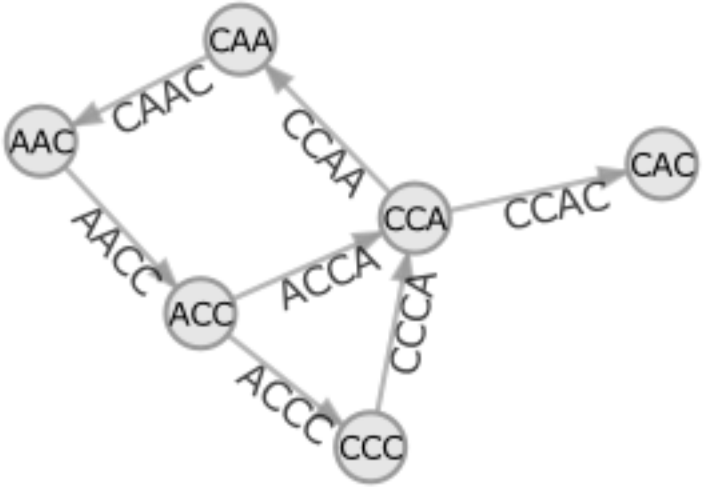
De novo transcriptome assembly

How it works, limitations and how to tell if your assembly is any good

De novo

- Latin expression meaning "from the beginning," "afresh," "anew," "beginning again."
- De novo, a term for any method that makes predictions about biological features using only a computational model without extrinsic comparison to existing data

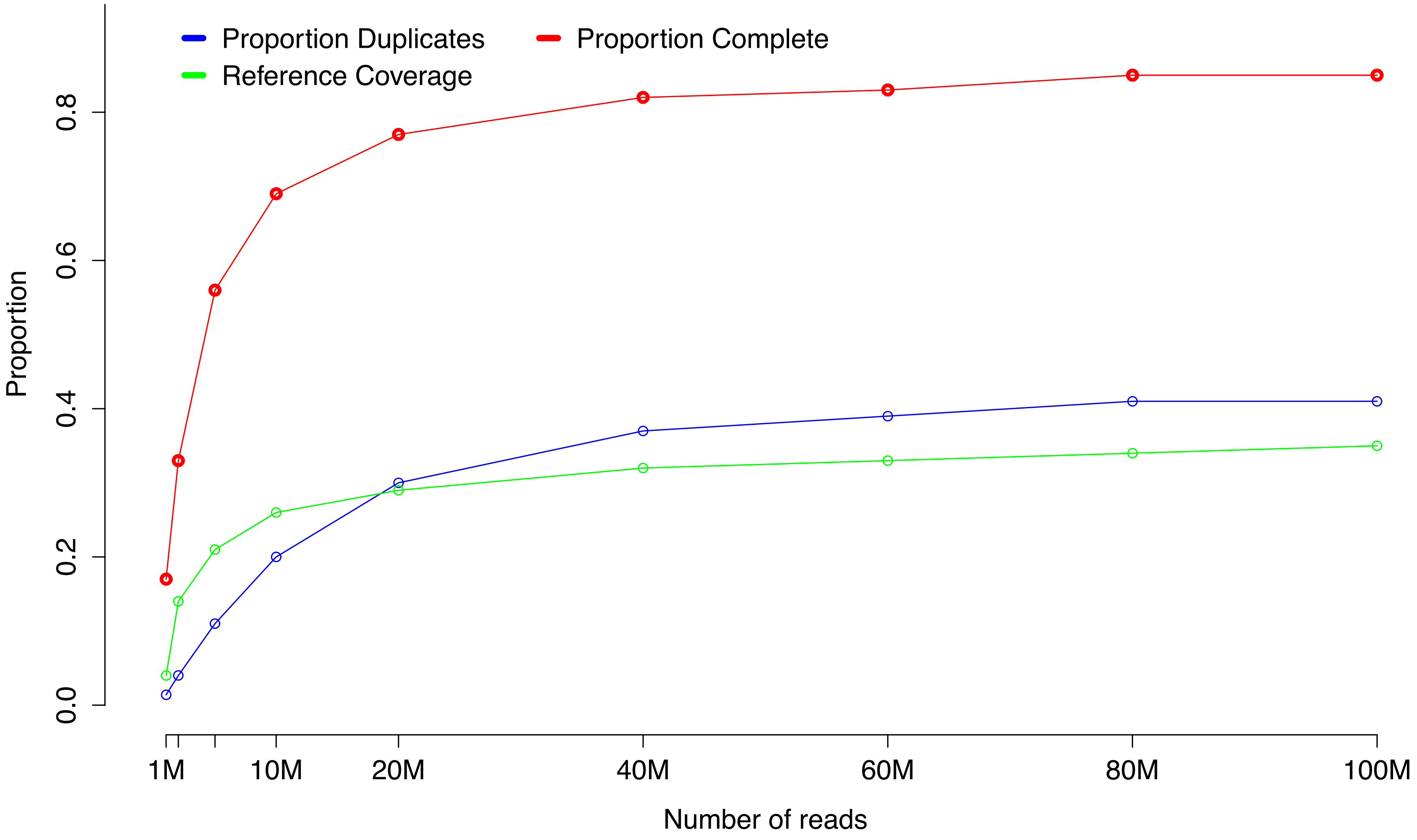


- No reference genome, then you must assemble your reads into genes

Transcriptome vs Genome Assembly

- Transcriptome is easier
  - Smaller total volume of bases
  - Less low complexity regions
- Transcriptome is harder
  - Alternative splicing
  - Expression variability between tissues/cells
    - Difficult/impossible to fully sample all transcripts
    - Exponentially distributed coverage levels – assembler must work on both high and low depth of coverage regions
- Because of the very unique properties of transcriptome assemblies, it is important to use an assembler meant for transcriptomes (not genomes)

How many reads?

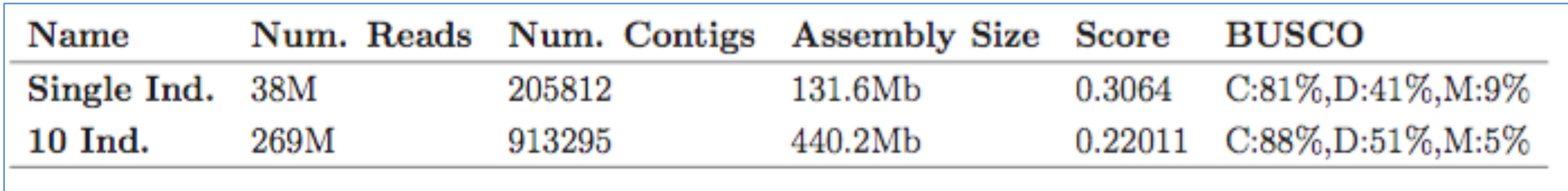


MacManes 2016 BioRxiv

How many individuals? How many tissues?

- Preferably a single individual – don't introduce more heterozygosity
- Preferably many tissues, development stages, stress
  - Tissues, developmental stages and environmental conditions all turn on/off transcription
  - Diversify libraries to try to sample as many transcripts as possible

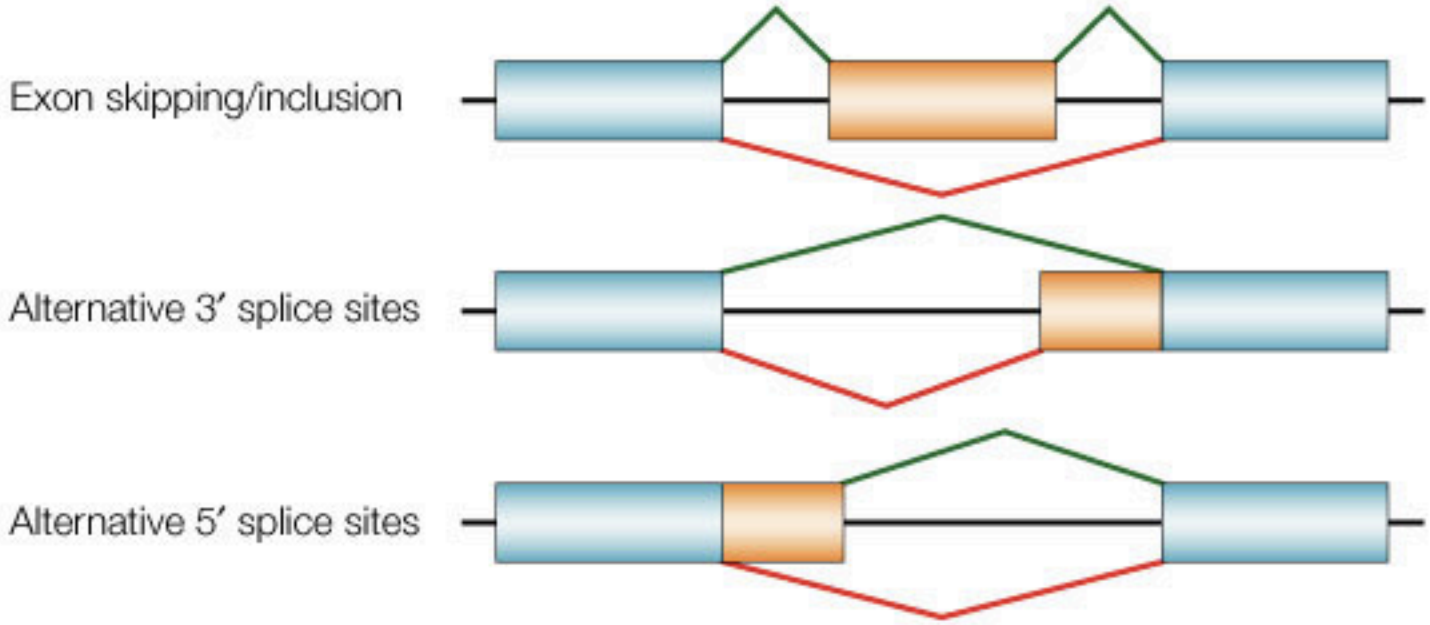
MacManes et al 2016 (bioRxiv)



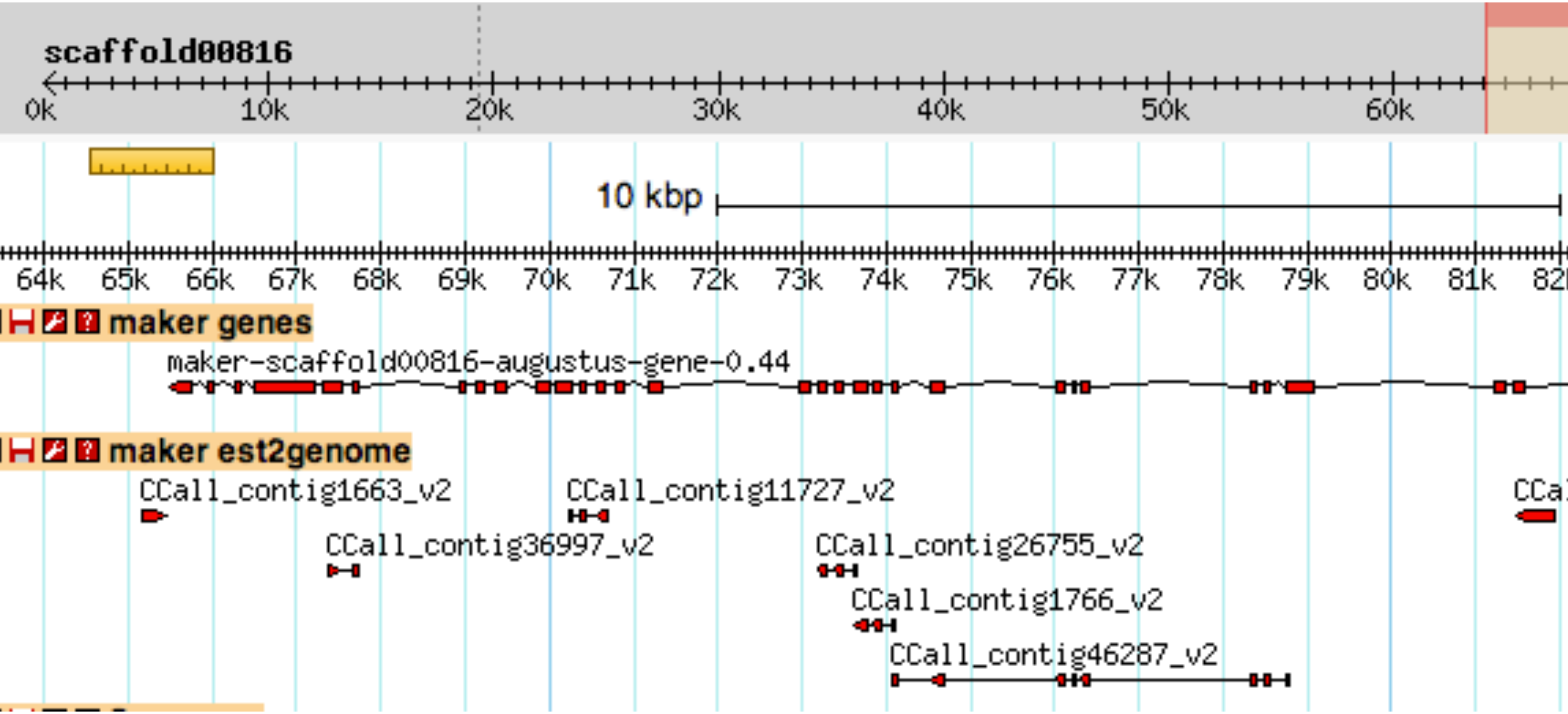
Name	Num. Reads	Num. Contigs	Assembly Size	Score	BUSCO
Single Ind.	38M	205812	131.6Mb	0.3064	C:81%,D:41%,M:9%
10 Ind.	269M	913295	440.2Mb	0.22011	C:88%,D:51%,M:5%

Problems with *de novo* assemblies

- Results
  - Highly fragmented assemblies
  - Chimeras (can be biological, experimental or computational)
  - Paralogs, alleles and alternative splicing variants combined or fragmented



Chestnut

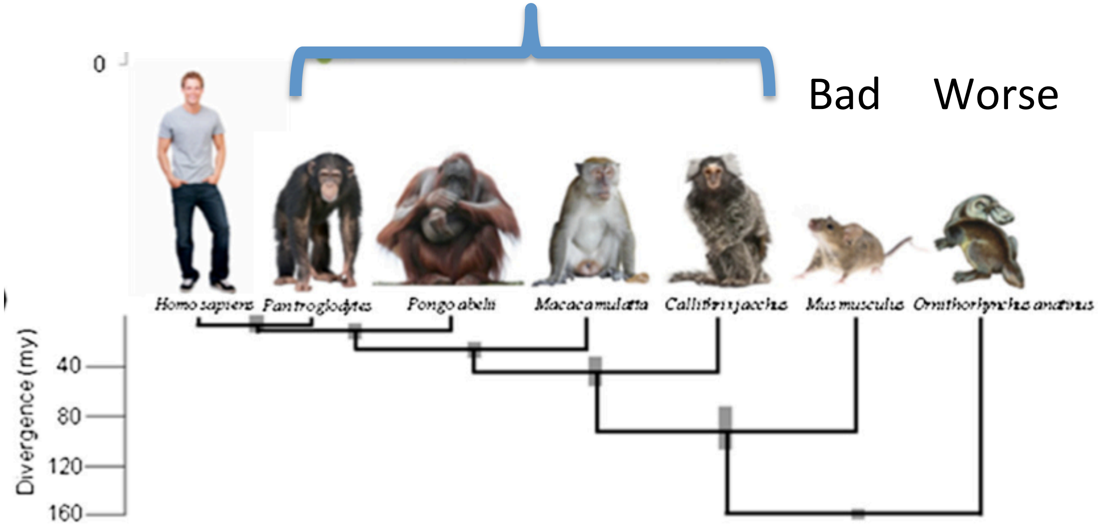


scaffold00816
0k
10k
20k
30k
40k
50k
60k
10 kbp
64k
65k
66k
67k
68k
69k
70k
71k
72k
73k
74k
75k
76k
77k
78k
79k
80k
81k
82
maker genes
maker-scaffold00816-augustus-gene-0.44
maker est2genome
CCall_contig1663_v2
CCall_contig11727_v2
CCa
CCall_contig36997_v2
CCall_contig26755_v2
CCall_contig1766_v2
CCall_contig46287_v2

Avoid an entirely *de novo* assembly - Use a distant or close relative

- Is there a close relative with a sequenced genome?
- How close is close enough?
    - Align then assemble
    - Assemble then align

Reasonable Distant References



Bad Worse

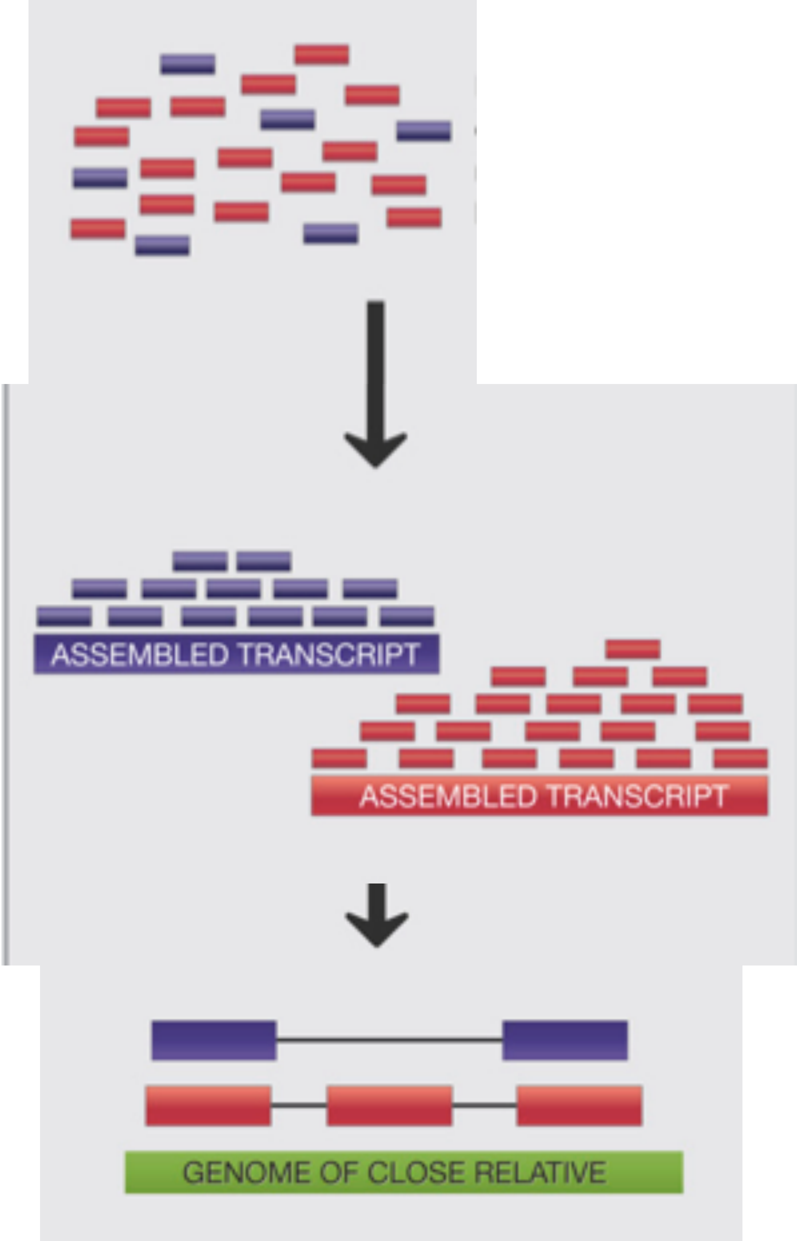
*Likely to vary by phylogenetic neighborhood

Hornett and Wheat. Quantitative RNA-Seq analysis in non-model species: assessing transcriptome assemblies as a scaffold and the utility of evolutionary divergent genomic reference species. BMC Genomics 2012, 13:361

Assemble then align

- First, assemble
- Next, align to a close relative

- Main Problems:
  - Fragmented assemblies – gene pieces are scattered in a different consensus pieces
  - More difficult to sort out gene family members
- Main Advantages:
  - Alignment to a close relative can identify exon/exon boundaries (sort out alternative splicing)
  - Less bias – can discover novel gene sequences



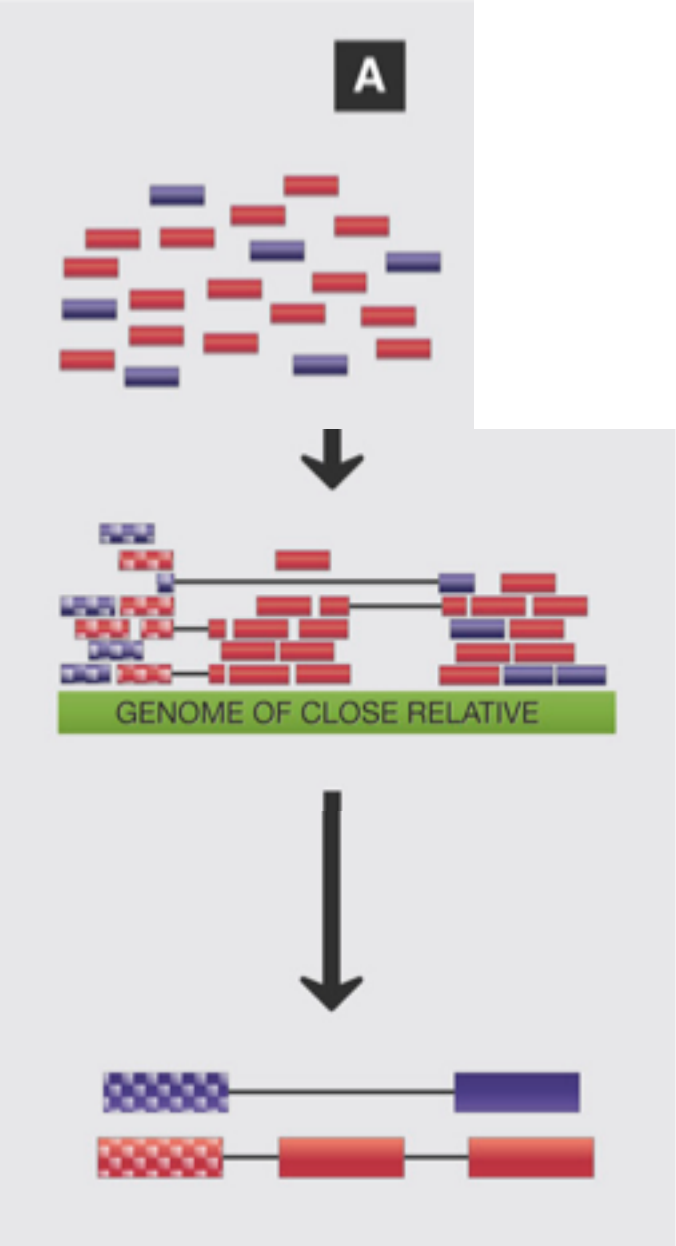
Weber et al. Strategies for transcriptome analysis in nonmodel plants Am J Bot. 2012 99(2): 267-276

Align then assemble

- First, map reads to (distant) reference
- Next, do local assemblies for each gene

- Main Problems
  - Read alignment may be poor due to lack of sequence similarity
  - Gene family expansion/contraction

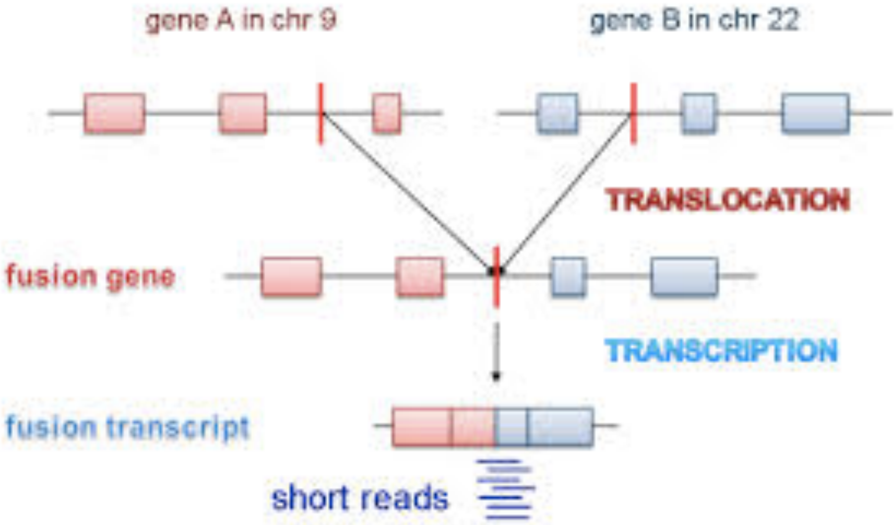
- Main Advantage
  - Transcript assembly is less likely to be fragmented
  - Even where it is fragmented, you can identify all the fragments that originate from a single locus



Weber et al. Strategies for transcriptome analysis in nonmodel plants Am J Bot. 2012 99(2): 267-276

De novo transcriptome assemblies

- Completely reference free
- What are they useful for?
  - Transcriptome characterization
  - Enabling proteomics experiments
  - Candidate gene discovery
  - Marker discovery/ development
  - Cancer or other tissues where fusion events are important
  - Metatranscriptomics – surveying microbiota

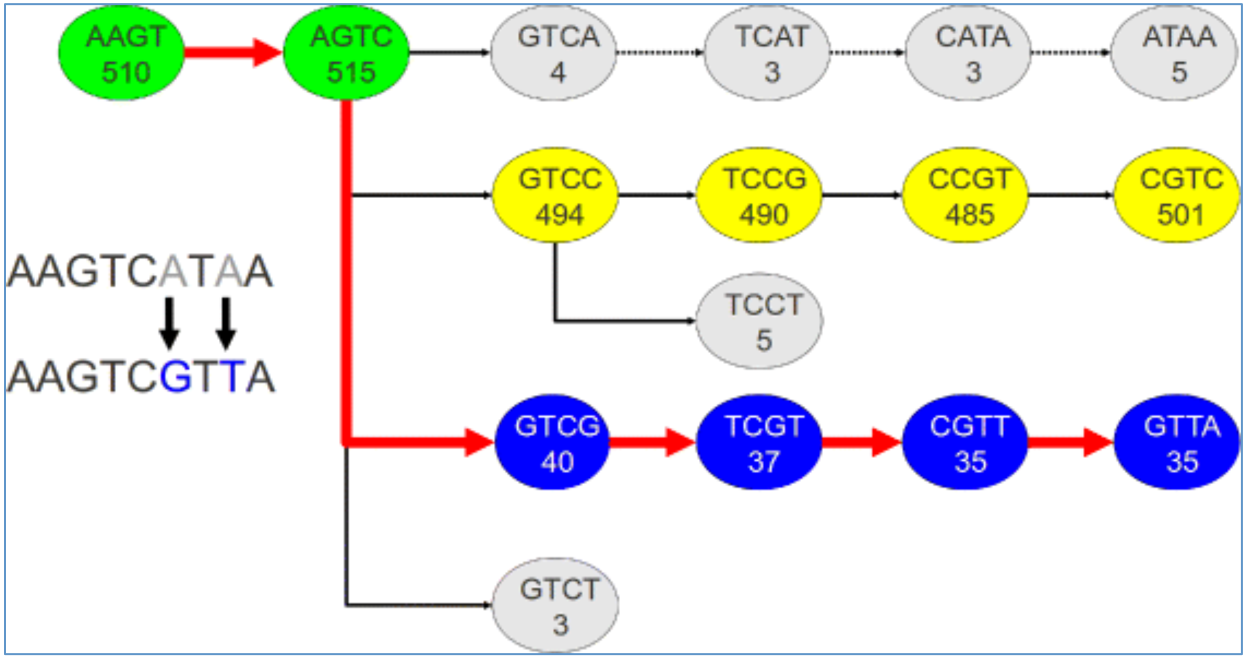


Steps

- Gentle trimming
- Error correction
- Assembly
- ORF Identification
- Annotation

Error Correcting

- Can substantially improve contig assembly
- Software:
  - Rcorrector
  - Bfc



Song and Florea 2015

- Uses k-mer strategy
- Very rare k-mers are likely to contain an error

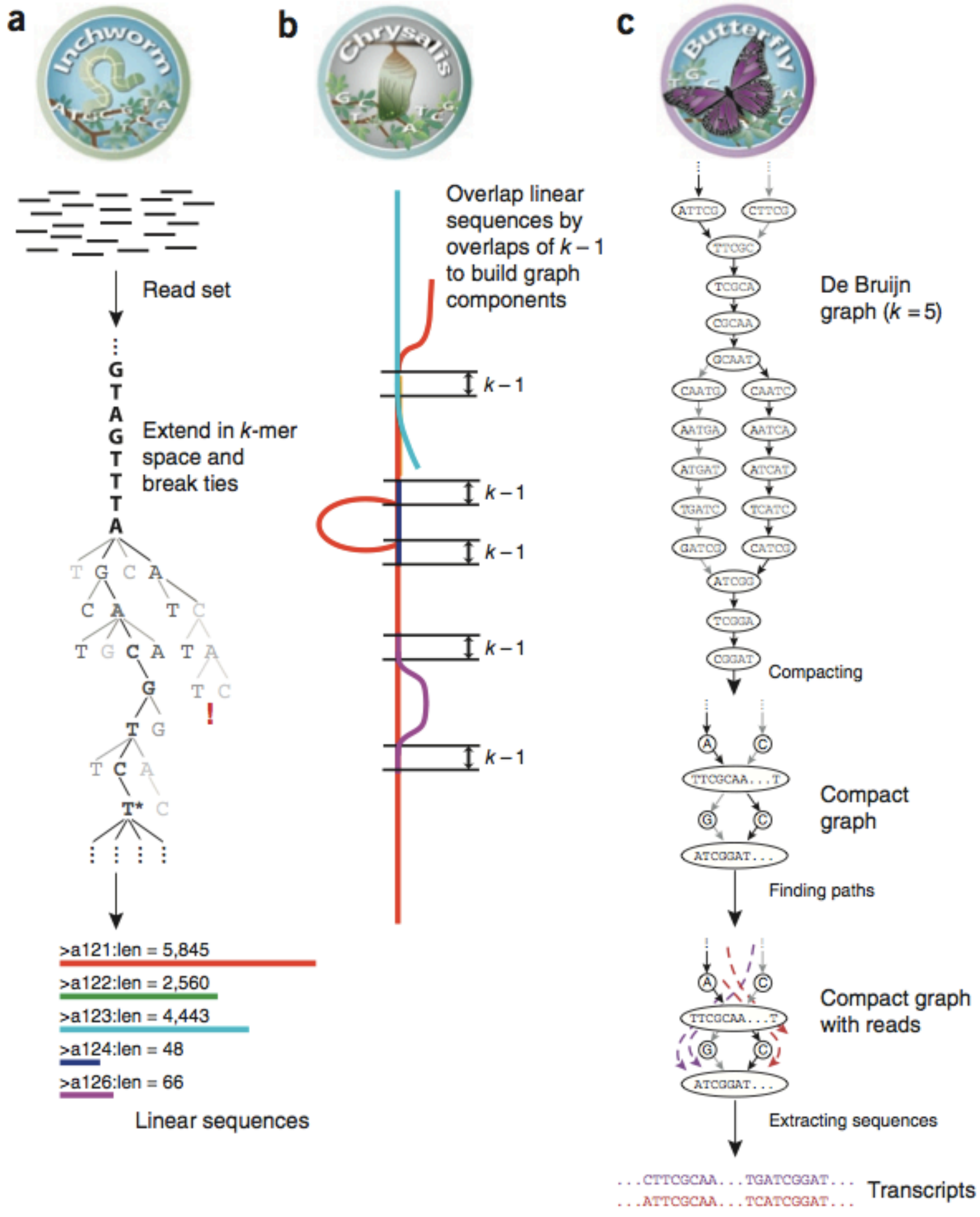
Assembly Software

- Trinity is most common and usually considered most accurate
  - Smith-Unna et al 2015
  - Li et al 2014
- SOAP de novo trans
- Trans-Abyss
- Oases

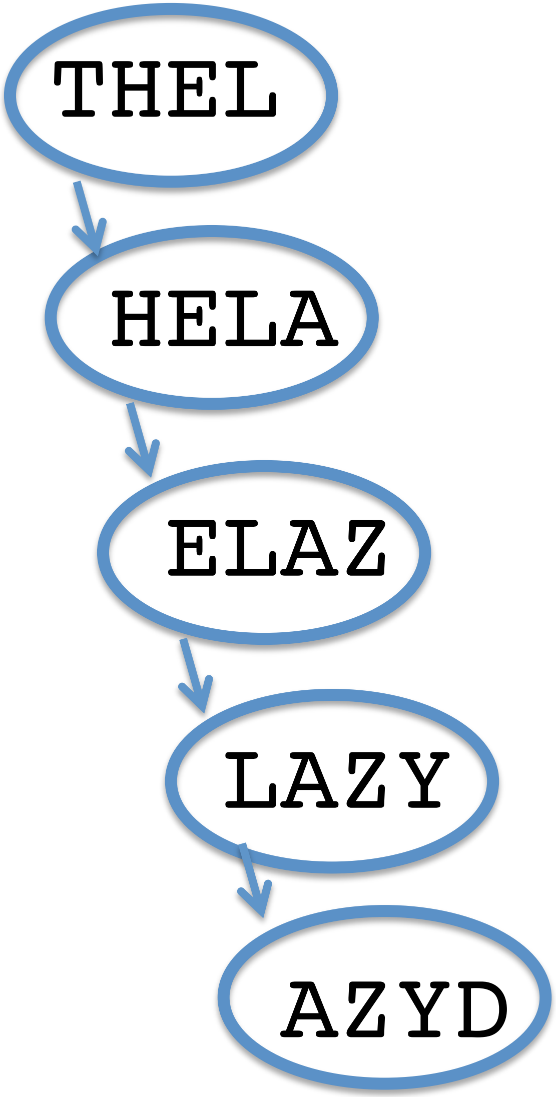
Trinity strategy

Three stages

1. Inchworm
2. Crysalis
3. Butterfly



THELAZYBROWNDOG



THEL

HELA

ELAZ

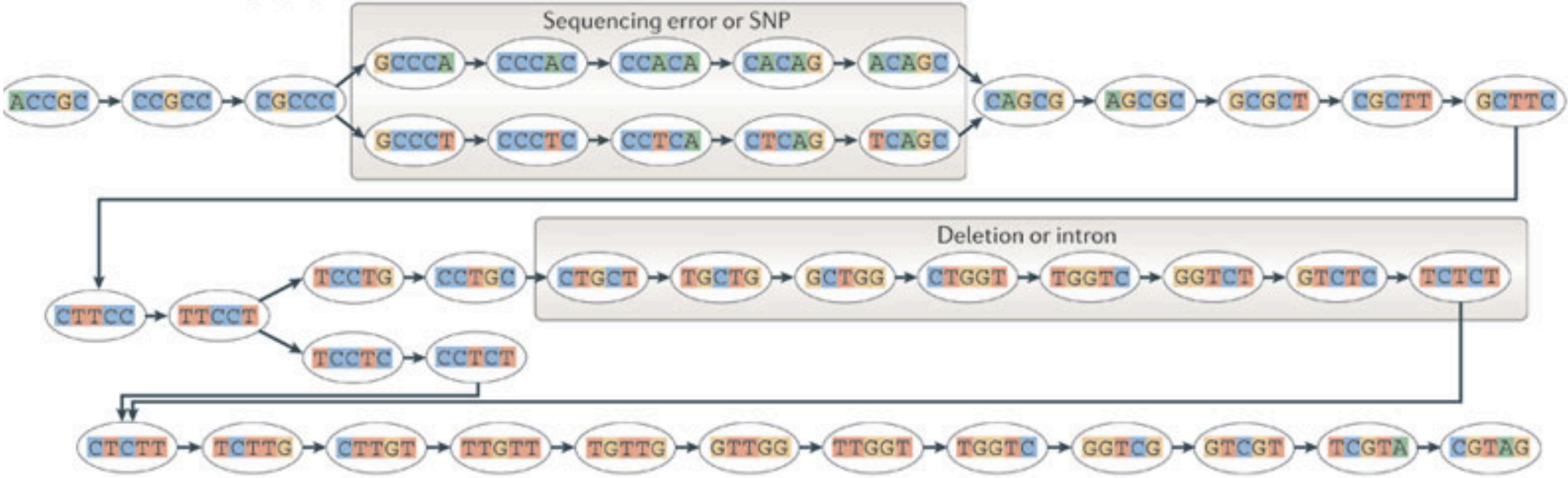
LAZY

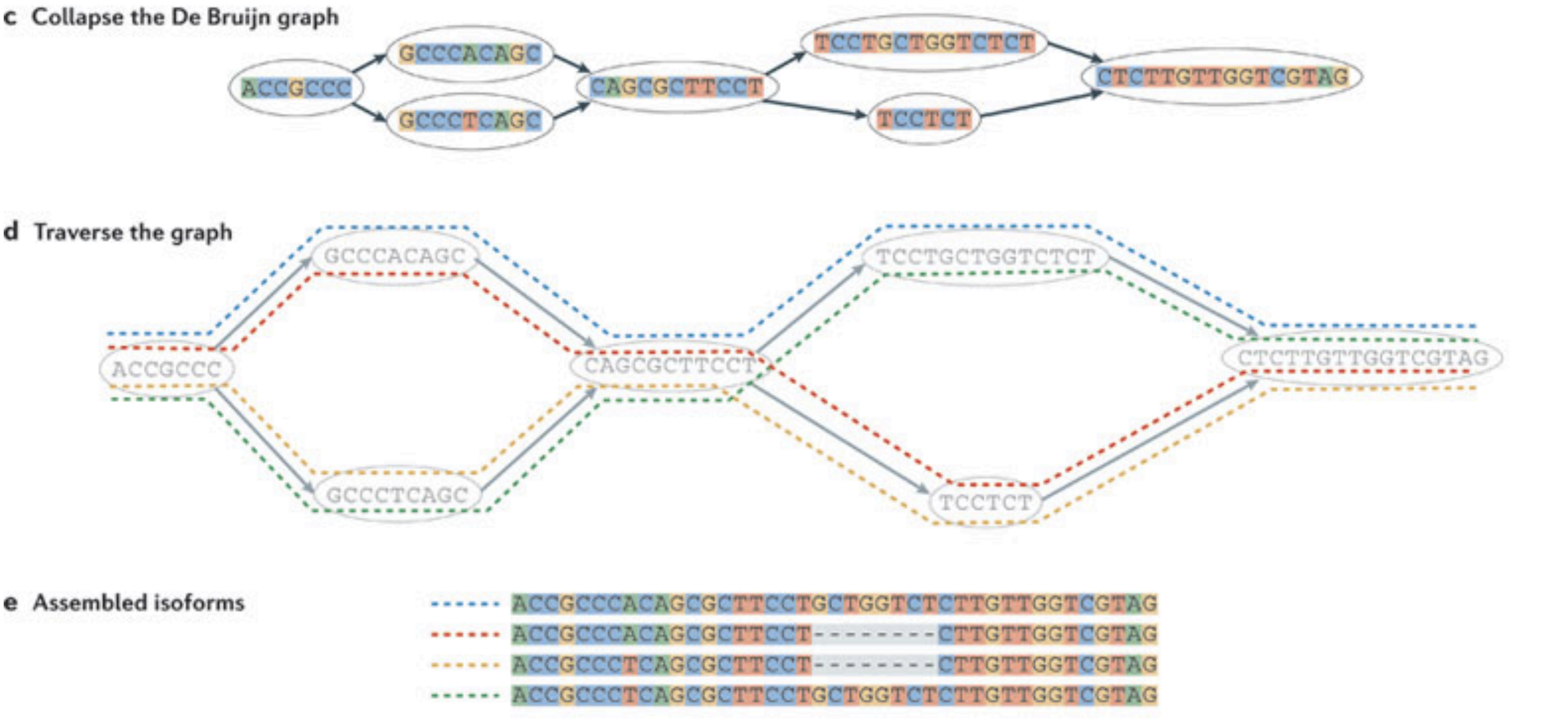
AZYD

De Bruijn Graph

Directed acyclic graph of kmers

Generate the De Bruijn graph



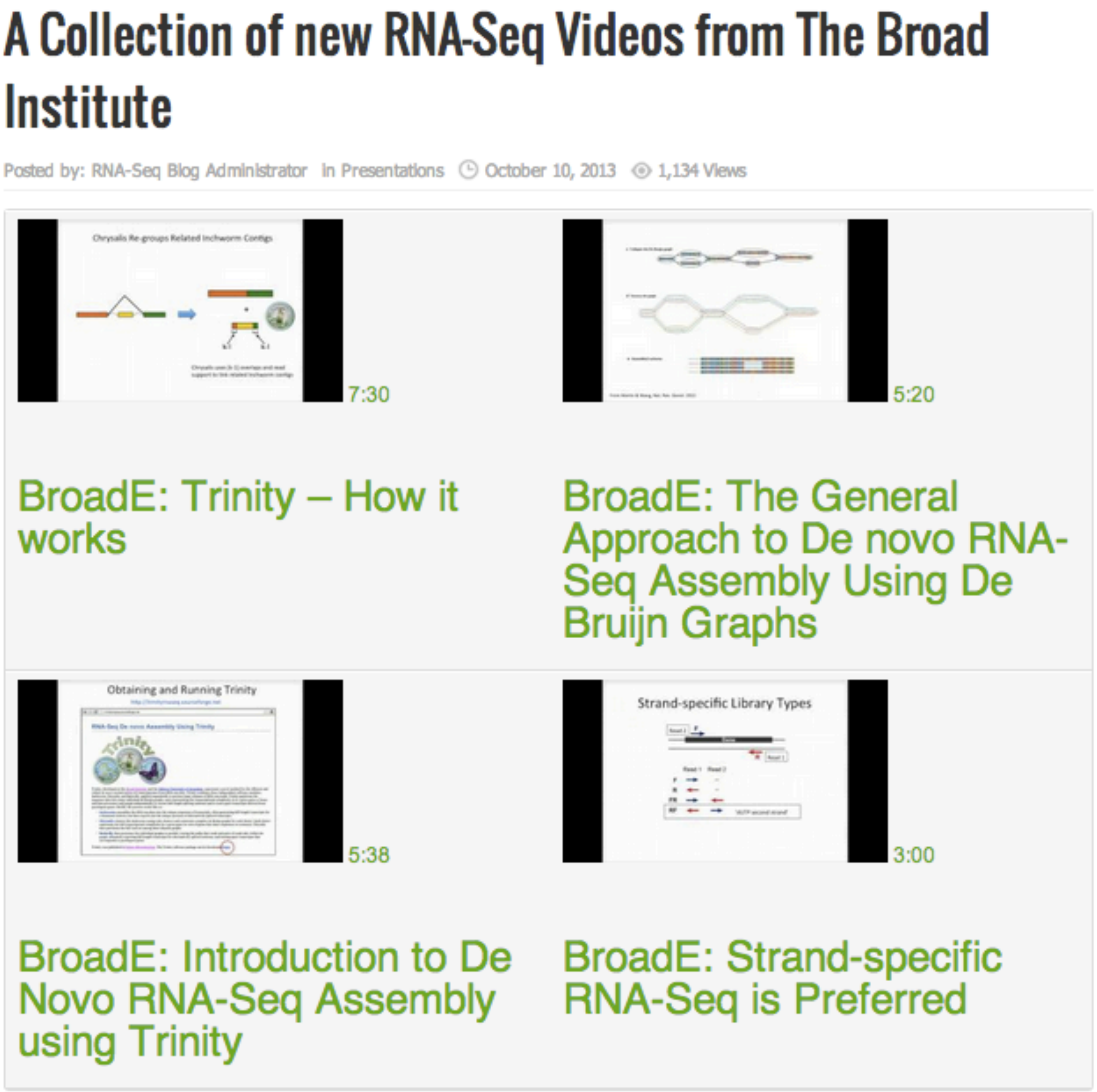


c Collapse the De Bruijn graph
ACCGCCC
GCCCACAGC
GCCCTCAGC
CAGCGCTTCCT
TCCTGCTGGTCTCT
TCCTCT
CTCTTGTTGGTCGTAG
d Traverse the graph
GCCCACAGC
ACCGCCC
GCCCTCAGC
CAGCGCTTCCT
TCCTGCTGGTCTCT
TCCTCT
CTCTTGTTGGTCGTAG
e Assembled isoforms
ACCGCCCACAGCGCTTCCTGCTGGTCTCTTGTTGGTCGTAG
ACCGCCCACAGCGCTTCCT--------CTTGTTGGTCGTAG
ACCGCCCTCAGCGCTTCCT--------CTTGTTGGTCGTAG
ACCGCCCTCAGCGCTTCCTGCTGGTCTCTTGTTGGTCGTAG

Nature Reviews | Genetic

Videos!

http://www.rna-seqblog.com/a-collection-of-new-rna-seq-videos-from-the-broad-institute/



A Collection of new RNA-Seq Videos from The Broad Institute

Posted by: RNA-Seq Blog Administrator In Presentations October 10, 2013 1,134 Views

Chrysalis Re-groups Related Inchworm Contigs
7:30
5:20

BroadE: Trinity – How it works

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

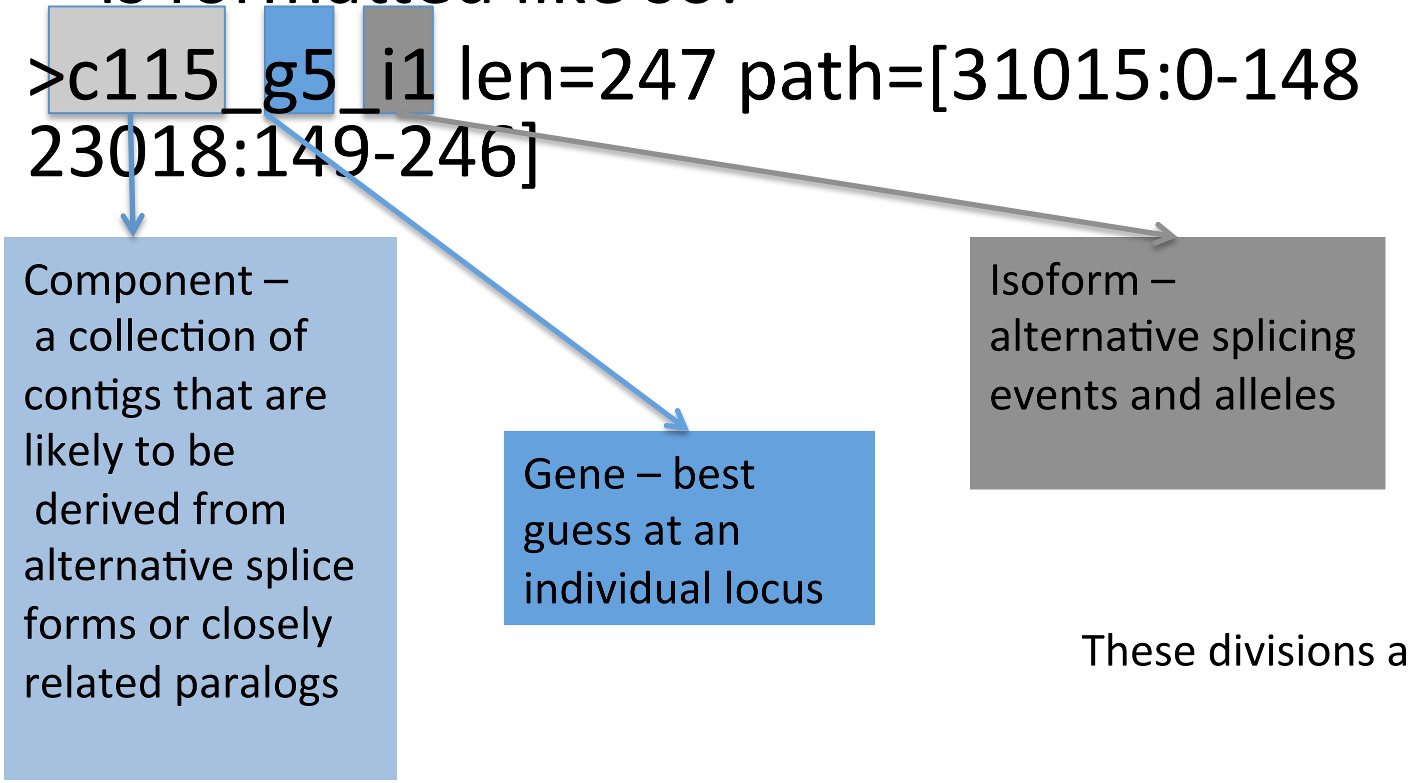
Obtaining and Running Trinity
5:38
Strand-specific Library Types
3:00

BroadE: Introduction to De Novo RNA-Seq Assembly using Trinity

BroadE: Strand-specific RNA-Seq is Preferred

Trinity output – deciphering the naming

- An example Fasta entry for one of the transcripts is formatted like so:



>c115_g5_i1 len=247 path=[31015:0-148 23018:149-246]

Component – a collection of contigs that are likely to be derived from alternative splice forms or closely related paralogs

Gene – best guess at an individual locus

Isoform – alternative splicing events and alleles

These divisions are guesses only!

II. Improving the assembly and checking quality

How to figure out if your assembly is good

- BUSCO
  - Benchmarking Universal Single-Copy Orthologs
  - based on evolutionarily informed expectations of gene content from near-universal single-copy orthologs selected from OrthoDB.
  - Use to assess completeness of transcriptome
  - http://busco.ezlab.org/

How to figure out if your assembly is good

- Map reads back and see what % are captured in the assembly
- Transrate
  - analyses a transcriptome assembly in three key ways:
    - by inspecting the contig sequences
    - by mapping reads to the contigs and inspecting the alignments
    - by aligning the contigs against proteins or transcripts from a related species and inspecting the alignments

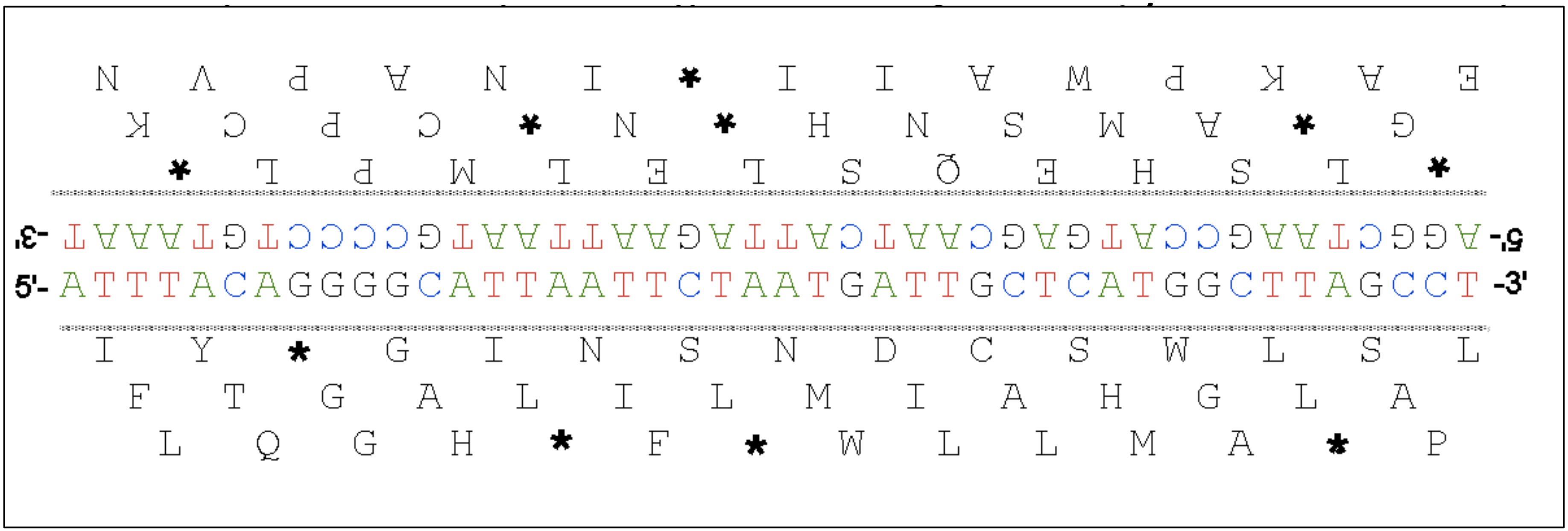
transrate

understand your transcriptome assembly

III. De novo transcriptome sequencing – after assembly

ORF Finding - TransDecoder

- Searches all frames for ORFs, start codons and stop codons
- Maximizes length and log-likelihood score of ORF
- a single transcript can report multiple ORFs (allowing for operons, chimeras, etc).



3'- TAAATGTCCCCGTAATTAAGATTACTAACGAGTACCGAATCGGA -5'
5'- ATTTACAGGGGCATTAATTCTAATGATTGCTCATGGCTTAGCCT -3'
I Y * G I N S N D C S W L S L
F T G A L I L M I A H G L A
L Q G H * F * W L L M A * P