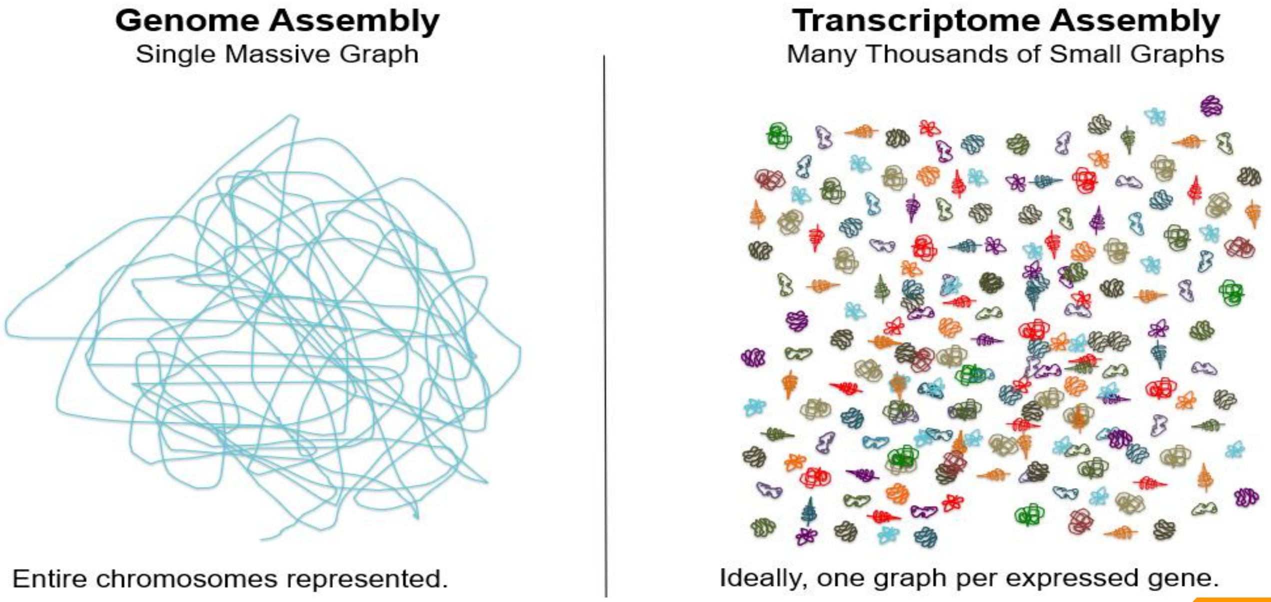
Trinity: A tool for full-length transcriptome **assembly** from RNA-Seq data without a **reference genome**

2021.10.06
Advanced Bioinformatics 1
Presenter : Catherine Apio

nature
biotechnology

Introduction

- What is assembly?



Genome Assembly

Single Massive Graph

Entire chromosomes represented.

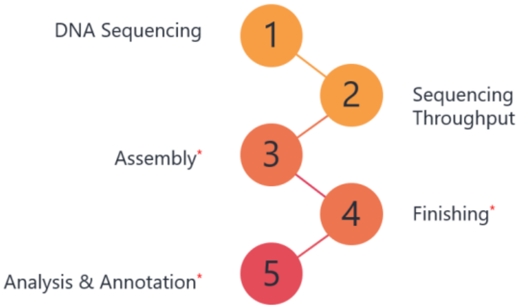
Transcriptome Assembly

Many Thousands of Small Graphs

Ideally, one graph per expressed gene.

Introduction

◆ What is assembly



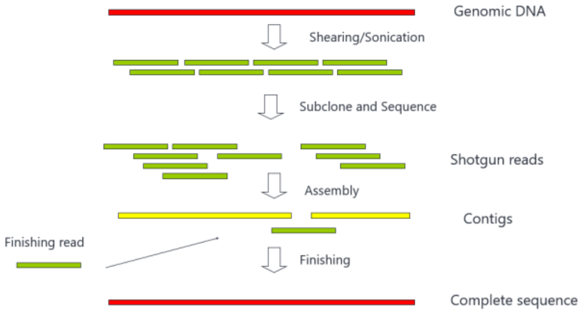
DNA Sequencing 1

2 Sequencing Throughput

Assembly* 3

4 Finishing*

Analysis & Annotation* 5



Genomic DNA

Shearing/Sonication

Subclone and Sequence

Shotgun reads

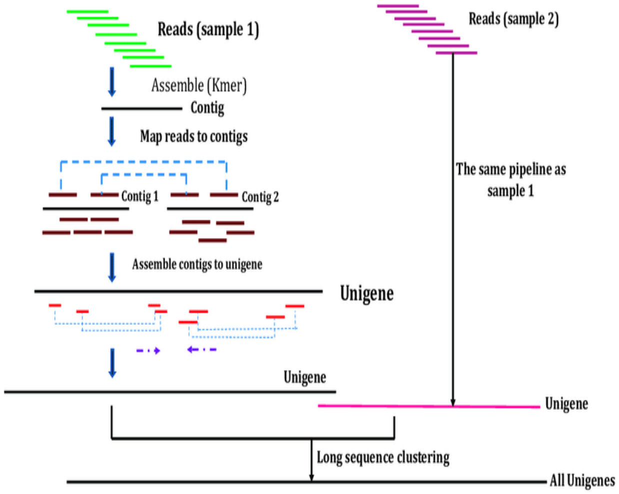
Assembly

Contigs

Finishing read

Finishing

Complete sequence



Reads (sample 1)

Assemble (Kmer)

Contig

Map reads to contigs

Contig 1

Contig 2

Assemble contigs to unigene

Unigene

Unigene

Reads (sample 2)

The same pipeline as sample 1

Unigene

Long sequence clustering

All Unigenes

Introduction

- Challenges of transcriptome assembly;

Low or high transcript coverage

Uneven transcript coverage along length

Sequencing errors

Chimeric transcript

Alternative splicing and repetitions across genes

Introduction

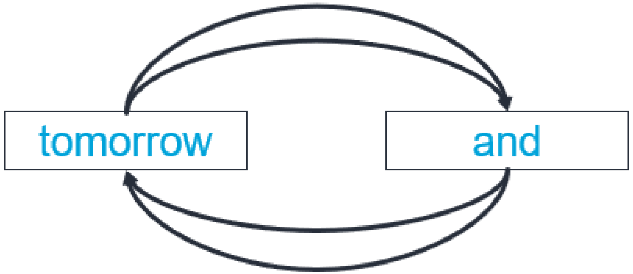
- Alternative computational strategies for transcriptome reconstruction;
  - ✓ Mapping-first approaches (Scripture, Cufflinks)
  - ✓ Assembly-first methods (ABySS, SOAPdenovo, Oases)

Introduction

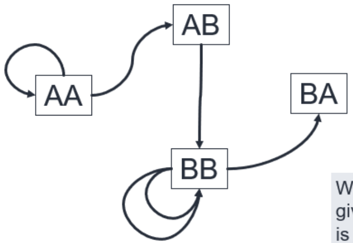
Mapping- first approaches	Assembly-first methods
Align reads to a reference (unannotated) genome, merge overlapping sequences	Use the reads to assemble transcripts directly
Maximum sensitivity (correct reference genome)	Do not require read-reference genome
Complicated; splicing, sequence errors, lack or incomplete reference genome	Good for gapped, highly fragmented or altered sequences
More progress	Less progress; solved by de Bruijn graph

de Bruijn graph

“tomorrow and tomorrow and tomorrow”



- Genome: AAABBBBA
- 3-mers: AAA, AAB, ABB, BBB, BBB, BBA
- 2-mers: AA, AA, AA, AB, AB, BB, BB, BB, BB, BB, BB, BA



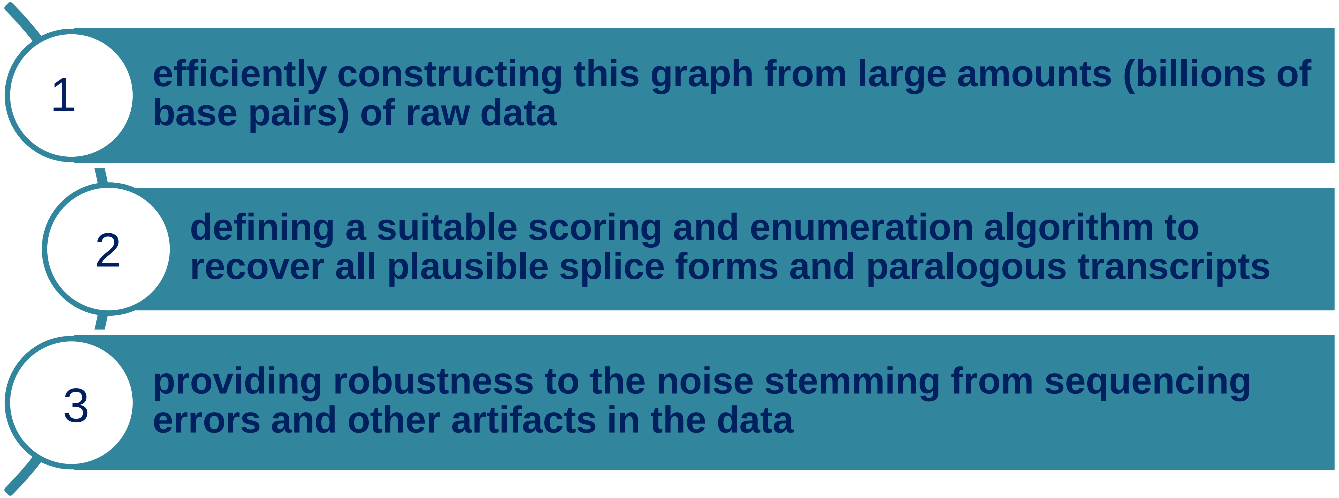
AAABBBBA

Q: Can we reconstruct the genome from the De Brujin graph?

Walking across each edge exactly once gives a reconstruction of the genome. This is an Eulerian walk.

de Bruijn graph

- Challenges of de Bruijn graphs to *de novo* assembly of RNA-Seq data;



1. **efficiently constructing this graph from large amounts (billions of base pairs) of raw data**
2. **defining a suitable scoring and enumeration algorithm to recover all plausible splice forms and paralogous transcripts**
3. **providing robustness to the noise stemming from sequencing errors and other artifacts in the data**

Trinity

- ➢ A method for the efficient and robust *de novo* reconstruction of transcriptomes, consisting of three software modules;
  - ✓ Inchworm - assembles contigs
  - ✓ Chrysalis - builds de Bruijn graph
  - ✓ Butterfly - resolves

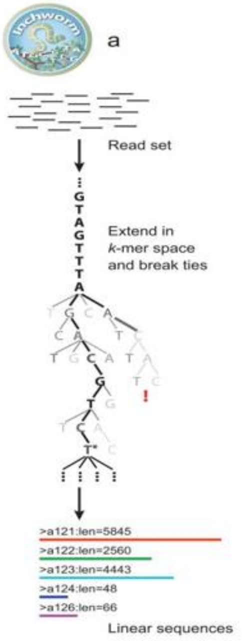
- ➢ Trinity was evaluated using;
  - ✓ Micro-organism (fission yeast; S. pombe)
  - ✓ Mammal (mouse)
  - ✓ Insect (whitefly) – no genome yet

Inchworm

- ➢ Inchworm assembles reads using a greedy *k-mer* based approach for fast and efficient transcript assembly.

- ➢ Recovers only a single (best) representative (owing to alternative splicing, gene duplication or allelic variation)

- ➢ Inchworm efficiently reconstructs linear transcript contigs in six steps



Inchworm

- Steps;

i. constructs a *k-mer* dictionary from all sequence reads (k = 25)

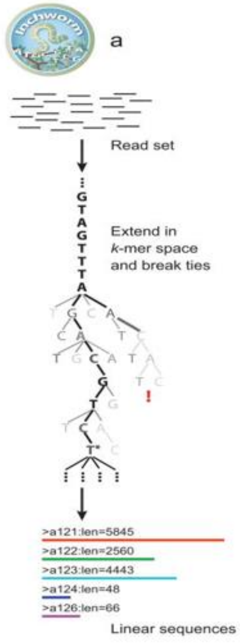
ii. removes likely error-containing *k-mers* from the *k-mer* dictionary (low-complexity and singleton *k-mers* excluded)

iii. selects the most frequent *k-mer* in the dictionary to seed a contig assembly

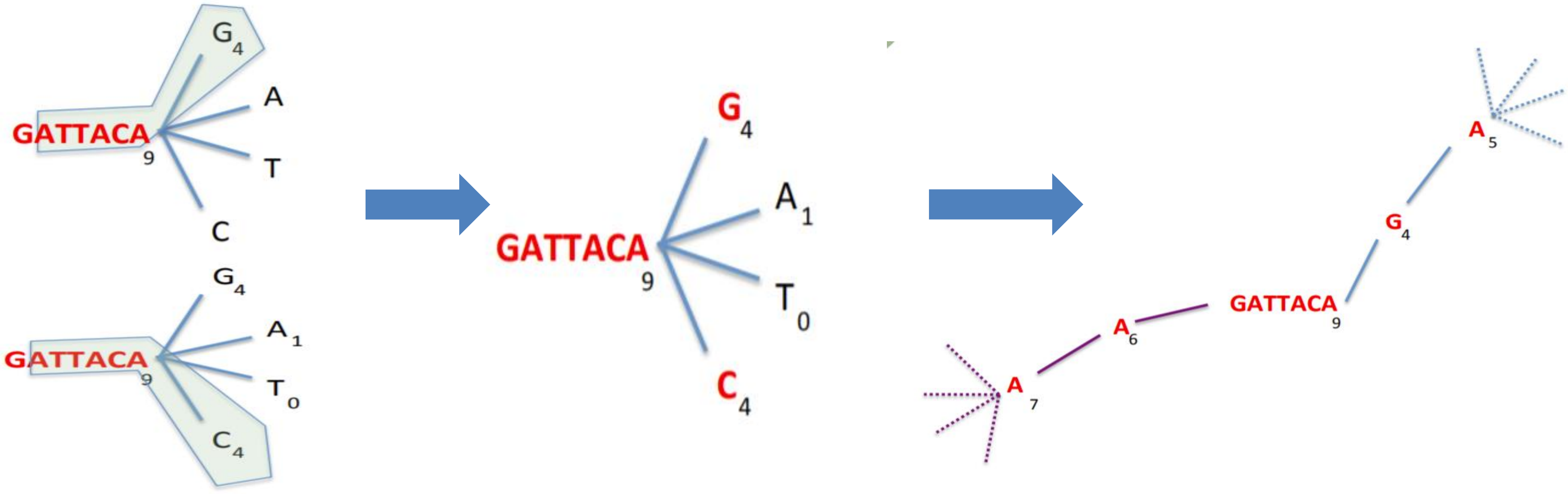
iv. Extension of seed with highest k-mer with k-1 overlap and concatenation

v. extends the sequence in either direction until no further extension

vi. repeats steps iii–v, starting with the next most abundant *k-mer*, until the entire *k-mer* dictionary has been exhausted



Inchworm



Report contig: ....AAGATTACAGA....

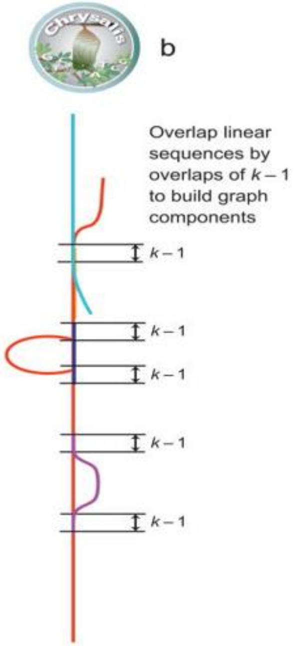
Chrysalis

- Chrysalis clusters minimally overlapping Inchworm contigs into sets of connected components

- Constructs complete de Bruijn graphs for each component.

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

- Chrysalis works in three phases.


Chrysalis
b
Overlap linear sequences by overlaps of $k-1$ to build graph components
$k-1$
$k-1$
$k-1$
$k-1$
$k-1$

Chrysalis

- The phases are;

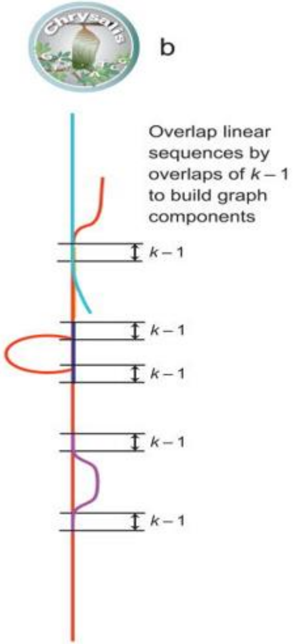
1 It recursively groups Inchworm contigs into connected components;

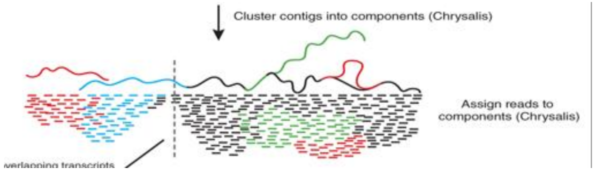
k – 1 bases overlap, (*k* – 1)/2 base across junctions

2 It builds a de Bruijn graph for each component;

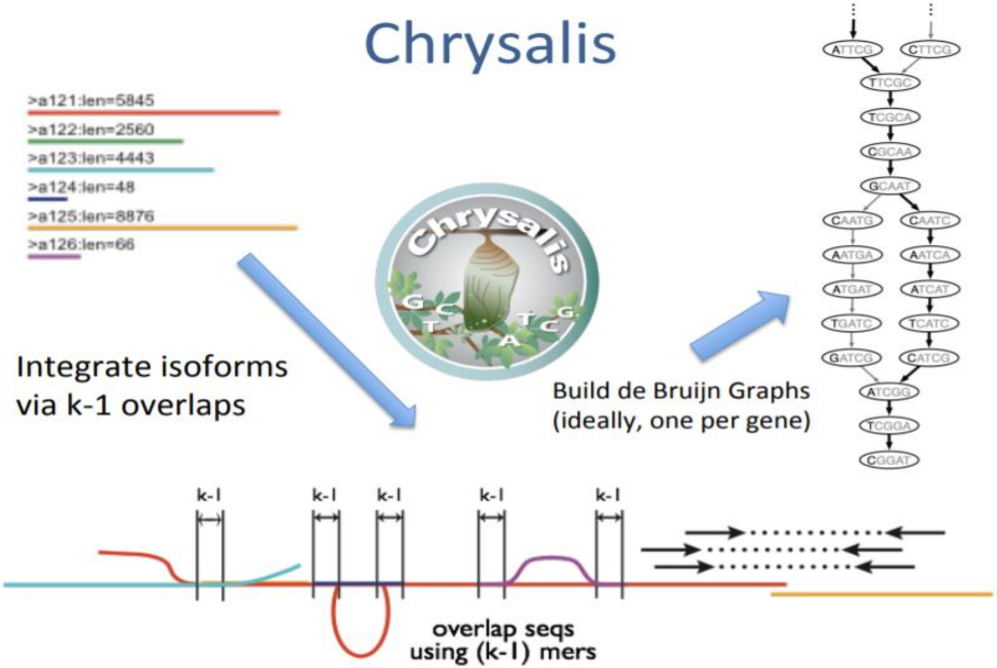
k – 1 nodes, k edges and weighs edges

3 Assigns reads with largest k-mers to components


Chrysalis
b
Overlap linear sequences by overlaps of $k-1$ to build graph components
$k-1$
$k-1$
$k-1$
$k-1$
$k-1$


Cluster contigs into components (Chrysalis)
Assign reads to components (Chrysalis)
verlapping transcripts

Chrysalis

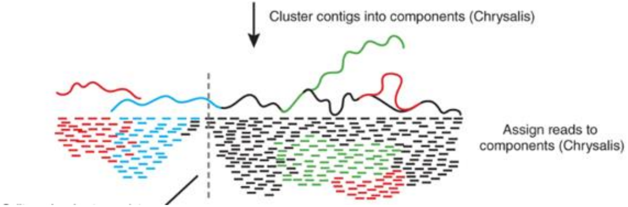


Chrysalis

Integrate isoforms via k-1 overlaps

Build de Bruijn Graphs (ideally, one per gene)

overlap seqs using (k-1) mers



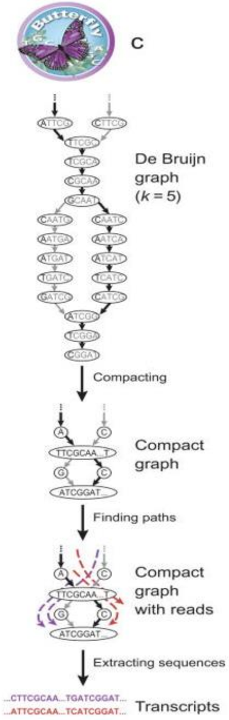
Cluster contigs into components (Chrysalis)

Assign reads to components (Chrysalis)

Butterfly

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

- It consists of 2 parts;
  - ✓ Graph simplification
  - ✓ Plausible path scoring

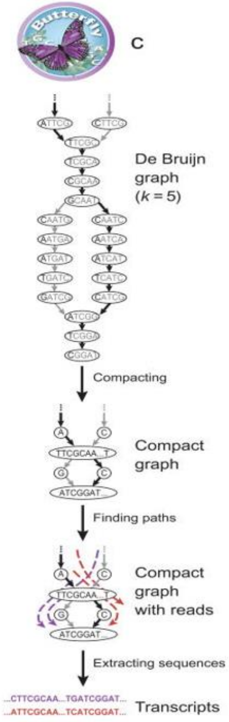


Butterfly

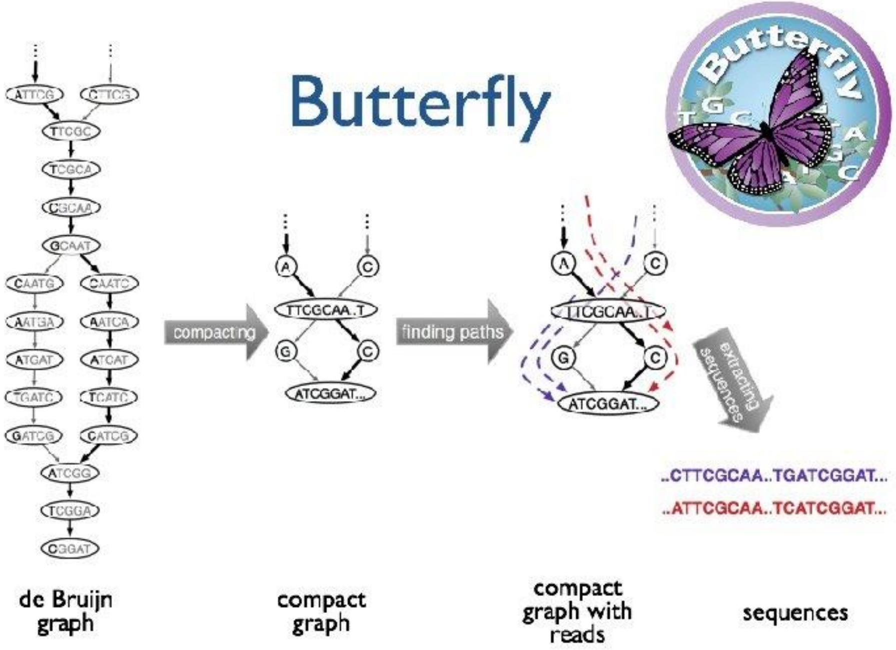
- During graph simplification, Butterfly iterates between;

  - merging consecutive nodes in linear paths in the de Bruijn graph to form nodes that represent longer sequences

  - pruning edges that represent minor deviations supported by fewer reads (sequencing errors)

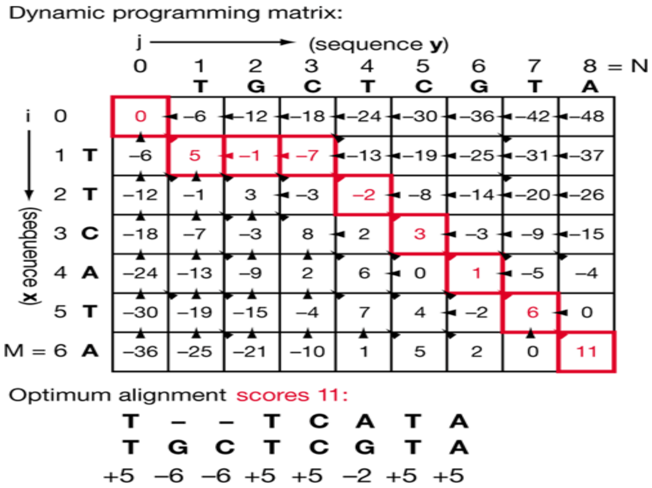


Butterfly


Butterfly
Butterfly
ATTCG
CTTCG
TTCGC
TCGCA
CGCAA
GCAAT
CAATG
CAATC
AATGA
AATCA
ATGAT
ATCAT
TGATC
TCATC
GATCG
CATCG
ATCGG
TCGGA
CGGAT
compacting
A
C
TTCGCAA..T
G
C
ATCGGAT...
finding paths
A
C
TTCGCAA..T
G
C
ATCGGAT...
extracting sequences
..CTTCGCAA..TGATCGGAT...
..ATTCGCAA..TCATCGGAT...
de Bruijn graph
compact graph
compact graph with reads
sequences

Butterfly

- In plausible, Butterfly identifies paths that are supported by actual reads and read pairs, using a dynamic programming procedure.



Dynamic programming matrix:

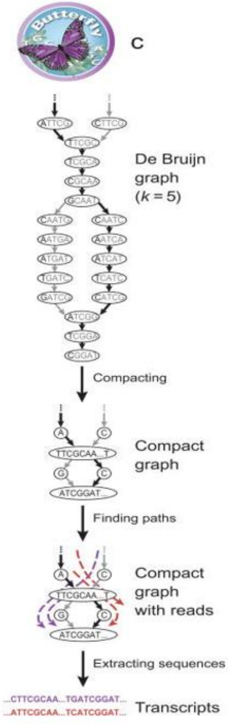
j → (sequence **y**)

i ↓ (sequence **x**)

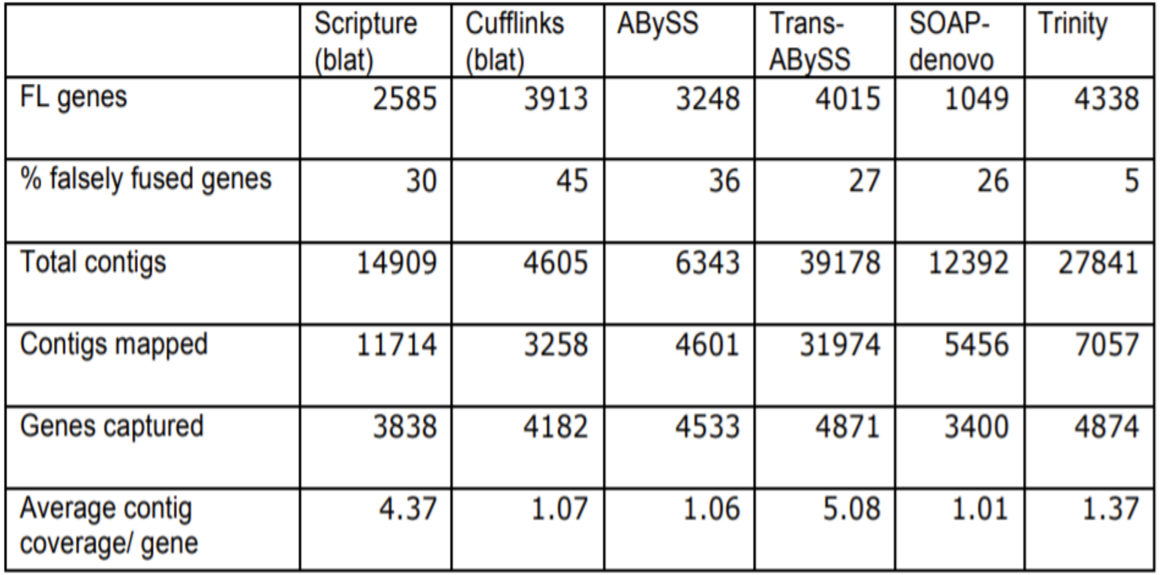
		0	1	2	3	4	5	6	7	8 = N
			T	**G**	**C**	**T**	**C**	**G**	**T**	**A**
i 0		0	–6	–12	–18	–24	–30	–36	–42	–48
1	**T**	–6	5	–1	–7	–13	–19	–25	–31	–37
2	**T**	–12	–1	3	–3	–2	–8	–14	–20	–26
3	**C**	–18	–7	–3	8	2	3	–3	–9	–15
4	**A**	–24	–13	–9	2	6	0	1	–5	–4
5	**T**	–30	–19	–15	–4	7	4	–2	6	0
M = 6	**A**	–36	–25	–21	–10	1	5	2	0	11

Optimum alignment scores 11:

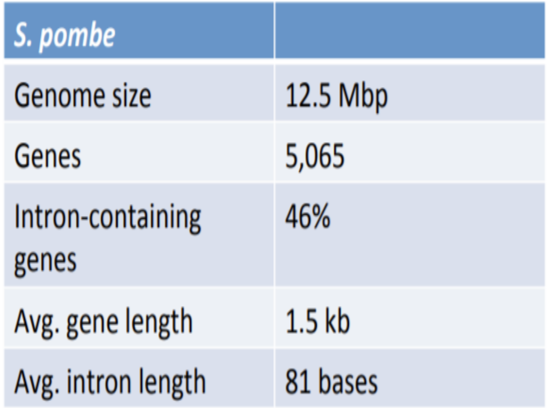
```
 T   –   –   T   C   A   T   A
 T   G   C   T   C   G   T   A
+5  –6  –6  +5  +5  –2  +5  +5
```



Results

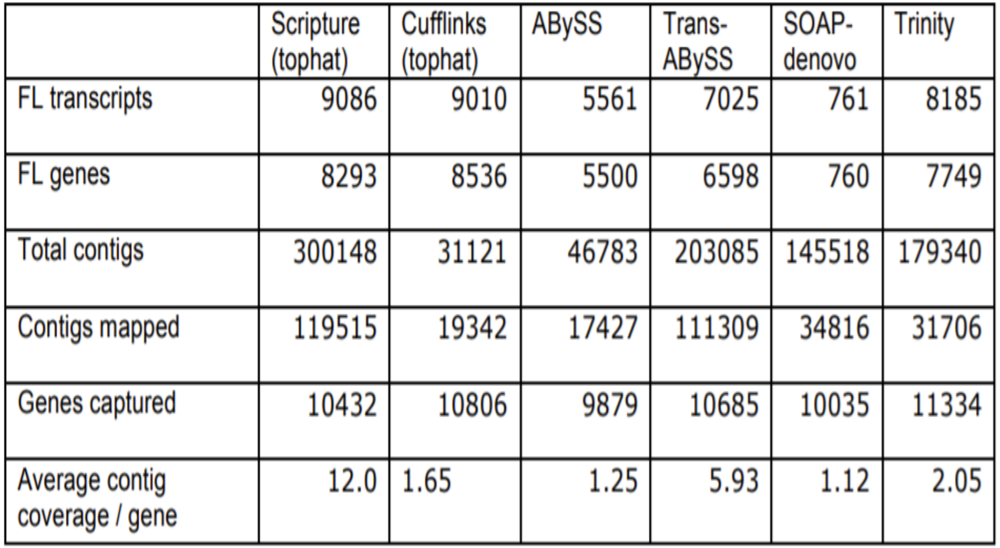


	Scripture (blat)	Cufflinks (blat)	ABySS	Trans-ABySS	SOAP-denovo	Trinity
FL genes	2585	3913	3248	4015	1049	4338
% falsely fused genes	30	45	36	27	26	5
Total contigs	14909	4605	6343	39178	12392	27841
Contigs mapped	11714	3258	4601	31974	5456	7057
Genes captured	3838	4182	4533	4871	3400	4874
Average contig coverage/ gene	4.37	1.07	1.06	5.08	1.01	1.37

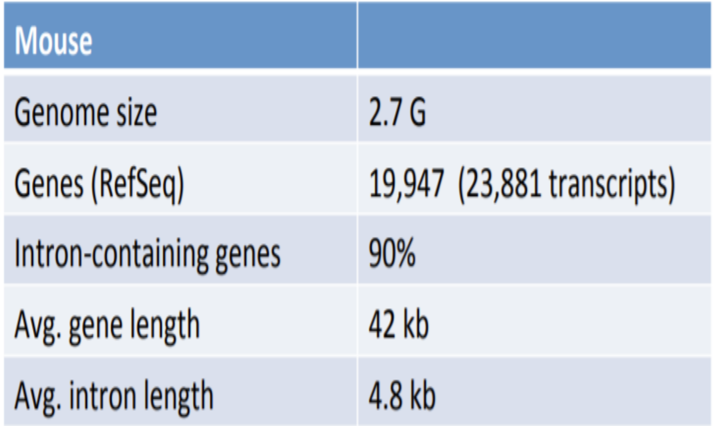


S. pombe	
Genome size	12.5 Mbp
Genes	5,065
Intron-containing genes	46%
Avg. gene length	1.5 kb
Avg. intron length	81 bases

Results

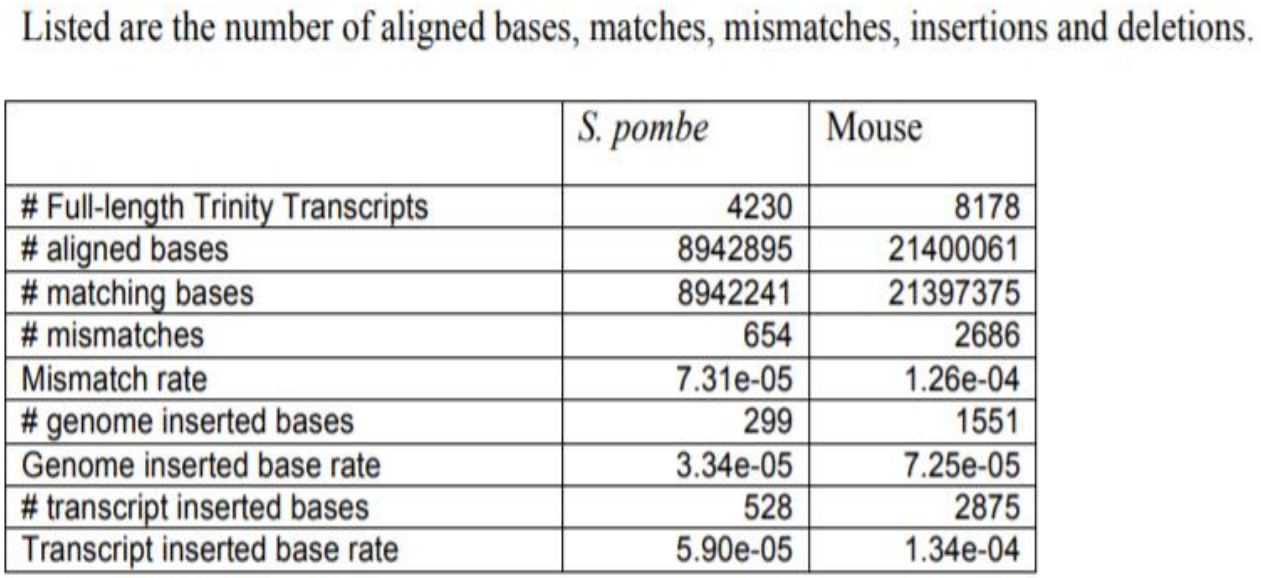


	Scripture (tophat)	Cufflinks (tophat)	ABySS	Trans-ABySS	SOAP-denovo	Trinity
FL transcripts	9086	9010	5561	7025	761	8185
FL genes	8293	8536	5500	6598	760	7749
Total contigs	300148	31121	46783	203085	145518	179340
Contigs mapped	119515	19342	17427	111309	34816	31706
Genes captured	10432	10806	9879	10685	10035	11334
Average contig coverage / gene	12.0	1.65	1.25	5.93	1.12	2.05



Mouse	
Genome size	2.7 G
Genes (RefSeq)	19,947 (23,881 transcripts)
Intron-containing genes	90%
Avg. gene length	42 kb
Avg. intron length	4.8 kb

Results

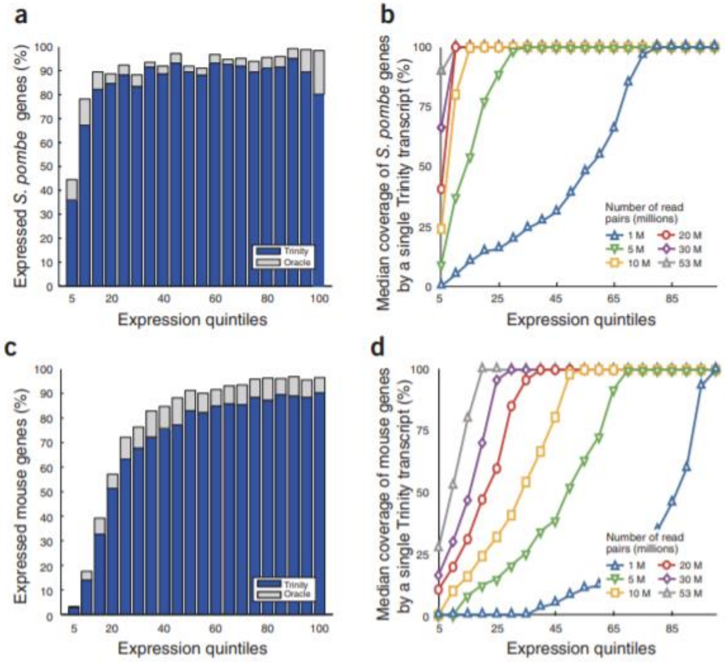


Listed are the number of aligned bases, matches, mismatches, insertions and deletions.

	S. pombe	Mouse
# Full-length Trinity Transcripts	4230	8178
# aligned bases	8942895	21400061
# matching bases	8942241	21397375
# mismatches	654	2686
Mismatch rate	7.31e-05	1.26e-04
# genome inserted bases	299	1551
Genome inserted base rate	3.34e-05	7.25e-05
# transcript inserted bases	528	2875
Transcript inserted base rate	5.90e-05	1.34e-04

Results

Figure 2 Trinity correctly reconstructs the majority of full-length transcripts in fission yeast and mouse. (**a**,**c**) The fraction of genes that are fully reconstructed and in the Oracle Set in different expression quintiles (5% increments) in fission yeast (50 M pairs assembly) (**a**) and the fraction of genes that have at least one fully reconstructed transcript and are in the Oracle Set in different expression quintiles in mouse (53 M pairs assembly) (**c**). Each bar represents a 5% quintile of read coverage for genes expressed. Gray bars show the remaining fraction of transcripts that are in the Oracle Set but not fully reconstructed. For example, ~36% of the *S. pombe* transcripts at the bottom 5% of expression levels are fully reconstructed by Trinity; ~45% of the transcripts in this quintile are in the Oracle Set. (**b**,**d**) Curves show the median values for coverage (as fraction of length of reference transcripts) by the longest corresponding Trinity-assembled transcript, according to expression quintiles in yeast (**b**) and mouse (**d**), depending on the number of read pairs that went into each assembly.

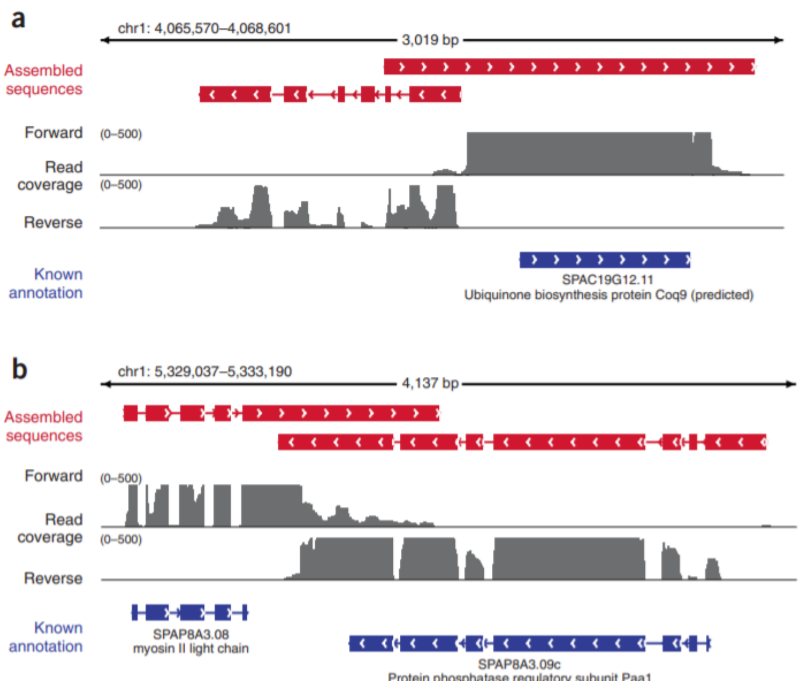


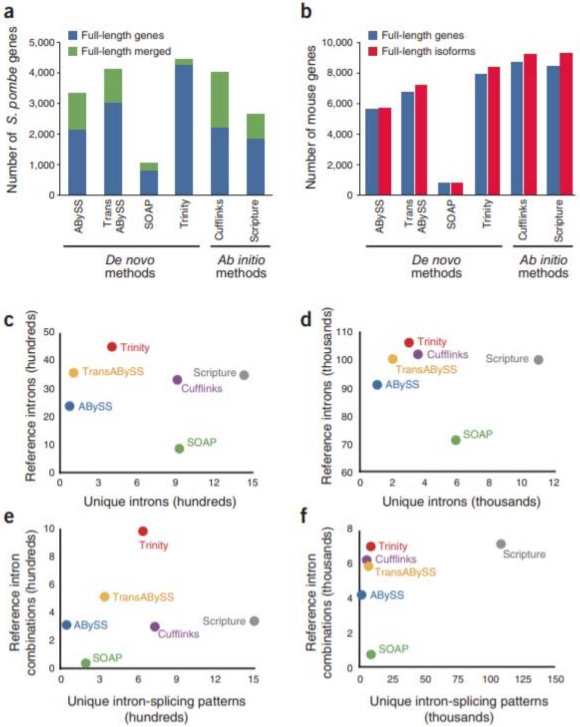
yeast

mouse

and singleton *k*-mers (appearing only once); (iv) extends the seed in each direction by finding the highest occurring *k*-mer with a *k* – 1 overlap with the current contig terminus and concatenating its terminal base to the growing

Results


a
chr1: 4,065,570–4,068,601
3,019 bp
Assembled sequences
Forward
(0–500)
Read coverage
(0–500)
Reverse
Known annotation
SPAC19G12.11
Ubiquinone biosynthesis protein Coq9 (predicted)
b
chr1: 5,329,037–5,333,190
4,137 bp
Assembled sequences
Forward
(0–500)
Read coverage
(0–500)
Reverse
Known annotation
SPAP8A3.08
myosin II light chain
SPAP8A3.09c
Protein phosphatase regulatory subunit Paa1


a
Number of S. pombe genes
0
1,000
2,000
3,000
4,000
5,000
Full-length genes
Full-length merged
ABySS
Trans ABySS
SOAP
Trinity
Cufflinks
Scripture
De novo methods
Ab initio methods
b
Number of mouse genes
0
2,000
4,000
6,000
8,000
10,000
Full-length genes
Full-length isoforms
ABySS
Trans ABySS
SOAP
Trinity
Cufflinks
Scripture
De novo methods
Ab initio methods
c
Reference introns (hundreds)
0
10
20
30
40
50
Trinity
TransABySS
ABySS
Scripture
Cufflinks
SOAP
0
3
6
9
12
15
Unique introns (hundreds)
d
Reference introns (thousands)
60
70
80
90
100
110
Trinity
Cufflinks
TransABySS
ABySS
Scripture
SOAP
0
2
4
6
8
10
12
Unique introns (thousands)
e
Reference intron combinations (hundreds)
0
2
4
6
8
10
Trinity
TransABySS
ABySS
Cufflinks
Scripture
SOAP
0
3
6
9
12
15
Unique intron-splicing patterns (hundreds)
f
Reference intron combinations (thousands)
0
2
4
6
8
Trinity
Cufflinks
TransABySS
ABySS
Scripture
SOAP
0
25
50
75
100
125
150
Unique intron-splicing patterns (thousands)

Conclusion

- Early successes achieved in exploring Trinity de novo transcriptome assemblies for downstream analyses
  - – differential expression, SNPs, and gene content studies
- Transcriptome assembly is an attractive alternative, **but not a substitute** to a genome assembly.
  - – Clear limitations (e.g. genes must be expressed!).

THANKS!

Any questions?