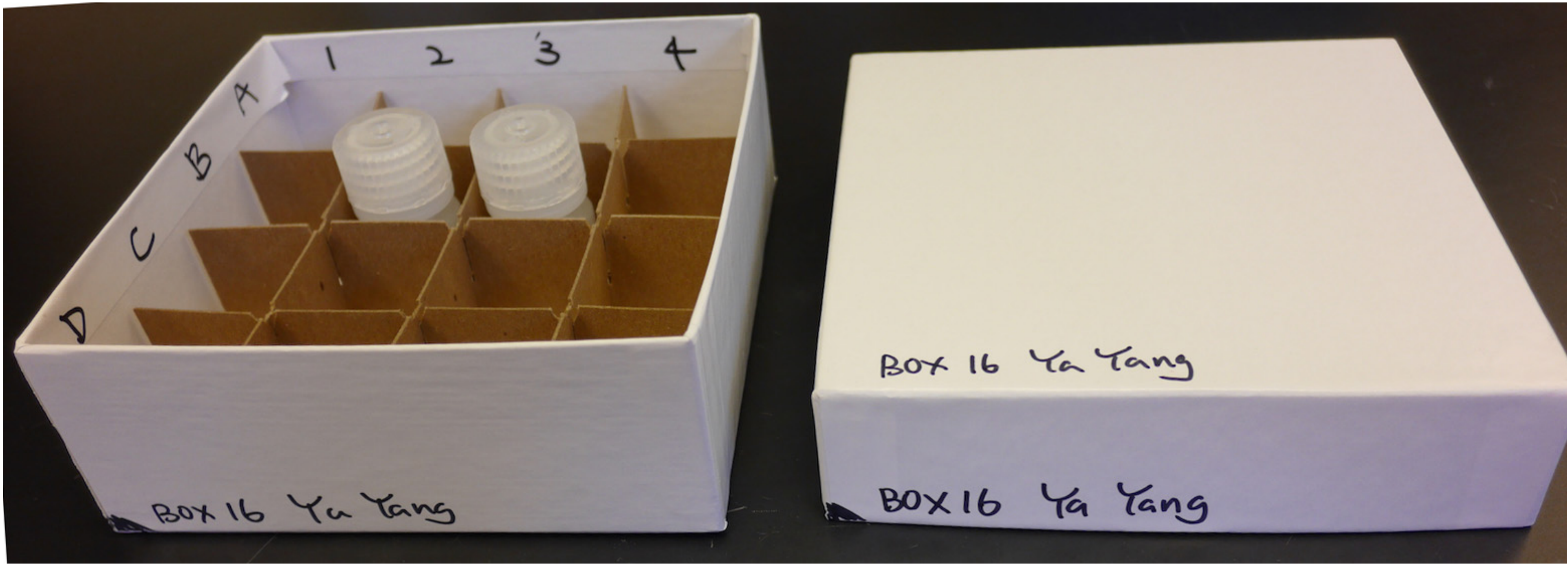
RNA extraction: QIAGEN RNeasy Plant Mini Kit or PureLink Plant RNA Reagent (streamlined CTAB)

DNase digestion

Quality control by Bioanalyzer

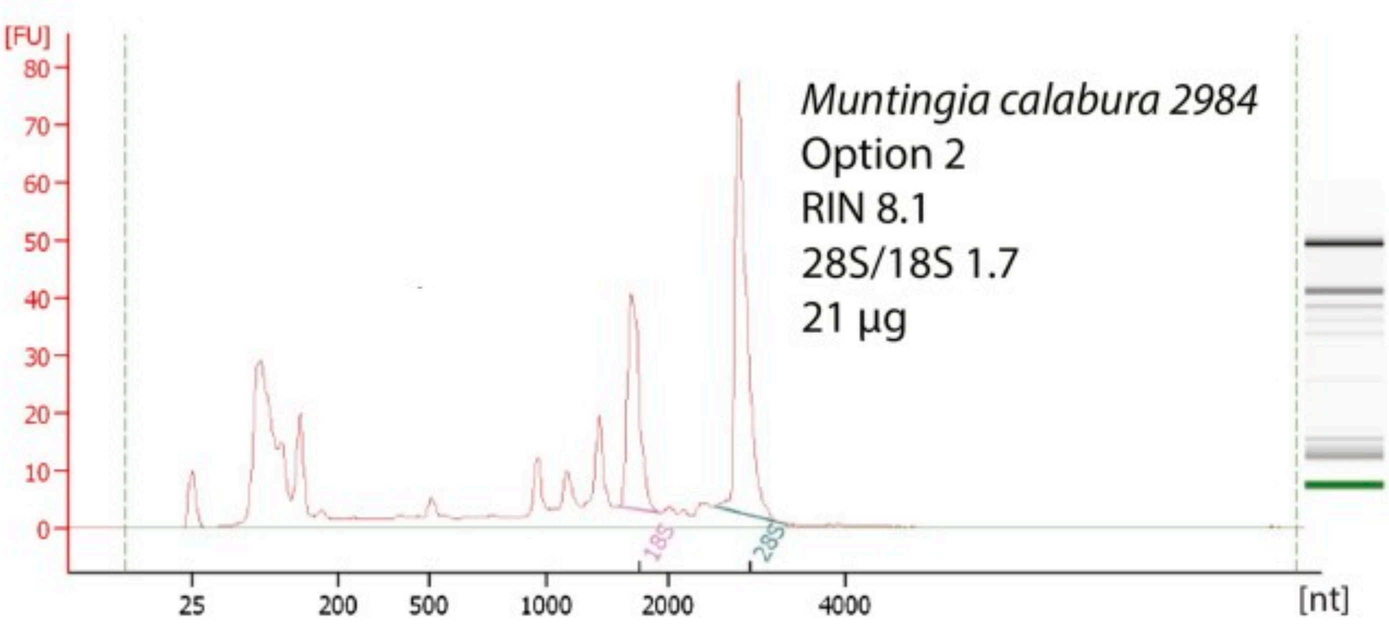
Jordon-Thaden *et al.*, 2015 APPS

See Yang at el. APPS 2017 for field, lab, and sample curation protocols



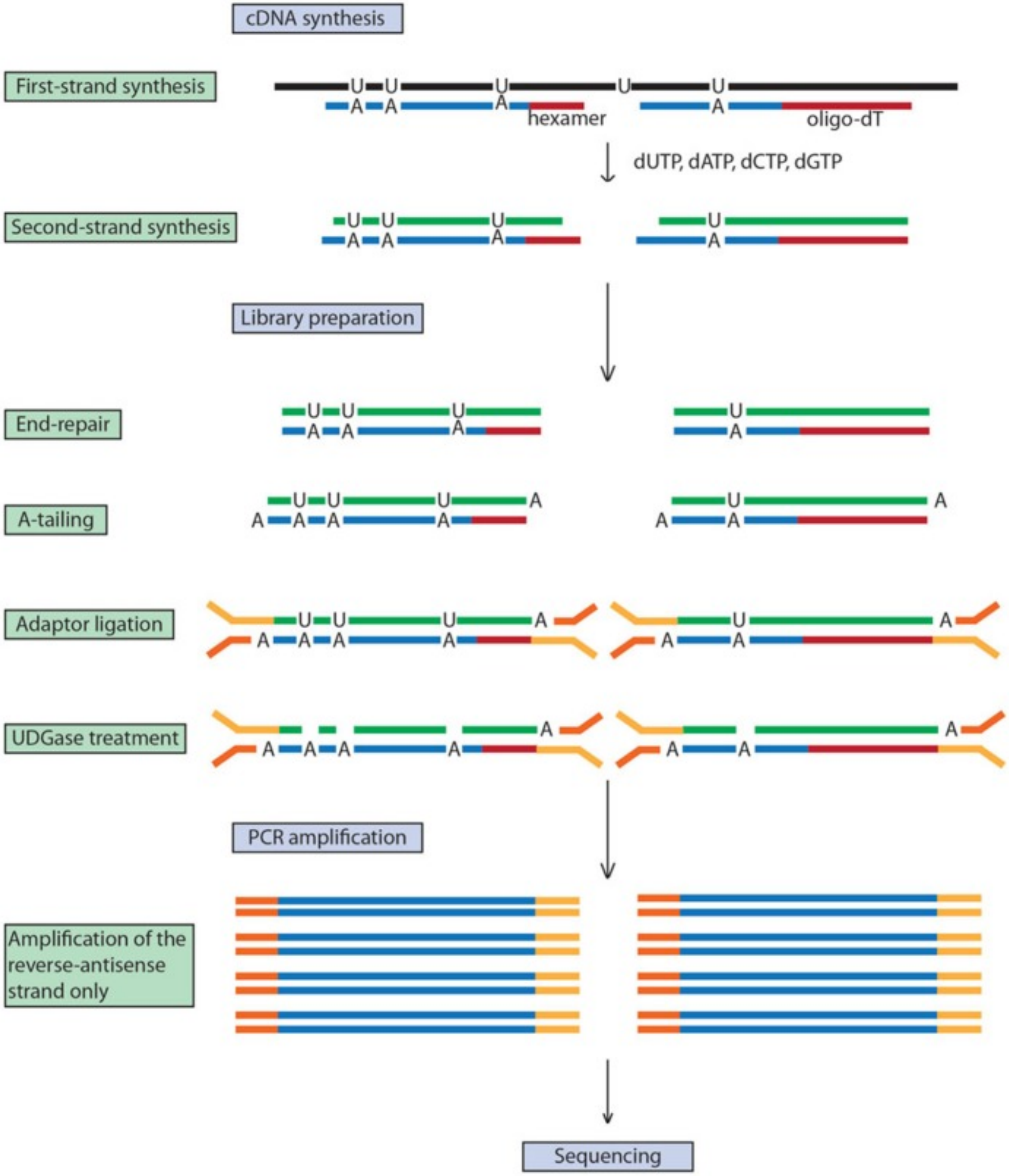
Library preparation: KAPA kit or outsource

- Poly-A enrichment to enrich mRNA
- Or alternatively, RiboMinus to reduce rRNA



Jordon-Thaden *et al.*, 2015 APPS

Library preparation



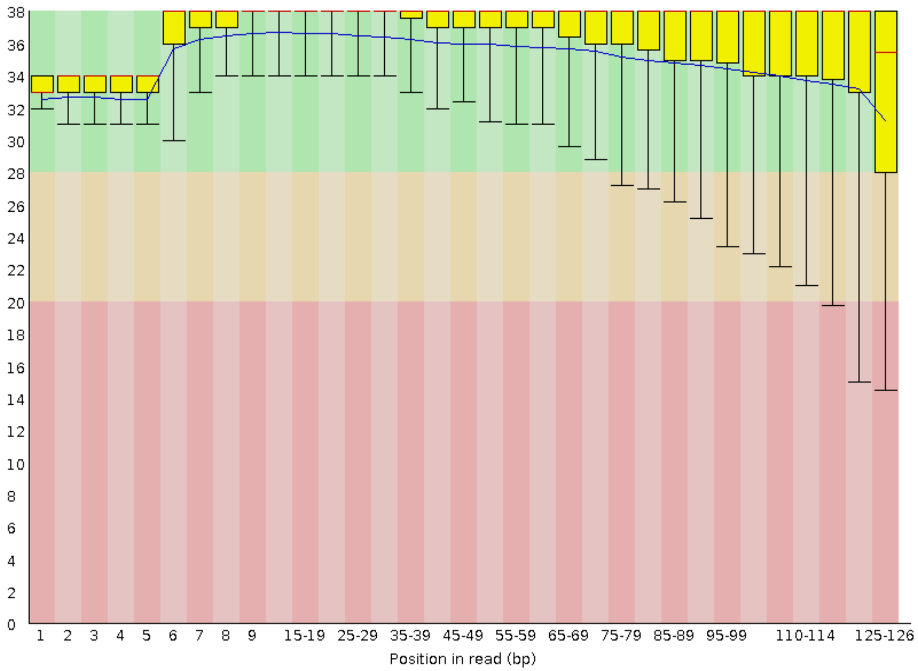
- Stranded mRNA library prep

Martin *et al.*, 2013 Front. Plant Sci.

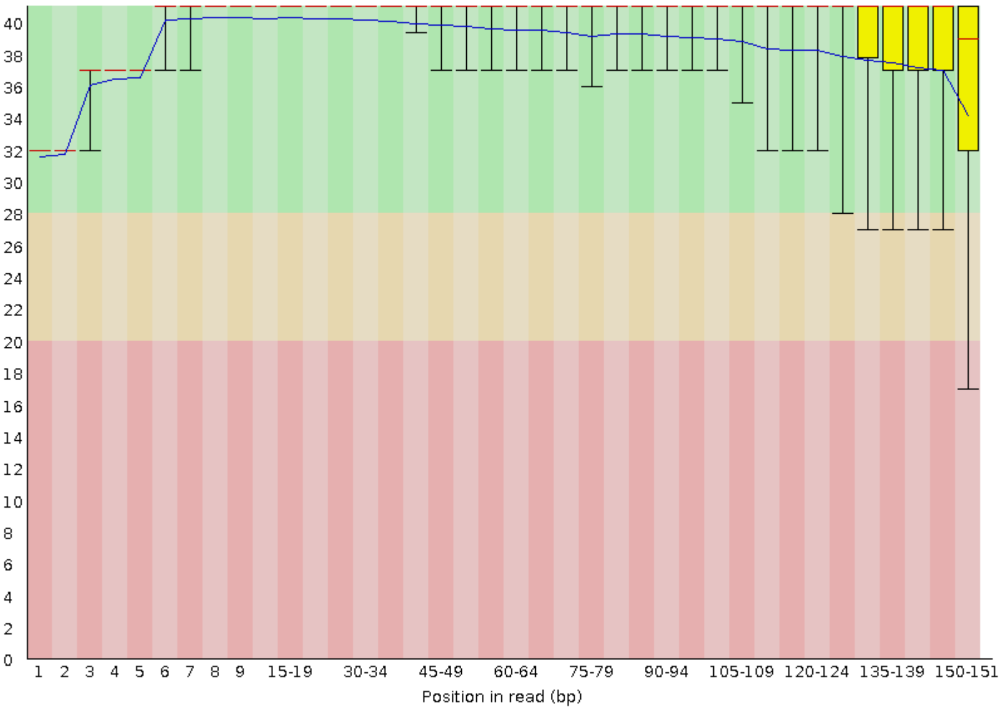
Choice of sequencing platforms

- Illumina HiSeq2500/4000: our workhorse the past few years
- Illumina NextSeq
- Illumina NovaSeq: much cheaper but not practical for phylotranscriptomics
- Multiplex to aim for 25–35 million read pairs

HiSeq2500V4
high-throughput mode



HiSeq4000



Read processing

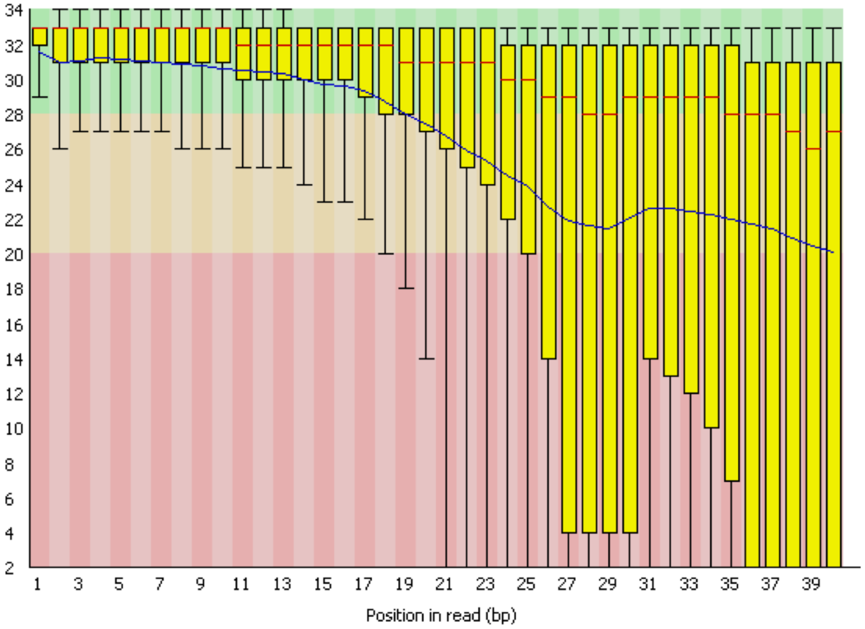
- Random sequencing error correction with Rcorrector
- Remove sequencing adapters and low quality sequences with Trimmomatic
- Filter organelle reads (cpDNA, mtDNA or both) with Bowtie2 and assemble with Fast-Plast
- Run FastQC to check read quality and detect over-represented reads
- Remove over-represented sequences

Quality trimming of raw reads

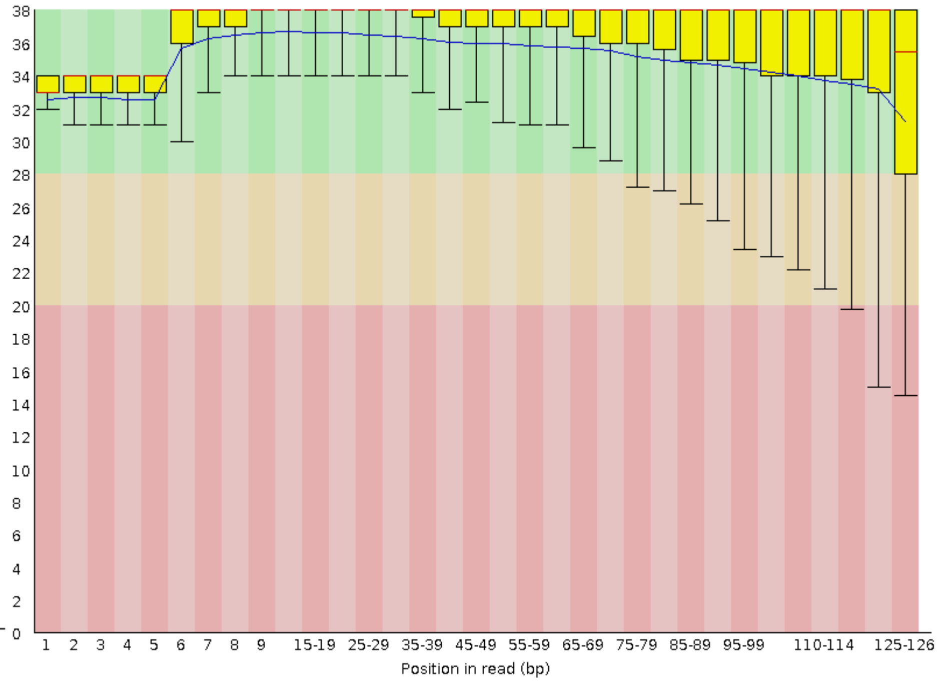
- Optimal trimming parameters are dependent on your purpose (recover more complete or more accurate assemblies).

- With the latest Illumina platforms, the short answer is gentle trimming is usually good

Visualize quality of reads using FastQC



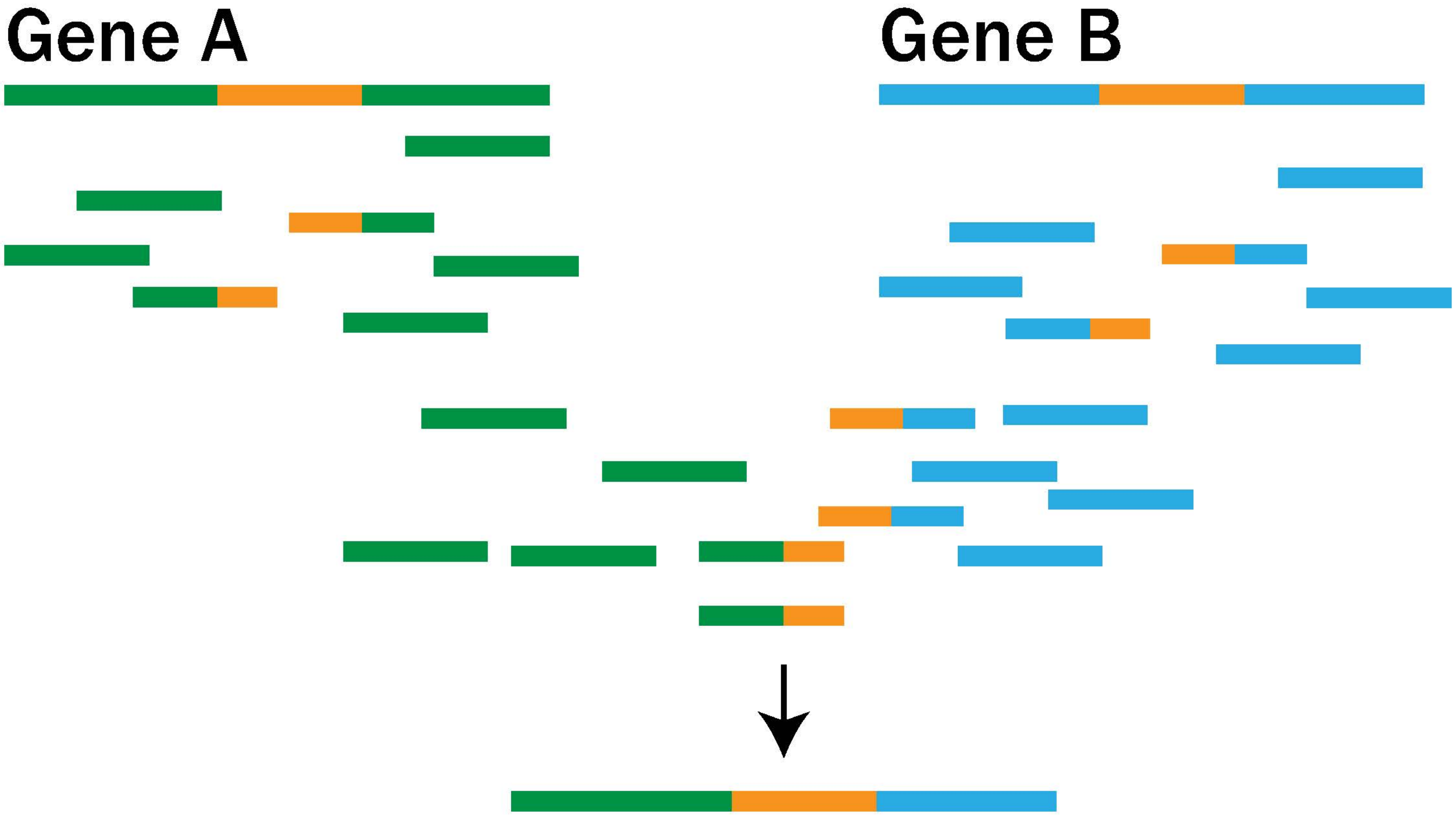
Problematic



Good

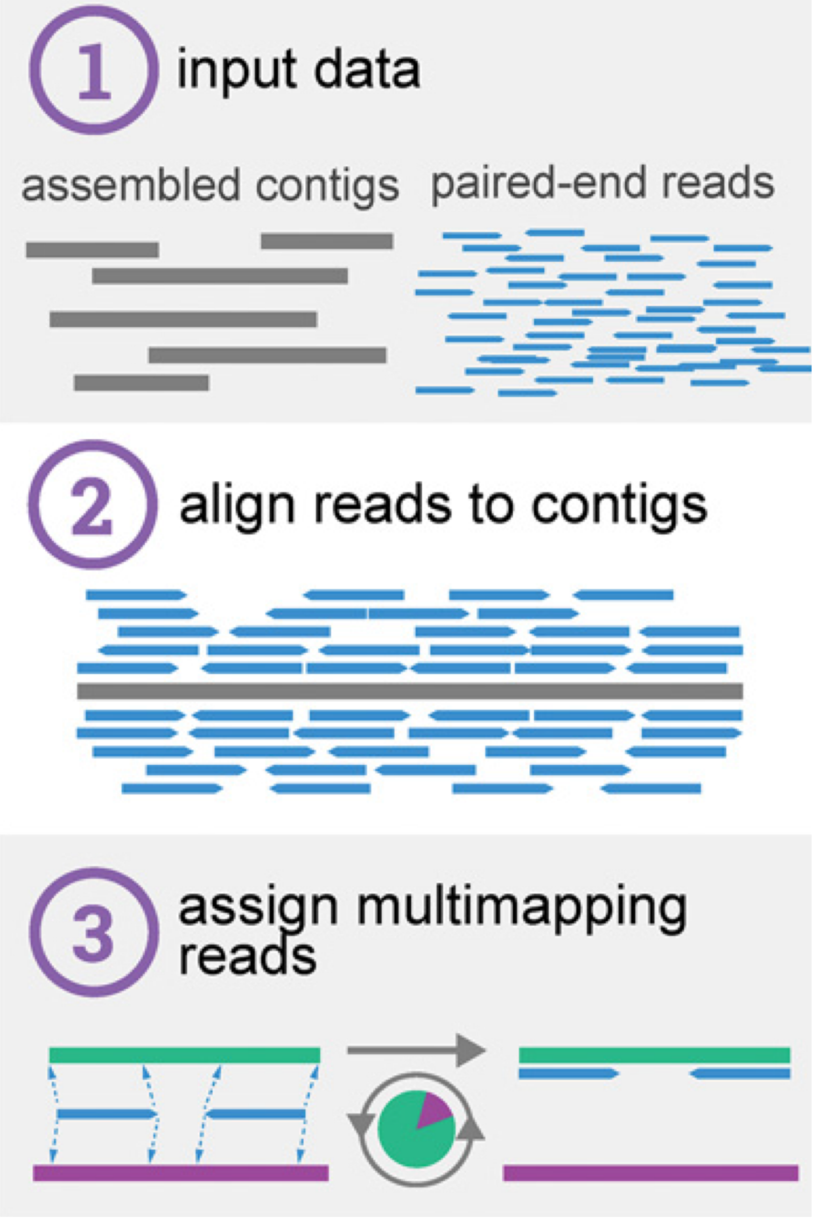
de novo assembly with Trinity

Chimeric transcripts



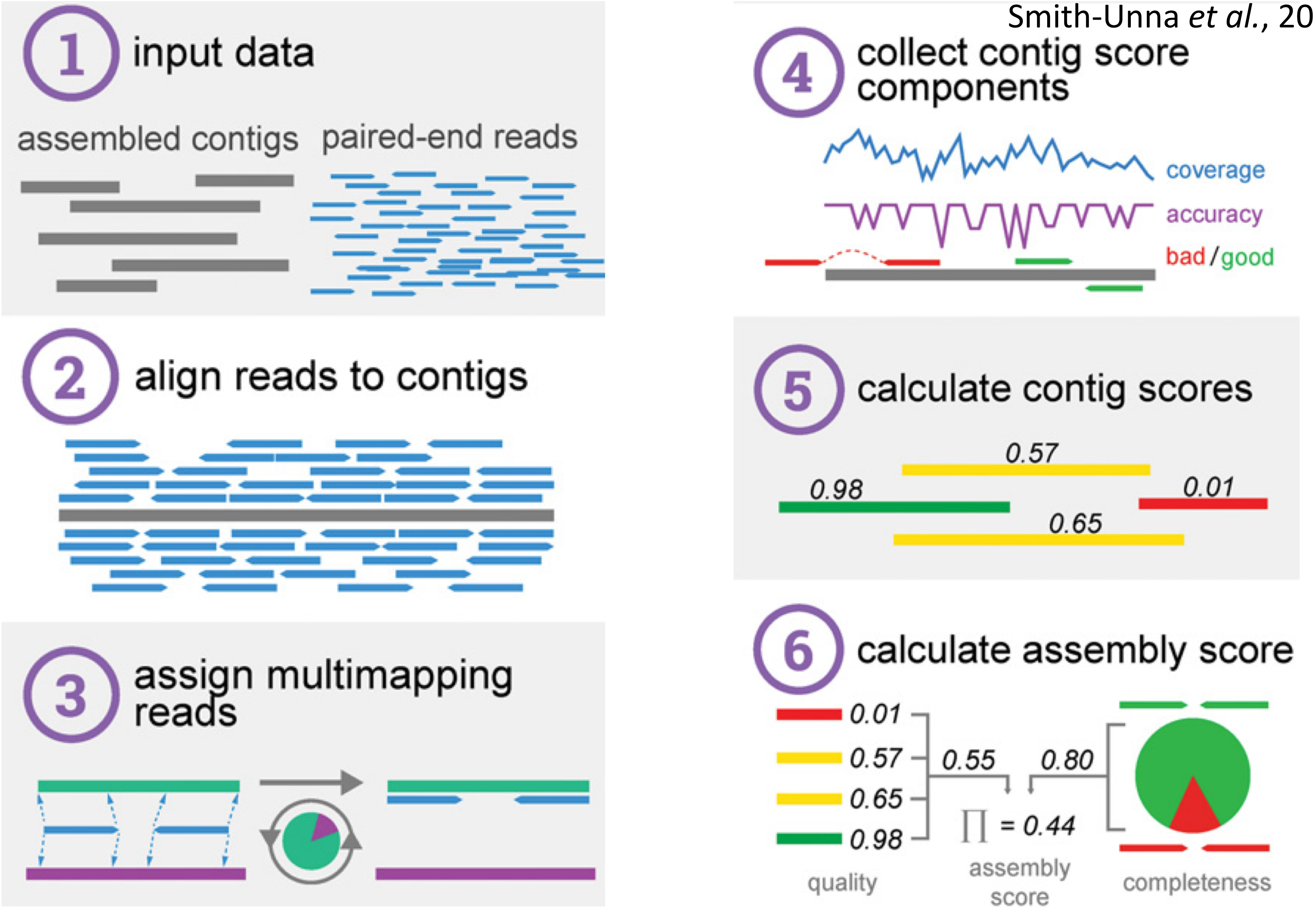
Evaluating assembly by TransRate

Smith-Unna *et al.*, 2016



Evaluating assembly by TransRate

Smith-Unna *et al.*, 2016



Remove transcripts with low support by TransRate

Remove chimeric (Yang and Smith 2013)

Transcript clustering with Corset

- Corset clusters transcripts from the same putative gene based on reads share
- Trinity tend to over cluster. Corset is more accurate. However, for species with polyploidy during the past few years neither work well
- Extract one representative transcript per gene.

TransDecoder for translation

- Build your own BLAST database to guide detection of open reading frames

Arabidopsis thaliana + proteomes from species closely related to your study group

“The **quality of the input data** is more important in determining the quality of a *de novo* assembly than the choice of assembly method that is used. “

Smith-Unna *et al.*, 2016 *Genome Research*

The most important slide

- Transcriptomics is an extremely active research area
- “Best practice” differs by plant group and data set
- Invest time to get familiar with command line and regular expression. Unit test on subset of data