SciLifeLab

Karolinska Institutet · KTH Vetenskap och Konst · Stockholms universitet · Uppsala Universitet

Transcriptome and isoform reconstruction with short reads

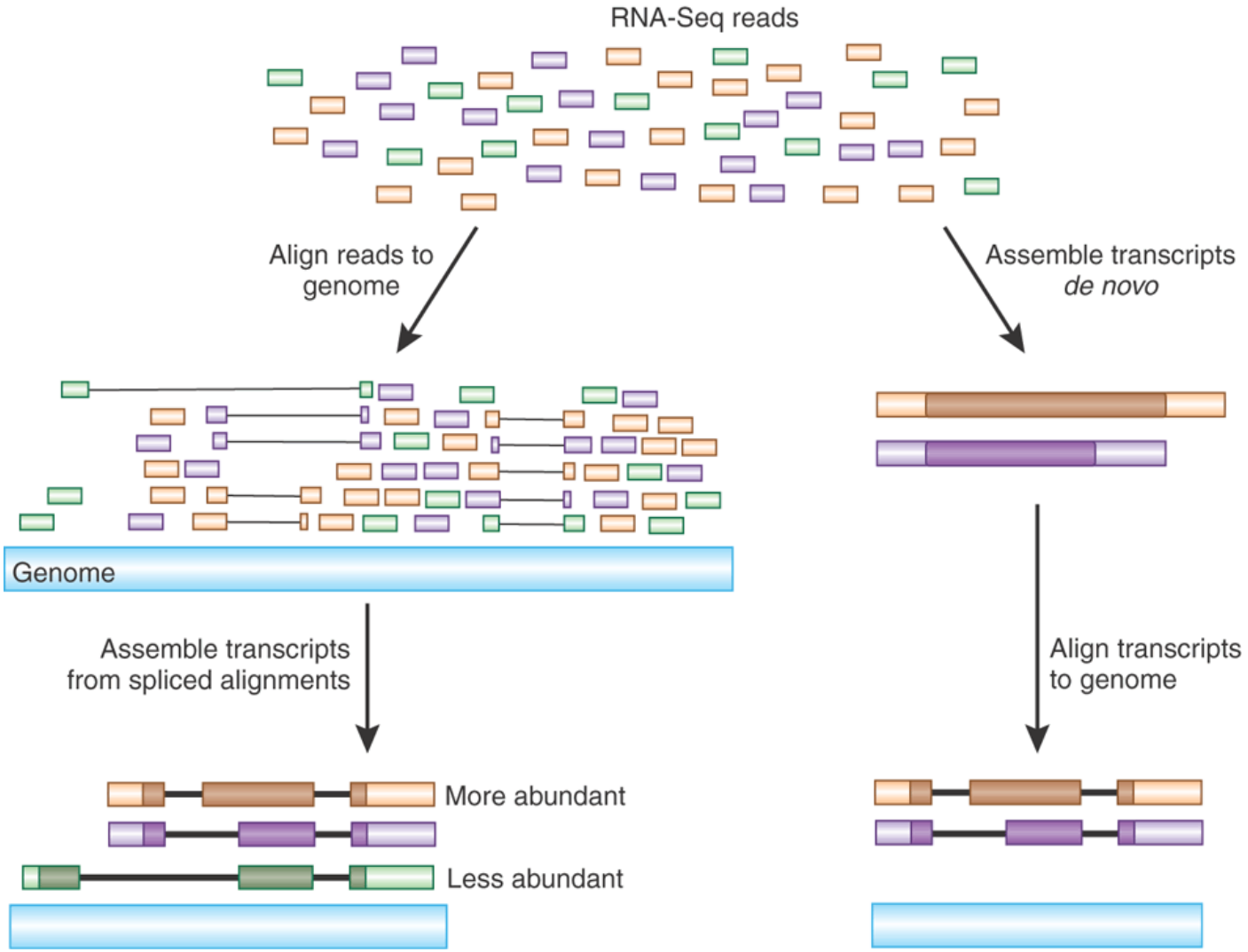
Manfred Grabherr

SciLifeLab RNAseq workshop

November 2017

Transcriptome assembly

SciLifeLab



Haas and Zody, Nature Biotechnology 28, 421–423 (2010)

Reference-based assembly

SciLifeLab

Case study:
The transcriptome of the domestic dog

An improved canine genome and a comprehensive catalogue of coding genes and non-coding transcripts. Hoeppner MP et al. PLoS One 2014 Mar 13;9(3):e91172

Reference-based assembly

SciLifeLab

- Why dogs?
  - Shared environment with humans for > 10.000 years
  - Affected by cancer or heart disease
  - Breed-specific disease
- New genome release in 2011 (canFam3.1)
  - 85 Mb of additional sequences integrated
  - 99.8% of euchromatic portion of genome covered, high quality
- Annotation: not so good
  - Mostly homology-based
  - Almost no isoform information

Reference-based assembly

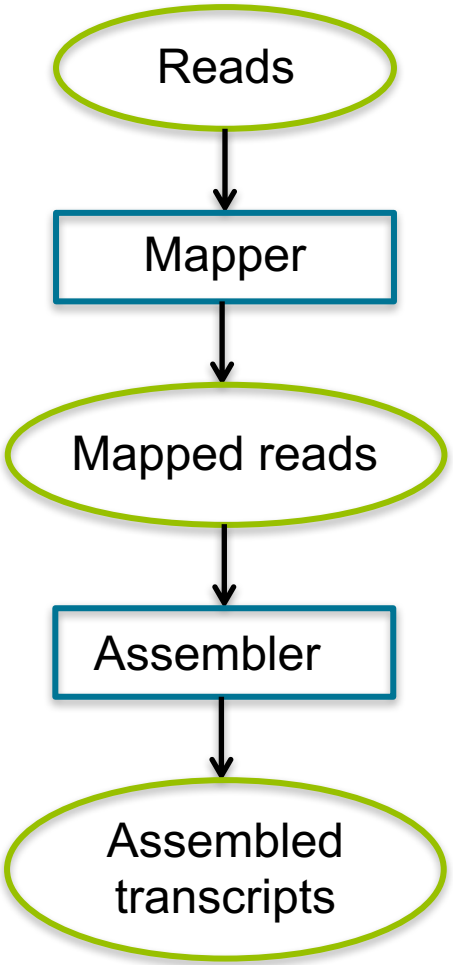
SciLifeLab

- 10 tissues at great depth (30-100 million paired-end reads)
  - blood, brain, heart, kidney, liver, lung, ovary, skeletal muscle, skin, and testis
- 2 sets of libraries
  - strand-specific dUTP with poly-A selection: captures protein coding genes and other transcripts transcribed by polymerase II
  - duplex-specific nuclease (DSN): targets all RNAs, reduces the levels of the highly abundant ribosomal transcripts through normalization

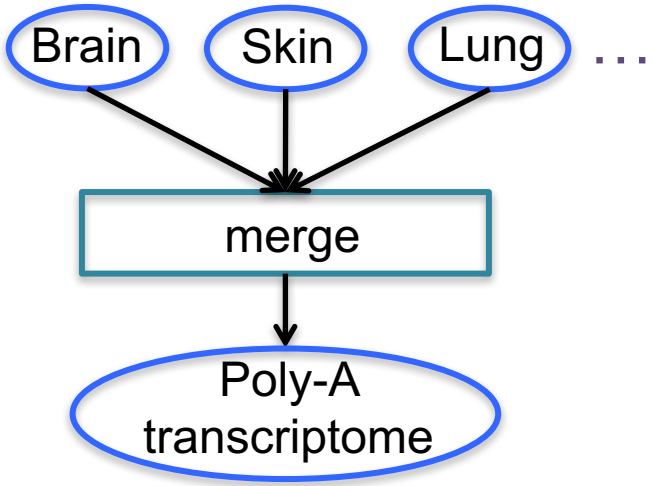
Reference-based assembly

SciLifeLab

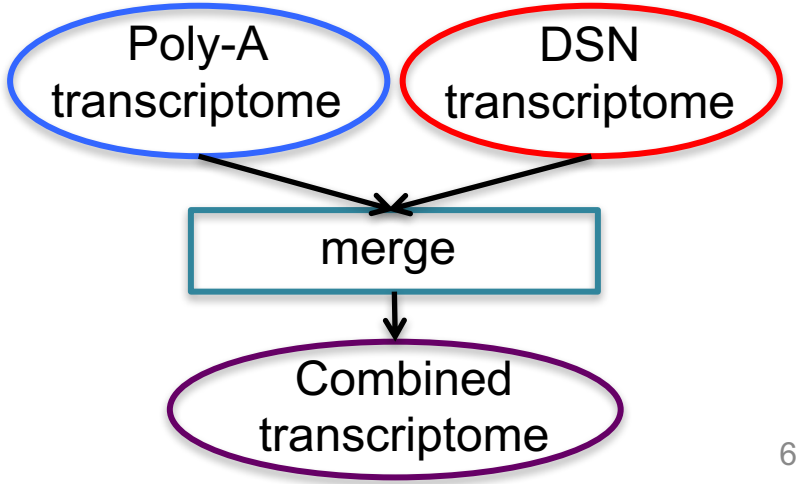
For each sample (tissue/library)



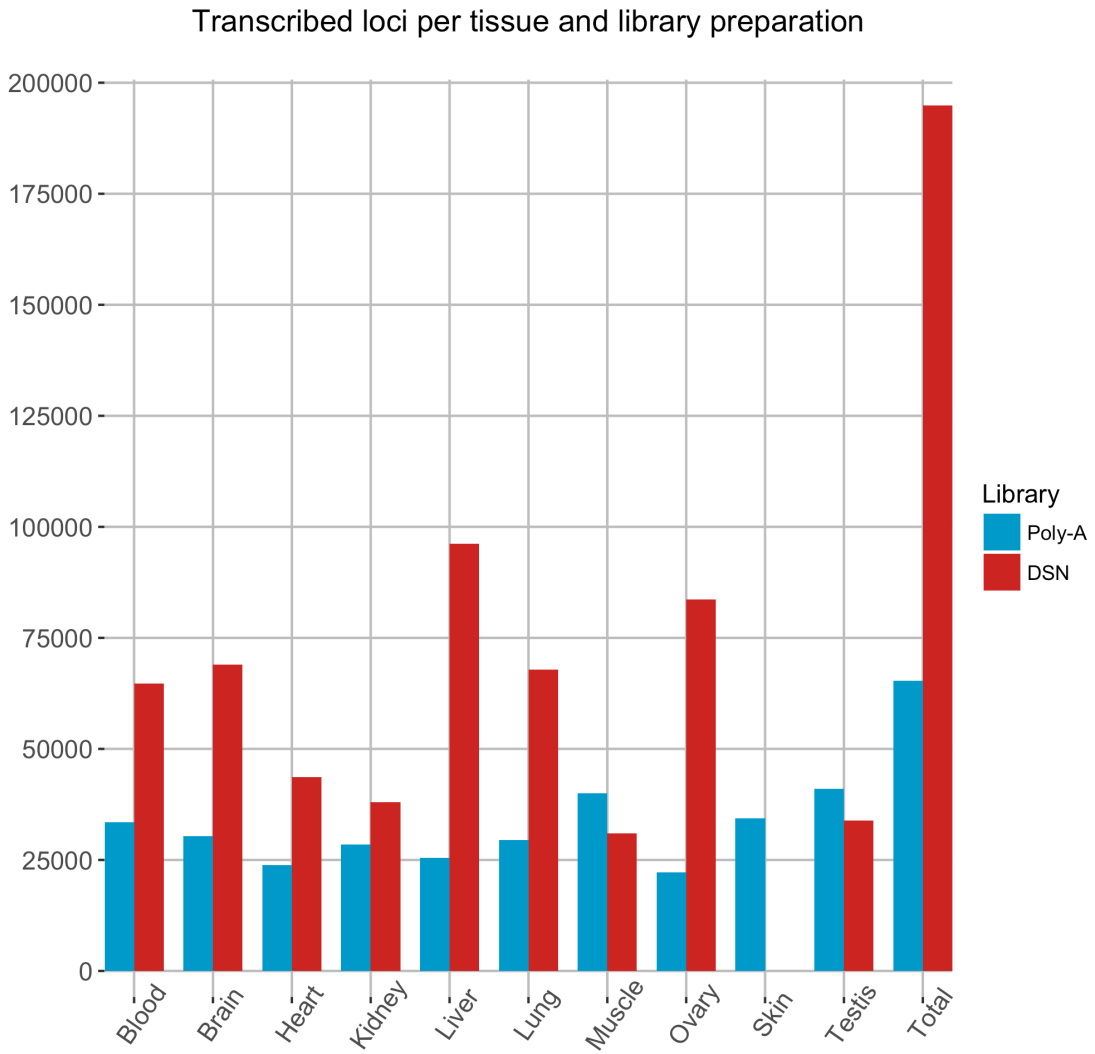
For each library (Poly-A, DSN)



Final transcriptome

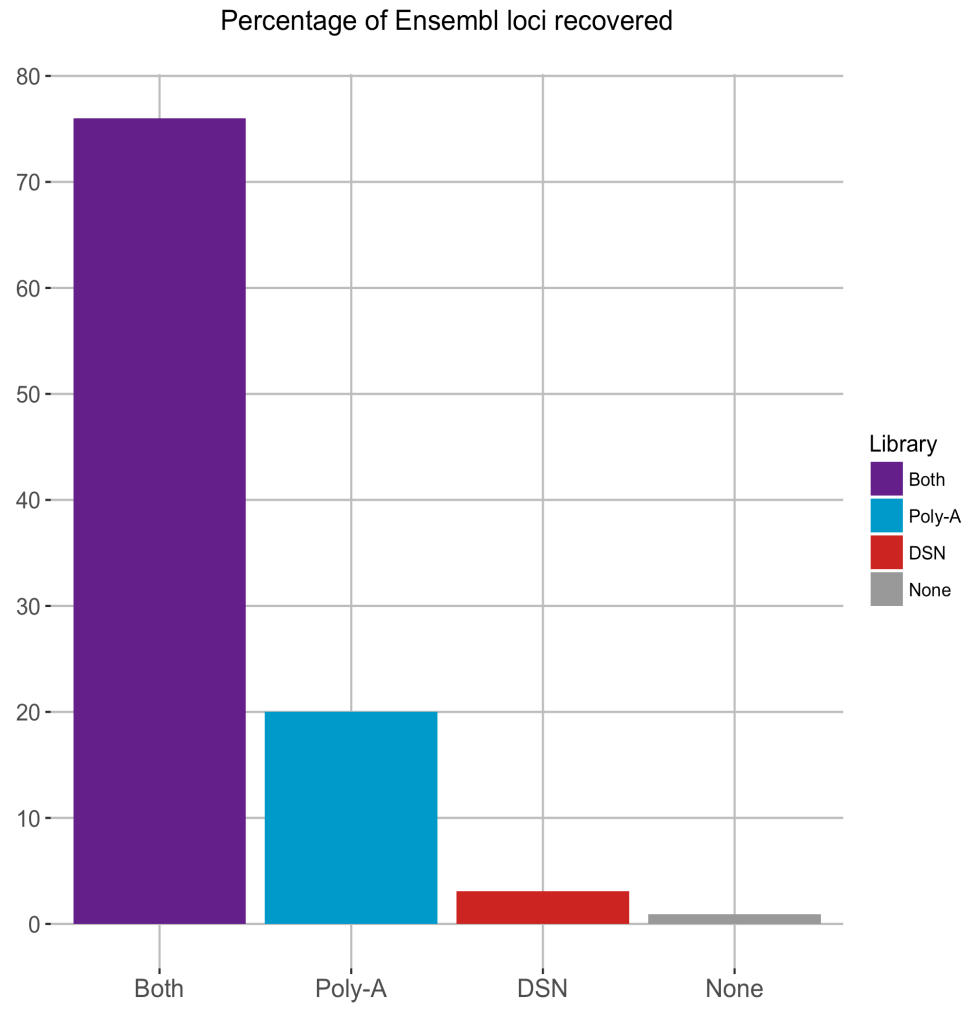


Reference-based assembly



- DSN recovers more transcripts
- Poly-A: highest number in testis, then muscle
- Poly-A: heart and muscle share 88% of loci
- Mean transcript length:
  - Poly-A: 3169 bp
  - DSN: 1485 bp

Reference-based assembly



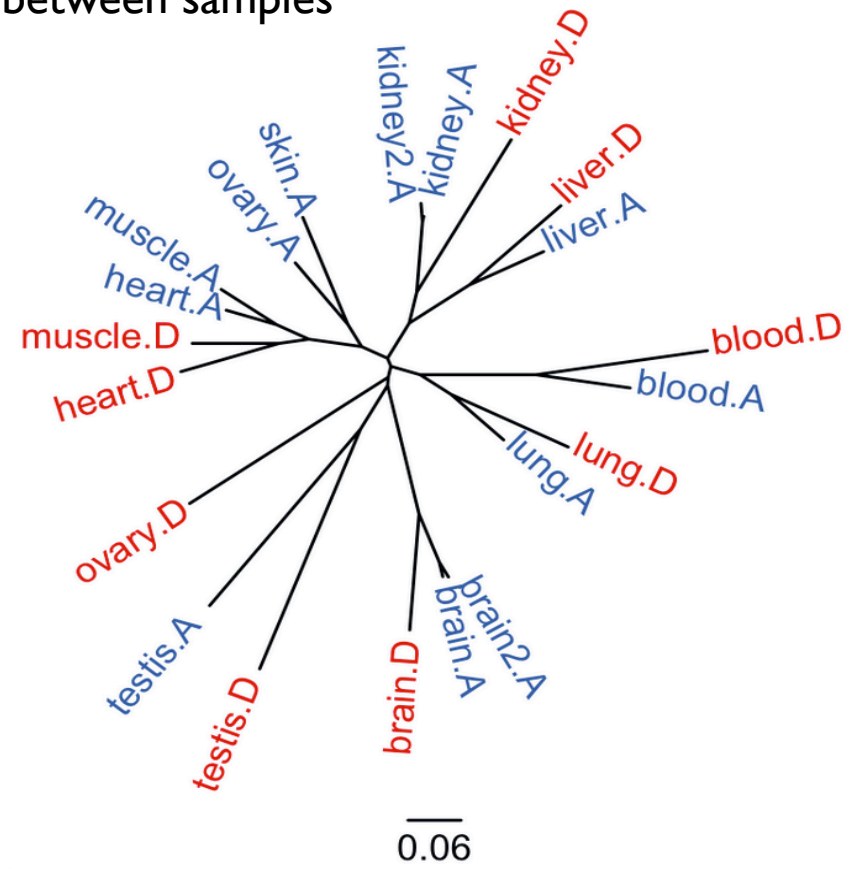
- Ensembl build 64: 19,856 annotated loci
- Combined Poly-A + DSN: 174,336 loci
- Majority located in introns of known genes and transcribed in the same sense
  - potential byproducts of incomplete splicing
- Many located outside of known features, seem independently transcribed

Reference-based assembly

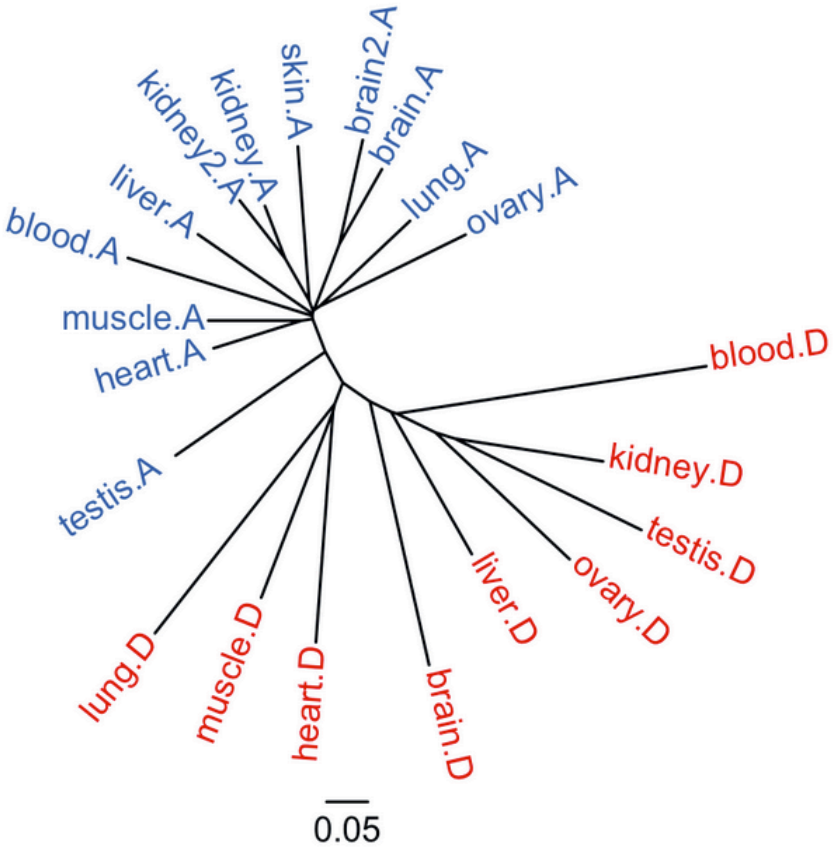
SciLifeLab

Distance trees of expression profile

Neighbor-joining trees based on the correlation between expression values (FPKM>1.0) between samples



Protein-coding genes with RNA-Seq support



Intergenic and uncharacterized single exons

An improved canine genome and a comprehensive catalogue of coding genes and non-coding transcripts. Hoeppner MP et al. PLoS One 2014 Mar 13;9(3):e91172

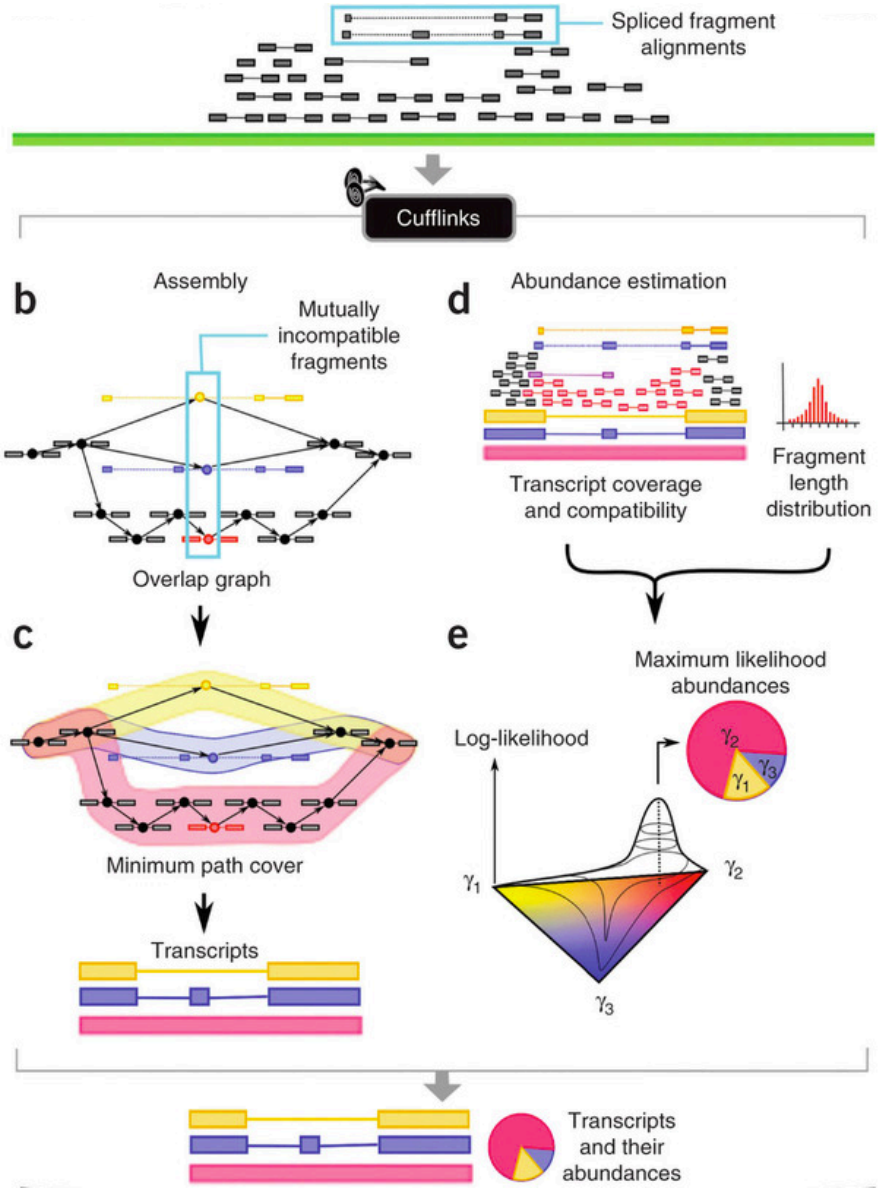
Reference-based assembly

SciLifeLab

Cufflinks

From the "Tuxedo" protocol

Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Trapnell C. et al. Nature Biotechnology 28, 511–515 (2010)

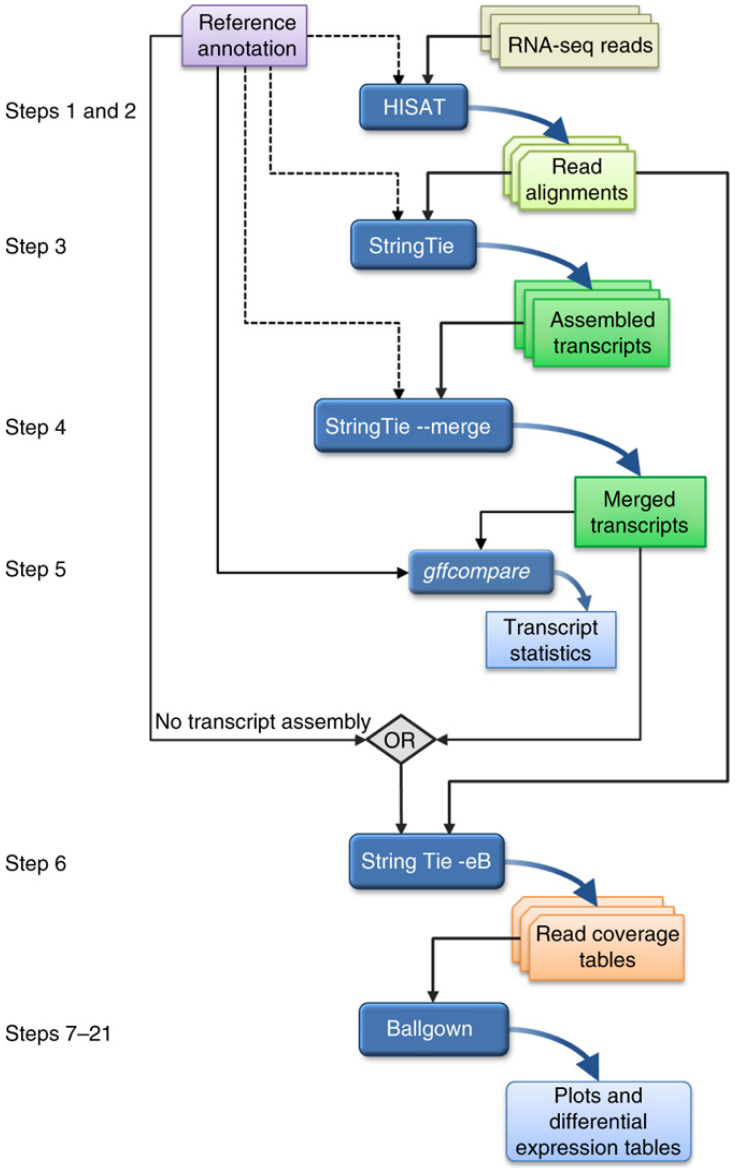


Reference-based assembly

SciLifeLab

The "new Tuxedo" protocol

Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. Pertea M. et al. Nature protocol 11, 1650–1667 (2016)



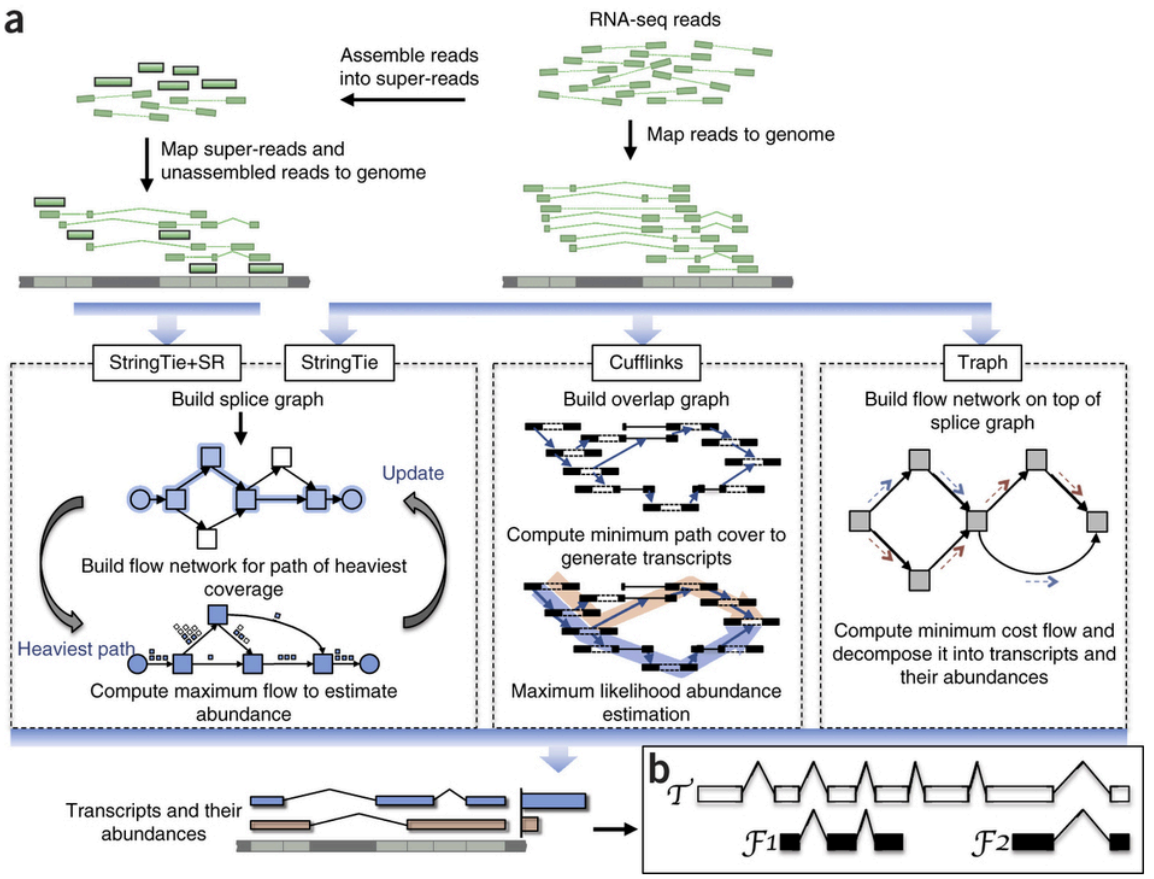
Reference-based assembly

SciLifeLab

StringTie

From the "new Tuxedo" protocol

StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Pertea M.. et al. Nature Biotechnology 33, 290–295 (2015)

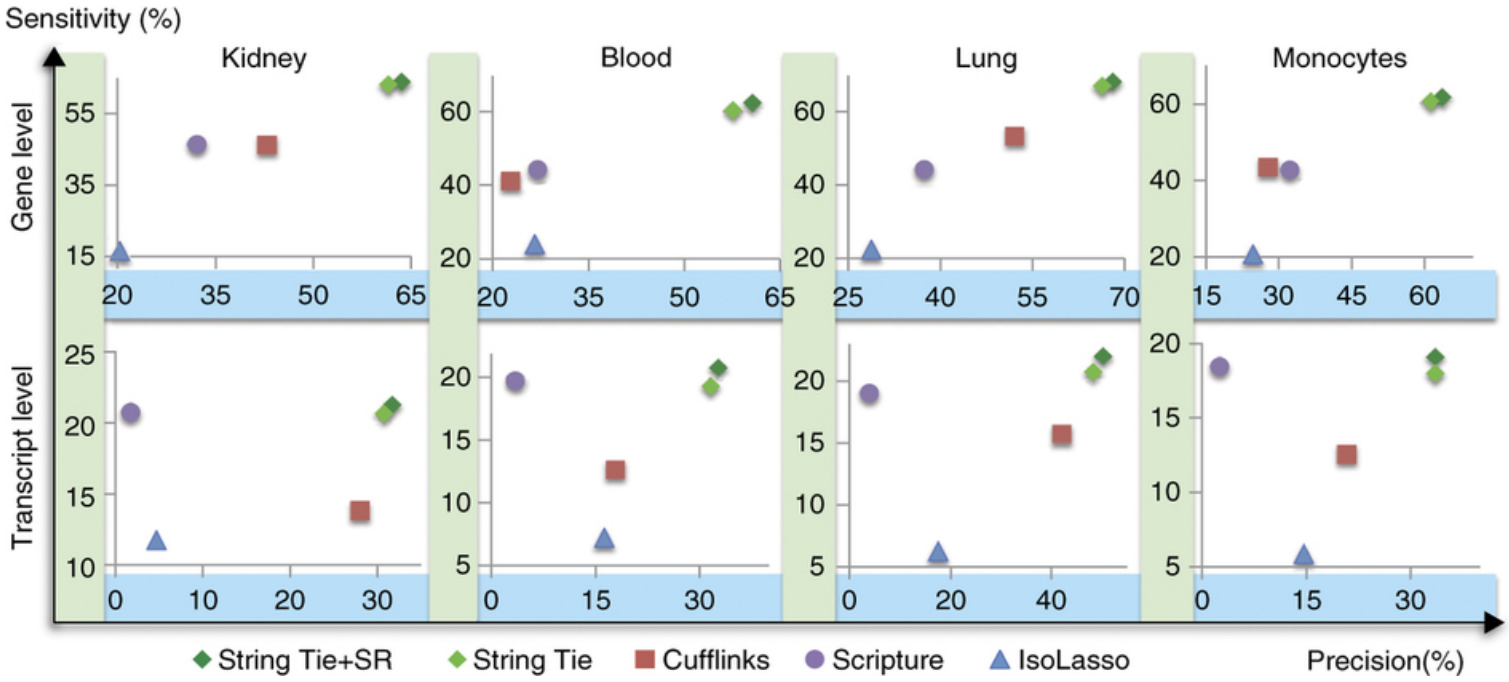


Reference-based assembly

SciLifeLab

StringTie

Fig.3: Accuracy of transcript assemblers at assembling known genes, measured on real data sets from four different tissues (RefSeq, UCSC or Ensembl human gene databases)



Sensitivity (genes): % of genes for which a program got at least one isoform correct
Sensitivity (transcripts): % of known transcripts that were correctly assembled
Precision: % of all predicted genes/transcripts that match an annotated gene/transcript

Reference-based assembly

SciLifeLab

Take-home message

- Need a very good reference (genome most of the time)
- Can use existing annotation (GTF/GFF file)
- Can detect novel transcripts

De novo assembly

SciLifeLab

- Most used programs (latest release date):
  - Trinity (March 2016)
  - SOAPdenovo-Trans (July 2013)
  - Trans-ABySS (August 2016)
  - Velvet+Oases (March 2015)
- Originally SOAPdenovo, ABySS and Velvet for de novo genome assembly
- "SOAPdenovo-Trans incorporates the error-removal model from Trinity and the robust heuristic graph traversal method from Oases."
- All based on de Bruijn graph

Trinity, Grabherr et al. 2011

De novo assembly

SciLifeLab

De novo transcriptome assembly databases for the butterfly orchid *Phalaenopsis equestris*

Data in Brief

De novo transcriptome assembly of mangosteen (*Garcinia mangostana* L.) fruit

De novo Transcriptome Analysis Reveals Distinct Defense Mechanisms by Young and Mature Leaves of *Hevea brasiliensis* (Para Rubber Tree)

De novo transcriptome assembly and analysis of differentially expressed genes of two barley genotypes reveal root-zone-specific responses to salt exposure

***De Novo* Sequencing and Analysis of Lemongrass Transcriptome Provide First Insights into the Essential Oil Biosynthesis of Aromatic Grasses**

De novo transcriptome assembly of two contrasting pumpkin cultivars

Identification of novel and useful EST-SSR markers from *de novo* transcriptome sequence of wheat (*Triticum aestivum* L.)

De Novo Transcriptome Assembly and Characterization for the Widespread and Stress-Tolerant Conifer *Platycladus orientalis*

De novo* Assembly of Leaf Transcriptome in the Medicinal Plant *Andrographis paniculata

Transcriptome sequencing and de novo characterization of Korean endemic land snail, *Koreanohadra kurodana* for functional transcripts and SSR markers

***De Novo* Assembly of the Transcriptome of *Turritopsis*, a Jellyfish that Repeatedly Rejuvenates**

Transcriptome of the Caribbean stony coral *Porites astreoides* from three developmental stages

De novo transcriptome assembly of the marine gastropod *Reishia clavigera* for supporting toxic mechanism studies

The *De Novo* Transcriptome and Its Functional Annotation in the Seed Beetle *Callosobruchus maculatus*

De Novo Transcriptome Analysis of the Common New Zealand Stick Insect *Clitarchus hookeri* (Phasmatodea) Reveals Genes Involved in Olfaction, Digestion and Sexual Reproduction

Characterization and analysis of a *de novo* transcriptome from the pygmy grasshopper *Tetrix japonica*

Optimizing Hybrid *de Novo* Transcriptome Assembly and Extending Genomic Resources for Giant Freshwater Prawns (*Macrobrachium rosenbergii*): The Identification of Genes and Markers Associated with Reproduction

De Novo Transcriptome Analysis of Two Seahorse Species (*Hippocampus erectus* and *H. mohnikei*) and the Development of Molecular Markers for Population Genetics

De Novo assembly and annotation of the freshwater crayfish *Astacus astacus* transcriptome

De novo assembly

SciLifeLab

The de Bruijn graph

CTTGGAACAATATGAATTGGCAAT
ATTGGCAATTGACTTTTGCCGTAAT
CCGTAATCCGGCATATCTGGATA

Kmers (k = 7)

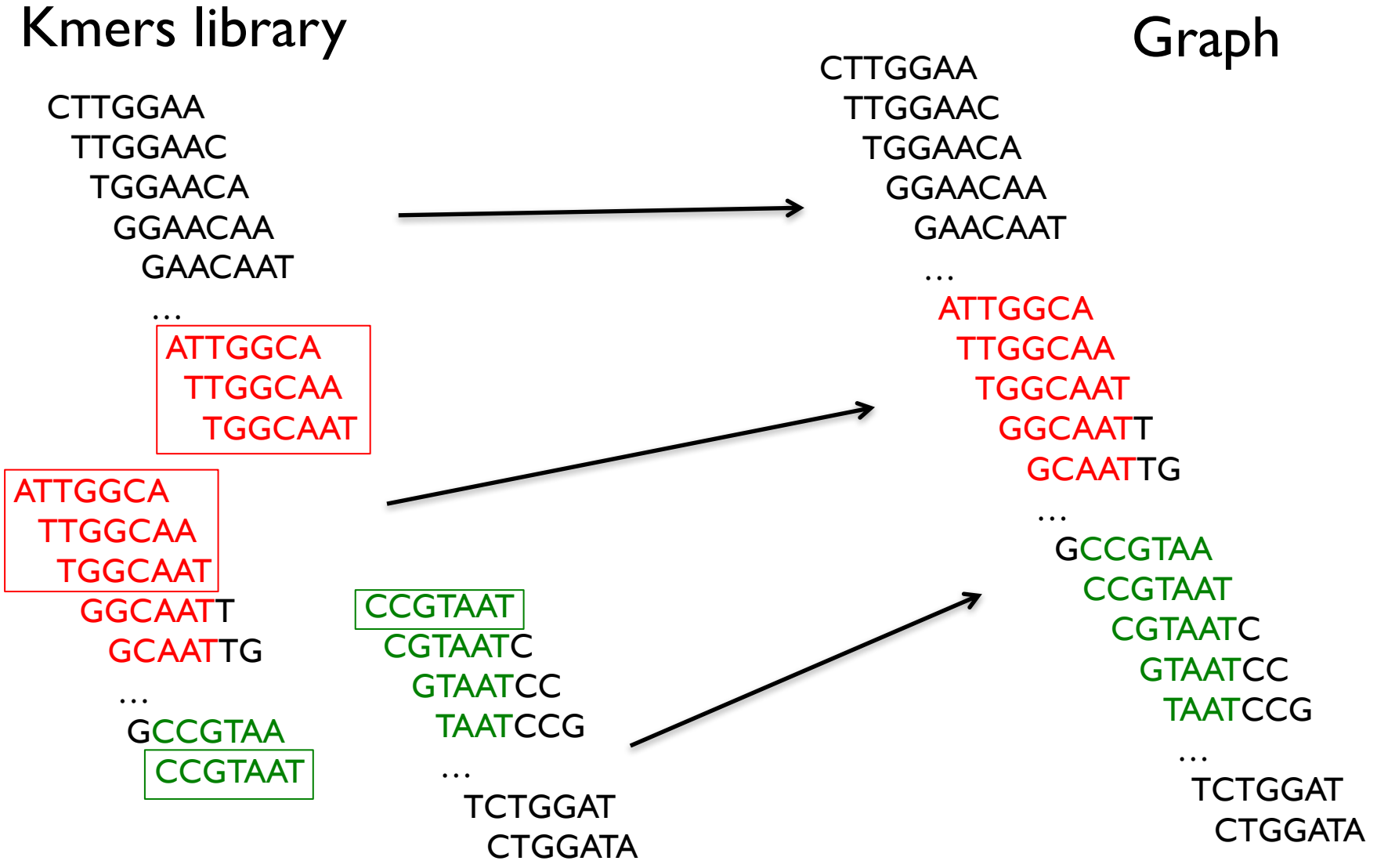
CTTGGAA
TTGGAAC
TGGAACA
GGAACAA
GAACAAT
…
ATTGGCA
TTGGCAA
TGGCAAT

ATTGGCA
TTGGCAA
TGGCAAT
GGCAATT
GCAATTG
…
GCCGTAA
CCGTAAT

CCGTAAT
CGTAATC
GTAATCC
TAATCCG
…
TCTGGAT
CTGGATA

De novo assembly

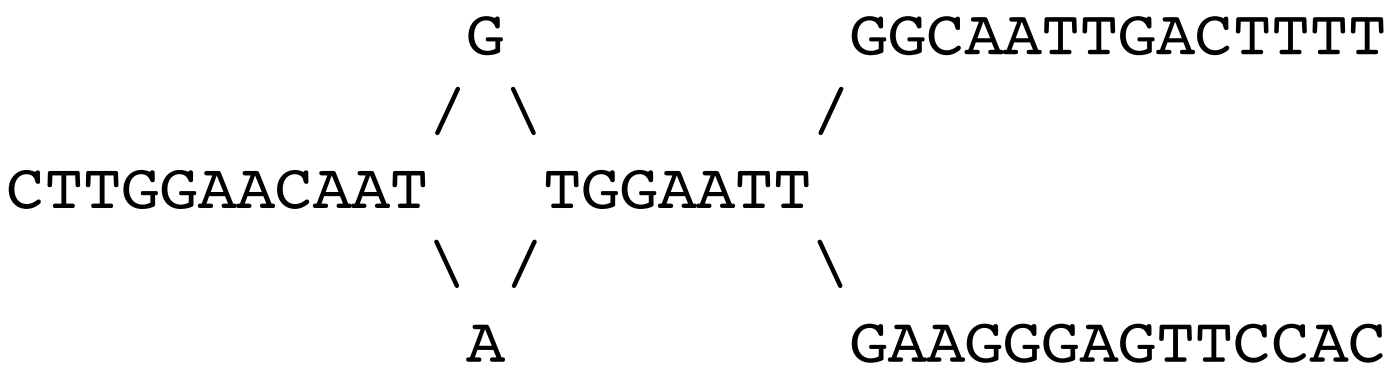
SciLifeLab



De novo assembly

SciLifeLab

Graphs can have nodes and edges

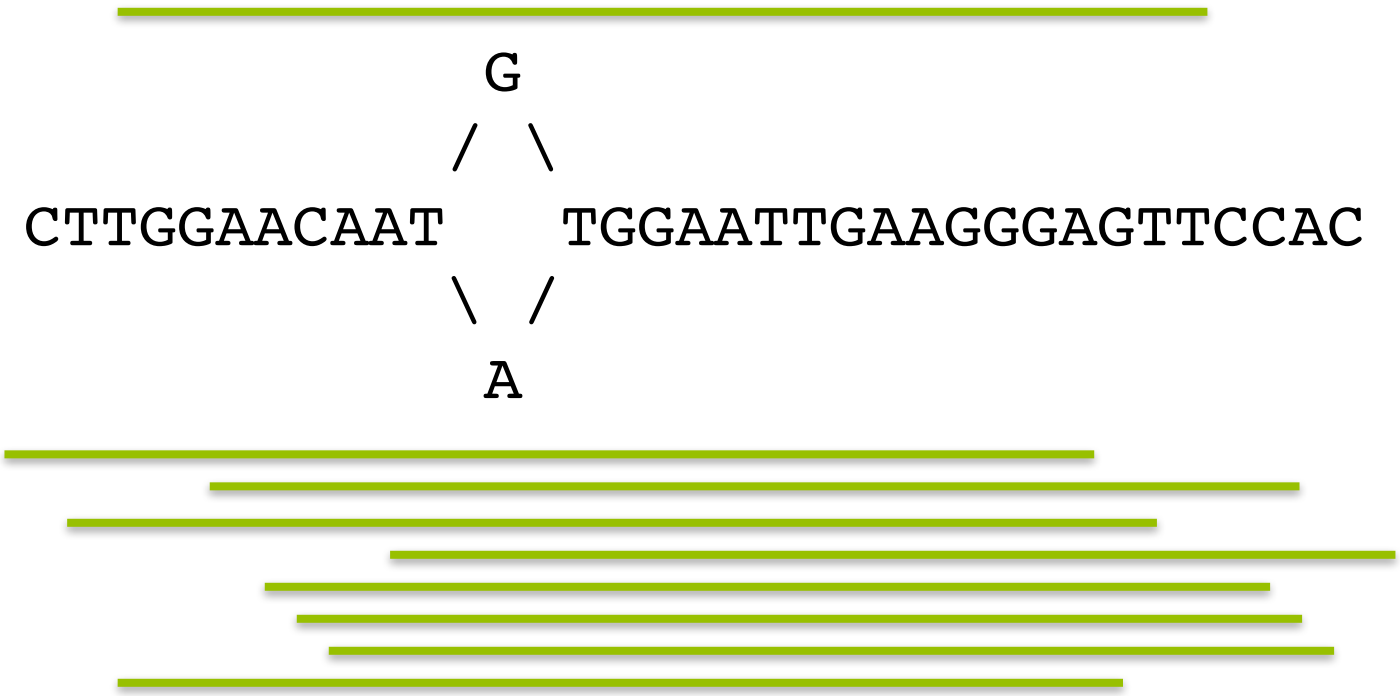


De novo assembly

SciLifeLab

Differences between programs:

- Kmer length
- Removing edges



De novo assembly

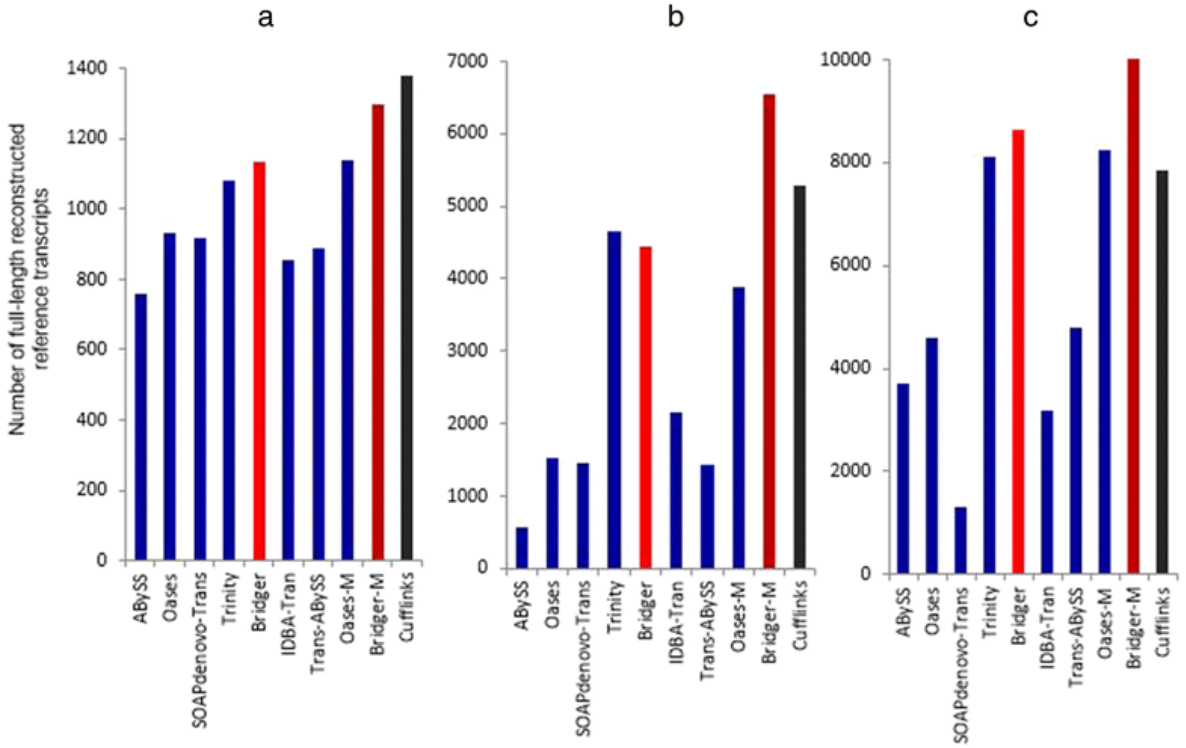
SciLifeLab

METHOD | OPEN ACCESS

Bridger: a new framework for *de novo* transcriptome assembly using RNA-seq data

Zheng Chang†, Guojun Li†, Juntao Liu, Yu Zhang, Cody Ashby, Deli Liu, Carole L Cramer and Xiuzhen Huang
† Contributed equally

Genome Biology 2015 16:30 | DOI: 10.1186/s13059-015-0596-2 | © Chang et al.; licensee BioMed Central. 2015
Received: 22 May 2014 | Accepted: 23 January 2015 | Published: 11 February 2015



Number of full-length reconstructed reference transcripts for (a) dog, (b) human, and (c) mouse

De novo assembly

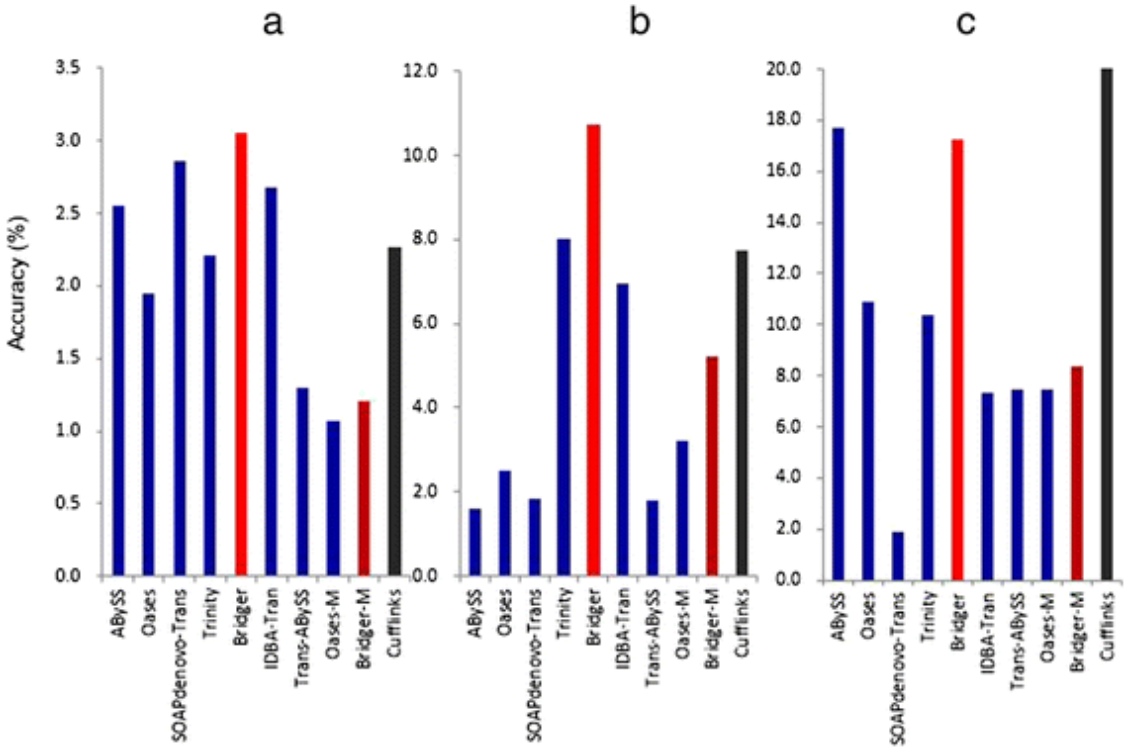
SciLifeLab

METHOD | OPEN ACCESS

Bridger: a new framework for *de novo* transcriptome assembly using RNA-seq data

Zheng Chang†, Guojun Li†, Juntao Liu, Yu Zhang, Cody Ashby, Deli Liu, Carole L Cramer and Xiuzhen Huang
† Contributed equally

Genome Biology 2015 16:30 | DOI: 10.1186/s13059-015-0596-2 | © Chang et al.; licensee BioMed Central. 2015
Received: 22 May 2014 | Accepted: 23 January 2015 | Published: 11 February 2015



Accuracy for (a) dog, (b) human, and (c) mouse [the most reference transcripts by the least candidate transcripts]

De novo assembly

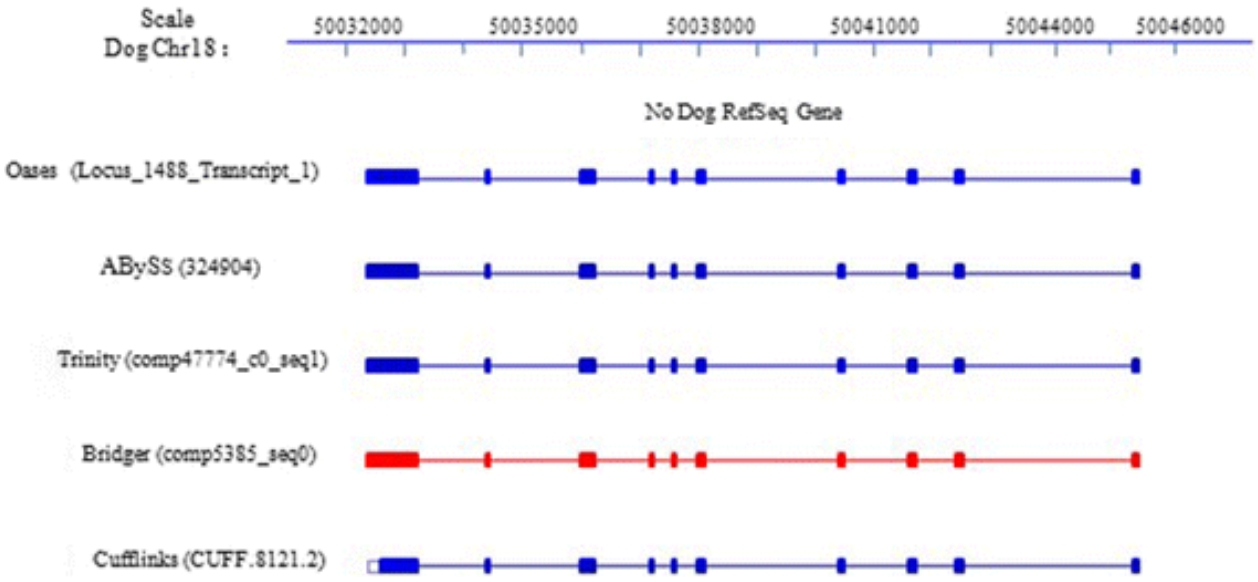
SciLifeLab

METHOD | OPEN ACCESS

Bridger: a new framework for *de novo* transcriptome assembly using RNA-seq data

Zheng Chang†, Guojun Li† ✉, Juntao Liu, Yu Zhang, Cody Ashby, Deli Liu, Carole L Cramer and Xiuzhen Huang ✉
† Contributed equally

Genome Biology 2015 16:30 | DOI: 10.1186/s13059-015-0596-2 | © Chang et al.; licensee BioMed Central. 2015
Received: 22 May 2014 | Accepted: 23 January 2015 | Published: 11 February 2015



A novel gene containing 10 exons was assembled by all assemblers. Interestingly, all *de novo* assemblers captured longer UTR than the reference-based assembler Cufflinks

De novo assembly

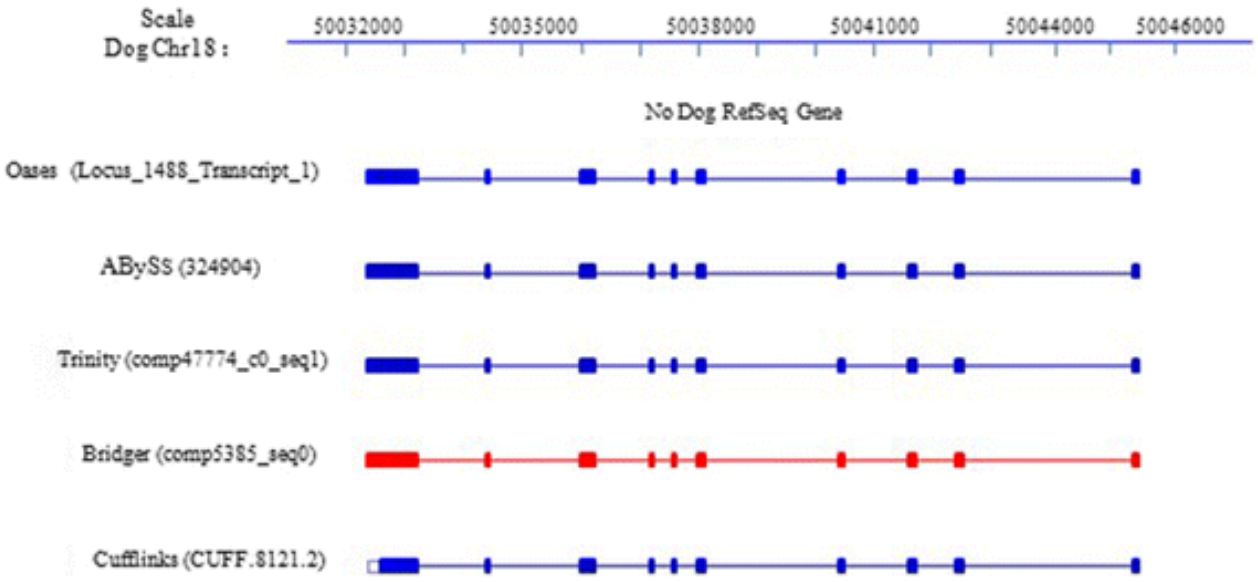
SciLifeLab

METHOD | OPEN ACCESS

Bridger: a new framework for *de novo* transcriptome assembly using RNA-seq data

Zheng Chang†, Guojun Li†, Juntao Liu, Yu Zhang, Cody Ashby, Deli Liu, Carole L Cramer and Xiuzhen Huang
†Contributed equally

Genome Biology 2015 16:30 | DOI: 10.1186/s13059-015-0596-2 | © Chang et al.; licensee BioMed Central. 2015
Received: 22 May 2014 | Accepted: 23 January 2015 | Published: 11 February 2015



A novel gene containing 10 exons was assembled by all assemblers. Interestingly, all *de novo* assemblers captured longer UTR than the reference-based assembler Cufflinks

De novo assembly

SciLifeLab

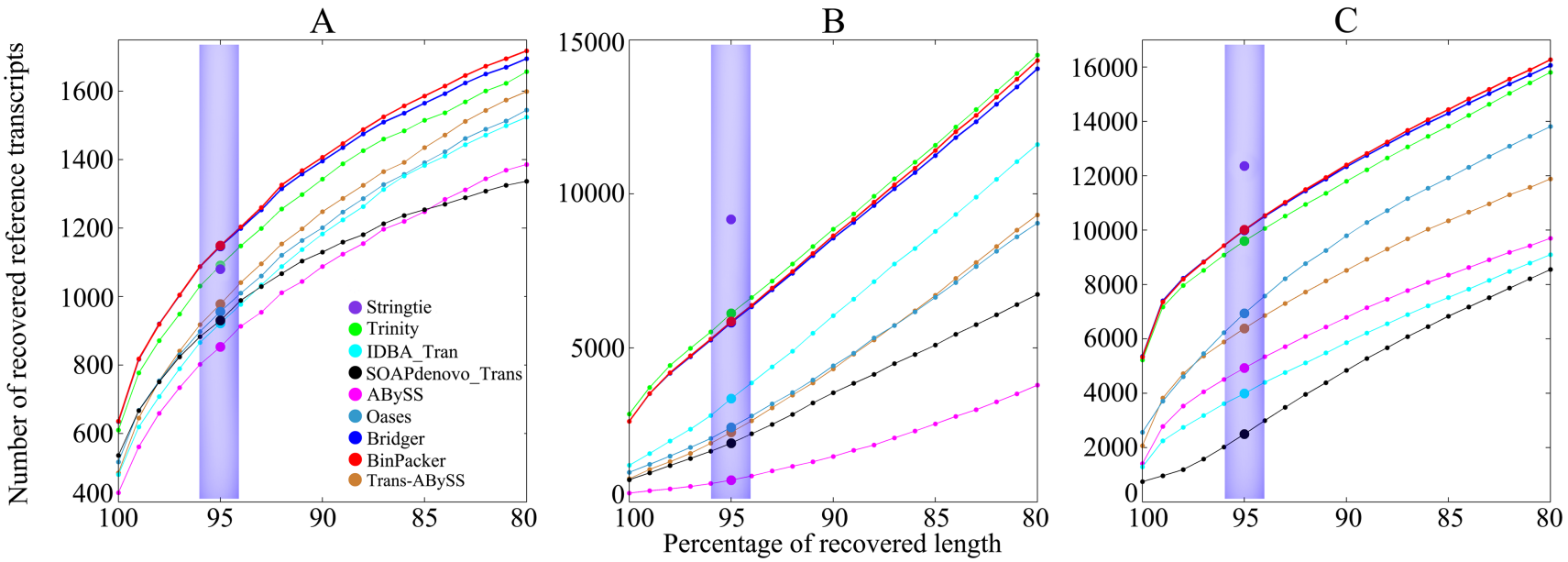
OPEN ACCESS PEER-REVIEWED

RESEARCH ARTICLE

BinPacker: Packing-Based *De Novo* Transcriptome Assembly from RNA-seq Data

Juntao Liu, Guojun Li, Zheng Chang, Ting Yu, Bingqiang Liu, Rick McMullen, Pengyin Chen, Xiuzhen Huang

Published: February 19, 2016 • http://dx.doi.org/10.1371/journal.pcbi.1004772



Comparison of recovered reference sensitivity and its distribution against recovered sequence length rates (sequence identity) ranging from 80% to 100% on (A) dog, (B) human and (C) mouse datasets.

De novo assembly

SciLifeLab

The next-next generation of transcriptome assemblers:
- Overlap based
- Take advantage of longer reads
- Accommodate 3rd gen sequencing (e.g. MinION)
- Metatranscriptomes

Under development: ***Ananas***
…from the people who brought you Trinity, coming to a server near you soon!

SciLifeLab

Take-home message

- No reference needed
- Many programs available
- Lots of potential transcripts. Filter!

Combining both methods

SciLifeLab

Improvement of genome assembly completeness and identification of novel full-length protein-coding genes by RNA-seq in the giant panda genome

Meili Chen, Yibo Hu, Jingxing Liu, Qi Wu, Chenglin Zhang, Jun Yu, Jingfa Xiao ✉, Fuwen Wei ✉ & Jiayan Wu ✉

Scientific Reports **5**, Article number: 18019 (2015)
doi:10.1038/srep18019

Received: 05 May 2015
Accepted: 10 November 2015
Published online: 11 December 2015

Combining both methods

SciLifeLab

- Background
  - 1st *de novo* assembled genome based solely on short reads (Li et al., Nature 463, 2010)
  - 23,408 genes annotated on the basis of a homology search with human and dog genes and *ab initio* methods
- RNA-seq: 12 tissues
  - liver, stomach, small intestine, colon, pallium and testis from 1 male adult
  - pituitary gland, skeletal muscle, tongue, ovary and skin from 1 female adult

Combining both methods

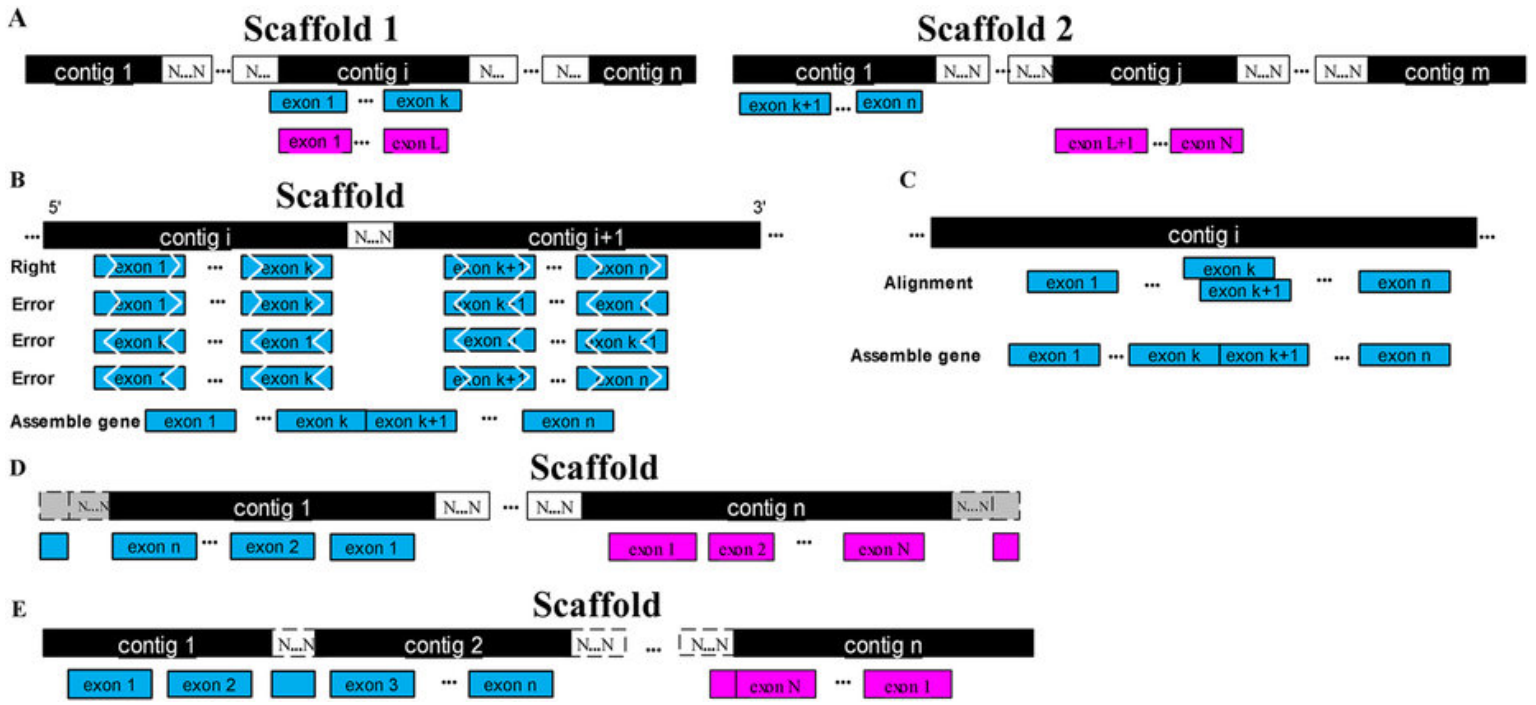
SciLifeLab

- Reference-based:
  - Transcripts reconstruction: Cufflinks (alignment: TopHat)
- De novo:
  - Transcripts reconstruction: Trinity
- 24 assemblies (12 tissues * 2 methods)
  - Merge the 12 transcriptomes for each method
  - Merge the 2 method transcriptomes

Combining both methods

SciLifeLab

Improvement of genome assembly

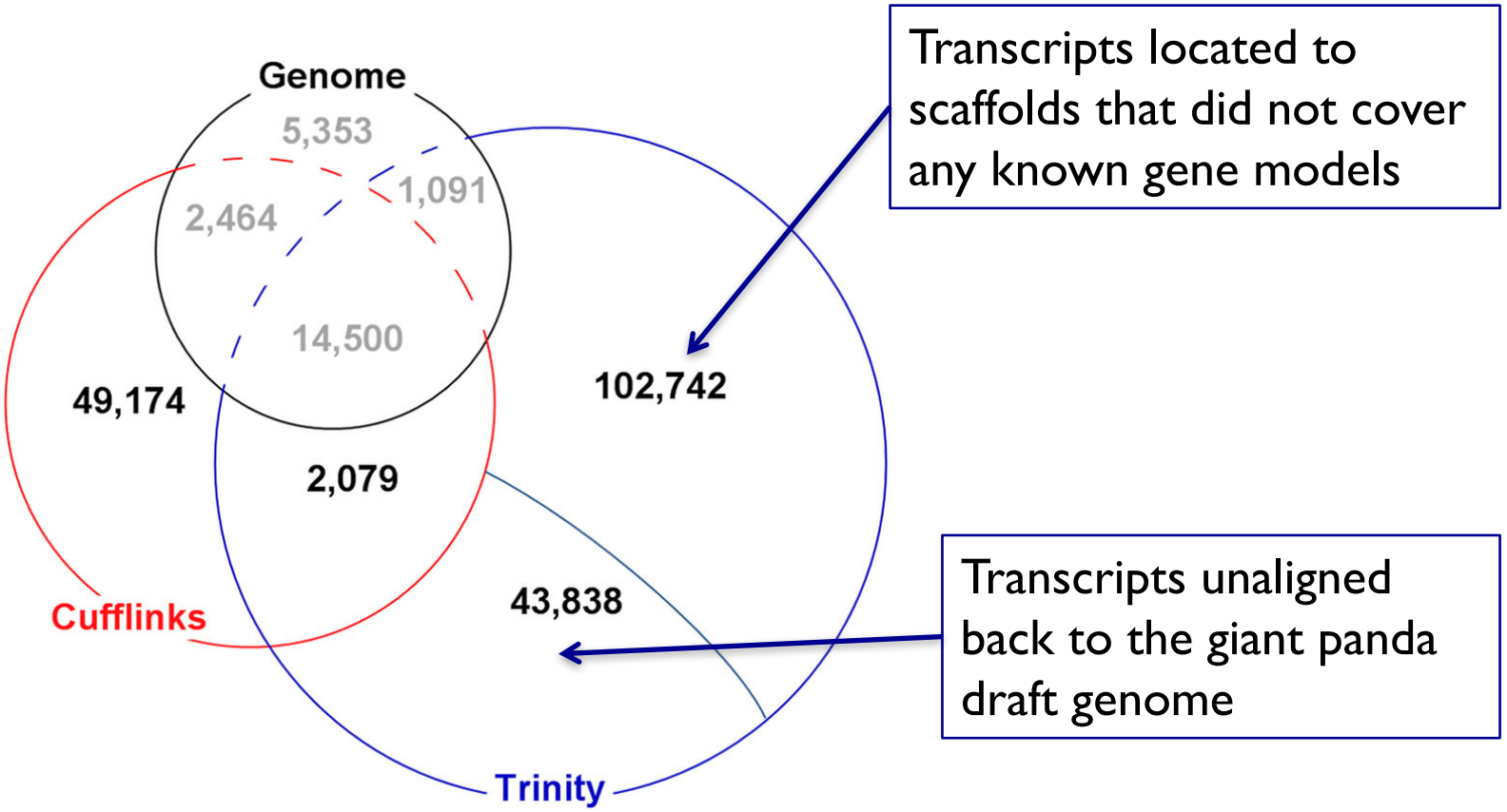


(A) Scaffolding improvement; (B) Scaffolding inconsistencies; (C) Nest assembly errors; (D) Boundary extensions; (E) Gap closure

Combining both methods

SciLifeLab

Transcriptome reconstruction



49,174 + 2,079 + 43,838 + 102,742 = 197,833 potential novel transcripts!

Combining both methods

SciLifeLab

Validation of candidate novel protein-coding genes

- ORF detection (Augustus)
  - 197,833 novel transcripts => 28,522 potential novel protein-coding genes
- Homology search (blast) – 3 categories
  - 551 (1.93%) ***homology-based genes*** that were similar to known proteins in the nr database and known cDNA sequences in the nt database;
  - 6,290 (22.03%) ***unknown genes*** that were similar to EST sequences in dbEST but had no protein or cDNA homology information;
  - 12,575 (44.09%) ***hypothetical genes*** that had a complete ORF but no known homologs.
  - 9,106 ORFs were filtered out (no start or stop codon, too short CDS…)

Combining both methods

SciLifeLab

Validation of candidate novel protein-coding genes

- Protein domain search on 19,416 ORFs (InterProScan)
  - 409 out of 551 *homology-based genes*
  - 5,112 out of 6,290 *unknown genes*
  - 7,981 out of 12,575 *hypothetical genes*
- Proteomic analysis in 5 tissues
  - 12,043 peptide hits
  - 1,691 novel protein-coding genes characterized with at least 1 peptide

Combining both methods

SciLifeLab

Take-home message

- Useful if the reference is incomplete
- Can help improving the reference
- Can help annotating the reference
- Need to filter the results!

Thank you for listening!

SciLifeLab

Questions?