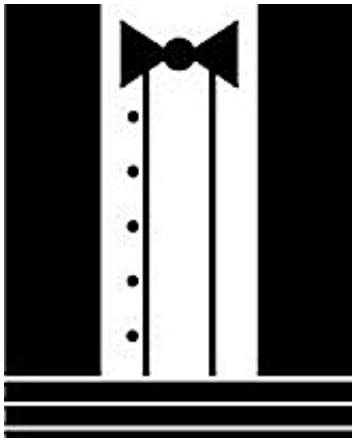


Transcript Reconstruction from RNA-Seq

May 2013

Brian Haas

Broad Institute

Transcript Reconstruction from RNA-Seq Reads

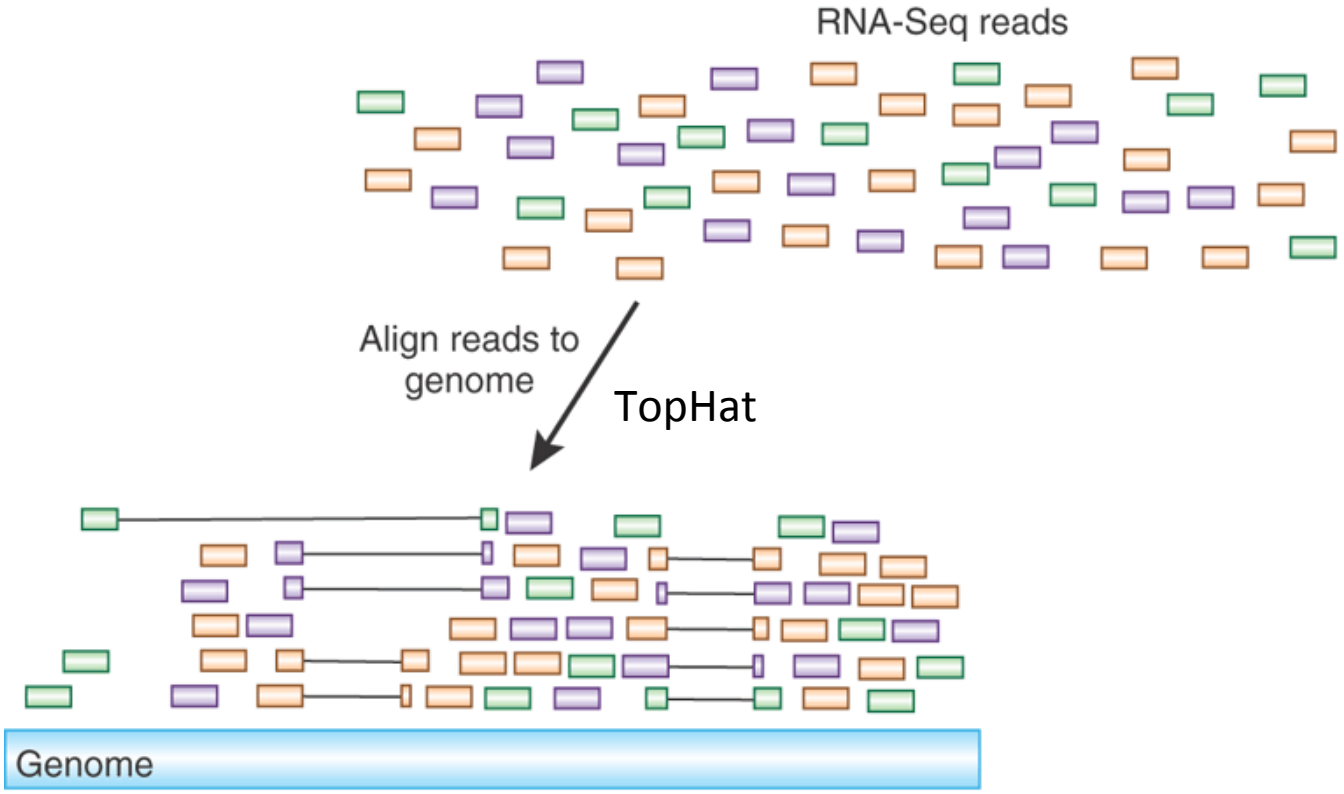
Advancing RNA-Seq analysis

Brian J Haas & Michael C Zody

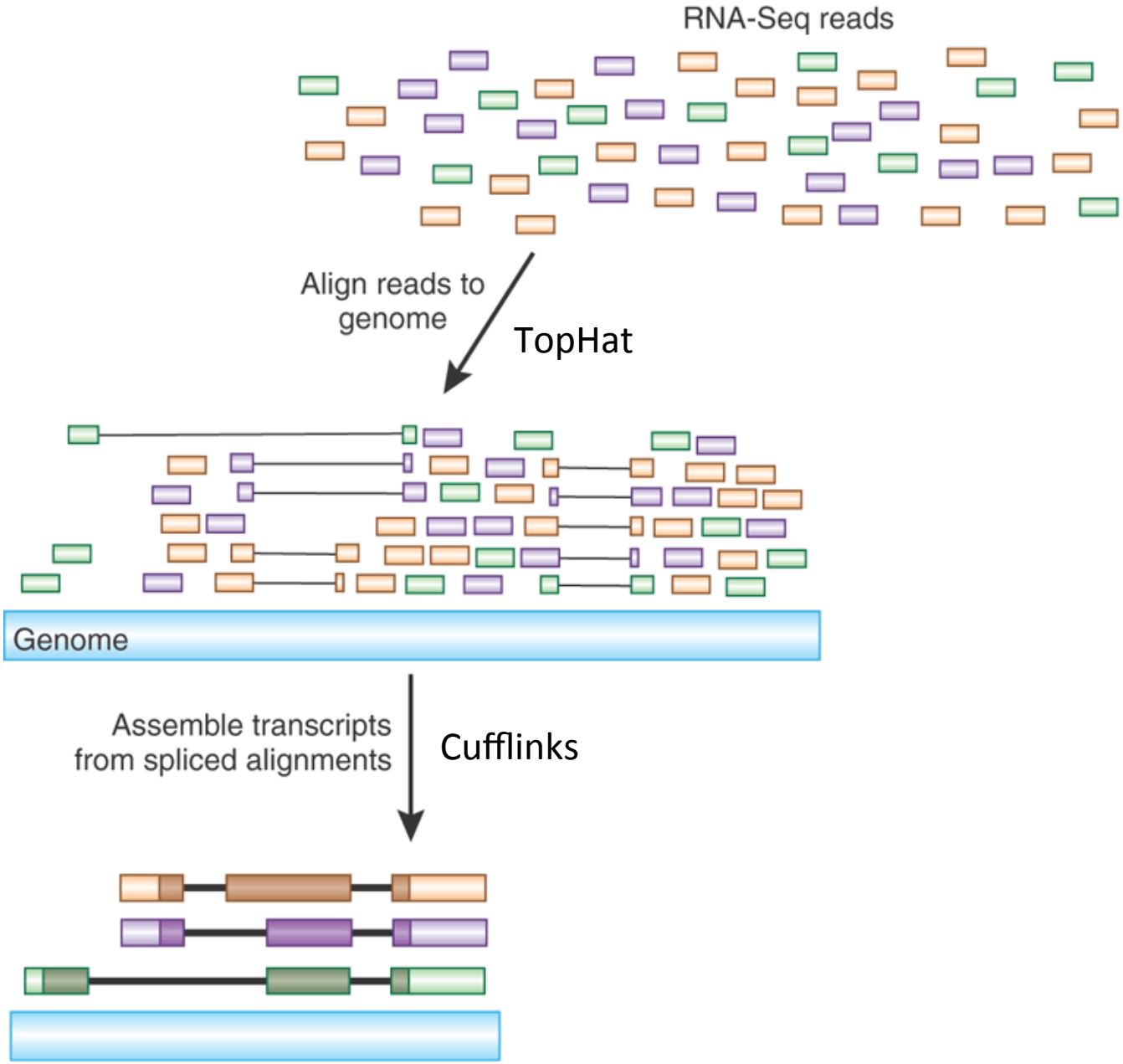
Nature Biotech, 2010

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

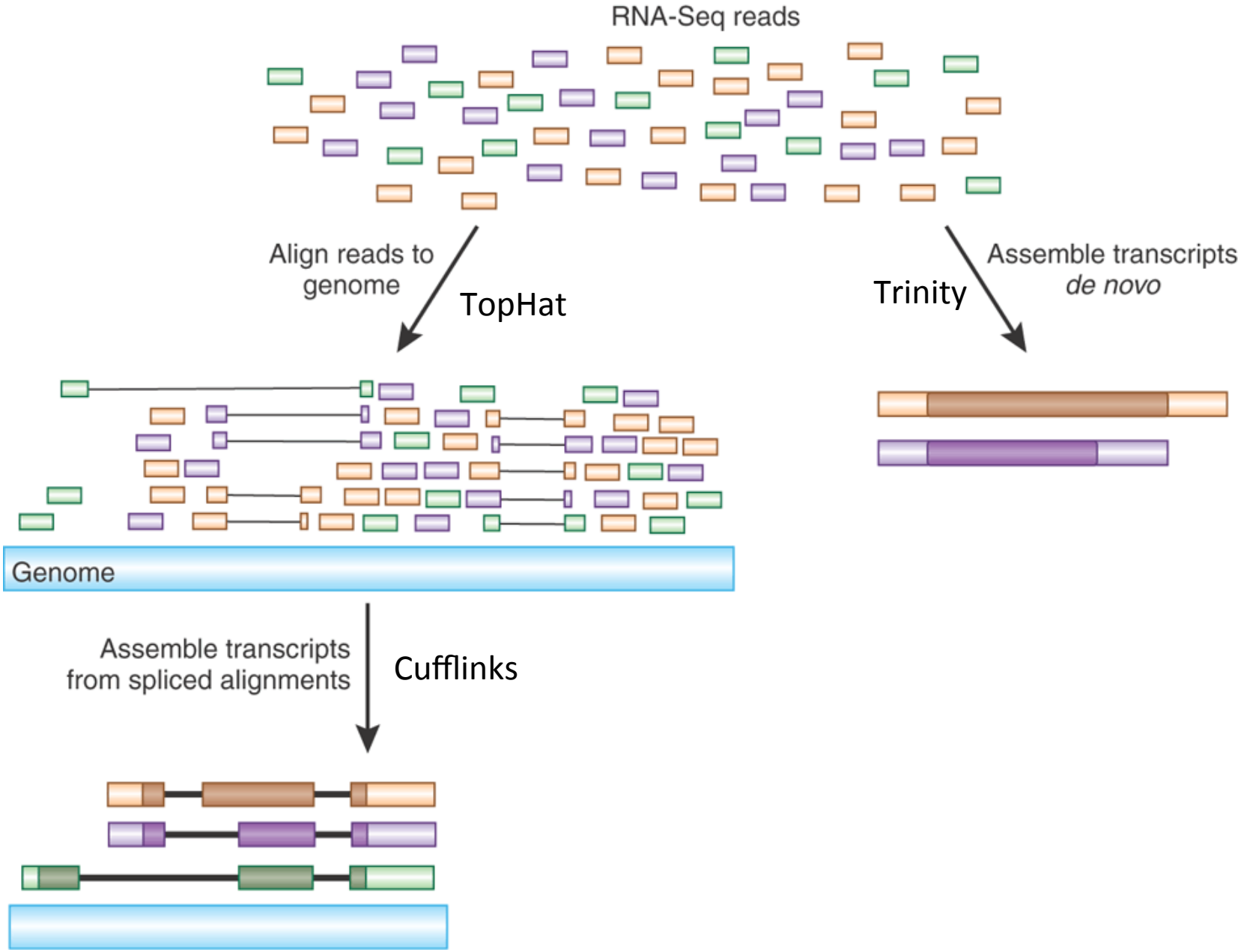
Transcript Reconstruction from RNA-Seq Reads



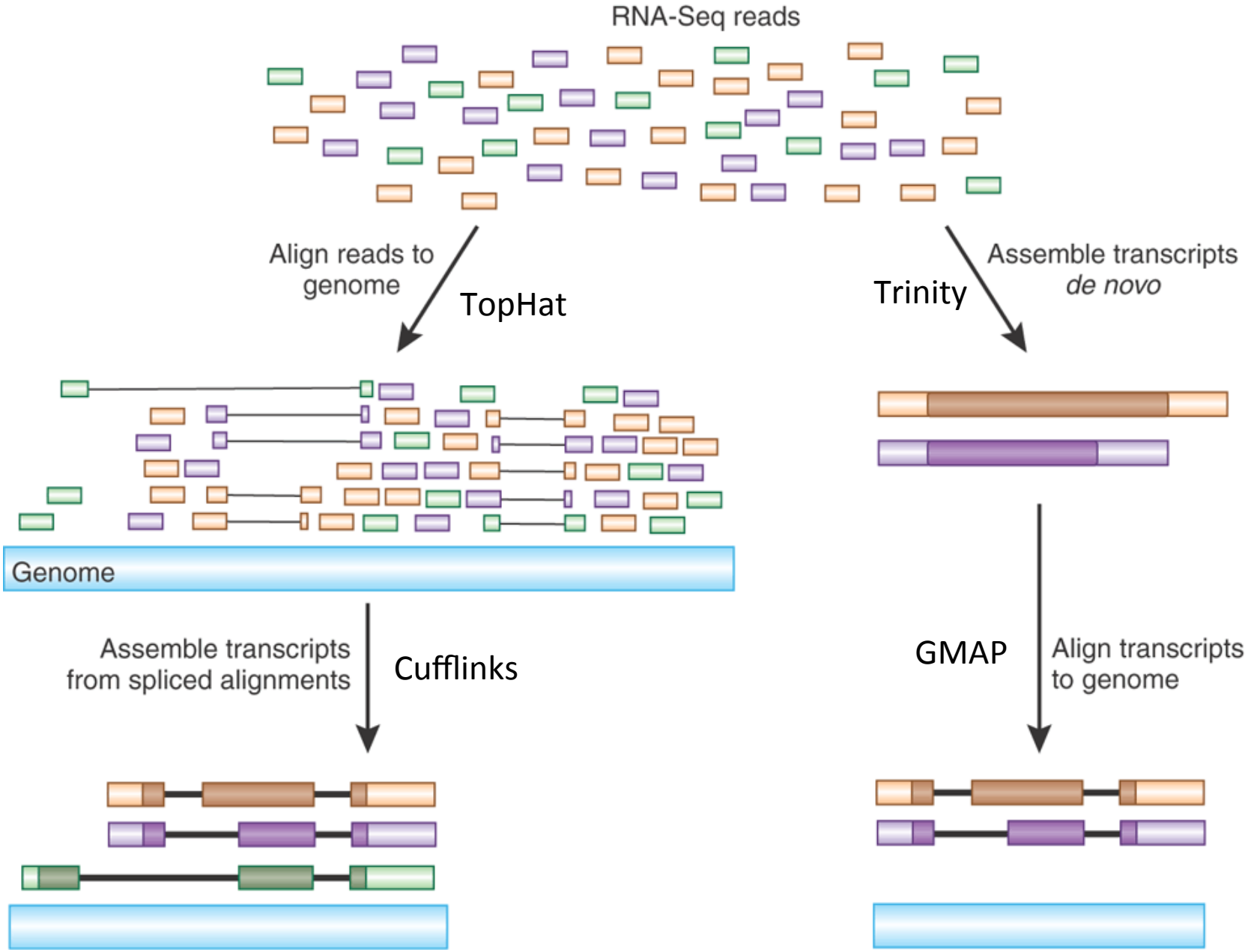
Transcript Reconstruction from RNA-Seq Reads



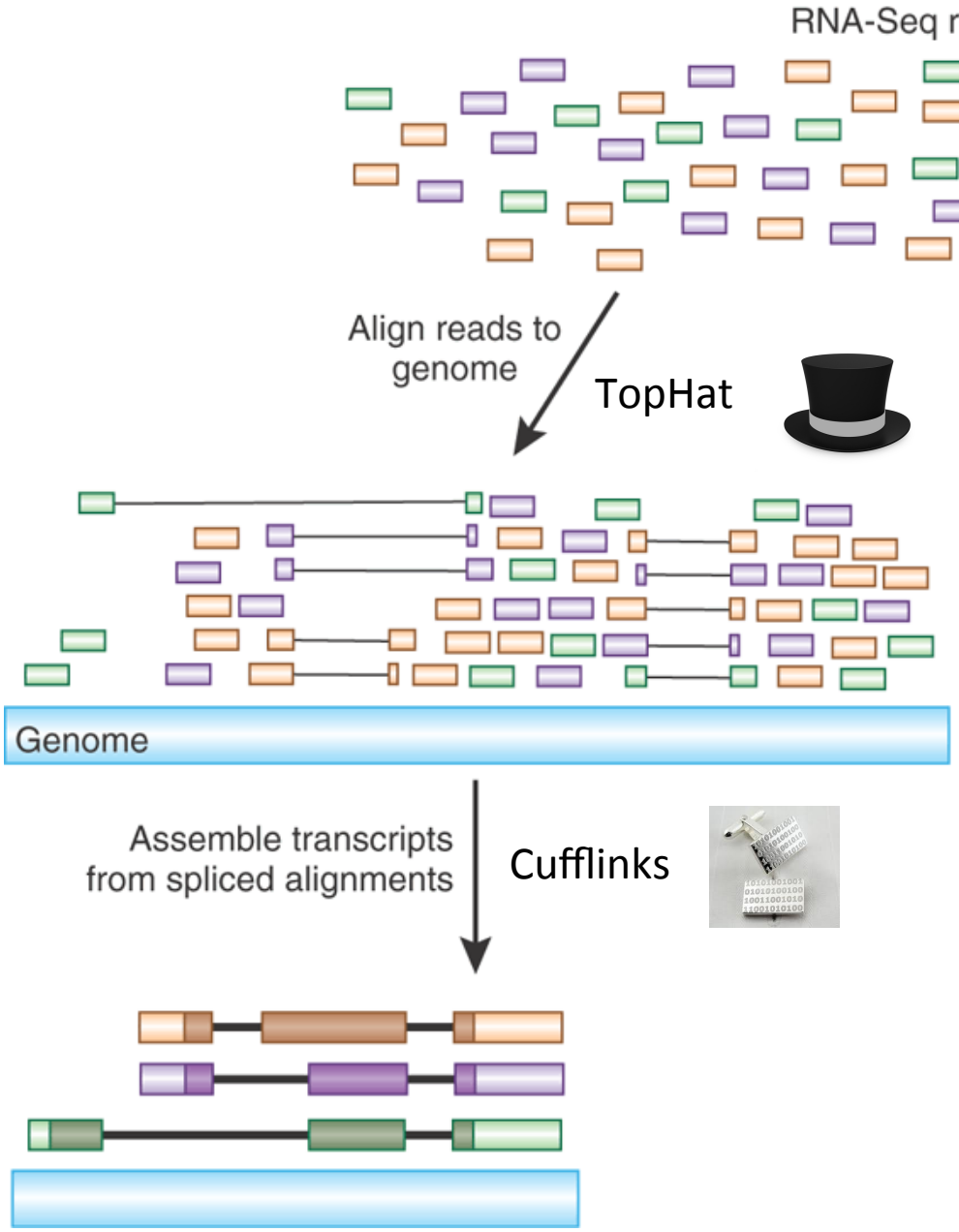
Transcript Reconstruction from RNA-Seq Reads



Transcript Reconstruction from RNA-Seq Reads



Transcript Reconstruction from RNA-Seq Reads



The Tuxedo Suite:

End-to-end Genome-based RNA-Seq Analysis Software Package

NATURE PROTOCOLS | PROTOCOL

Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks

Cole Trapnell, Adam Roberts, Loyal Goff, Geo Pertea, Daehwan Kim, David R Kelley, Harold Pimentel, Steven L Salzberg, John L Rinn & Lior Pachter

Affiliations | Contributions | Corresponding author

Nature Protocols **7**, 562–578 (2012) | doi:10.1038/nprot.2012.016
Published online 01 March 2012

Overview of the Tuxedo Software Suite

Bowtie (fast short-read alignment)

TopHat (spliced short-read alignment)

Cufflinks (transcript reconstruction from alignments)

Cuffdiff (differential expression analysis)

CummeRbund (visualization & analysis)

Slide courtesy of Cole Trapnell

Tuxedo development team


Daehwan Kim
Johns Hopkins


Ben Langmead
Johns Hopkins


Ali Mortazavi
Caltech


Barbara Wold
Caltech


Geo Pertea
Johns Hopkins


Steven Salzberg
Johns Hopkins


Lior Pachter
UC Berkeley


Cole Trapnell
Harvard


Brian Williams
Caltech


Mihai Pop
U. Maryland


Jeltje van Baren
MBARI


Loyal Goff
Harvard

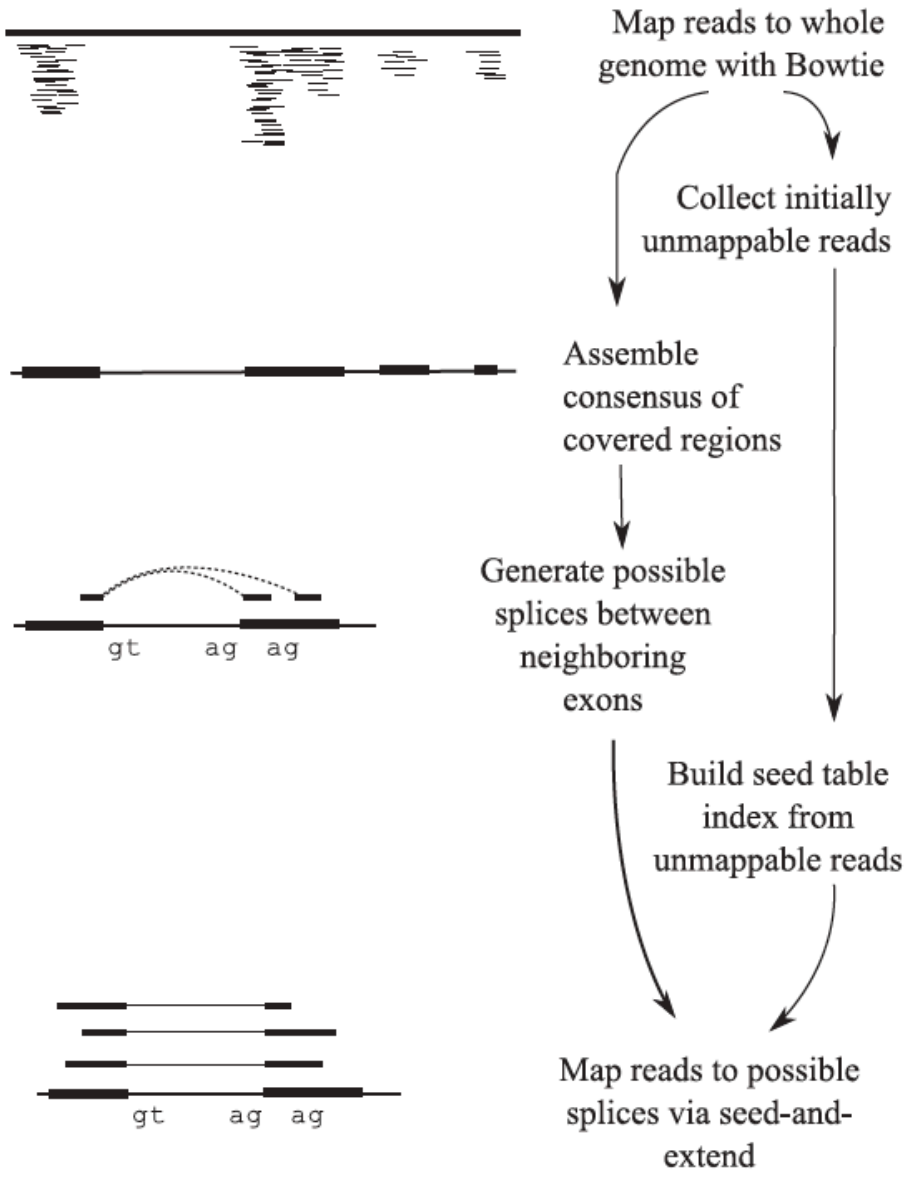

Adam Roberts
UC Berkeley


John Rinn
Harvard


Dave
Hendrickson
Harvard


Manolis Kellis
MIT

The TopHat Pipeline



From Trapnell, Pachter, & Salzberg. Bioinformatics. 2009

Alignments are reported in a compact representation: SAM format

```
0       61G9EAAXX100520:5:100:10095:16477
1       83
2       chr1
3       51986
4       38
5       46M
6       =
7       51789
8       -264
9       CCCAAACAAGCCGAACTAGCTGATTTGGCTCGTAAAGACCCGGAAA
10      ###CB?=ADDBCBCDEEFFDEFFFDEFFGDBEFGEDGCFGFGGGGG
11      MD:Z:67
12      NH:i:1
13      HI:i:1
14      NM:i:0
15      SM:i:38
16      XQ:i:40
17      X2:i:0
```

SAM format specification: http://samtools.sourceforge.net/SAM1.pdf

Alignments are reported in a compact representation: SAM format

```
0     61G9EAAXX100520:5:100:10095:16477   (read name)
1     83   (FLAGS stored as bit fields; 83 = 00001010011 )
2     chr1   (alignment target)
3     51986  (position alignment starts)
4     38
5     46M  (Compact description of the alignment in CIGAR format)
6     =
7     51789
8     -264
9     CCCAAACAAGCCGAACTAGCTGATTTGGCTCGTAAAGACCCGGAAA   (read sequence, oriented according to the forward alignment)
10    ###CB?=ADDBCBCDEEFFDEFFFDEFFGDBEFGEDGCFGFGGGGG   (base quality values)
11    MD:Z:67
12    NH:i:1
13    HI:i:1
14    NM:i:0
15    SM:i:38                                          (Metadata)
16    XQ:i:40
17    X2:i:0
```

SAM format specification: http://samtools.sourceforge.net/SAM1.pdf

Alignments are reported in a compact representation: SAM format

```
0       61G9EAAXX100520:5:100:10095:16477  (read name)
1       83  (FLAGS stored as bit fields; 83 = 00001010011 )
2       chr1  (alignment target)
```

Still not compact enough…
Millions to billions of reads takes up a lot of space!!

Convert SAM to binary – BAM format.

```
15      SM:i:38
16      XQ:i:40
17      X2:i:0
```

SAM format specification: http://samtools.sourceforge.net/SAM1.pdf

Samtools

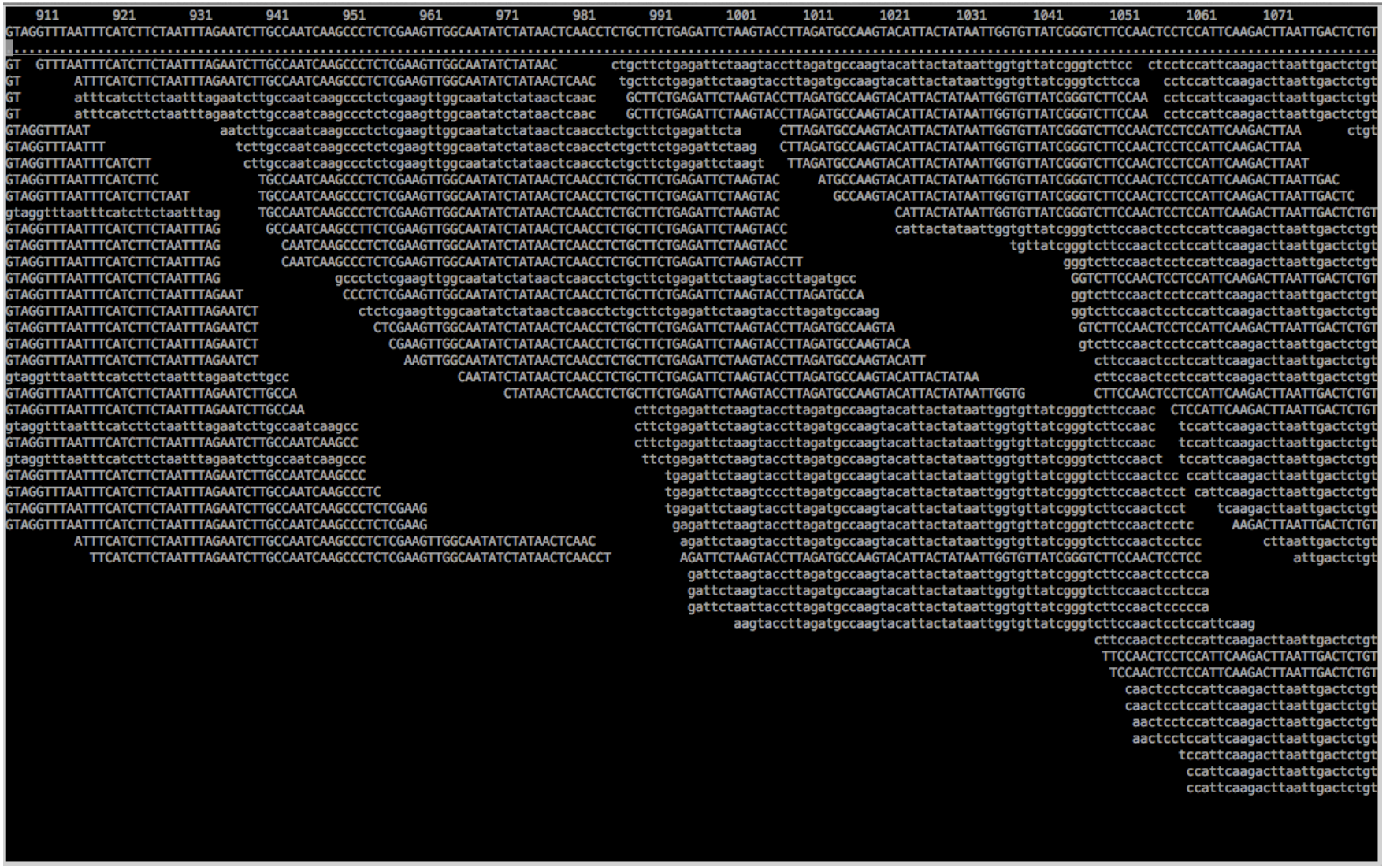
- Tools for
  - converting SAM <-> BAM
  - Viewing BAM files (eg. samtools view file.bam | less)
  - Sorting BAM files, and lots more:

```
Program: samtools (Tools for alignments in the SAM format)
Version: 0.1.18 (r982:295)

Usage:   samtools <command> [options]

Command: view        SAM<->BAM conversion
         sort        sort alignment file
         mpileup     multi-way pileup
         depth       compute the depth
         faidx       index/extract FASTA
         tview       text alignment viewer
         index       index alignment
         idxstats    BAM index stats (r595 or later)
         fixmate     fix mate information
         flagstat    simple stats
         calmd       recalculate MD/NM tags and '=' bases
         merge       merge sorted alignments
         rmdup       remove PCR duplicates
         reheader    replace BAM header
         cat         concatenate BAMs
         targetcut   cut fosmid regions (for fosmid pool only)
         phase       phase heterozygotes
```

Visualizing Alignments of RNA-Seq reads

Text-based Alignment Viewer

% samtools tview alignments.bam target.fasta



IGV

www.broadinstitute.org/igv/

Integrative Genomics Viewer

- Home
- Downloads
- Documents
  - Hosted Genomes
  - FAQ
  - IGV User Guide
  - File Formats
  - Release Notes
  - Credits
- Contact

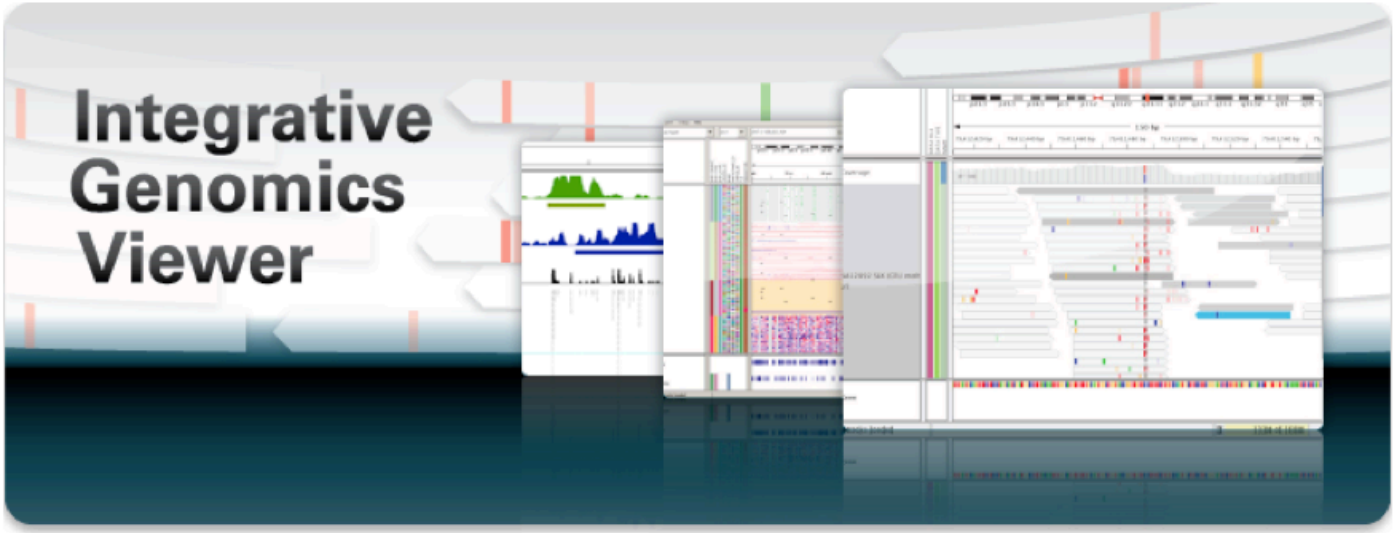
Search website

search

Broad Home
Cancer Program

BROAD INSTITUTE

Home



What's New

July 3, 2012. Soybean (Glycine max) and Rat (rn5) genomes have been updated.

April 20, 2012. IGV 2.1 has been released. See the release notes for more details.

April 19, 2012. See our new IGV paper in Briefings in Bioinformatics.

Overview

Citing IGV

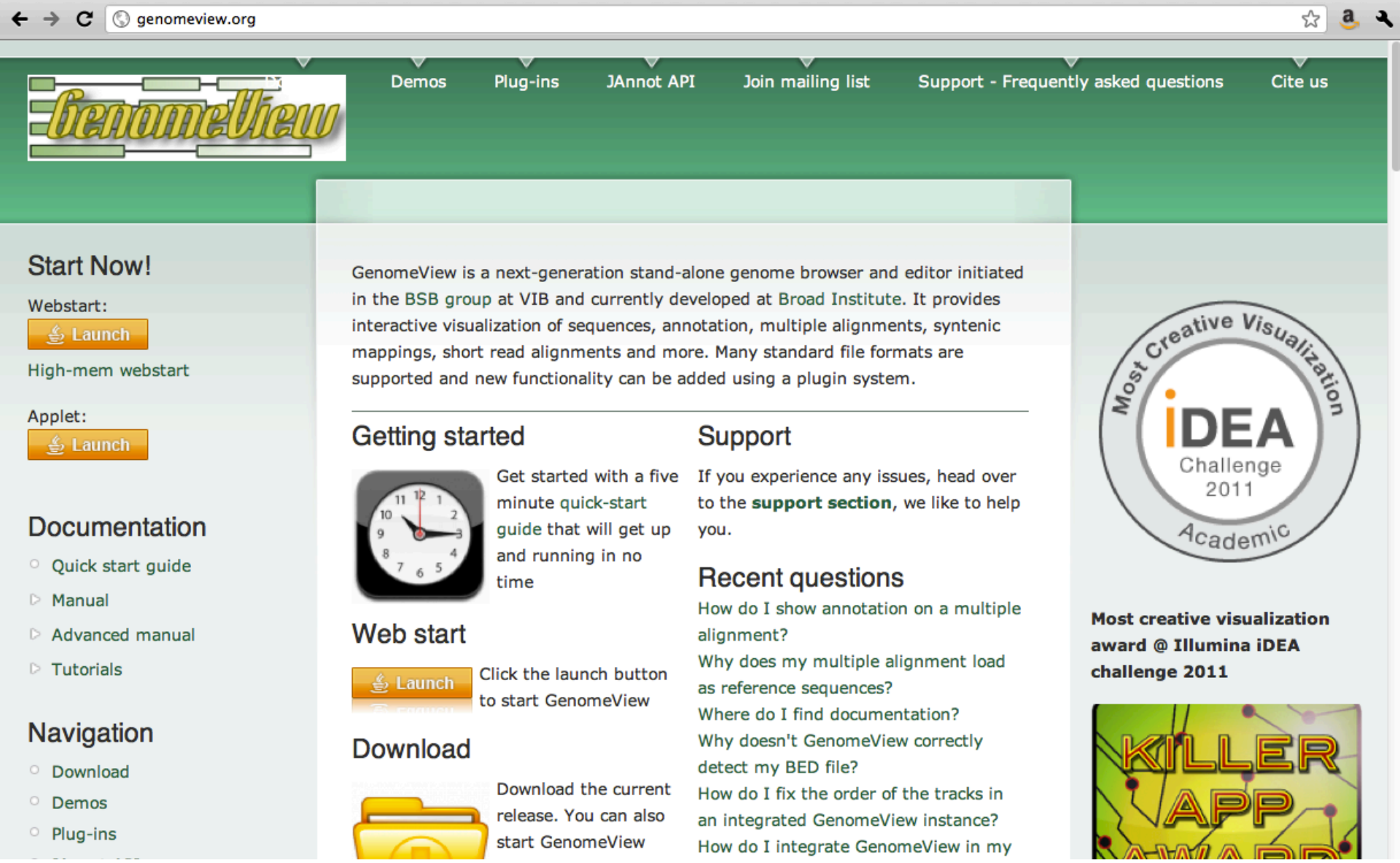
To cite your use of IGV in your publication:

James T. Robinson, Helga Thorvaldsdóttir, Wendy Winckler, Mitchell Guttman, Eric S. Lander, Gad Getz, Jill P. Mesirov. **Integrative Genomics Viewer**. Nature Biotechnology 29, 24–26 (2011), or

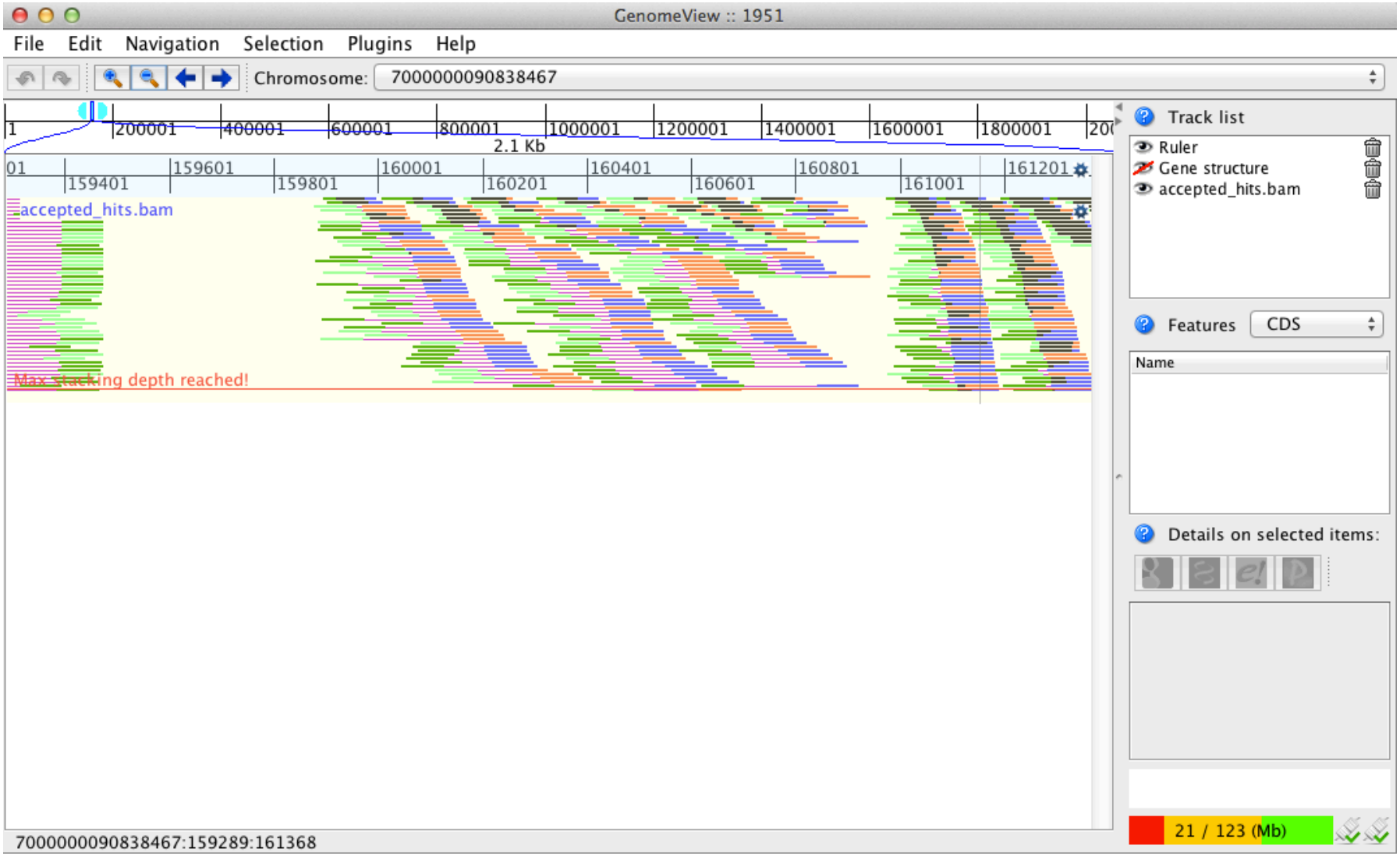
Helga Thorvaldsdottir, James T. Robinson, Jill P. Mesirov. **Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration.**

IGV: Viewing Tophat Alignments

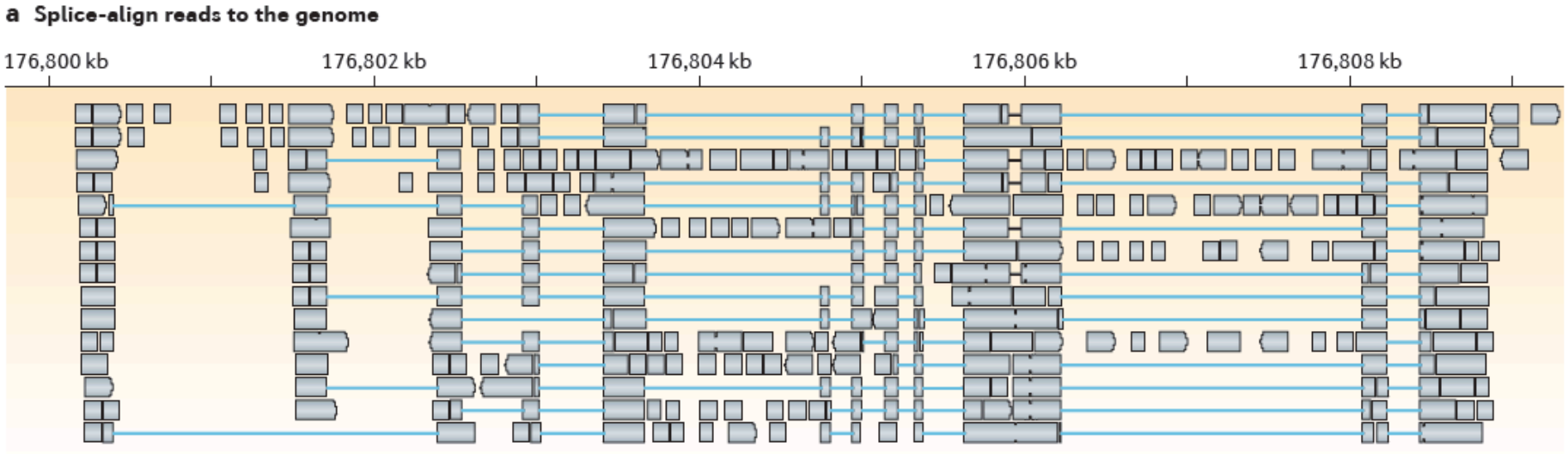
GenomeView


genomeview.org
GenomeView
Demos
Plug-ins
JAnnot API
Join mailing list
Support - Frequently asked questions
Cite us
Start Now!
Webstart:
Launch
High-mem webstart
Applet:
Launch
Documentation
Quick start guide
Manual
Advanced manual
Tutorials
Navigation
Download
Demos
Plug-ins
GenomeView is a next-generation stand-alone genome browser and editor initiated in the BSB group at VIB and currently developed at Broad Institute. It provides interactive visualization of sequences, annotation, multiple alignments, syntenic mappings, short read alignments and more. Many standard file formats are supported and new functionality can be added using a plugin system.
Getting started
Get started with a five minute quick-start guide that will get up and running in no time
Web start
Launch
Click the launch button to start GenomeView
Download
Download the current release. You can also start GenomeView
Support
If you experience any issues, head over to the support section, we like to help you.
Recent questions
How do I show annotation on a multiple alignment?
Why does my multiple alignment load as reference sequences?
Where do I find documentation?
Why doesn't GenomeView correctly detect my BED file?
How do I fix the order of the tracks in an integrated GenomeView instance?
How do I integrate GenomeView in my
Most Creative Visualization
iDEA
Challenge
2011
Academic
Most creative visualization award @ Illumina iDEA challenge 2011
KILLER
APP

GenomeView: viewing TopHat alignments

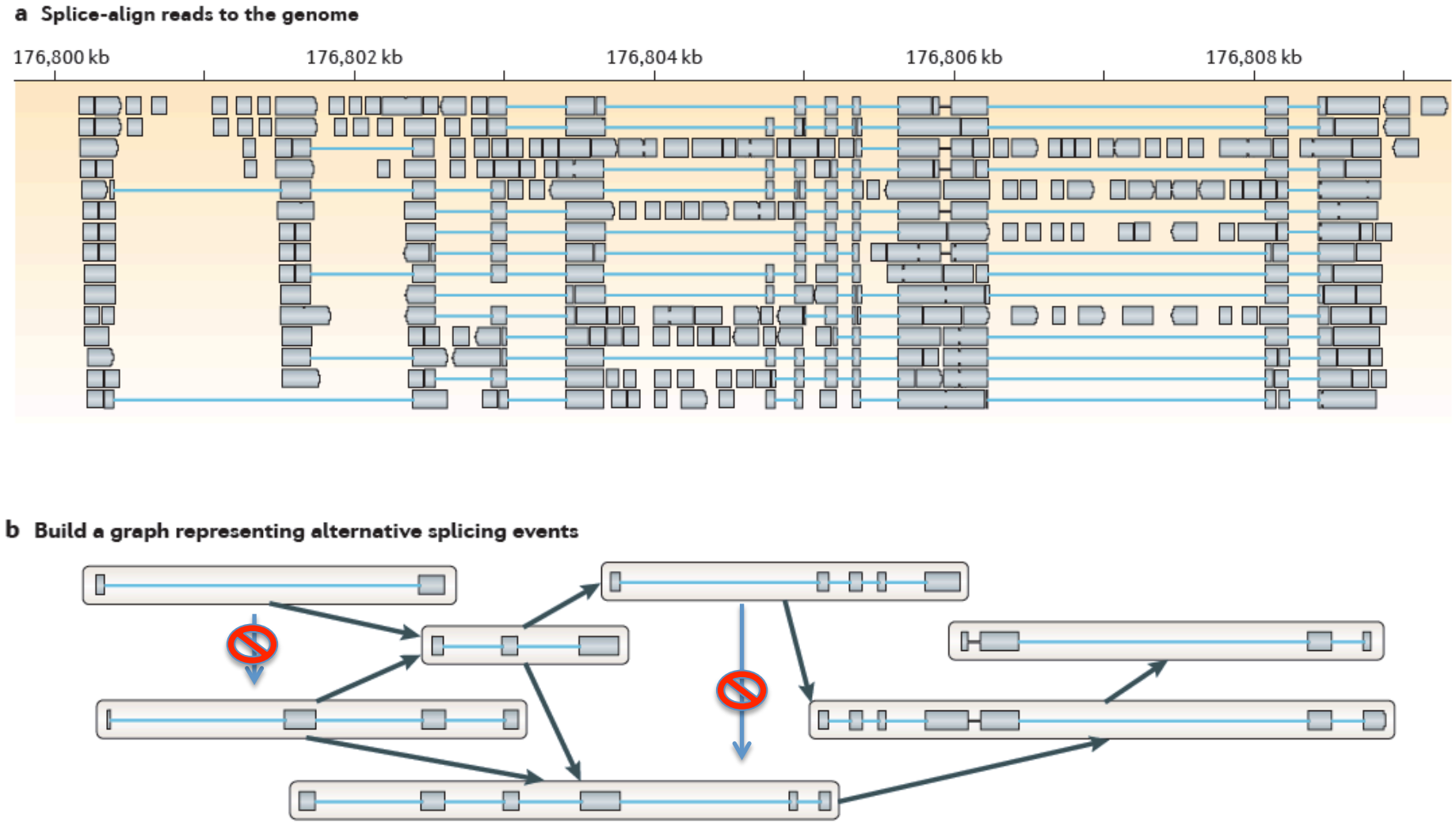


Transcript Reconstruction Using Cufflinks



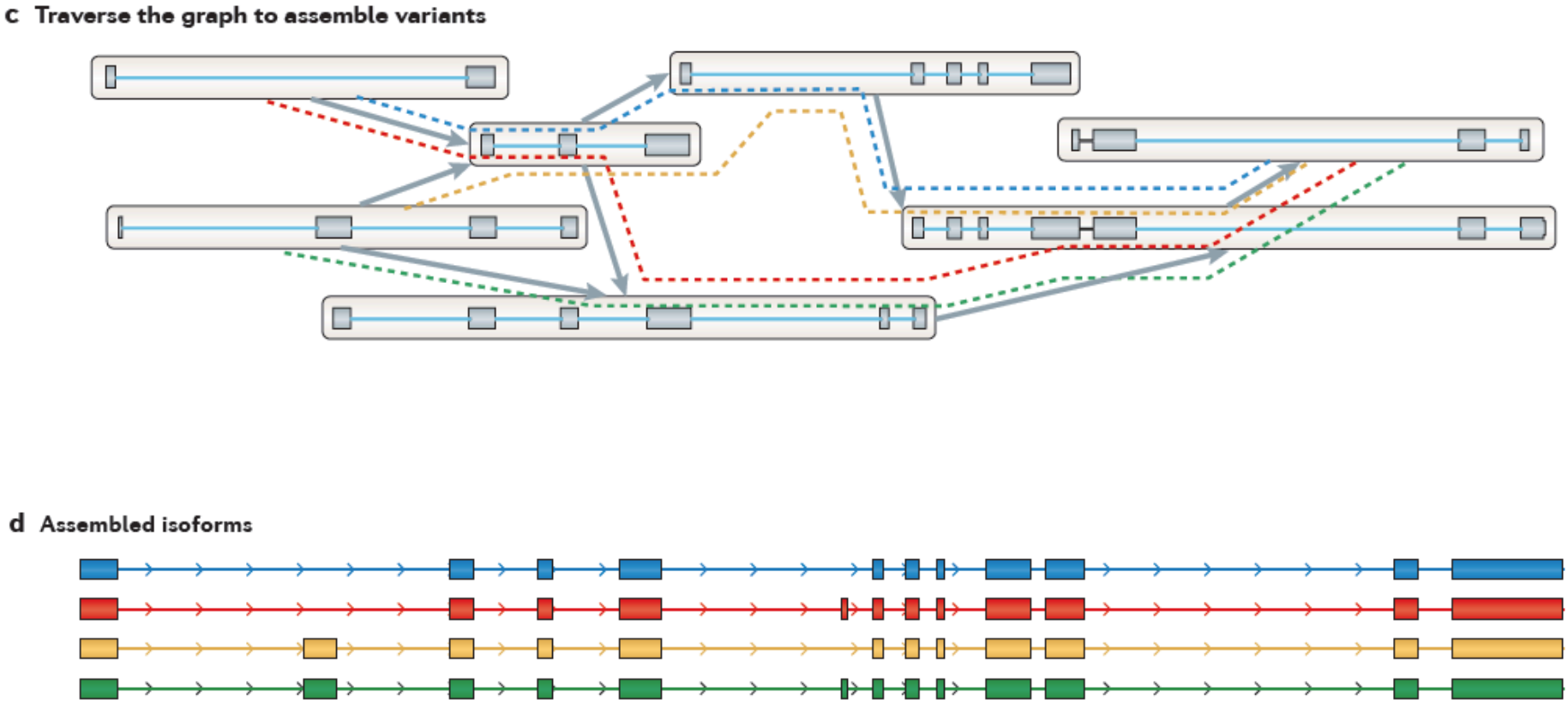
From Martin & Wang. Nature Reviews in Genetics. 2011

Transcript Reconstruction Using Cufflinks



From Martin & Wang. Nature Reviews in Genetics. 2011

Transcript Reconstruction Using Cufflinks



From Martin & Wang. Nature Reviews in Genetics. 2011

Transcript Structures in **GTF** Format

(tab-delimited fields per line shown transposed to a column format here)

0	7000000090838467	(genomic contig identifier)
1	Cufflinks	
2	transcript	
3	101	(left coordinate)
4	5716	(right coordinate)
5	1000	
6	+	(strand)
7	.	
8	gene_id "CUFF.1"; transcript_id "CUFF.1.1"; FPKM "378.0239937260“	(annotations)

0	7000000090838467
1	Cufflinks
2	exon
3	101
4	5716
5	1000
6	+
7	.
8	gene_id "CUFF.1"; transcript_id "CUFF.1.1"; exon_number "1"; FPKM "378.0239937260“

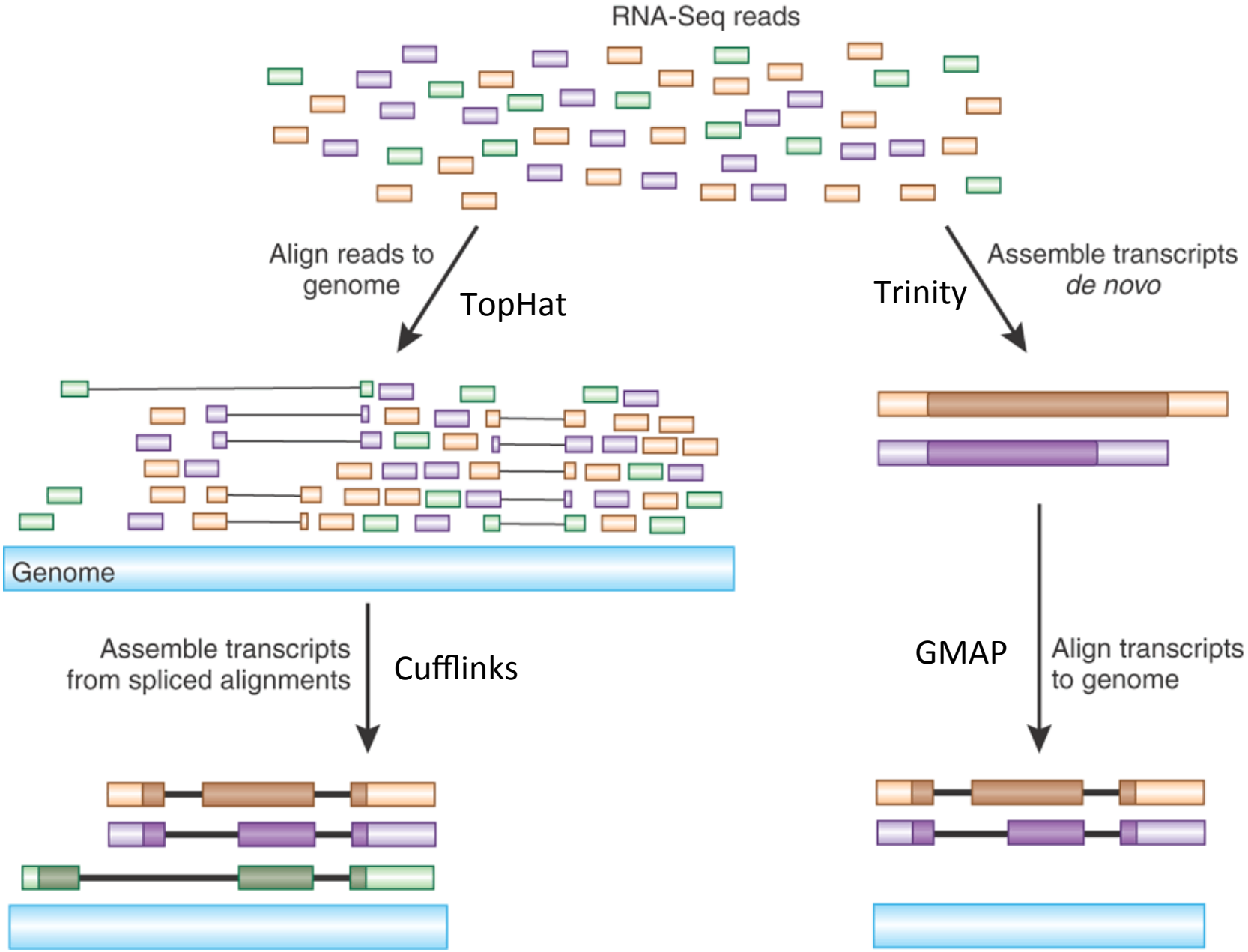
De novo transcriptome assembly

No genome required

Empower studies of non-model organisms

- expressed gene content
- transcript abundance
- differential expression

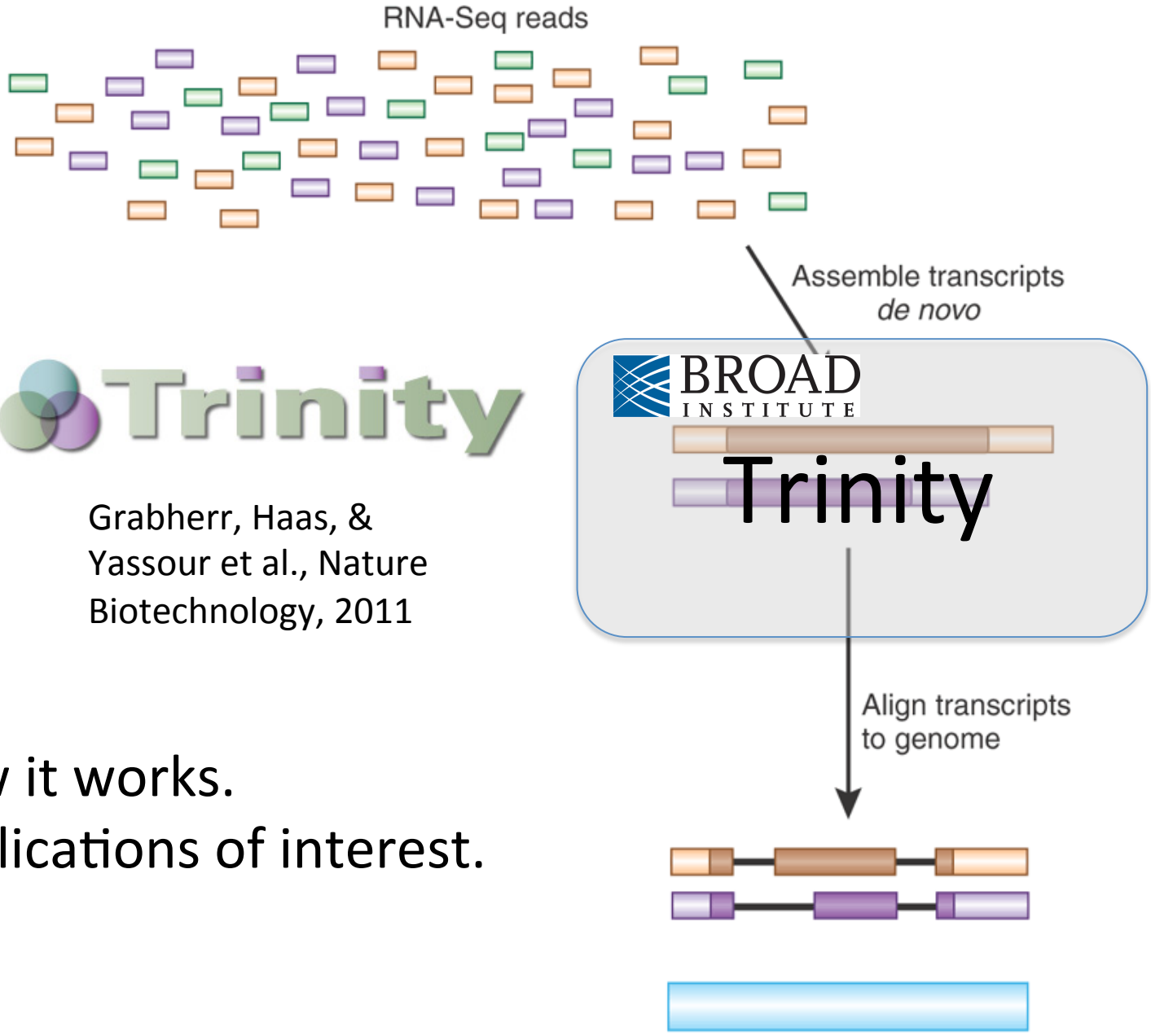
Transcript Reconstruction from RNA-Seq Reads



Transcript Reconstruction from RNA-Seq Reads

Trinity

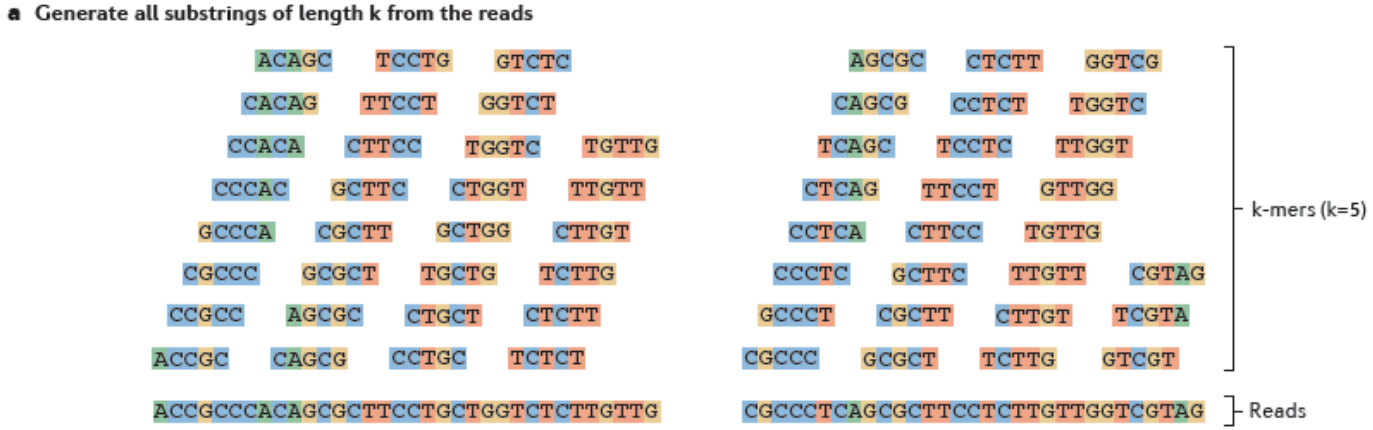
Grabherr, Haas, &
Yassour et al., Nature
Biotechnology, 2011

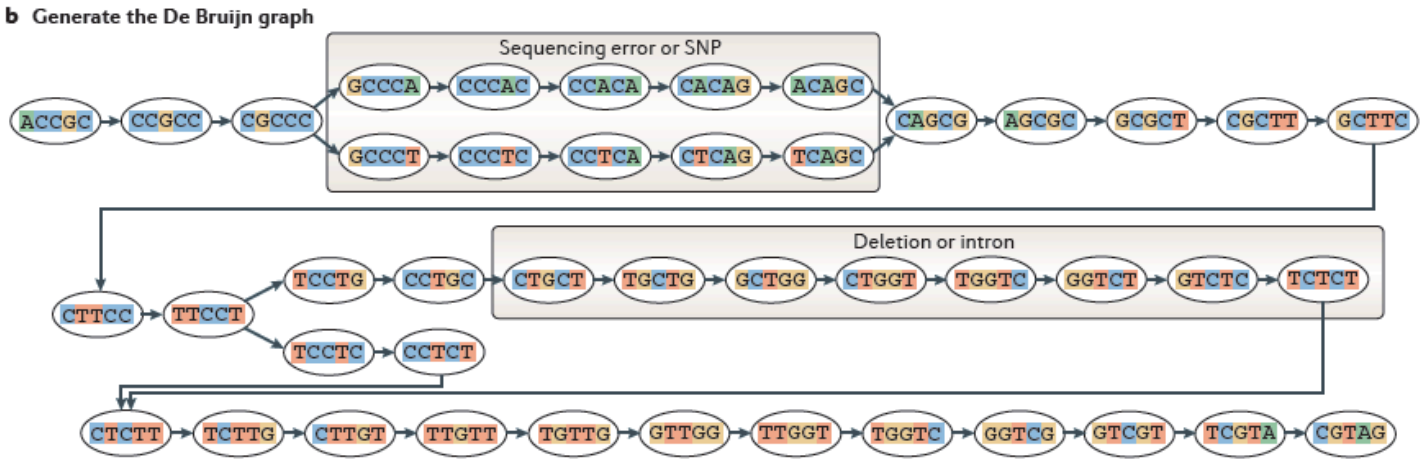


- How it works.
- Applications of interest.

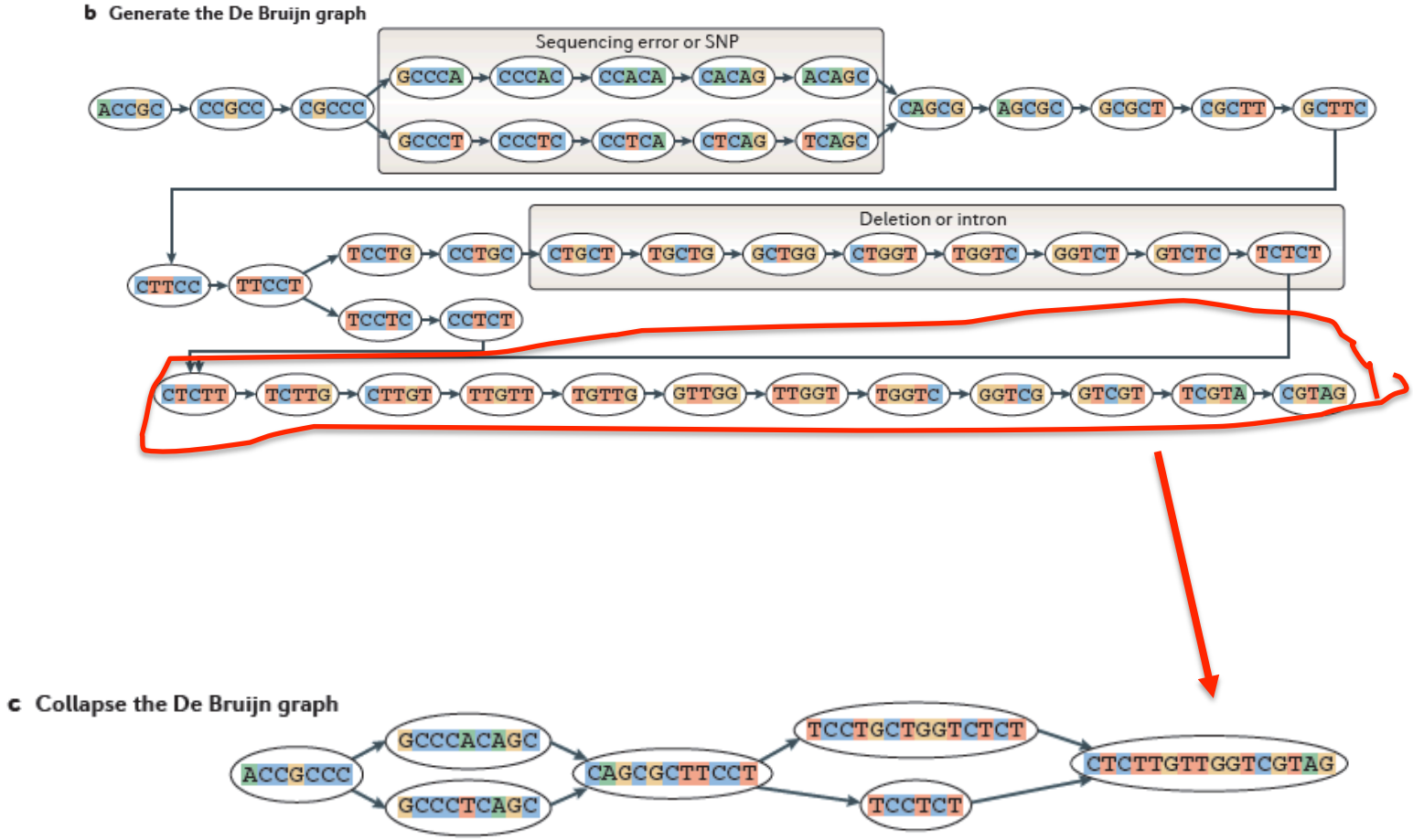
The General Approach to *De novo* RNA-Seq Assembly Using De Bruijn Graphs

Sequence Assembly via De Bruijn Graphs

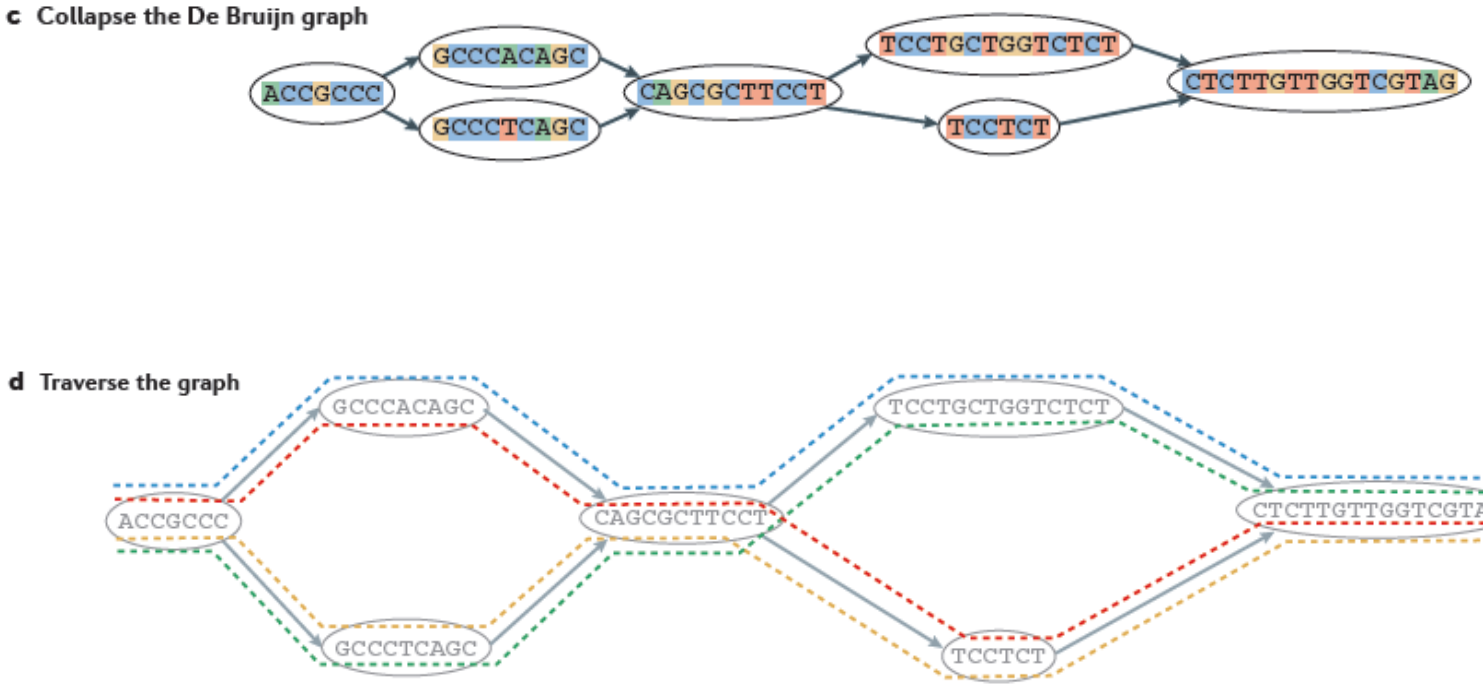




From Martin & Wang, Nat. Rev. Genet. 2011



From Martin & Wang, Nat. Rev. Genet. 2011



From Martin & Wang, Nat. Rev. Genet. 2011

Contrasting Genome and Transcriptome Assembly

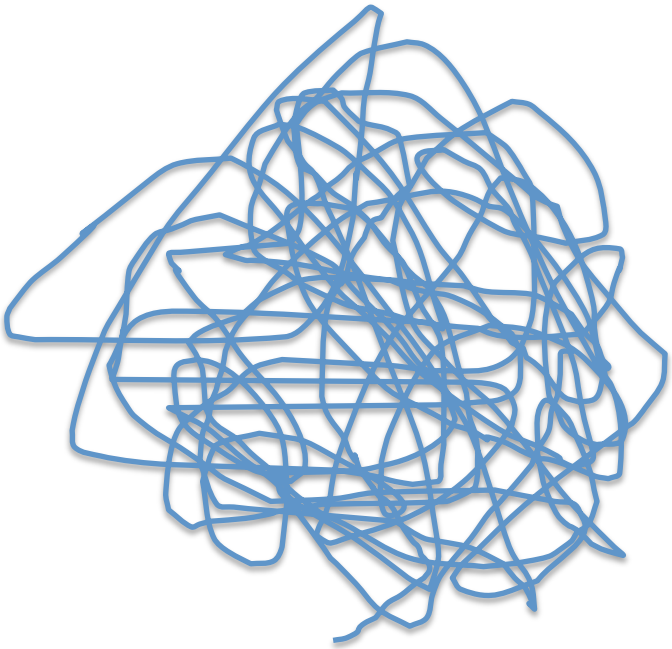
Genome Assembly

Transcriptome Assembly

Trinity Aggregates Isolated Transcript Graphs

Genome Assembly

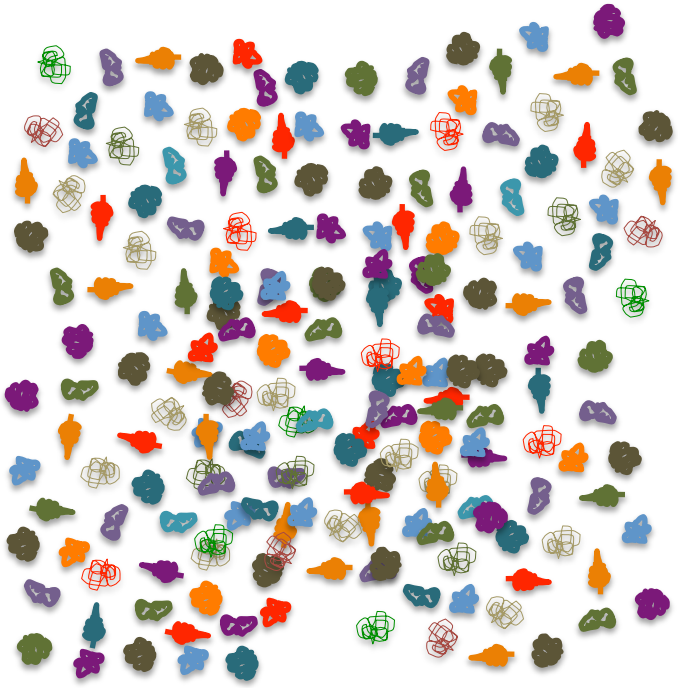
Single Massive Graph



Entire chromosomes represented.

Trinity Transcriptome Assembly

Many Thousands of Small Graphs



Ideally, one graph per expressed gene.

Trinity – How it works:

RNA-Seq reads → **Linear contigs** → **de-Bruijn graphs** → **Transcripts + Isoforms**

>a121:len=5845
>a122:len=2560
>a123:len=4443
>a124:len=48
>a125:len=8876
>a126:len=66

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

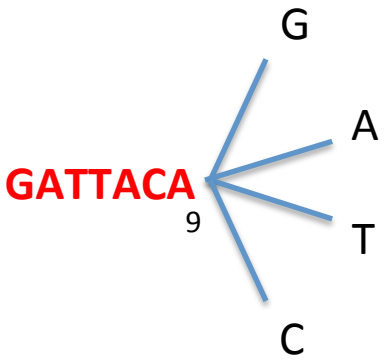
Thousands of disjoint graphs

Inchworm Algorithm

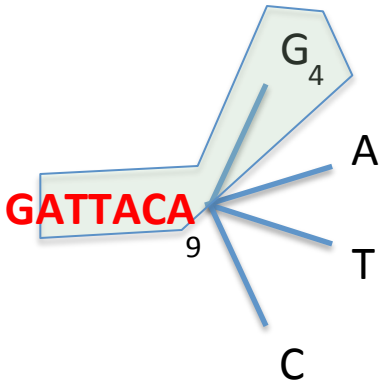
Decompose all reads into overlapping Kmers (25-mers)

Identify seed kmer as most abundant Kmer, ignoring low-complexity kmers.

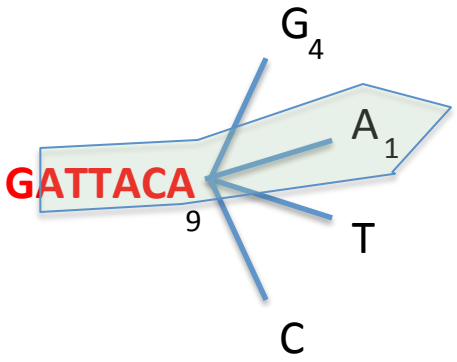
Extend kmer at 3' end, guided by coverage.



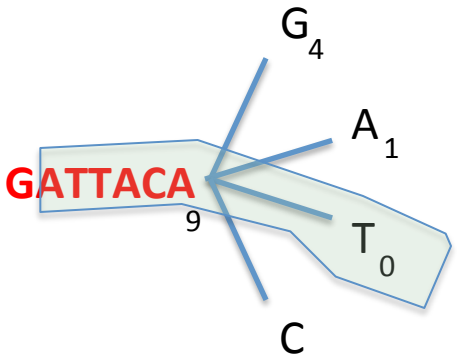
Inchworm Algorithm



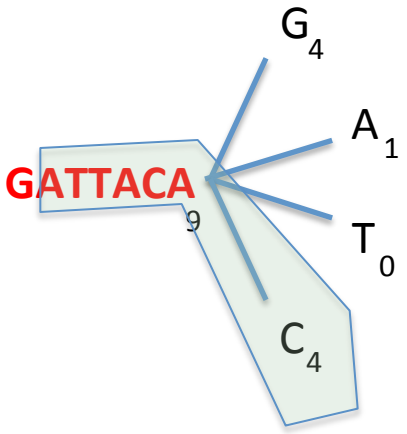
Inchworm Algorithm



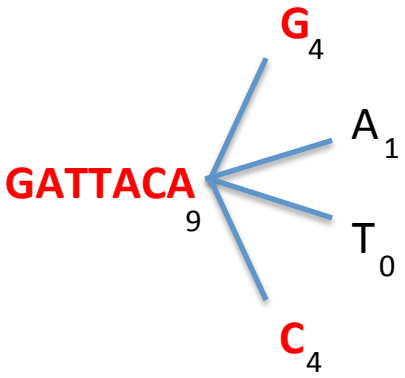
Inchworm Algorithm



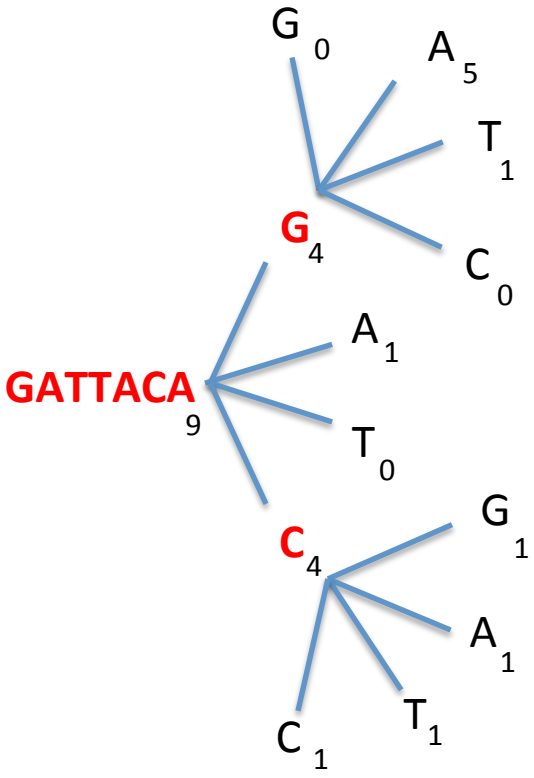
Inchworm Algorithm



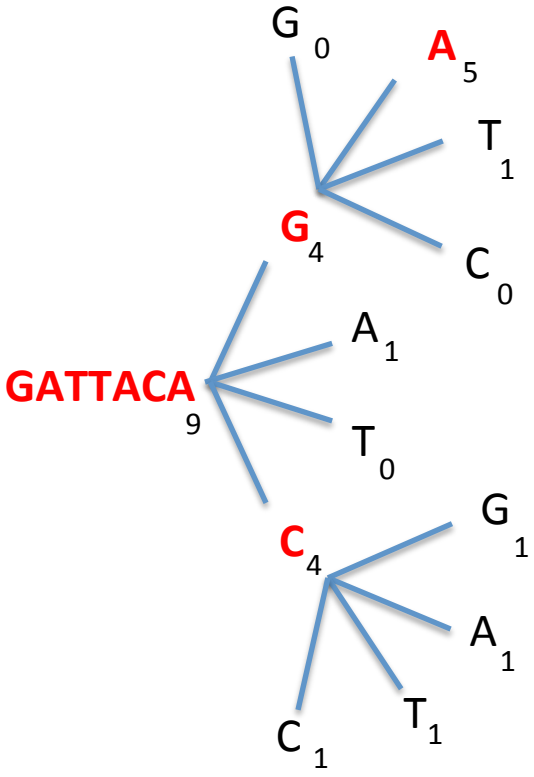
Inchworm Algorithm



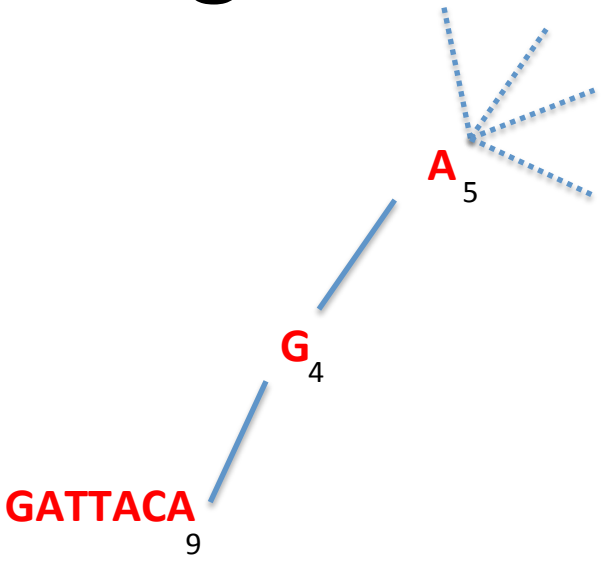
Inchworm Algorithm



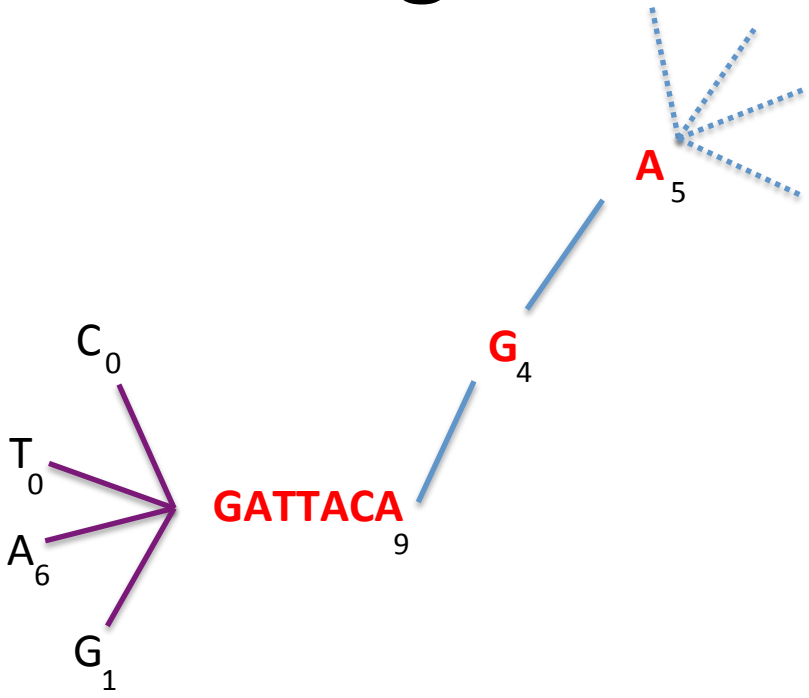
Inchworm Algorithm



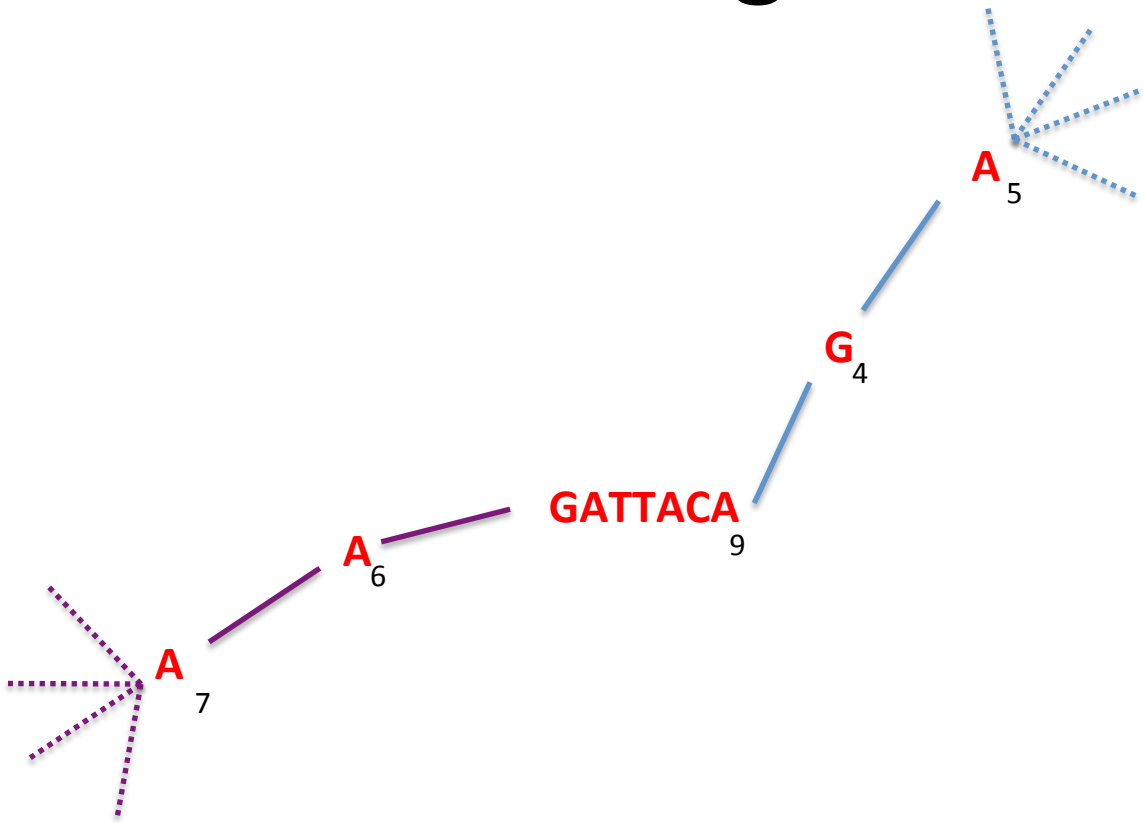
Inchworm Algorithm



Inchworm Algorithm



Inchworm Algorithm



Report contig: ….AAGATTACAGA….

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

Inchworm Contigs from Alt-Spliced Transcripts

Expressed isoforms

Isoform A

Isoform B

Inchworm Contigs from Alt-Spliced Transcripts

Expressed isoforms

Isoform A

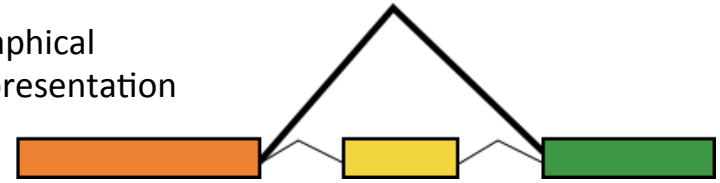
Isoform B

Expression

(low)

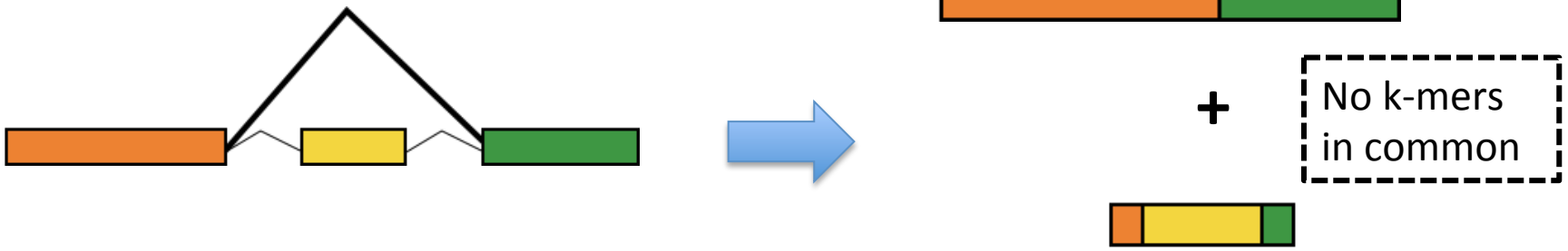
(high)

Graphical representation

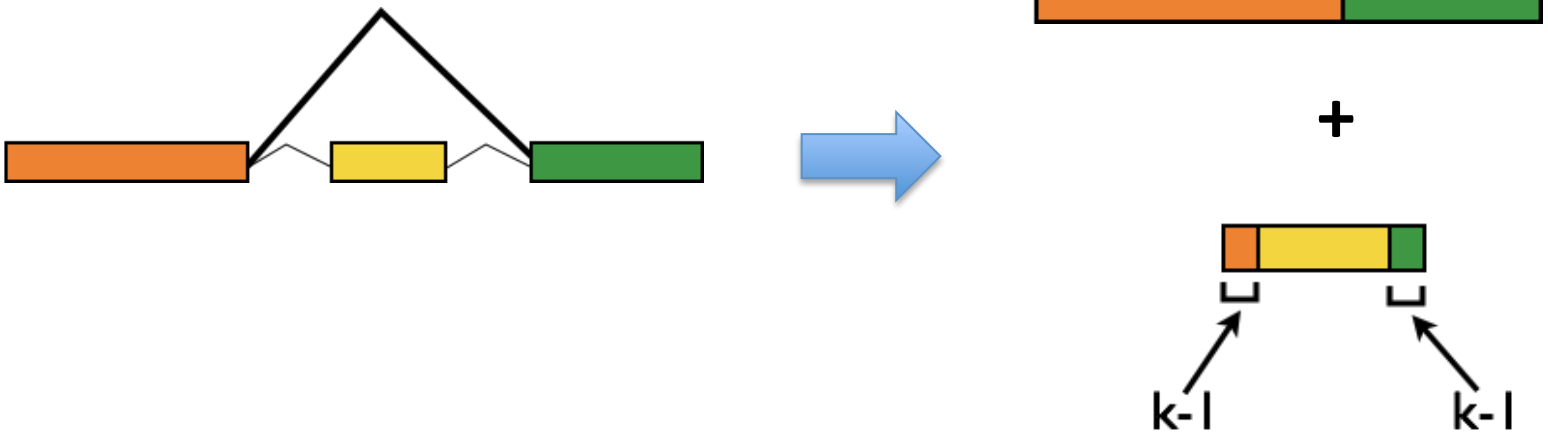


Inchworm Contigs from Alt-Spliced Transcripts

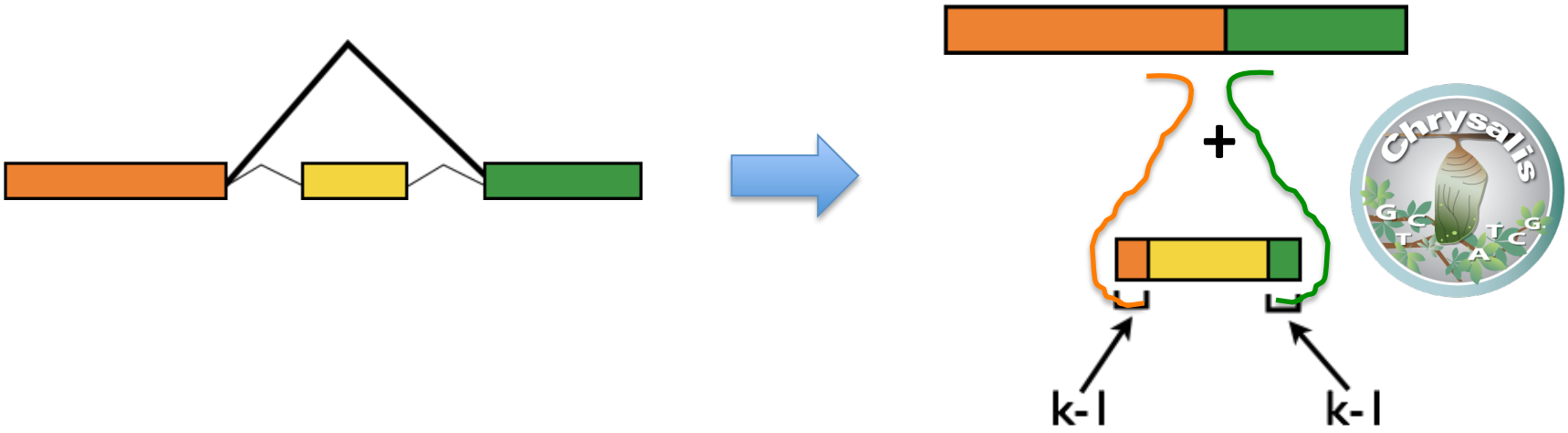
Inchworm Contigs from Alt-Spliced Transcripts



Inchworm Contigs from Alt-Spliced Transcripts



Chrysalis Re-groups Related Inchworm Contigs



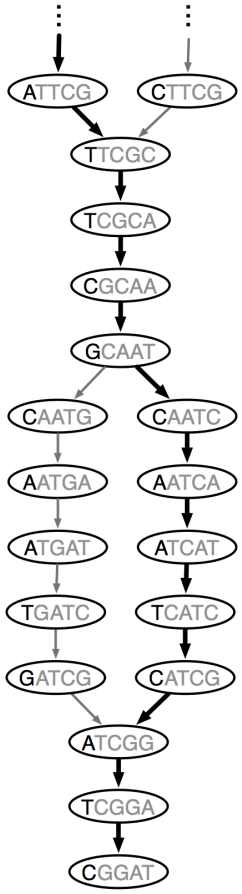
Chrysalis uses (k-1) overlaps and read support to link related Inchworm contigs

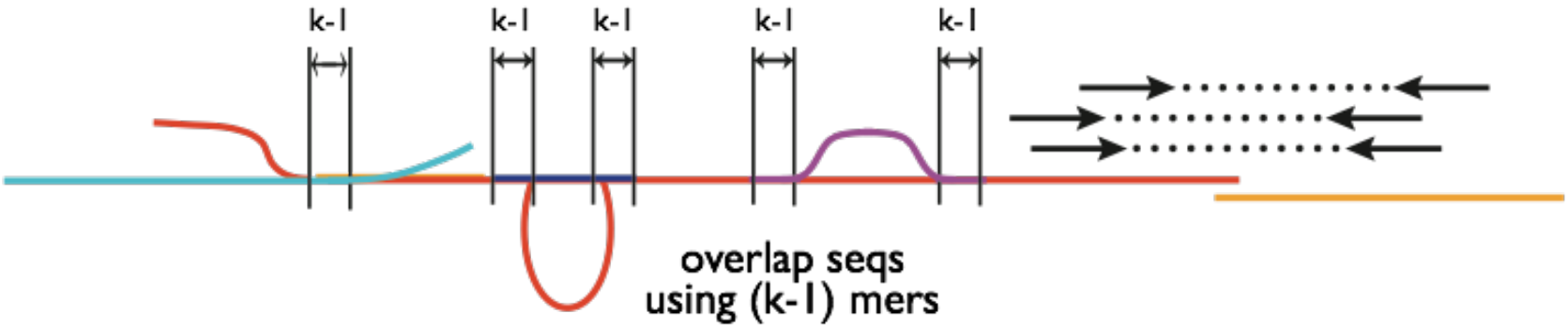
Chrysalis

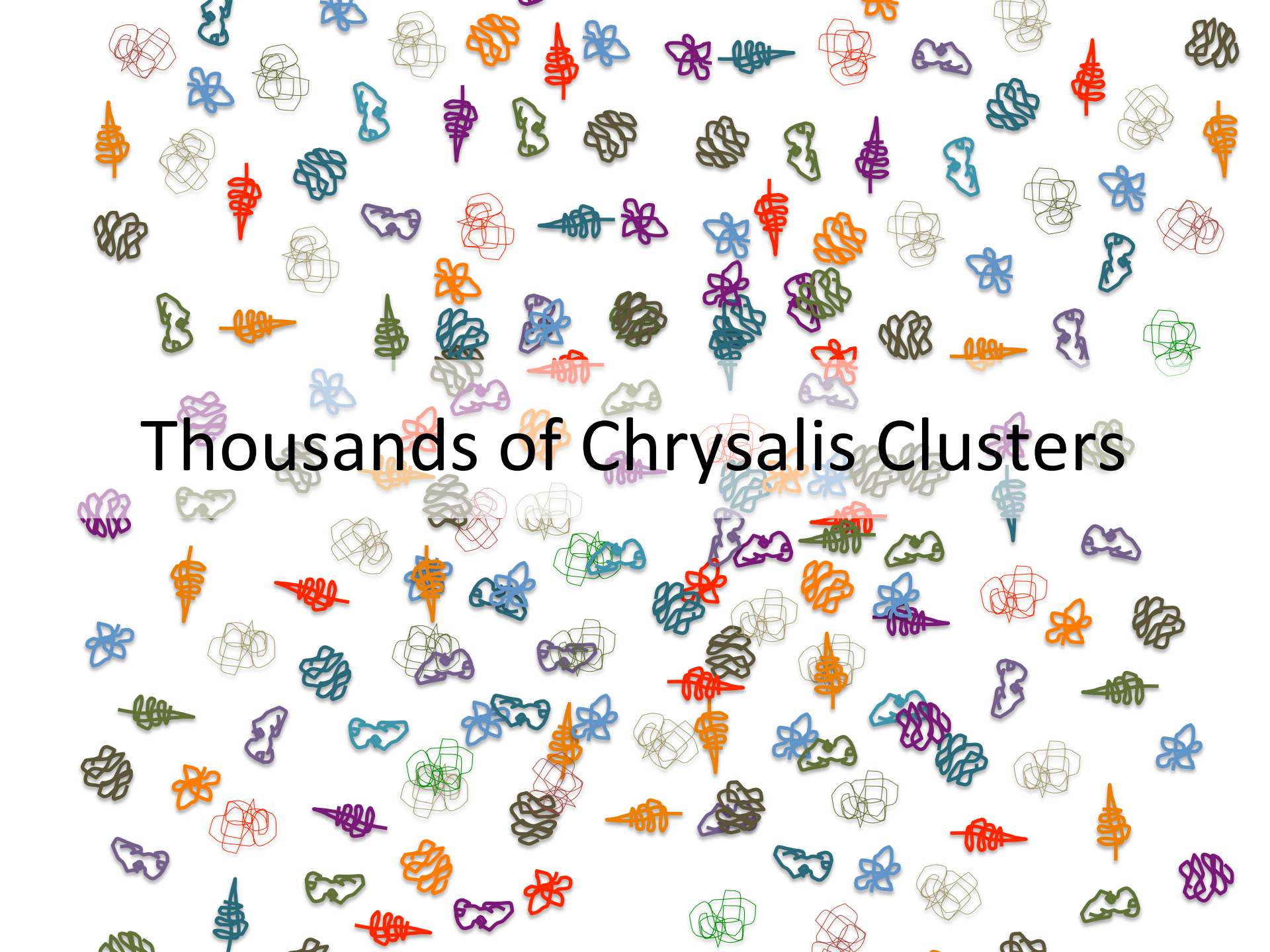
>a121:len=5845
>a122:len=2560
>a123:len=4443
>a124:len=48
>a125:len=8876
>a126:len=66

Integrate isoforms
via k-1 overlaps

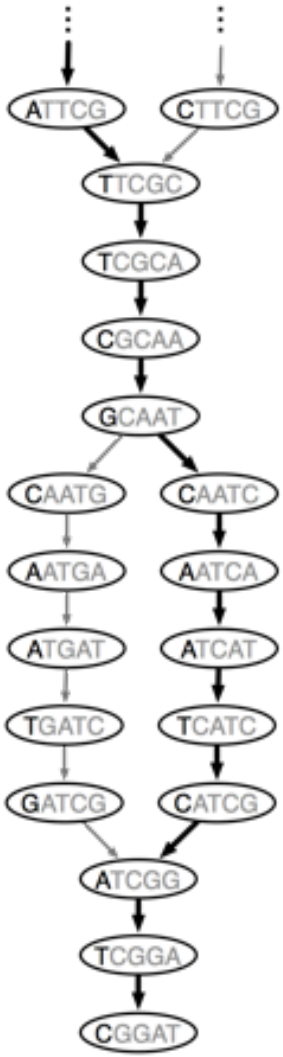
Build de Bruijn Graphs
(ideally, one per gene)





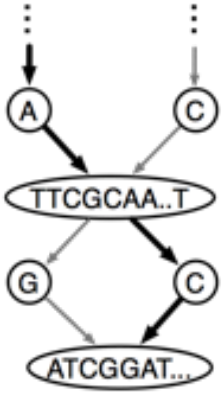

Thousands of Chrysalis Clusters

Butterfly



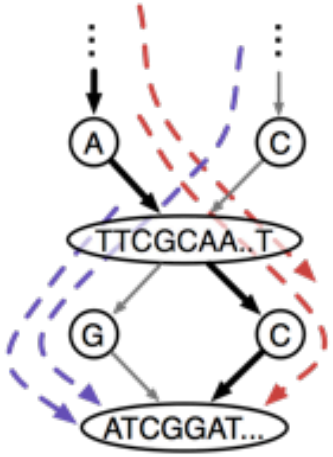
de Bruijn graph

compacting



compact graph

finding paths



compact graph with reads

extracting sequences

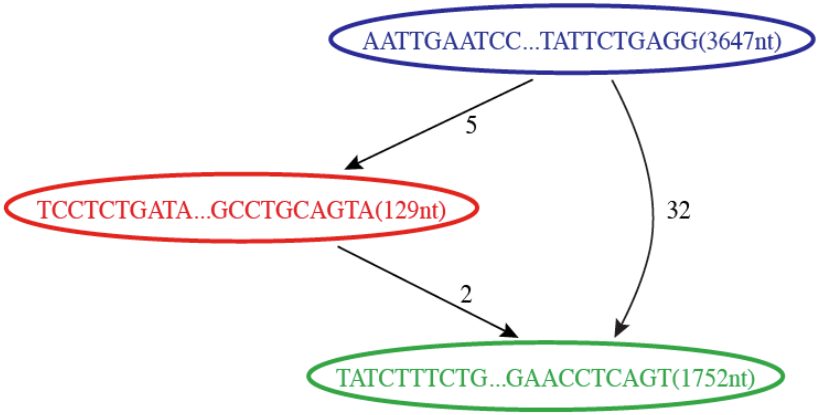
..CTTCGCAA..TGATCGGAT...

..ATTCGCAA..TCATCGGAT...

sequences (isoforms and paralogs)

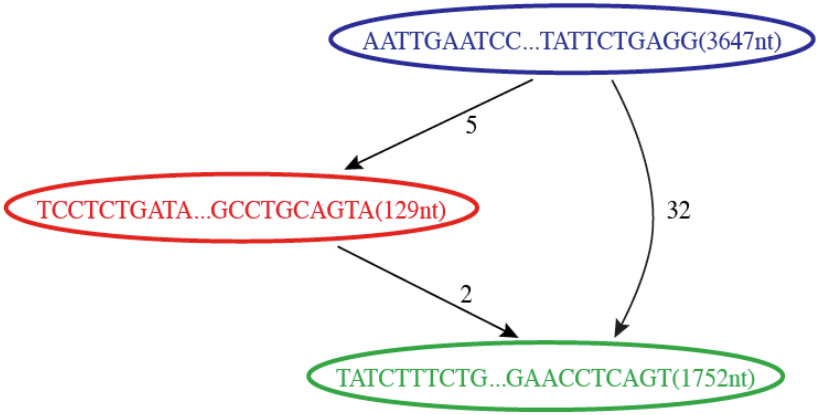
Butterfly Example 1:
Reconstruction of Alternatively Spliced Transcripts

Butterfly's Compacted Sequence Graph



Reconstruction of Alternatively Spliced Transcripts

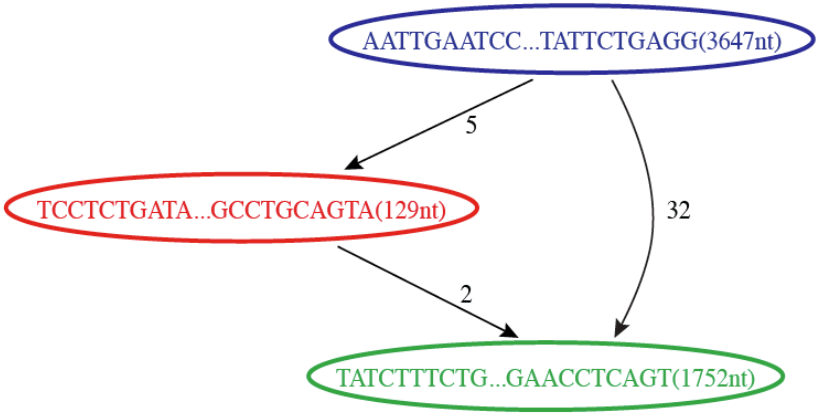
Butterfly's Compacted Sequence Graph



Reconstructed Transcripts

Reconstruction of Alternatively Spliced Transcripts

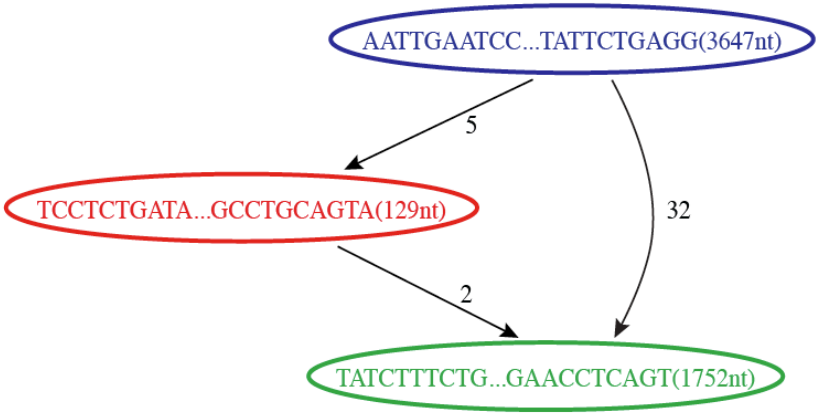
Butterfly's Compacted Sequence Graph



Reconstructed Transcripts

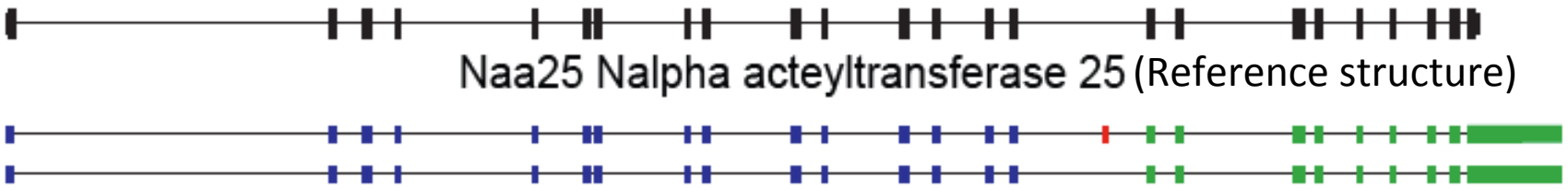
Reconstruction of Alternatively Spliced Transcripts

Butterfly's Compacted Sequence Graph



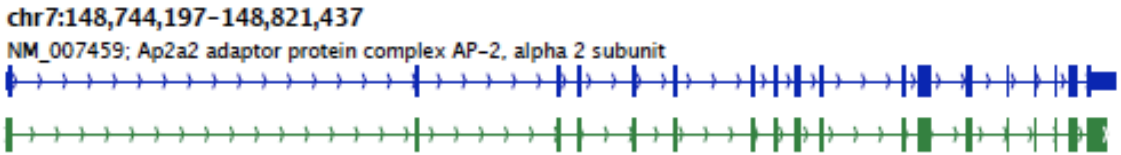
Reconstructed Transcripts

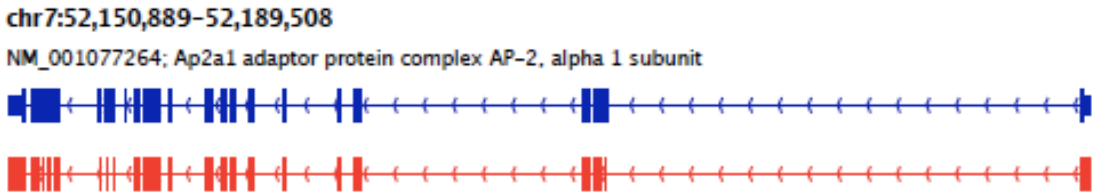
Aligned to Mouse Genome



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:

ex. Forward != reverse complement

(GGAA != TTCC)

Biologically: separate sense vs. antisense transcription

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

BROAD INSTITUTE

Comprehensive comparative analysis of strand-specific RNA sequencing methods

Joshua Z Levin[1,6], Moran Yassour[1–3,6], Xian Adiconis[1], Chad Nusbaum[1], Dawn Anne Thompson[1], Nir Friedman[3,4], Andreas Gnirke[1] & Aviv Regev[1,2,5]

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

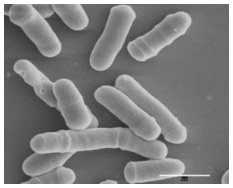
'dUTP second strand marking' identified as the leading protocol

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

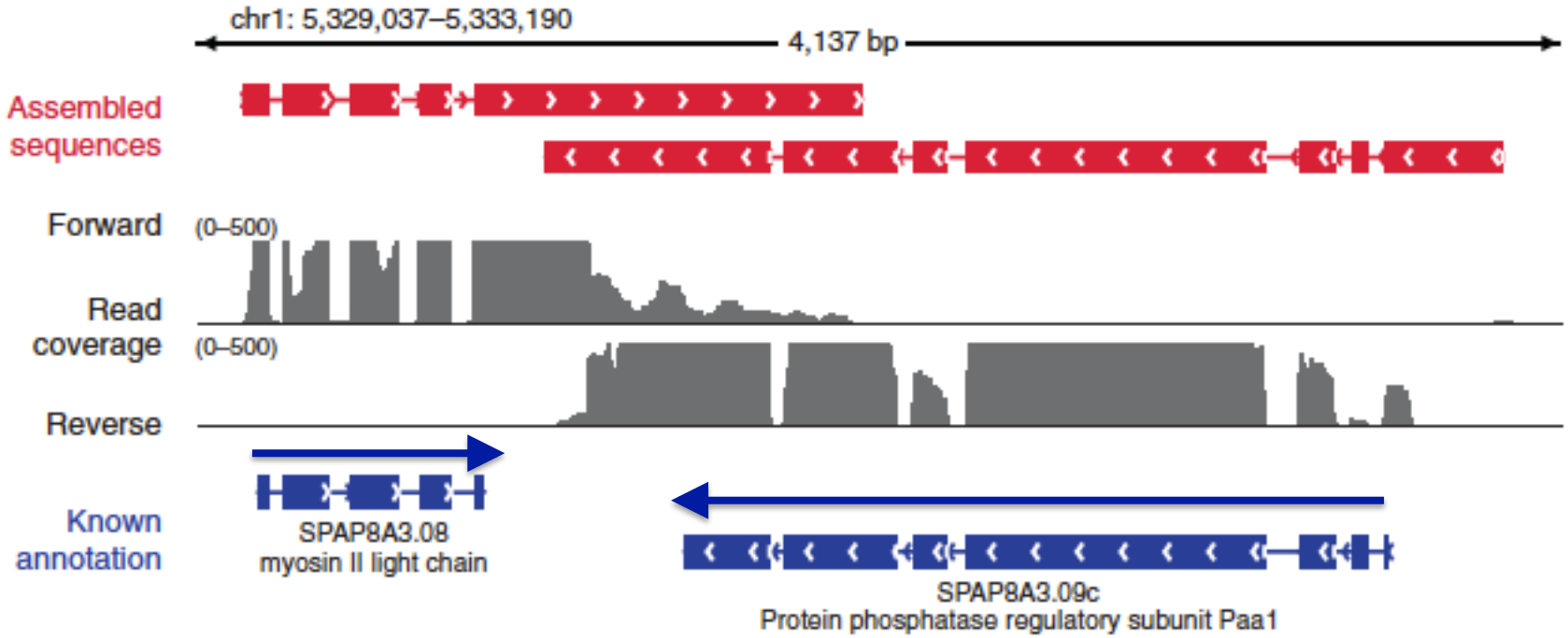
Nevertheless, direct information on the originating strand can substantially enhance the value of an RNA-seq experiment. For

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

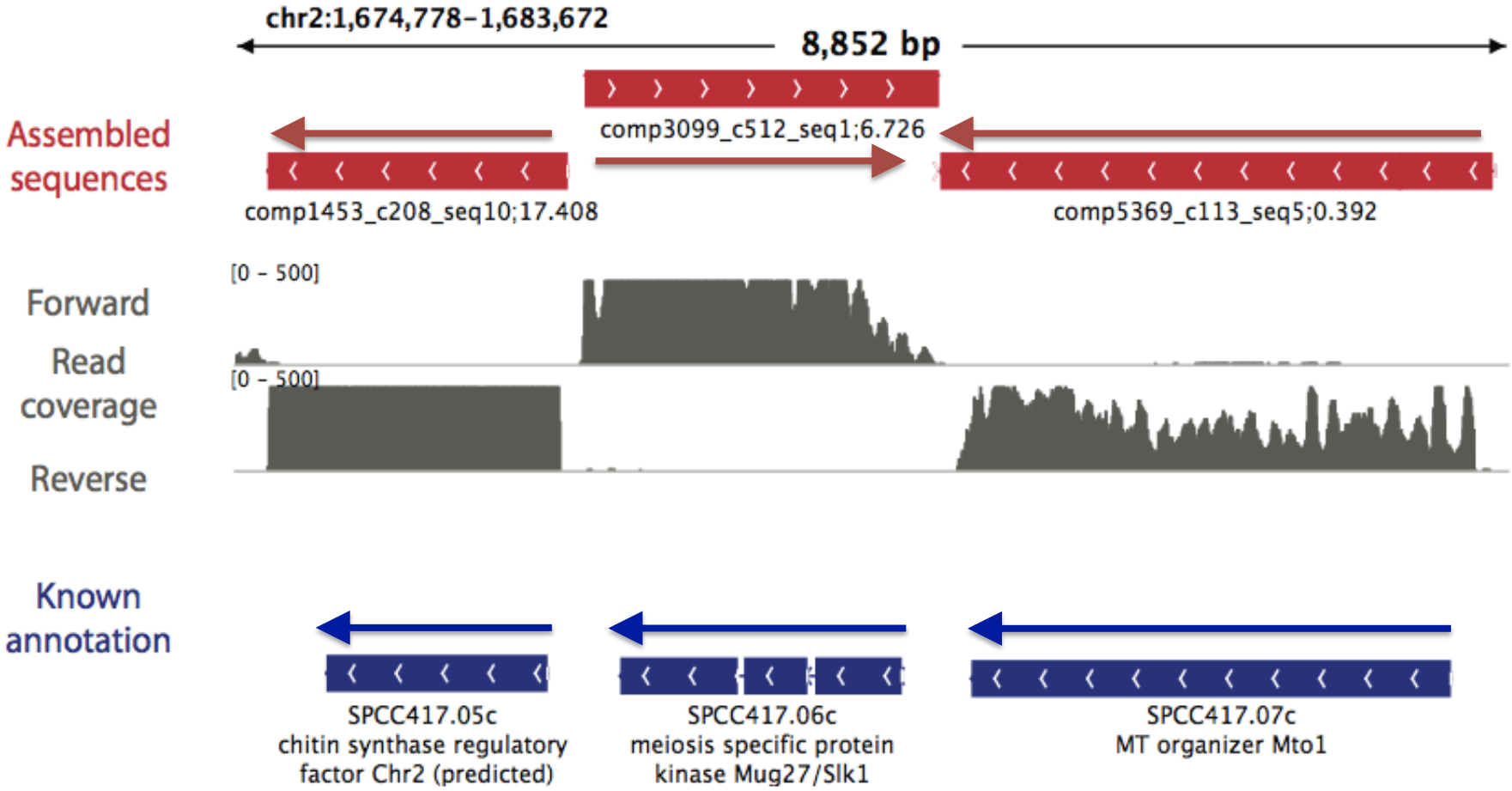
Overlapping UTRs from Opposite Strands

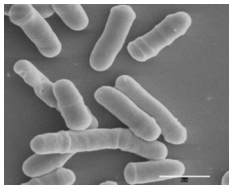


Schizosacharomyces pombe
(fission yeast)



Antisense-dominated Transcription



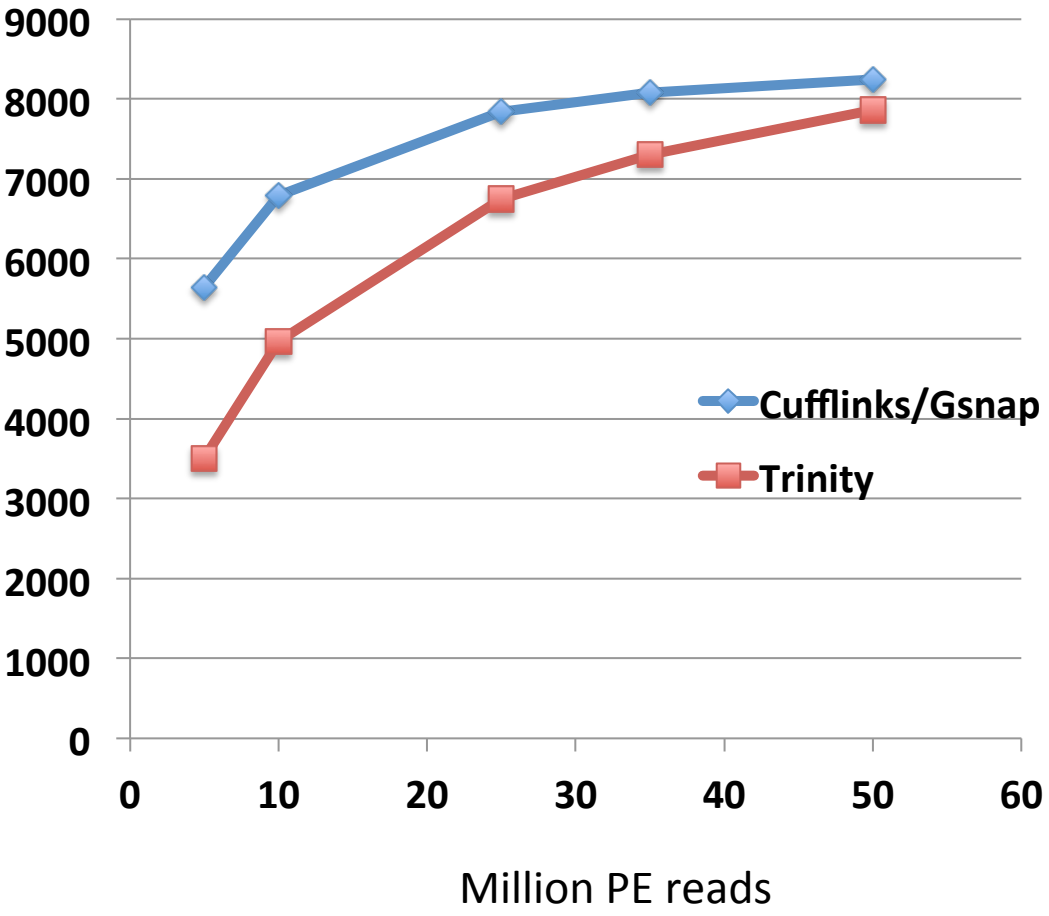


Trinity output: A multi-fasta file

Can align Trinity transcripts to genome scaffolds to examine intron/exon structures

(Trinity transcripts aligned using GMAP)

Improved reconstruction with deeper sequencing depth and Genome-based reconstruction is more sensitive than de novo methods



Mouse data

Summary

- Two paradigms for transcript reconstruction
  - Rna-seq alignment assembly
    - Tuxedo (tophat, cufflinks)
  - genome-free de novo read assembly
    - Trinity

- Best to pursue both strategies
  - Maximize sensitivity for genome-based transcript reconstruction + capture missing or ill-represented transcripts via de novo assembly.

Software Links

- Tuxedo
  - Tophat: http://tophat.cbcb.umd.edu/
  - Cufflinks: http://cufflinks.cbcb.umd.edu/
- Trinity
  - http://trinityrnaseq.sourceforge.net/
- Visualization
  - IGV: http://www.broadinstitute.org/igv/
  - Genomeview: http://genomeview.org
- GMAP
  - http://research-pub.gene.com/gmap/
- Samtools
  - http://samtools.sourceforge.net/

Papers of Interest

- Next generation transcriptome assembly
  - http://www.nature.com/nrg/journal/v12/n10/full/nrg3068.html
- Tuxedo protocol
  - http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3334321/
- Trinity
  - http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3571712/