


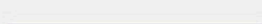
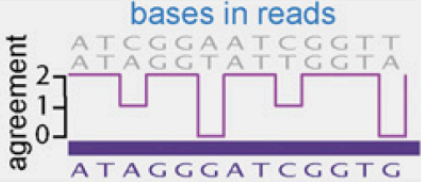



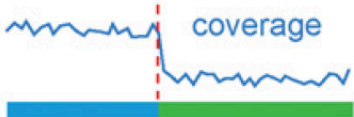

















Evaluating the quality of your transcriptome assembly



De novo Transcriptome Assembly is Prone to Certain Types of Errors

Error type	Transcripts	Assembly	Read evidence
Family collapse	geneAA  geneAB  geneAC  n=3	 n=1	 <p>bases in reads</p> <p>ATCGGAATCGGTT ATAGGTATTGGTA</p> <p>agreement</p> <p>ATAGGGATCGGTG</p>
Chimerism	 geneC  geneB n=2	 n=1	 <p>coverage</p>
Unsupported insertion	 n=1	 n=1	no reads align to insertion 
Incompleteness	 n=1	 n=1	read pairs align off end of contig 
Fragmentation	 n=1	 n=4	bridging read pairs 
Local misassembly	 n=1	 n=1	read pairs in wrong orientation 
Redundancy	 n=1	 n=3	all reads assign to best contig 



TransRate

1 input data

assembled contigs paired-end reads



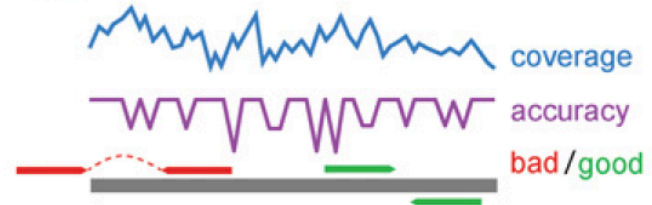
2 align reads to contigs



3 assign multimapping reads



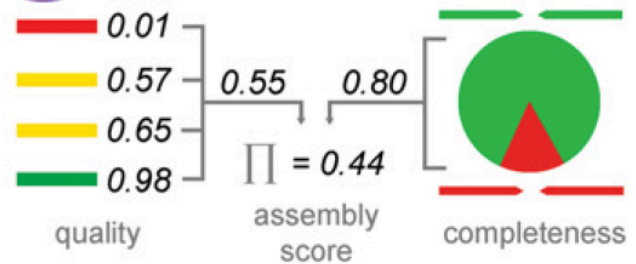
4 collect contig score components



5 calculate contig scores



6 calculate assembly score



Simple Quantitative and Qualitative Assembly Metrics

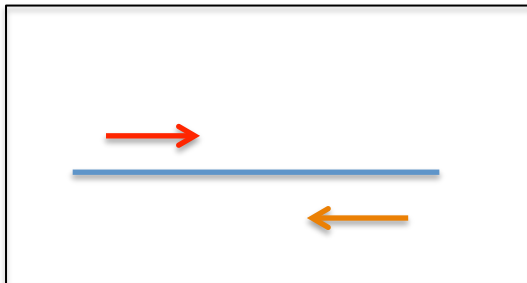
Read representation by assembly

Align reads to the assembled transcripts using Bowtie.

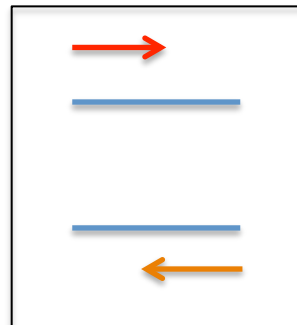
A typical 'good' assembly has ~80 % reads mapping to the assembly and ~80% are properly paired.

Given read pair:   Possible mapping contexts in the Trinity assembly are reported:

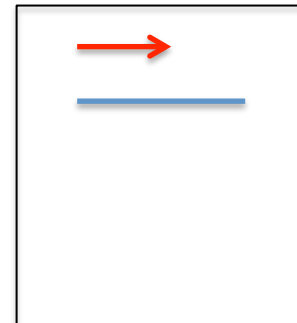
Proper pairs



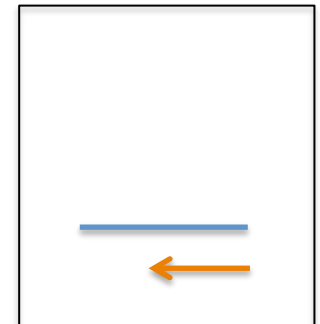
Improper pairs



Left only



Right only



Assembled transcript contig is only as good as its read support.

% samtools tview alignments.bam target.fasta

```
911 921 931 941 951 961 971 981 991 1001 1011 1021 1031 1041 1051 1061 1071
GTAGGTTAATTTTCATCTTCTAATTTAGAATCTTGCCAATCAAGCCCTCTCGAAGTTGGCAATATCTATAACTCAACCTCTGCTTCTGAGATTCTAAGTACCTTAGATGCCAAGTACATTACTATAATTGGTGTATCGGGTCTTCCAACCTCTCCATTCAAGACTTAATTGACTCTGT
-----
GT GTTTAATTTTCATCTTCTAATTTAGAATCTTGCCAATCAAGCCCTCTCGAAGTTGGCAATATCTATAAC ctgcttctgagattctaagtaccttagatgccaagtagcattactataaattgggtttatcgggtcttcc ctctccattcaagacttaattgactctgt
GT ATTTTCATCTTCTAATTTAGAATCTTGCCAATCAAGCCCTCTCGAAGTTGGCAATATCTATAACTCAAC tgcttctgagattctaagtaccttagatgccaagtagcattactataaattgggtttatcgggtcttcca ctctccattcaagacttaattgactctgt
GT atttcatcttctaatttagaactctgccaatcaagccctctcgaagttggcaatatactataactcaac GCTTCTGAGATTCTAAGTACCTTAGATGCCAAGTACATTACTATAAATTGGTGTATCGGGTCTTCCAA ctctccattcaagacttaattgactctgt
GT atttcatcttctaatttagaactctgccaatcaagccctctcgaagttggcaatatactataactcaac GCTTCTGAGATTCTAAGTACCTTAGATGCCAAGTACATTACTATAAATTGGTGTATCGGGTCTTCCAA ctctccattcaagacttaattgactctgt
GTAGGTTAAT aatcttgccaatcaagccctctcgaagttggcaatatactataactcaacctctgcttctgagattcta CTTAGATGCCAAGTACATTACTATAAATTGGTGTATCGGGTCTTCCAACCTCTCCATTCAAGACTTAA ctgt
GTAGGTTAATTT tcttgccaatcaagccctctcgaagttggcaatatactataactcaacctctgcttctgagattctaag CTTAGATGCCAAGTACATTACTATAAATTGGTGTATCGGGTCTTCCAACCTCTCCATTCAAGACTTAA
GTAGGTTAATTTTCATCTT cttgccaatcaagccctctcgaagttggcaatatactataactcaacctctgcttctgagattctaag TTAGATGCCAAGTACATTACTATAAATTGGTGTATCGGGTCTTCCAACCTCTCCATTCAAGACTTAA
GTAGGTTAATTTTCATCTTC TGCCAATCAAGCCCTCTCGAAGTTGGCAATATCTATAACTCAACCTCTGCTTCTGAGATTCTAAGTAC ATGCCAAGTACATTACTATAAATTGGTGTATCGGGTCTTCCAACCTCTCCATTCAAGACTTAAATTGAC
GTAGGTTAATTTTCATCTTCTAAT TGCCAATCAAGCCCTCTCGAAGTTGGCAATATCTATAACTCAACCTCTGCTTCTGAGATTCTAAGTAC GCCAAGTACATTACTATAAATTGGTGTATCGGGTCTTCCAACCTCTCCATTCAAGACTTAAATTGACTC
gtaggtttaatttcattctctaatttag TGCCAATCAAGCCCTCTCGAAGTTGGCAATATCTATAACTCAACCTCTGCTTCTGAGATTCTAAGTAC CATTACTATAAATTGGTGTATCGGGTCTTCCAACCTCTCCATTCAAGACTTAAATTGACTCTGT
GTAGGTTAATTTTCATCTTCTAATTTAG GCCAATCAAGCCCTCTCGAAGTTGGCAATATCTATAACTCAACCTCTGCTTCTGAGATTCTAAGTACC cactactataaattgggtttatcgggtcttccaactctccattcaagacttaattgactctgt
GTAGGTTAATTTTCATCTTCTAATTTAG CAATCAAGCCCTCTCGAAGTTGGCAATATCTATAACTCAACCTCTGCTTCTGAGATTCTAAGTACC tgttatcgggtcttccaactctccattcaagacttaattgactctgt
GTAGGTTAATTTTCATCTTCTAATTTAG CAATCAAGCCCTCTCGAAGTTGGCAATATCTATAACTCAACCTCTGCTTCTGAGATTCTAAGTACC ttgttatcgggtcttccaactctccattcaagacttaattgactctgt
GTAGGTTAATTTTCATCTTCTAATTTAG gccctctcgaagttggcaatatactataactcaacctctgcttctgagattctaagtagcttagatgcc GGCTCTCCAACCTCTCCATTCAAGACTTAAATTGACTCTGT
GTAGGTTAATTTTCATCTTCTAATTTAGAAT CCTCTCGAAGTTGGCAATATCTATAACTCAACCTCTGCTTCTGAGATTCTAAGTACCTTAGATGCCA
GTAGGTTAATTTTCATCTTCTAATTTAGAAT cttcgaagttggcaatatactataactcaacctctgcttctgagattctaagtagcttagatgccaag ggcttccaactctccattcaagacttaattgactctgt
GTAGGTTAATTTTCATCTTCTAATTTAGAATCT CTTGCAAGTTGGCAATATCTATAACTCAACCTCTGCTTCTGAGATTCTAAGTACCTTAGATGCCAAGTAC GTCTTCCAACCTCTCCATTCAAGACTTAAATTGACTCTGT
GTAGGTTAATTTTCATCTTCTAATTTAGAATCT CGAAGTTGGCAATATCTATAACTCAACCTCTGCTTCTGAGATTCTAAGTACCTTAGATGCCAAGTACAC gttctccaactctccattcaagacttaattgactctgt
GTAGGTTAATTTTCATCTTCTAATTTAGAATCT AAGTTGGCAATATCTATAACTCAACCTCTGCTTCTGAGATTCTAAGTACCTTAGATGCCAAGTACATT ctccaactctccattcaagacttaattgactctgt
gtaggtttaatttcattctctaatttagaactctgcca CAATATCTATAACTCAACCTCTGCTTCTGAGATTCTAAGTACCTTAGATGCCAAGTACATTACTATAA ctccaactctccattcaagacttaattgactctgt
GTAGGTTAATTTTCATCTTCTAATTTAGAATCTTGCCA CTATAACTCAACCTCTGCTTCTGAGATTCTAAGTACCTTAGATGCCAAGTACATTACTATAAATTGGTGT CTTCCAACCTCTCCATTCAAGACTTAAATTGACTCTGT
GTAGGTTAATTTTCATCTTCTAATTTAGAATCTTGCCA cttctgagattctaagtagccttagatgccaagtagcattactataaattgggtttatcgggtcttccaac CTCCATTCAAGACTTAAATTGACTCTGT
gtaggtttaatttcattctctaatttagaactctgccaatcaagcc cttctgagattctaagtagccttagatgccaagtagcattactataaattgggtttatcgggtcttccaac tccattcaagacttaattgactctgt
GTAGGTTAATTTTCATCTTCTAATTTAGAATCTTGCCA cttctgagattctaagtagccttagatgccaagtagcattactataaattgggtttatcgggtcttccaac tccattcaagacttaattgactctgt
gtaggtttaatttcattctctaatttagaactctgccaatcaagccc ttctgagattctaagtagccttagatgccaagtagcattactataaattgggtttatcgggtcttccaact tccattcaagacttaattgactctgt
GTAGGTTAATTTTCATCTTCTAATTTAGAATCTTGCCA tgagattctaagtagccttagatgccaagtagcattactataaattgggtttatcgggtcttccaactcc ccattcaagacttaattgactctgt
GTAGGTTAATTTTCATCTTCTAATTTAGAATCTTGCCA agattctaagtagccttagatgccaagtagcattactataaattgggtttatcgggtcttccaactctct cattcaagacttaattgactctgt
GTAGGTTAATTTTCATCTTCTAATTTAGAATCTTGCCA agattctaagtagccttagatgccaagtagcattactataaattgggtttatcgggtcttccaactctcc AAGACTTAAATTGACTCTGT
GTAGGTTAATTTTCATCTTCTAATTTAGAATCTTGCCA ATTTTCATCTTCTAATTTAGAATCTTGCCAATCAAGCCCTCTCGAAGTTGGCAATATCTATAACTCAAC cttaattgactctgt
TCATCTTCTAATTTAGAATCTTGCCAATCAAGCCCTCTCGAAGTTGGCAATATCTATAACTCAAC attgactctgt
gattctaagtagccttagatgccaagtagcattactataaattgggtttatcgggtcttccaactctcca
gattctaagtagccttagatgccaagtagcattactataaattgggtttatcgggtcttccaactctcca
aagtaccttagatgccaagtagcattactataaattgggtttatcgggtcttccaactctcca
aagtaccttagatgccaagtagcattactataaattgggtttatcgggtcttccaactctccattcaag
ctccaactctccattcaagacttaattgactctgt
TTCCAACCTCTCCATTCAAGACTTAAATTGACTCTGT
TCCAACCTCTCCATTCAAGACTTAAATTGACTCTGT
caactctccattcaagacttaattgactctgt
caactctccattcaagacttaattgactctgt
aactctccattcaagacttaattgactctgt
aactctccattcaagacttaattgactctgt
tccattcaagacttaattgactctgt
ccattcaagacttaattgactctgt
ccattcaagacttaattgactctgt
```

IGV

← → ↻ www.broadinstitute.org/igv/ ☆ a



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

Search website

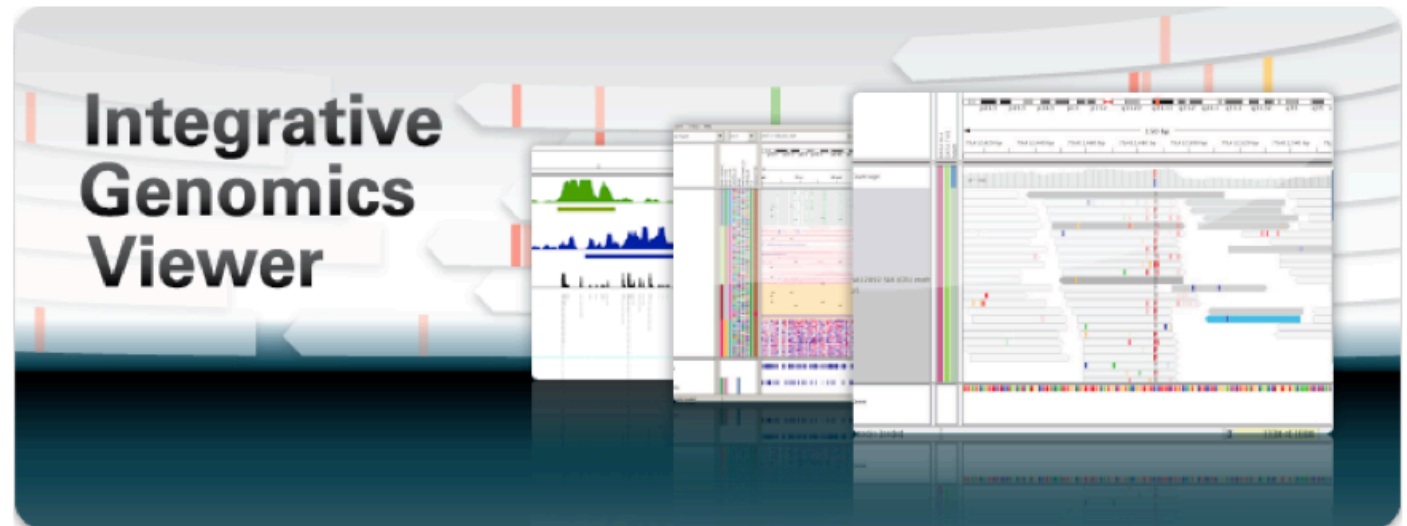
search

[Broad Home](#)
[Cancer Program](#)



© 2012 Broad Institute

Home



What's New



July 3, 2012. Soybean (*Glycine max*) and Rat (*m5*) 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:

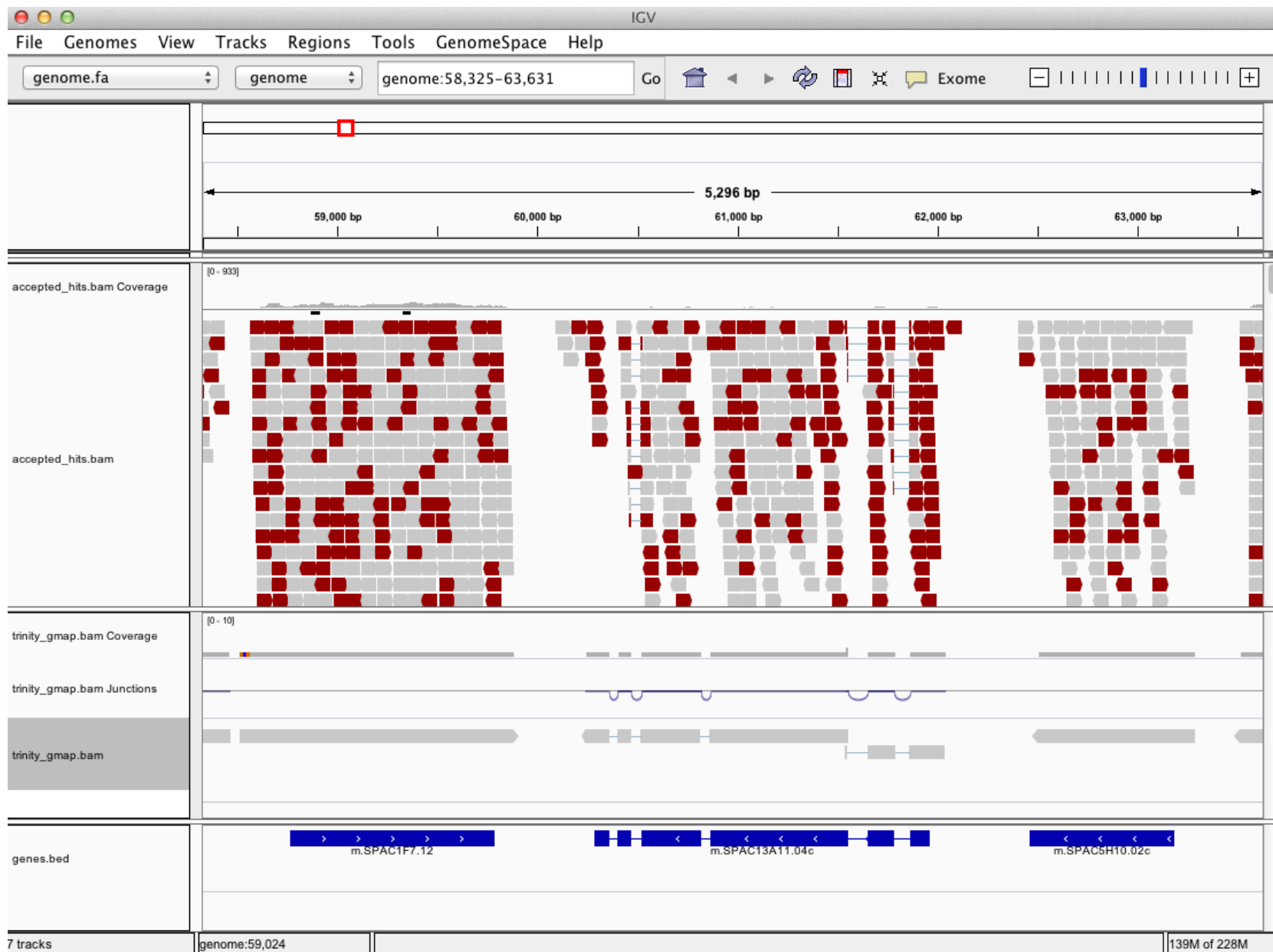
James T. Robinson, Helga Thorvaldsdóttir, Wendy Winckler, Mitchell Guttman, Eric S. Lander, Gad Getz, Jill P. Mesirov. [Integrative Genomics Viewer](#). *Nature Biotechnology* 29, 24–26 (2011), or

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

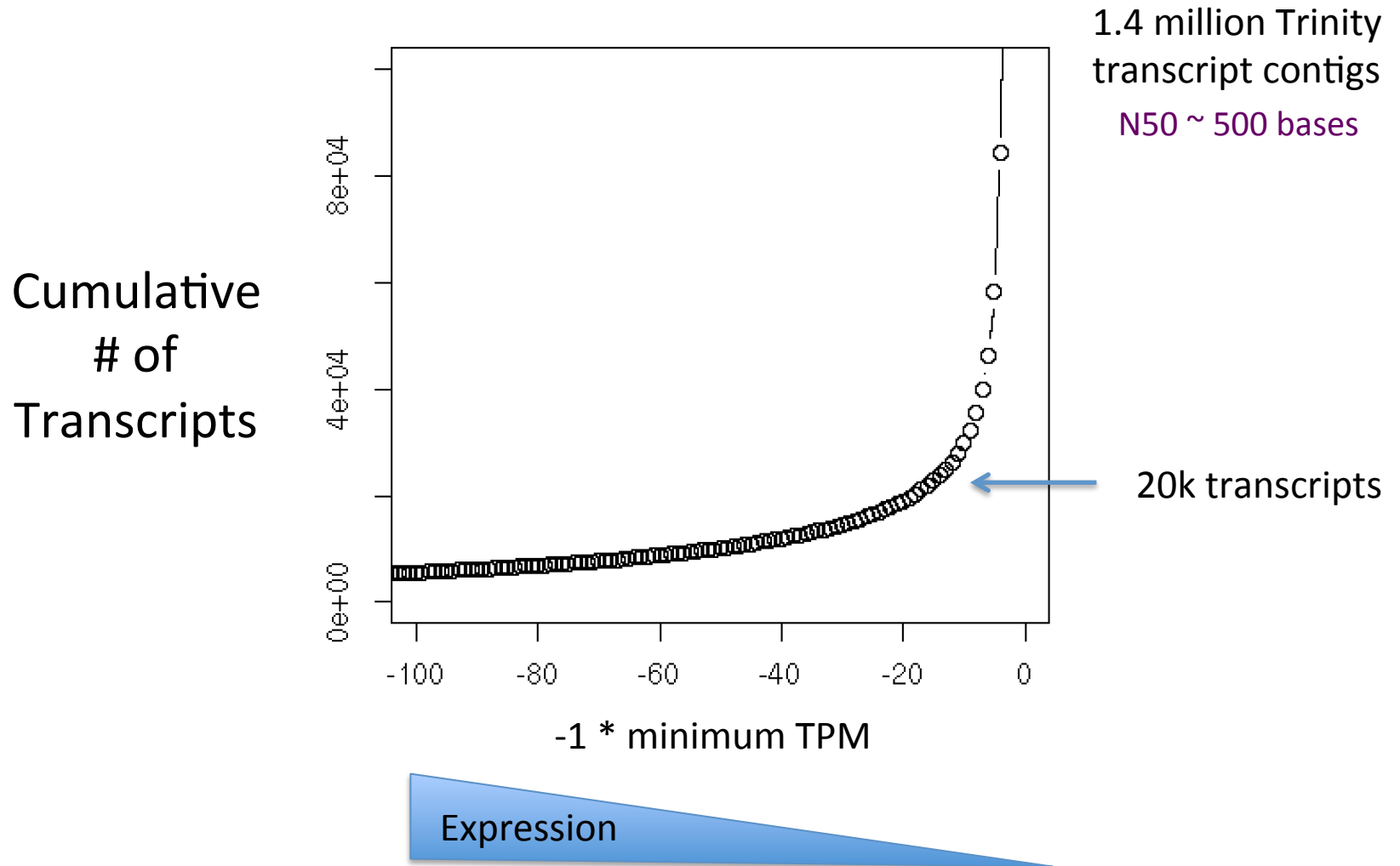
Can Examine Transcript Read Support Using IGV



Can align Trinity transcripts to genome scaffolds to examine intron/exon structures (Trinity transcripts aligned to the genome using GMAP)

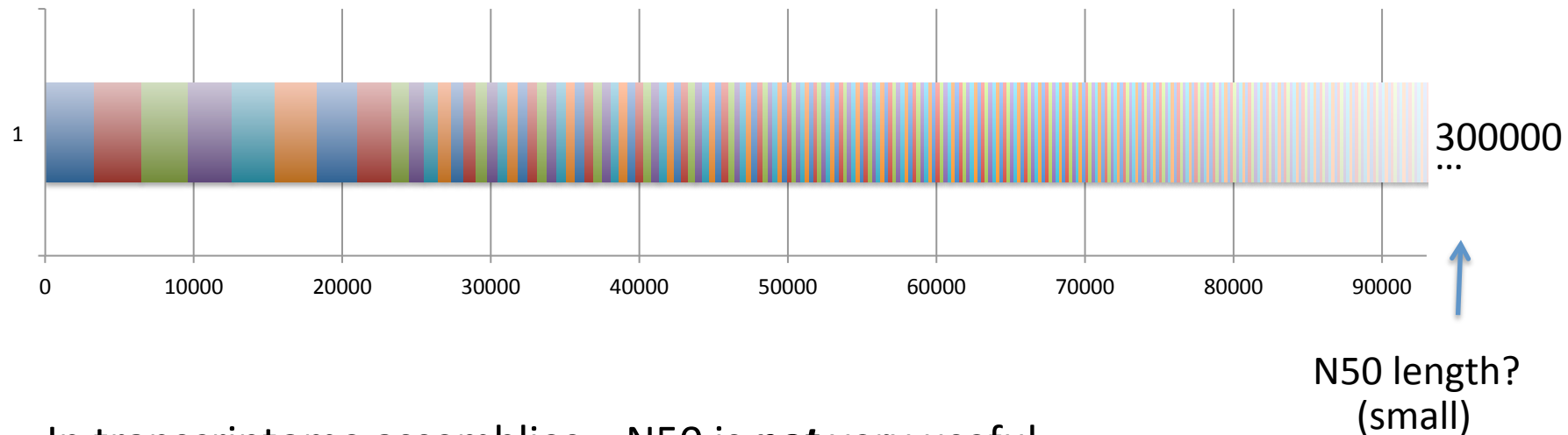


Often, most assembled transcripts are *very* lowly expressed
(How many 'transcripts & genes' are there really?)



* Salamander transcriptome

N50 Calculation for *Transcriptome Assemblies*??

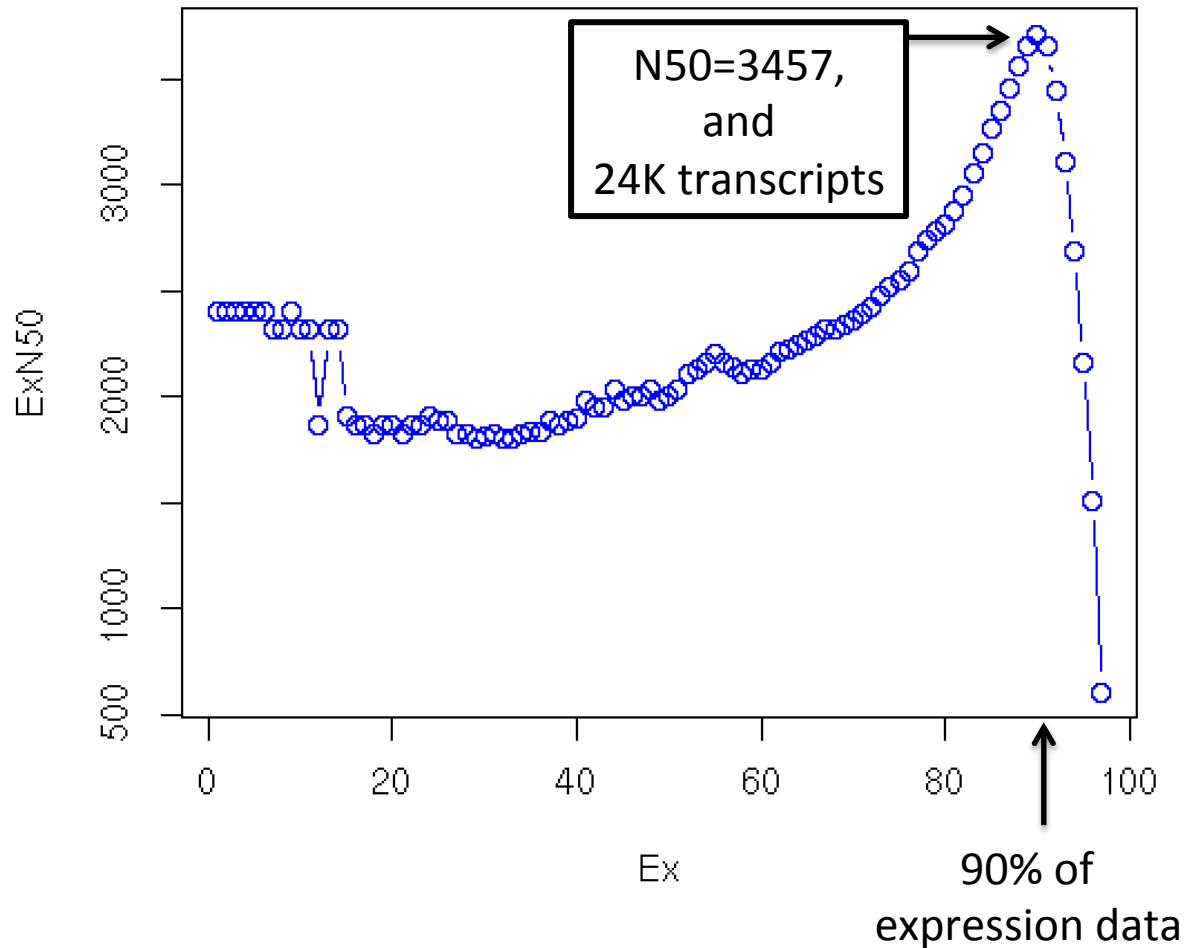


In transcriptome assemblies – N50 is *not* very useful.

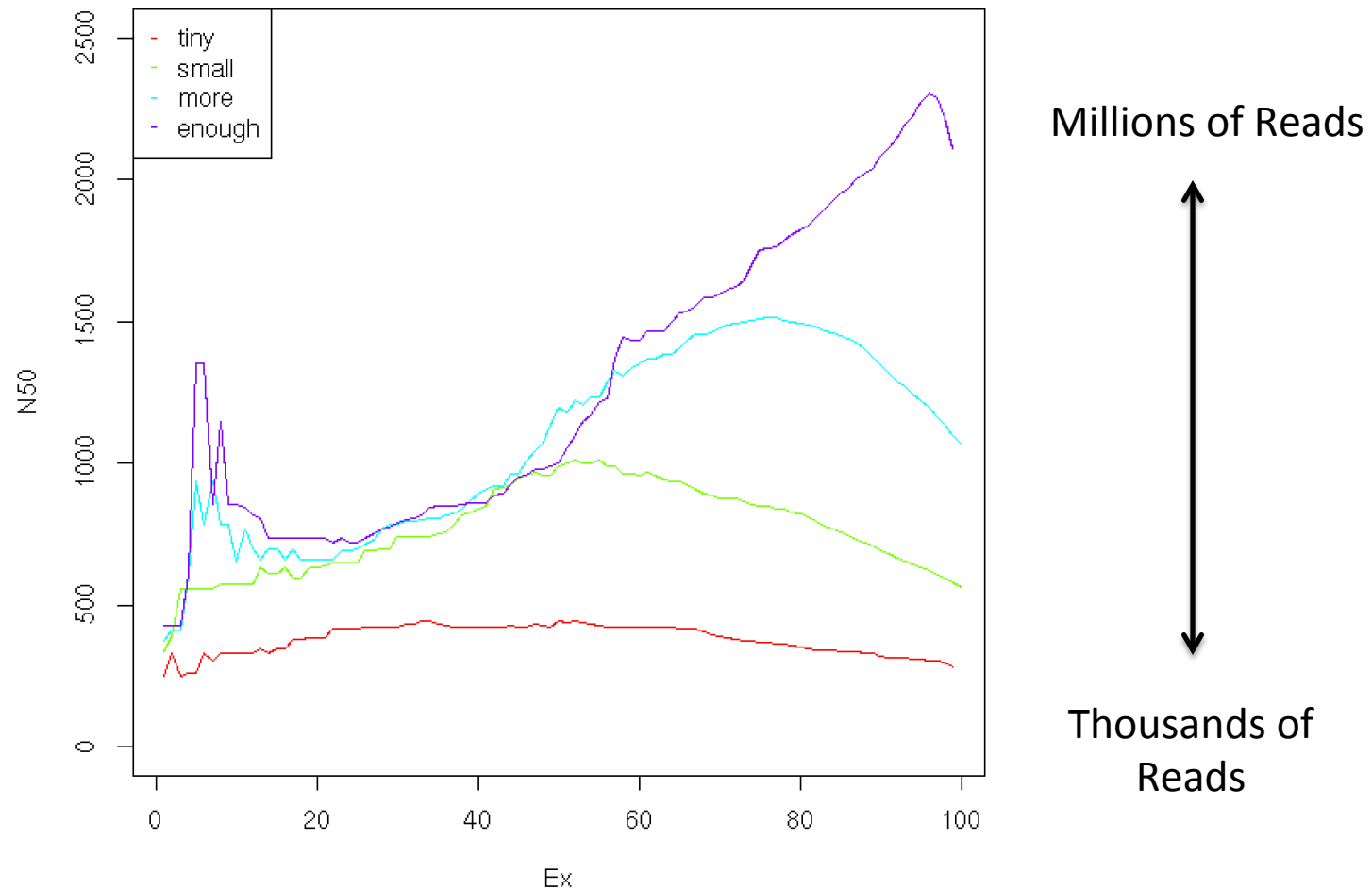
- Overzealous isoform annotation for long transcripts drives higher N50
- Very sensitive reconstruction for short lowly expressed transcripts drives lower N50

Compute N50 Based on the Top-most Highly Expressed Transcripts (ExN50)

- Sort contigs by expression value, descendingly.
- Compute N50 given minimum % total expression data thresholds => ExN50



ExN50 Profiles for Different Trinity Assemblies Using Different Read Depths

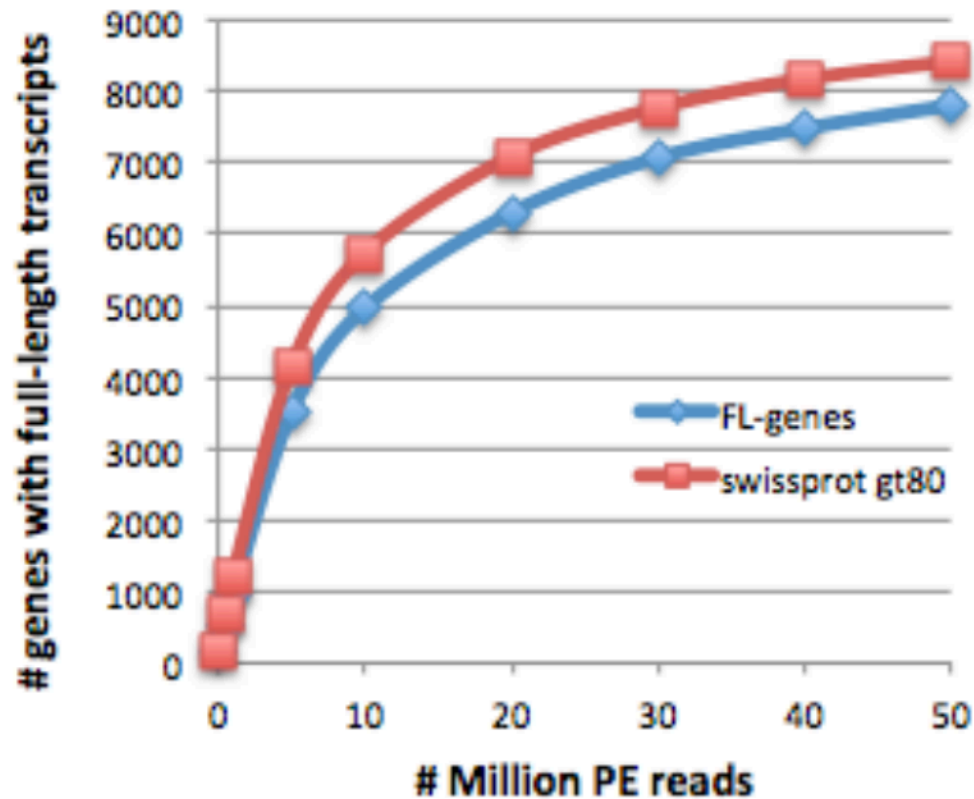


Note shift in ExN50 profiles as you assemble more and more reads.

* Candida transcriptome

Evaluating the quality of your transcriptome assembly

Full-length Transcript Detection via BLASTX



Have you
sequenced
deeply
enough?

* Mouse transcriptome



Assessing genome assembly and annotation completeness with **B**enchmarking **U**niversal **S**ingle-**C**opy **O**rthologs

About BUSCO

BUSCO v2 provides quantitative measures for the assessment of genome assembly, gene set, and transcriptome completeness, based on evolutionarily-informed expectations of gene content from near-universal single-copy orthologs selected from [OrthoDB v9](#).

BUSCO assessments are implemented in open-source software, with a large selection of lineage-specific sets of Benchmarking Universal Single-Copy Orthologs. These conserved orthologs are ideal candidates for large-scale phylogenomics studies, and the annotated BUSCO gene models built during genome assessments provide a comprehensive gene predictor training set for use as part of genome annotation pipelines.



Assessing genome assembly and annotation completeness with Benchmarking Universal Single-Copy Orthologs

#Summarized BUSCO benchmarking for file: Trinity.fasta
#BUSCO was run in mode: trans

Summarized benchmarks in BUSCO notation:
C:88%[D:53%],F:4.5%,M:7.3%,n:3023

Representing:

- 1045 Complete Single-copy BUSCOs
- 1617 Complete Duplicated BUSCOs
- 139 Fragmented BUSCOs
- 222 Missing BUSCOs
- 3023 Total BUSCO groups searched

Detonate: Which assembly is better?

“RSEM-EVAL [sic] uses a novel probabilistic model-based method to compute the joint probability of both an assembly and the RNA-Seq data as an evaluation score.”

$$\text{score}_{\text{RSEM-EVAL}}(A) = \log P(A, D)$$

“the RSEM-EVAL score of an assembly is defined as the log joint probability of the assembly A and the reads D used to construct it”

$$\begin{aligned} \log P(A, D) &= \log \int_{\Lambda} P(D|A, \Lambda)P(A|\Lambda)P(\Lambda)d\Lambda \\ &\approx \underbrace{\log P(D|A, \Lambda_{\text{MLE}})}_{\text{likelihood}} + \underbrace{\log P(A|\Lambda_{\text{MLE}})}_{\text{assembly prior}} \\ &\quad - \underbrace{\frac{1}{2}(M+1)\log N}_{\text{BIC penalty}}, \end{aligned}$$

Detonate: Which assembly is better?

“RSEM-EVAL [sic] uses a novel probabilistic model-based method to compute the joint probability of both an assembly and the RNA-Seq data as an evaluation score.”

$$\text{score}_{\text{RSEM-EVAL}}(A) = \log P(A, D)$$

“the RSEM-EVAL score of an assembly is defined as the log joint probability of the assembly A and the reads D used to construct it”

$$\begin{aligned} \log P(A, D) &= \log \int_{\Lambda} P(D|A, \Lambda)P(A|\Lambda)P(\Lambda)d\Lambda \\ &\approx \underbrace{\log P(D|A, \Lambda_{\text{MLE}})} + \underbrace{\log P(A|\Lambda_{\text{MLE}})} \end{aligned}$$

Bigger Score = Better Assembly

$$- \underbrace{\frac{1}{2}(M+1)\log N}_{\text{BIC penalty}}$$

Detonate: Which assembly is better?

“RSEM-EVAL [sic] uses a novel probabilistic model-based method to compute the joint probability of both an assembly and the RNA-Seq data as an evaluation score.”

