Transcriptome Assembly

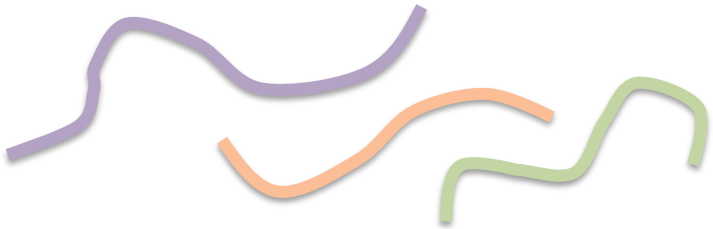
CSHL SEQTEC 2018

Brian Haas

Broad Institute

RNA-Seq Empowers Transcriptome Studies

Extract RNA, convert to cDNA

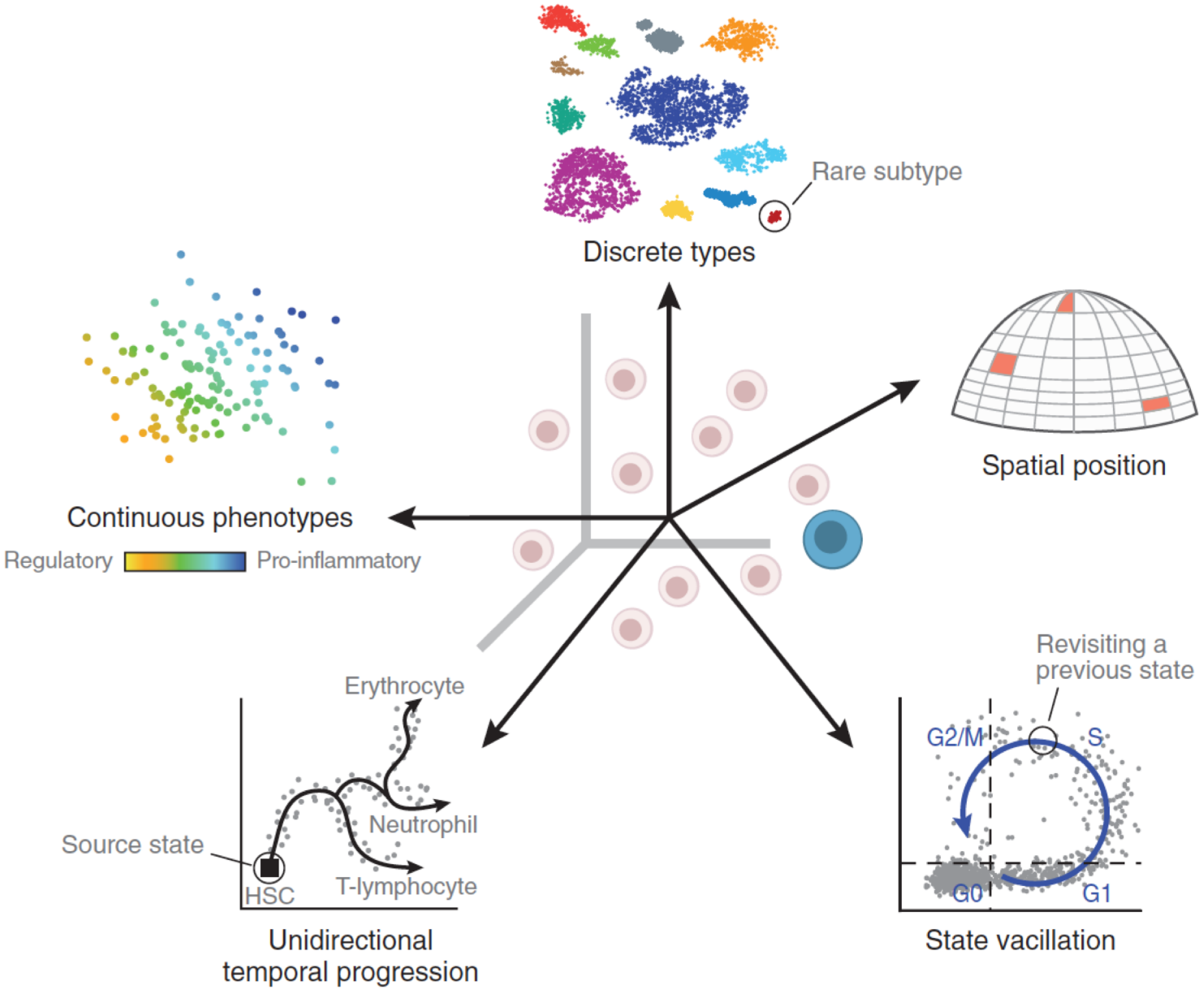


Next-gen Sequencer
(pick your favorite)

RNA-Seq Empowers Many Facets of Biological Investigations

- Transcript identification (ie. which genes active)
- Expression Levels
- Alternative splicing isoforms
- Allelic variants
- Mutations
- Fusion Transcripts
- RNA-editing

RNA-Seq is Empowering Discovery at Single Cell Resolution



Wagner, Regev, and Yosef. NBT 2016

Generating RNA-Seq: *How to Choose?*

Many different instruments hit the scene in the last decade

Illumina

454

SOLiD

Helicos

Ion Torrent

Pacific Biosciences

Oxford Nanopore

Slide courtesy of Joshua Levin, Broad Institute.

Generating RNA-Seq: *How to Choose?*

Popular choices for RNA-Seq today

[Current RNA-Seq workhorse]

Illumina

Ion Torrent

[Full-length single molecule sequencing]

Pacific Biosciences

[Newly emerging technology for full-length single molecule sequencing]

Oxford Nanopore

Generating RNA-Seq: *How to Choose?*

Platform	Project Firefly 2018	MiniSeq	MiSeq	Next Seq 550	HiSeq 2500 RR	Hiseq 2500 V3	HiSeq 2500 V4	HiSeq 4000	HiSeq X	Nova Seq S1 2018	Nova Seq S2	Nova Seq S4	5500 XL	318 HiQ 520	Ion 530	Ion Proton P1	PGM HiQ 540	RS P6-C4	Sequel	R&D end 2018	Smidg ION RnD	Mini ION R9.5	Grid ION X5	Prome thION RnD	Prome thION theor etical	QiaGen Gene Reader	BGI SEQ 500	BGI SEQ 50	#
Reads: (M)	4	25	25	400	600	3000	4000	5000	6000	3300	6600	20000	1400	3-5	15-20	165	60-80	5.5	38.5	--	--	--	--	--	--	400	1600	1600	--
Read length: (paired-end*)	150*	150*	300*	150*	100*	100*	125*	150*	150*	150*	150*	150*	60	200 400	200 400	200	200	15K	12K	32K	--	--	--	--	--	--	100*	50	--
Run time: (d)	0.54	1	2	1.2	1.125	11	6	3.5	3	1.66	1.66	1.66	7	0.37	0.16	--	0.16	4.3	--	--	--	2	2	2	--	--	1	0.4	--
Yield: (Gb)	1	7.5	15	120	120	600	1000	1500	1800	1000	2000	6000	180	1.5	7	10	12	12	5	150	4	8	40	2400	11000	80	200	8	--
Rate: (Gb/d)	1.85	7.5	7.5	100	106.6	55	166	400	600	600	1200	3600	30	5.5	50	--	93.75	2.8	--	--	--	4	20	1200	5500	--	200	20	--
Reagents: ($K)	0.1	1.75	1	5	6.145	23.47	29.9	--	--	--	--	--	10.5	0.6	--	1	1.2	2.4	--	1	--	0.5	1.5	--	--	0.5	--	--	--
per-Gb: ($)	100	233	66	50	51.2	39.1	31.7	20.5	7.08	18	15	5.8	58.33	--	--	100	--	200	80	6.6	--	62.5	37.5	20	4.3	--	--	--	--
hg-30x: ($)	12000	28000	8000	5000	6144	4692	3804	2460	849.6	1800	1564	700	7000	--	--	12000	--	24000	9600	1000	--	7500	4500	2400	500	--	600	--	
Machine: ($)	30K	49.5K	99K	250K	740K	690K	690K	900K	1M	999K	999K	999K	595K	50K	65K	243K	242K	695K	350K	350K	--	--	125K	75K	75K	--	200K	--	--
#Page maintained by http://twitter.com/albertvilella http://tinyurl.com/ngslytics #Editable version: http://tinyurl.com/ngsspecsshared																													
#curl "https://docs.google.com/spreadsheets/d/1GMMfhyLK0-q8Xklo3YxlWaZA5vVMuhU1kg41g4xLkXc/export?gid=4&format=csv" | grep -v '^#' | grep -v '^"' | column -t -s\, | less -S																													

*Not all shown at scale

Generating RNA-Seq: *How to Choose?*

Platform	Project Firefly 2018	MiniSeq	MiSeq	Next Seq 550	HiSeq 2500 RR	Hiseq 2500 V
Reads: (M)	4	25	25	400	600	300
Read length: (paired-end*)	150*	150*	300*	150*	100*	100
Run time: (d)	0.54	1	2	1.2	1.125	1
Yield: (Gb)	1	7.5	15	120	120	60
Rate: (Gb/d)	1.85	7.5	7.5	100	106.6	5
Reagents: ($K)	0.1	1.75	1	5	6.145	23.4
per-Gb: ($)	100	233	66	50	51.2	39.
hg-30x: ($)	12000	28000	8000	5000	6144	469
Machine: ($)	30K	49.5K	99K	250K	740K	690

#Page maintained by http://twitter.com/albertvilella http://t

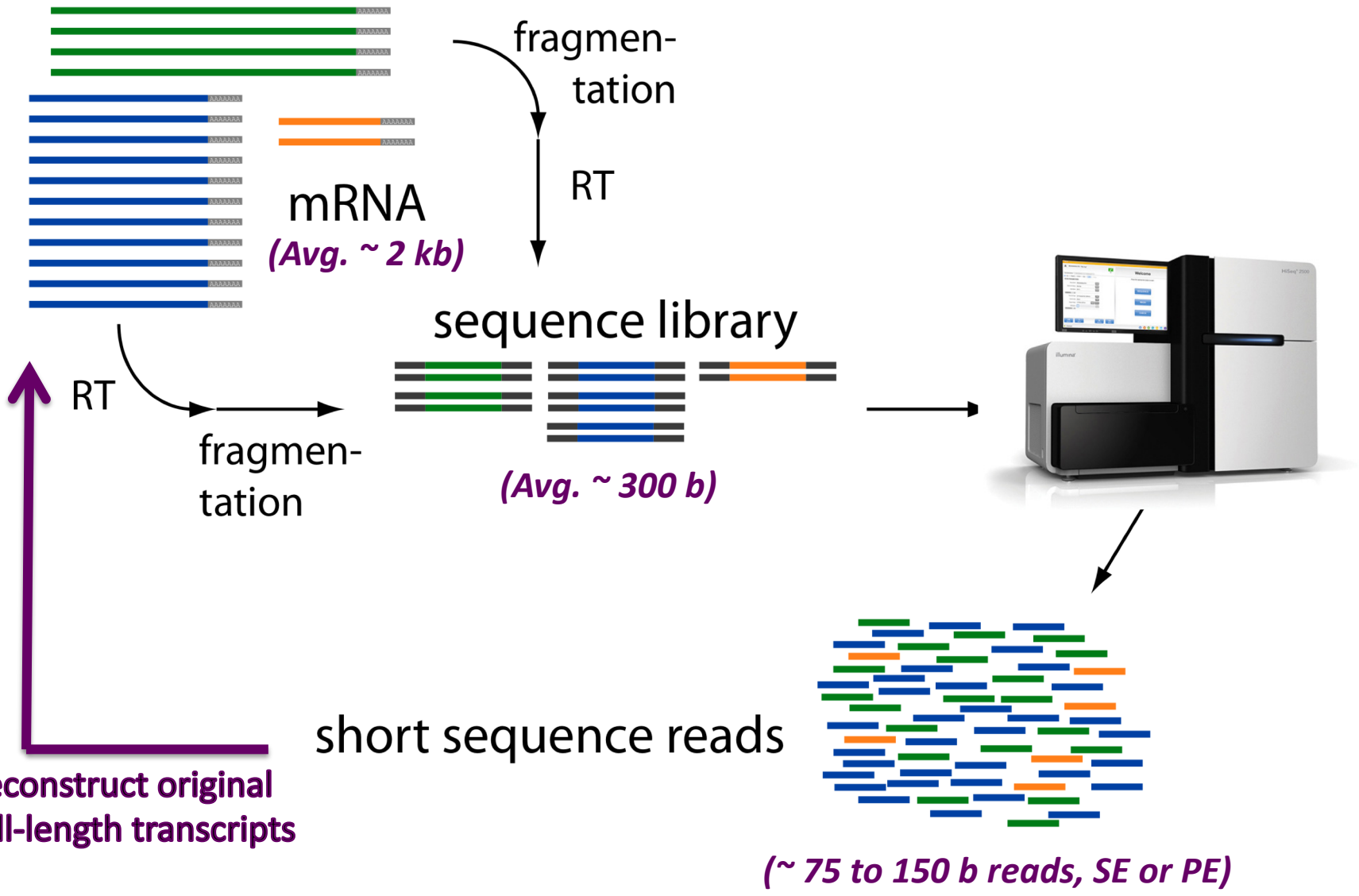
#curl "https://docs.google.com/spreadsheets/d/1GMMfhyLK0-q

g	Mini ION R9.5	Grid ION X5	Prome thION RnD	Prome thION theor etical	QiaGen Gene Reader	BGI SEQ 500	BGI SEQ 50	#
--	--	--	--	--	400	1600	1600	--
--	--	--	--	--	--	100*	50	--
--	2	2	2	--	--	1	0.4	--
4	8	40	2400	11000	80	200	8	--
--	4	20	1200	5500	--	200	20	--
--	0.5	1.5	--	--	0.5	--	--	--
--	62.5	37.5	20	4.3	--	--	--	--
--	7500	4500	2400	500	--	600	--	
--	--	125K	75K	75K	--	200K	--	--

"What I especially like about this baby is this little drawer where I can keep my lunch."

Thx Joshua Levin, for the cartoon. ☺

RNA-Seq Challenge: Transcript Reconstruction



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

Transcript Reconstruction from RNA-Seq Reads

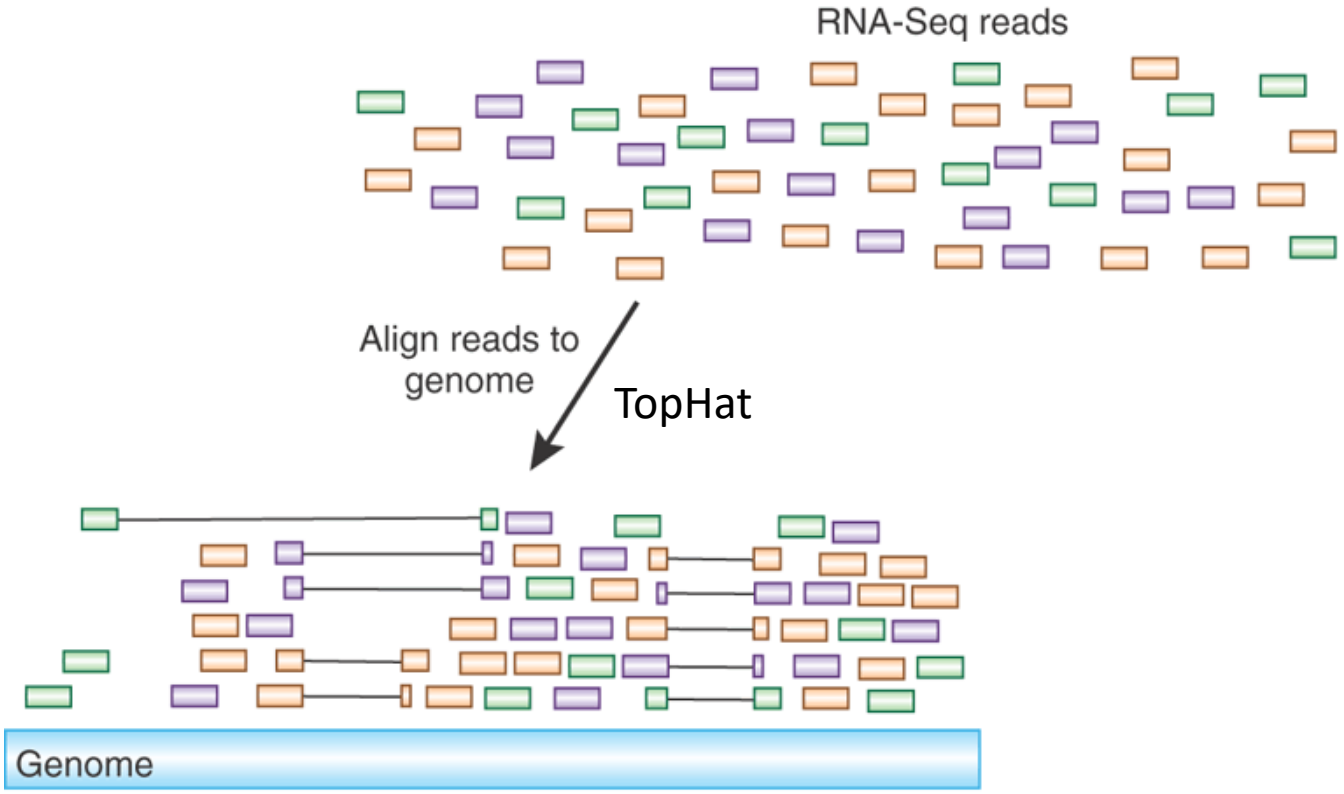
Advancing RNA-Seq analysis

Brian J Haas & Michael C Zody

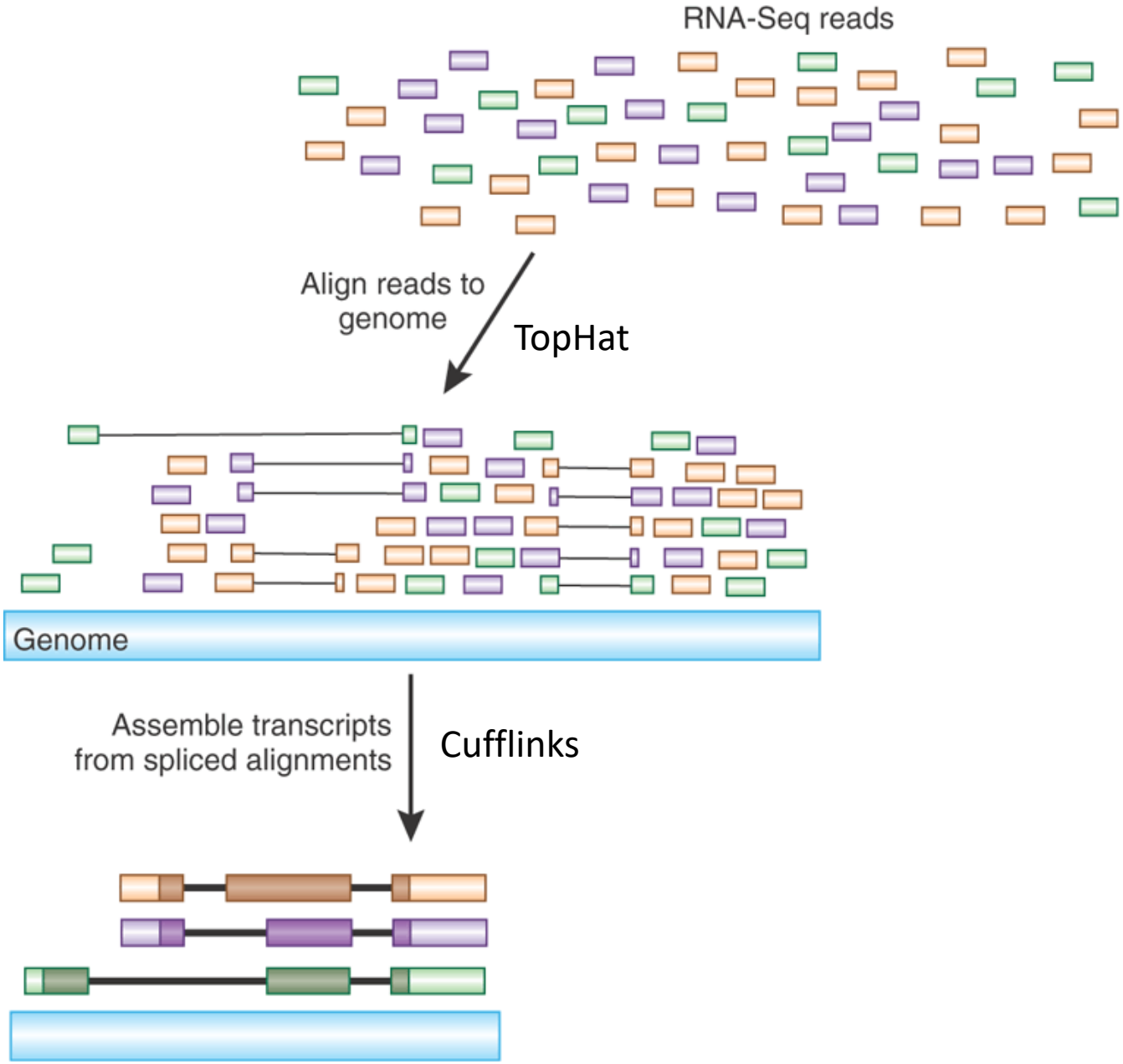
Nature Biotech, 2010

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

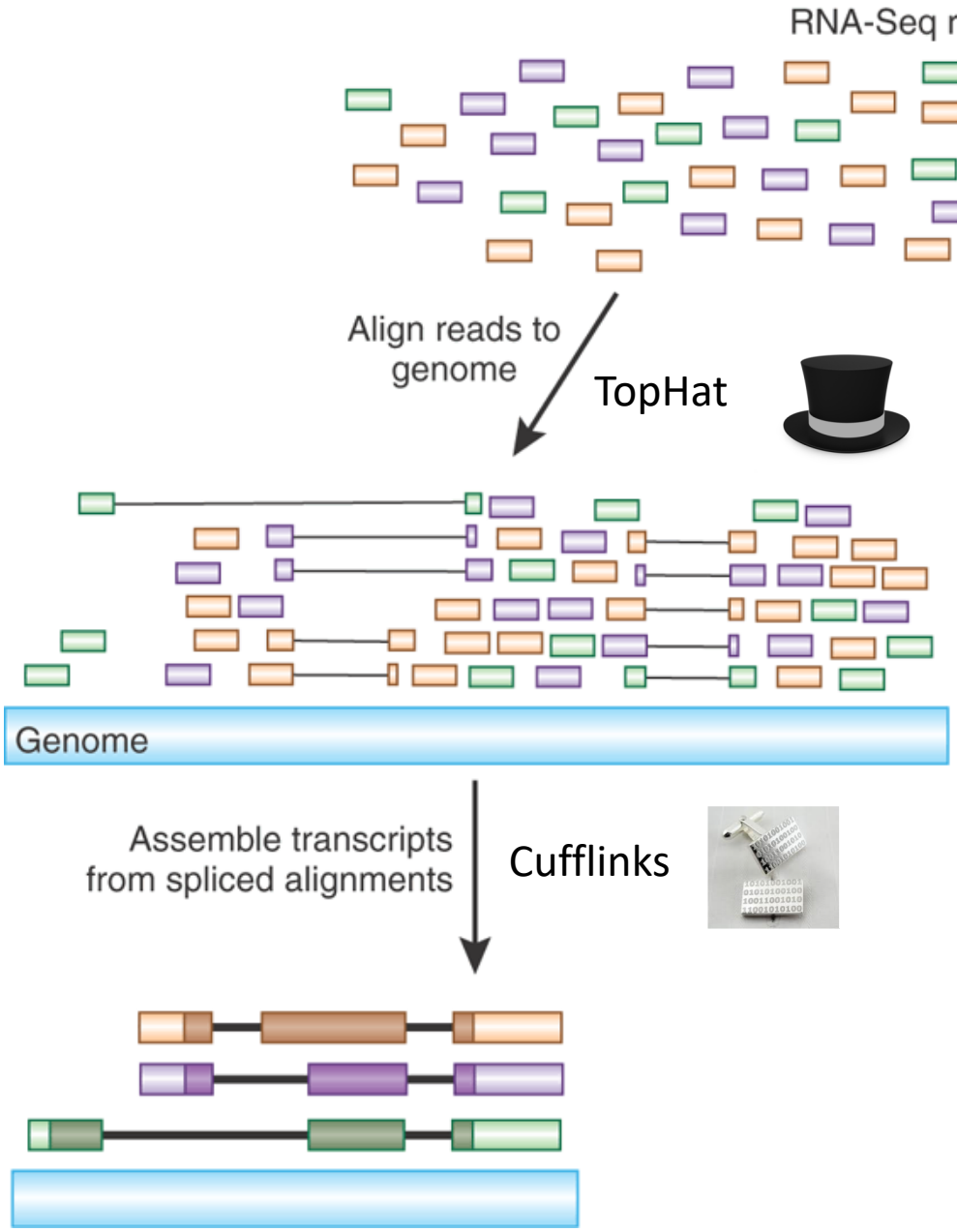
Transcript Reconstruction from RNA-Seq Reads



Transcript Reconstruction from RNA-Seq Reads



Transcript Reconstruction from RNA-Seq Reads



The Tuxedo Suite:

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

NATURE PROTOCOLS | PROTOCOL

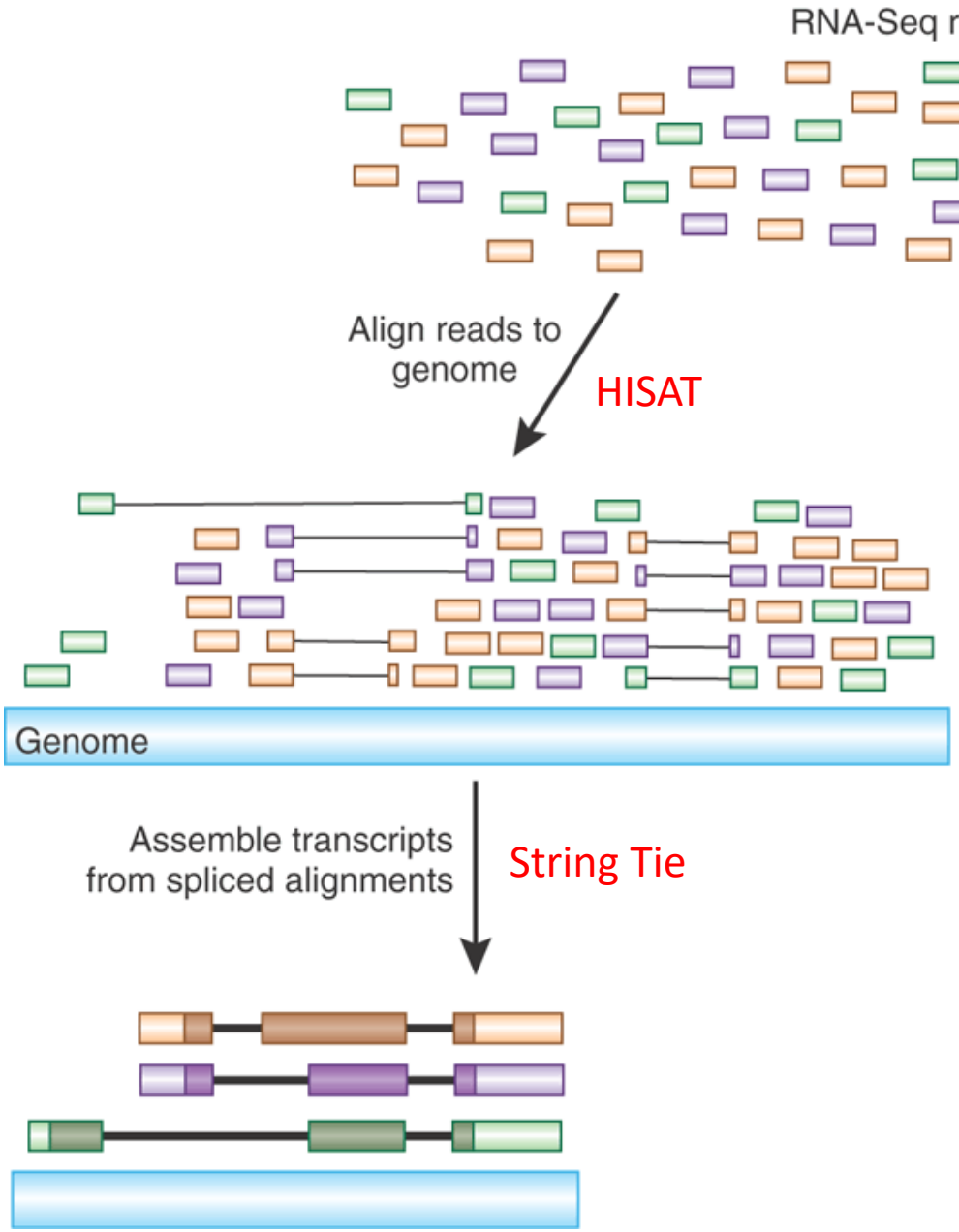
Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks

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

Transcript Reconstruction from RNA-Seq Reads



The "New Tuxedo" Suite:

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

NATURE PROTOCOLS | PROTOCOL

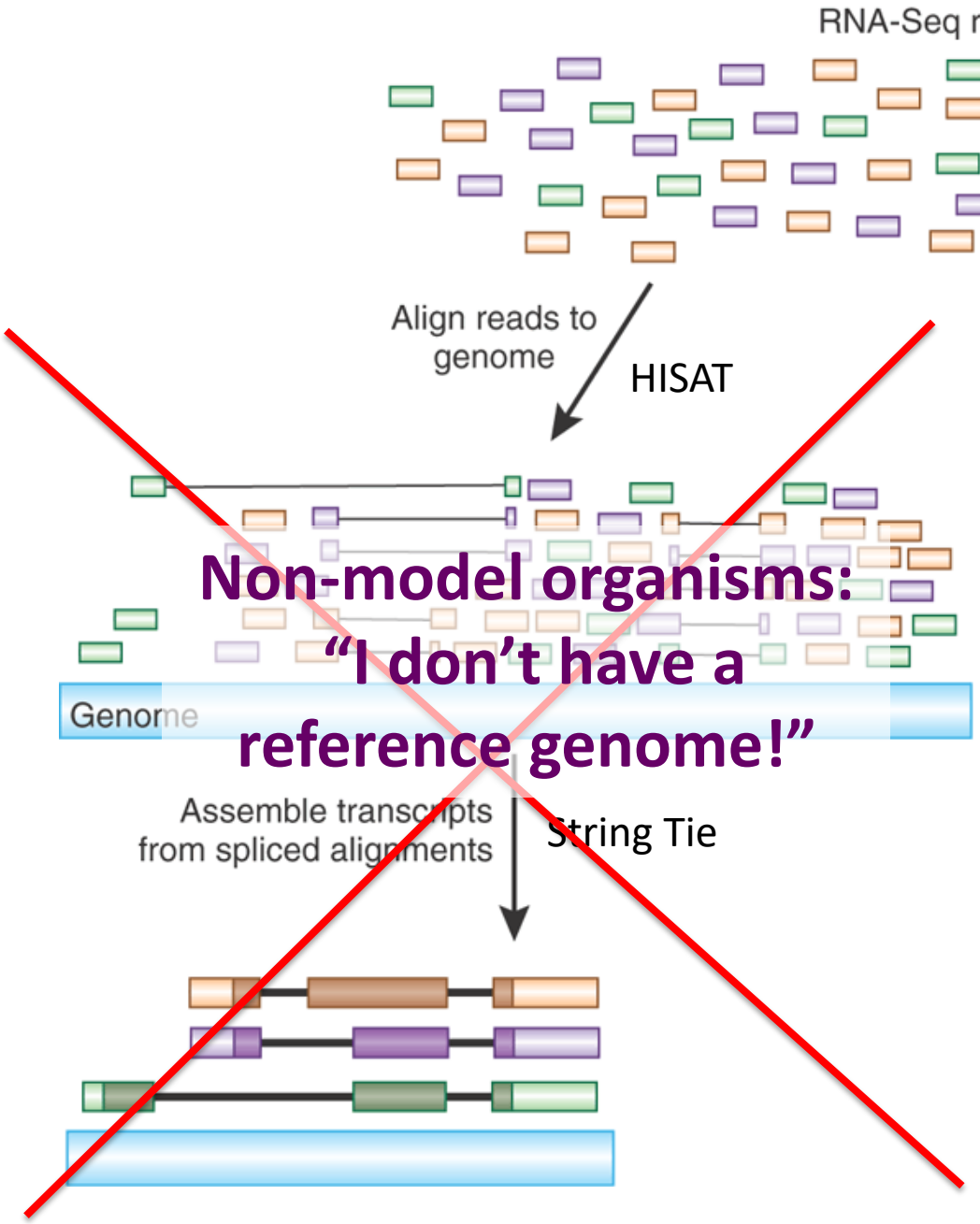
Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown

Mihaela Pertea, Daehwan Kim, Geo M Pertea, Jeffrey T Leek & Steven L Salzberg

Affiliations | Contributions | Corresponding author

Nature Protocols **11**, 1650–1667 (2016) | doi:10.1038/nprot.2016.095
Published online 11 August 2016

Transcript Reconstruction from RNA-Seq Reads



Non-model organisms: "I don't have a reference genome!"

The "New Tuxedo" Suite:

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

NATURE PROTOCOLS | PROTOCOL

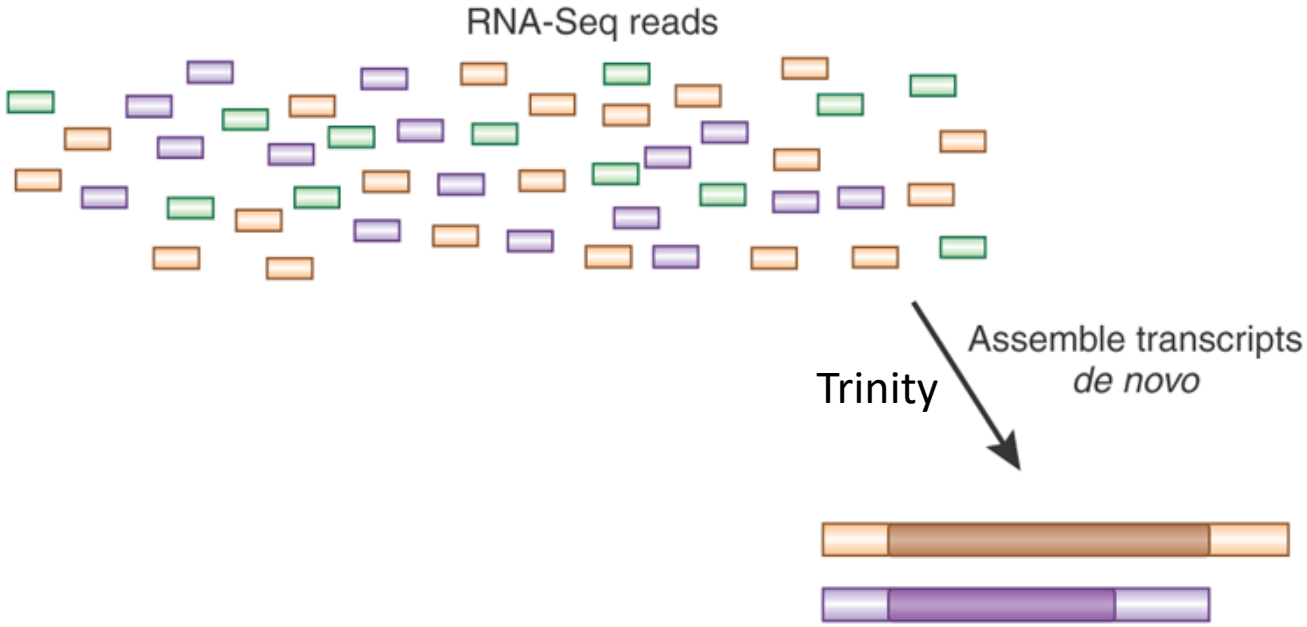
Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown

Mihaela Pertea, Daehwan Kim, Geo M Pertea, Jeffrey T Leek & Steven L Salzberg

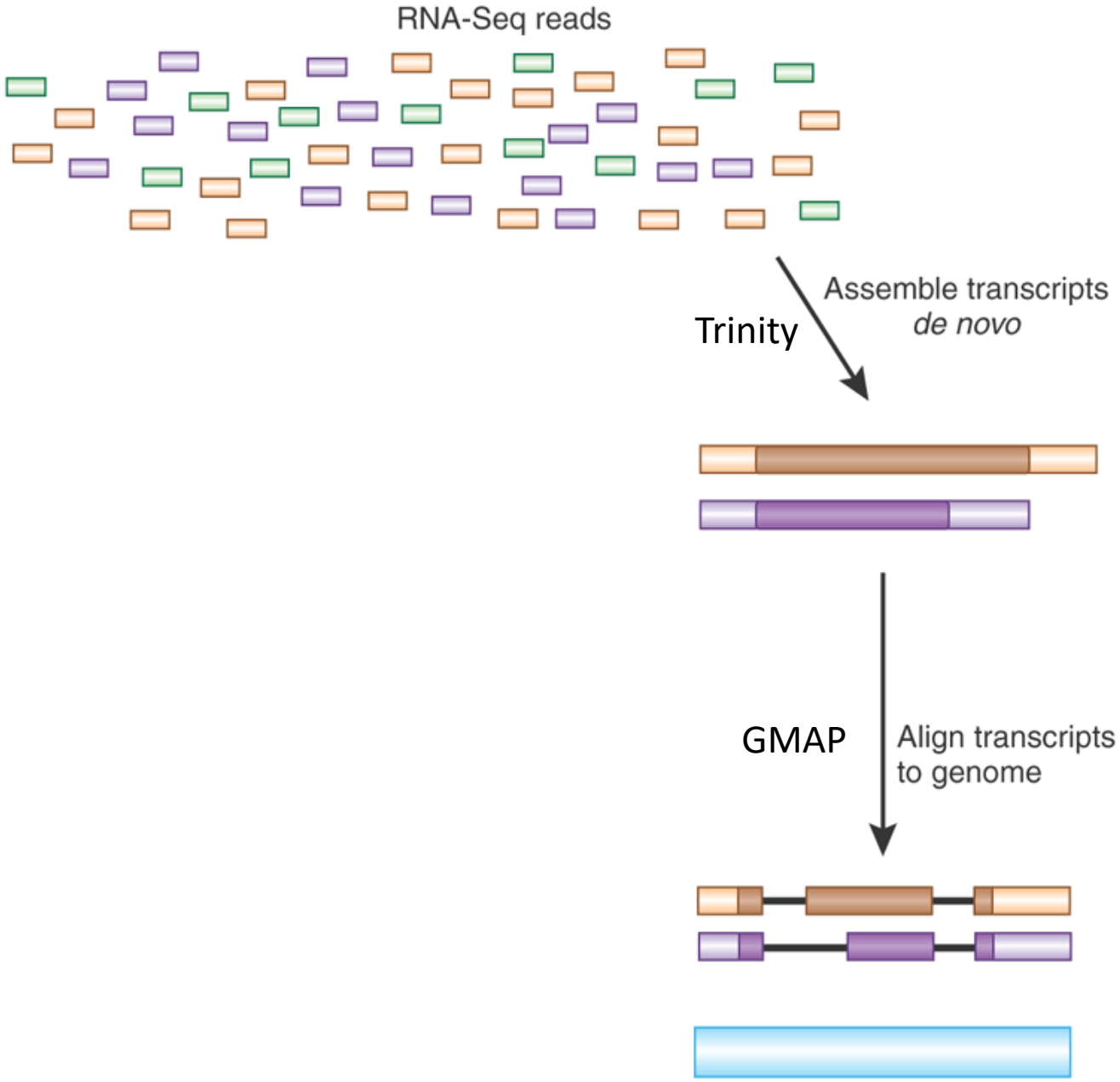
Affiliations | Contributions | Corresponding author

Nature Protocols **11**, 1650–1667 (2016) | doi:10.1038/nprot.2016.095
Published online 11 August 2016

Transcript Reconstruction from RNA-Seq Reads



Transcript Reconstruction from RNA-Seq Reads



Transcript Reconstruction from RNA-Seq Reads

End-to-end **Transcriptome**-based
RNA-Seq Analysis
Software Package

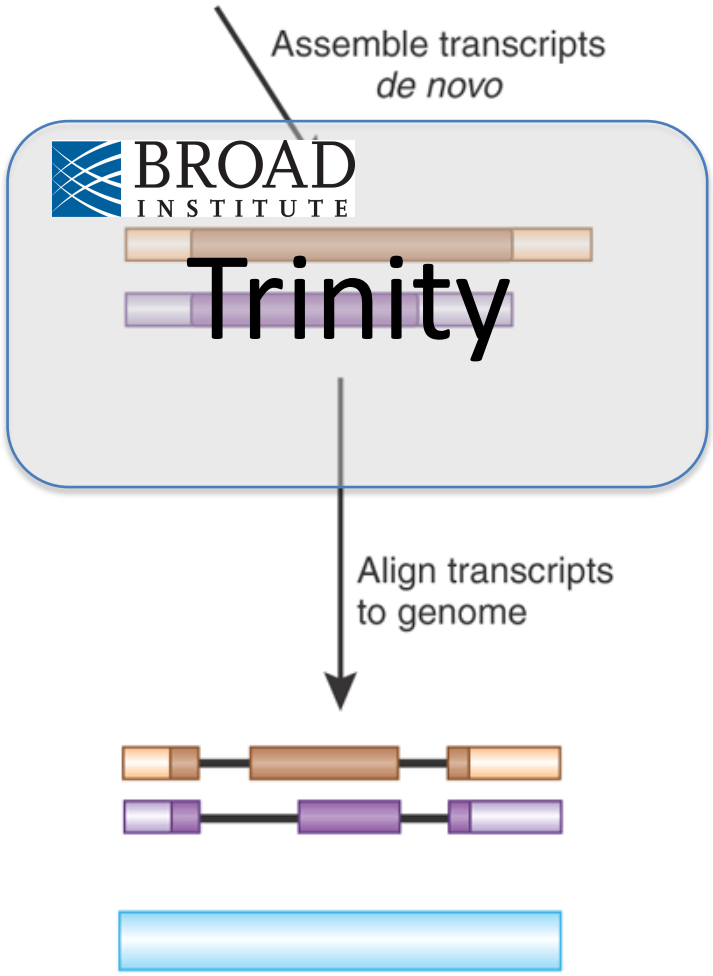
NATURE PROTOCOLS | PROTOCOL

De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis

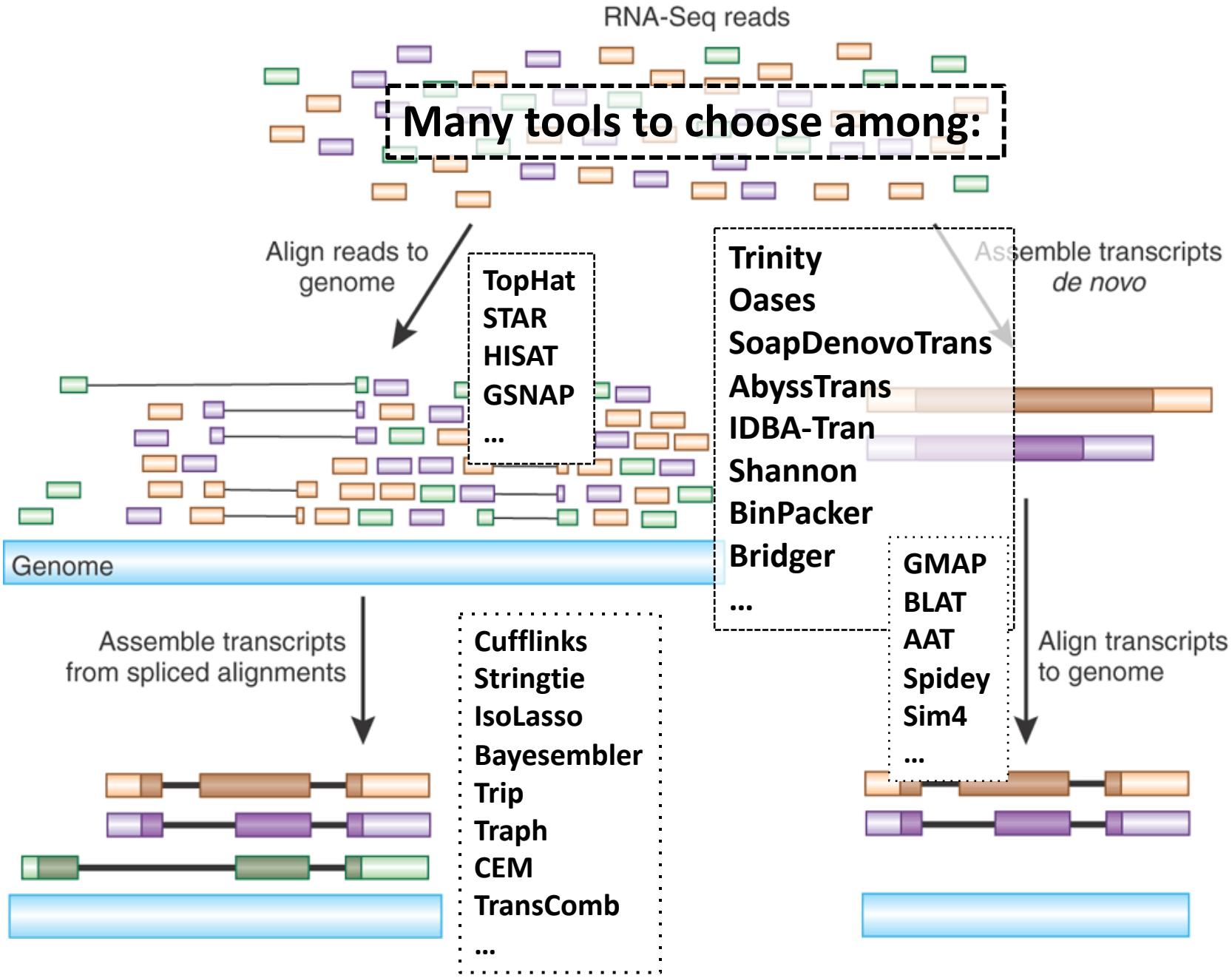
Brian J Haas, Alexie Papanicolaou, Moran Yassour, Manfred Grabherr, Philip D Blood, Joshua Bowden, Matthew Brian Couger, David Eccles, Bo Li, Matthias Lieber, Matthew D MacManes, Michael Ott, Joshua Orvis, Nathalie Pochet, Francesco Strozzi, Nathan Weeks, Rick Westerman, Thomas William, Colin N Dewey, Robert Henschel, Richard D LeDuc, Nir Friedman & Aviv Regev

Affiliations | Contributions | Corresponding authors

Nature Protocols **8**, 1494–1512 (2013) | doi:10.1038/nprot.2013.084
Published online 11 July 2013



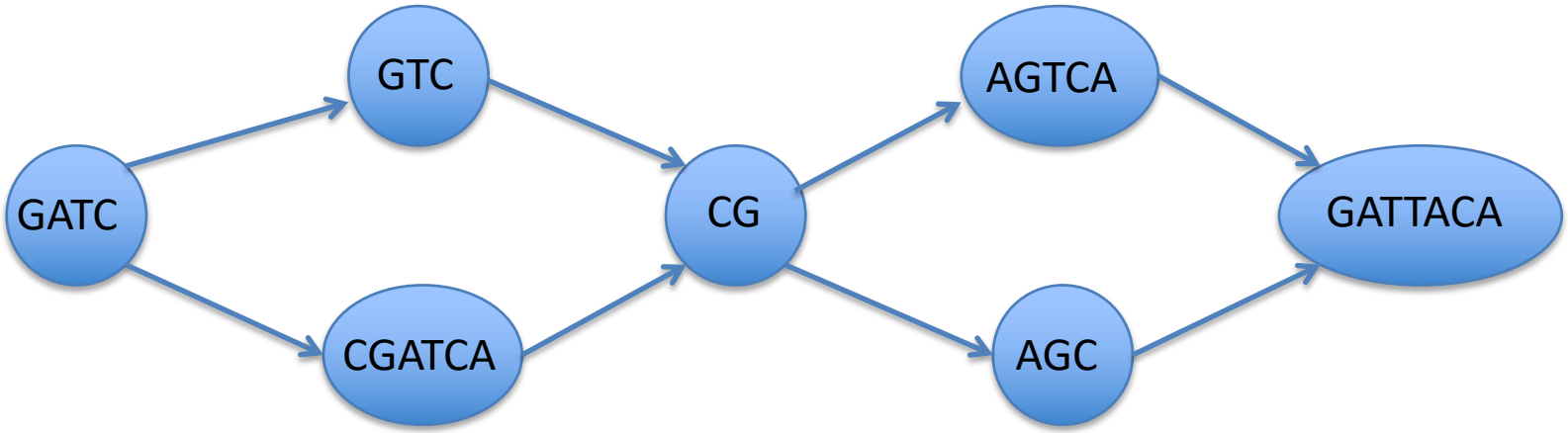
Transcript Reconstruction from RNA-Seq Reads



Graph Data Structures Commonly Used For Assembly

- Sequence
- Order
- Orientation (+, -)
- Overlap

Reads to Graph

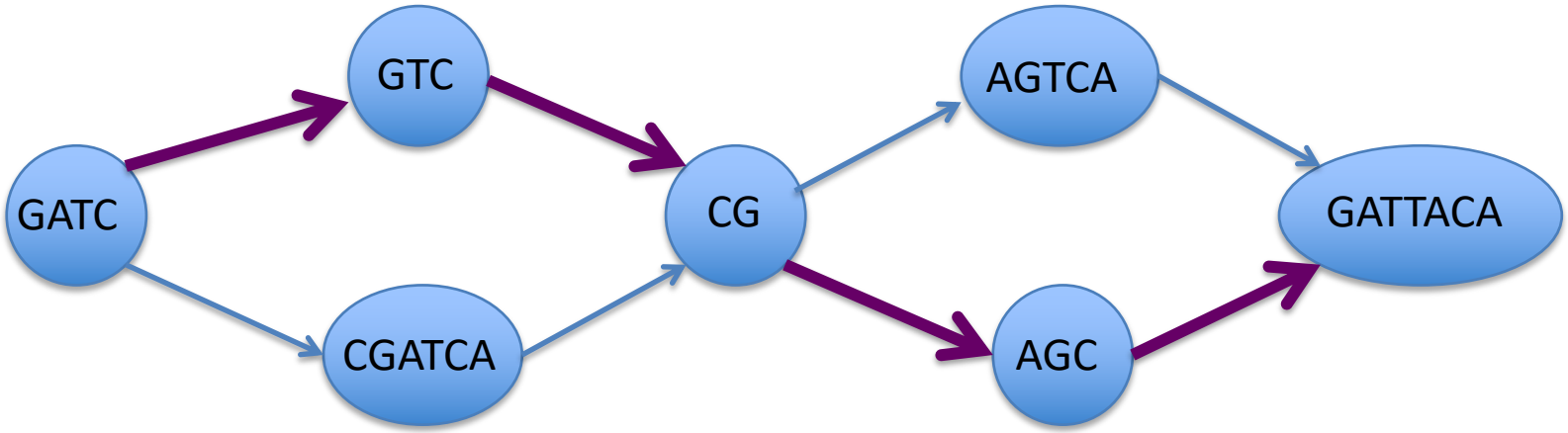


Nodes = sequence (+/-)
Edges = order, overlap

Graph Data Structures Commonly Used For Assembly

- Sequence
- Order
- Orientation (+, -)
- Overlap

Reads to Graph



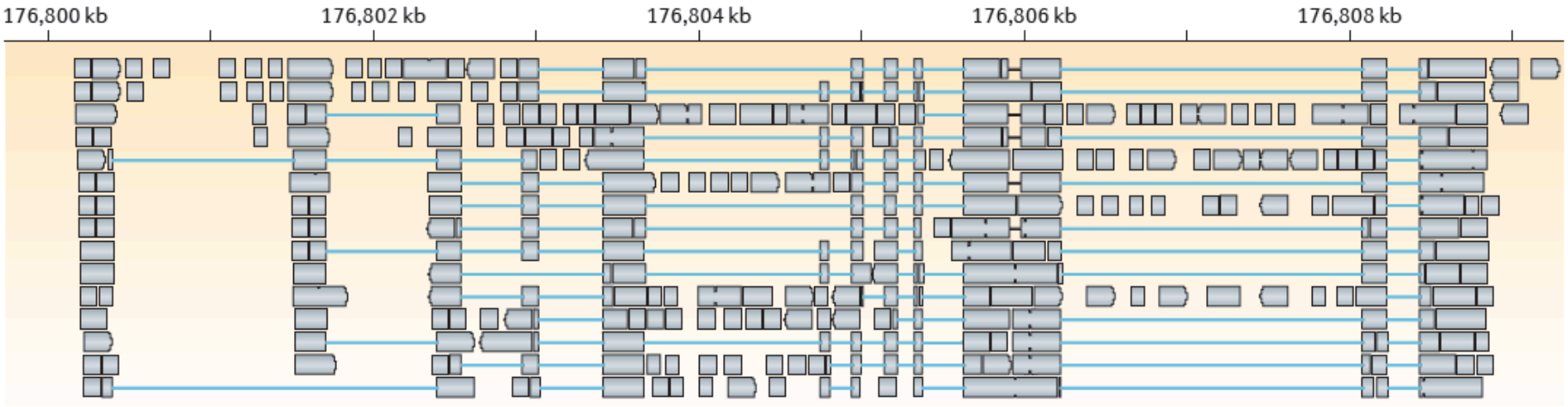
GATCGTCCGAGCGATTACA

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

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

Genome-Guided Transcript Reconstruction

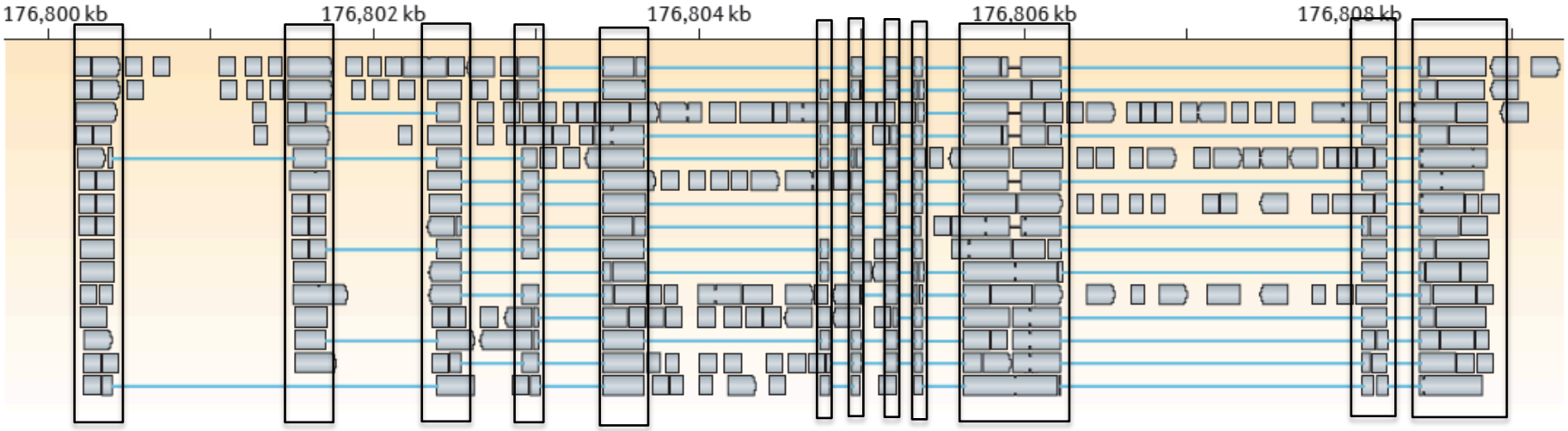
Splice-align reads to the genome



From Martin & Wang. Nature Reviews in Genetics. 2011

Genome-Guided Transcript Reconstruction

Splice-align reads to the genome

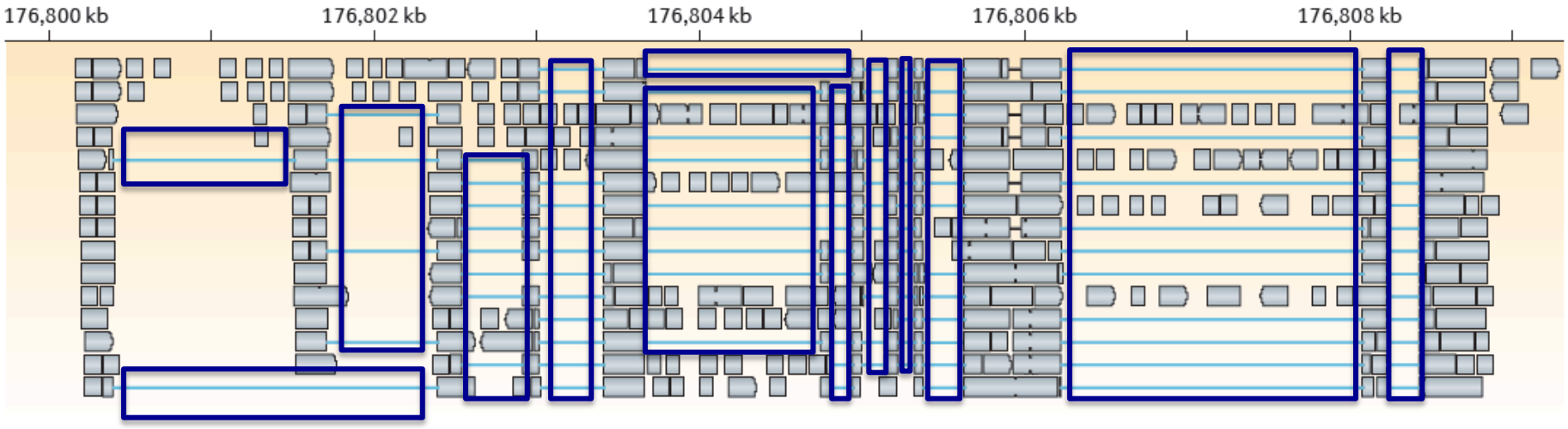


Alignment segment piles => exon regions

From Martin & Wang. Nature Reviews in Genetics. 2011

Genome-Guided Transcript Reconstruction

Splice-align reads to the genome

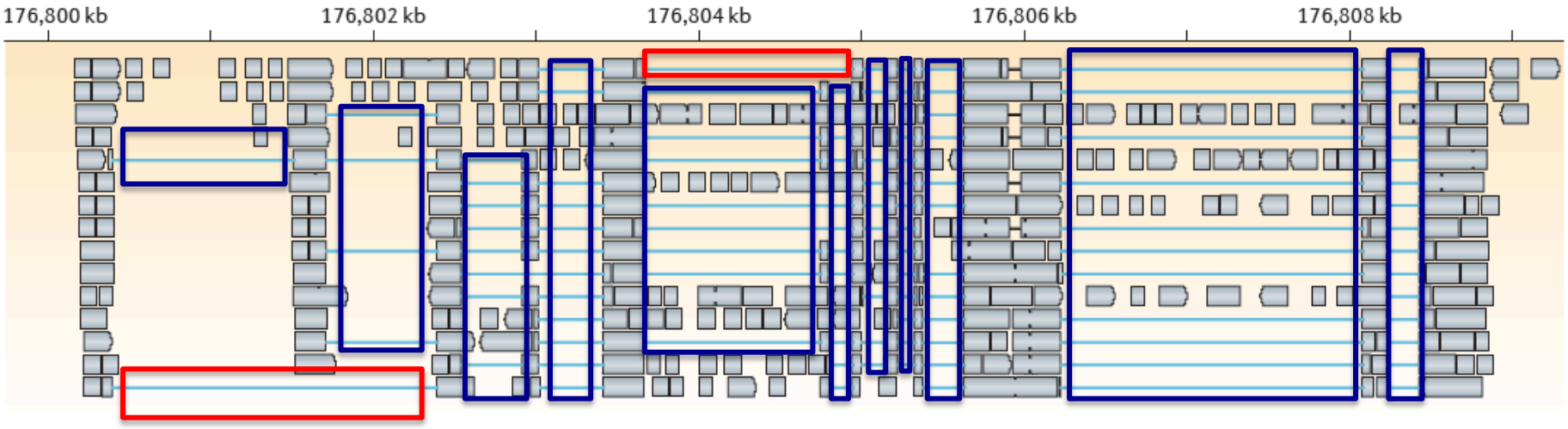


Large alignment gaps => introns

From Martin & Wang. Nature Reviews in Genetics. 2011

Genome-Guided Transcript Reconstruction

Splice-align reads to the genome

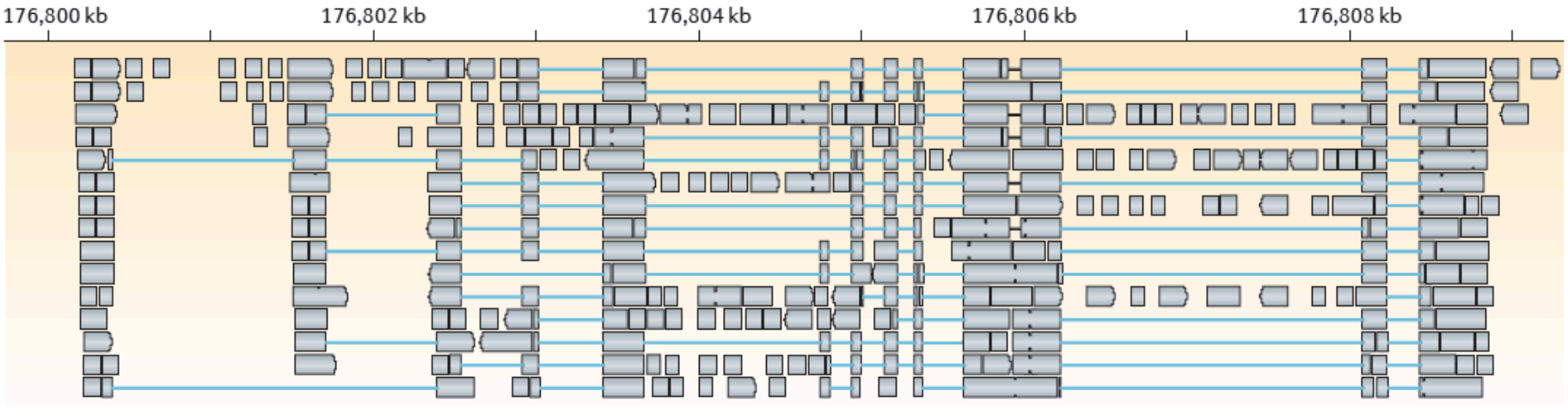


Overlapping but different introns = evidence of alternative splicing

From Martin & Wang. Nature Reviews in Genetics. 2011

Genome-Guided Transcript Reconstruction

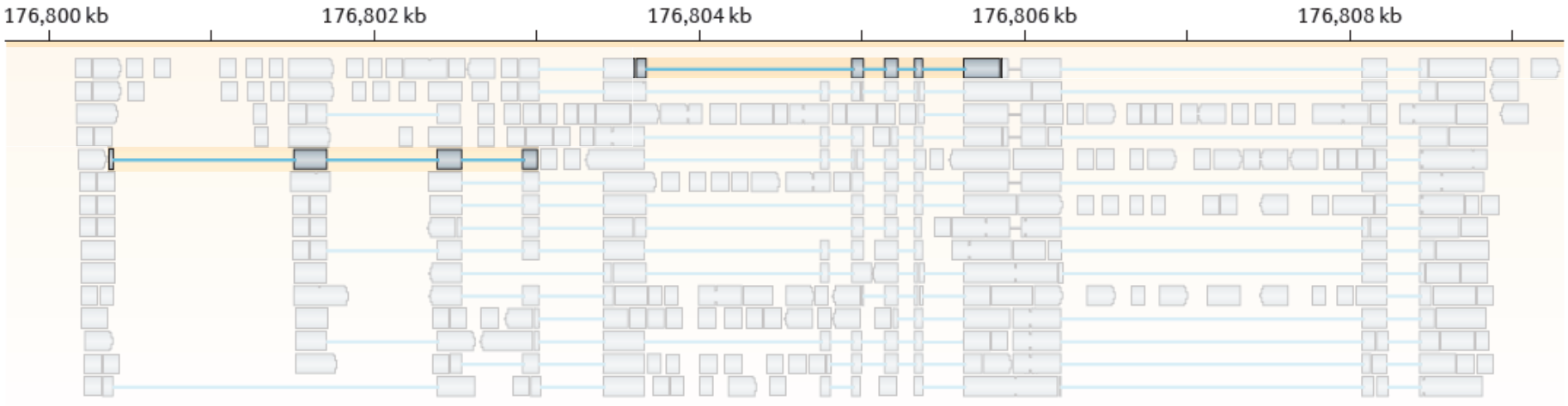
Splice-align reads to the genome



From Martin & Wang. Nature Reviews in Genetics. 2011

Genome-Guided Transcript Reconstruction

Splice-align reads to the genome

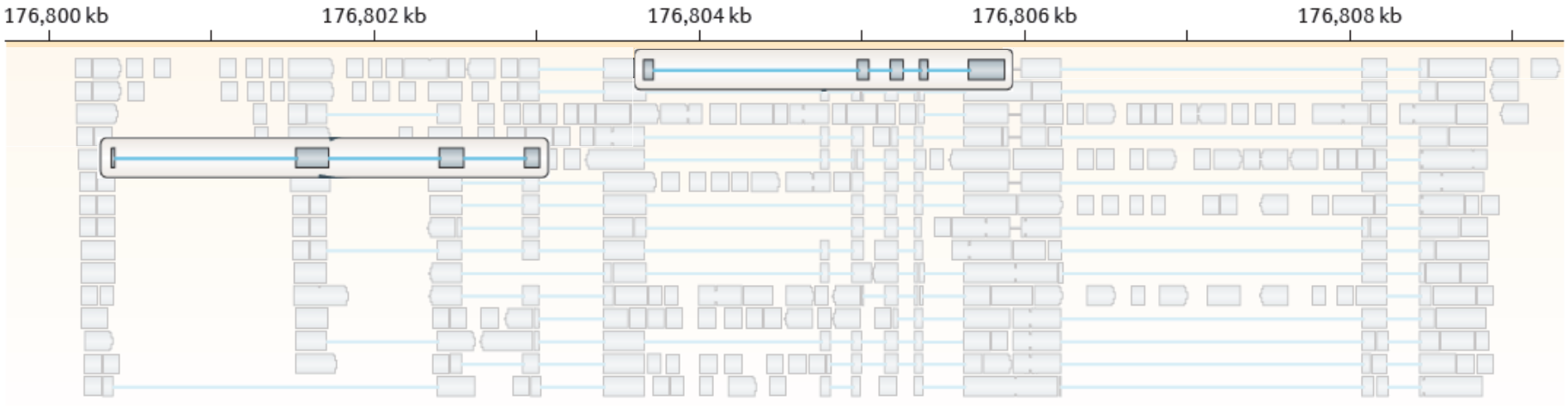


Individual reads can yield multiple exon and intron segments (splice patterns)

From Martin & Wang. Nature Reviews in Genetics. 2011

Genome-Guided Transcript Reconstruction

Splice-align reads to the genome

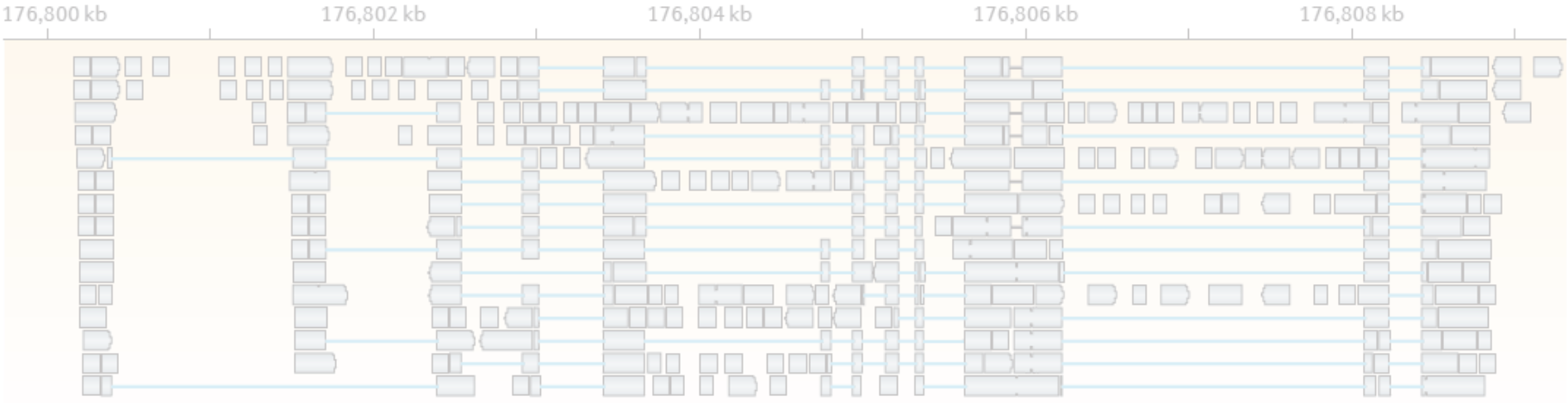


Nodes = unique splice patterns

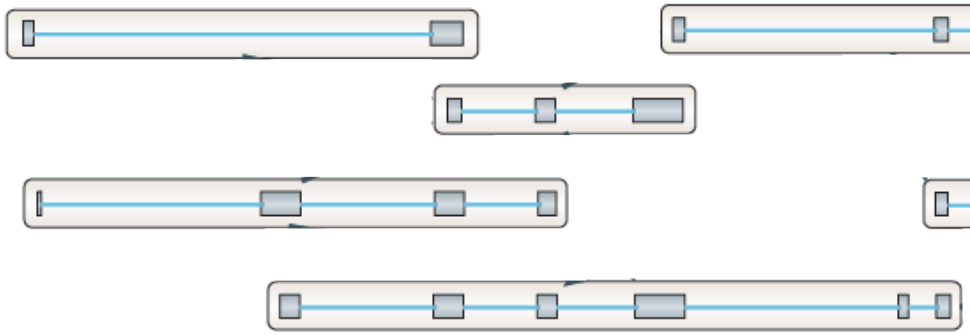
From Martin & Wang. Nature Reviews in Genetics. 2011

Genome-Guided Transcript Reconstruction

Splice-align reads to the genome



Construct graph from unique splice patterns of aligned reads.

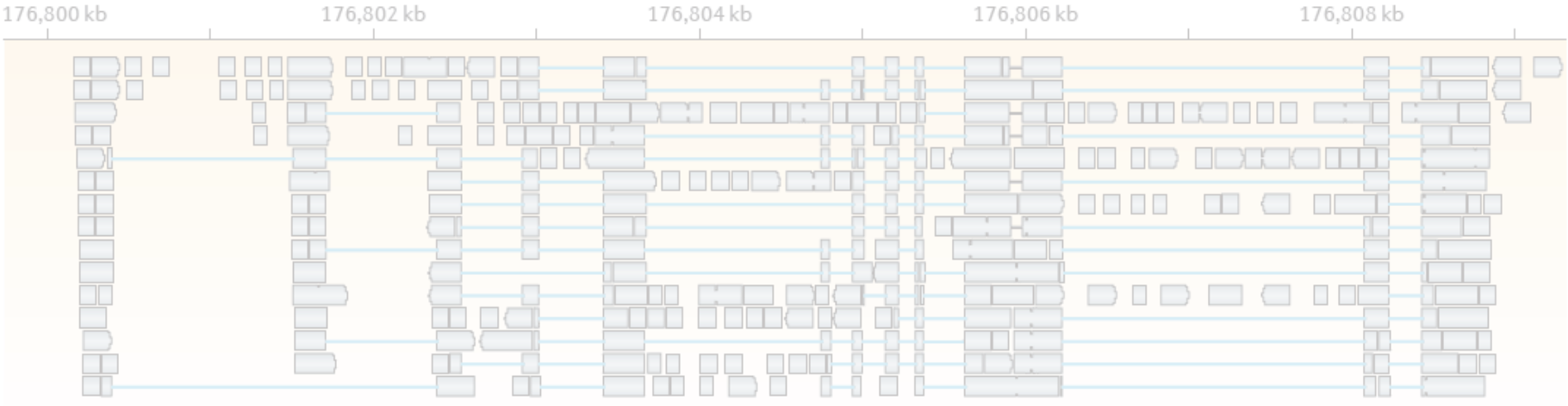


Nodes = unique splice patterns

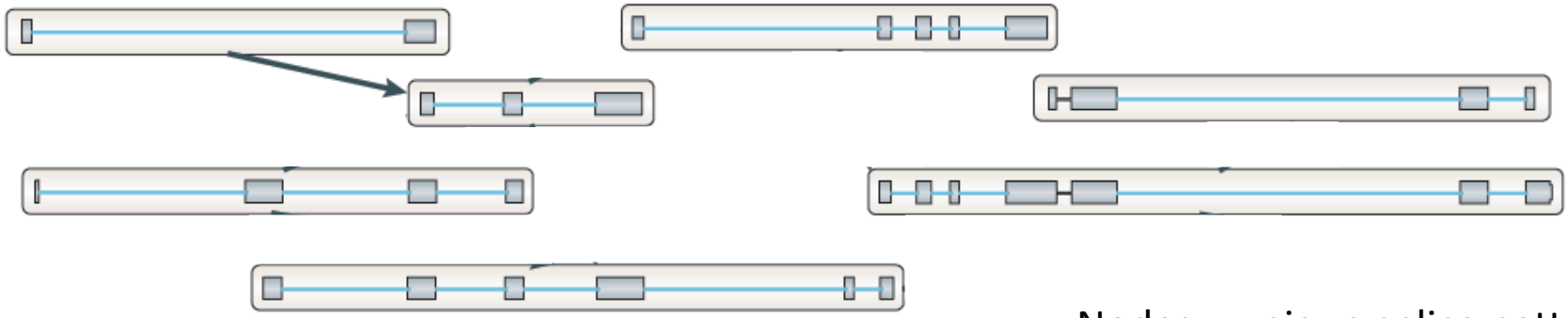
From Martin & Wang. Nature Reviews in Genetics. 2011

Genome-Guided Transcript Reconstruction

Splice-align reads to the genome



Construct graph from unique splice patterns of aligned reads.

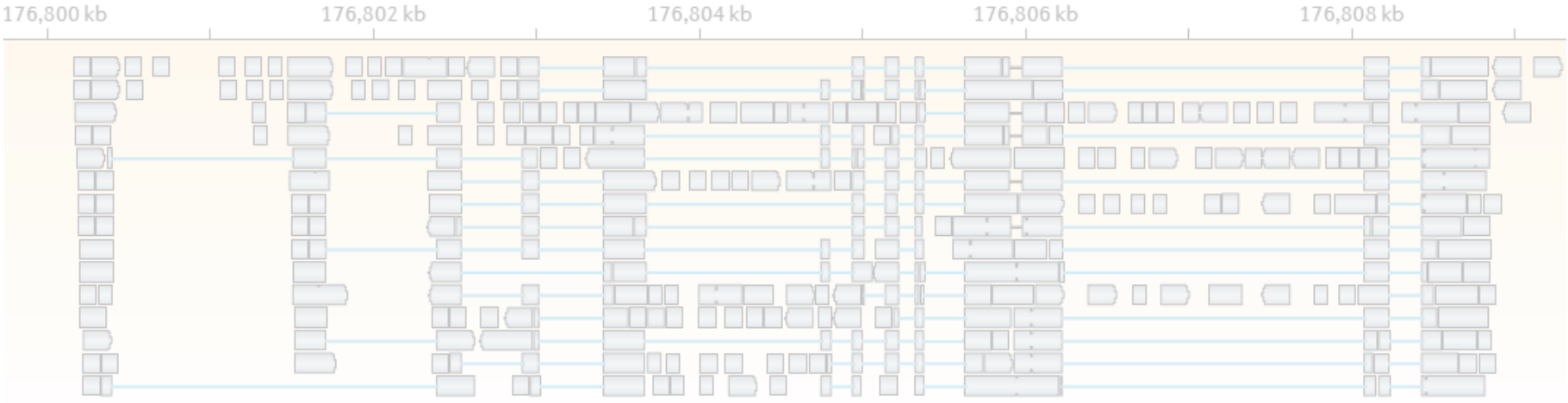


Nodes = unique splice patterns
Edges = compatible patterns

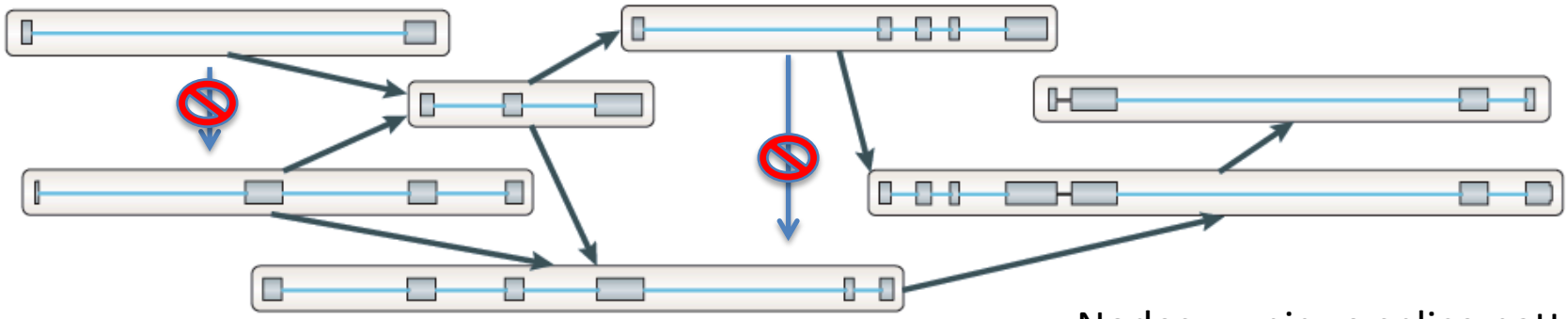
From Martin & Wang. Nature Reviews in Genetics. 2011

Genome-Guided Transcript Reconstruction

Splice-align reads to the genome



Construct graph from unique splice patterns of aligned reads.

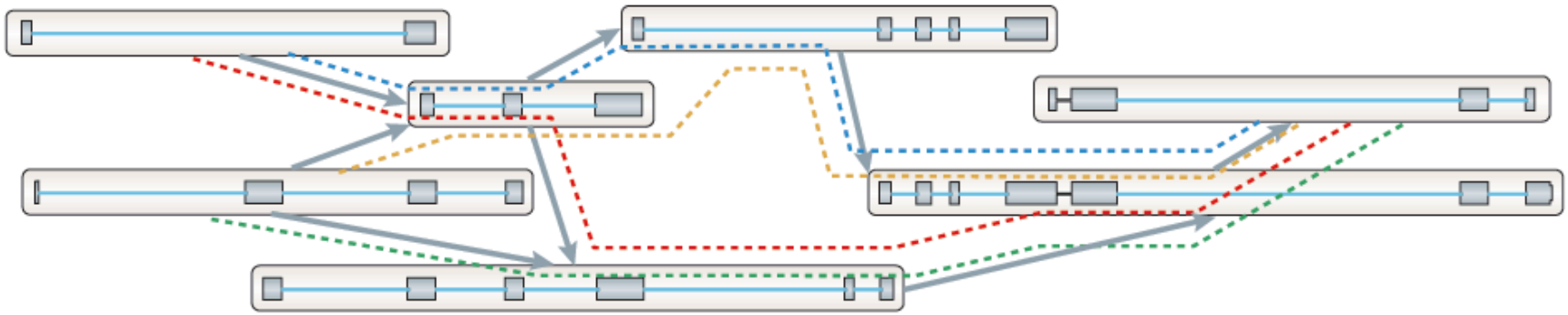


Nodes = unique splice patterns
Edges = compatible patterns

From Martin & Wang. Nature Reviews in Genetics. 2011

Genome-Guided Transcript Reconstruction

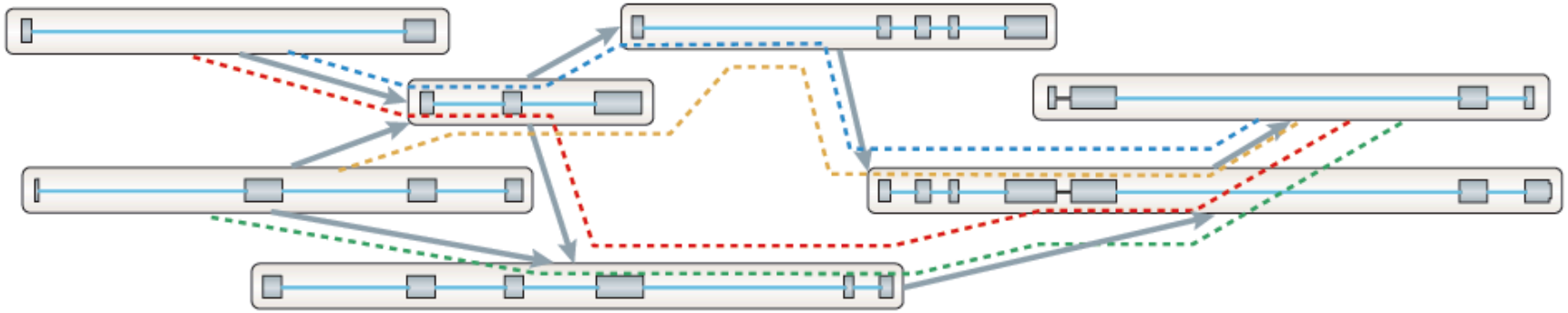
Traverse paths through the graph to assemble transcript isoforms



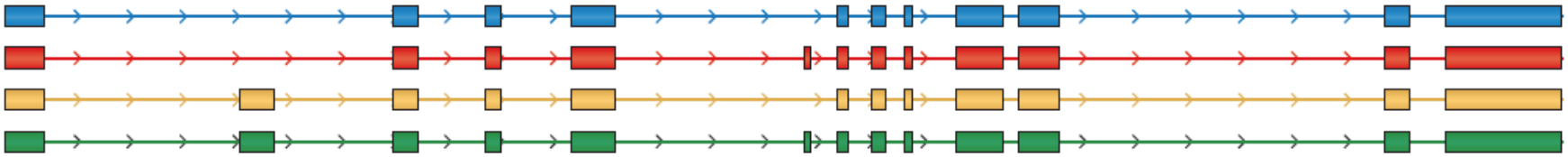
From Martin & Wang. Nature Reviews in Genetics. 2011

Genome-Guided Transcript Reconstruction

Traverse paths through the graph to assemble transcript isoforms



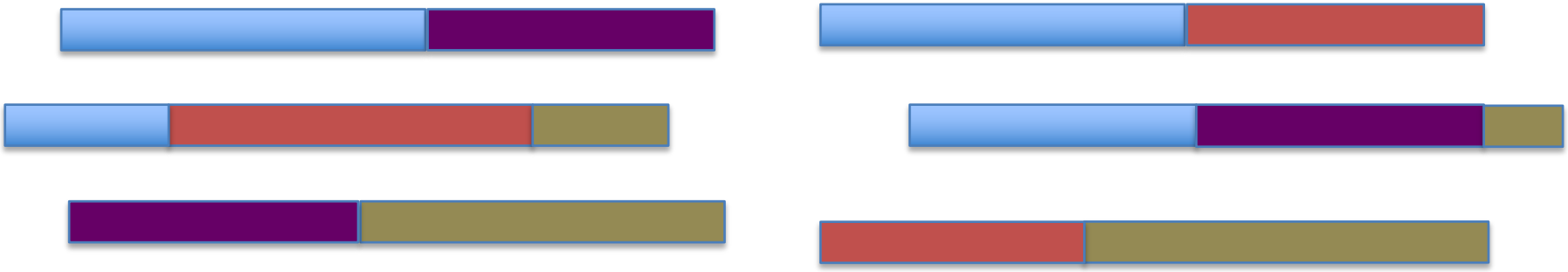
Reconstructed isoforms



From Martin & Wang. Nature Reviews in Genetics. 2011

What if you don’t have a high quality reference genome sequence?

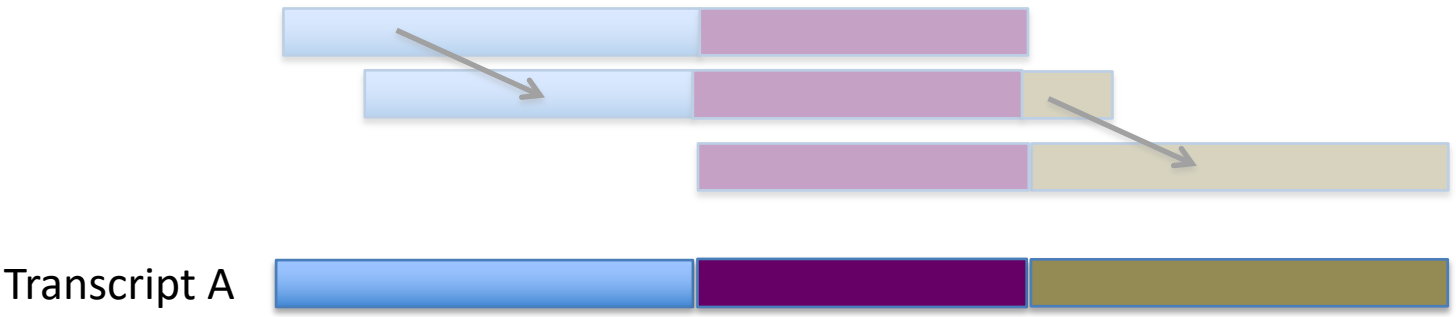
Read Overlap Graph: Reads as nodes, overlaps as edges



Read Overlap Graph: Reads as nodes, overlaps as edges

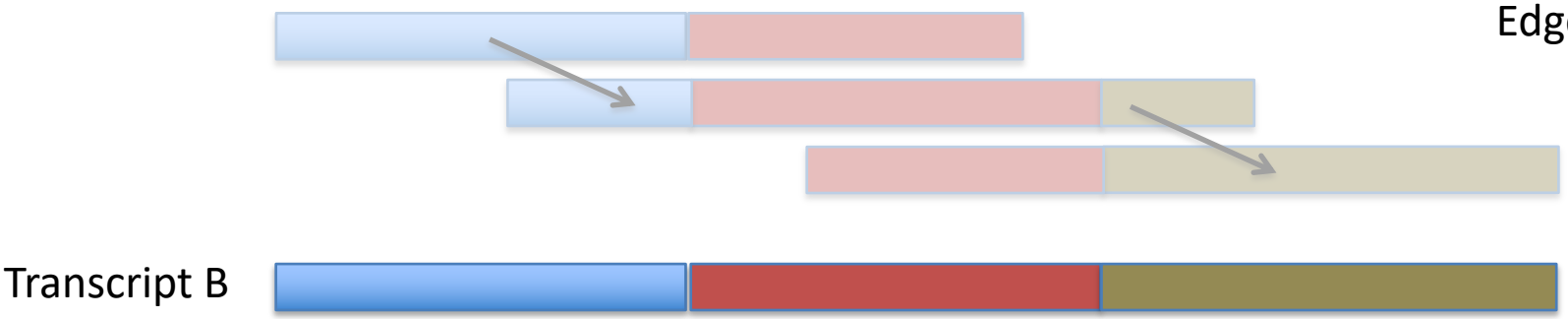
Node = read
Edge = overlap

Read Overlap Graph: Reads as nodes, overlaps as edges

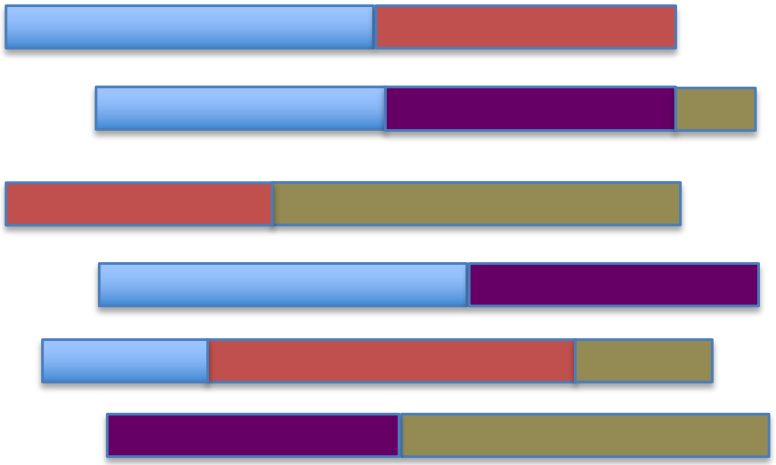


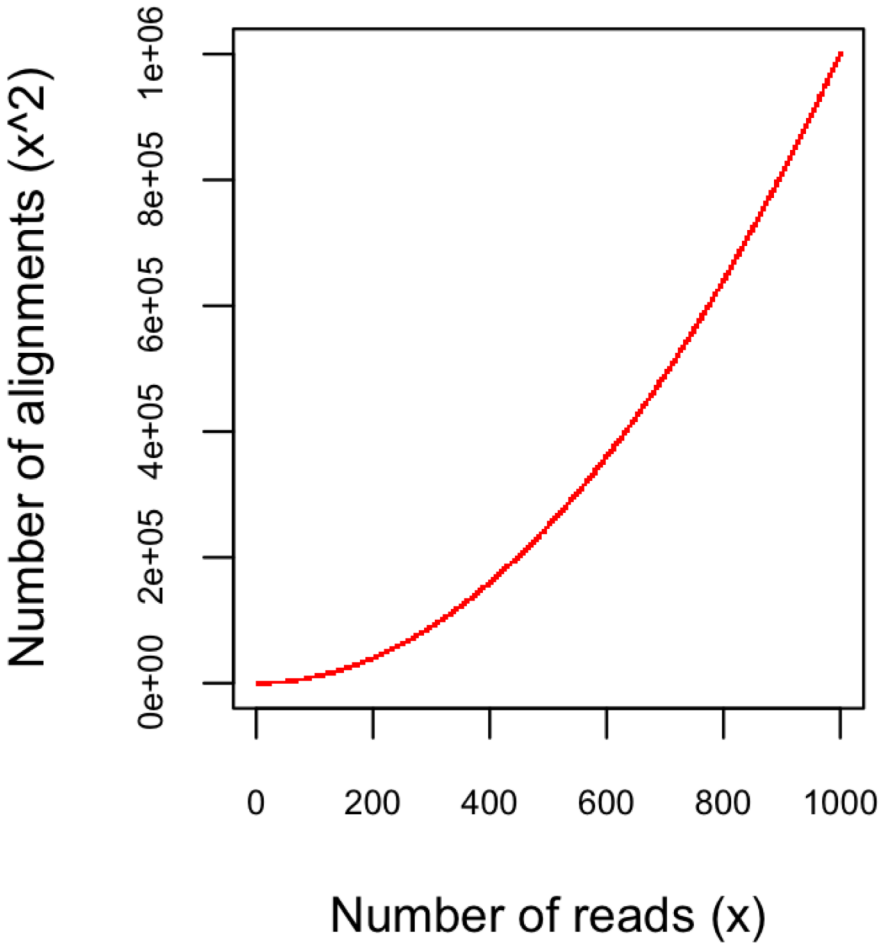
Generate consensus sequence where reads overlap

Node = read
Edge = overlap



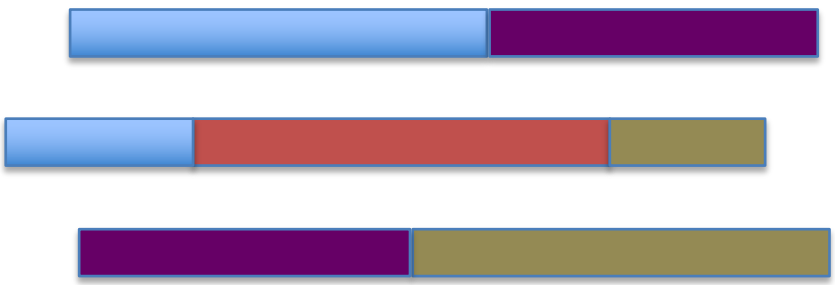
Finding pairwise overlaps between *n* reads involves ~ n^2 comparisons.



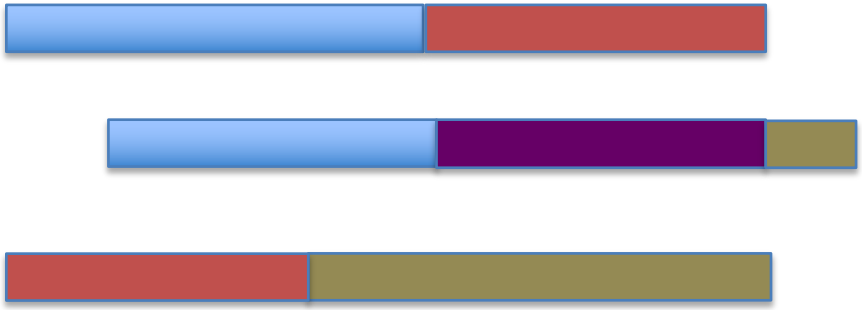


Impractical for typical RNA-Seq data (50M reads)

No genome to align to…



De novo assembly required

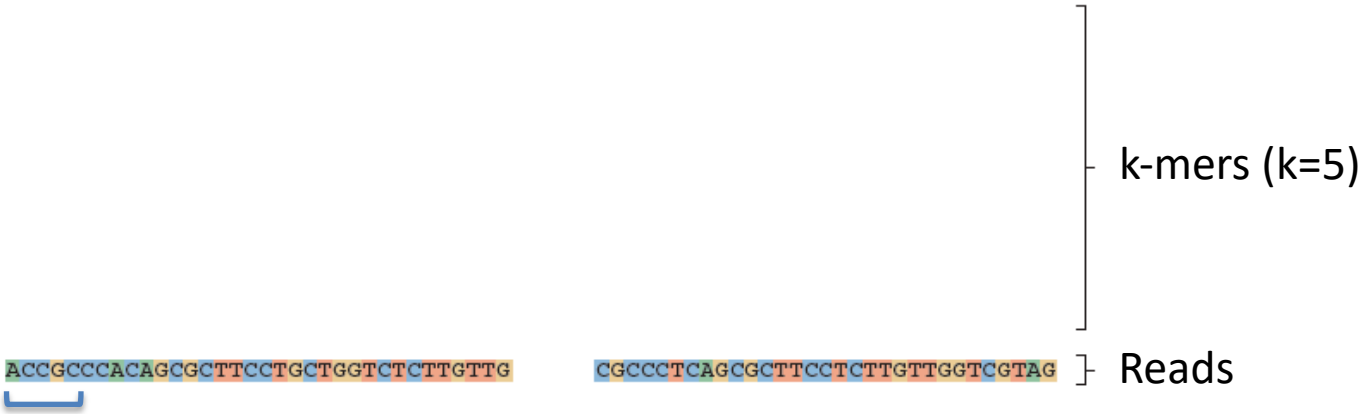


Want to avoid n^2 read alignments to define overlaps

Use a de Bruijn graph

Sequence Assembly via de Bruijn Graphs

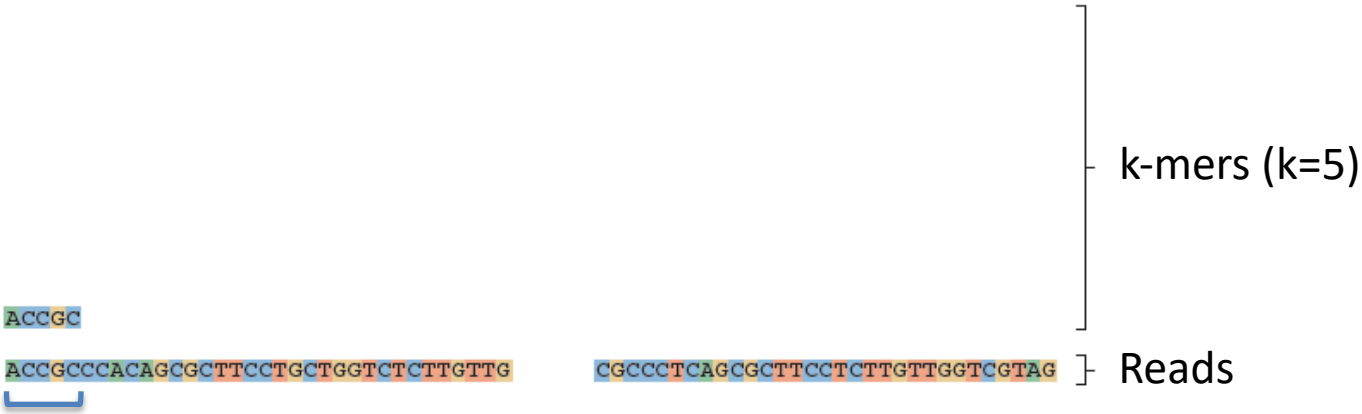
Generate all substrings of length k from the reads



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

Sequence Assembly via De Bruijn Graphs

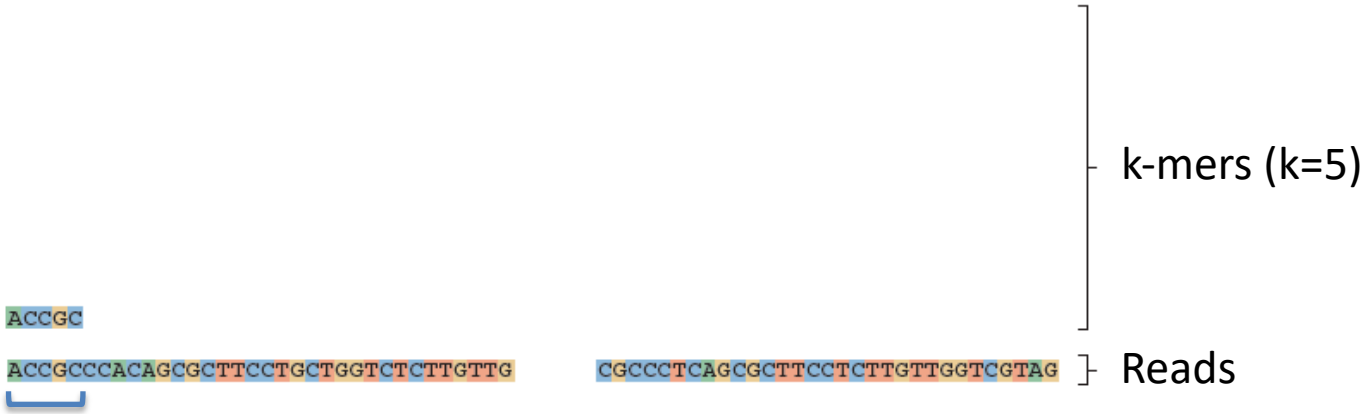
Generate all substrings of length k from the reads



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

Sequence Assembly via De Bruijn Graphs

Generate all substrings of length k from the reads



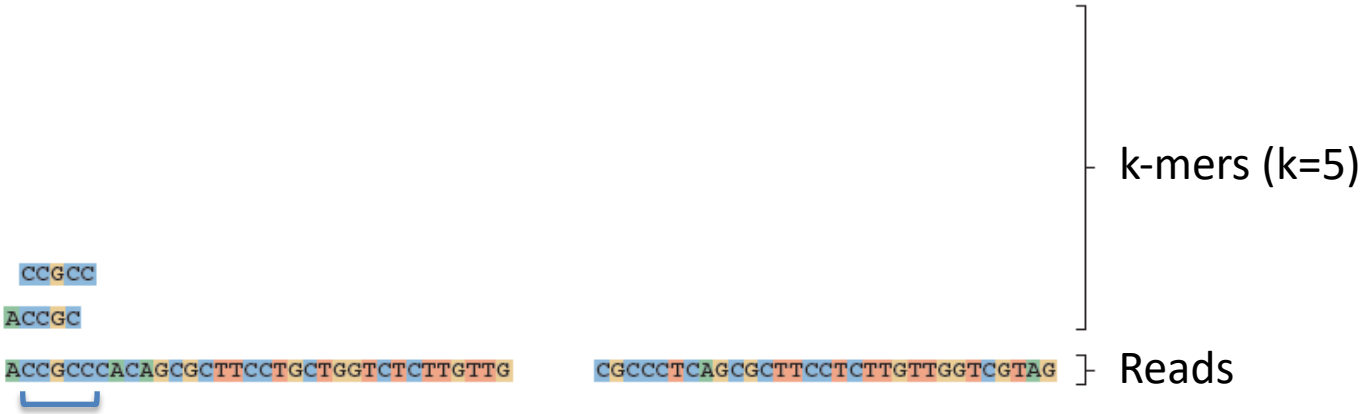
Construct the de Bruijn graph

Nodes = unique k-mers

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

Sequence Assembly via De Bruijn Graphs

Generate all substrings of length k from the reads



Construct the de Bruijn graph

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

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

Sequence Assembly via De Bruijn Graphs

Generate all substrings of length k from the reads

Construct the de Bruijn graph

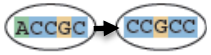
From Martin & Wang, Nat. Rev. Genet. 2011

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

Sequence Assembly via De Bruijn Graphs

Generate all substrings of length k from the reads

Construct the de Bruijn graph

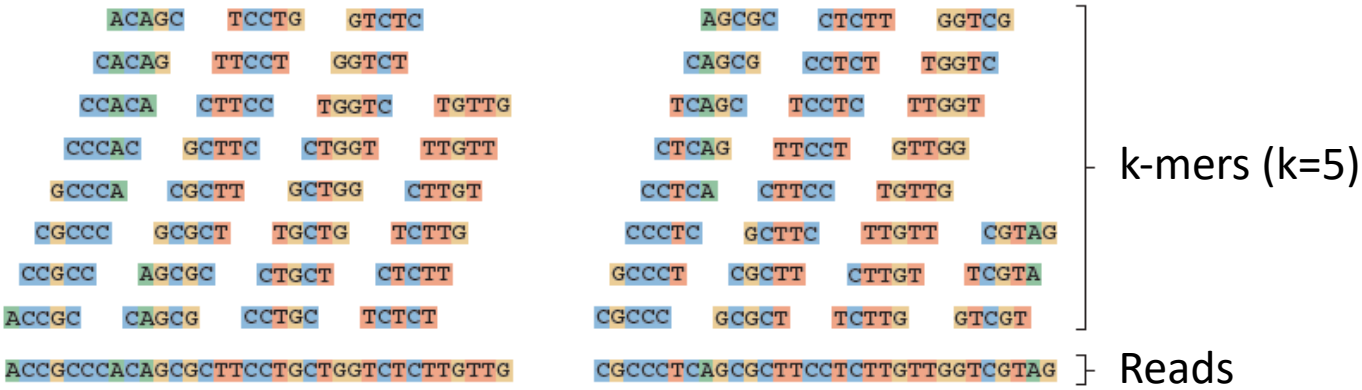


From Martin & Wang, Nat. Rev. Genet. 2011

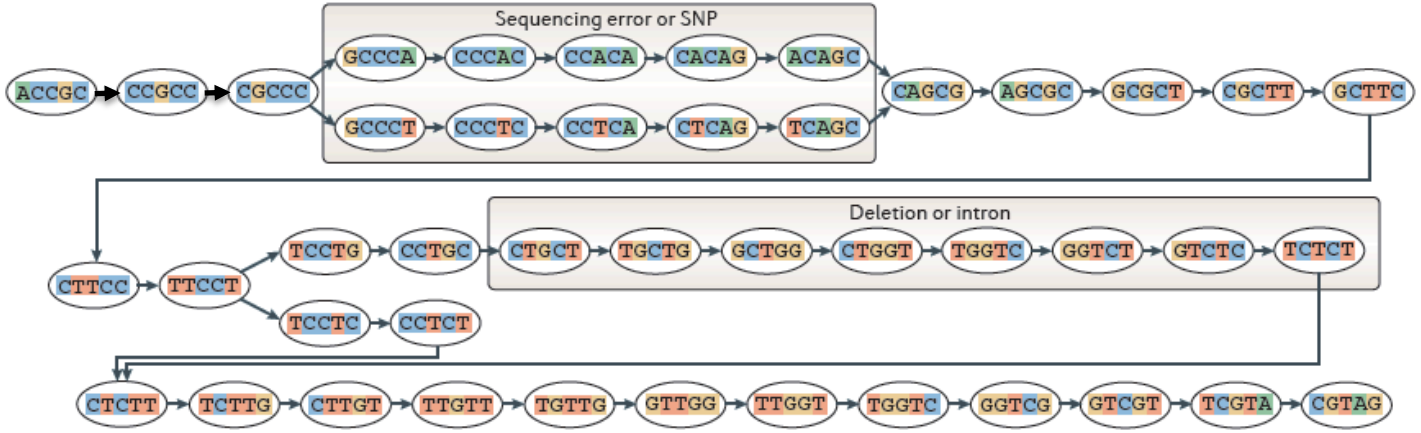
Nodes = unique k-mers
Edges = overlap by (k-1)

Sequence Assembly via De Bruijn Graphs

Generate all substrings of length k from the reads



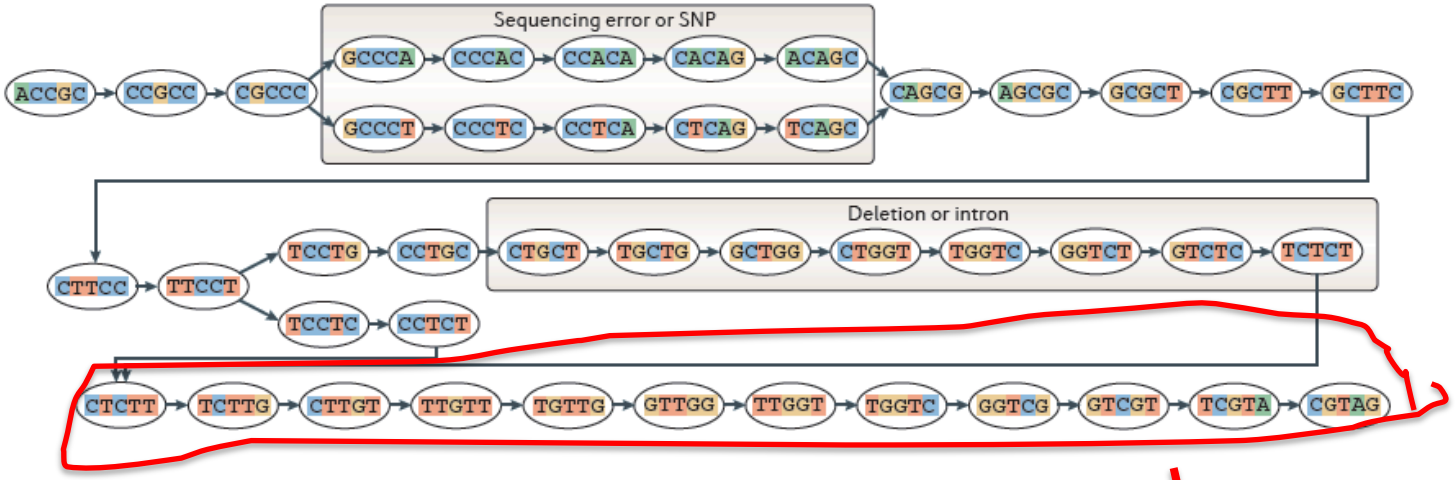
Construct the de Bruijn graph



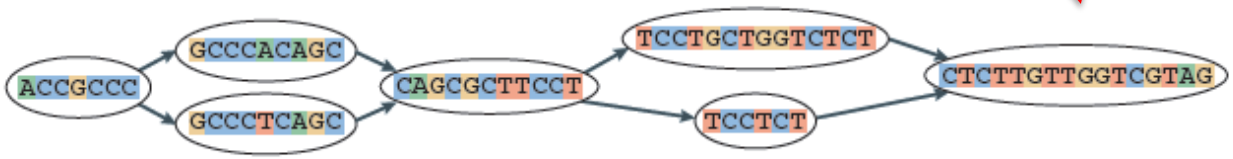
Nodes = unique k-mers
Edges = overlap by (k-1)

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

Construct the de Bruijn graph

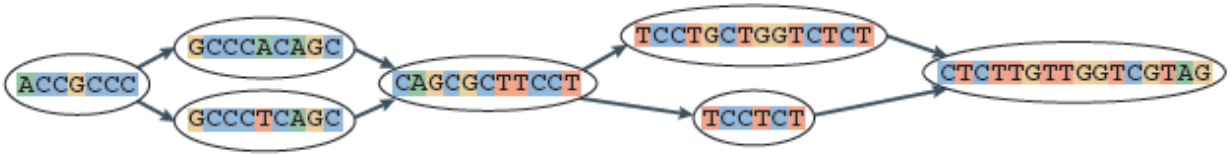


Collapse the de Bruijn graph

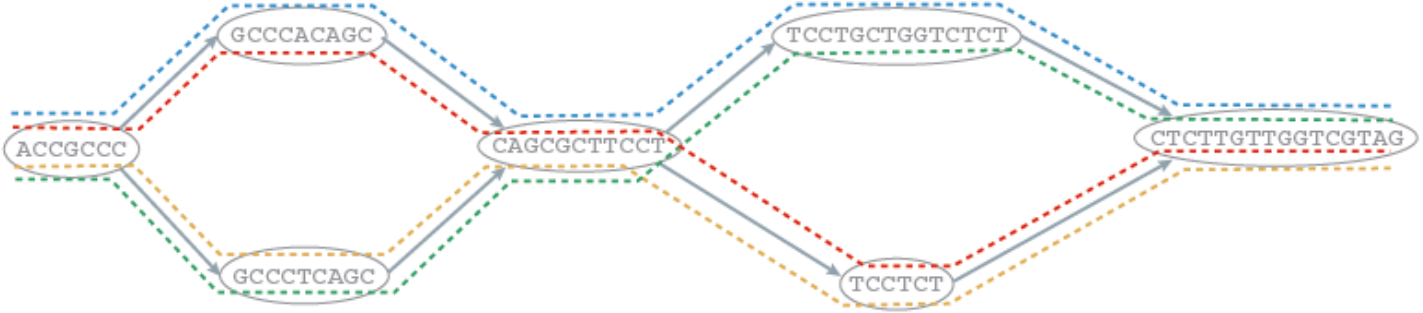


From Martin & Wang, Nat. Rev. Genet. 2011

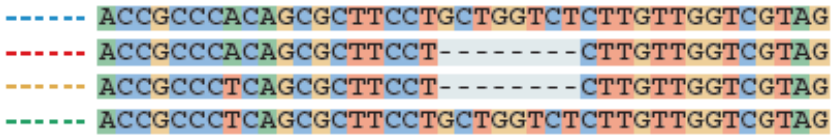
Collapse the de Bruijn graph



Traverse the graph



Assemble Transcript Isoforms



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

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 – 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...

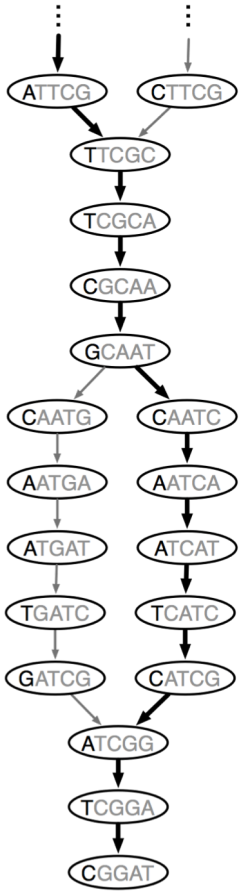
Thousands of disjoint graphs

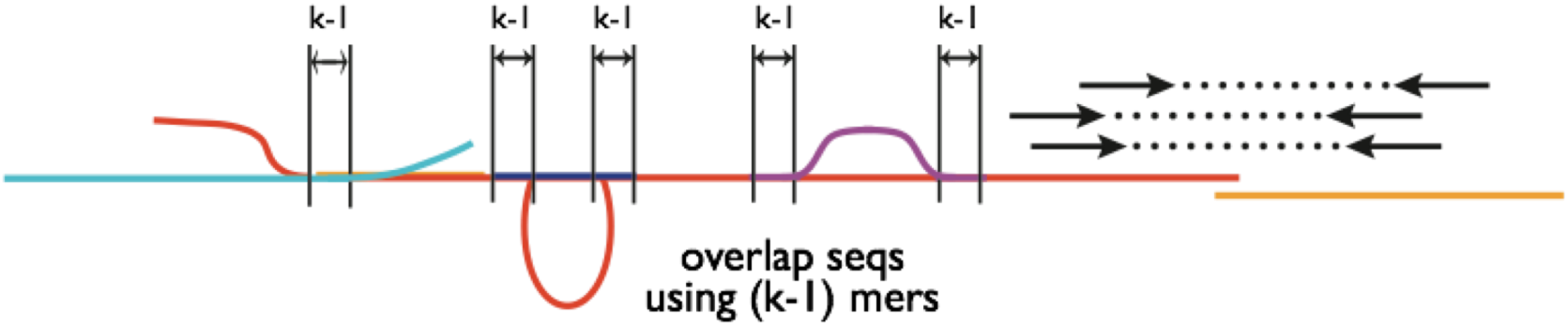
Chrysalis

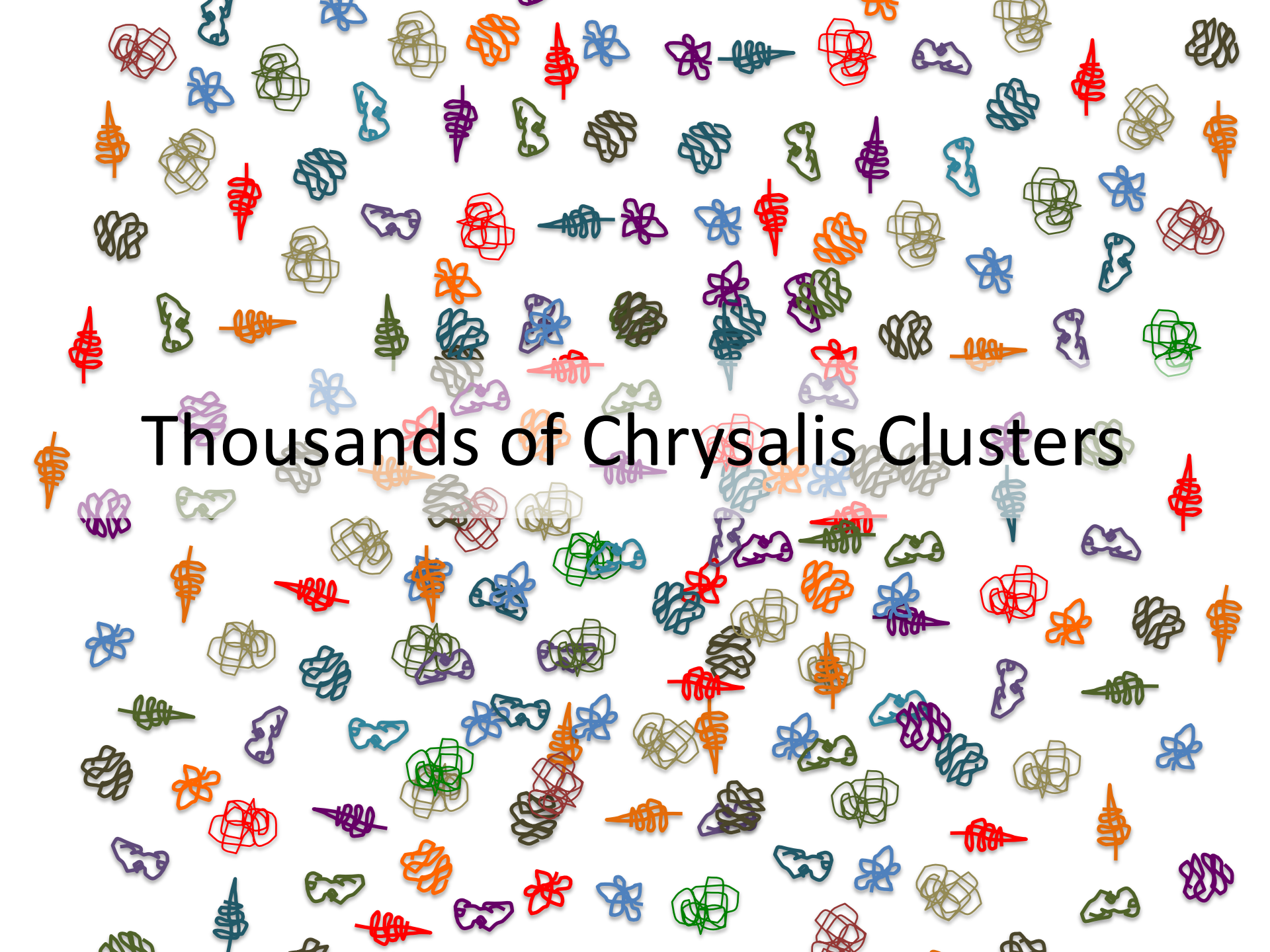
>a121:len=5845
>a122:len=2560
>a123:len=4443
>a124:len=48
>a125:len=8876
>a126:len=66

Integrate isoforms
via k-1 overlaps

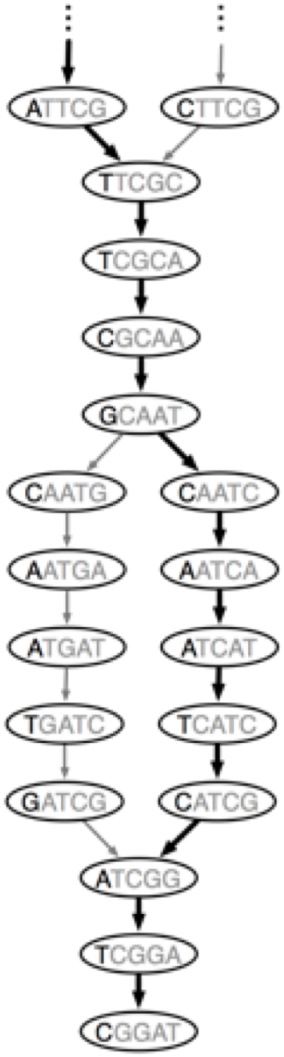
Build de Bruijn Graphs
(ideally, one per gene)





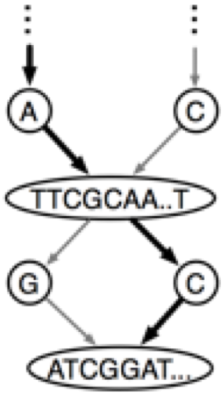

Thousands of Chrysalis Clusters

Butterfly



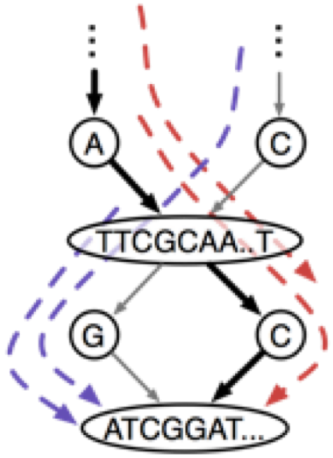
de Bruijn graph

compacting



compact graph

finding paths



compact graph with reads

extracting sequences

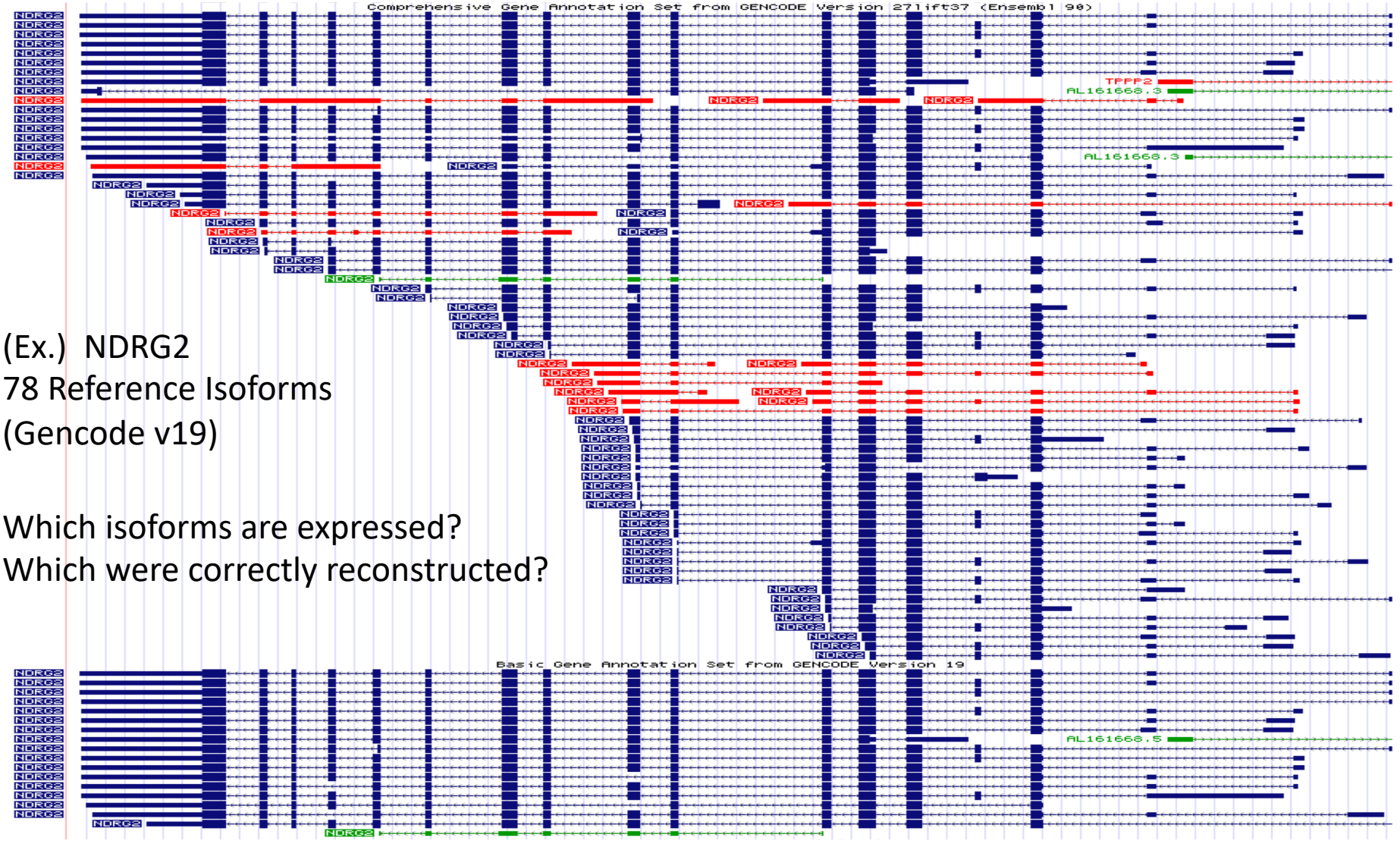
..CTTCGCAA..TGATCGGAT...

..ATTCGCAA..TCATCGGAT...

sequences (isoforms and paralogs)

Trinity output: A multi-fasta file

Reconstructing the correct isoforms from short reads can be quite challenging.



(Ex.) NDRG2
78 Reference Isoforms
(Gencode v19)

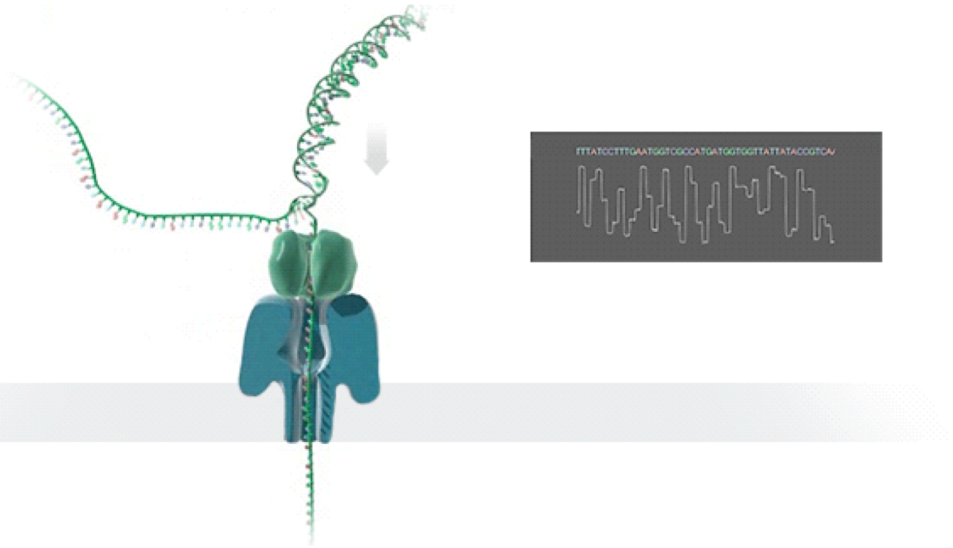
Which isoforms are expressed?
Which were correctly reconstructed?

Long reads to the rescue

PACBIO®

Single Molecule Real Time (SMRT) Sequencing Technology

Oxford NANOPORE Technologies



Want to learn more about Trinity?

nature protocols

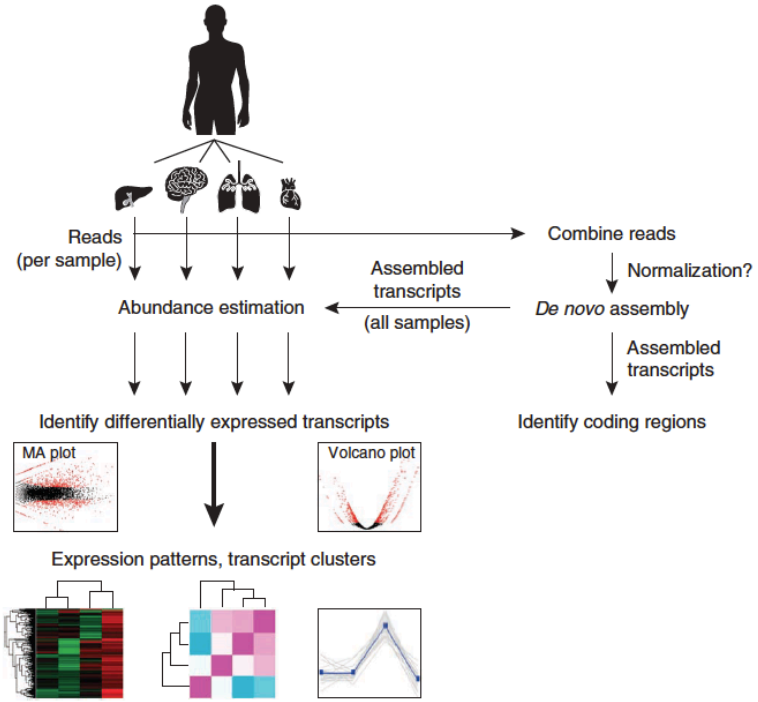
NATURE PROTOCOLS | PROTOCOL

De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis

Brian J Haas, Alexie Papanicolaou, Moran Yassour, Manfred Grabherr, Philip D Blood, Joshua Bowden, Matthew Brian Couger, David Eccles, Bo Li, Matthias Lieber, Matthew D MacManes, Michael Ott, Joshua Orvis, Nathalie Pochet, Francesco Strozzi, Nathan Weeks, Rick Westerman, Thomas William, Colin N Dewey, Robert Henschel, Richard D LeDuc, Nir Friedman & Aviv Regev

Affiliations | Contributions | Corresponding authors

Nature Protocols **8**, 1494–1512 (2013) | doi:10.1038/nprot.2013.084
Published online 11 July 2013



RNA-Seq De novo Assembly Using Trinity

Quick Guide for the Impatient

Trinity assembles transcript sequences from Illumina RNA-Seq data.

Download Trinity here.

Build Trinity by typing 'make' in the base installation directory.

Assemble RNA-Seq data like so:

```
Trinity --seqType fq --left reads_1.fq --right reads_2.fq --CPU 6 --max_memory 20G
```

Find assembled transcripts as: 'trinity_out_dir/Trinity.fasta'

Use the documentation links in the right-sidebar to navigate this documentation, and contact our Google group for technical support.

Workshop Practical

- Genome-guided Transcript Reconstruction
  - Tuxedo2 approach: Hisat2 & Stringtie

- Genome-free de novo reconstruction
  - Trinity

To begin, visit:

https://github.com/trinityrnaseq/CSHLadvSeqTech2017/wiki