RNASEQ WITHOUT A REFERENCE

Experimental Design

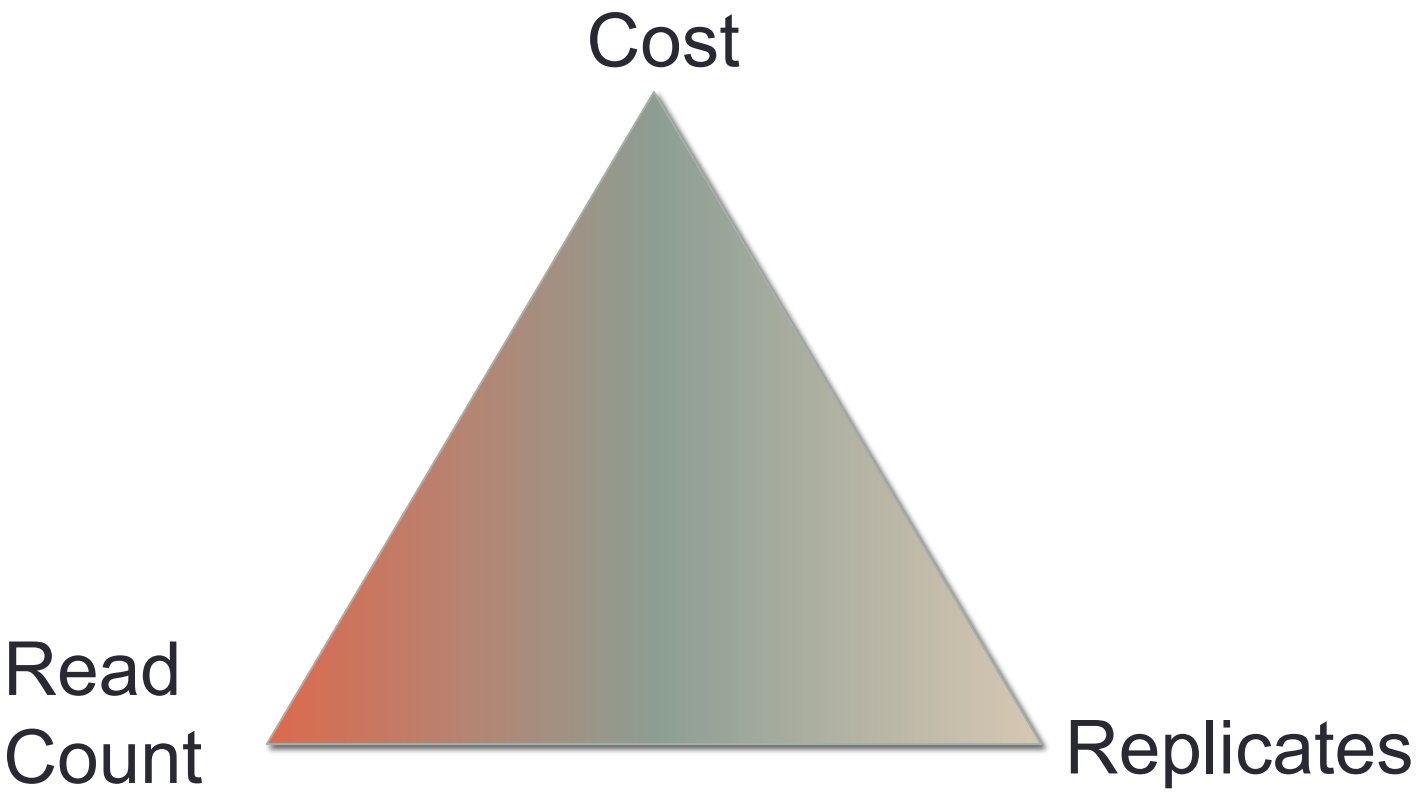
Assembly in Non-Model Organisms

And other (hopefully useful) Stuff

Meg Staton
mstaton1@utk.edu
University of Tennessee
Knoxville, TN

I. Project Design

Things you need to know BEFORE you begin

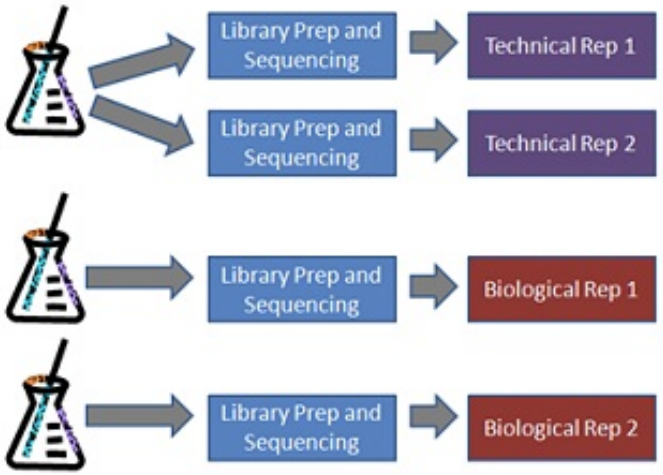


Pro Tip: Who is your resident statistician? Buy them a coffee and make friends.

Replicates – What?

- Biological Replicates – independent biological sample, processed separately and barcoded
- Technical Replicates – independent library construction or sequencing of the same biological sample

- Technical reproducibility is very good for RNASeq
- Biological variation is much greater!
- Different genes have different variances and are potentially subject to different errors and biases.



"Thinking About RNA Seq Experimental Design for Measuring Differential Gene Expression: The Basics" http://gkno2.tumblr.com/post/24629975632/thinking-about-rna-seq-experimental-design-for

Marioni, J.C., et al. (2008) RNA-seq: An assessment of technical reproducibility and

Replicates – How many?

- beyond a depth of 10 million reads, replicates provide more statistical power than depth for detecting differential gene expression

RNA-seq differential expression studies: more sequence or more replication?

Yuwen Liu[1,2], Jie Zhou[1,3] and Kevin P. White[1,2,3,*]

- Many people say at least 3 – this enables the t-test
- What if one fails?
- (Fishers exact test can utilize no replicates)

Replicates – Software?

- Both EdgeR and DeSeq will calculate variance from replicates (but neither do a t-test)
- From the horse's mouth:
- "to use something like a t test, you need enough replicates to estimate a variance for each gene. With two groups of five samples, you are already entering the regime there this should work well. For comparison, also try a tool that pools information from several genes to get better confidence in variance estimates, such as our DESeq or the Smyth group's edgeR. Of course, we like to claim that DESeq is better than edgeR, and for only two or three replicates, I do think so, but for five or more replicates, edgeR's "moderation" feature really pays off. So, even though I don't like admitting this, for your set-up [of 5 replicates per treatment], edgeR should work better than DESeq."

  -Simon Anders on SeqAnswers

http://seqanswers.com/forums/showthread.php?t=10410

Replicates – And Blocks?

- Randomized Block Design
- Randomize - assigning individuals at random to treatments in an experiment
- Blocking - Experimental units are grouped into homogeneous clusters in an attempt to improve the comparison of treatments
  - Example – all organisms from the same location are “blocks”, multiple locations used
  - Example - each block is a cultivar, with individuals from that cultivar randomly assigned to a treatment

Read Count - How to Decide?

- Standards, Guidelines and Best Practices for RNA-Seq
- V1.0 (June 2011)
- The ENCODE Consortium
- What are you trying to do?
  - Compare two mRNA samples for differential expression (30M PE per sample)
  - Discover novel elements, perform more precise quantification, especially of lowly expressed transcripts (100-200M PE per sample)
- What resources do you already have?
  - Well assembled and annotated genomes – single ends, shorter reads
  - De novo – longer reads, paired ends
- What is being published in your community?

http://encodeproject.org/ENCODE/protocols/dataStandards/ENCODE_RNAseq_Standards_V1.0.pdf

Read Count – How to Decide? (cont.)

- Blogosphere disagrees
  - Need half the coverage, double the replicates!
- Current experiments indicate that we are NOT discovering significantly more transcripts with a hiSeq run vs a miSeq run. (At least not transcripts that look like genes)
  - A deep biased view

Scotty – You need more power!

- Scotty is a web service to plan RNA-Seq experiments that measure differential gene expression.
- Prototype data required
  - Pilot data -at least two replicates of either control or treatment
  - Pre-loaded data

Scotty – up to $20k

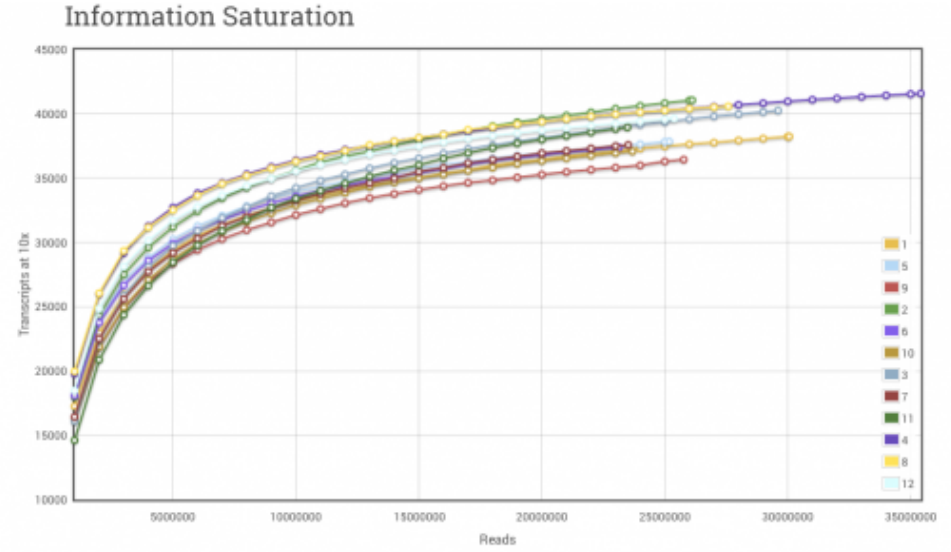
User Inputs Used in the Analysis

Control columns in pilot data: 3
Test columns in pilot data: 3
Cost per replicate, control: $200
Cost per replicate, test: $200
Cost per million reads: $23
Alignment Rate: 90%
Maximum cost of experiment: $20000
Percentage of genes detected: 50
At p value cutoff: 0.01
For the following true fold change: 2
Maximum percentage of genes with low-powered (biased) measurements: 50

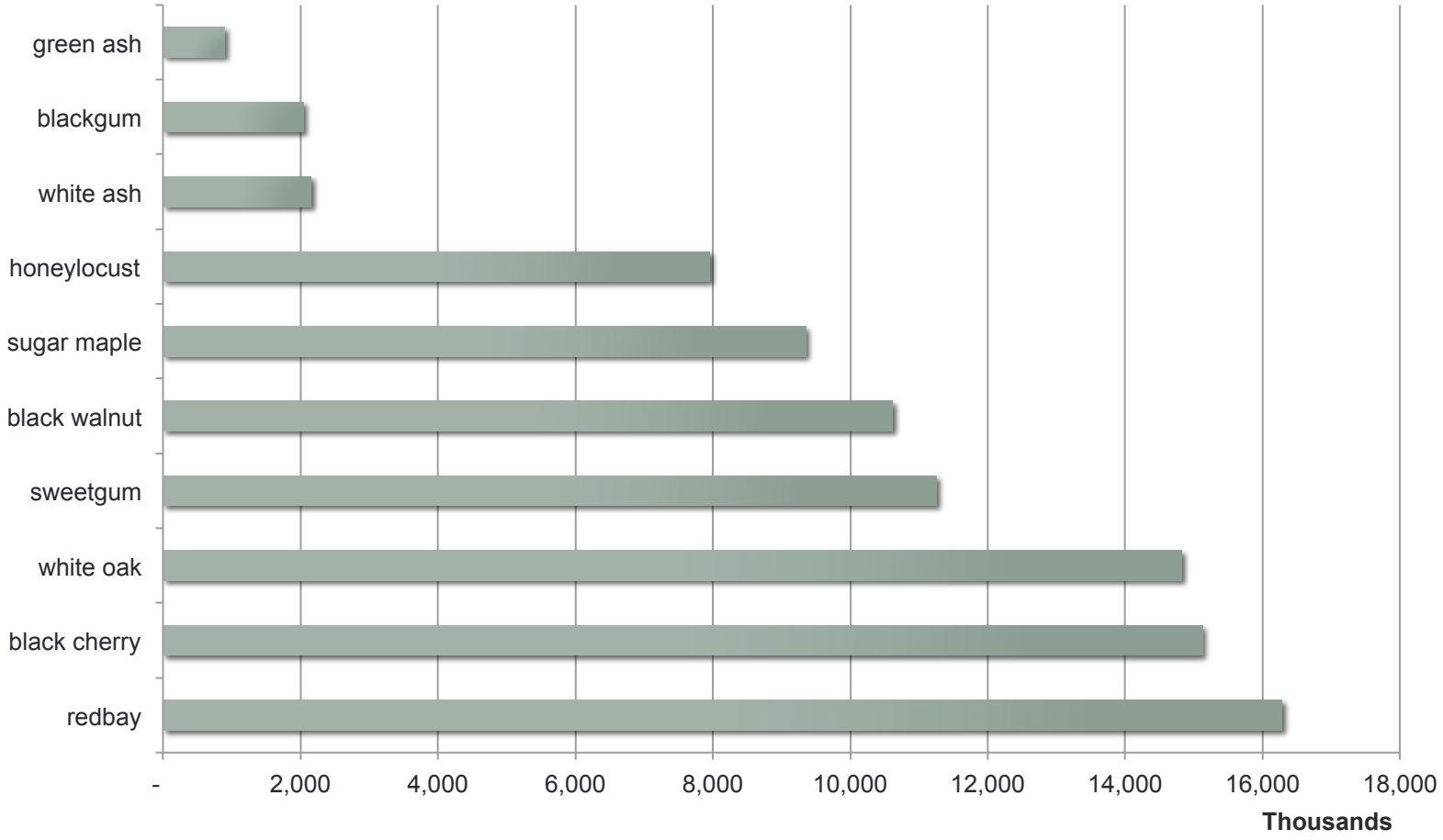
- Least expensive: 6 replicates sequenced to a depth of 12 million reads aligned to genes per replicate. $5,712
- Most powerful: 20 replicates sequenced to a depth of 34 million reads aligned to genes per replicate. $19,640

When is enough enough?

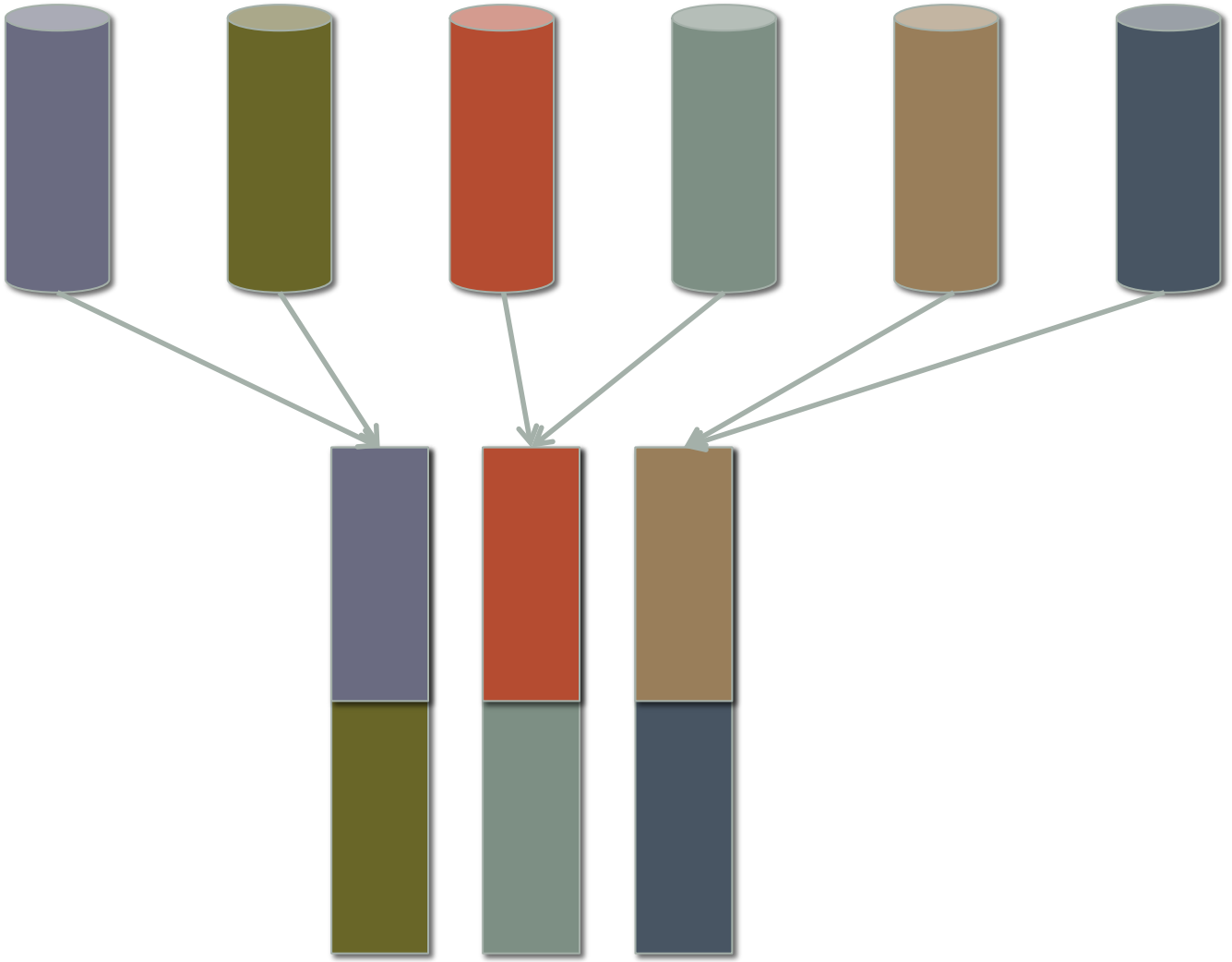
- Average coverage is not a helpful metric for RNASeq – relative expression varies by orders of magnitude
- Saturation curves
  - the number of transcript references with < 10 read alignments
- New Discovery Rate
  - How many new genes are being discovered with each additional slice of data?



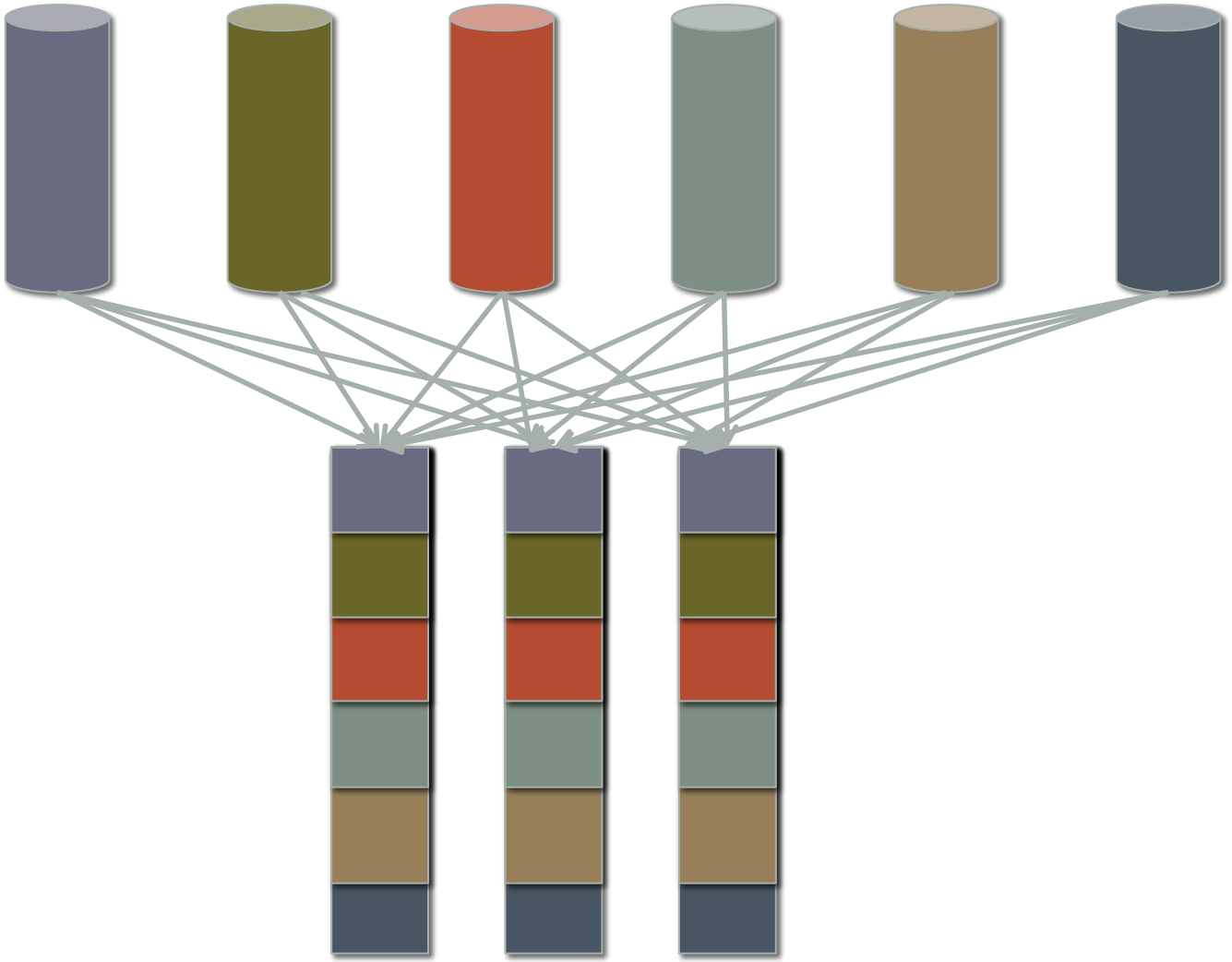
Reality: Variation in # of Sequences



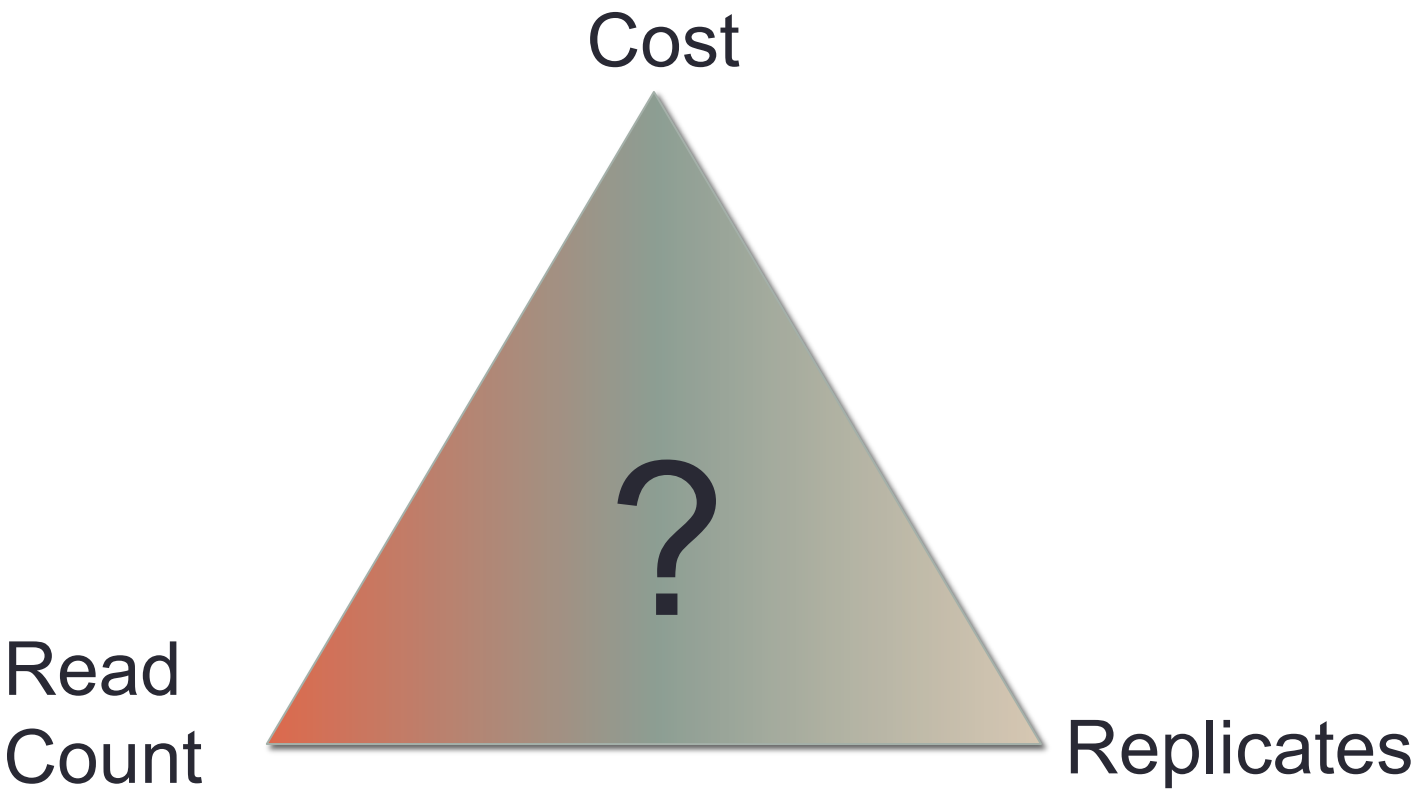
Spreading libraries across lanes?



Spreading libraries across lanes?
Balanced Block Design



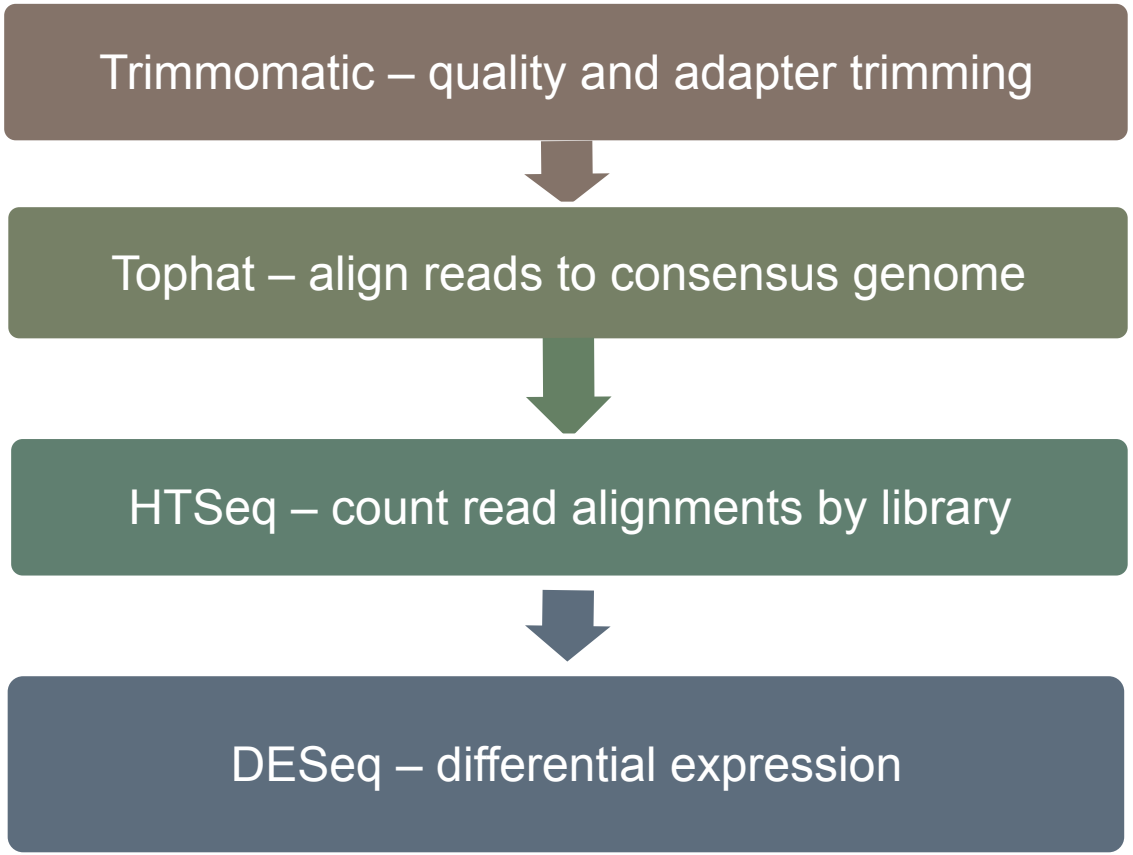
What's right for your experiment?



II. De novo transcriptome sequencing - assembly

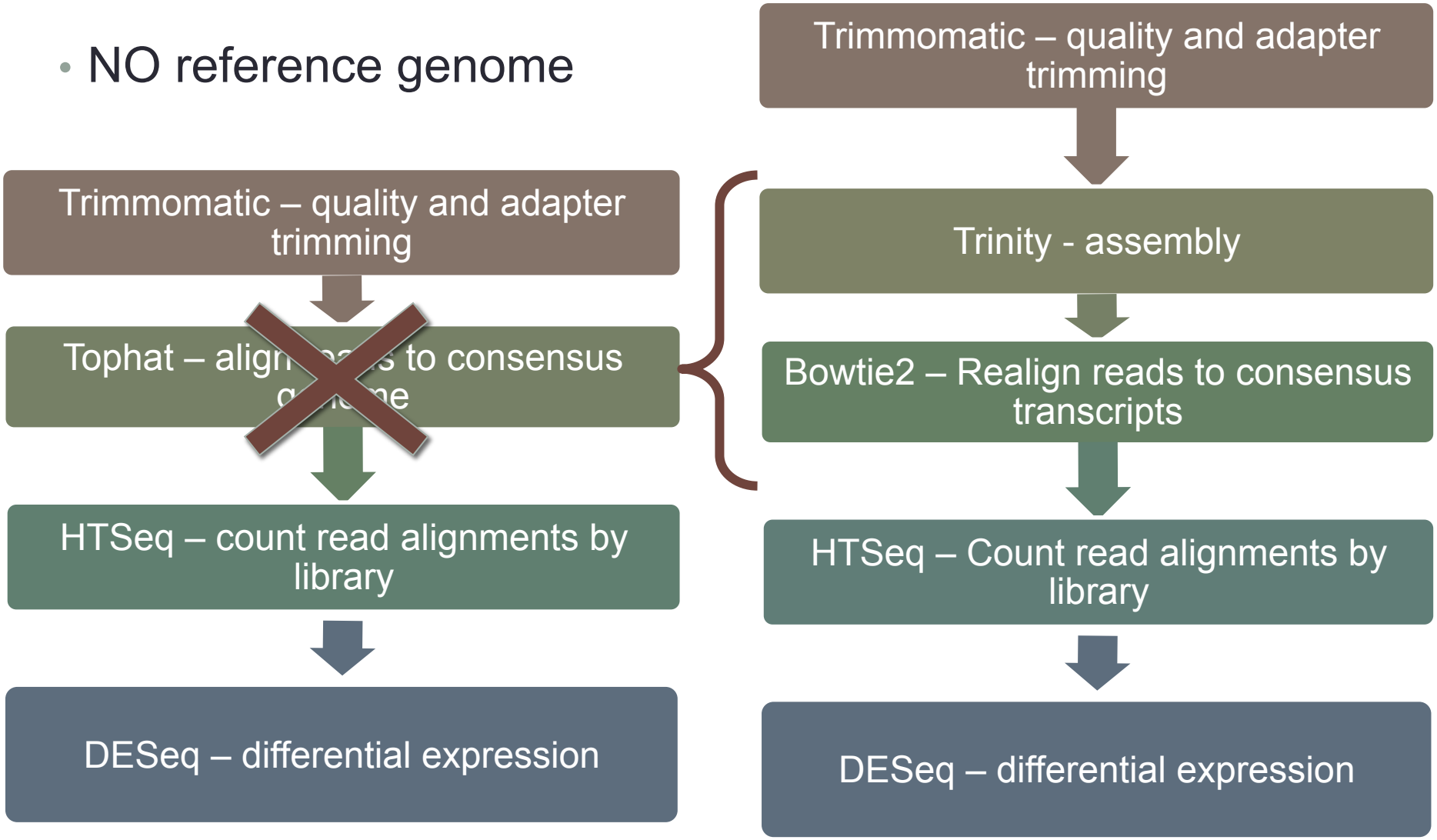
Model Organism

- reference genome



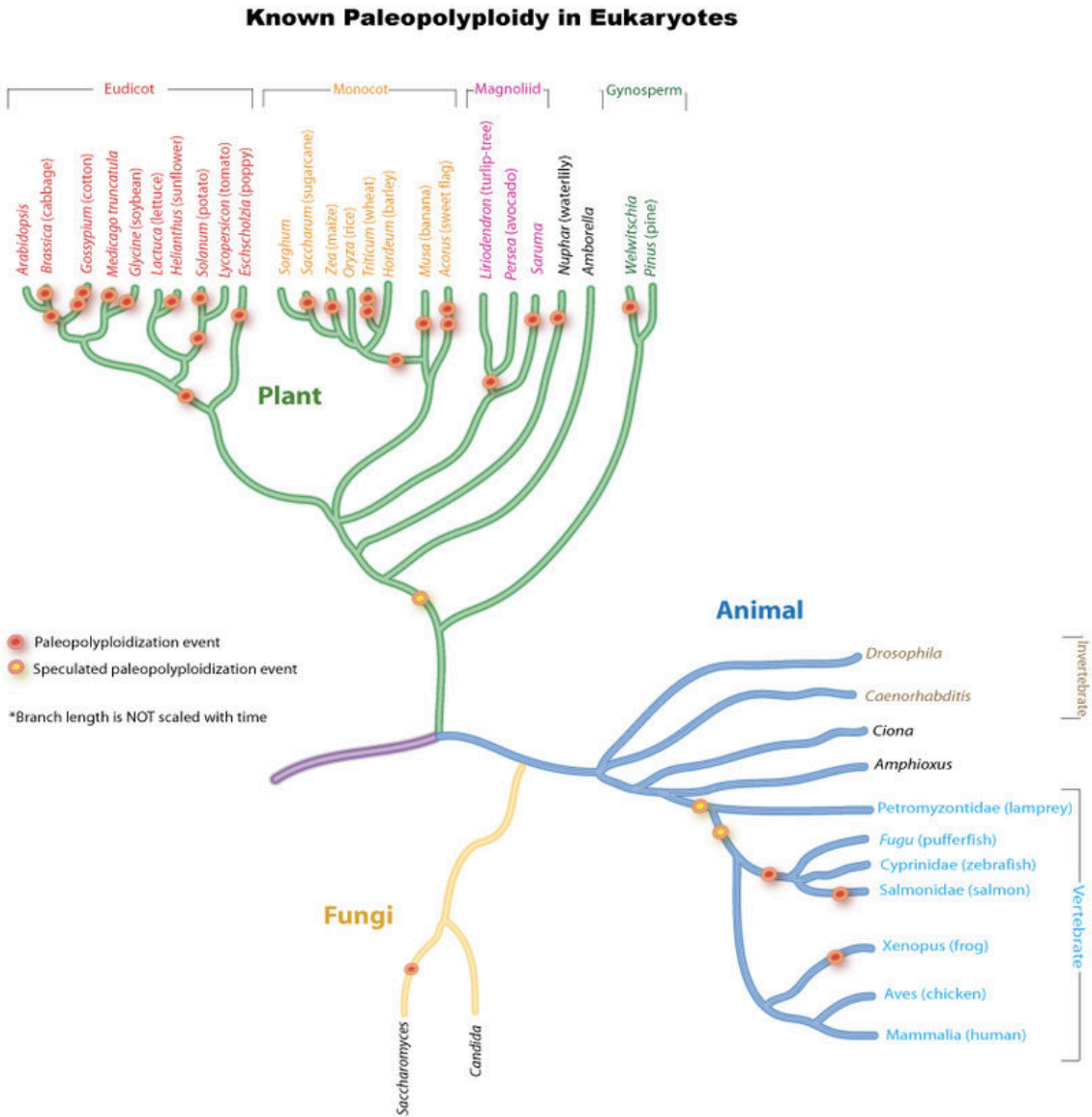
Non-model Organism

- NO reference genome



Problems with *de novo* assemblies

- plant species have larger and more complex genome sizes and structures than animal species
- tremendous diversity in both size and structure
- From a plant perspective
  - Polyploidy
  - Gene family proliferation
  - Heterozygosity
  - Repetitive element proliferation



Why plants are difficult (cont)

- Our best reference, *Arabidopsis thaliana* has a genome that underwent a 30% reduction in genome size and at least nine rearrangements in the short time since its divergence to *Arabidopsis lyrata*
- Maize pan genome - Intraspecific variations of as much as 38.8% from the average of 5.5 pg/2n nucleus driven by LTR retrotransposon expansion
- Conifer genome sizes
  - Loblolly pine 22Gb (7x bigger than human)
  - largest genome contains roughly 60,000,000,000 more base pairs than the smallest genome

- Often these difficulties make transcriptome sequencing more attractive than whole genome sequencing!

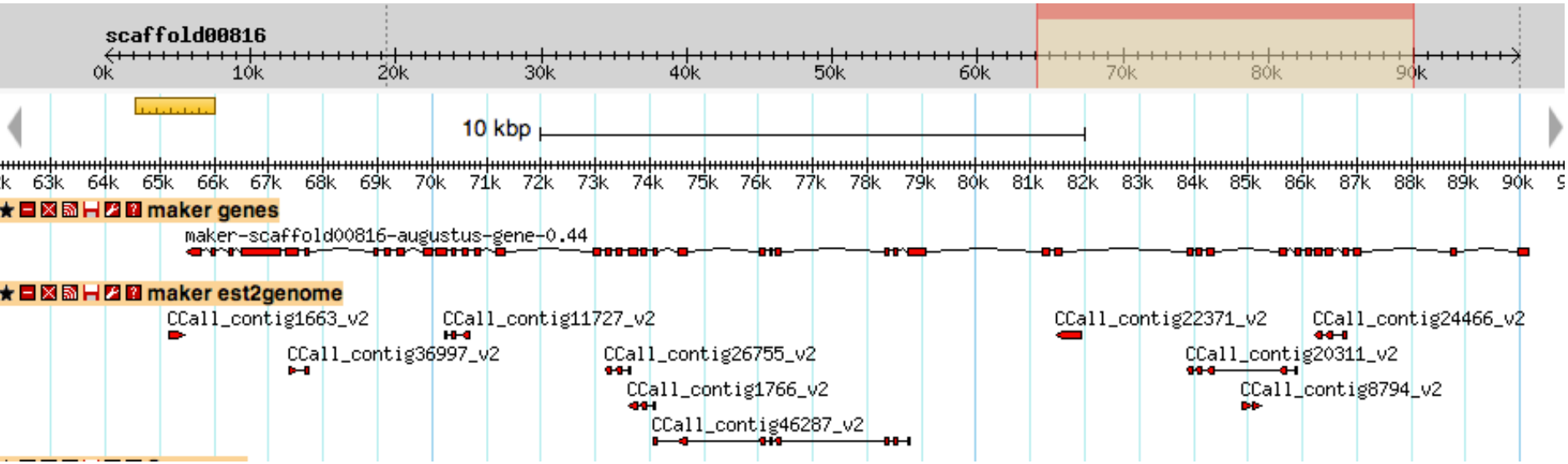
Problems with *de novo* assemblies

- Results
  - Highly fragmented assemblies
  - Chimeras:
  - Paralogs, alleles and alternative splicing variants mushed together or fragmented

- “Metrics based upon contig lengths (e.g. mean, median, N50) do not provide quantitative insights into how much of the target species transcriptome is represented in the de novo TA.”

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

Chestnut


scaffold00816
0k
10k
20k
30k
40k
50k
60k
70k
80k
90k
10 kbp
63k
64k
65k
66k
67k
68k
69k
70k
71k
72k
73k
74k
75k
76k
77k
78k
79k
80k
81k
82k
83k
84k
85k
86k
87k
88k
89k
90k
maker genes
maker-scaffold00816-augustus-gene-0.44
maker est2genome
CCall_contig1663_v2
CCall_contig11727_v2
CCall_contig22371_v2
CCall_contig24466_v2
CCall_contig36997_v2
CCall_contig26755_v2
CCall_contig20311_v2
CCall_contig1766_v2
CCall_contig8794_v2
CCall_contig46287_v2

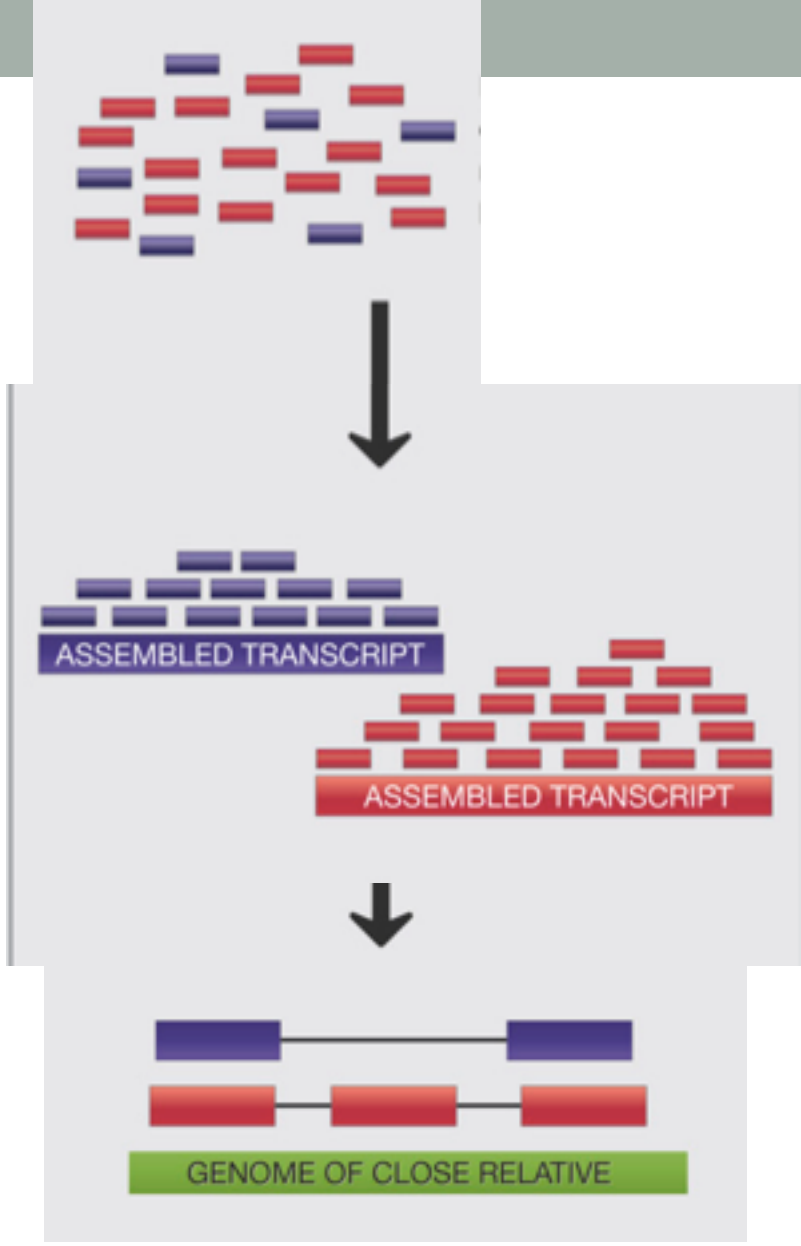
De novo transcriptome assemblies

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

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

Green Ash (*Fraxinus pennsylvanica*)

	Trimmed reads	Trimmed bases
ozone project (miseq)	5,151,500	750,286,276
Tissues (miseq)	21,362,330	2,926,958,573
Tissues (hiseq)	442,863,286	42,122,511,244
Stress (miseq)	27,470,000	3,650,984,673
Stress (hiseq)	350,952,104	35,411,991,796
Data	**847,799,220**	**84,862,732,562**

	Green Ash
transcripts	107,611
peptides	52,899
% ORF discovery	49%

55 libraries
Plus
41 technical replicates

Ash Genome

- Richard Buggs' lab

  Queen Mary, University of London (QMUL)
- British Ash Tree Genome Project

  Fraxinus excelsior
- 89,285 scaffolds, with an N50 of 99 kbp, and total size of 875 Mbp

- 36,944 genes
- 36,893 proteins

	Green Ash
transcripts	107,611
peptides	52,899
% ORF discovery	49%

Ash Genome

From perspective of the genome

- 36,893 proteins from genome
  - 29,782 have a match to our RNASeq proteome (81%)
- 36,944 genes from genome
  - 35,298 have a match to our RNASeq transcripts (96%)

From perspective of the transcriptome

- 52,899 proteins from RNASeq
  - 47,657 have a match to the genome proteins (90%)
- 107,611 transcripts from RNASeq
  - 80,628 have a match to the genome transcripts (75%)

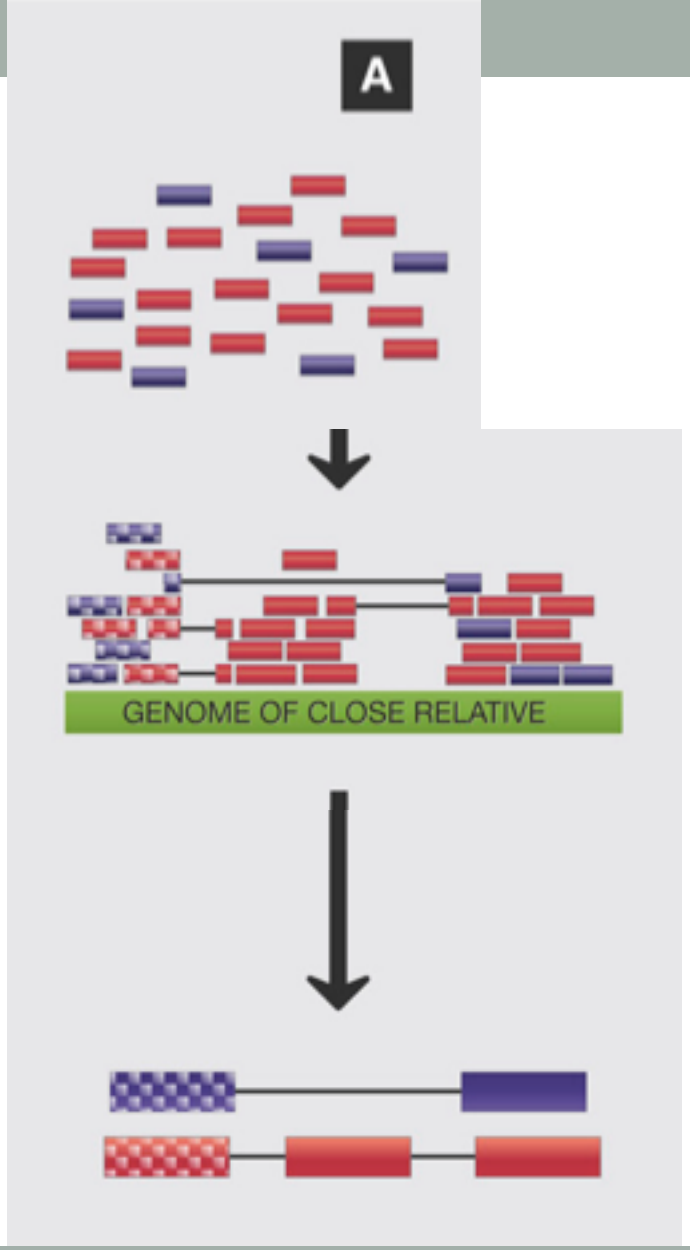
BLAST 1e-10

Align then assemble

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

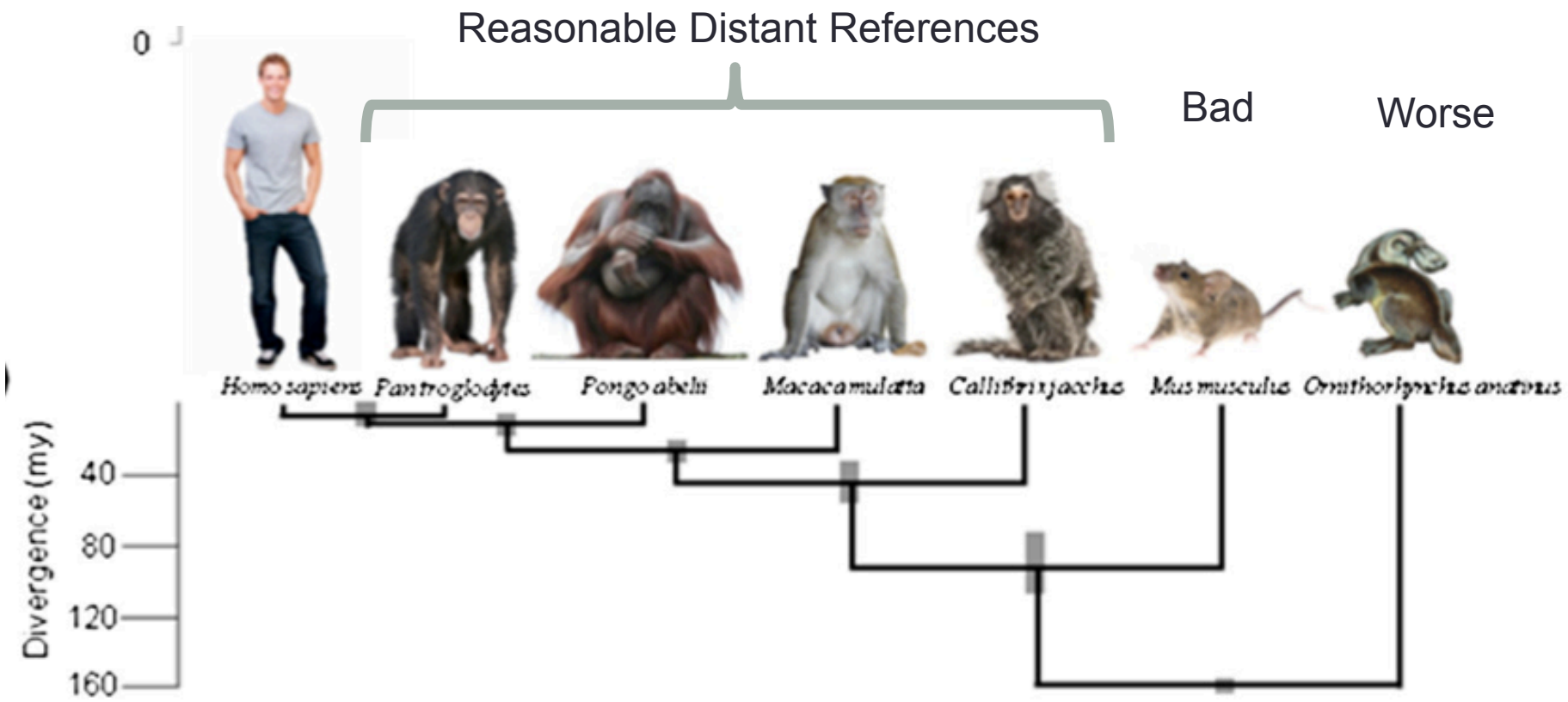
- Main Problems
  - Read alignment may be poor due to lack of sequence similarity
  - Gene family expansion/contraction

- 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

Assemble then align



*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

De novo transcriptome assemblies

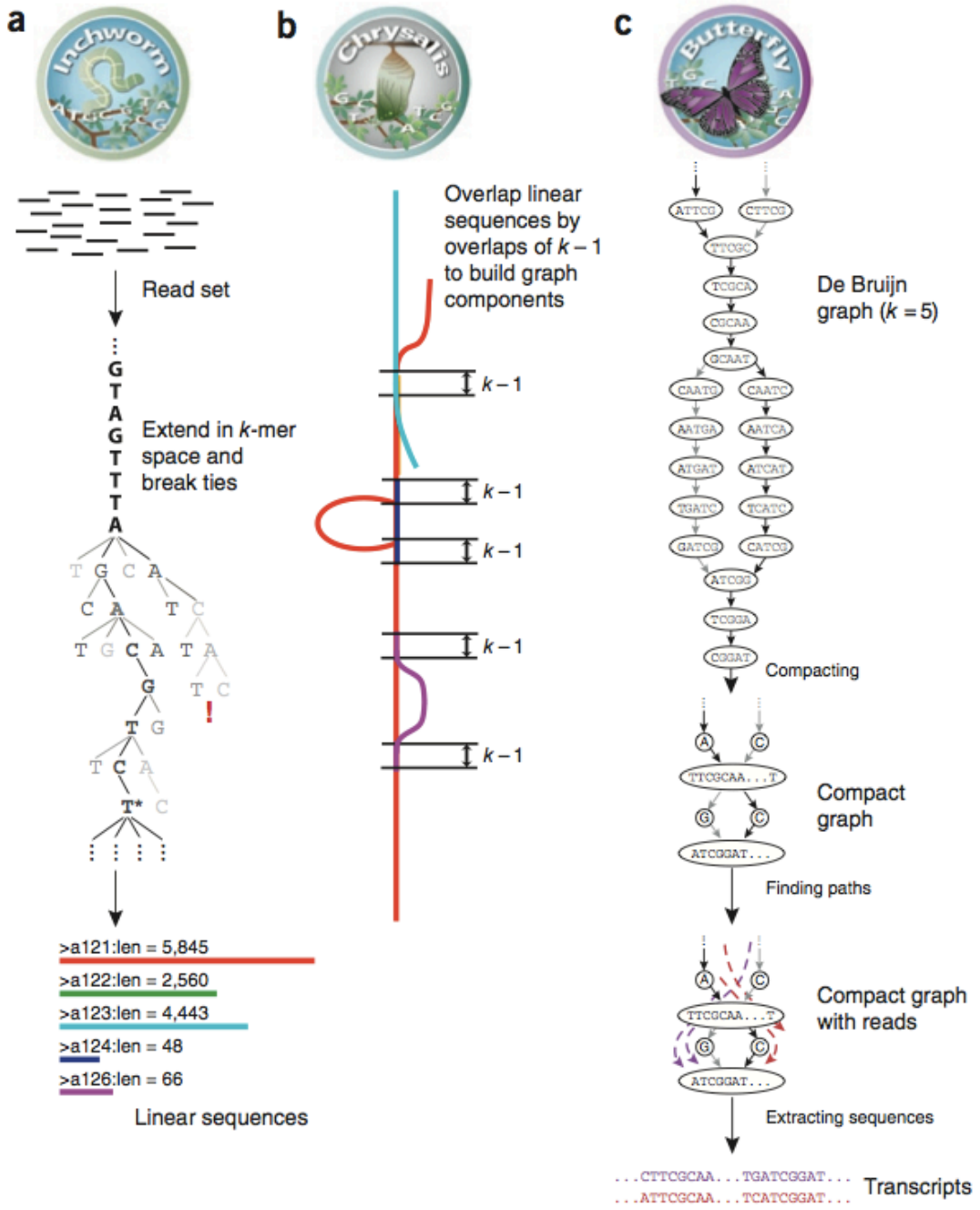
- Completely reference free
- What are they useful for?
  - Transcriptome characterization
    - Whats there?
    - Resource Building
  - Enabling proteomics experiments
  - Candidate gene discovery
  - Targeted sequencing
  - Marker discovery/development
    - Sequence parents of a cross
    - SNP array
    - (May want to consider genotyping by sequencing/restriction site associated DNA techniques instead)

De novo transcriptome assemblies

- What to do if you want/need differential expression data?
  - Long reads
  - Paired ends, possibly with different insert sizes
  - Analyze gene families for differential expression instead of individual genes
  - Alter the parameters of your assembler
    - Merge at a lower level of heterozygosity – 98% or 97%
  - Utilize a closely related relative with a sequenced genome

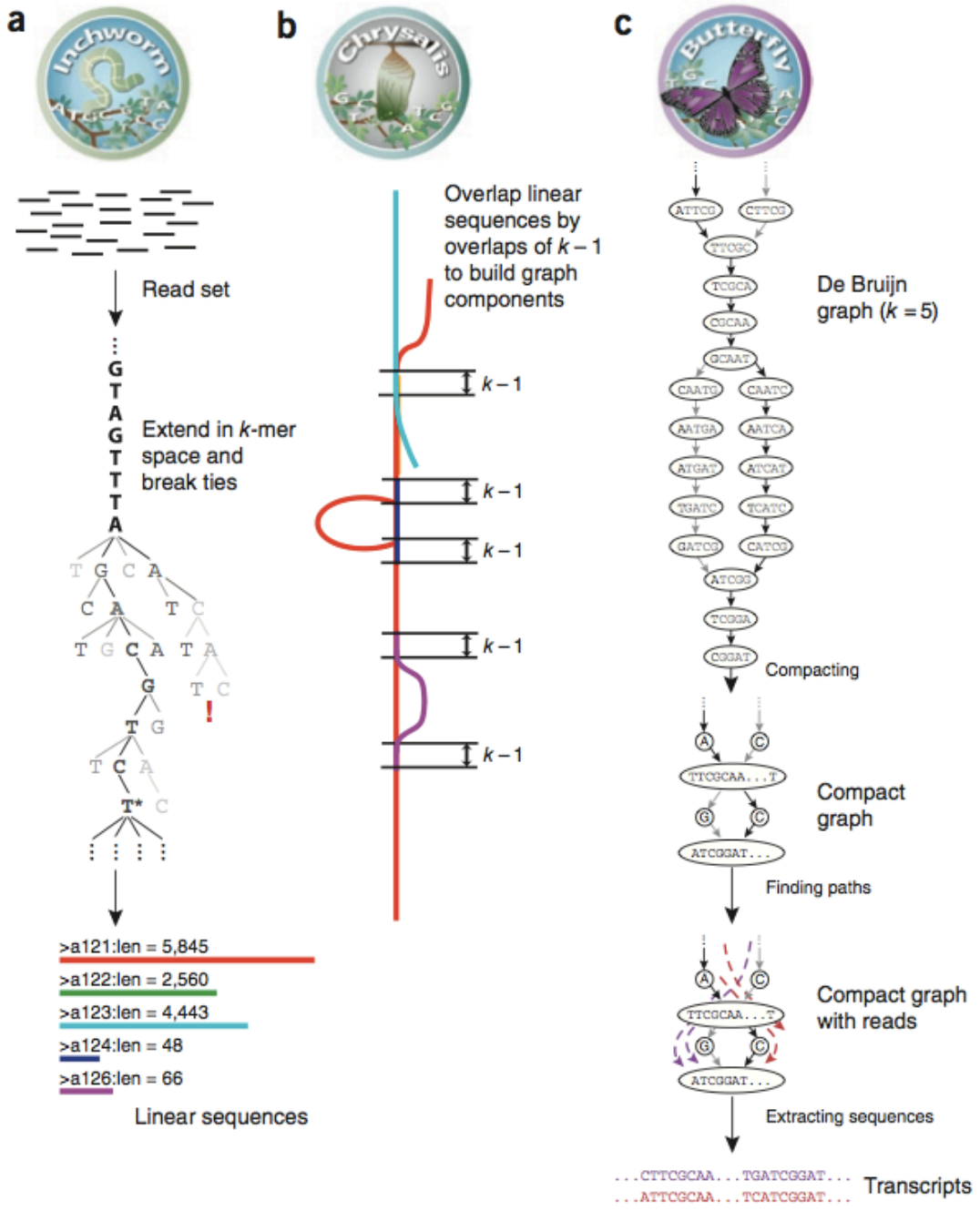
Trinity strategy

Inchworm assembles the RNA-seq data into the unique sequences of transcripts, often generating full-length transcripts for a dominant isoform, but then reports just the unique portions of alternatively spliced transcripts.



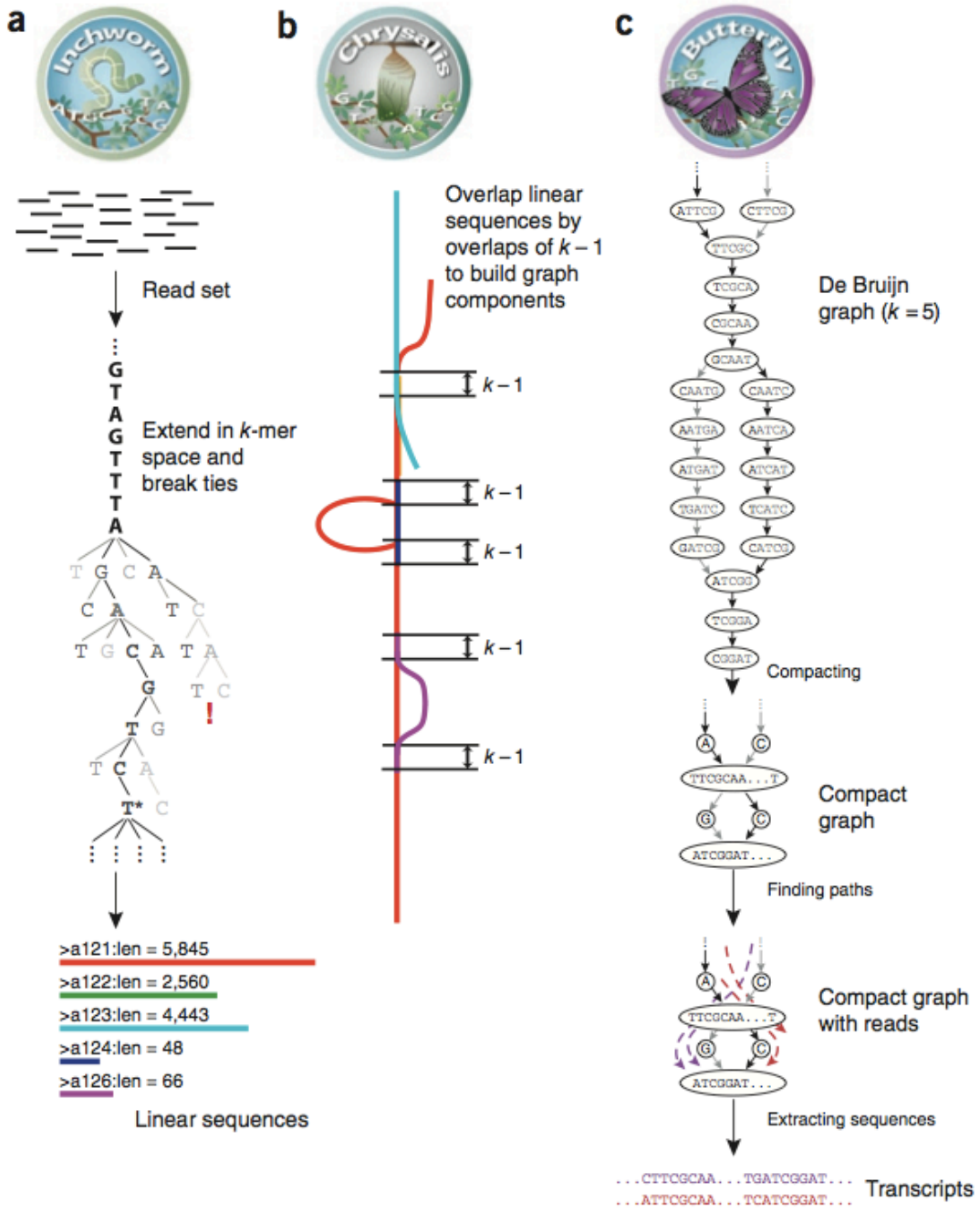
Trinity strategy

Chrysalis clusters the Inchworm contigs into clusters and constructs complete de Bruijn graphs for each cluster. Each cluster represents the full transcriptonal complexity for a given gene (or sets of genes that share sequences in common). Chrysalis then partitions the full read set among these disjoint graphs.



Trinity strategy

Butterfly then processes the individual graphs in parallel, tracing the paths that reads and pairs of reads take within the graph, ultimately reporting full-length transcripts for alternatively spliced isoforms, and teasing apart transcripts that corresponds to paralogous genes.



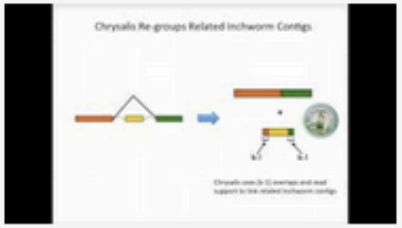
Trinity

Videos!

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

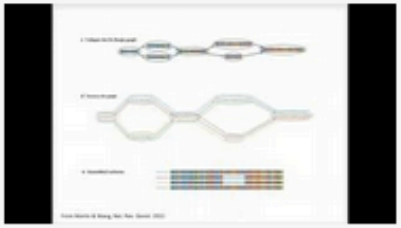
A Collection of new RNA-Seq Videos from The Broad Institute

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



7:30

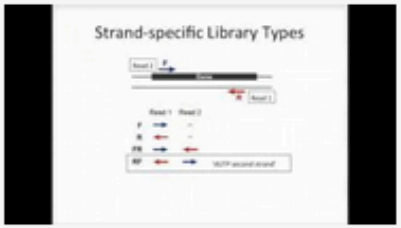
BroadE: Trinity – How it works


5:20

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

5:38

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



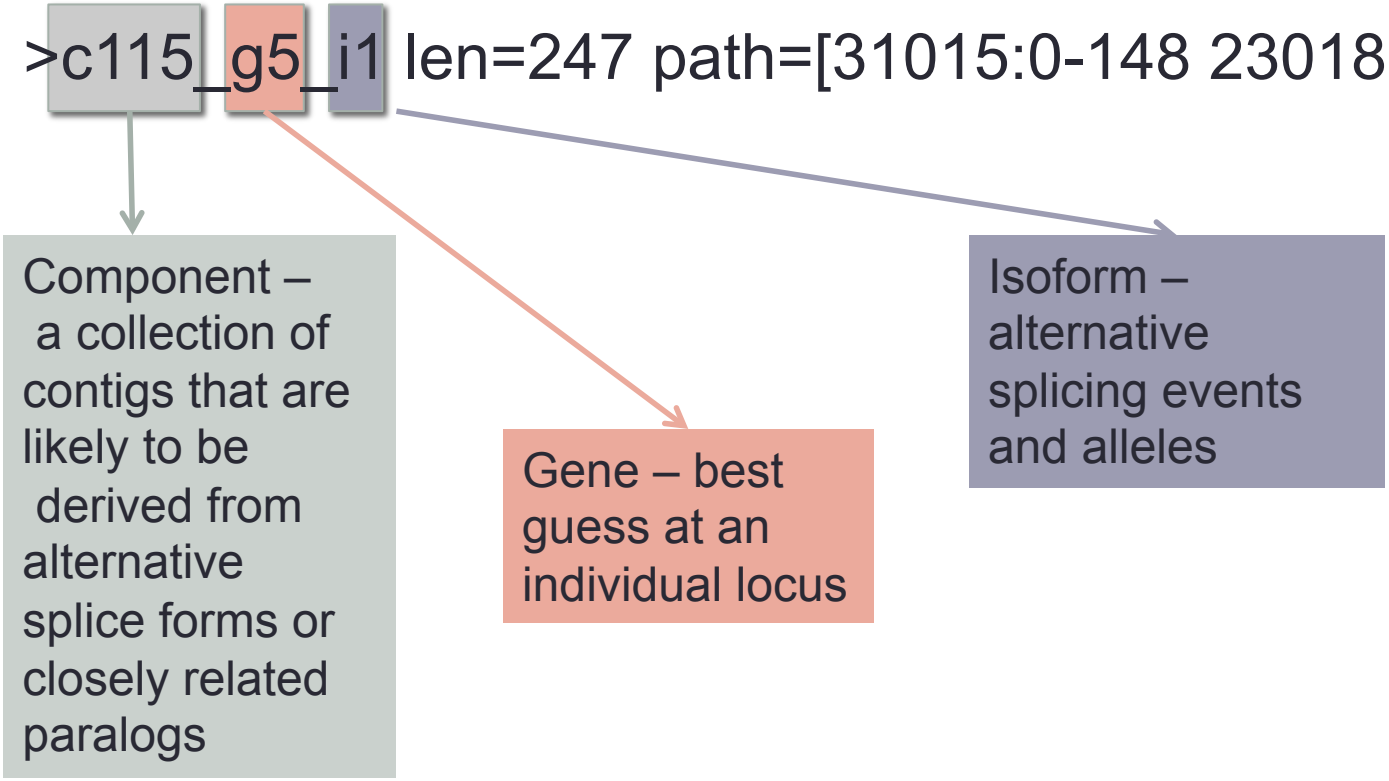
3:00

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]



II. De novo transcriptome sequencing – after assembly

Trinity TransDecoder and Coverage

- Maximizes length and likelihood score of ORF
- Optionally, looks for a putative peptide that has a match to a Pfam domain

- Full-length transcript analysis for model and non-model organisms using BLAST+
- Perl script analyze_blastPlus_topHit_coverage.pl

hit_pct_cov_bin	count_in_bin	>bin_below
100	3242	3242
90	268	3510
80	186	3696
70	202	3898
60	216	4114
50	204	4318
40	164	4482
30	135	4617
20	76	4693
10	0	4693
0	0	4693

Functional Annotation

InterProScan →

- 329,311 annotations
- 45,893 transcripts have at least one annotation (87%)
- 234,546 GO term assignments
- 29,666 transcripts with go terms (56%)

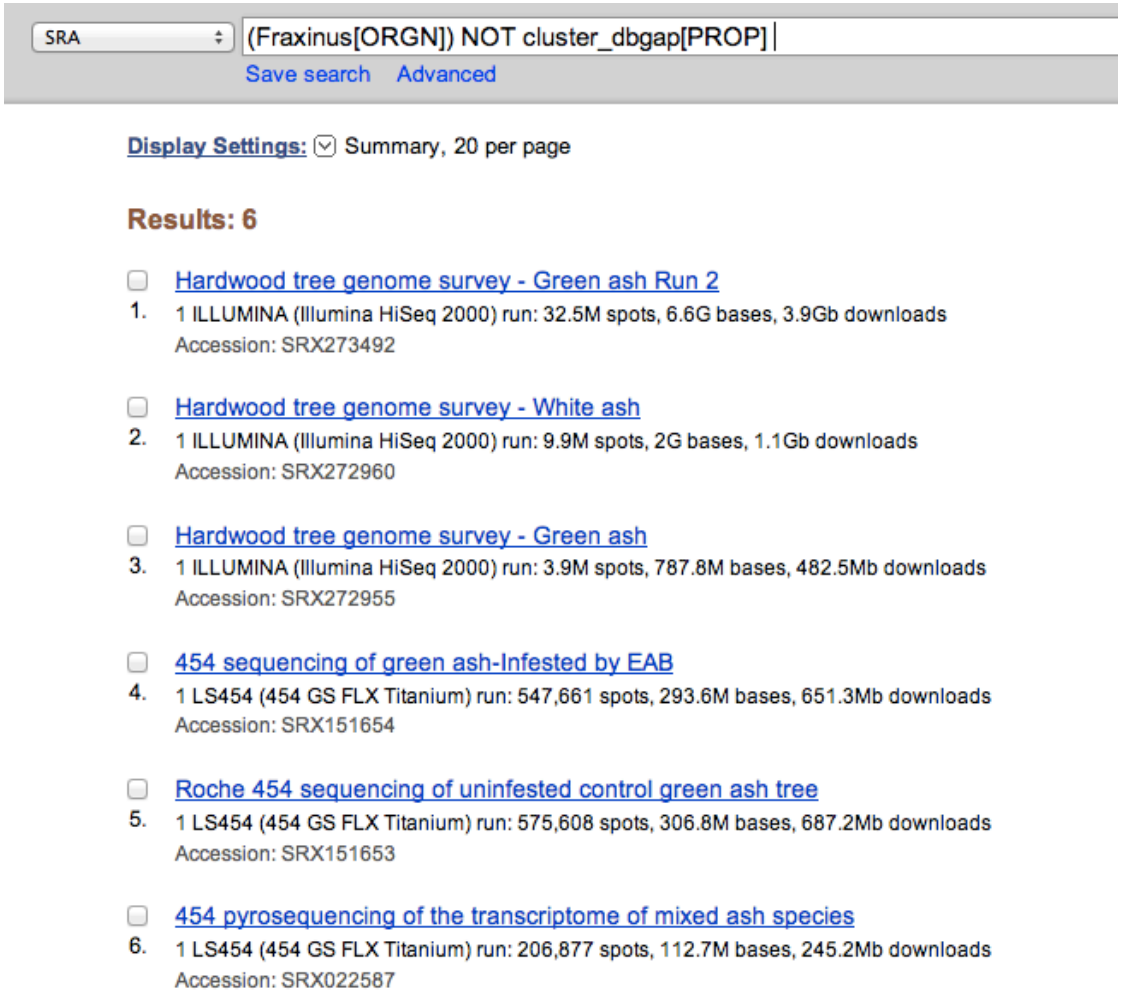
Software:

72,706 PANTHER
51,025 Pfam
41,391 Gene3d
39,965 SUPERFAMILY
27,246 TMHMM
26,189 ProSiteProfiles
20,835 PRINTS
20,078 SMART
8,267 Coils
5,780 TIGR-FAM
2,456 SignalP_EUK
1,224 PIRSF
825 HAMAP

Making data public

- NCBI Short Read Archive
  - stores raw sequence data from "next-generation" sequencing technologies including 454, IonTorrent, Illumina, SOLiD, Helicos and Complete Genomics. In addition to raw sequence data, SRA now stores alignment information in the form of read placements on a reference sequence.

NCBI SRA



SRA format – includes data and metadata

Convert using the SRA Toolkit (linux, mac and windows versions available)

http://www.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?cmd=show&f=software&m=software&s=software

Upload to SRA

Gather information

Why did you perform your analysis?

- Project title and abstract
- Aims and objectives
- Organism(s) sequenced
- Optional: Funding sources, publications, etc.

What did you sequence?

- Descriptive sample information
- Tabular format is ideal
- Examples: Organism(s), age(s), gender(s), location data, cell line(s), etc.

How did you sequence your samples?

- Sequencing methods
- Kits used
- Instrument model(s)

What is your data file format?

- Files in acceptable format(s): BAM, FASTQ, etc.
- MD5 checksum for each file
- Minimum of 1 unique dataset per sample

BioProject → **SRA Study**

- A description of the research effort
- "Why" you sequenced your samples

BioSample → **SRA Sample**

- A description of biologically or physically unique specimens
- "What" you sequenced

SRA Experiment

- A description of a sample-specific sequencing library
- "How" you performed the sequencing
- Multiple Experiments can "point" to a single Sample, but not vice-versa

SRA Run

- All files linked to a Run are "merged" into a single dataset
- Files are converted to SRA format
- Files submitted by FTP or Aspera once steps 1 and 2 are complete

1 2 3

NSF Hardwood Genomics Project