

RNA-Seq

Bioinformatics Applications (PLPTH813)

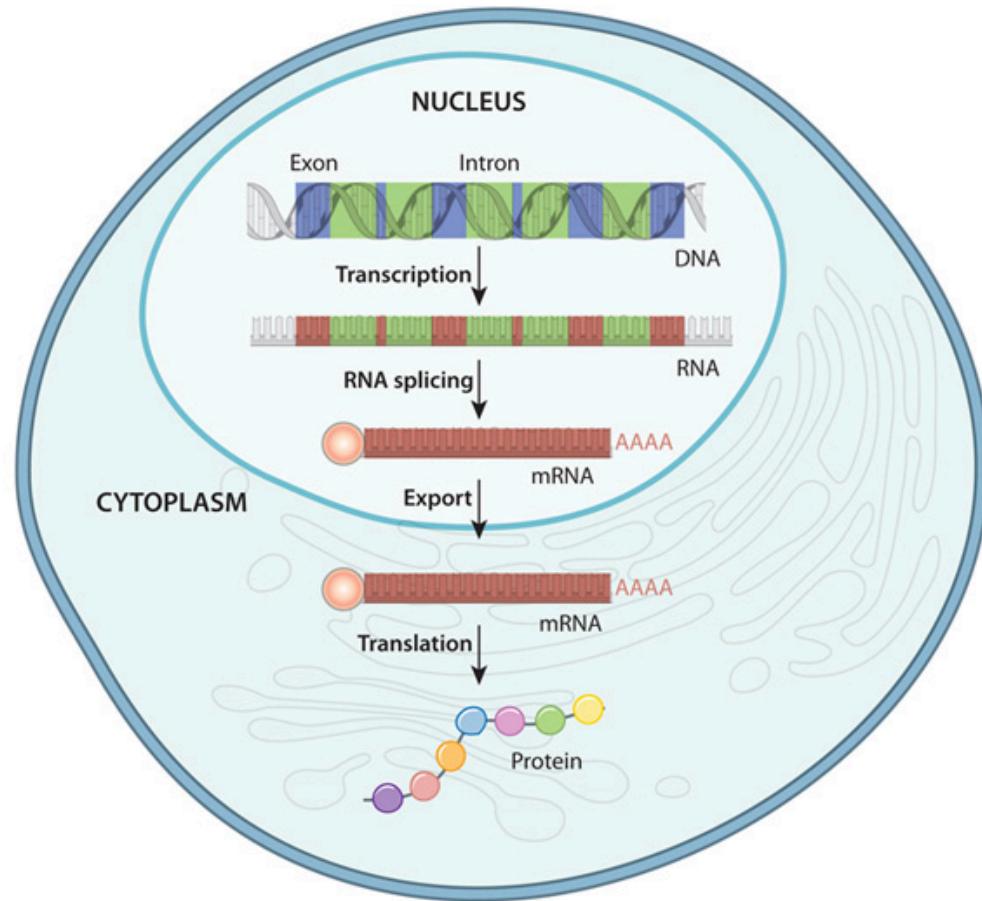
Sanzhen Liu

3/30/2017

Outline

- Introduction of RNA-Seq
- RNA-Seq procedure
- Reference guided assembly
- RNA-Seq *de novo* assembly
- PacBio Iso-seq

Transcriptome



DNA to protein in eukaryote

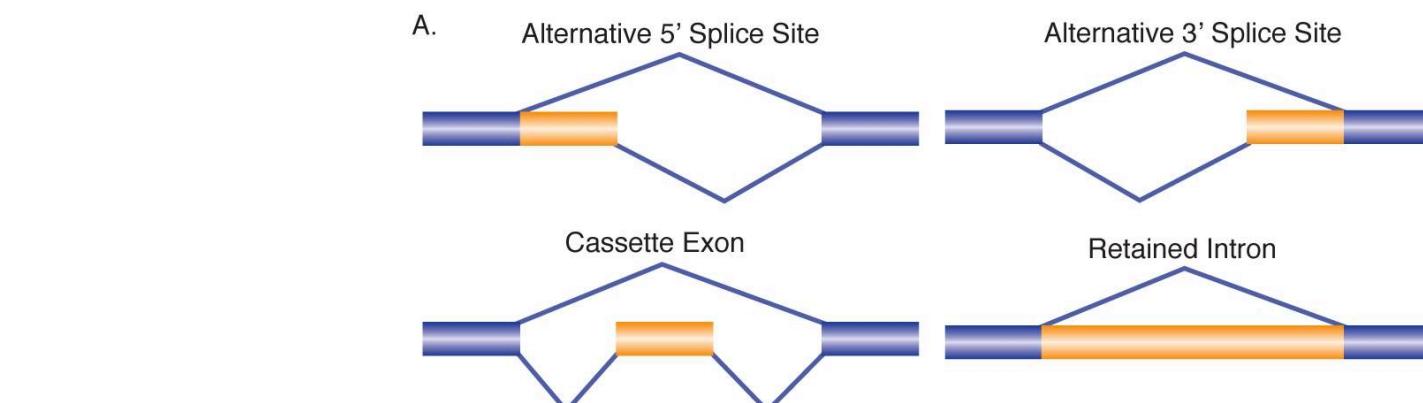
nature.com/scitable/topicpage/gene-expression-14121669

Alternative splicing

Genome

Pre-mRNA 5' exon intron AAAAAA 3'

mRNA transcript



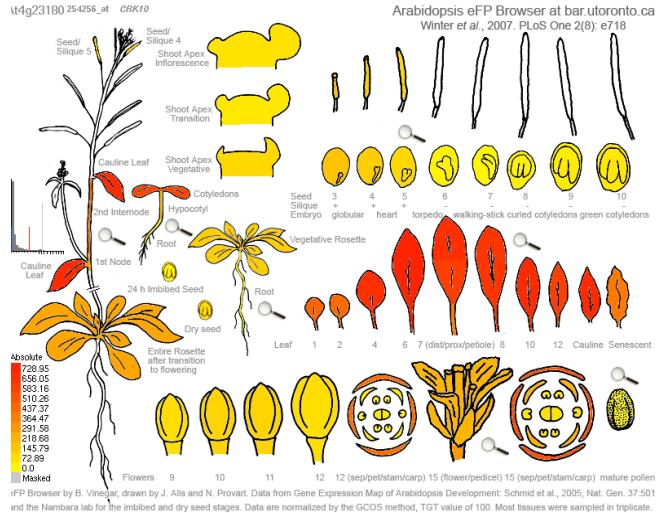
Complexity of transcriptome

In many eukaryotic organisms, the majority of genes are alternatively spliced to produce multiple transcript isoforms.

In humans, for example, there is evidence for alternative splicing of more than 95% of genes, with an average of more than five isoforms per gene.

- Tilgner et al. (2014) PNAS

Transcriptome analysis



Expression profiles in different tissues



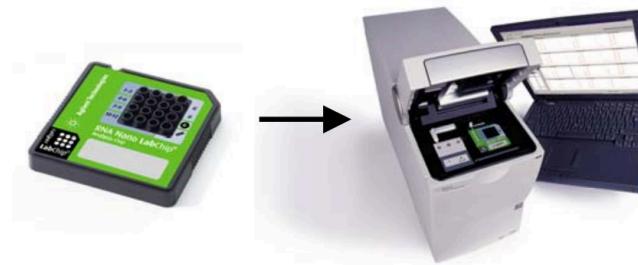
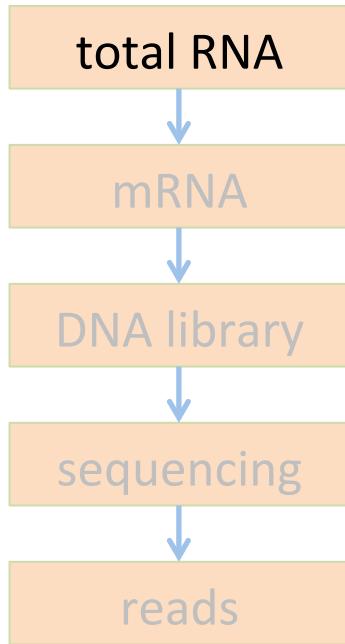
Response to biotic stress

1. What are sequences of transcripts?
2. What is the expression level of each transcript?

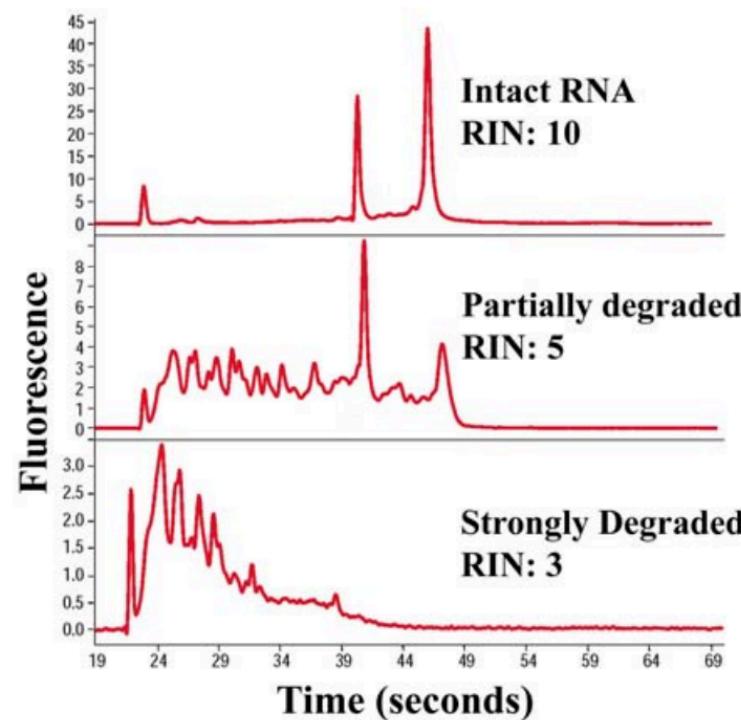
RNA-Seq addresses both questions pretty well

total RNA

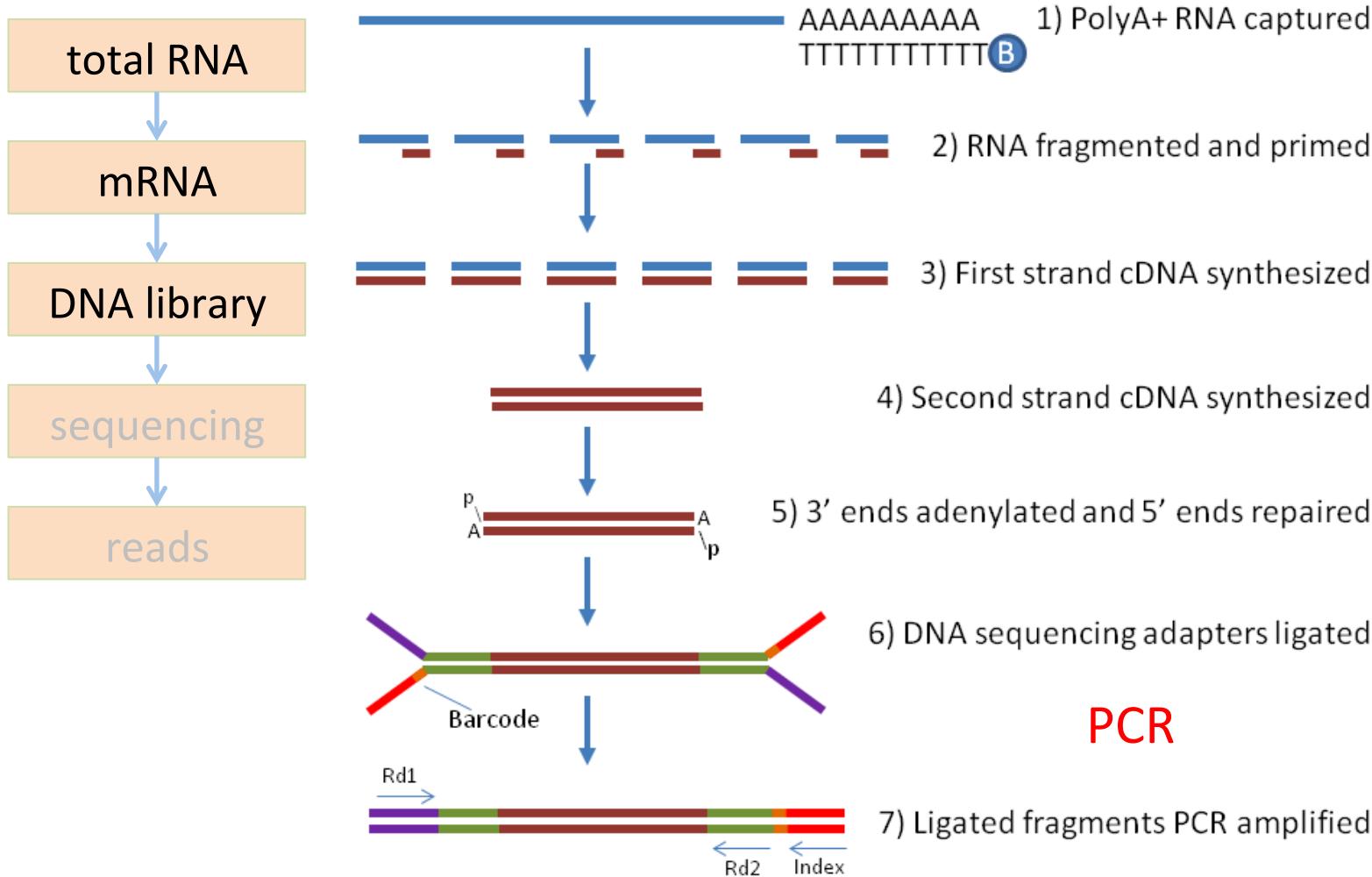
RNA-Seq



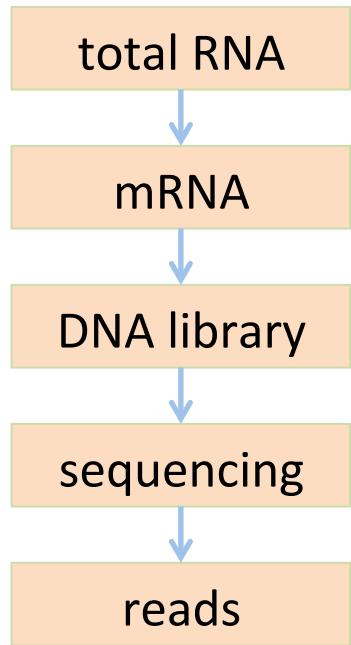
RIN: RNA Integrity Number



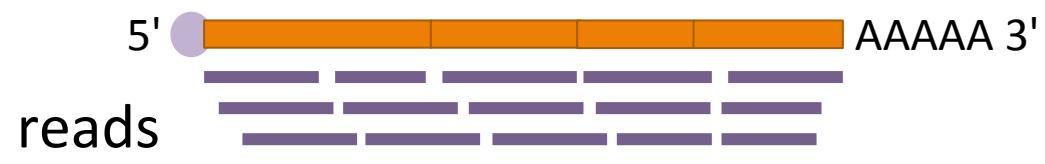
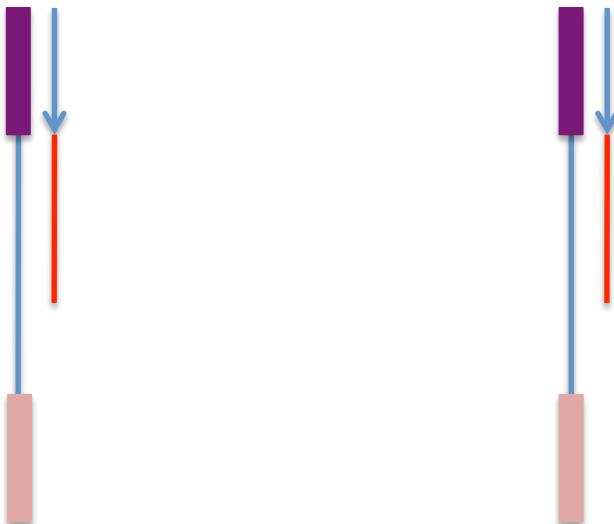
Illumina RNA-Seq library preparation



Illumina sequencing



Single-end reads Paired-end reads

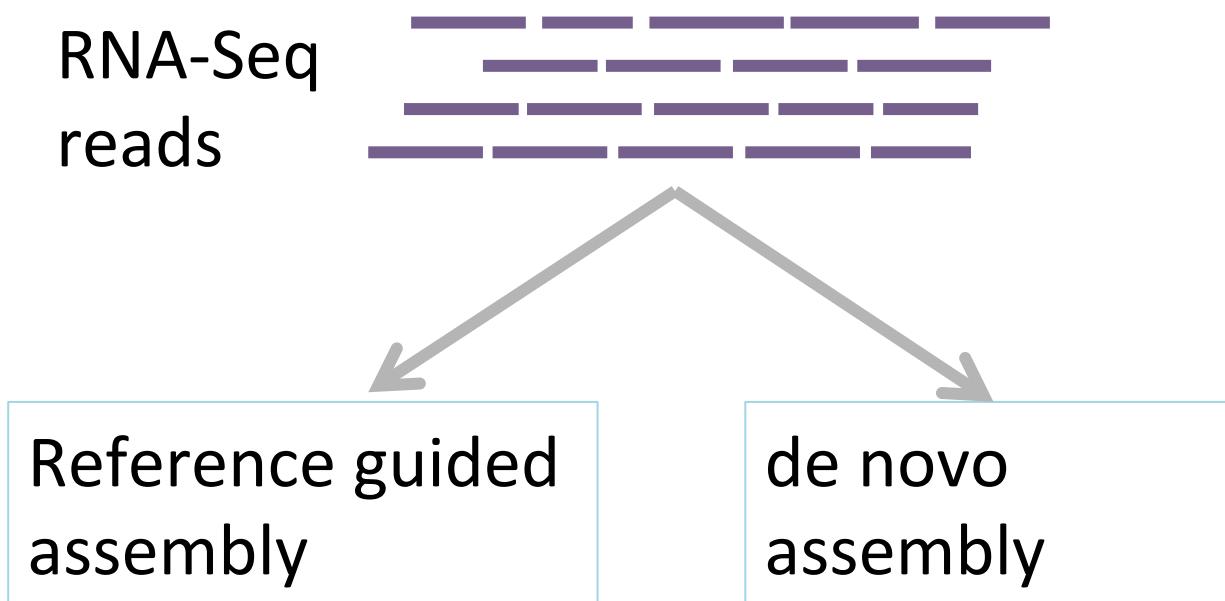


How to reconstruct full-length transcripts from short reads?

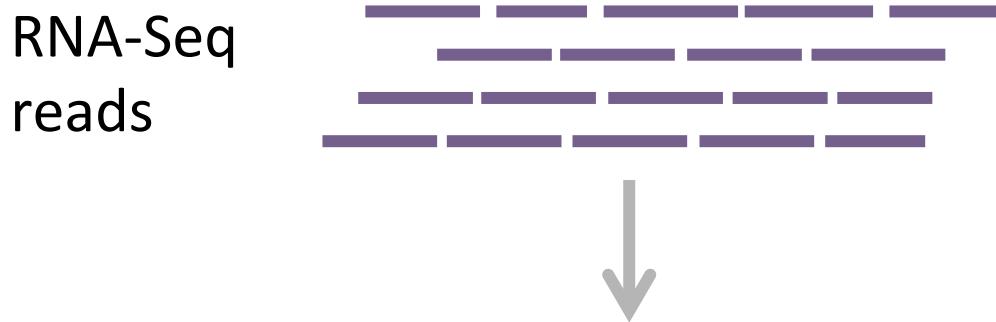
Challenges of reconstruction of full-length transcripts (transcriptome assembly)

1. Sequencing errors
2. Repeats in different genes
3. Various coverage on different transcripts
4. Read coverage may be uneven across the transcript's length, owing to sequencing biases
5. Alternative splicing greatly complicates transcriptome assembly

Two main strategies for transcriptome assembly



Reference-guided transcriptome assembly

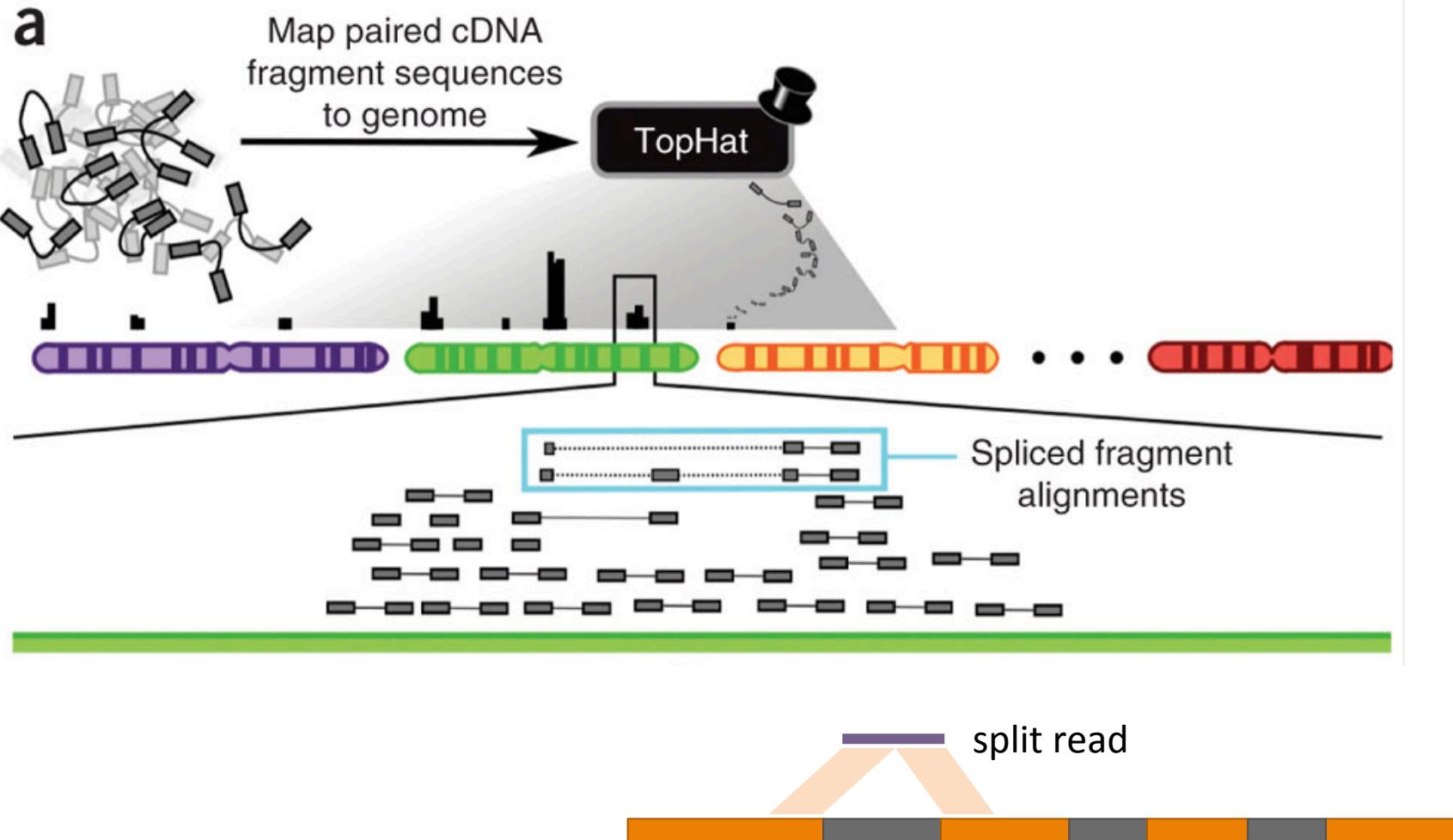


Reference guided assembly (Cufflinks)

1. alignment to the reference genome
2. assembly based on alignments

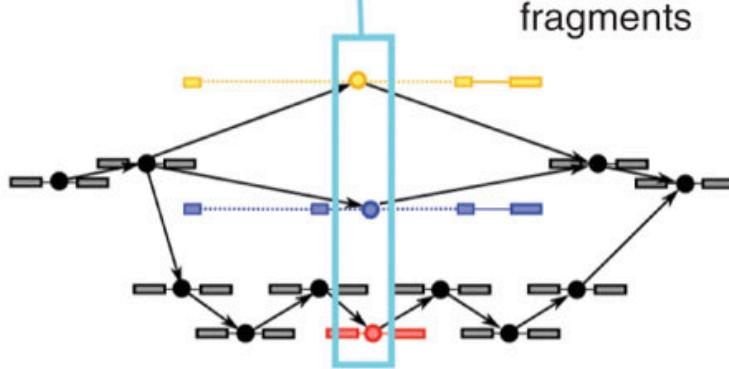
Question: Using a genome as the reference, what difference between RNA-Seq alignments and DNA-Seq alignments?

Cufflinks - spliced alignments

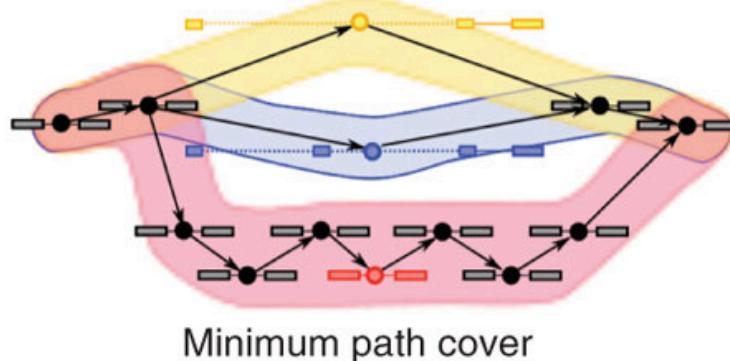


b

Assembly

Mutually
incompatible
fragments

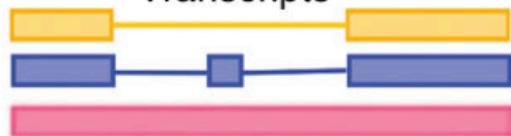
Overlap graph

**c**

Minimum path cover



Transcripts



Cufflinks - assembly

Incompatible fragments are fragments originated from distinct spliced mRNA isoforms

Overlapping graph is made up with compatible fragments (nodes) and alignments overlap in the genome

Isoforms are assembled from the overlap graph

Pros and cons of reference-guided assembly

Theoretically, reference-guided assembly promises **maximum sensitivity**

The quality of assembly is **depended on correct read-to-reference alignments**, which are complicated by splicing, sequencing errors and the lack or incompleteness of many reference genomes

Alignments are even more complicated when **the species used to construct a reference genome is distant from the species** that is used for constructing a reference genome

De novo assembly

De novo transcriptome assembly is to perform an assembly from scratch, which does not rely on read-reference alignments.

Important when the genomic sequence is not available, is gapped, highly fragmented or substantially different with the reference genome.

Trinity

a



b



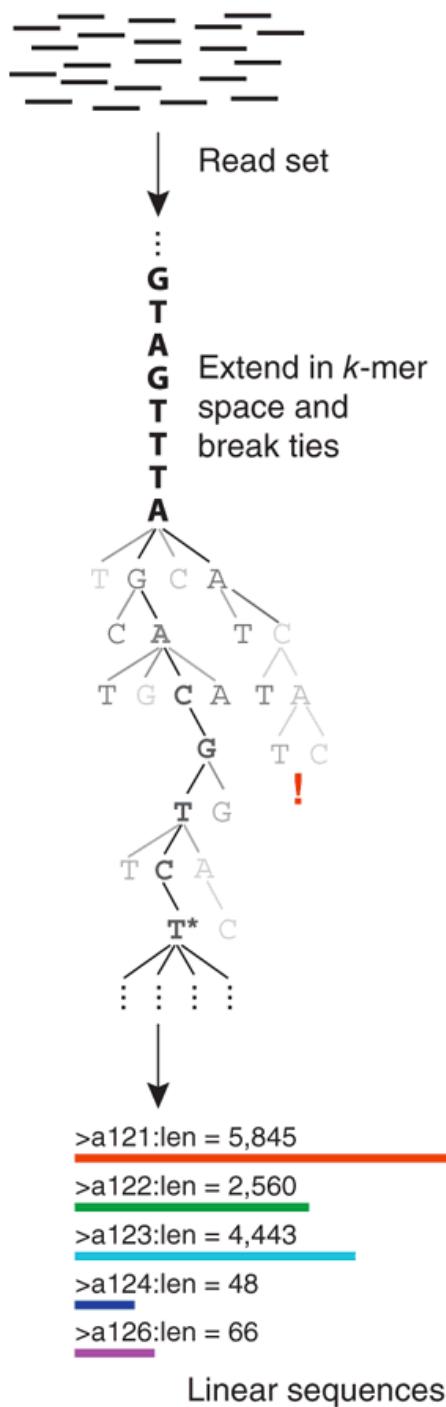
c



step 1: *Inchworm* assembles reads into unique sequences of transcripts (**reads to contigs**)

step 2: *Chrysalis* clusters related contigs (**contigs to clusters and then components**)

step 3: *Butterfly* analyzes the paths taken by reads in the context of the corresponding de Bruijn graph and reports all plausible transcripts (**components + reads to transcripts**)

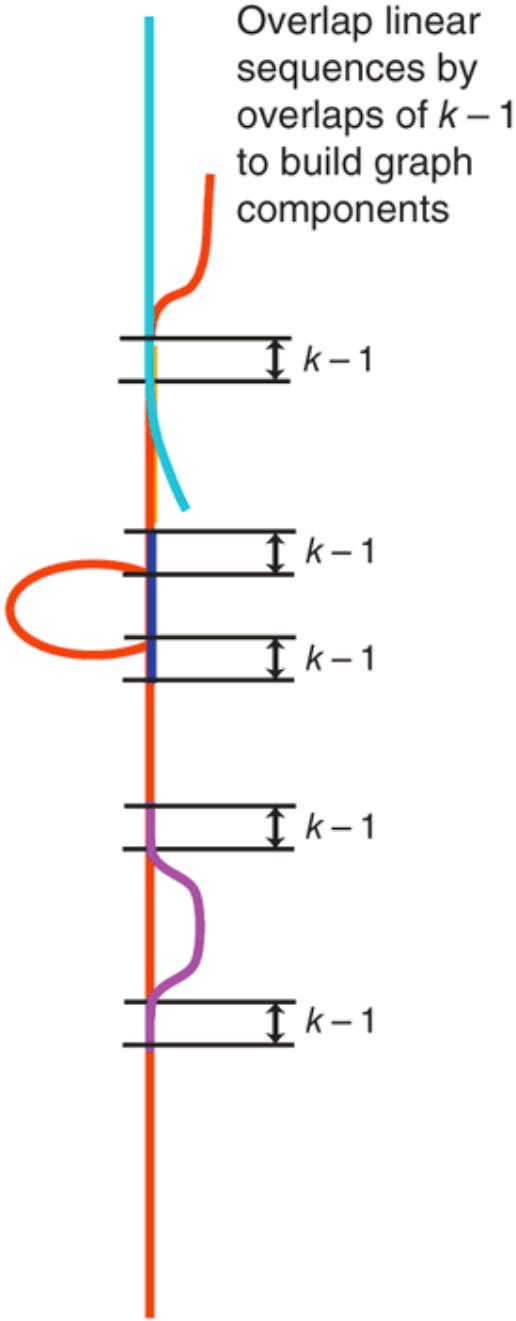


Trinity - Inchworm

step 1: Inchworm assembles reads into the unique sequences of transcripts.

Inchworm uses a k -mer-based approach for transcript assembly, recovering only a single (best) representative for a set of alternative variants that share k -mers

The contigs alone do not capture the full complexity of the transcriptome

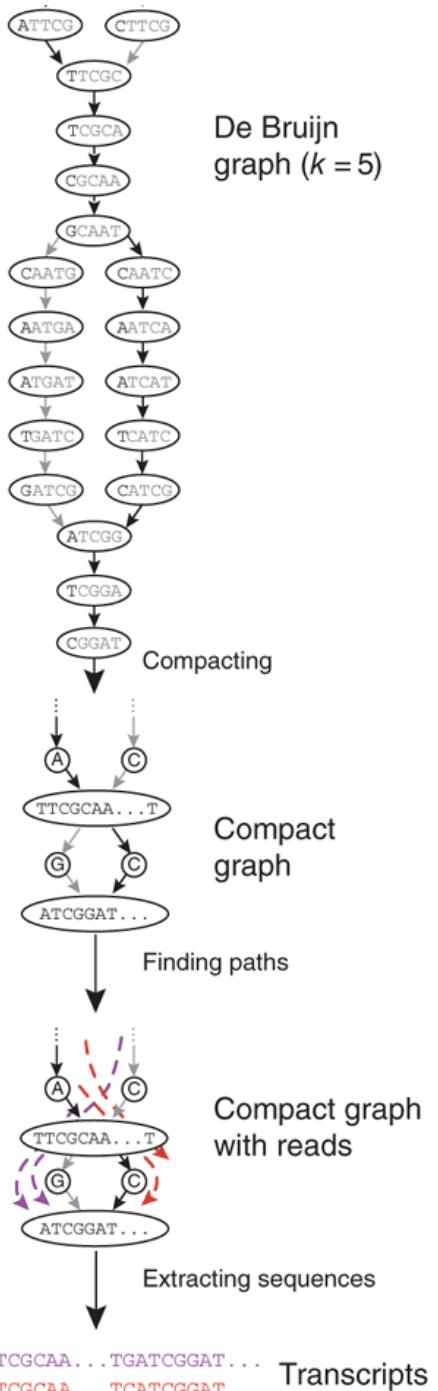


Trinity - Chrysalis

step 2: **Chrysalis clusters related contigs** that correspond to portions of alternatively spliced transcripts or otherwise unique portions of paralogous genes.

Chrysalis then **constructs a de Bruijn graph** for each cluster of related contigs

Determine **components**. Each component defines a collection of contigs that are likely to be derived from alternative splice forms or closely related paralogs.



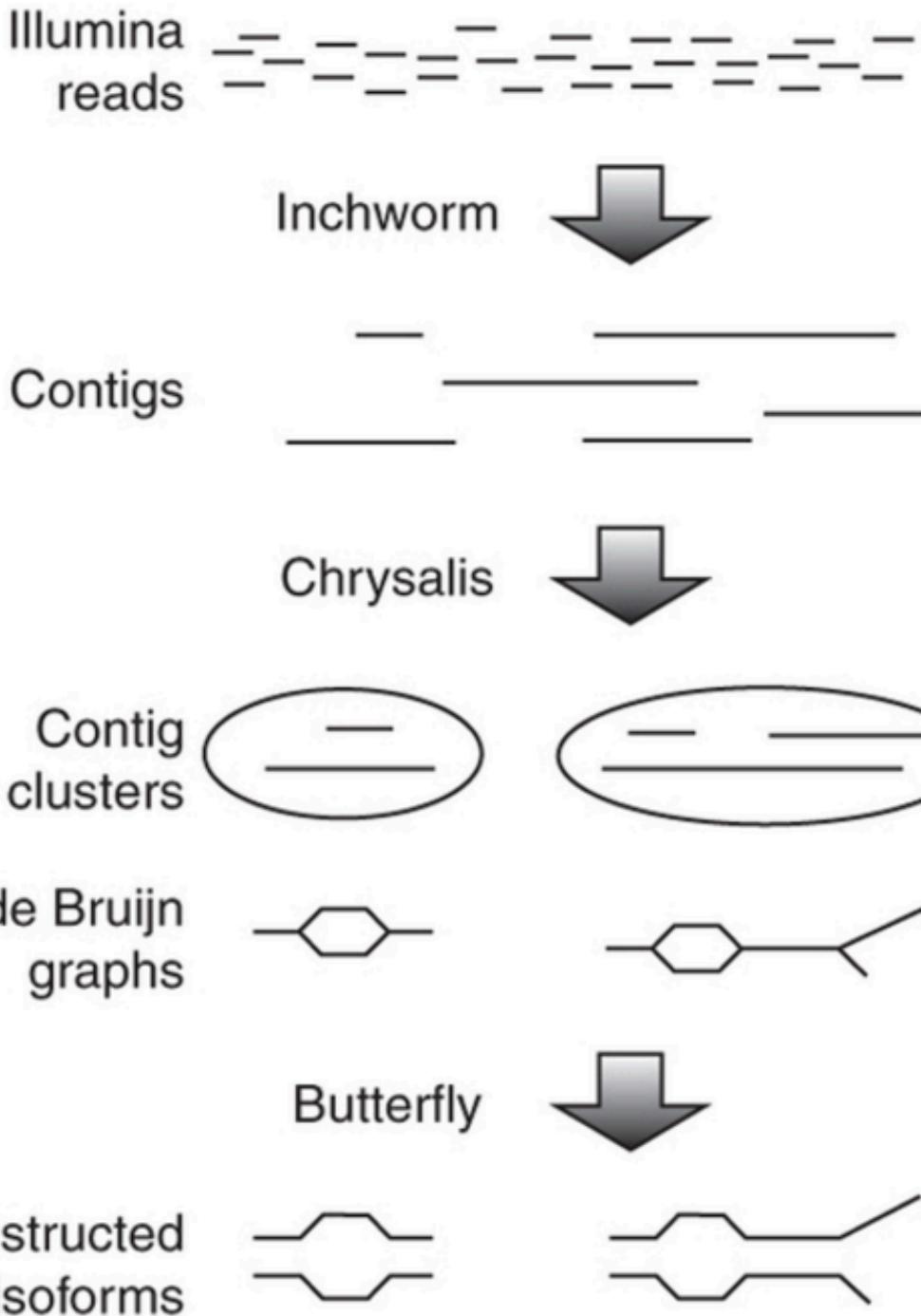
Trinity - Butterfly

step 3: Butterfly reconstructs plausible, full-length, linear transcripts by reconciling the individual de Bruijn graphs generated by Chrysalis **with the original reads and paired ends.**

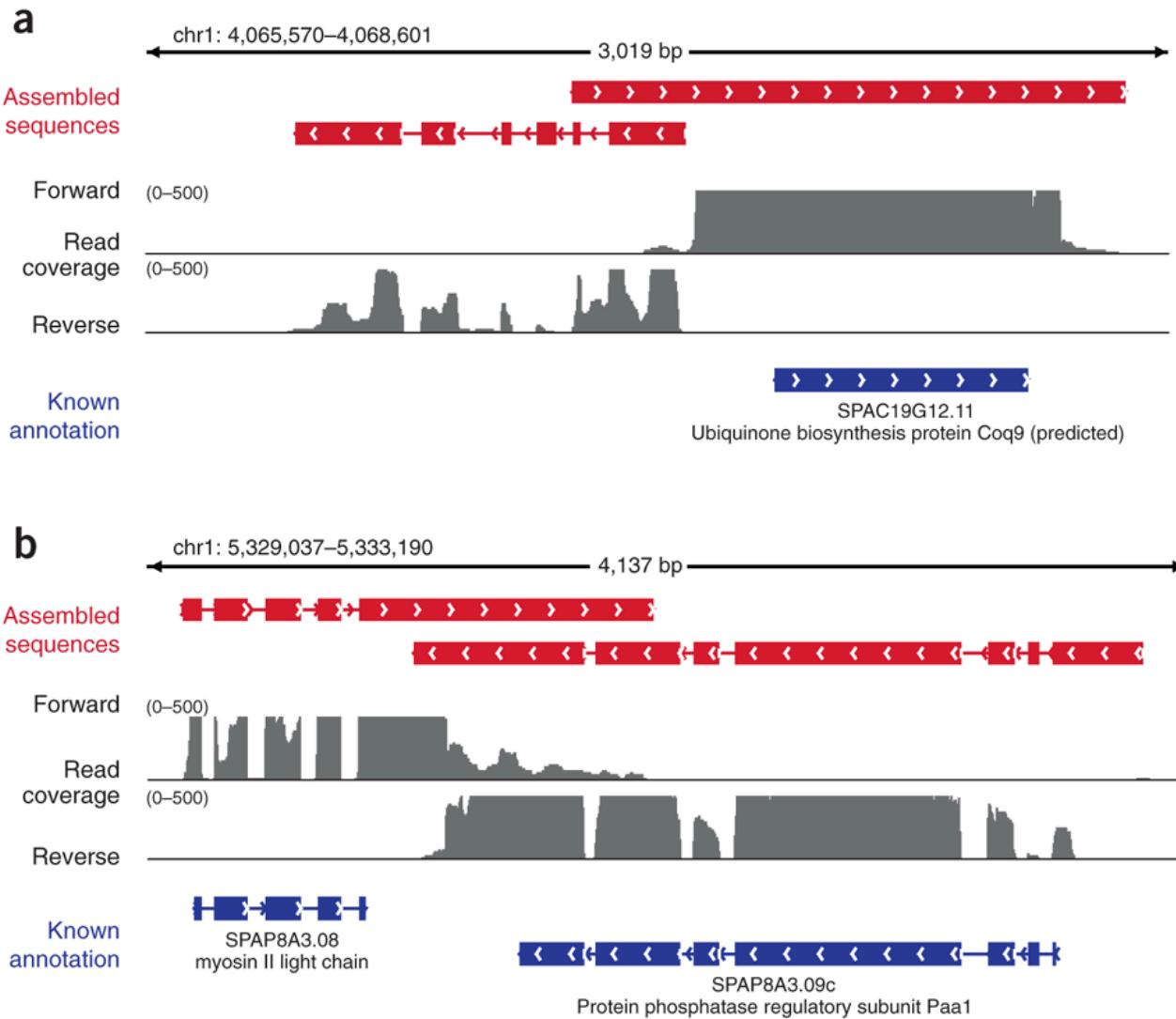
Butterfly resolves alternatively spliced isoforms and transcripts derived from paralogous genes.

...CTTCGCAA...TGATCGGAT... Transcripts
...ATTAGCAA...TCATCGGAT...

Trinity summary



De novo assembly could correct wrong annotations



Note

Trinity also provides an option to perform reference-guided transcriptome assembly,

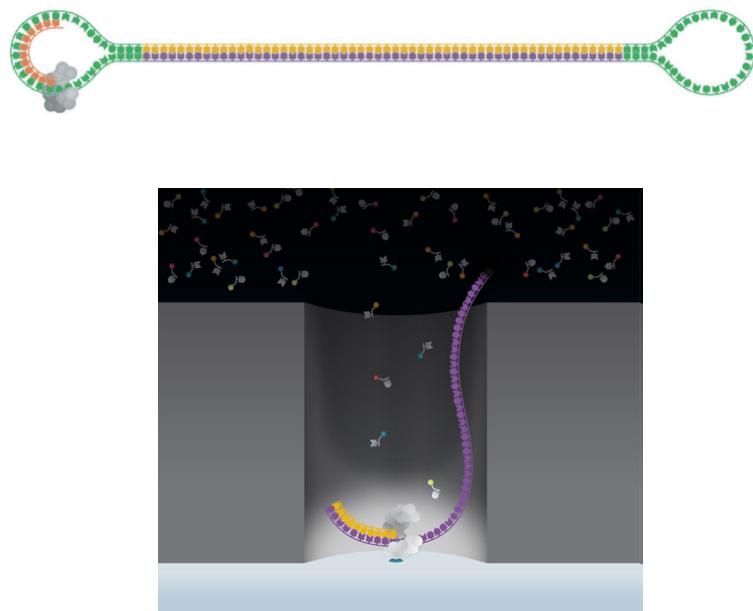
<https://github.com/trinityrnaseq/trinityrnaseq/wiki/Genome-Guided-Trinity-Transcriptome-Assembly>

Outline

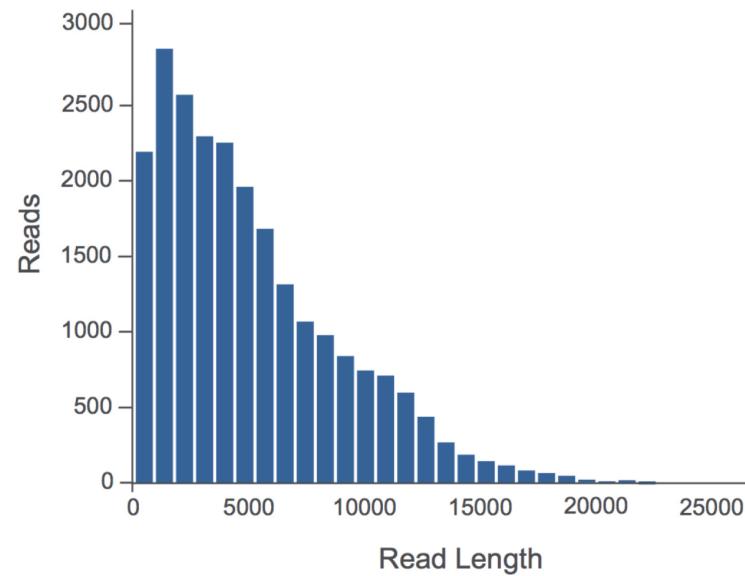
- Introduction of RNA-Seq
- RNA-Seq procedure
- Reference guided assembly
- RNA-Seq *de novo* assembly
- PacBio Iso-seq

PacBio long reads for RNA sequencing (Iso-Seq)

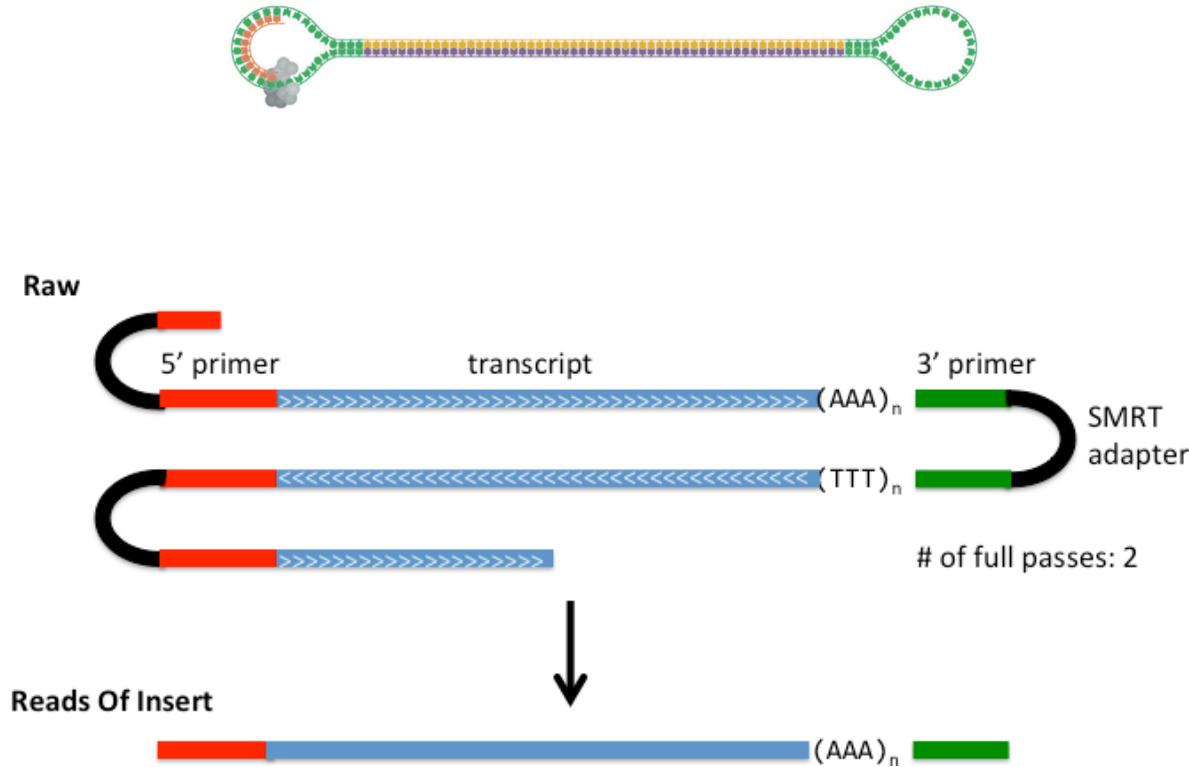
- High-quality, single-molecule, circular-consensus (CCS)



Read Length Distribution



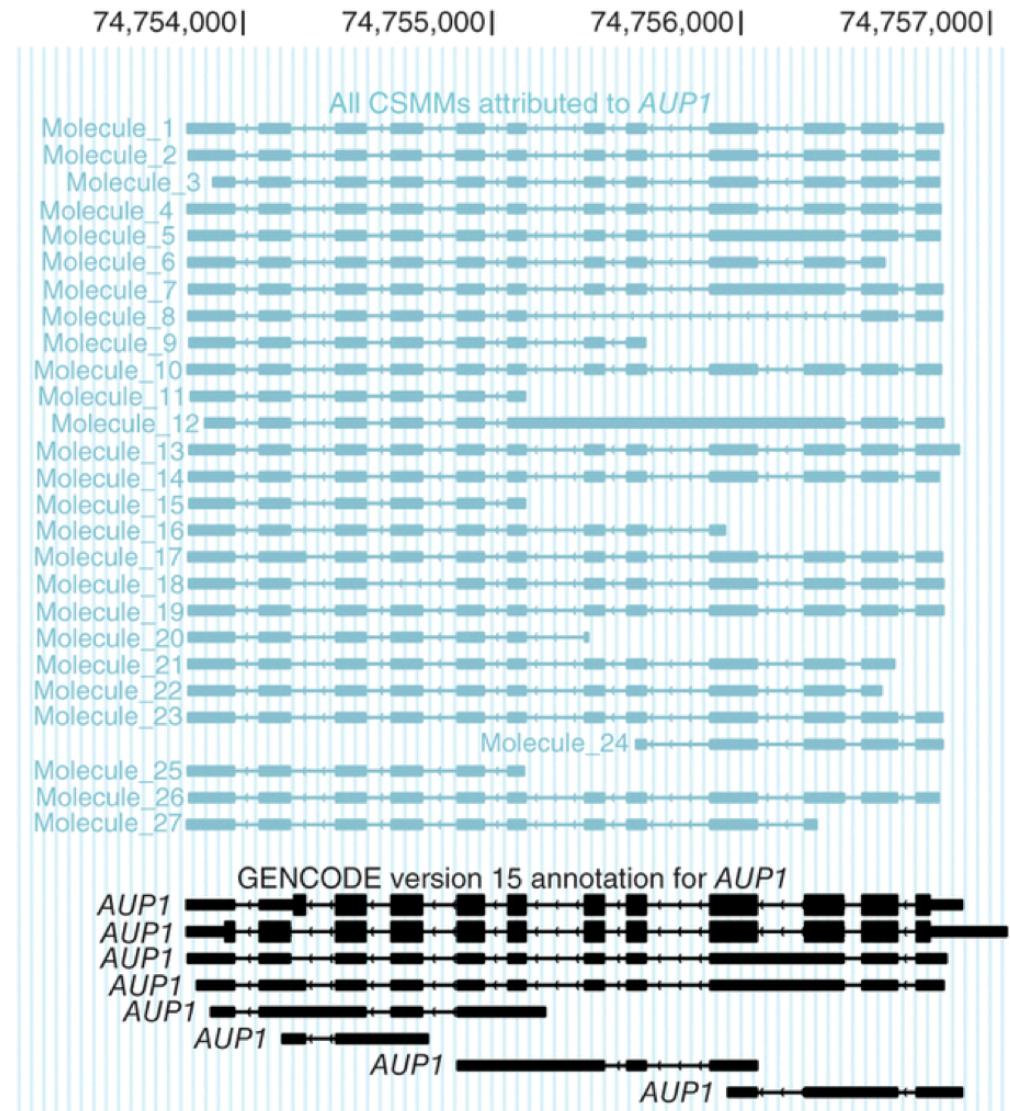
Multiple passes improve sequence quality



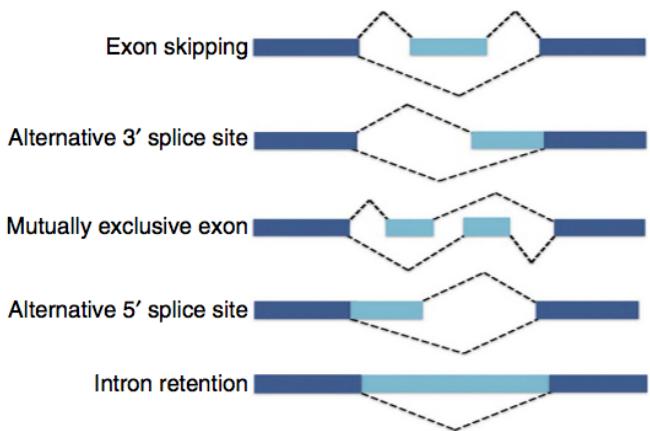
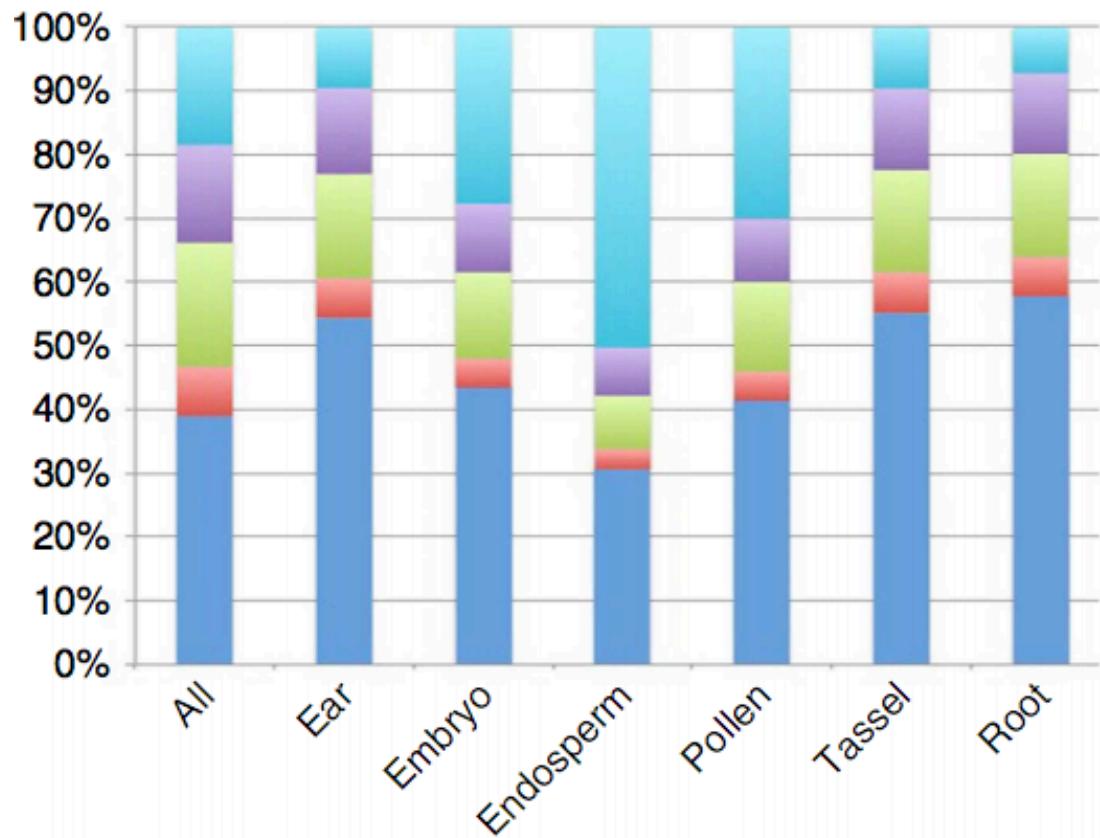
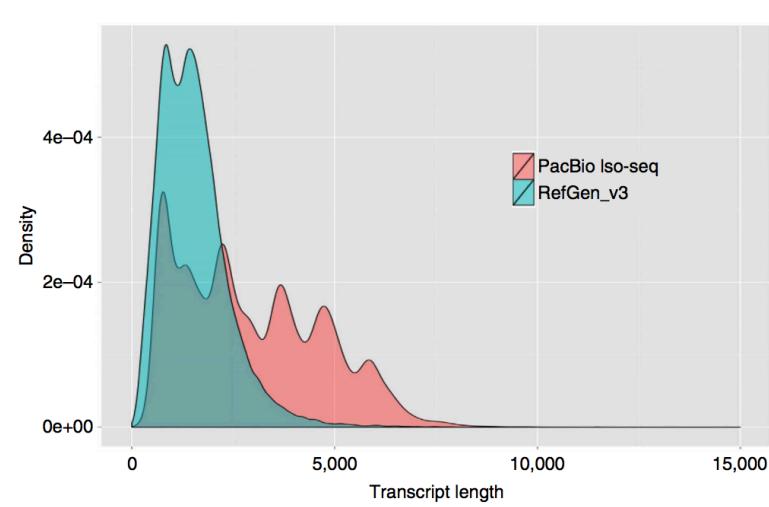
https://github.com/PacificBiosciences/cDNA_primer/wiki/Understanding-PacBio-transcriptome-data

Long reads to sequence full-length cDNA

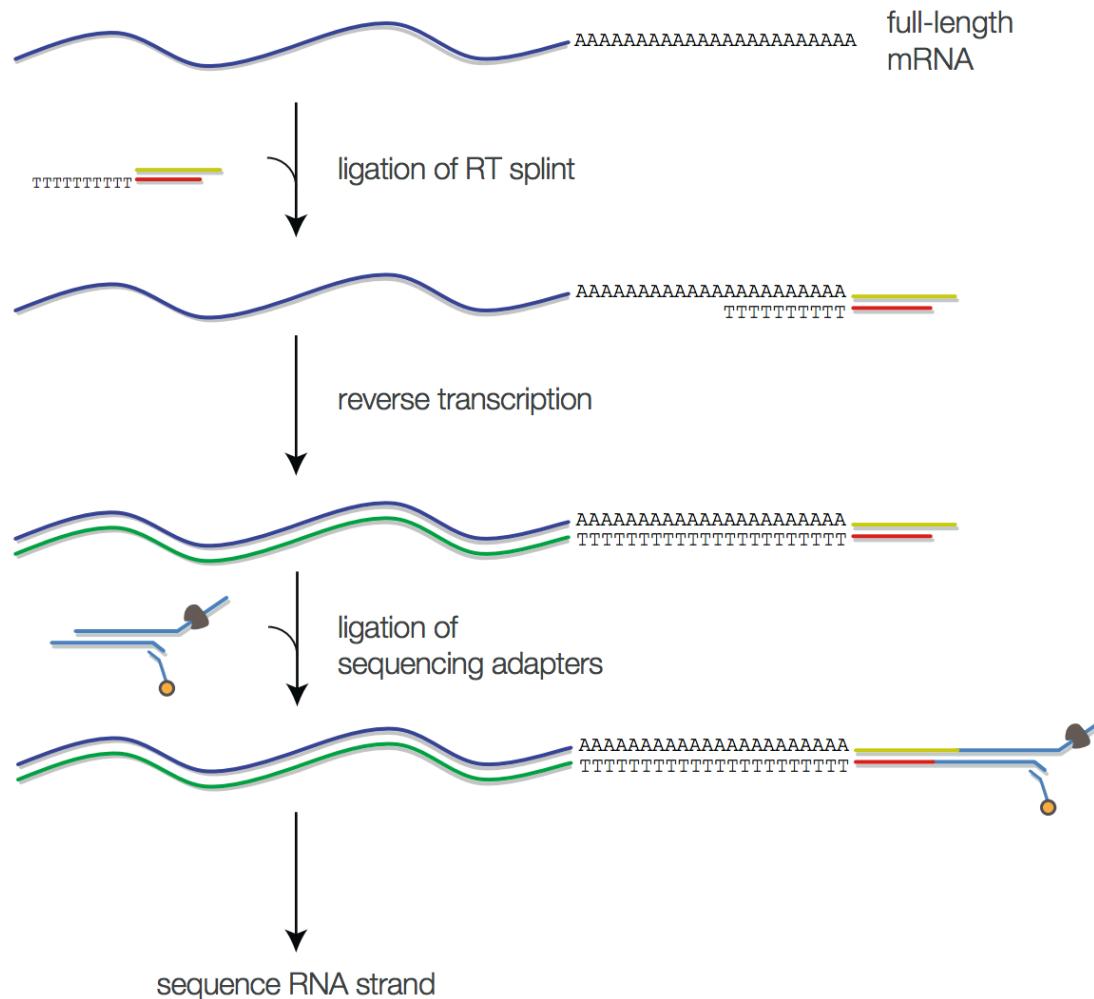
- The majority of reads represent all splice sites of original transcripts
- Isoforms can be monitored at a single-molecule level without amplification or fragmentation



maize – Iso-Seq



Nanopore – direct RNA sequencing



References

Cufflinks paper: Trapnell C et al., Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol.* 2010, 28, 511–5.

Tophat-Cufflinks-Cuffdiff-CummeRbund protocol: Trapnell C et al., Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat Protocols.* 2012, 7, 562–578.

Trinity paper: Grabherr MG et al., Full-length transcriptome assembly from RNA-seq data without a reference genome. *Nat Biotechnol.* 2011, 29:644-52.

Protocol for using Trinity: Haas BJ et al., De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat Protoc.* 2013, 8:1494-512.

Performance tuning of Trinity: Henschel R et al., Trinity RNA-Seq assembler performance optimization. XSEDE 2012 Proceedings of the 1st Conference of the Extreme Science and Engineering Discovery Environment: Bridging from the eXtreme to the campus and beyond.