RNA-Seq

Bioinformatics Applications (PLPTH813)

Sanzhen Liu

3/30/2017

Schedule

6-Apr	Th 10:30-11:20	RNA-Seq and RNA-Seq differential expression
11-Apr	Tu 10:30-11:20	RNA-Seq differential expression
13-Apr	Th 10:30-11:20	Co-expression network (Hairong Wei; MTU)
18-Apr	Tu 9:30-10:20	Genome assembly (Illumina)
20-Apr	Th 12:30-1:30	Paper presentation - RNA-Seq (3 students)

Projects

- In-class mini projects (from reads to results)

1. RNA-Seq
2. SNP discovery

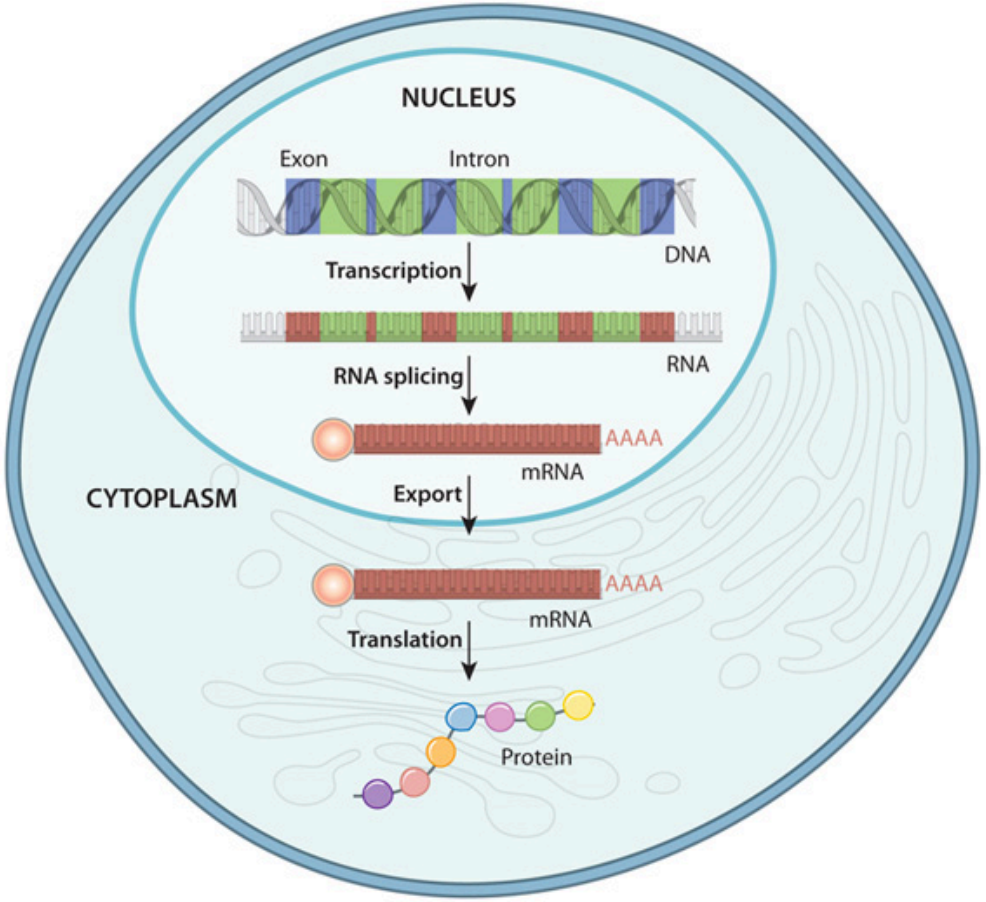
- a little more Unix and some tips for data analysis

- project reports (5-10 minutes from each person)

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

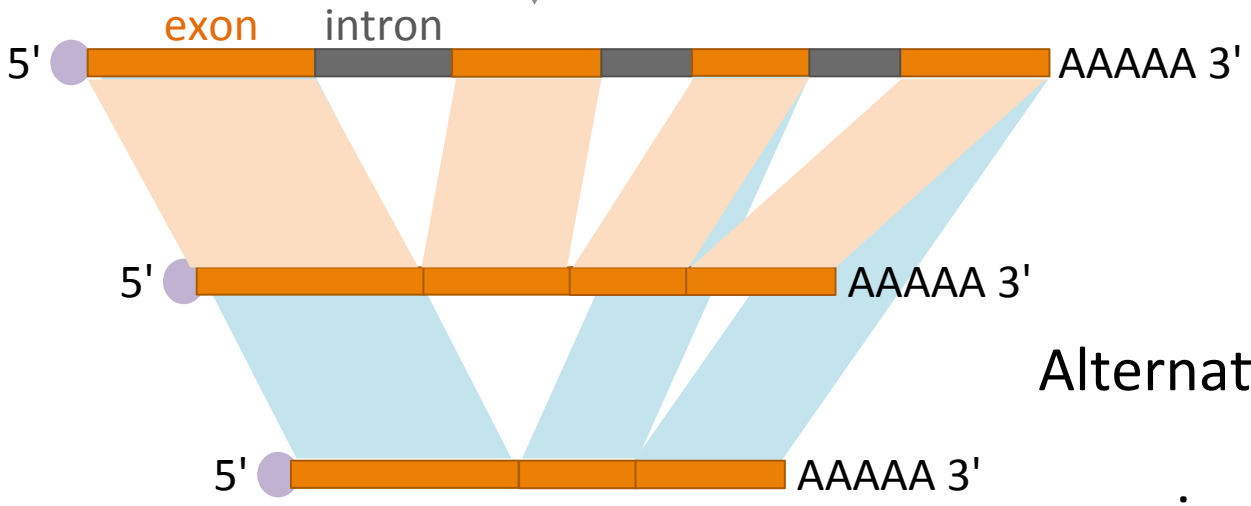
Alternative splicing

Genome

transcription

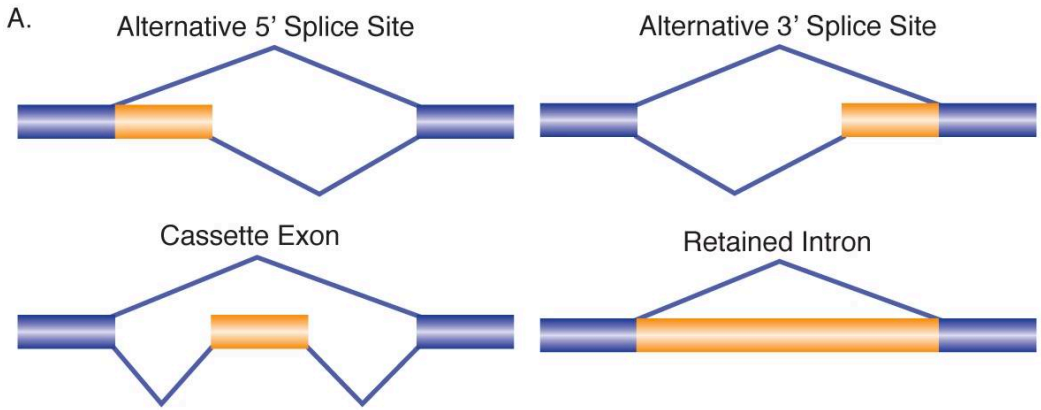
Pre-mRNA

mRNA transcript



Alternative splicing

isoforms



Curr Opin Genet Dev. 21(4): 373–379 (2011)

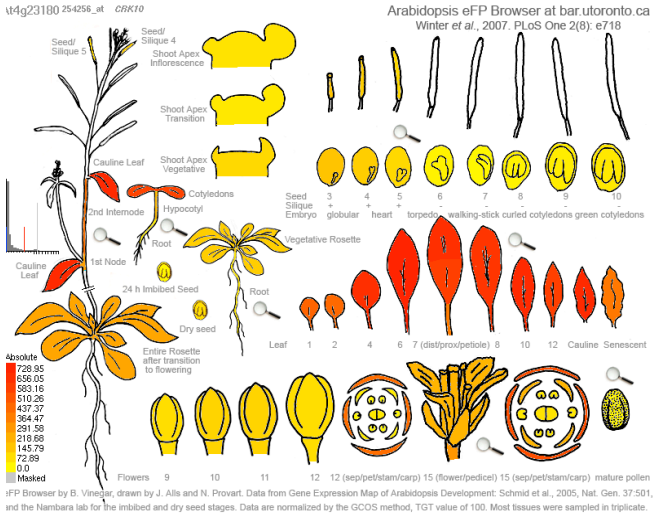
Complexity of transcriptome

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

RNA-Seq

total RNA

↓

mRNA

↓

DNA library

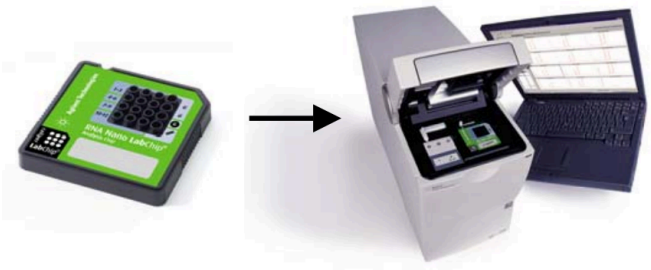
↓

sequencing

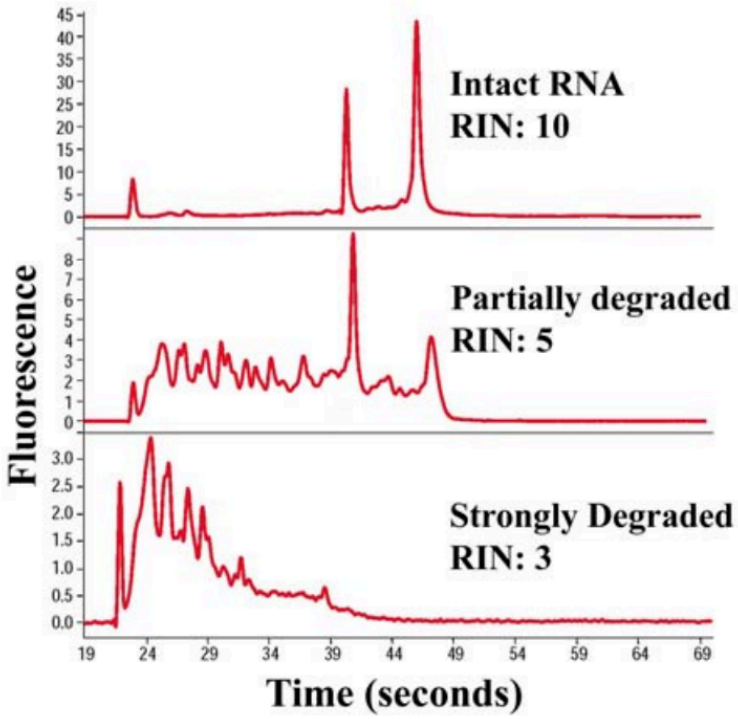
↓

reads

total RNA

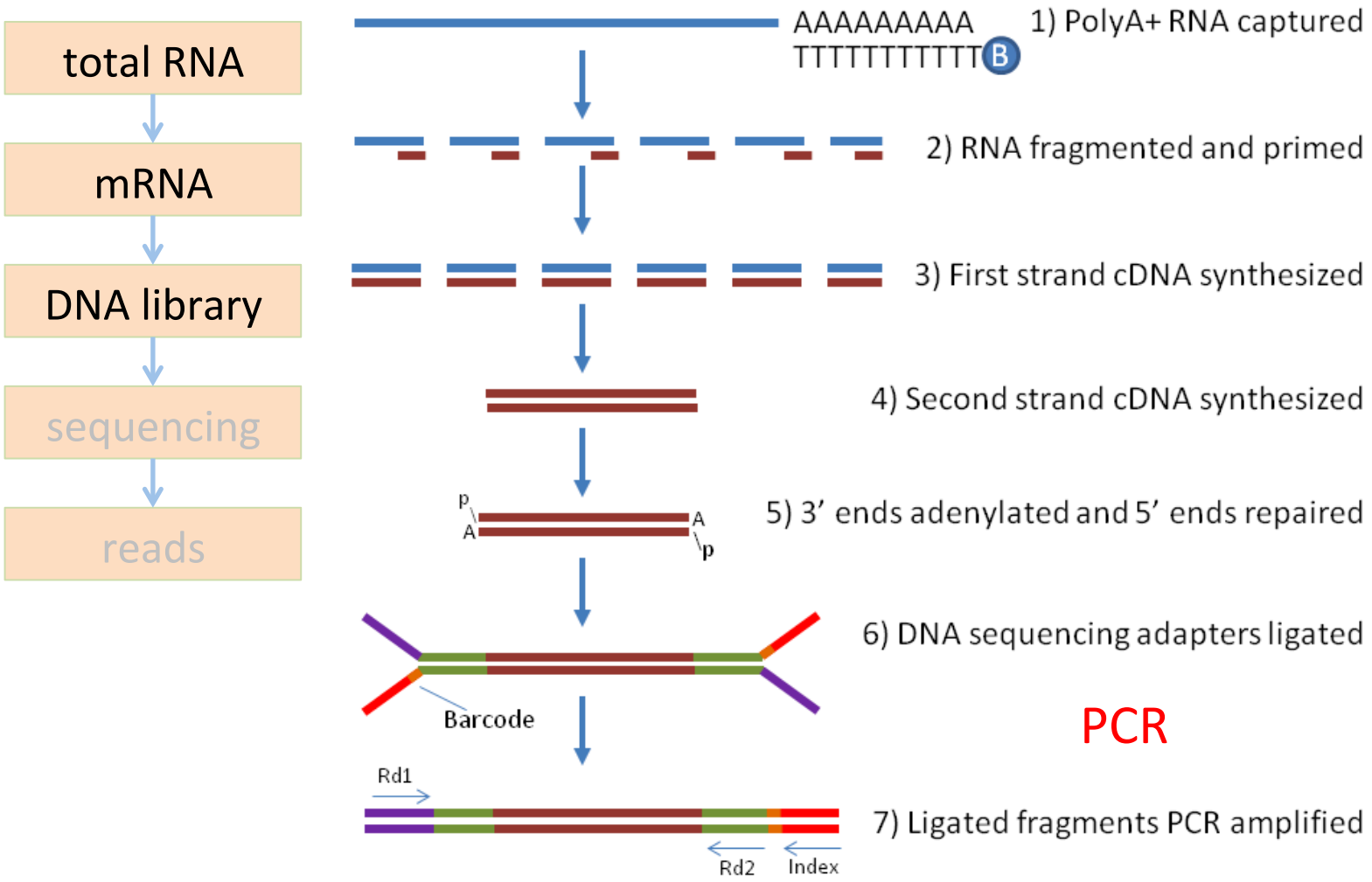


RIN: RNA Integrity Number



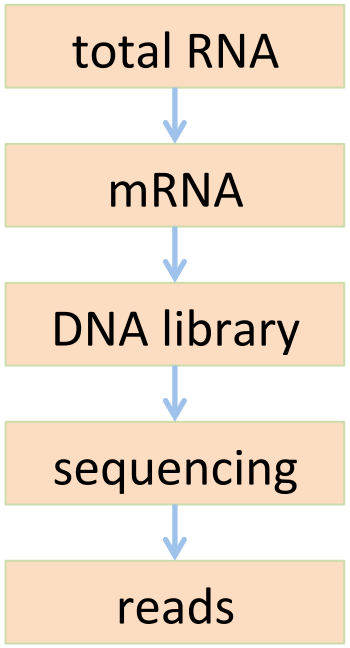
http://www.agilent.com/about/newsroom/lsca/background/rna_integrity.pdf

Illumina RNA-Seq library preparation



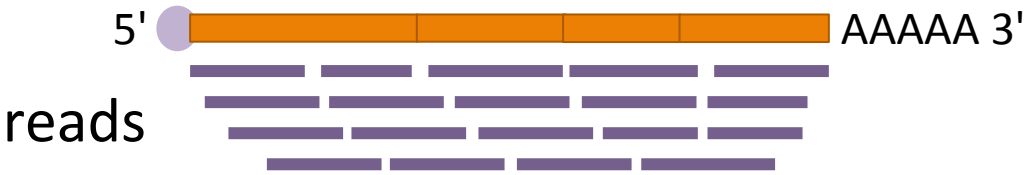
labome.com/method/RNA-seq-Using-Next-Generation-Sequencing.html

Illumina sequencing



Single-end reads

Paired-end reads

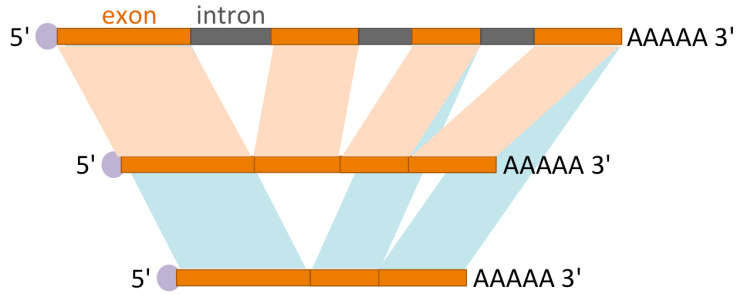


How to reconstruct full-length transcripts from short reads?

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

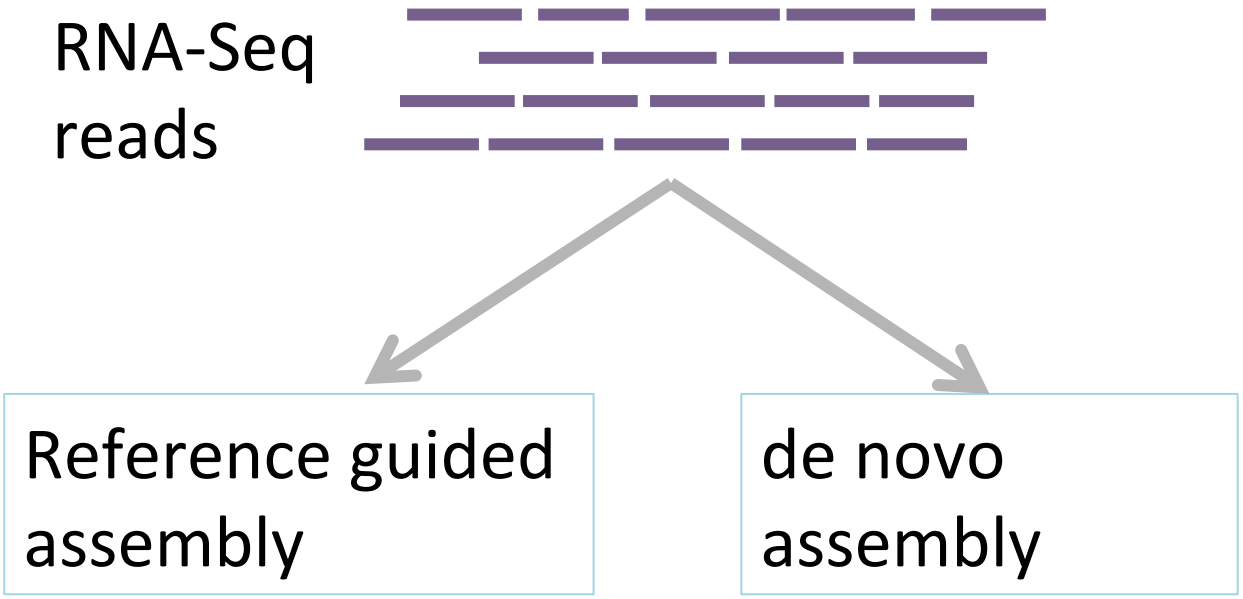
1. Sequencing errors
2. Repeats in different genes

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



Nature Biotechnology 29, 644–652 (2011)

Two main strategies for transcriptome assembly



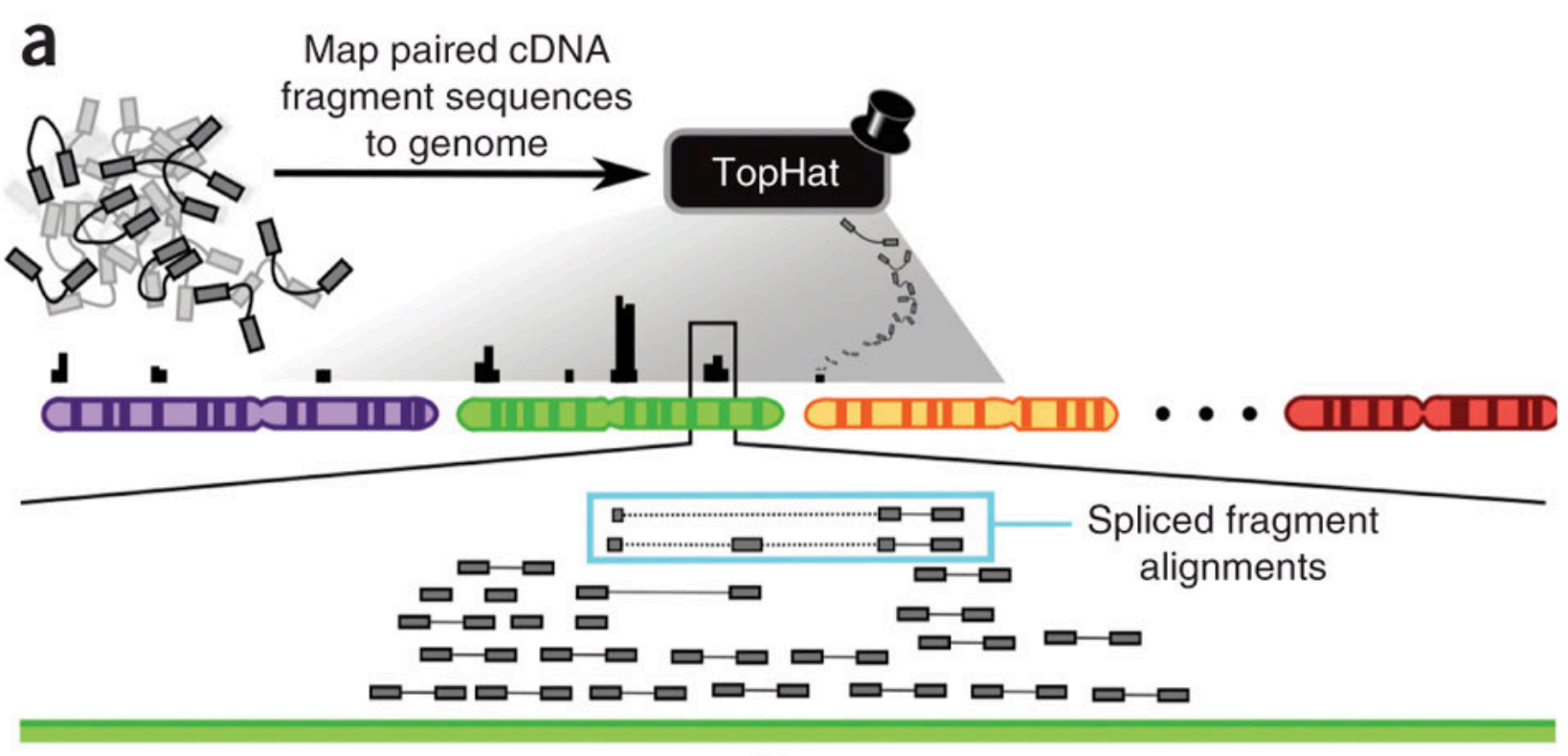
Reference-guided transcriptome assembly

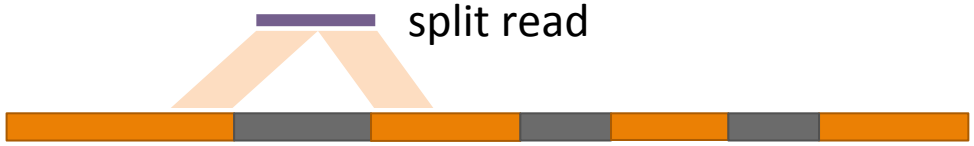
RNA-Seq reads

Reference guided assembly (Cufflinks)
1. alignment to the reference genome
2. assembly based on alignments

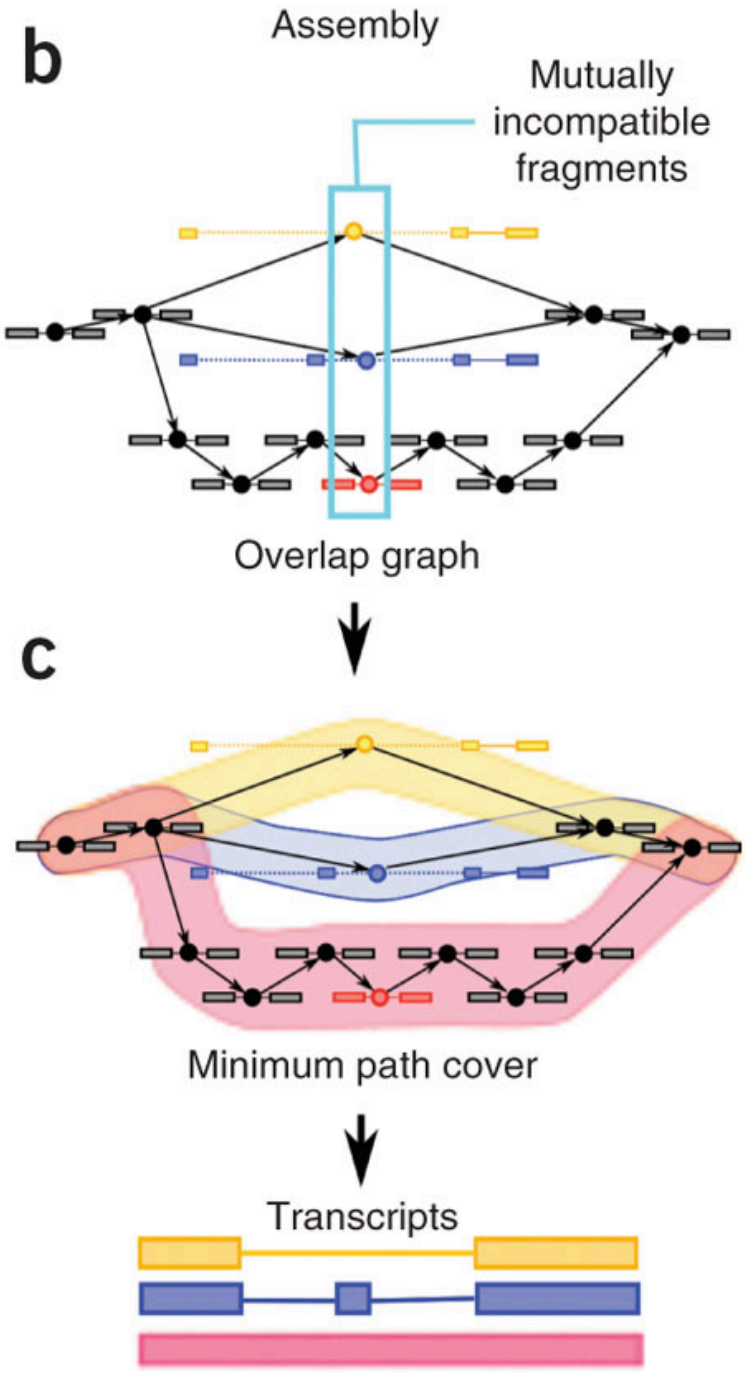
Question: Using a genome as the reference, what difference between RNA-Seq alignments and DNA-Seq alignments?

Cufflinks - spliced alignments





Nature Biotechnology 28, 511–515 (2010)



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

Nature Biotechnology 28, 511–515 (2010)

Pros and cons of reference-guided assembly

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

The quality of assembly is **depended on the accuracy of 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

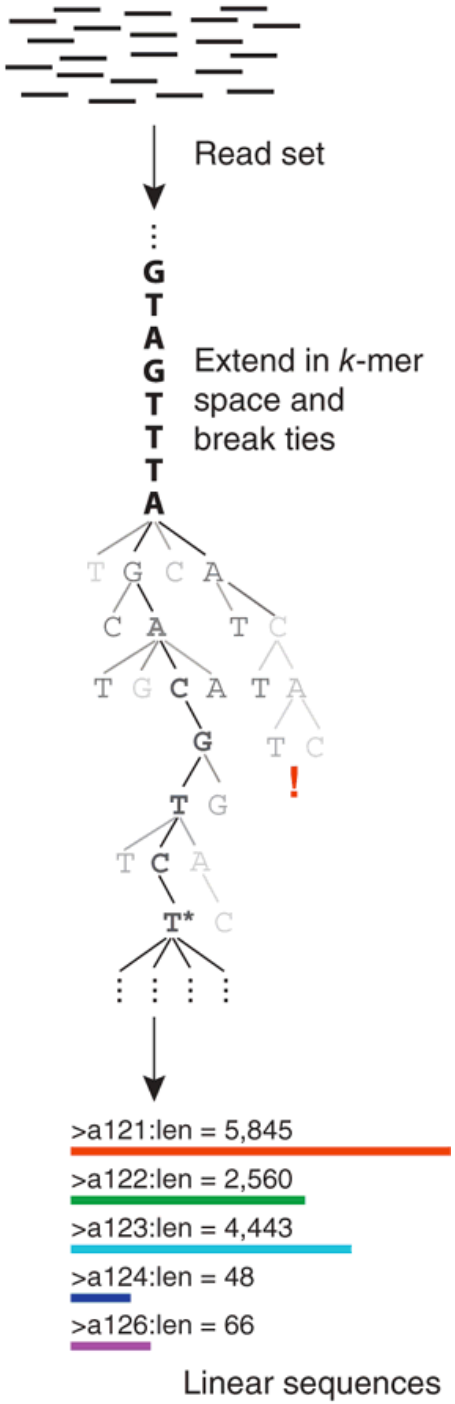
step 1: *Inchworm* assembles reads into unique sequences of transcripts (**reads to contigs**)

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**)

Nature Biotechnology 29, 644–652 (2011)

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

k-mer

- **exact substrings** of length *k* (k-mer) extracted from input reads.

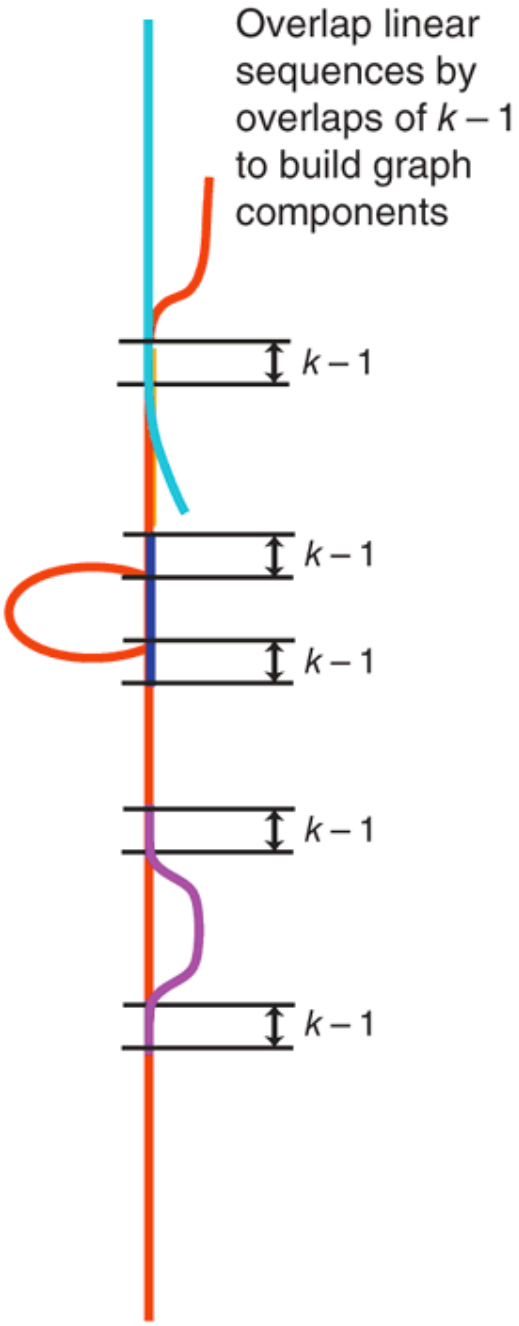
(k=3)

reads A T G G C G T

k-mer (k=3)

1.ATG
2.TGG
3.GGC
4.GCG
5.CGT

Compeau et al., 2011. Nature Biotechnology 29: 987-991.



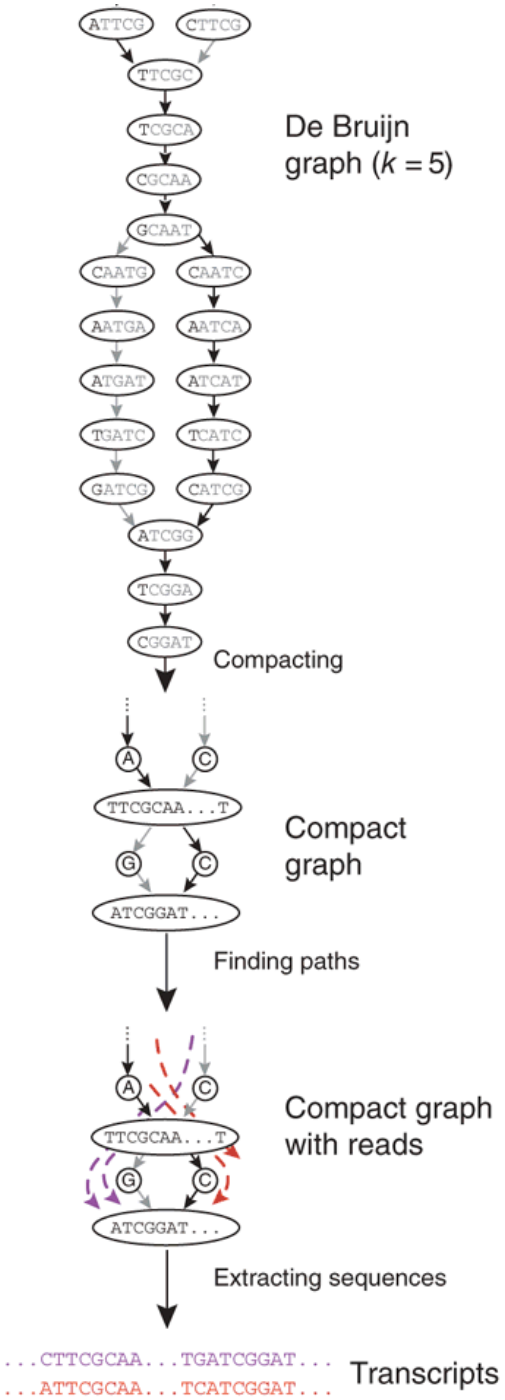
Trinity - Chrysalis

step 2: **Chrysalis clusters related contigs**

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.

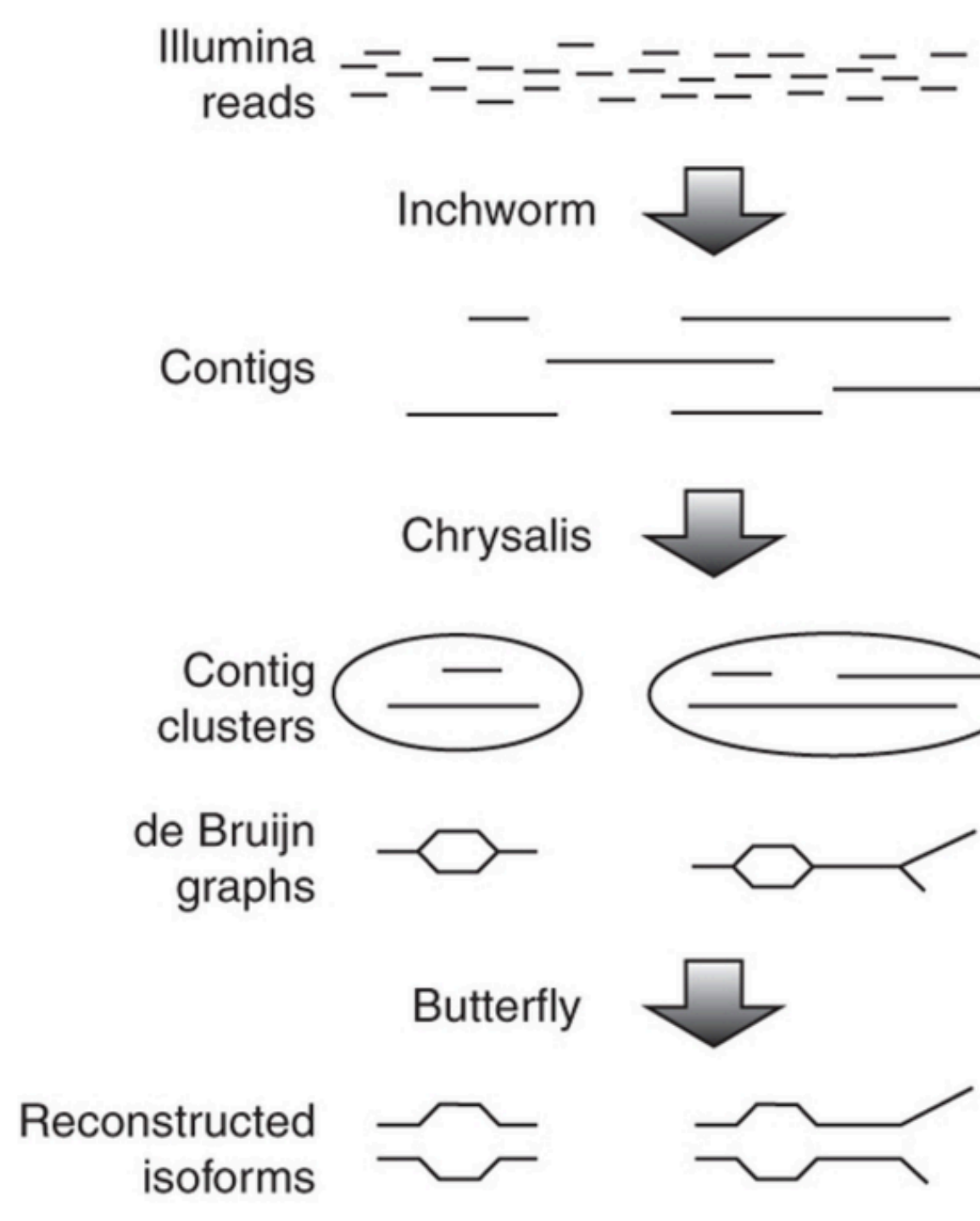
Trinity - Butterfly



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.

Trinity summary

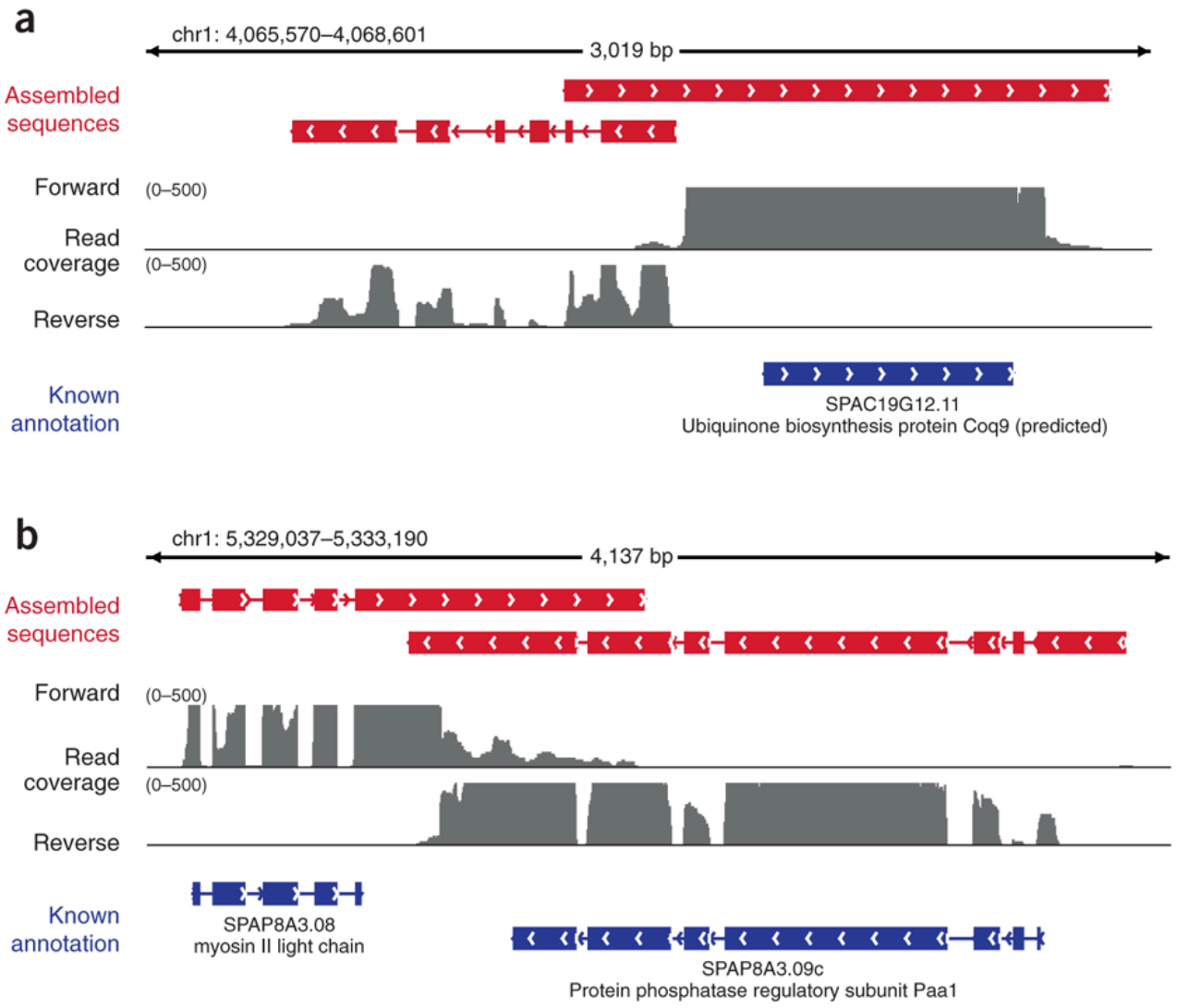


reads to contigs

contigs to cluster/graph

graph to transcripts

De novo assembly could correct wrong annotations



Note

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

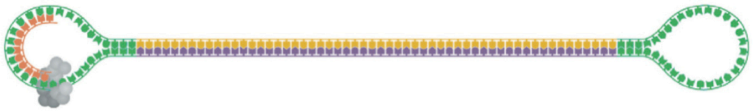
https://github.com/trinityrnaseq/trinityrnaseq/wiki/Genome-Guided-Trinity-Transcriptome-Assembly

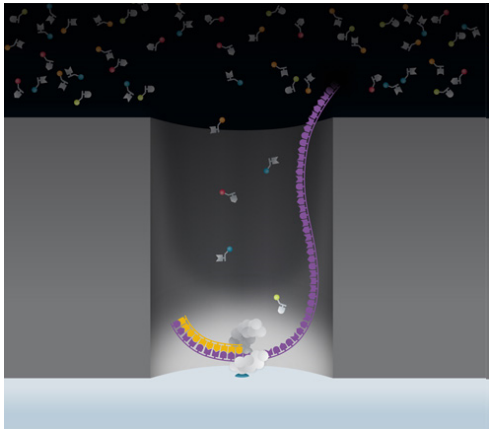
Outline

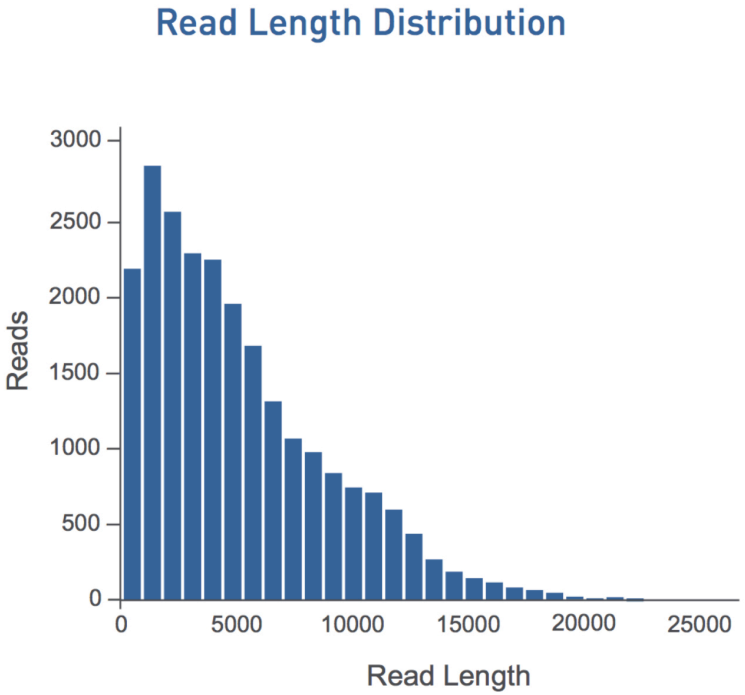
- Introduction of RNA-Seq
- RNA-Seq procedure
- Reference guided assembly
- RNA-Seq *de novo* assembly
- PacBio Iso-seq

PacBio long reads for RNA sequencing (Iso-Seq)

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

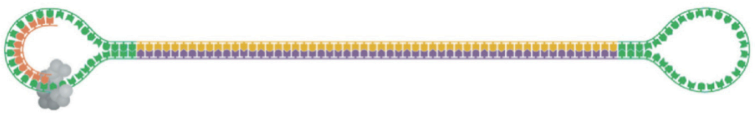


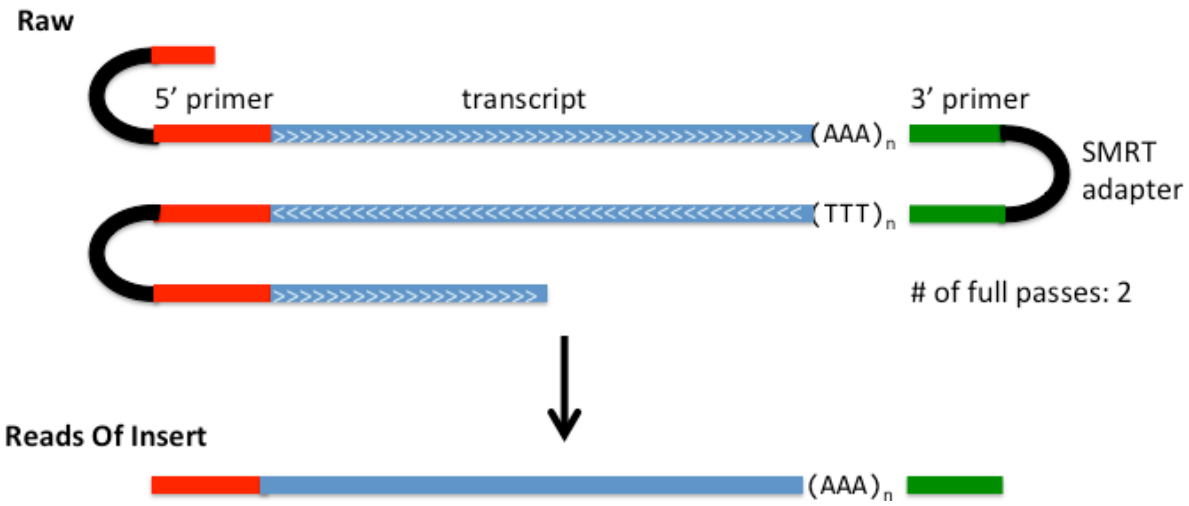




http://flxlexblog.wordpress.com/2013/02/11/applications-for-pacbio-circular-consensus-sequencing/

Multiple passes improve sequence quality

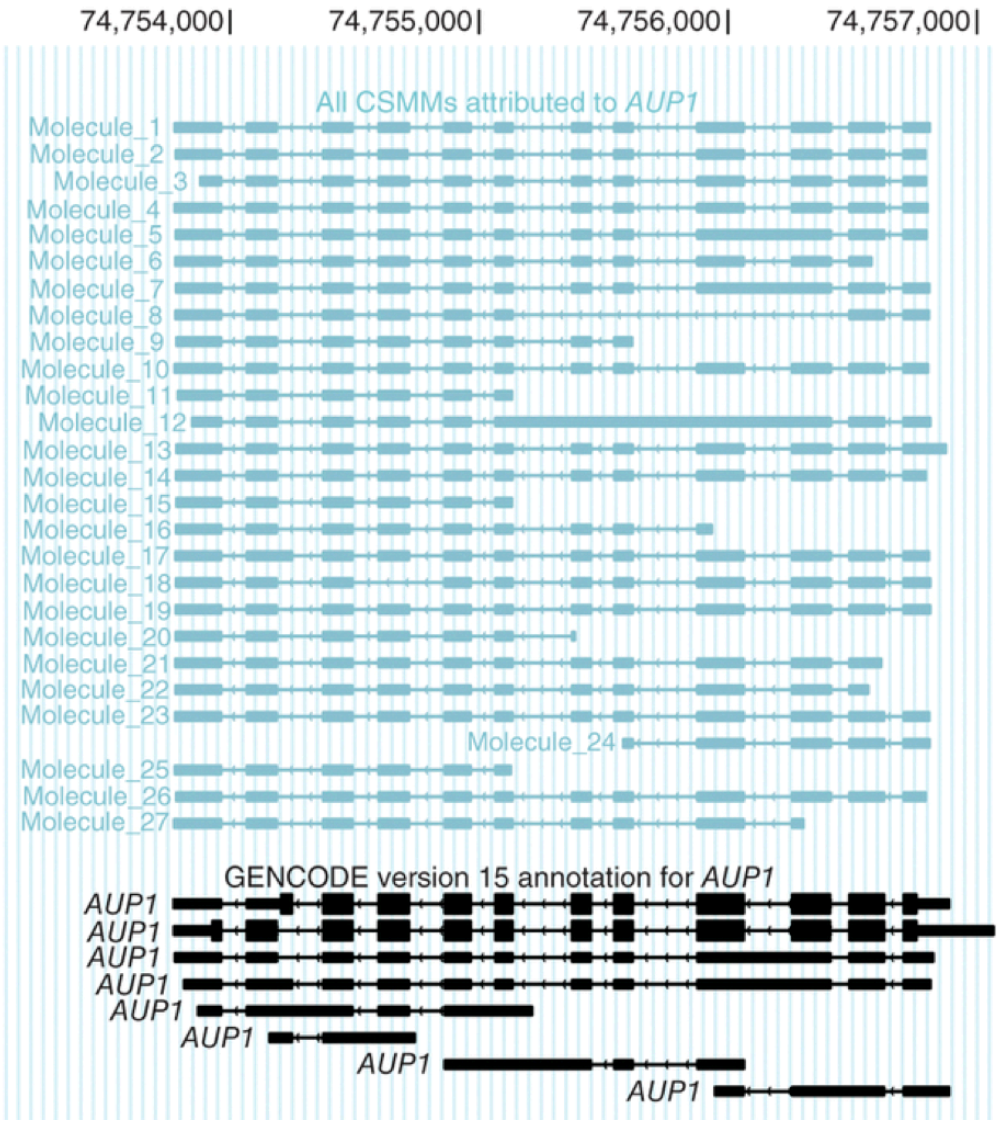




https://github.com/PacificBiosciences/cDNA_primer/wiki/Understanding-PacBio-transcriptome-data

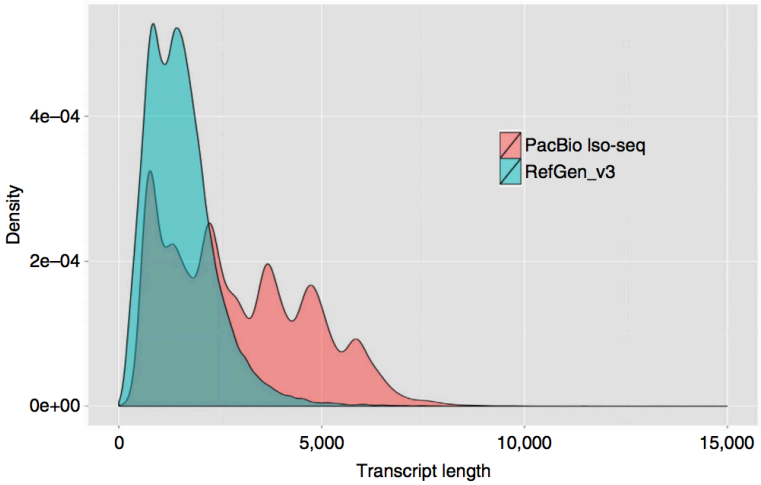
Long reads to sequence full-length cDNA

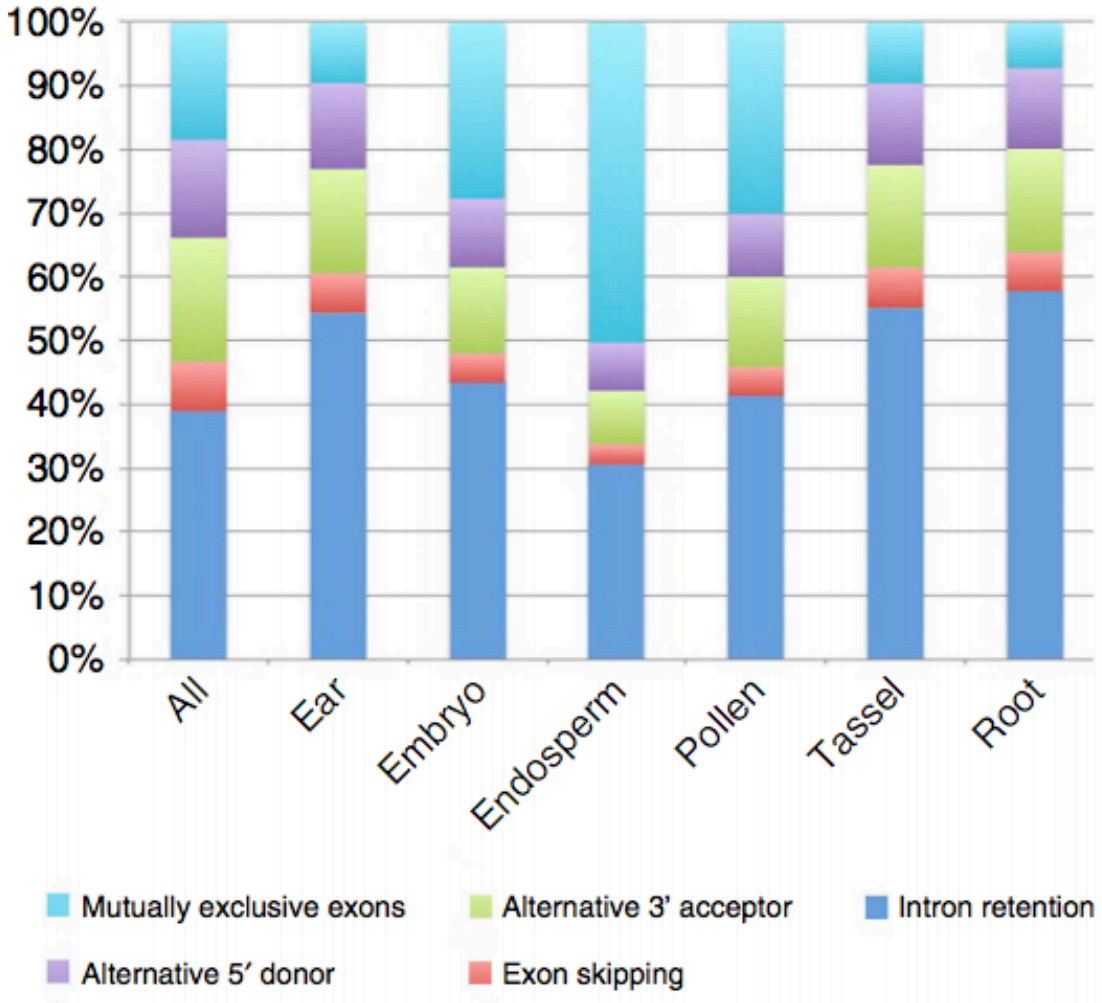
- The majority of reads represent all splice sites of original transcripts
- Isoforms can be monitored at a single-molecule level without amplification or fragmentation

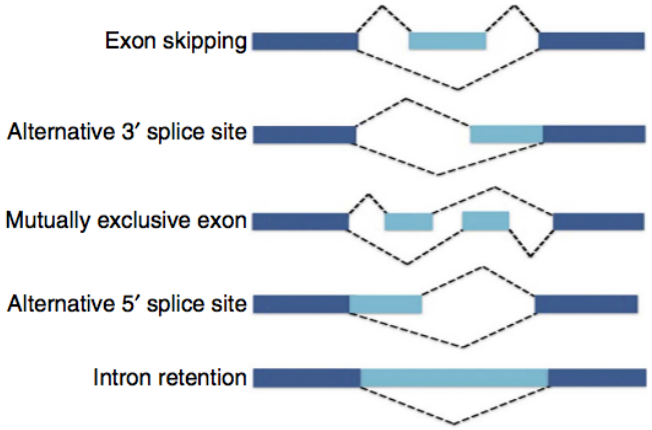


Sharon et al. Nature Biotechnology 31, 1009–1014 (2013)

maize – Iso-Seq

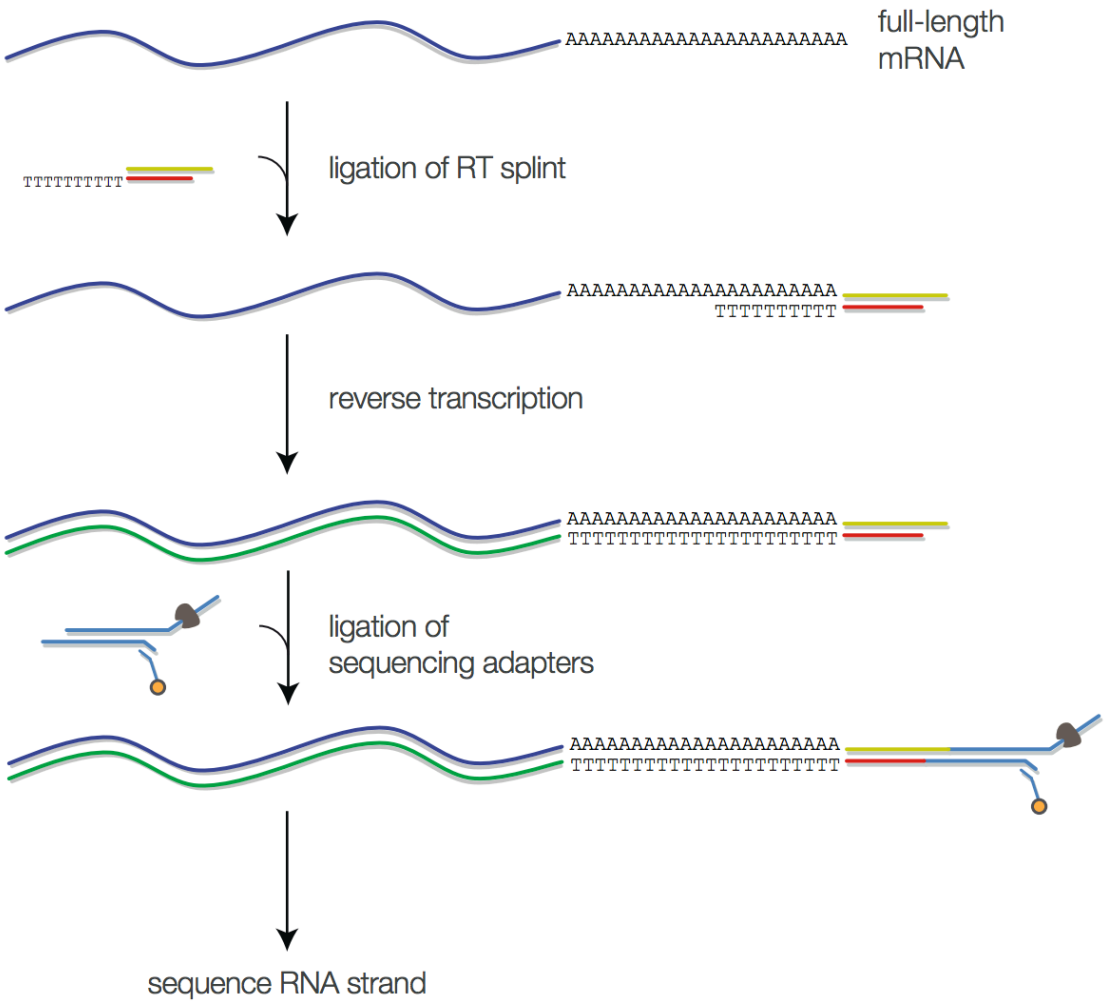






Wang et al., Nature Communications, 2016, 7:11708

Nanopore – direct RNA sequencing



http://biorxiv.org/content/early/2016/08/12/068809

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.