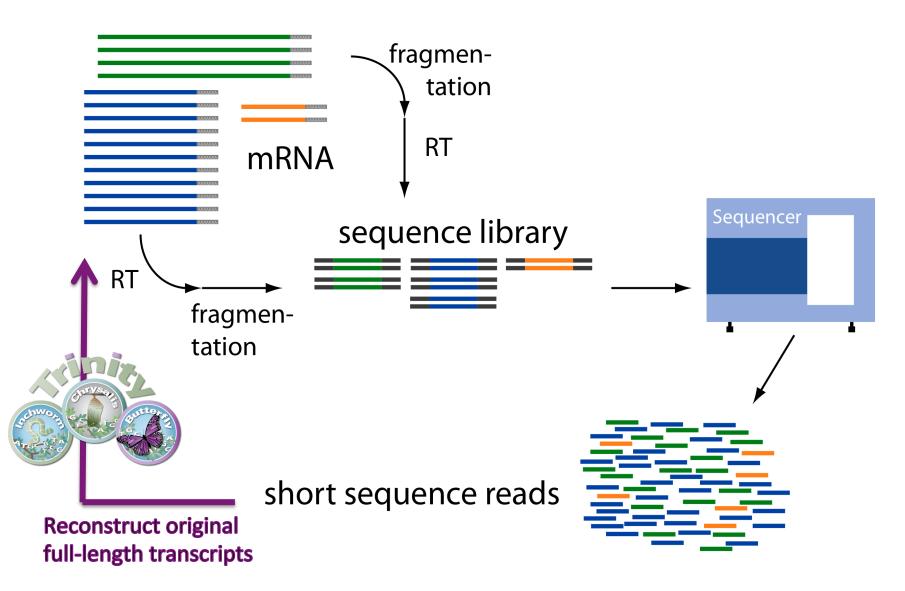
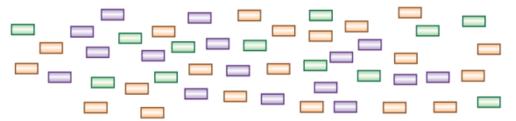
Assembly Required



From: http://www2.fml.tuebingen.mpg.de/raetsch/members/research/transcriptomics.html

RNA-Seq reads

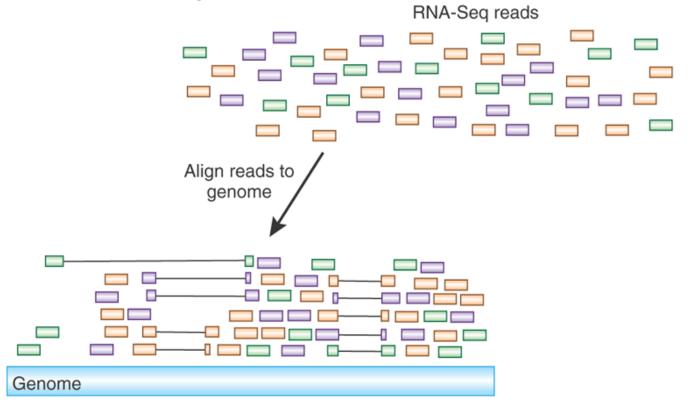


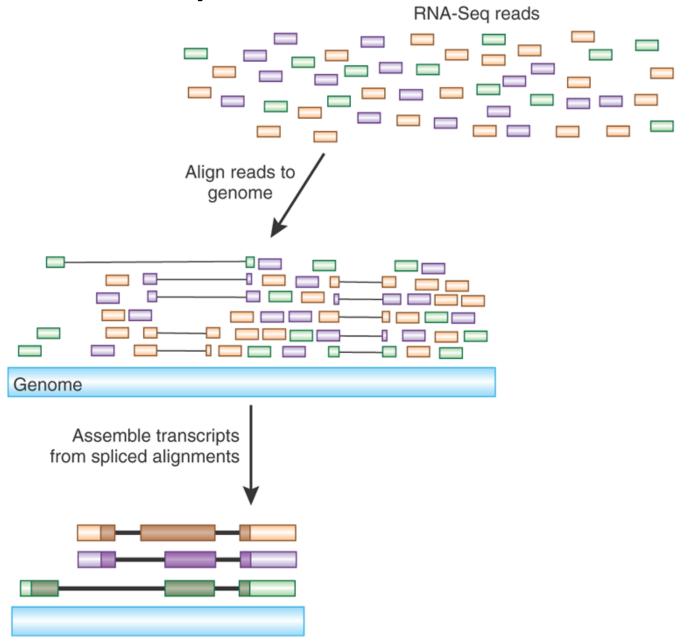
Advancing RNA-Seq analysis

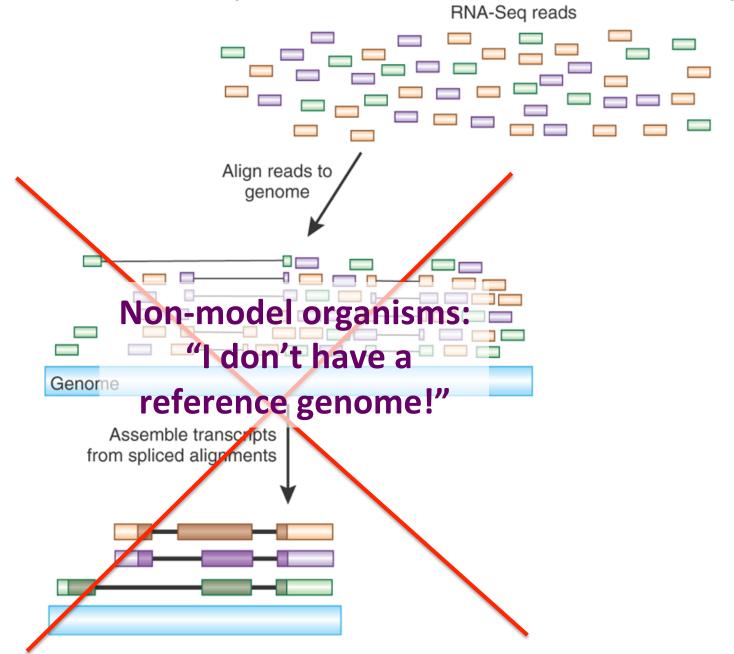
Brian J Haas & Michael C Zody

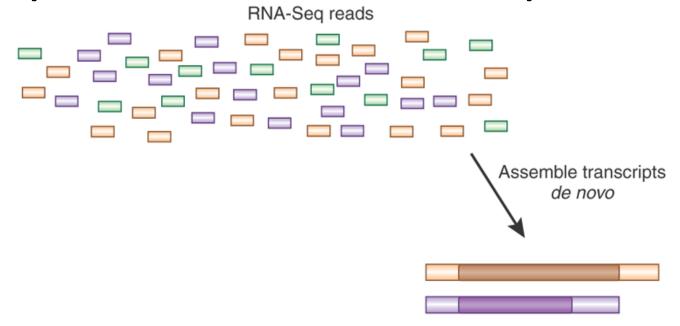
Nature Biotech, 2010

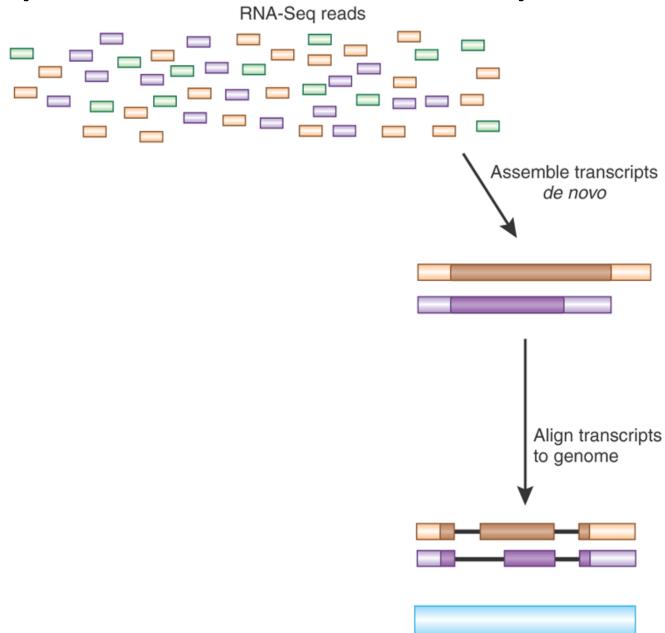
New methods for analyzing RNA-Seq data enable *de novo* reconstruction of the transcriptome.

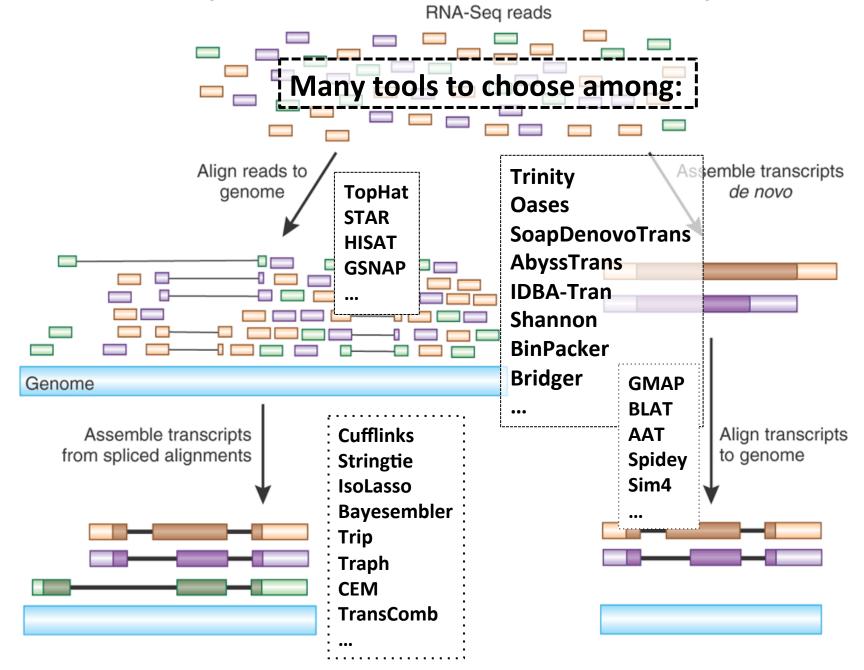




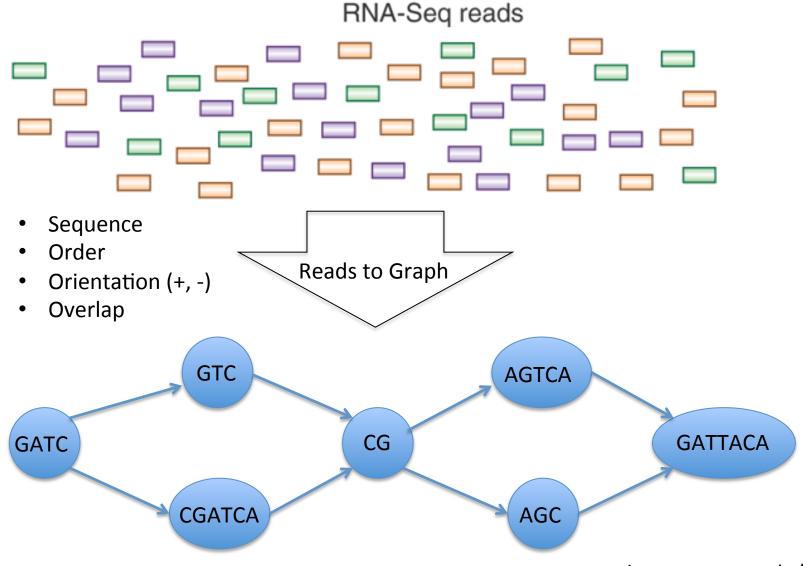






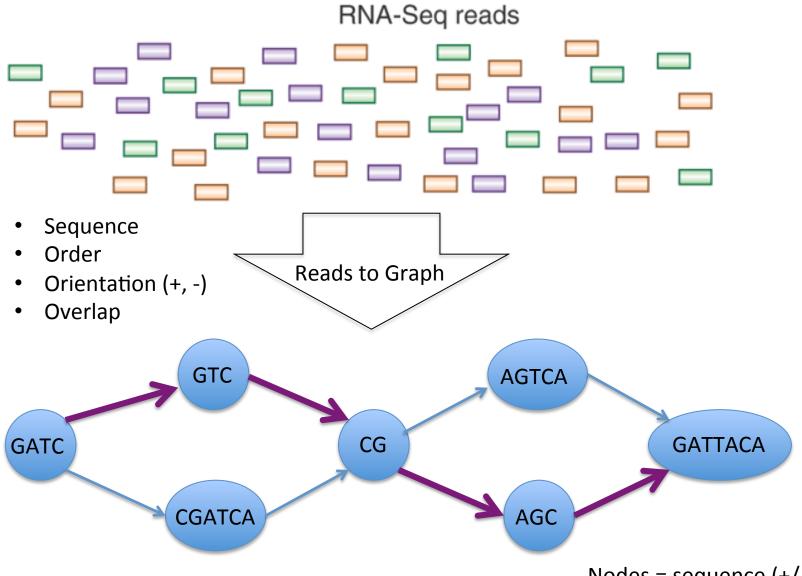


Graph Data Structures Commonly Used For Assembly



Nodes = sequence (+/-) Edges = order, overlap

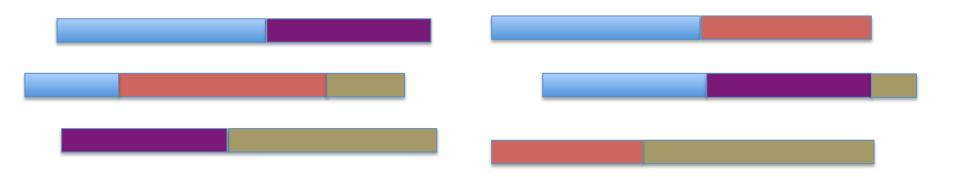
Graph Data Structures Commonly Used For Assembly



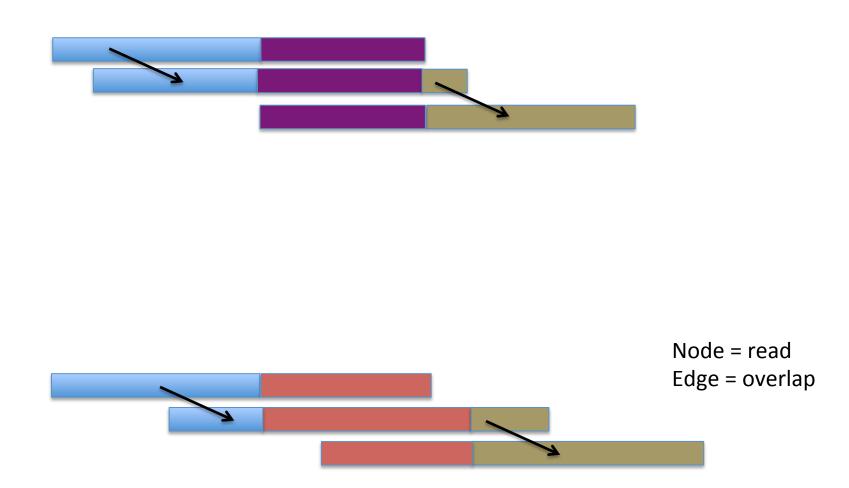
GATCGTCCGAGCGATTACA

Nodes = sequence (+/-) Edges = order, overlap

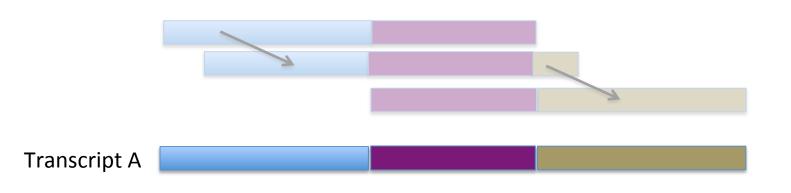
Read Overlap Graph: Reads as nodes, overlaps as edges



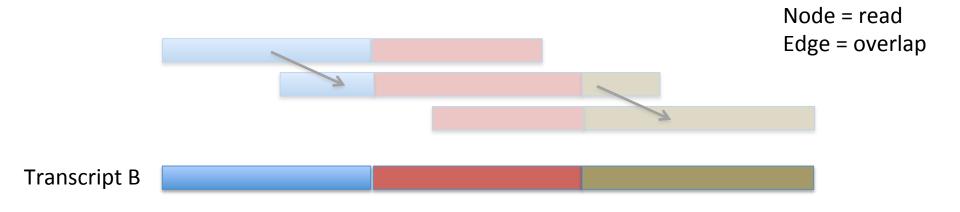
Read Overlap Graph: Reads as nodes, overlaps as edges



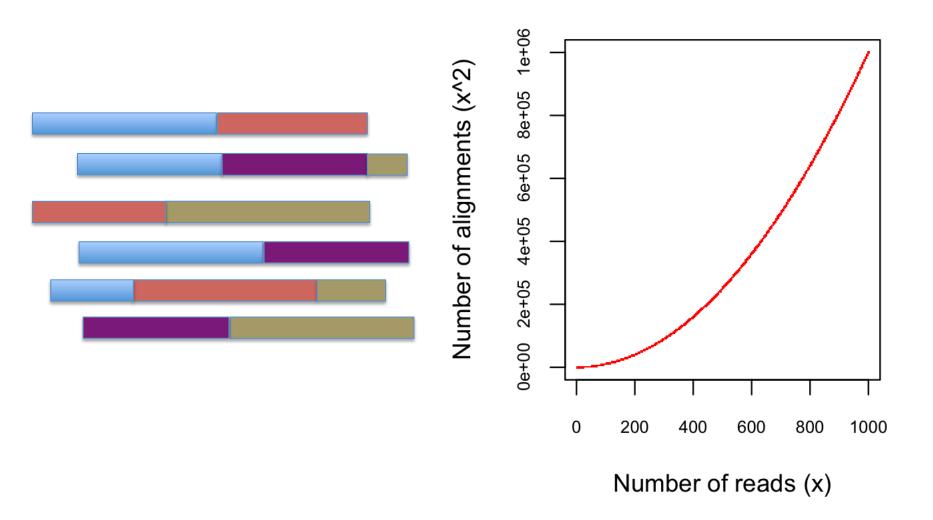
Read Overlap Graph: Reads as nodes, overlaps as edges



Generate consensus sequence where reads overlap

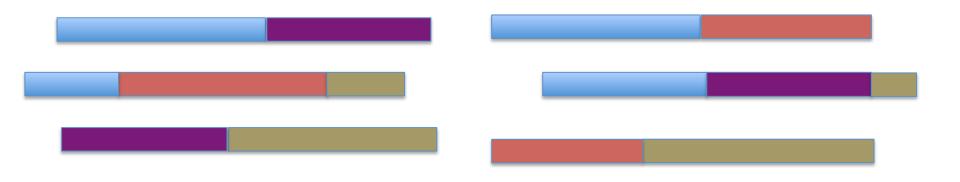


Finding pairwise overlaps between n reads involves $\sim n^2$ comparisons.



Impractical for typical RNA-Seq data (50M reads)

No genome to align to... De novo assembly required



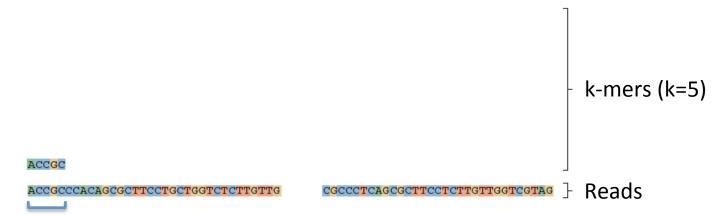
Want to avoid n^2 read alignments to define overlaps

Use a de Bruijn graph

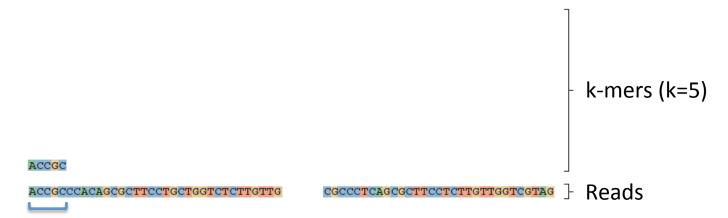
Generate all substrings of length k from the reads



Generate all substrings of length k from the reads

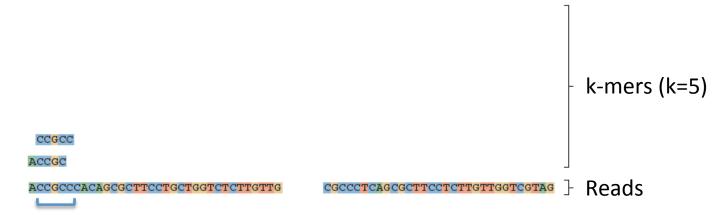


Generate all substrings of length k from the reads





Generate all substrings of length k from the reads





Generate all substrings of length k from the reads

```
- k-mers (k=5)

(k-1) overlap

CCGCC

ACCGC

ACCGC

ACCGCC

ACCGCC

ACCGCCCACAGCGCTTCCTGTTGTTC

CGCCCTCAGCGCTTCCTTGTTGGTCGTAG

Reads
```



Generate all substrings of length k from the reads

```
- k-mers (k=5)

(k-1) overlap

CCGCC

ACCGC

ACCGC

ACCGCC

ACCGCC

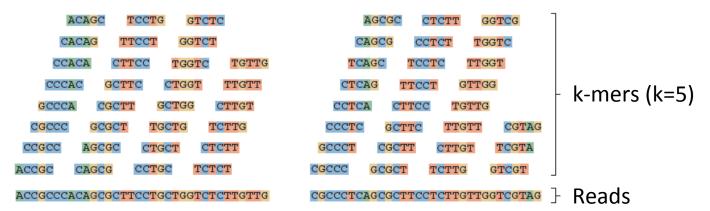
ACCGCCCACAGCGCTTCCTGTTGTTC

CGCCCTCAGCGCTTCCTTGTTGGTCGTAG

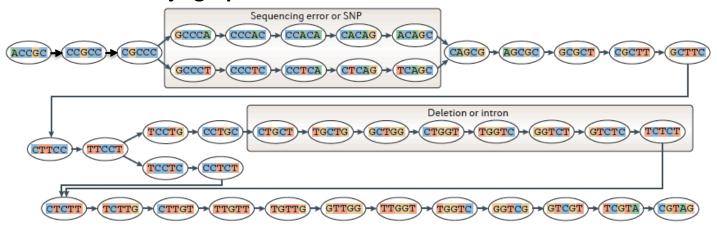
Reads
```



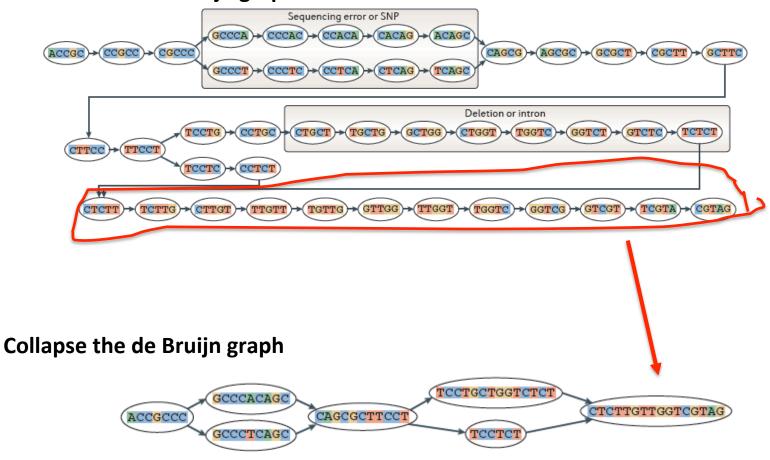
Generate all substrings of length k from the reads



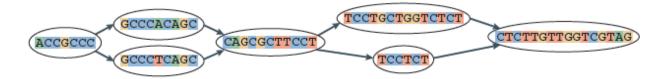
Construct the de Bruijn graph



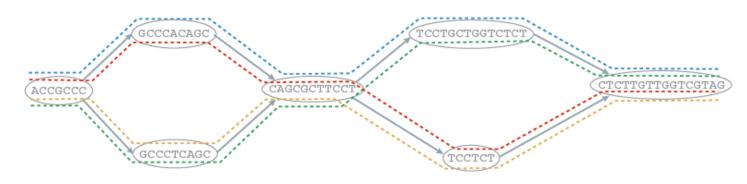
Nodes = unique k-mers Edges = overlap by (k-1)



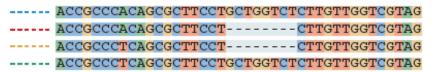
Collapse the de Bruijn graph



Traverse the graph



Assemble Transcript Isoforms



Contrasting Genome and Transcriptome *De novo* Assembly

Genome Assembly

- Uniform coverage
- Single contig per locus
- Assemble small numbers of large Mb-length chromosomes
- Double-stranded data

Transcriptome Assembly

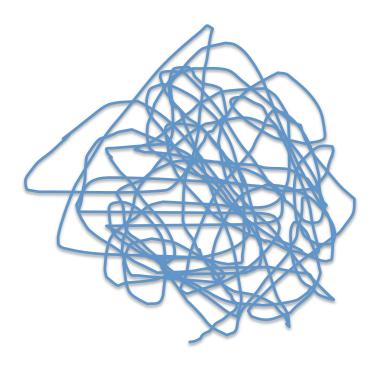
- Exponentially distributed coverage levels
- Multiple contigs per locus (alt splicing)
- Assemble many thousands of Kb-length transcripts
- Strand-specific data available



Trinity Aggregates Isolated Transcript Graphs

Genome Assembly

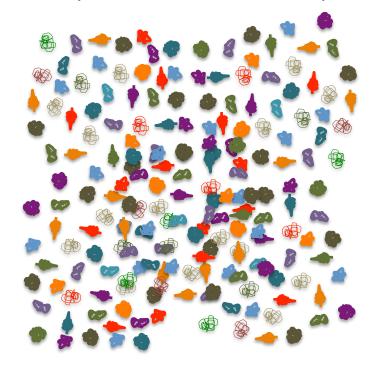
Single Massive Graph



Entire chromosomes represented.

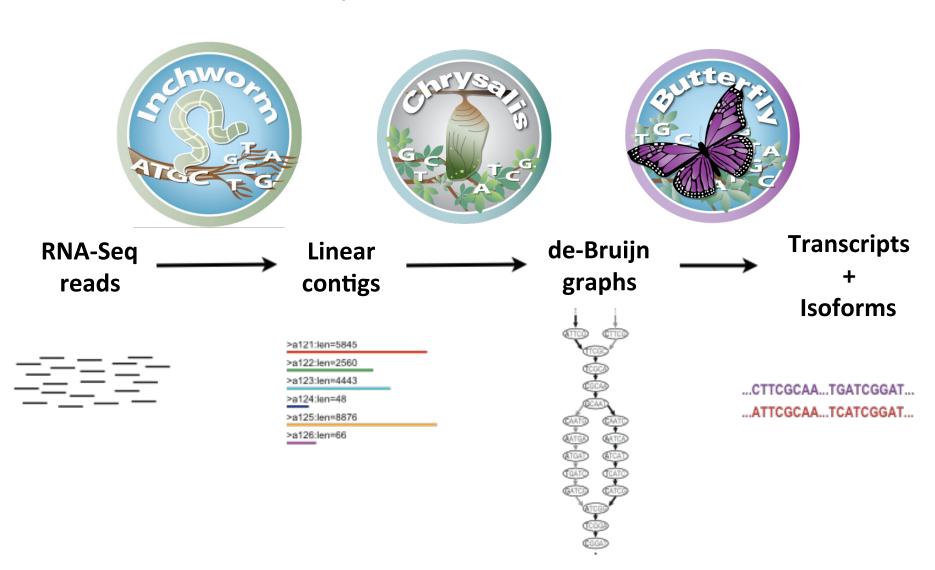
Trinity Transcriptome Assembly

Many Thousands of Small Graphs



Ideally, one graph per expressed gene.

Trinity – How it works:



Thousands of disjoint graphs



Decompose all reads into overlapping Kmers => hashtable(kmer, count)

Read: AATGTGAAAACTGGATTACATGCTGGTATGTC...

AATGTGA

ATGTGAA Overlapping kmers of length (k)

TGTGAAA

• • •

Kmer Catalog (hashtable)

Kmer	Count among all reads
AATGTGA	4
ATGTGAA	2
TGTGAAA	1
GATTACA	9



- Decompose all reads into overlapping Kmers => hashtable(kmer, count)
- Identify seed kmer as most abundant Kmer, ignoring low-complexity kmers.

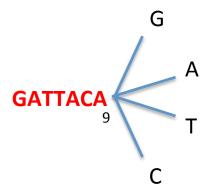
GATTACA

Kmer Catalog (hashtable)

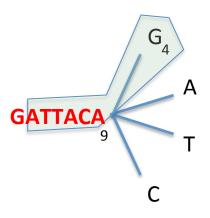
Kmer	Count among all reads
AATGTGA	4
ATGTGAA	2
TGTGAAA	1
GATTACA	9



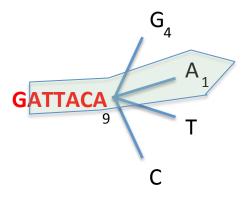
- Decompose all reads into overlapping Kmers => hashtable(kmer, count)
- Identify seed kmer as most abundant Kmer, ignoring low-complexity kmers.
- Extend kmer at 3' end, guided by coverage.



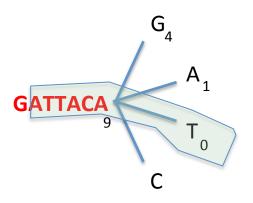




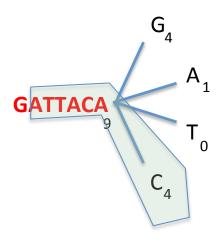




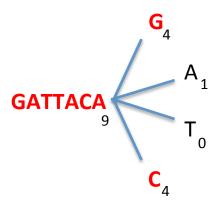




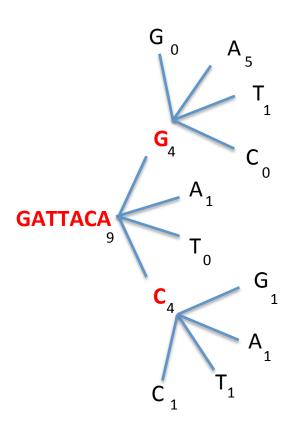




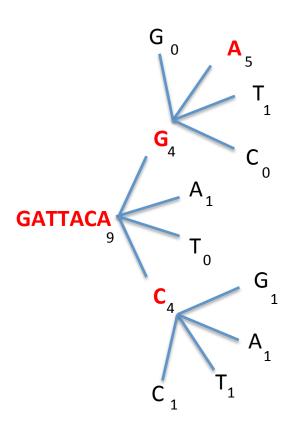




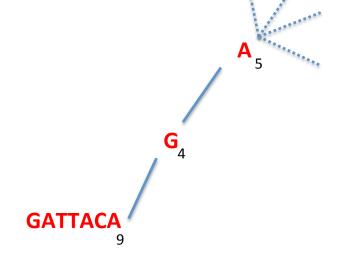




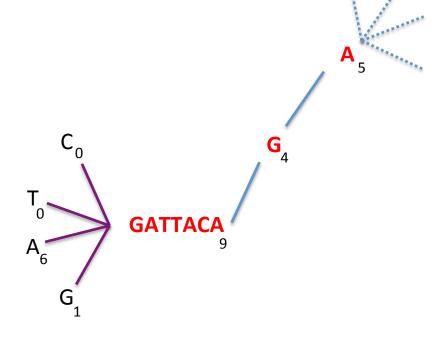




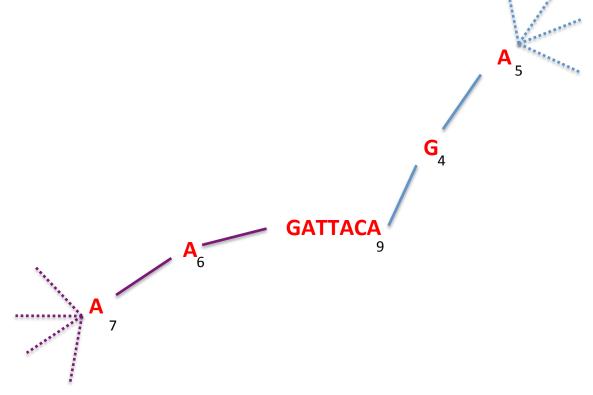












Report contig:AAGATTACAGA....

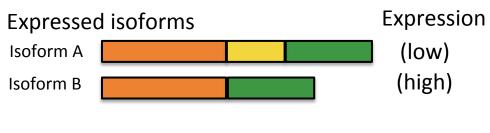
Remove assembled kmers from catalog, then repeat the entire process.

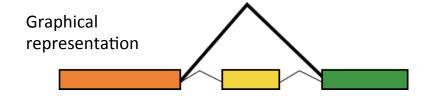


Expressed isoforms

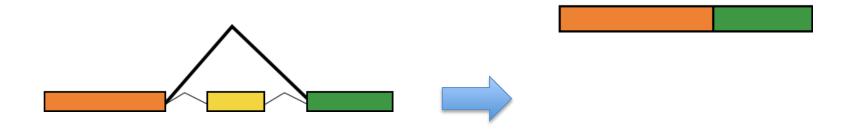




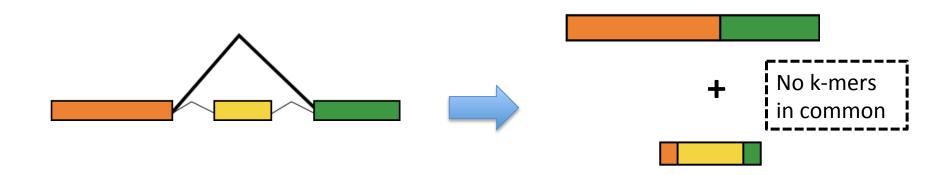




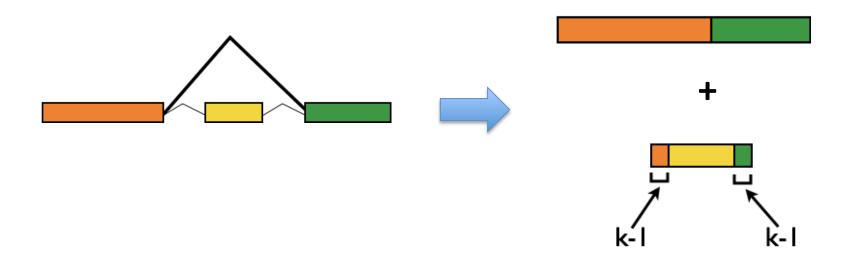




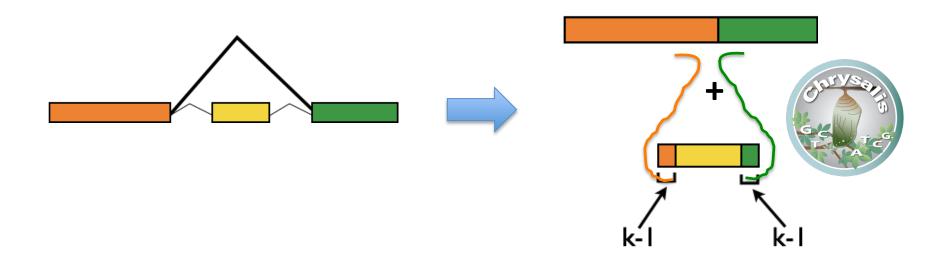








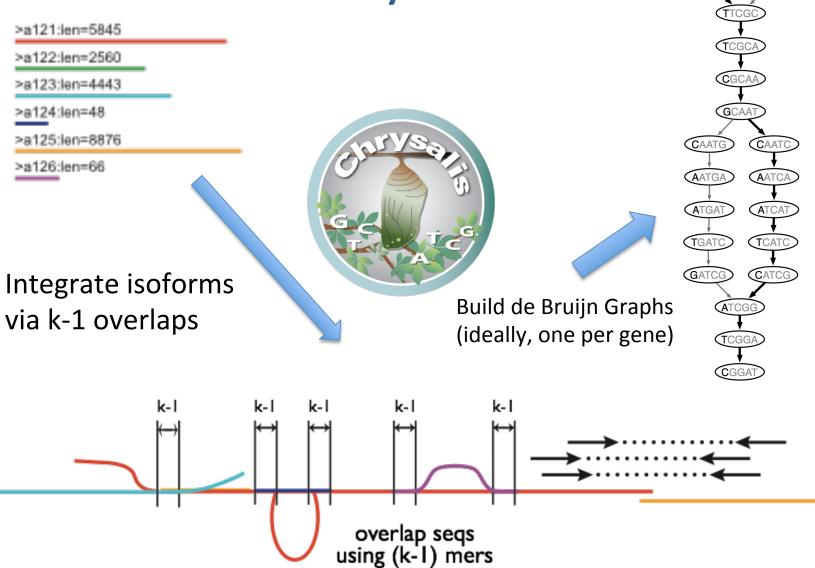
Chrysalis Re-groups Related Inchworm Contigs

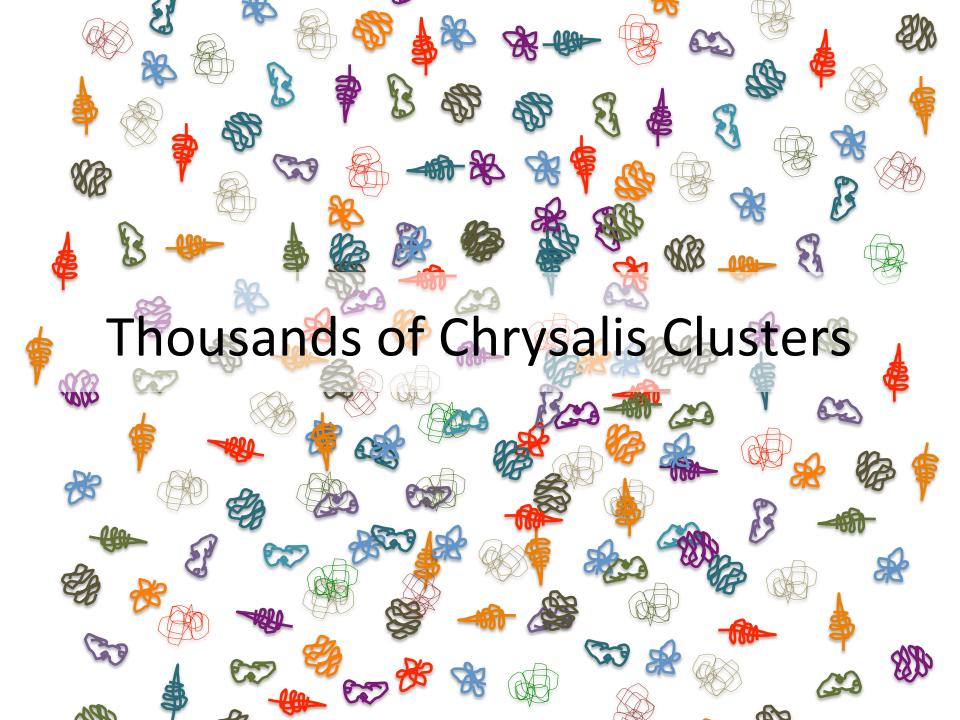


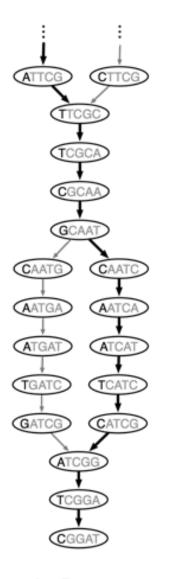
Chrysalis uses (k-1) overlaps and read support to link related Inchworm contigs



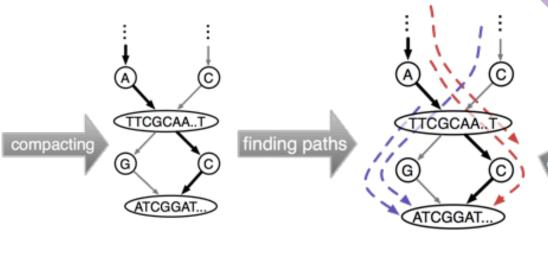
CTTC







Butterfly



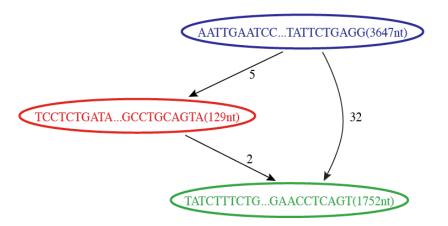
..CTTCGCAA..TGATCGGAT...
..ATTCGCAA..TCATCGGAT...

de Bruijn graph compact graph compact graph with reads

sequences (isoforms and paralogs)

Butterfly Example 1: Reconstruction of Alternatively Spliced Transcripts

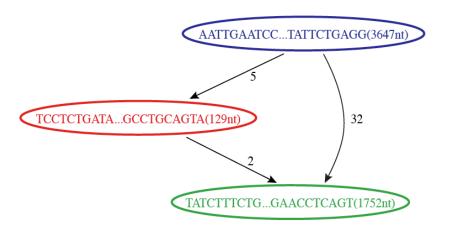
Butterfly's Compacted Sequence Graph





Reconstruction of Alternatively Spliced Transcripts

Butterfly's Compacted Sequence Graph

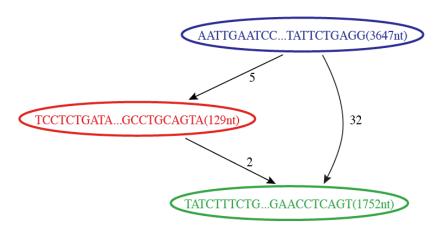


Reconstructed Transcripts



Reconstruction of Alternatively Spliced Transcripts

Butterfly's Compacted Sequence Graph



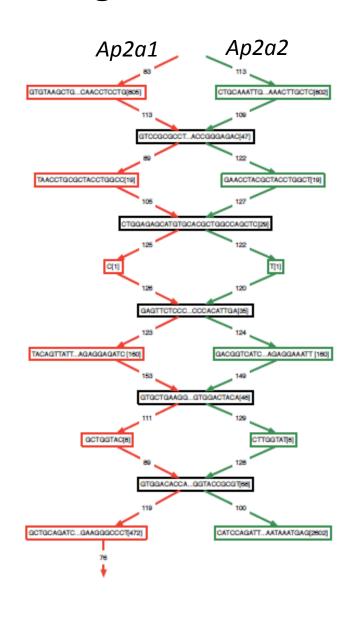
Reconstructed Transcripts



Reconstruction of Alternatively Spliced Transcripts

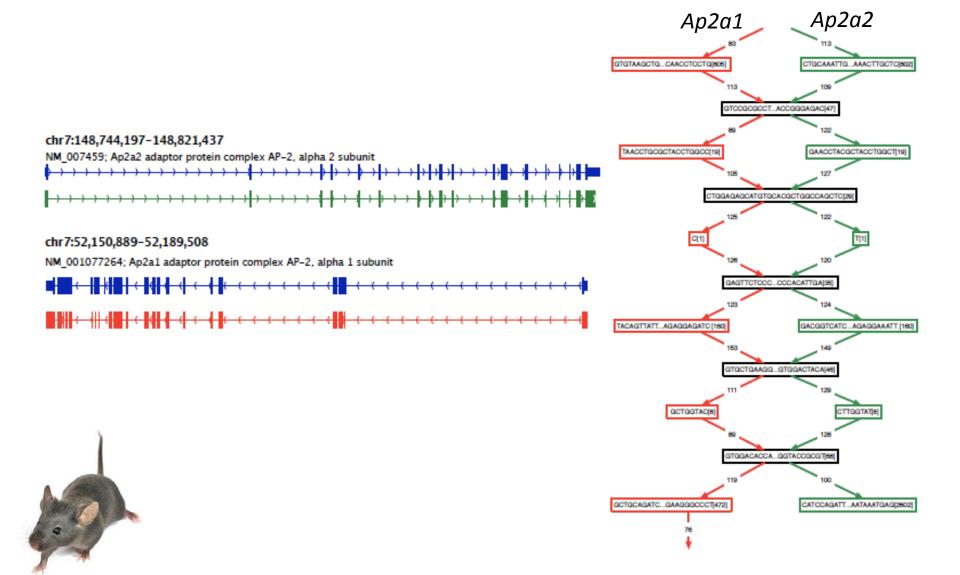
AATTGAATCC...TATTCTGAGG(3647nt) **Butterfly's Compacted** TCCTCTGATA...GCCTGCAGTA(129nt) 32 Sequence Graph TATCTTTCTG...GAACCTCAGT(1752nt) Reconstructed Transcripts Aligned to Mouse Genome Naa25 Nalpha acteyltransferase 25 (Reference structure)

Butterfly Example 2: Teasing Apart Transcripts of Paralogous Genes





Teasing Apart Transcripts of Paralogous Genes



Strand-specific RNA-Seq is Preferred

Computationally: fewer confounding graph structures in de novo assembly:

ex. Forward != reverse complement

(GGAA != TTCC)

Biologically: separate sense vs. antisense transcription

NATURE METHODS | VOL.7 NO.9 | SEPTEMBER 2010 |



Comprehensive comparative analysis of strand-specific RNA sequencing methods

Joshua Z Levin^{1,6}, Moran Yassour^{1-3,6}, Xian Adiconis¹, Chad Nusbaum¹, Dawn Anne Thompson¹, Nir Friedman^{3,4}, Andreas Gnirke¹ & Aviv Regev^{1,2,5}

Strand-specific, massively parallel cDNA sequencing (RNA-seq) is a powerful tool for transcript discovery, genome annotation

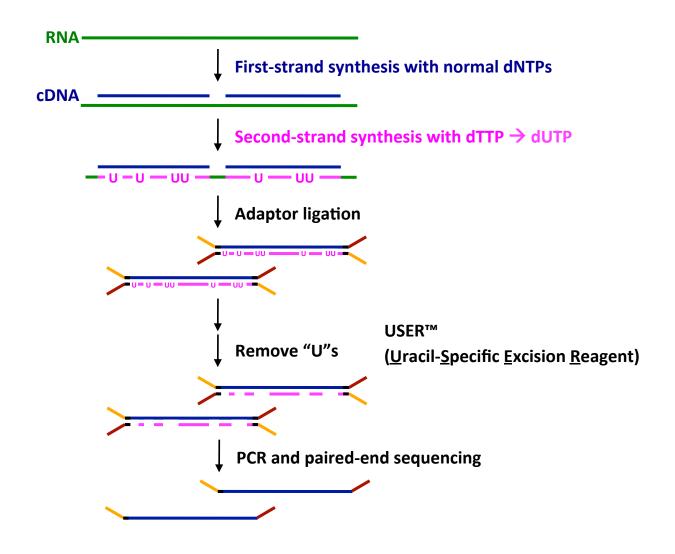
Nevertheless, direct information on the originating strand can substantially enhance the value of an RNA-seq experiment. For

'dUTP second strand marking' identified as the leading protocol

computational pipeline to compare library quality metrics from any RNA-seq method. Using the well-annotated Saccharomyces cerevisiae transcriptome as a benchmark, we compared seven library-construction protocols, including both published and

transcribed strand of other noncoding Kivrs, demarcate the exact boundaries of adjacent genes transcribed on opposite strands and resolve the correct expression levels of coding or noncoding overlapping transcripts. These tasks are particularly challenging in small microbial genomes, prokaryotic and eukaryotic, in which

dUTP 2nd Strand Method: Our Favorite



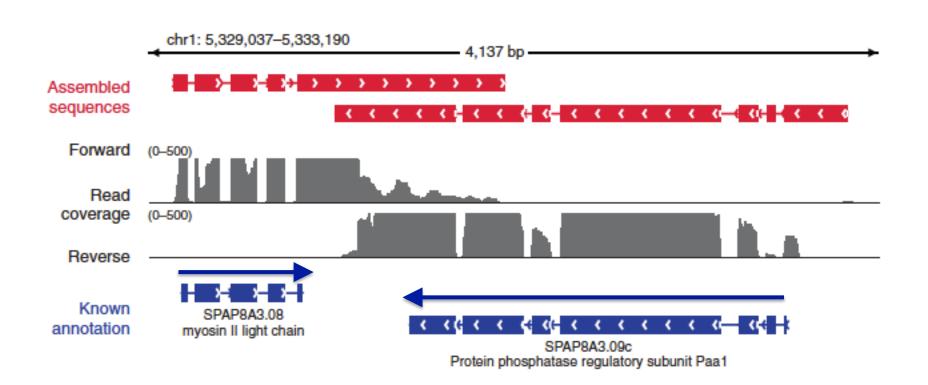
Modified from Parkhomchuk et al. (2009) Nucleic Acids Res. 37:e123

Slide courtesy of Joshua Levin, Broad Institute.

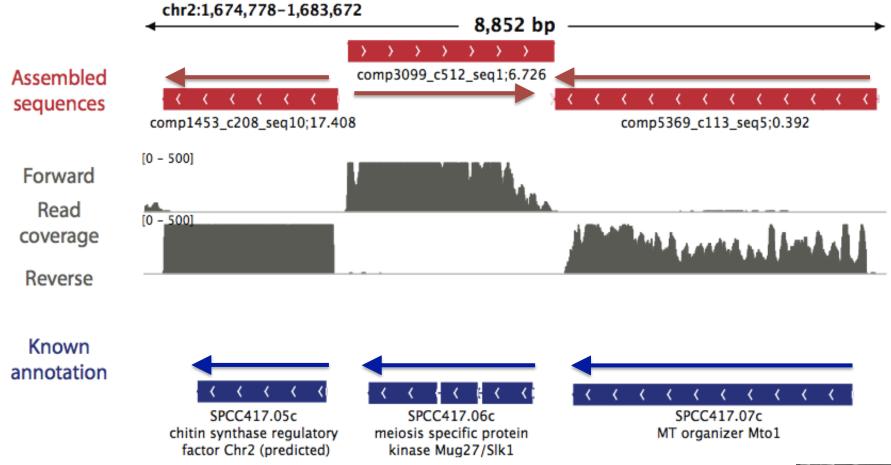
Overlapping UTRs from Opposite Strands



Schizosacharomyces pombe (fission yeast)



Antisense-dominated Transcription





Trinity output: A multi-fasta file

>comp0 c0 seq1 len=5528 path=[1:0-3646 10775:3647-3775 3648:3776-5527]

TECACTCTCCATCATGTGGGAGATACTACAGAGAGGACTATCCGTCCACAGCAAGAAGTAACTTGAACCCGATTCCTGCTTTTGCAAAGTCTTGACTTAGGATCTCAGTAGTAGTAGTAGTAGTAAAAAGCAGGGATTCTTTTTTTCAGTCT ACCTATCTCAAAATGTAAGAATGTAGAATGTAGAATGTACATTTAGTAAGAAATCAGGAAGTAACAGGGAGGTGAACCCCACCATGACATGTTATTTCTCAACAAGACCAGTGGAGGCCCTACATGTTAGAGAAGACCAGG TTTTTGAATCCCAGACGTTACGATAAGAATGCAATGCAGTGCTGCTGCGAGCAGTCCATGGGAAAGACCAGTCCTCACCAAGTCATCTTTTCACCTTACAGTTACTCTCAGGAATAAAGTGACAGGAACAAGAACAGA AGTGAGAGAGAGATTCAGACACAAAACAGTCAGGGAAAGCGCTGTCGGAGCTCGGCATGACATAATCAAGAGCAGTTTTCATCTTCTCGCAGACCAGCCTCTTAAGCTGGAGGCTTAGGGAACAGGCCACCAACCTTAG GTGAACAACATGAATACACCCTGATGCAGTTCTTAAGTGTCAACAGGACACCAACATCAGGCCCATTATAAAACATACCTTTTTCAACCTAAAAACTTAAAACCTAGGTTAAAACCCATATTTAATCCTTCCATCCCAACG ACAACCTCTTCCAACAGCTTTTCCACCTTTCCCAGAAGACCAGCGGGGCTCAGTCTCACGAGCTTTACGAAGCTGCTCAGGGCTGTGCTCAGAAGTTCTTCCTCAAAGTGGGCAGAACTTGGACTTCTCCTCACTGCCCTGAG GETGASCTCTCCCAGGAGCGGGGCTGGGGATGTCTTGGCTGCGGGCTGTGAAGCCAAAACTGAGCACCTGTGGTCATGTGGTGCCCTTGGGCACGTACTCAGCCACATCTCCCTTGAGCCTCTGGGGTCCAT CCCGGTTCTGTGGGATGGCACAAGGAACCTGCCACTGAGGGAAGGGTCCTGCTCTGGGTCTCCTACACCTGCTTTTGCCTCATACGGCTACAGTATGGGCTACCCAACCCAAACTCCCACTGCATGATGAAAAAGCAAC TEACAGTAACTGGACACCCAAAGGATGACAGAATAGTCTCAACGAAGAAGACCAGATTCTCTAGGACTGCAGGTCTTCACATTGCCATCTGTAACTTCTAAGAGGTCCCCTTTACATGTCTGAAGACACCTTT CARTAGITCTAGIAAGATGCAAGACTATTTCTATTCGTTCCTGTATATCCACTGTACCCCTCTAAGCCACTGGTGTCCAAGCTCATACATGTCGCTGACAGGAACAGATGGAAGGACTGGACTGGCACTGACACCAGACCAGAAC TTGCTTCAAGTAGAAGGTCTAACAGCATCCGCTCAGTGCGCACATTGCGCAAAATGGAGGAGTTTTCAGCCTGCTCCTAAAAGCCGATAAACCTCGGATCTTCCTAACAGCTGGAACAATGCACCACATTTCTAACAATGATCAT TOTAL OFFICE TOTAL TAXABLE CONTROL OF THE CONTROL O AGAGATECACCACTGGTTCAAATGCACCCAGCAGCAGCAGTAGATCGGAACAAGTAGGAACTTGAACTAGGGCTTGGGTTGAGGCTTTTTCCAGCAGGAGTCAGGGCCTGCCAGACCGCAGACTCCTCACCT GETTETETECATACATCAATGAGCACATGAACAGCGAGCAGCAGCAGTAATAGTETGAGAACTGCAACTCTGTTTCAAACAACAGGGGGGCCCAAACTCCAGTCGTGGTACCTTAACATCAGTTCTTTTGACCACATCCAG TACTTCCCCTTCTAAGGAGTGTTCCCCCTCAGCAGGCGGGTCCAGGCTTCCTCAATCAGTCGAAAGAGAGCAGAATAAGTCAGATAGAACTGCCAGTCATCTGAGTTTTTCAGAAGGGGGTCGGGGAAAGGGGGGT TOCACTOCOGCACTTGCTCACCTTGTTACATGGCCATGCATTTTTTTCCCCAACTTCTCACTTGTCACTTCTCCTCTAATTTTCCCCTAATGACATCCAGGGCTTCCTGGTACTTTCCCCAAGCGCTCCAGA GCTCATCACTGACCAAAAGTAGGGTTTTTGGGGGACAATCTTATACAGAGCCATGCCAGCCTGCATCTTCTTGTACTCGCCCACTCTGGCATAGGCCATGAAGAGGTGAGAGTGGTACTCCTCGCTGTTTGGGAA GAGGCCCCGGTCGTTAGGGTCCTGCACATGGCCCCCGCGTCGCCATGATGACAAGCGCAGAACCTCAGT

>comp0 c0 seq2 len=5399 path=[1:0-3646 3648:3647-5398]

AATTGAATCCCTTTTGTATGGAAAAGCTGAAAGGCATGAAGGCATATACAGATTGGATGGGGGGTGATTGCAAAATATAATGCAAATTTGGAACAATTAGAAAATATACAAAATTGATGGCAGCACCCTAGGTTTG TECACTCTCCATCATGTGGGGGTACTACAGAGGGCTATCCGTCCACCAGCAAGAAGTAACTTGACCTGACTTTTCTACCAGTGTTTTTCTCAGTCT AGTTATCTCAAAATGTAAGAATTAGATCTGATTGAAATGCTACATTTAGTAAGAAATCAGCAAGTAACAGGGAAGTGTAACCCCACCATGACATTATTTGTCAACAAGACCAGTGGAGGCCCTACATGTTAGAGCAGG TTTGTGAATCCCAGACAGTTACGATAAAGAATGCAATGGTGTGCTGCTGCAGCAGTCCATGGGAAAGACCAGTCCTTCACCATGTTACAGCTTACAGTTACTCTCCAGGAATAAAGTGACAGGGAACAAGAACAAG ACTGAGAGAGAGTTCAGACACAAAACAGTCAGGGAAAGCGCTGTCGGAGCTCGGCATGACATAATCAAGAGGCAGTTTTCATCTTCTCGCAGACCAGCCTCTTAAGCTGGAGGCTTAGGGAACAGGCCACCAACCTTAG GTGRACARCATGRATACACCCTGRTGCRGTTCTTAAGTGTCRACAGGGCCCAACATCAGGCCCATTATRARACCTTTTTCAACCTARAACCTTRARACCTAGGTTAAAACCCATTTTAATCCTTCCATCCCARAG ACAACCTCTTCCAACAGCTTTCCCACGTTCCCAGAAGACCAGCGGGGCTCAGTCTCAGCAGTTTACGAAGCTCACGGTGGGCTGTGCTCAGATTTCTTCCTCAAAGTGGGCAGAACTTGGACTTCCTCACTGCCCTGAG CCCGGTTCTGTGGGGTGGCACAAGGAACCTGCCACTGAGGGAAGGGTCCTGCTCCTACACCTGCTTTTGCCTCATACGGCTACAGTATGGGCTACCCAAACCCAAACTCCCACTGCATGATGAAAAAGCAAC TEACASTARTSGACACCCARAGGATGACAGARATAGTCTCRACGARGAGACCCGGATCTCTCTAGGACTGCGGTCTCACATTGCCATCTGTAACTTCTAAGACGCTCCCTTTACATGTGCGAACACCTTT CANTAGTTCTAGTAAAGATGCAAGACTATTTCTATTCGTTCCTGTATATCCACTGTACCCTCTAAGCCACTGGTGCTAAGCTCATACATGTCCCTGACAGGACCAGATGGAAGCACTAGACAGCACTAGCACAGAAC TTGCTTCAAGTAGAAGGTCTAACAGCATCCGCTCAGTGCGCAAAATGCAGAGAGAATTATTCAGCCTGTTCCTAAAAGCGATAACTCTCGCAAATGCACCATATTTGTAAGCTTGAAATATTATT TETRAGGTATETTETGGTTGGAGTGGAAAAACCTGAGTGCAAAGTTACAGGACTGGGAGGCAGCAGCATACTGACCCAGGGATGCAGCATACCGGGTCAGAAGATAGCCAATGGTGTCGTGCTTGGATGTGCCTTAGCATC AACAACTCCAAGGAGCTGGTTAATGAATTGTGTGCACTGGGCGGCAAGGGCCAAGGGCCAACGACCTCACGTCTTGTGAAGCACCTTATCCCCGAACTTCTTGAAATACTGGAACATTAACTCTTCAGGGTCAC GENGAGCAGTGTACTTCCCCTTCTAAGGAGTGTTCCCCCTCAGCAGGCGGCGTCCAGGCTTCCTCAATCAGTCGAAAGACAGAATCGAAATAGGTCAGAATAGAACTGCCAGTCATCTGAGTTTTTCAGAAGGAGGGGGCGTCG GGANAGGCCCTTGCACCTCGCCCACCTCCCACCTCCCACCTCCCCACCTCCCCACCTCCCCACCTCCCCAACTTCCCAACTTCCCCAACTTCCCAACTTCCCAACTTCCCAACTTCCCAACTTCCCAACTTCCAACTTCCCAACTTCAACTTCCAACTTCCAACTTCAACTTCAACTTCCAACTTCAACTTCAACTTCCAACTTCCAACTTCA AGGGCTCCAGAATCATGTAATAAAGTTCAACCTCAGCCTCTAGCCTGTGTCCTCCCCCTCACCATCTTCTCCGCCAGGGGCAGAAACATGGTTTTGGAGAGGTTCTCATCCGGCGGATATAGAT TOCATGATCAGCTCATCACTGACCAAAAGTAGTAGTAGGGGTTTTTGGGGACAATCTTATACAGAGCCATGCCGCCTGCTGCTTCTTGTACTCCCCCCACTCTGGCATAGGCCATGAAGAGGTGAGAGTGGTACTCCTC TGAAAGCTTCTTCCTGCTGCCAGTTCTCTGTAAACCAATGGCCTTGAGAACCTTTGCACAGGGAGACCTTTGTGTTCTTCAACAGTTTATCAGCTTGATATTGCCATTATTATTATTATTATCAGCAAGATAATGG TARATEGGCCGGAGGCGCCGGTCGTTAGGCTCCTGCACATGGCCCCGGGGTCGCCATGATGACAAGCGCAGAACCTCAGT

Trinity assembly practical