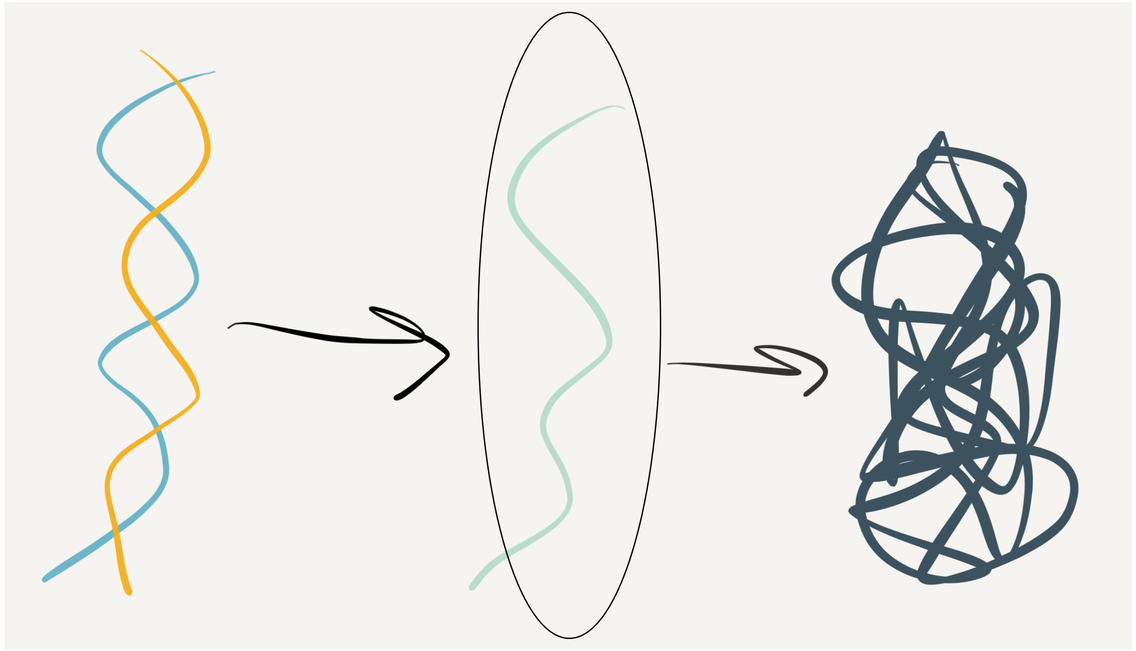
RNAseq in Conservation

March 28, 2023

What is RNAseq?

- sequencing the RNA of an organism at a given point in time
    - captures a moment of gene expression
- Tissue-specific
- Time-specific
- A bit finicky



Why might we want transcripts?

- In recent years - transcriptomes have been a more affordable way of getting a reduced-representation of the genome - less true now
    - good for phylogenetics, population structure, genetic diversity, detecting natural selection, gene family investigations
- Good tool for genome annotation
- Allows you to do functional genomic study (to a point)
- Can use it to assess gene expression (in certain circumstances)

Transcriptomics in endangered species can be tricky

- Only capture RNA at the moment of collection
- Often requires sacrifice of the individual
    - Wait until you find a dead one?
        - … must be relatively fresh
        - will only get the RNA that was being expressed leading up to death

Three main approaches to RNAseq

Short (Illumina):

- require assembly
- inexpensive
- might be fine depending on application
- less high-quality RNA is okay
- need to make cDNA

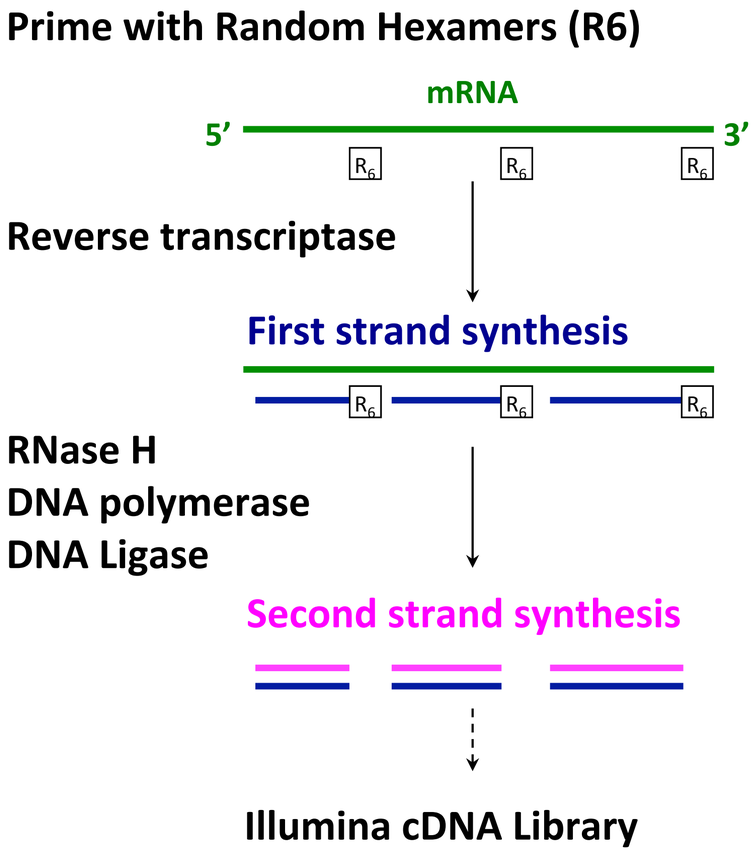
Long PacBio IsoSeq

- no assembly
- getting less expensive
- better isoform detection
- discover long non-coding RNA
- need to make cDNA

Long Nanopore direct RNAseq

- no assembly
- avoid amplification biases
- can look at RNA modifications - epitranscriptome
- sequence RNA directly

RNA-Seq: How do we make cDNA?



Slide courtesy of Joshua Levin, Broad Institute.

Transcriptomes != Genomes

Genome

- One large assembly per chromosome
- Single contig per locus
- Double-stranded
- Uniform coverage

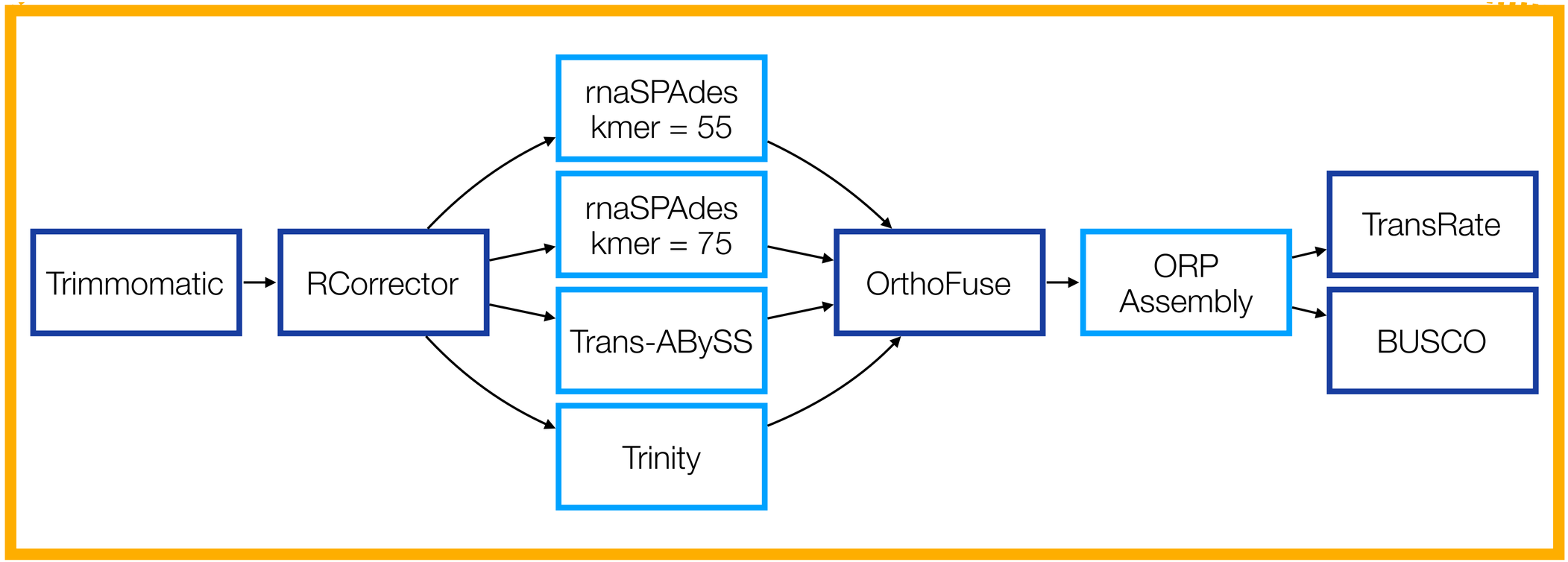
Transcriptome

- Thousands of small assemblies
- Multiple contigs per locus (alternative splicing)
- Strand-specific
- Exponentially distributed coverage levels

Transcriptome assembly is not always straightforward

- Not a deterministic process
- kmer tradeoffs
    - Long kmers yield better quality common transcripts and transcripts with repetitive regions
    - Short kmers are better at picking up rare transcripts
- Good strategy is multiple assemblies

The Oyster River Protocol



ORP Table

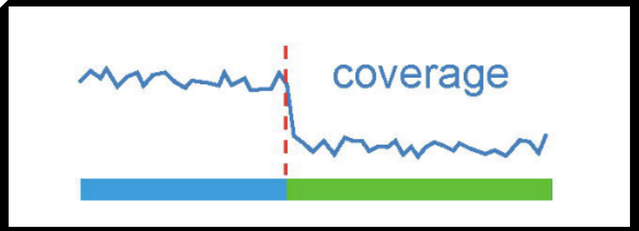
In the event that the ORP is impossible to install…

a merged assembly is still a good idea.

- Trans-ABySS is a solid option
    - can do multiple assemblies at different kmer values
    - has a built-in merge command
- Trinity is a solid option
    - has lots of backend commands so you can easily continue the analysis
    - watch out for huge files if it fails
    - good busco scores

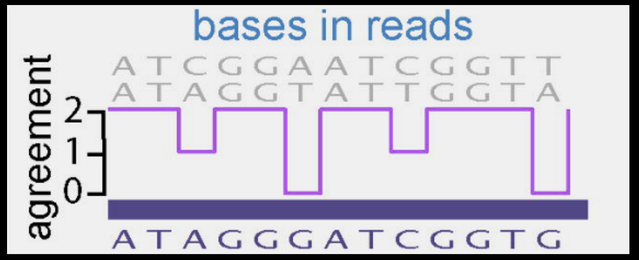
Tutorial

Chimerism

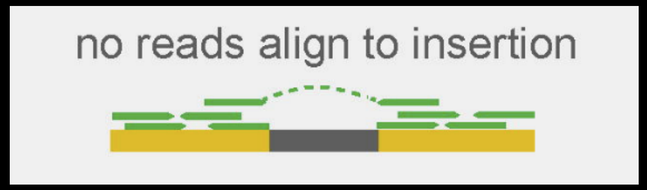


Incompleteness

Family Collapse

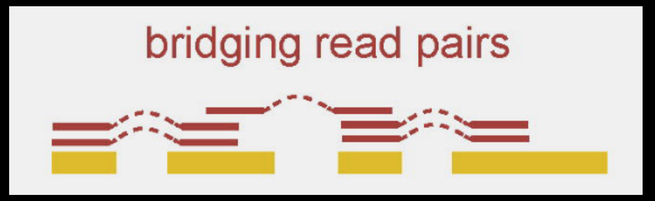


Unsupported Insertion



Redundancy

Fragmentation



Local Misassembly

Smith-Unna, R., Boursnell, C., Patro, R., Hibberd, J. M., & Kelly, S. (2016). TransRate: reference-free quality assessment of de novo transcriptome assemblies. *Genome research*, *26*(8), 1134-1144.