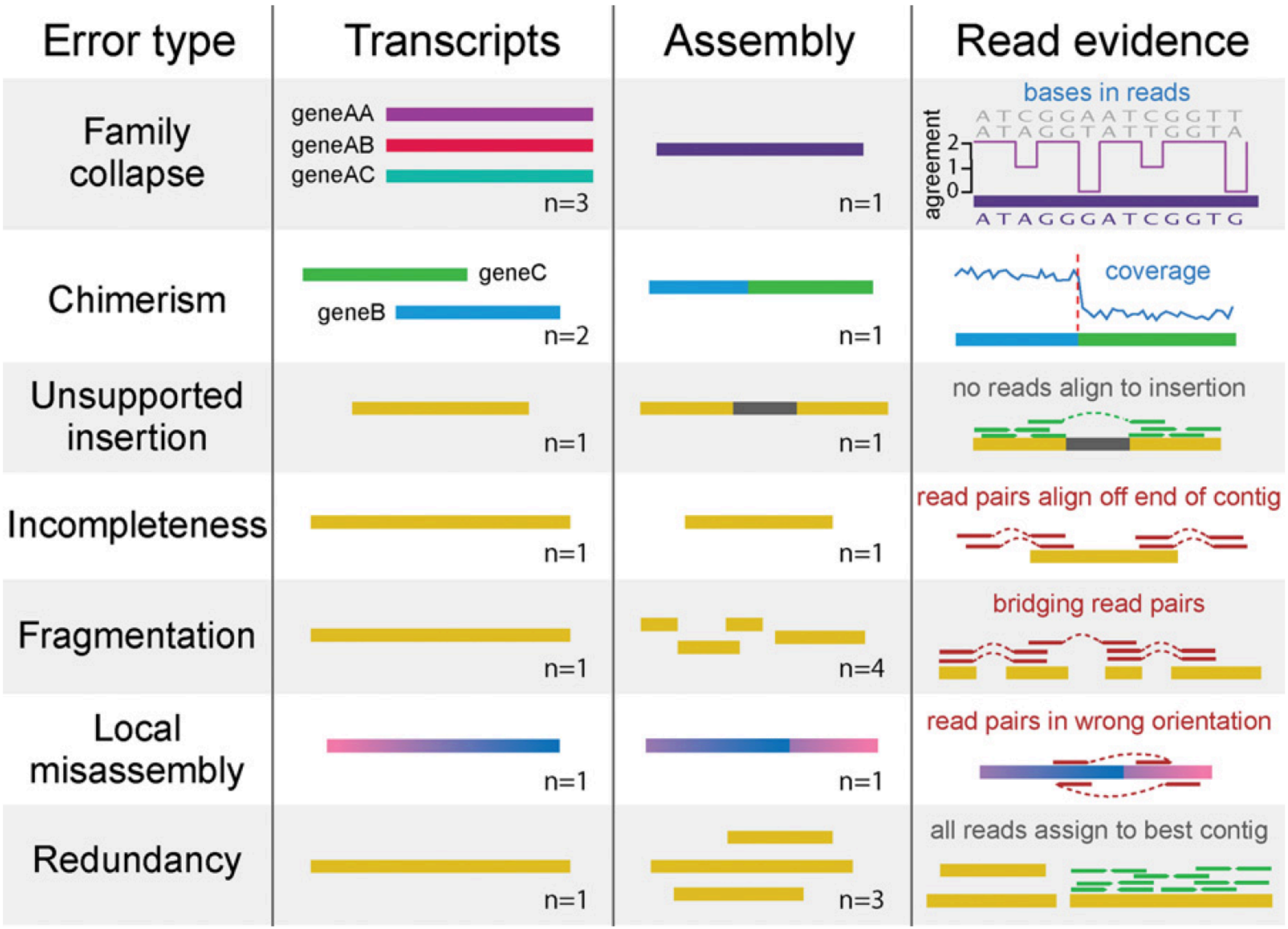
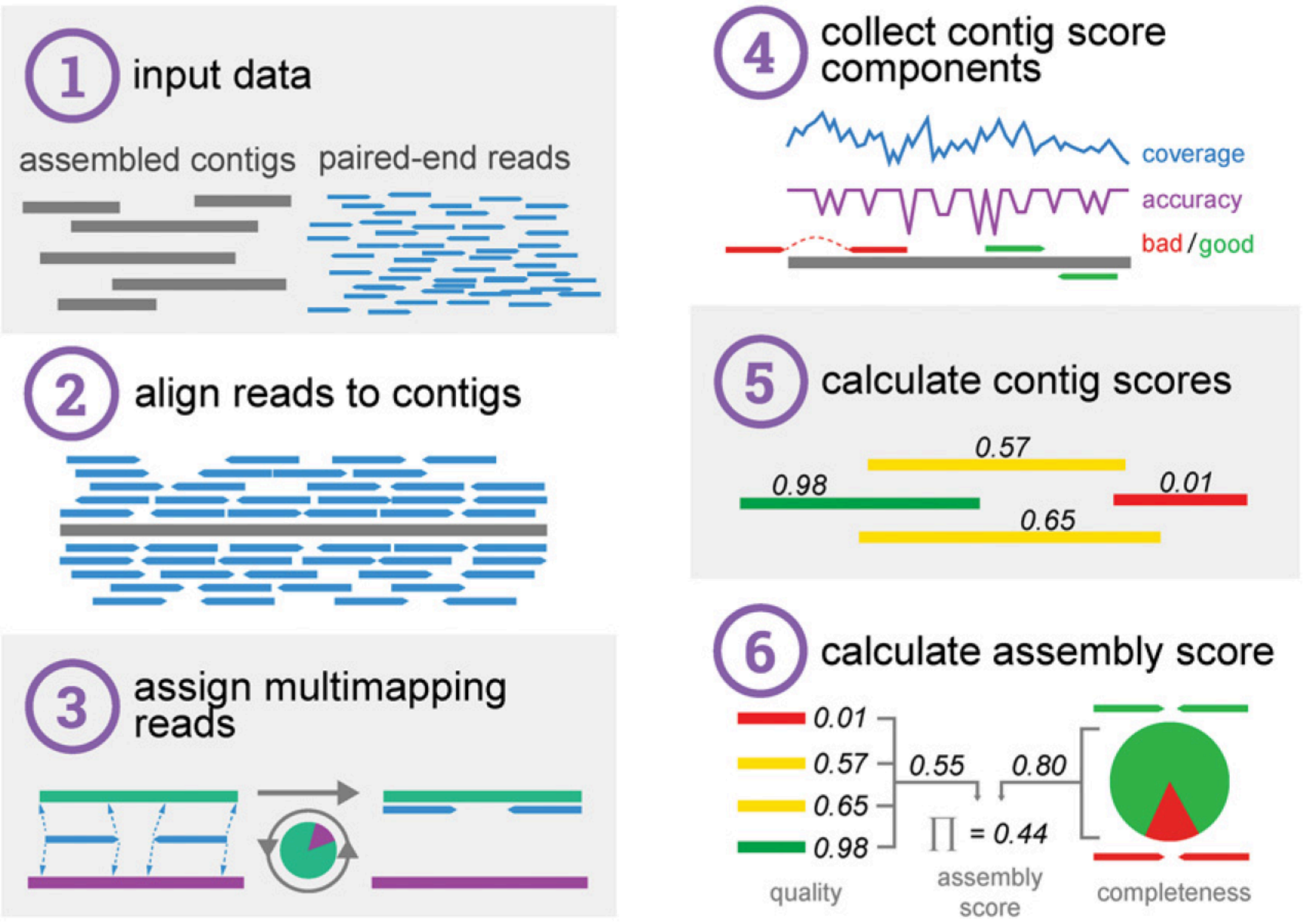
Evaluating the quality of your transcriptome assembly

De novo Transcriptome Assembly is Prone to Certain Types of Errors



Smith-Unna et al. Genome Research, 2016

TransRate



Smith-Unna et al. Genome Research, 2016

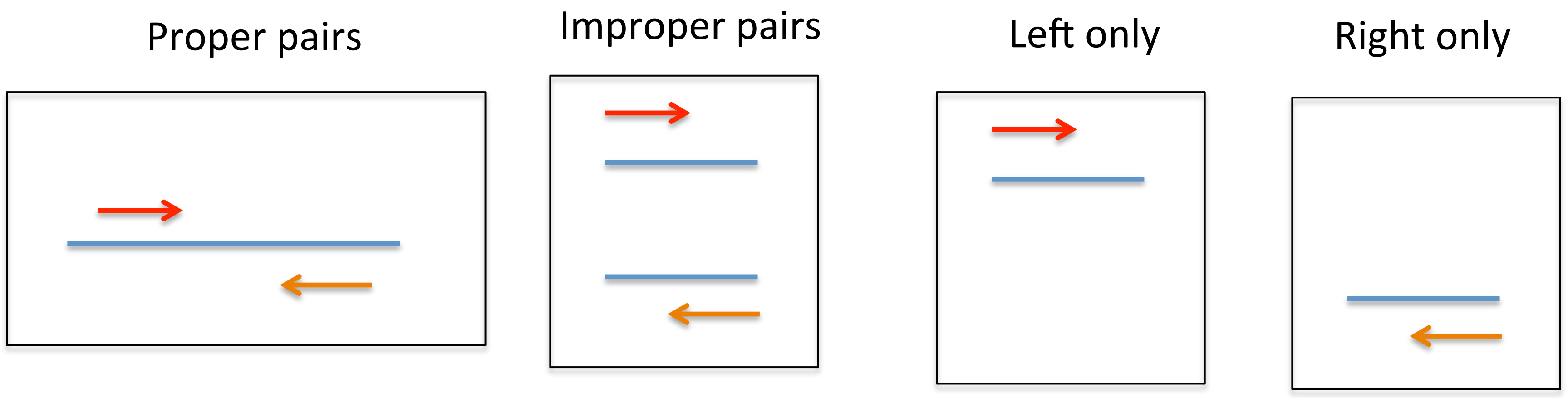
Simple Quantitative and Qualitative Assembly Metrics

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

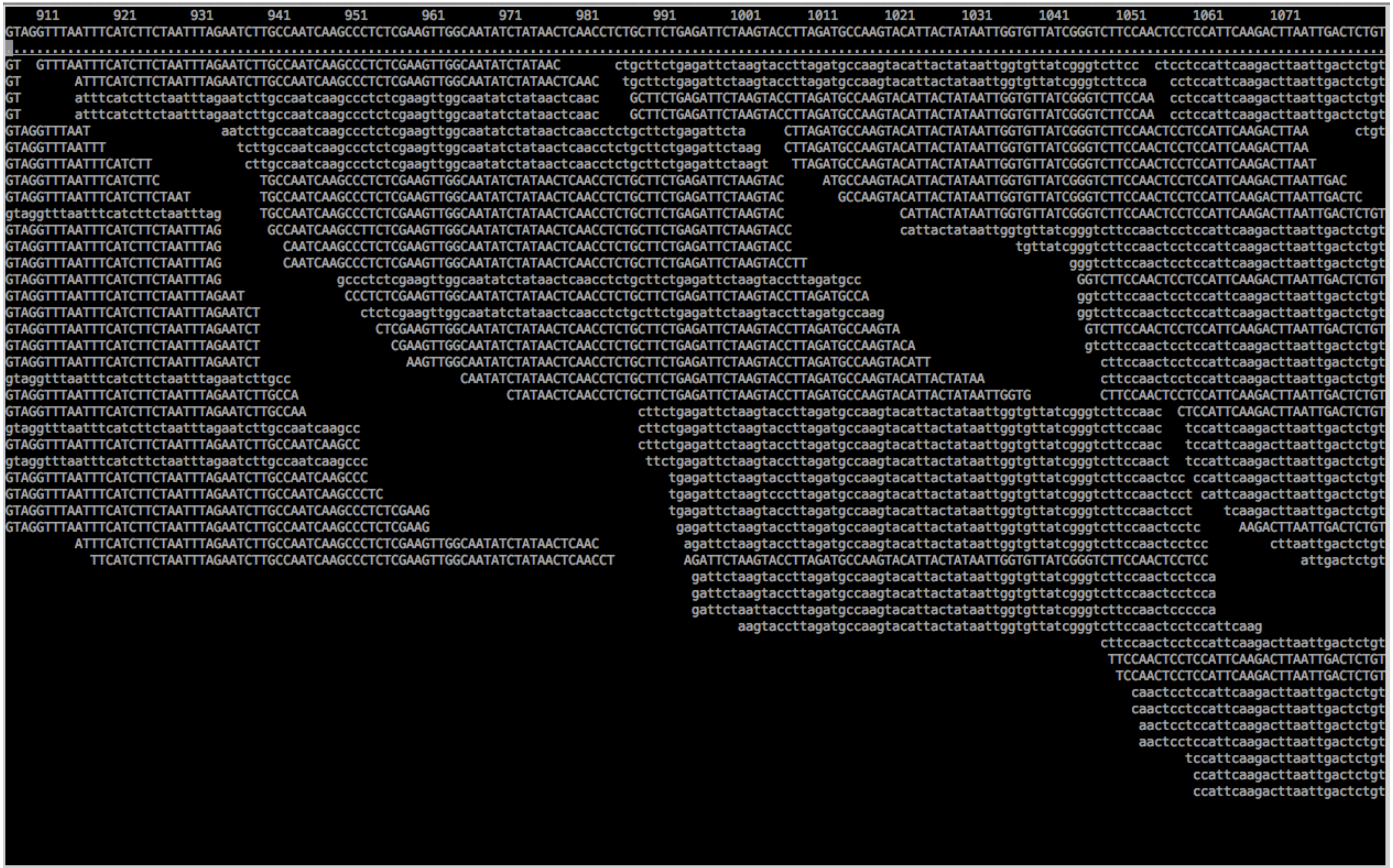
Improper pairs

Left only

Right only

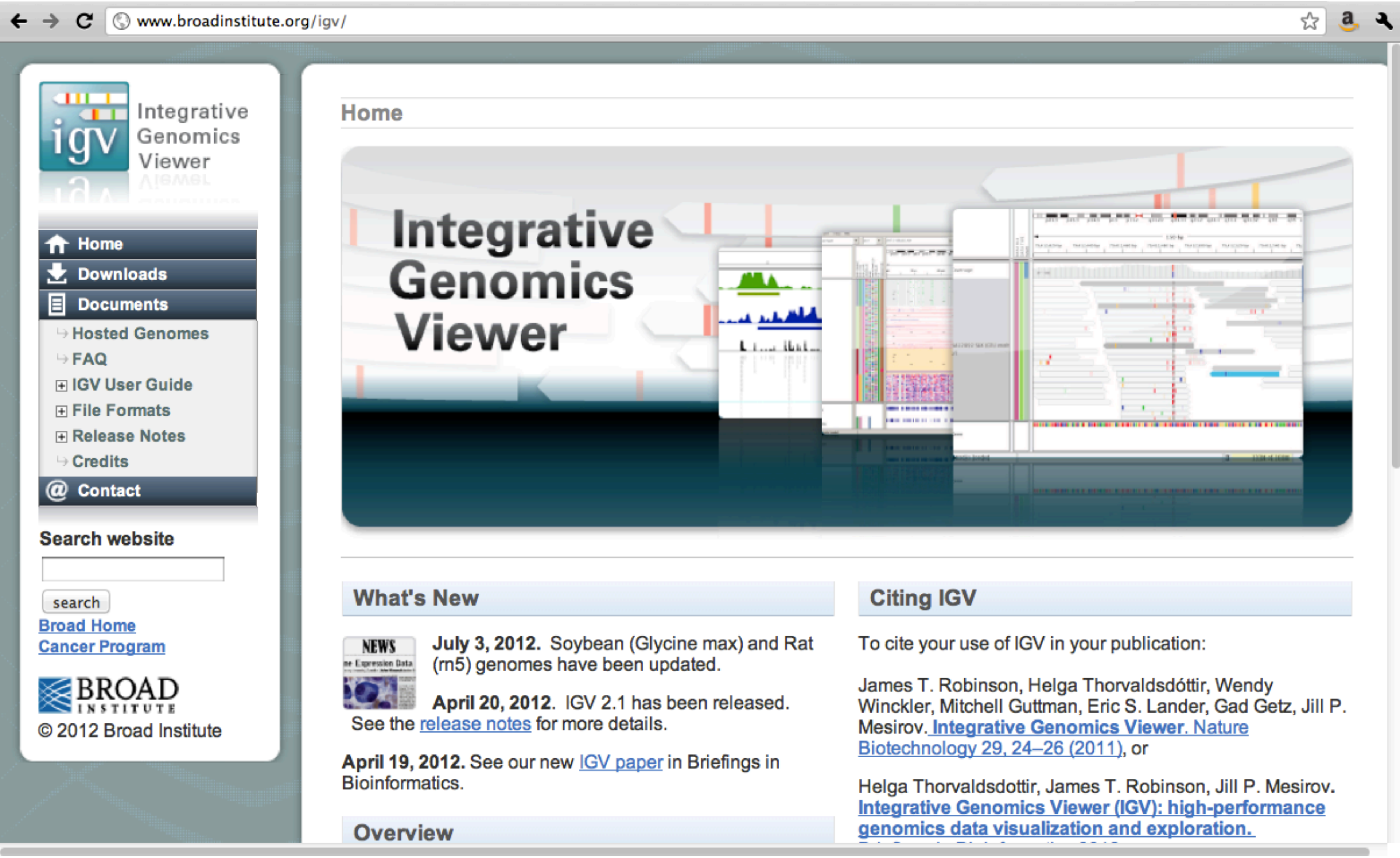
Assembled transcript contig is only as good as its read support.

% samtools tview alignments.bam target.fasta



```
    911       921       931       941       951       961       971       981       991      1001      1011      1021      1031      1041      1051      1061      1071
GTAGGTTTAATTTCATCTTCTAATTTAGAATCTTGCCAATCAAGCCCTCTCGAAGTTGGCAATATCTATAACTCAACCTCTGCTTCTGAGATTCTAAGTACCTTAGATGCCAAGTACATTACTATAATTGGTGTTATCGGGTCTTCCAACTCCTCCATTCAAGACTTAATTGACTCTGT
...................................................................................................................................................................................
GT  GTTTAATTTCATCTTCTAATTTAGAATCTTGCCAATCAAGCCCTCTCGAAGTTGGCAATATCTATAAC       ctgcttctgagattctaagtaccttagatgccaagtacattactataattggtgttatcgggtcttcc  ctcctccattcaagacttaattgactctgt
GT       ATTTCATCTTCTAATTTAGAATCTTGCCAATCAAGCCCTCTCGAAGTTGGCAATATCTATAACTCAAC   tgcttctgagattctaagtaccttagatgccaagtacattactataattggtgttatcgggtcttcca   cctccattcaagacttaattgactctgt
GT       atttcatcttctaatttagaatcttgccaatcaagccctctcgaagttggcaatatctataactcaac    GCTTCTGAGATTCTAAGTACCTTAGATGCCAAGTACATTACTATAATTGGTGTTATCGGGTCTTCCAA  cctccattcaagacttaattgactctgt
GT       atttcatcttctaatttagaatcttgccaatcaagccctctcgaagttggcaatatctataactcaac    GCTTCTGAGATTCTAAGTACCTTAGATGCCAAGTACATTACTATAATTGGTGTTATCGGGTCTTCCAA  cctccattcaagacttaattgactctgt
GTAGGTTTAAT                 aatcttgccaatcaagccctctcgaagttggcaatatctataactcaacctctgcttctgagattcta     CTTAGATGCCAAGTACATTACTATAATTGGTGTTATCGGGTCTTCCAACTCCTCCATTCAAGACTTAA      ctgt
GTAGGTTTAATTT                 tcttgccaatcaagccctctcgaagttggcaatatctataactcaacctctgcttctgagattctaag   CTTAGATGCCAAGTACATTACTATAATTGGTGTTATCGGGTCTTCCAACTCCTCCATTCAAGACTTAA
GTAGGTTTAATTTCATCTT            cttgccaatcaagccctctcgaagttggcaatatctataactcaacctctgcttctgagattctaagt   TTAGATGCCAAGTACATTACTATAATTGGTGTTATCGGGTCTTCCAACTCCTCCATTCAAGACTTAAT
GTAGGTTTAATTTCATCTTC             TGCCAATCAAGCCCTCTCGAAGTTGGCAATATCTATAACTCAACCTCTGCTTCTGAGATTCTAAGTAC     ATGCCAAGTACATTACTATAATTGGTGTTATCGGGTCTTCCAACTCCTCCATTCAAGACTTAATTGAC
GTAGGTTTAATTTCATCTTCTAAT         TGCCAATCAAGCCCTCTCGAAGTTGGCAATATCTATAACTCAACCTCTGCTTCTGAGATTCTAAGTAC       GCCAAGTACATTACTATAATTGGTGTTATCGGGTCTTCCAACTCCTCCATTCAAGACTTAATTGACTC
gtaggtttaatttcatcttctaatttag     TGCCAATCAAGCCCTCTCGAAGTTGGCAATATCTATAACTCAACCTCTGCTTCTGAGATTCTAAGTAC               CATTACTATAATTGGTGTTATCGGGTCTTCCAACTCCTCCATTCAAGACTTAATTGACTCTGT
GTAGGTTTAATTTCATCTTCTAATTTAG      GCCAATCAAGCCTTCTCGAAGTTGGCAATATCTATAACTCAACCTCTGCTTCTGAGATTCTAAGTACC              cattactataattggtgttatcgggtcttccaactcctccattcaagacttaattgactctgt
GTAGGTTTAATTTCATCTTCTAATTTAG        CAATCAAGCCCTCTCGAAGTTGGCAATATCTATAACTCAACCTCTGCTTCTGAGATTCTAAGTACC                             tgttatcgggtcttccaactcctccattcaagacttaattgactctgt
GTAGGTTTAATTTCATCTTCTAATTTAG        CAATCAAGCCCTCTCGAAGTTGGCAATATCTATAACTCAACCTCTGCTTCTGAGATTCTAAGTACCTT                                  gggtcttccaactcctccattcaagacttaattgactctgt
GTAGGTTTAATTTCATCTTCTAATTTAG               gccctctcgaagttggcaatatctataactcaacctctgcttctgagattctaagtaccttagatgcc                            GGTCTTCCAACTCCTCCATTCAAGACTTAATTGACTCTGT
GTAGGTTTAATTTCATCTTCTAATTTAGAAT             CCCTCTCGAAGTTGGCAATATCTATAACTCAACCTCTGCTTCTGAGATTCTAAGTACCTTAGATGCCA                           ggtcttccaactcctccattcaagacttaattgactctgt
GTAGGTTTAATTTCATCTTCTAATTTAGAATCT             ctctcgaagttggcaatatctataactcaacctctgcttctgagattctaagtaccttagatgccaag                         ggtcttccaactcctccattcaagacttaattgactctgt
GTAGGTTTAATTTCATCTTCTAATTTAGAATCT               CTCGAAGTTGGCAATATCTATAACTCAACCTCTGCTTCTGAGATTCTAAGTACCTTAGATGCCAAGTA                        GTCTTCCAACTCCTCCATTCAAGACTTAATTGACTCTGT
GTAGGTTTAATTTCATCTTCTAATTTAGAATCT                 CGAAGTTGGCAATATCTATAACTCAACCTCTGCTTCTGAGATTCTAAGTACCTTAGATGCCAAGTACA                      gtcttccaactcctccattcaagacttaattgactctgt
GTAGGTTTAATTTCATCTTCTAATTTAGAATCT                   AAGTTGGCAATATCTATAACTCAACCTCTGCTTCTGAGATTCTAAGTACCTTAGATGCCAAGTACATT                      cttccaactcctccattcaagacttaattgactctgt
gtaggtttaatttcatcttctaatttagaatcttgcc                      CAATATCTATAACTCAACCTCTGCTTCTGAGATTCTAAGTACCTTAGATGCCAAGTACATTACTATAA               cttccaactcctccattcaagacttaattgactctgt
GTAGGTTTAATTTCATCTTCTAATTTAGAATCTTGCCA                           CTATAACTCAACCTCTGCTTCTGAGATTCTAAGTACCTTAGATGCCAAGTACATTACTATAATTGGTG         CTTCCAACTCCTCCATTCAAGACTTAATTGACTCTGT
GTAGGTTTAATTTCATCTTCTAATTTAGAATCTTGCCAA                                           cttctgagattctaagtaccttagatgccaagtacattactataattggtgttatcgggtcttccaac  CTCCATTCAAGACTTAATTGACTCTGT
gtaggtttaatttcatcttctaatttagaatcttgccaatcaagcc                                    cttctgagattctaagtaccttagatgccaagtacattactataattggtgttatcgggtcttccaac   tccattcaagacttaattgactctgt
GTAGGTTTAATTTCATCTTCTAATTTAGAATCTTGCCAATCAAGCC                                    cttctgagattctaagtaccttagatgccaagtacattactataattggtgttatcgggtcttccaac   tccattcaagacttaattgactctgt
gtaggtttaatttcatcttctaatttagaatcttgccaatcaagccc                                    ttctgagattctaagtaccttagatgccaagtacattactataattggtgttatcgggtcttccaact  tccattcaagacttaattgactctgt
GTAGGTTTAATTTCATCTTCTAATTTAGAATCTTGCCAATCAAGCCC                                       tgagattctaagtaccttagatgccaagtacattactataattggtgttatcgggtcttccaactcc ccattcaagacttaattgactctgt
GTAGGTTTAATTTCATCTTCTAATTTAGAATCTTGCCAATCAAGCCCTC                                     tgagattctaagtcccttagatgccaagtacattactataattggtgttatcgggtcttccaactcct cattcaagacttaattgactctgt
GTAGGTTTAATTTCATCTTCTAATTTAGAATCTTGCCAATCAAGCCCTCTCGAAG                               tgagattctaagtaccttagatgccaagtacattactataattggtgttatcgggtcttccaactcct    tcaagacttaattgactctgt
GTAGGTTTAATTTCATCTTCTAATTTAGAATCTTGCCAATCAAGCCCTCTCGAAG                                gagattctaagtaccttagatgccaagtacattactataattggtgttatcgggtcttccaactcctc     AAGACTTAATTGACTCTGT
         ATTTCATCTTCTAATTTAGAATCTTGCCAATCAAGCCCTCTCGAAGTTGGCAATATCTATAACTCAAC           agattctaagtaccttagatgccaagtacattactataattggtgttatcgggtcttccaactcctcc        cttaattgactctgt
           TTCATCTTCTAATTTAGAATCTTGCCAATCAAGCCCTCTCGAAGTTGGCAATATCTATAACTCAACCT         AGATTCTAAGTACCTTAGATGCCAAGTACATTACTATAATTGGTGTTATCGGGTCTTCCAACTCCTCC            attgactctgt
                                                                                         gattctaagtaccttagatgccaagtacattactataattggtgttatcgggtcttccaactcctcca
                                                                                         gattctaagtaccttagatgccaagtacattactataattggtgttatcgggtcttccaactcctcca
                                                                                         gattctaattaccttagatgccaagtacattactataattggtgttatcgggtcttccaactccccca
                                                                                               aagtaccttagatgccaagtacattactataattggtgttatcgggtcttccaactcctccattcaag
                                                                                                                                              cttccaactcctccattcaagacttaattgactctgt
                                                                                                                                               TTCCAACTCCTCCATTCAAGACTTAATTGACTCTGT
                                                                                                                                                TCCAACTCCTCCATTCAAGACTTAATTGACTCTGT
                                                                                                                                                  caactcctccattcaagacttaattgactctgt
                                                                                                                                                  caactcctccattcaagacttaattgactctgt
                                                                                                                                                   aactcctccattcaagacttaattgactctgt
                                                                                                                                                   aactcctccattcaagacttaattgactctgt
                                                                                                                                                         tccattcaagacttaattgactctgt
                                                                                                                                                          ccattcaagacttaattgactctgt
                                                                                                                                                          ccattcaagacttaattgactctgt
```

IGV



© 2012 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

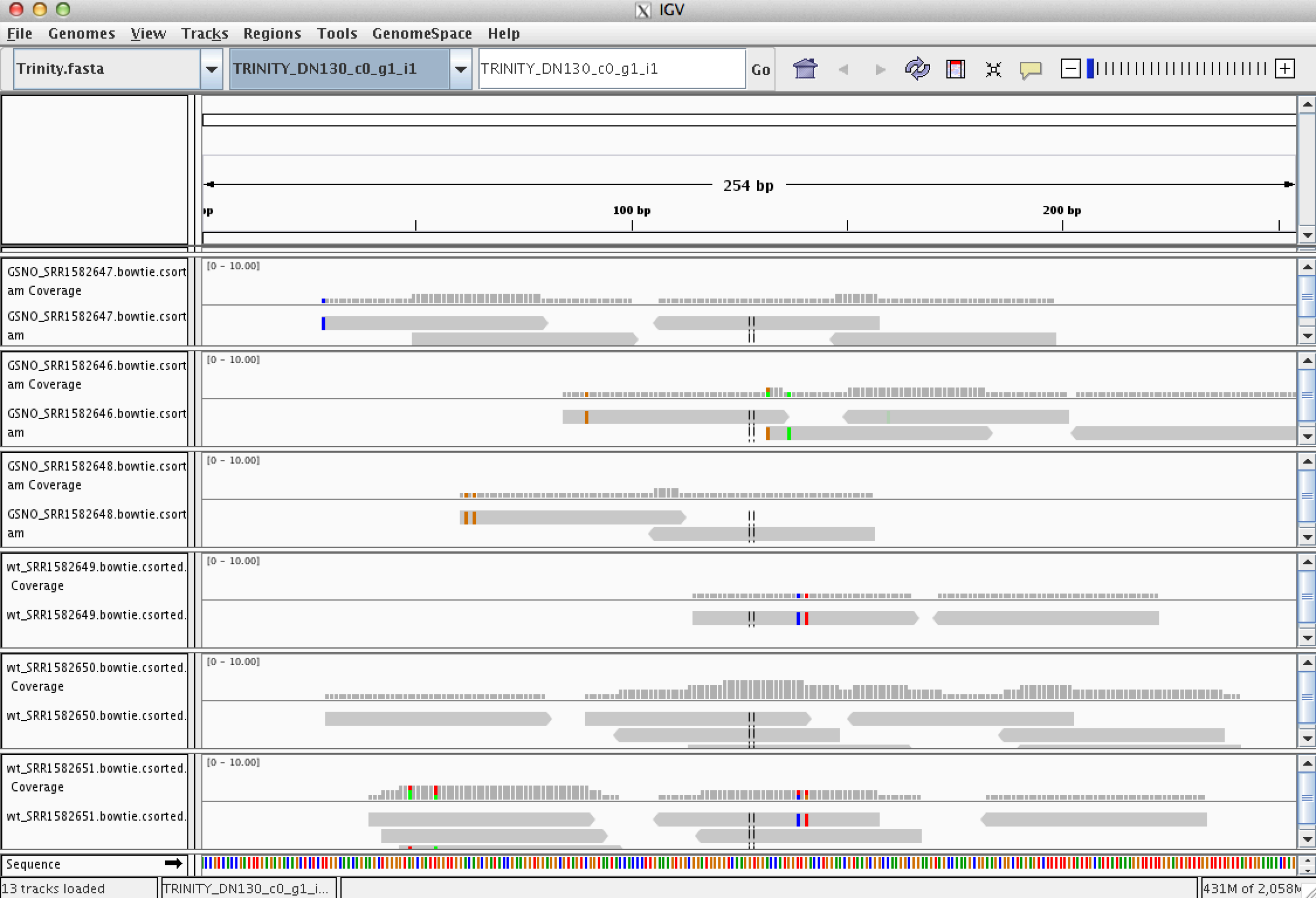
Citing IGV

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

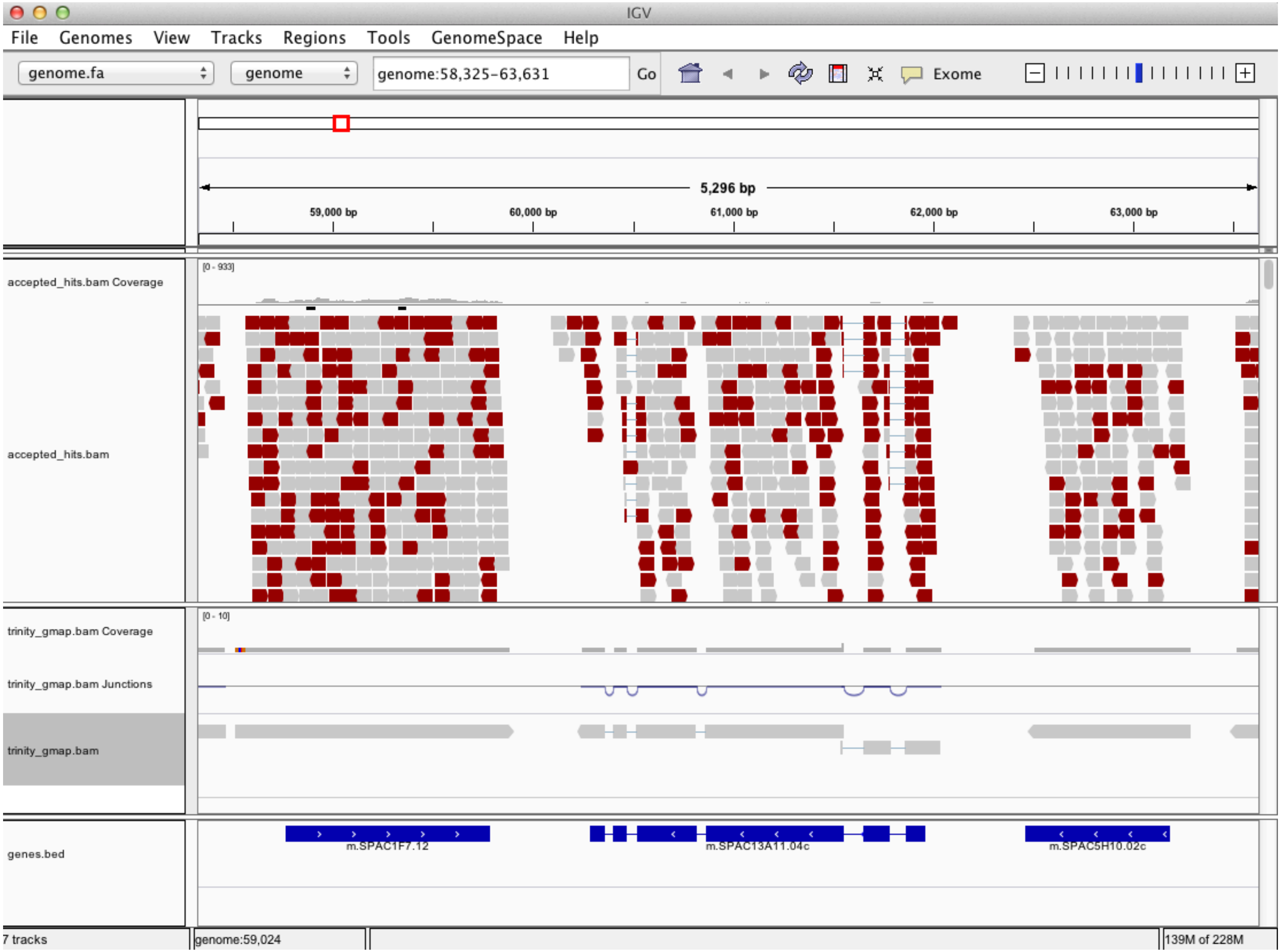
Helga Thorvaldsdottir, James T. Robinson, Jill P. Mesirov. **Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration.**

Can Examine Transcript Read Support Using IGV


IGV
File Genomes View Tracks Regions Tools GenomeSpace Help
Trinity.fasta
TRINITY_DN130_c0_g1_i1
TRINITY_DN130_c0_g1_i1
Go
254 bp
100 bp
200 bp
GSNO_SRR1582647.bowtie.csortam Coverage
GSNO_SRR1582647.bowtie.csortam
[0 - 10.00]
GSNO_SRR1582646.bowtie.csortam Coverage
GSNO_SRR1582646.bowtie.csortam
[0 - 10.00]
GSNO_SRR1582648.bowtie.csortam Coverage
GSNO_SRR1582648.bowtie.csortam
[0 - 10.00]
wt_SRR1582649.bowtie.csorted. Coverage
wt_SRR1582649.bowtie.csorted.
[0 - 10.00]
wt_SRR1582650.bowtie.csorted. Coverage
wt_SRR1582650.bowtie.csorted.
[0 - 10.00]
wt_SRR1582651.bowtie.csorted. Coverage
wt_SRR1582651.bowtie.csorted.
[0 - 10.00]
Sequence
13 tracks loaded
TRINITY_DN130_c0_g1_i...
431M of 2,058M

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

(Trinity transcripts aligned to the genome using GMAP)

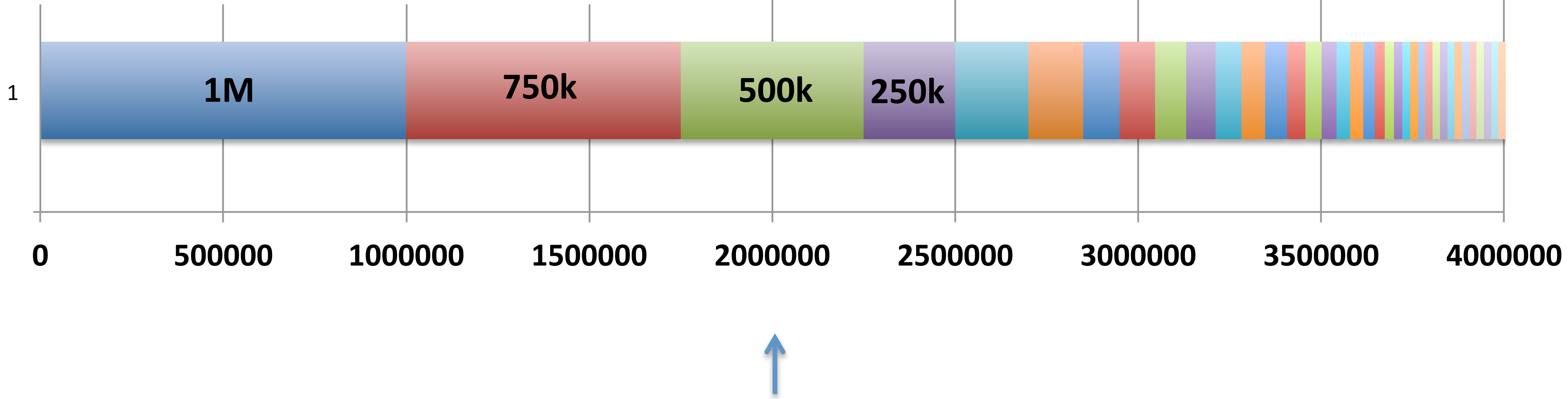


The Contig N50 statistic

“At least half of assembled bases are in contigs that are at least ***N50*** bases in length”

In genome assemblies – used often to judge ‘which assembly is better’

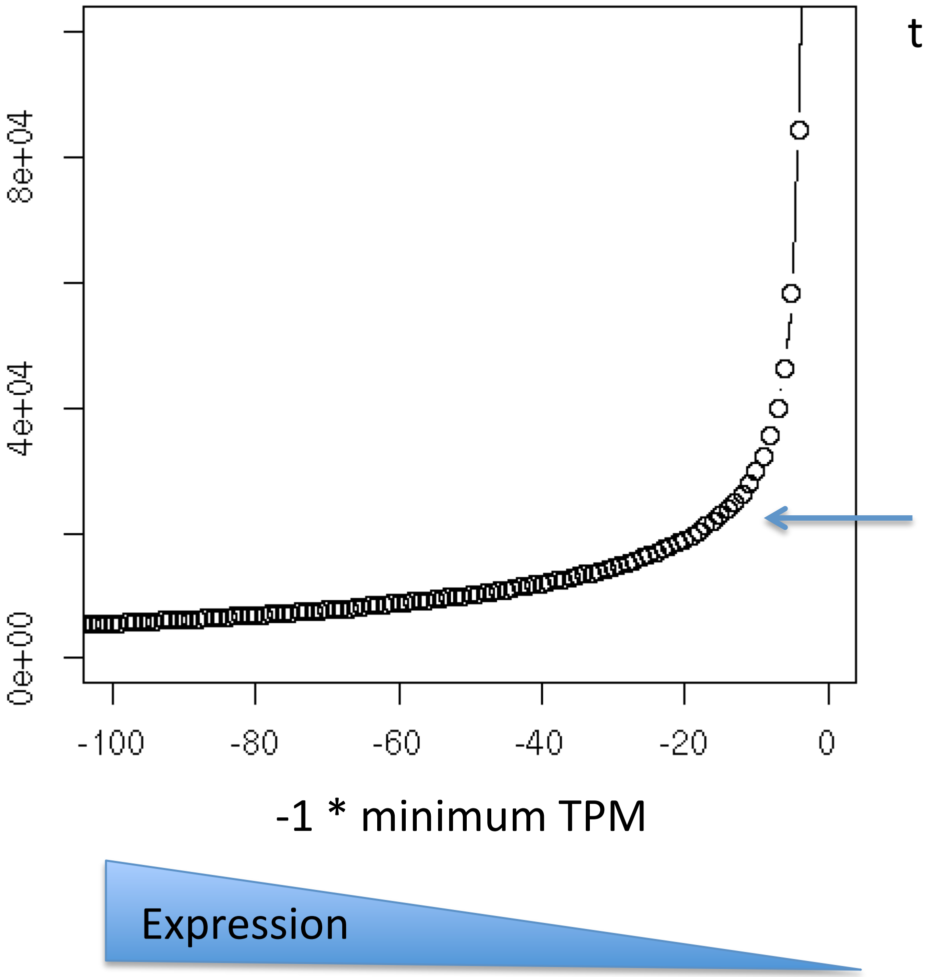
Assemblies ordered by length:



N50 contig length = 500k

Often, most assembled transcripts are *very* lowly expressed

(How many ‘transcripts & genes’ are there really?)



Cumulative # of Transcripts

-1 * minimum TPM

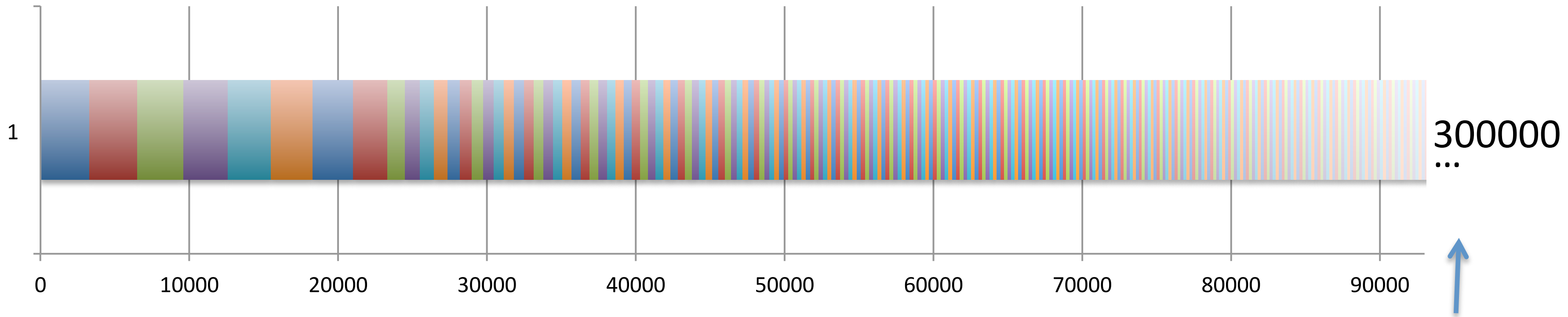
1.4 million Trinity transcript contigs

N50 ~ 500 bases

20k transcripts

* Salamander transcriptome

N50 Calculation for *Transcriptome* Assemblies??



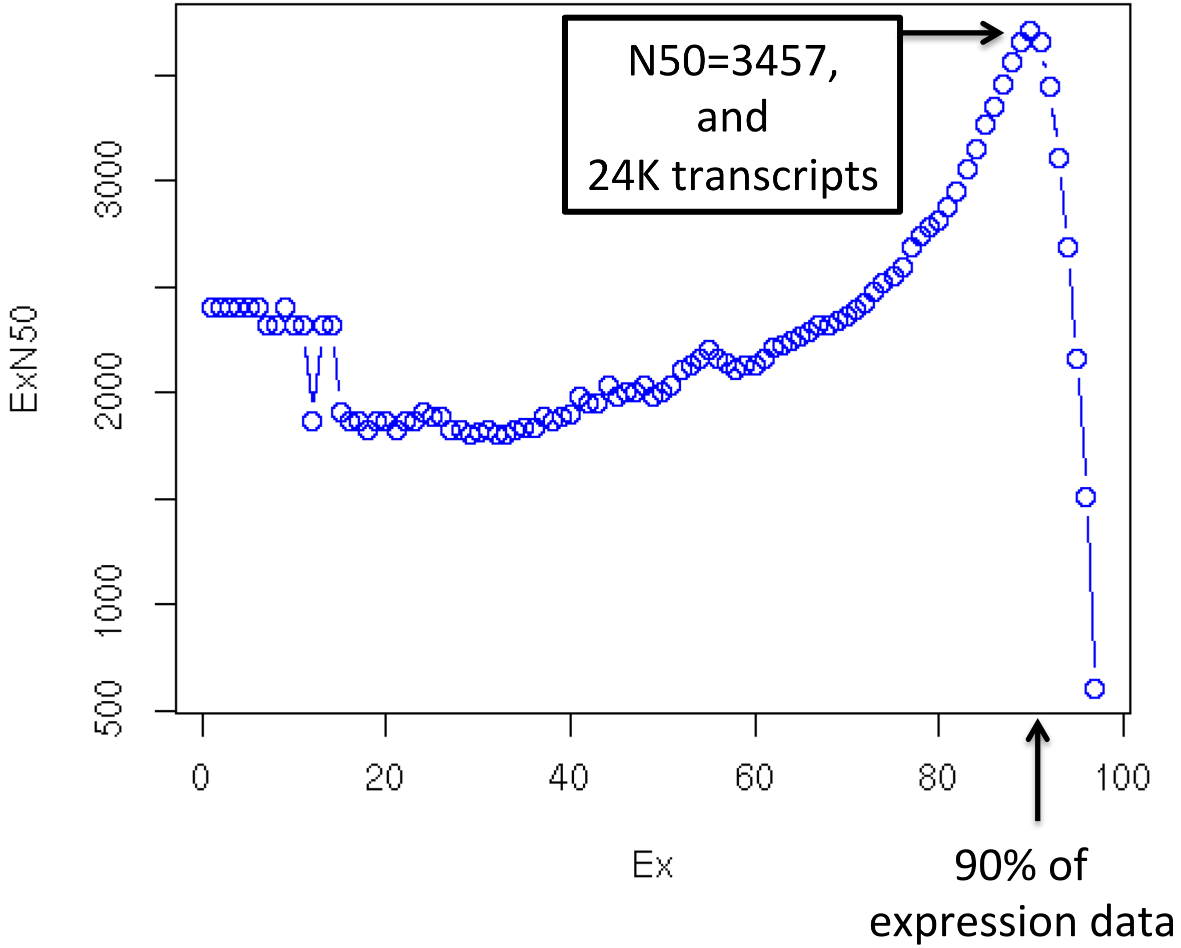
N50 length?
(small)

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

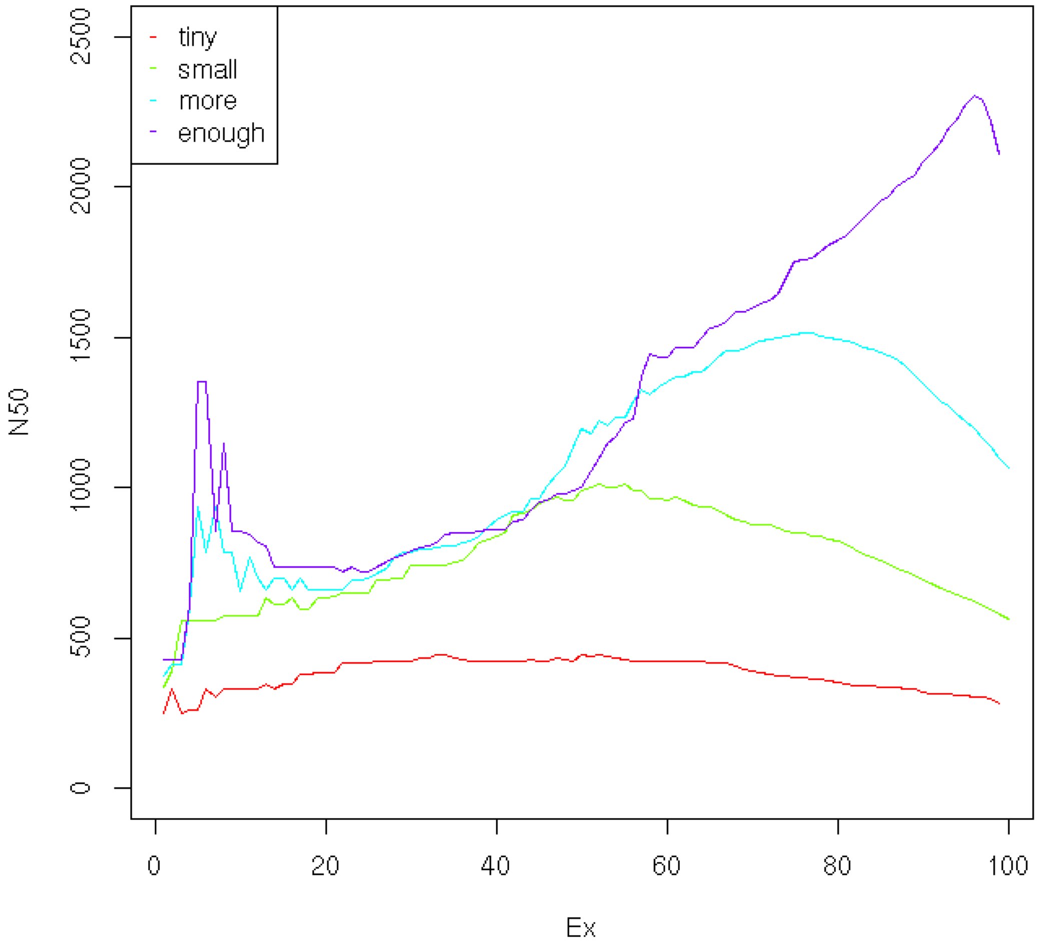
- Overzealous isoform annotation for long transcripts drives higher N50
- Very sensitive reconstruction for short lowly expressed transcripts drives lower N50

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



Millions of Reads

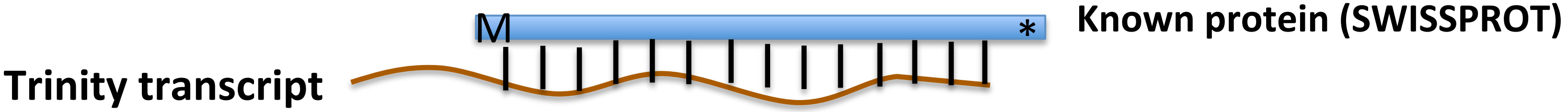
Thousands of Reads

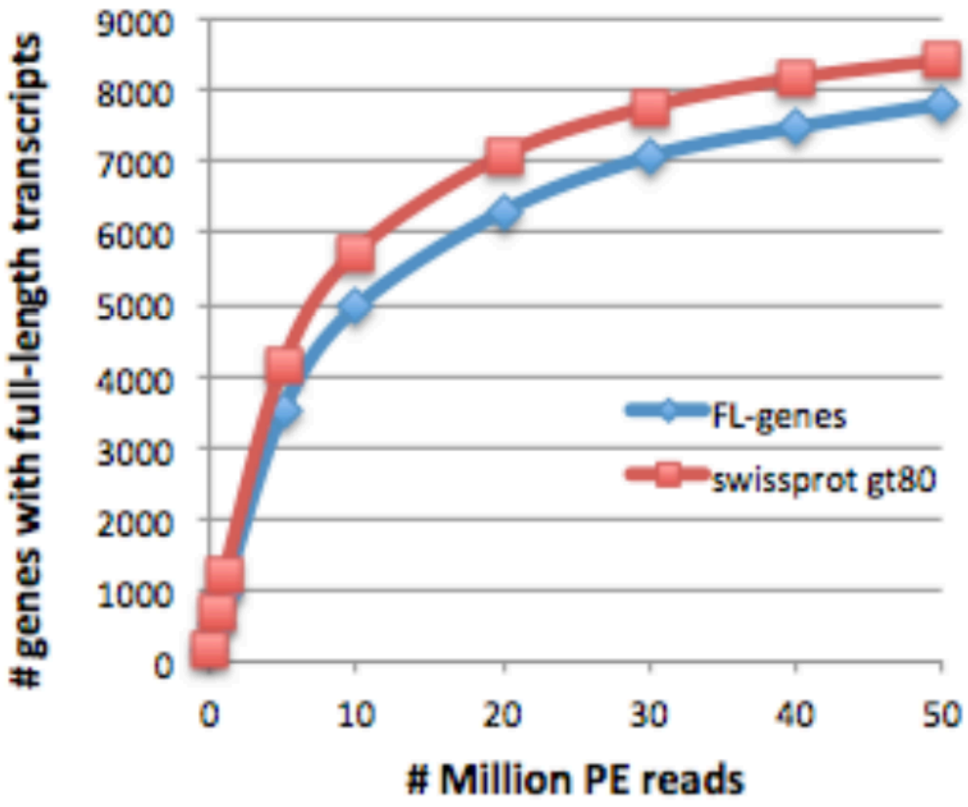
Note shift in ExN50 profiles as you assemble more and more reads.

* Candida transcriptome

Evaluating the quality of your transcriptome assembly

Full-length Transcript Detection via BLASTX

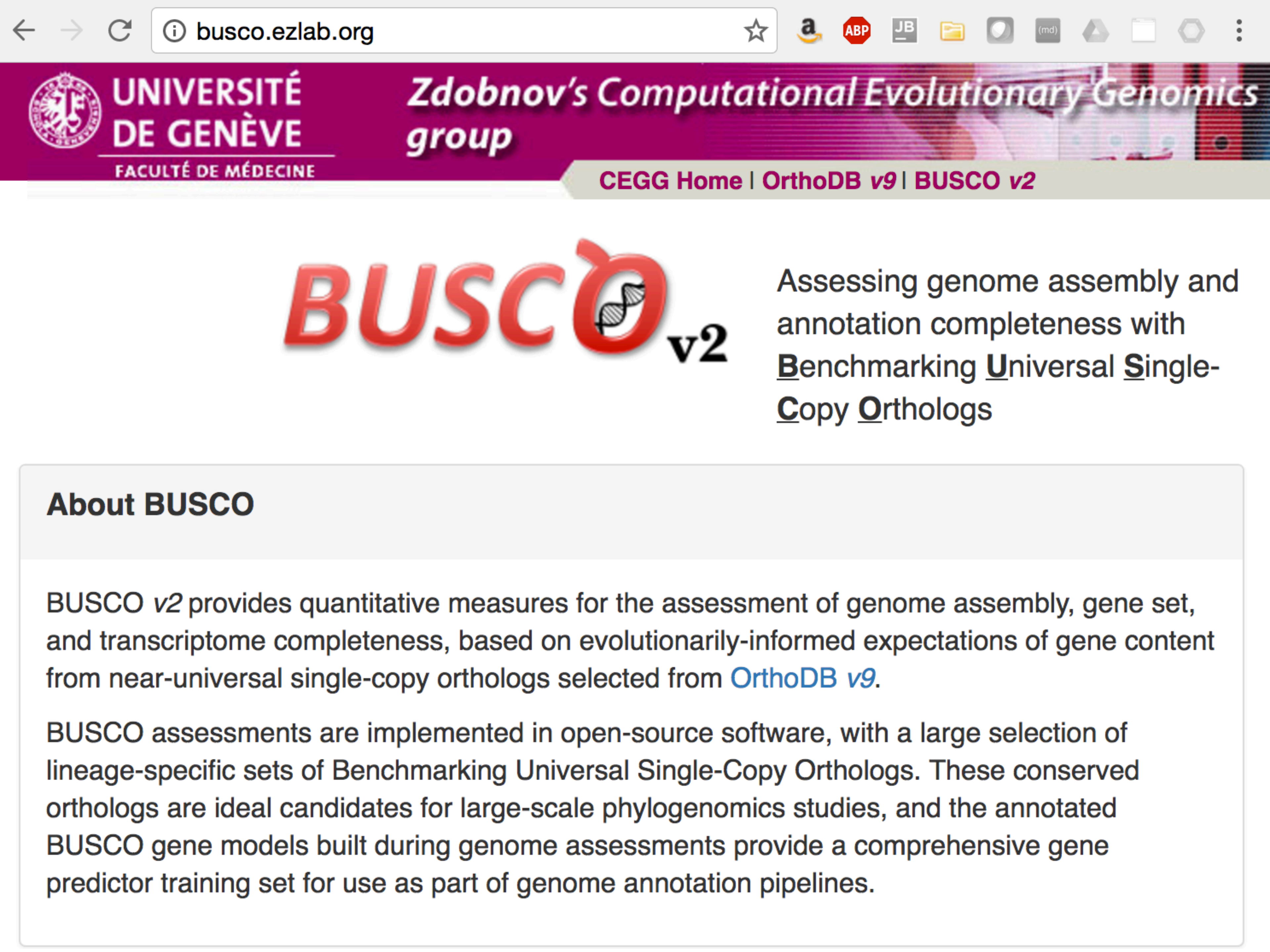




Have you sequenced deeply enough?

* Mouse transcriptome

Haas et al. Nat. Protoc. 2013

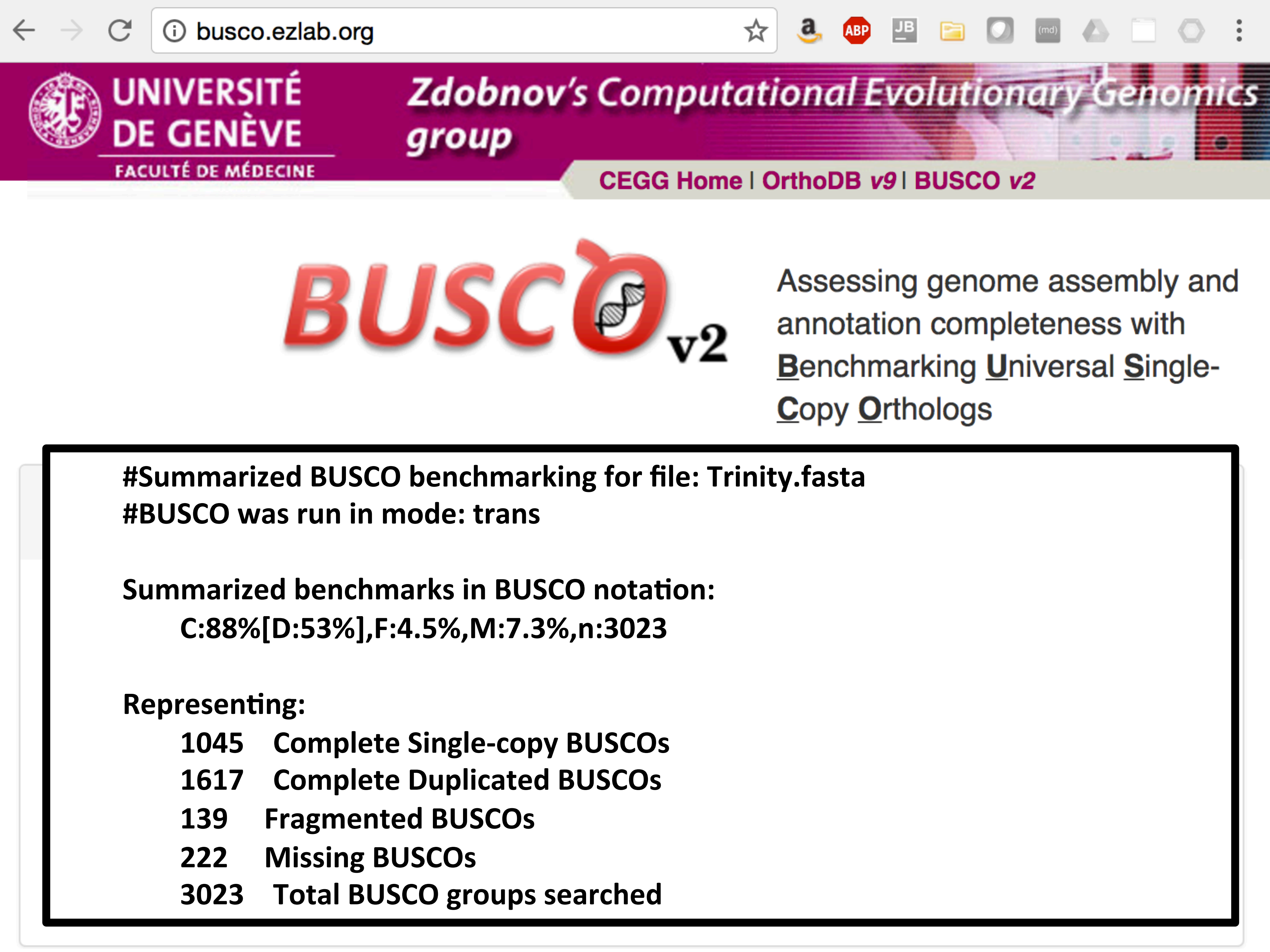

busco.ezlab.org
UNIVERSITÉ DE GENÈVE
FACULTÉ DE MÉDECINE
Zdobnov's Computational Evolutionary Genomics group
CEGG Home | OrthoDB v9 | BUSCO v2
BUSCO v2

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.


busco.ezlab.org
UNIVERSITÉ DE GENÈVE
FACULTÉ DE MÉDECINE
Zdobnov's Computational Evolutionary Genomics group
CEGG Home | OrthoDB v9 | BUSCO v2
BUSCO v2

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”

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

Li et al. **Evaluation of de novo transcriptome assemblies from RNA-Seq data**, Genome Biology 2014

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.”

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

“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”

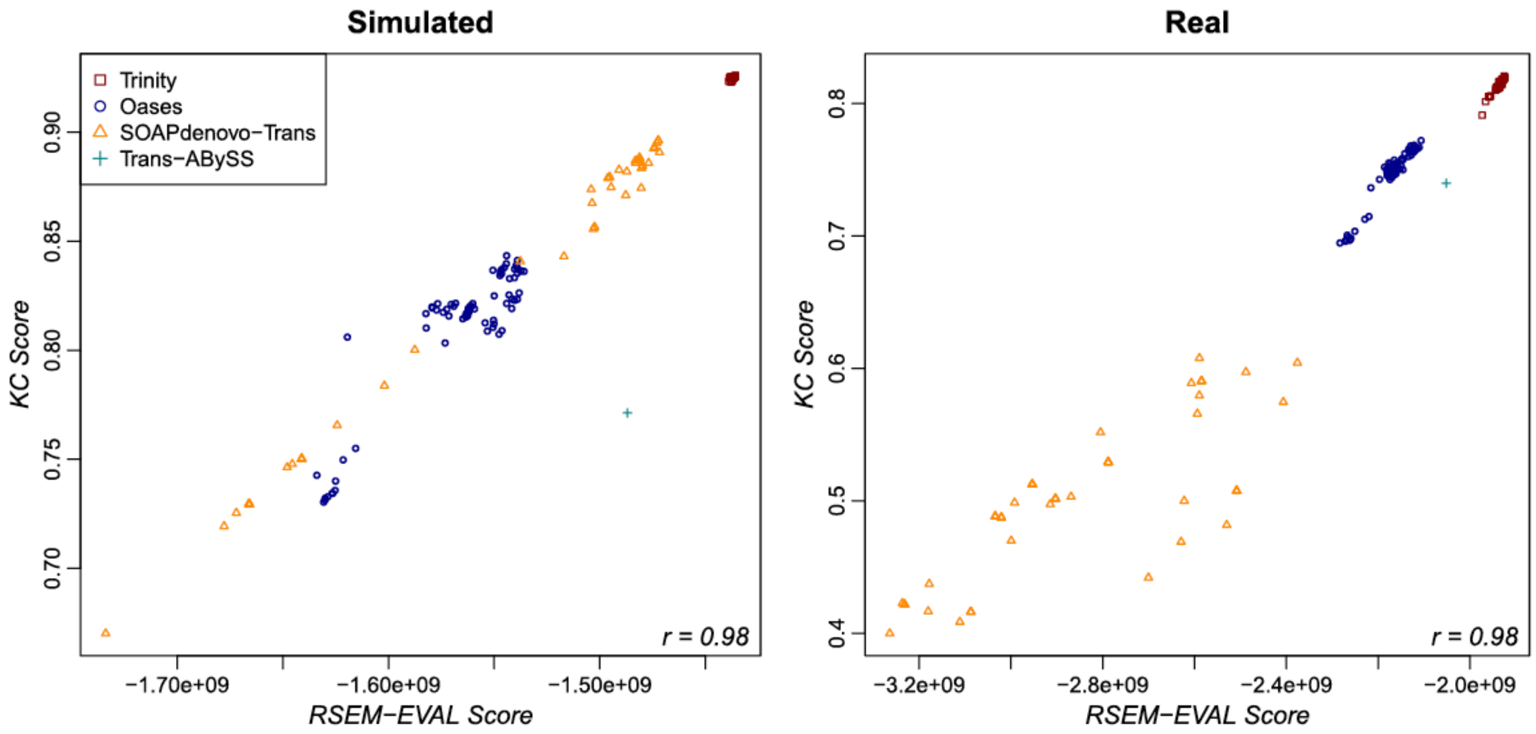
$$\begin{aligned}\log P(A,D) &= \log \int_{\Lambda} P(D|A,\Lambda)P(A|\Lambda)P(\Lambda)d\Lambda \\ &\approx \underbrace{\log P(D|A,\Lambda_{\mathrm{MLE}})}_{\text{likelihood}} + \underbrace{\log P(A|\Lambda_{\mathrm{MLE}})}_{\text{assembly prior}} \\ &\quad \underbrace{-\frac{1}{2}(M+1)\log N,}_{\text{BIC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

Bigger Score = Better Assembly

Li et al. **Evaluation of de novo transcriptome assemblies from RNA-Seq data**, Genome Biology 2014

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.”



Yay Trinity!!

Ref Genome –based metric

RSEM-EVAL Genome-free metric

Li et al. **Evaluation of de novo transcriptome assemblies from RNA-Seq data**, Genome Biology 2014