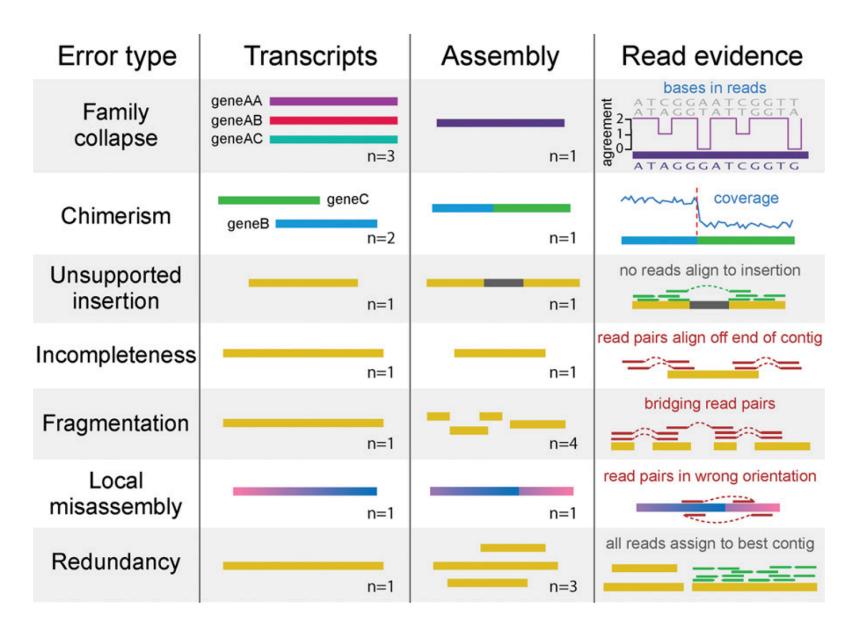
Evaluating the quality of your <u>transcriptome</u> assembly

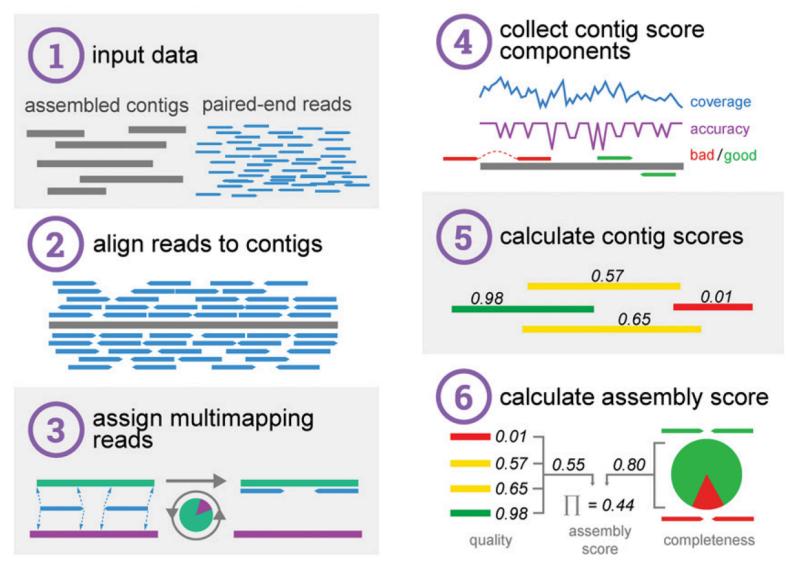


De novo Transcriptome Assembly is Prone to Certain Types of Errors



Smith-Unna et al. Genome Research, 2016





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

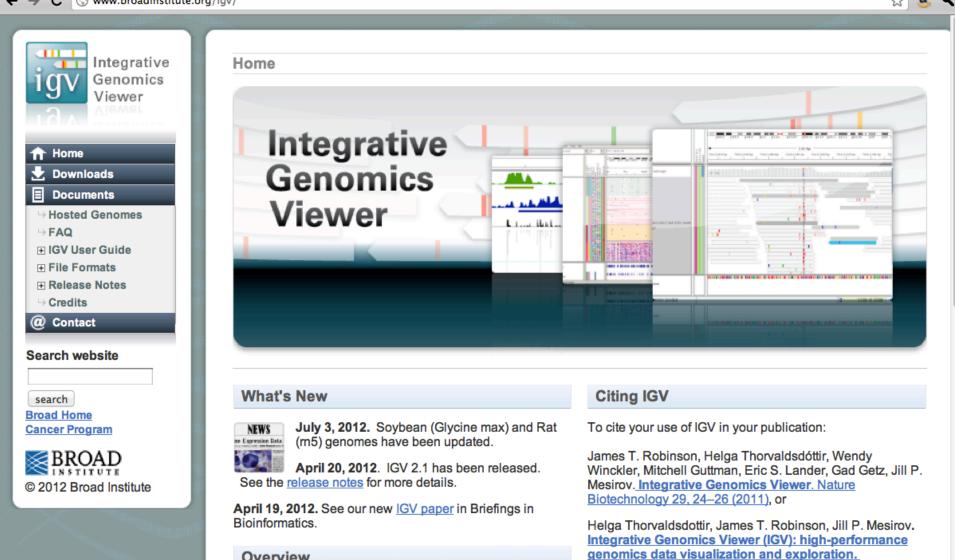
TAGGTTAATTCATCTTCTAATTTAGATCTTGCCAATCAAGCCCTCTCGAAGTTGGCAATATCTATAACTCAACCTCTGCTTCTGAGATTCTAAGTACCTTAGGTGCAATTACTATATTGGGTTATCGGGTCTTCCAACCCCCCCATTCAAGACTTAATTGACTCTG T GTTTAATTTCATCTTCTAATTTAGAATCTTGCCAATCAAGCCCTCTCGAAGTTGGCAATATCTATAACTCAAC T ATTCCATCTTCTAATTTAGAATCTGGCCAATCAAGCCCTCTCGAAGTTGGCAATATCTATAACTCAAC T atttaattcttaatttagaatttgccaattagatctggcaattaataattagatctagc GCTTCGGGATTCTAAGTCCCATTCAAGCCCTCTCGAAGTTGGCAATATCTATAACTCAAC T atttaattttagaatttgccaattagaattgccaaggcaattaataattagatcaac GCTTCGGGATTCTAAGTACCTTAGCATACCTTAACTGAAC T atttaattttagaatttgccaattagagccttctgaaggtggcaattactaataatcaac GCTTCGGGATTCTAAGTACCTTAGATGCCCAGTGCAGTG
GTTAATTCATCTTCTAATTCAAATCTGCCAATCAGeCCCTCTCGCAAGTTGCCAATATCTATAAC cdgettetgagatetetagatgecaagtagatettagatetggggtetetegggtettecgggtettetagggtettetggggtggtggggaastetagggatggetgggatggetgggatggetgggatggetgge
T ATTCATCTTCTATTTAGAATCTTGCCAATCAAGCCCTCTGGAATATCTATAACTCAAC tgtttgggattctaagtgccaagtagcattaattggtcttcaagggtttgcaatatctaagatctaagtggattgagaatatctaagtggattgagaatatctaagtggattggaatatctaagtggattggaatatctaagtggattggaatatctaagtggattggaatatctaagtggattggaatatctaagtggattggaatatctaagtggattggaatatctaagtggattggagatatctaagtggagattcaagtggagattcaagtggagattggagaatatctaagtgattggagatatctaagtggagattcaagtggagattcaagtggagattggagaatatctaagtggagattggagatatctaagtcggagttggagatatctaagtcggagttggagatatctaagtcggagttggagatatctaagtcggagttggagatatctaagtcggagttggagatatctaagtcggagttggagatatctaagtcggagttggagatatctaagtcggagttggagatatctaagtcggagttggagatatctaagtcggagttggagatatctaagtcggagttggagatatctaagtcggagttggagatatctaagtcggagttggagatatctaagtcggagttggagatatctaagtcggagttggagggag
T atttcatcttcaatttagaatcttgccaatcaagccctctcgaagttggcaatatctataactcaac GCTTCTGAGATTCTAAGTACCTTAGATGCCAAGTACATTACTATATTGGTGTTATCGGGTCTTCCAA cctccattcaaggacttaattgaatcagt TAGGTTTAAT TAGGTTTAAT TAGGTTTAATT atttgccaatcaagccctctgaagttggcaatatctataactcaac CTTCGAGATTCTAAGTACCATTACTATATTGGTGTTATGGGGTCTTCCAA cctccattcaaggacttgaattgactagt TAGGTTTAATT tcttgccaatcaagccctctgaagttggcaatatctataactcaacctctgcttctgaggattctaag TAGGTTTAATT tcttgccaatcaagccctctgaagttggcaatatctataactcaacctctgcttctgaggattctaag TAGGTTTAATT tcttgccaatcaagccctctgaagttggcaatatctataactcaacctctgcttctgaggattctaag TAGGTTTAATTTCATCTT cttgccaatcaagccctctgaagttggcaatatctataactcaacctctgcttctgaggattctaag TAGGTTTAATTTCATCTT TGCCAATCAAGCCCTTCGAAGTTGGCAATATCTATAATTGGTGTTAGGGGTCTTCCAACTCCTCCATCCA
T atttatcttcaatttagaatcttgccaatcaagccctctgaagttggcaatatctaaactcaac GCTTCTGAGATTCTAAGTACCTTAGATGCCAAGTACATTACTATATTGTGTGTTATCGGGTCTTCCAA cctccattcaaggccttaggagttggcaatatctaaactcagcttctggagttcaa CTTAGATGCCAAGTACATTACTATATTGGGTTATCGGGTCTTCCAACTCCCACTCCAAGAAGACTTAA TAGGTTTAATT ttgccaatcaagcccttcgaagttggcaatatctaaaccaagccttgtttgggagttcaag CTTAGATGCCAAGTACATTACTATAATTGGGTTATCGGGTCTTCCAACTCCACTCCCATTCAAGACTTAA TAGGTTTAATTT ttgccaatcaagcccttggaagttggcaatatctaaaccaagccttgtttggggattcaag CTTAGATGCCAAGTACATTACTATAATTGGGTTATCGGGTCTTCCAACTCCCACTCCCACTCAAGACTTAAT TAGGTTTAATTTCATCTTC TGCCAATCAAGCCCTCTGGAAGTTGGCAATATCTATAACTCAACCTCGGCTCTGAGATTCTAAGTAC TAGGTTTAATTTCATCTTCTAAT TGCCAATCAAGCCCTCTGGAAGTTGGCAATATCTATAACTCAACCTCGGCTCTGAGATTCTAAGTAC TAGGTTTAATTTCATCTTCTAATT GGCAATCAAGCCCTCTCGAAGTTGGCAATATCTATAACTCAACCTCGGTCTGGAGATTCTAAGTAC CAATCAAGCCCTCTCGAAGTTGGCAATATCTATAACTCAACCTCGGTCTGGAGATTCTAAGTAC CAATCAAGCCCTCTCGAAGTTGGCAATATCTATAACTCAACCTCGGTCTGGAGATTCTAAGTAC CAATCAAGCCCTCTCGAAGTTGGCAATATCTATAACTCAACCTCGGTCTGGAGATTCTAAGTAC GCCAATCAAGCCCTCTCGAAGTTGGCAATATCTATAACTCAACCTCGGTTCTGAGAATTCTAAGTAC GCCAATCAAGCCCTCTCGAAGTTGGCAATATCTATAACTCAACCTCGGTTCTGAGAATTCTAAGTAC GCCAATCAAGCCCTCTCGAAGTTGGCAATATCTATAACTCAACCTCGGTTCTGAGAATTCTAAGTAC GCCATCAAGCCCTCTCGAAGTTGGCAATATCTATAACTCAACCTCGGTTCTGAGATTCTAAGTAC GCCATCAAGCCCTCTCGAAGTTGGCAATATCTATAACTCAACCTCGGTTCTGAGATTCTAAGTACC tgggtttactggggtcttccaaetcctccattcaaggagttggcaatatctaagtgactcg TAGGTTTAATTTCATCTTCTAATTTAG GCCCCTCGGAAGTTGGCAATATCTATAACTCAACCTCGGTTCTGAGATTCTAAGTACC gGCCTTCCGACTCCCCCCCCGACGTGGCAATATCTATAACTCAACCTCGGTTCTGAGATTCTAAGTACCT gggtttactggggtcttccaaetcctccattcaaggacttaagtgacttgg TAGGTTTAATTTCATCTTCTAATTTAGGAATCT cCCCCCGCAGATATCTATAACTCAACCTCGGCTTCGGAGATTCTAAGTACCTAGGTGCCA ggctttccaaetcctccatccatggagttggcaatatctagactcaggttctgg TAGGTTTAATTTCATCTTCTAATTTAGGAATCT cCCCACTCCCCCCGACTATCTATAACTCAACCTCGGCTTCGGAGATTCTAAGTACCTAGGCCAAGTACCT GGCTTCCAACTCCCCCCCCCC
TAGGTTTAAT aatcttigccaatcaagccctctcgaagttiggcaatatctataactcaacctcgttcttgagattcta CTTAGATGCCAAGTACATTACTATAATTGGTGTTATCGGGTCTTCCAACTCCCCATTCAAGACTTAA ctgt TAGGTTTATTT tcttgccaatcaagcccctctgaagttiggcaatatctataactcaacctctgttctgagattctaag CTTAGATGCCAAGTACATTACTATAATTGGTGTTATCGGGTCTTCCAACTCCCCATTCAAGACTTAA tctg TAGGTTTATTTCATCTT tcttgccaatcaagcccctctgaagttiggcaatatctaacacctctgttctgagattctaagt CTTAGATGCCAAGTACATTACTATAATTGGTGTTATCGGGTCTTCCAACTCCTCCATTCAAGACTTAAT TAGGTTTAATTTCATCTTC TGCCAATCAAGCCCCTCGGAAGTTGGCAATATCTATAACTCAACCTCGCTTCTGAGATTCTAAGTAC ATGCCAATCAATACTAAGCCCCTCCGAAGTTGGCAATATCTATAACTCAACCTCGCTTCTGAGATTCTAAGTAC CATTAGATACATATAATTGGTGTTATCGGGGTCTTCCAACTCCTCCATTCAAGACCTTAATTGACTCAACCTCTGCTTCGAGATTGCTAAGTACC CATTAGATTAACTAAGCCCCTCCGAAGTTGGCAATATCTATAACTCAACCTCGCTTCTGAGATTCTAAGTAC CATTAGATGCCAAGTACCTTAATTAGTAGACC CATTAGATGCAAGGCGTTCCAACTCCCATTCAAGGACTTAATTGACTAGC TAGGTTTAATTTCATCTTCAATTTAG GCCAATCAAGCCCCTCTCGAAGTTGGCAATATCTATAACTCAACCTCGCTCTGGAGATTCTAAGTACC CATTAGATTAATTAAGTAGACC CATTAGATGCAAGGCCTTCCAACTCCCCATTCAAGGACTTAATTAGCTAGACCTCTGCAGATTGCTAAGTACC CATTAGATGCAAGGCCTTCCAACTCCCCATTCAAGGACCTTAAGTAGCC CATTAGATGCAAGGCCCTCCCAACTAGGGAATATCTATAACTCAACCTCTGCTTCTGGAGATTCTAAGTACC Cattactatattggtgttatcgggtttccaagtccatcctcattcaaggacttaattggacttgg TAGGTTTACATTAGTAGGCAATATCTATAACTCAACCTCGCGTCTGGAGATTCTAAGTACCT TGGCAATCAAGCCCCTCCGAAGTTGGCAATATCTATAACTCAACCTCGCGTCTGGAGATTCTAAGTACCT TGGCAATCAAGCCCCTCCGAAGTTGGCAATATCTATAACTCAACCTCGCGTCTGGAGATTCTAAGTACCT TGGCAATCAGGGCAATATGGAATACTATAACTCAACCCTCGCTCTGGAGATTCTA
TAGGTTTAATTT tcttgccaatcaagccctctgaagttggcaatatctataactcaacctctgcttctgagattctaag CTTAGATGCCAAGTACATTACTATAATTGGTGTTATCGGGTCTTCCAACTCCTCCATTCAAGACTTAA TAGGTTTAATTTCATCTT tctgccaatcaagccctctgaagttggcaatatctataactcaacctctgcttctgaagttctaagt TTAGATGCCAAGTACATTACTATAATTGGTGTTATCGGGTCTTCCAACTCCTCCATTCAAGACTTAATTGAGACTTAATTGAGTGTTACGGGTCTTCCAACTCCTCCATTCAAGACTTAATTGACTAGGTGTACGGGTCTTCCAACTCCTCCACTCCAAGACTTAATTGACT TAGGTTTAATTTCATCTTC TGCCAATCAAGCCCTCTCGAAGTTGGCAATATCTATAACTCAACCTCTGCTTCTGAGATTCTAAGTAC CATTACAATATGGTGTTACGGGTCTTCCCAACTCCACTCCAATTCAAGACTTAATTGACT TAGGTTTAATTTCATCTTCTAATTTAG GCCAATCAAGCCCTCTCGAAGTTGGCAATATCTATAACTCAACCTCTGCTTCTGAGATTCTAAGTAC CATTACAATATGGTGTTACGGGTCTTCCCAACTCCACTCCCACTCCAAGCACTTAATTGACTCAGGACTTATTGATAACTCAACCTCTGGTCTTGAGATTCTAAGTACC TAGGTTTAATTTCATCTTCTAATTTAG GCCAATCAAGCCCTCTCGAAGTTGGCAATATCTATAACTCAACCTCTGCTTCTGAGATTCTAAGTACC cattactataattggtgttatcgggtcttccaactcctactcaagacttagtggactattggactattggacttctggggtcttccaactcctcattcaagacttagtgacttgg TAGGTTTAATTTCATCTTCTAATTTAG GCCACTCGGAAGTTGGCAATATCTATAACTCAACCTCTGCTTCTGAGATTCTAAGTACCT ggtcttccaactcctcattcaagacttagtgacttgg TAGGTTTAATTTCATCTTCTAATTAG GCCACTCGGAAGTTGGCAATATCTATAACTCAACCTCTGCTTCTGAGATTCTAAGTACCT ggtcttccaactcctcattcaagacttagtgacttgg TAGGTTTAATTAGCATTCATAACTCGAAGTTGGCAATATCTATAACTCAACCTCTGCTCTGGAGATTCTAAGTACCTT ggtctccaactcctcatccaagacttagtgacttgg TAGGTTTAATTCATCTCTAATTAGA CCCCTCCGAAGTGGCAATATCTATAACTCAACCTCTGCGCTCTGGAGATTCTAAGTACCTT ggtctcccaactccctcatccaagacttagtggccattacttggattct
TAGGTTTAATTTCATCTT tdtgccaatcaagccctctcgaagttggcaatatctataactcaagcctctggtttggtatctaggtttcaggtttcaagtccaatcaat
TAGGTTTAATTTCATCTTCTĞCCAATCAAĞCCCTCTCĞAAĞTTĞĞCAATATCTATAACTCAACCTCTĞCTTCTĞAĞATTCTAAĞTACATGCCAAGTACATTACTATCGGGTCTTCCAAGTCCTCCAACTCCCTCC
TAGGTTTAATTTCATCTTCTAATTGCCAATCAAGCCCTCTCGAAGTTGGCAATATCTATAACTCAACCTCGGCTTCGAGATTCTAAGTACGCCAAGTACATTACTATAATTGGGTGTTATCGGGTCTTCCAAGTCCCCATTCAAGAGCTTAATTGACTCGtaggtttaatttcatttcaatttagTGCCAATCAAGCCCTCTCGAAGTTGGCAATATCTATAACTCAACCTCGGTCTGGAGATTCTAAGTACCATTACTATAATTGGTGTTATCGGGTCTTCCAAGTCCCCACTCCCATCAAGACTTGAAGACTTAACTGAAGCTCTAAGTACTAGGTTTAATTTCATCTTCTAATTTAGGCCAATCAAGCCCTTCTCGAAGTTGGCAATATCTATAACTCAACCTCGCTTCTGAGATTCTAAGTACCcattactataattggtgttatcgggtcttccaactcctccattcaagacttaattgactctgTAGGTTTAATTTCATCTTCTAATTTAGCCAATCAAGCCCTCTCGAAGTTGGCAATATCTATAACTCAACCTCGCTTCTGAGATTCTAAGTACCcattactataattggtgttatcgggtcttccaactcctccattcaagacttaattgactctgTAGGTTTAATTTCATCTTCTAATTTAGCAATCAAGCCCTCTCGAAGTTGGCAATATCTATAACTCAACCTCGCTTCTGAGATTCTAAGTACCggtcttccaactcctccattcaagacttaattgactctgTAGGTTTAATTTCATCTTCTAATTTAGgccctctcgaagttggcaatatctataactcaacctctgcttctgagattctaagtaccttagatgccGGTCTTCCAACTCCCCACTCCCAAGCTGACACTTAACTAAC
taggtttaatttcatcttctaatttagTGCCAATCAAGCCCTCTCGAAGTTGGCAATATCTATAACTCAACCTCGCTTCTGAAGTTCTAAGTACCATTACTATAATTGGGTGTTAATTGGGGCTCTCCAACTCCCCATTCAAGAGCTTGAAGTAGCCAAGATTCTATAACTCAACCTCGGCTTCTGAGATTCTAAGTACCTAGGTTTAATTTCATCTTCTAATTTAGGCCAATCAAGCCCTCTCGAAGTTGGCAATATCTATAACTCAACCTCGGCTCTGGAGATTCTAAGTACCtgtatcgggtcttccaactcccattcaagacttaattgactctgTAGGTTTAATTTCATCTTCTAATTTAGCAATCAAGCCCTCTCGAAGTTGGCAATATCTATAACTCAACCTCTGCTTCTGAGATTCTAAGTACCtgtatcgggtcttccaactcccattcaagacttaattgactctgTAGGTTTAATTTCATCTTCTAATTTAGGccAtTCAAGCCCTCTCGAAGTTGGCAATATCTATAACTCAACCTCTGCTTCTGAGATTCTAAGTACCtgtatcgggtcttccaactccccattcaagacttaattgactctgTAGGTTTAATTTCATCTTCTAATTTAGgcccttccgaagttggcaatatctataactcaacctctgcttctgagattctaagtagccGGTCTTCCAACTCCCCATTCCAAGCTCAAGCACTAATTGAACTCAACCTCGGCTCTGGAGATTCTAAGTACCTAAGTACCTAAGTACCTAAGTACCTAAGTACCTAAGTACCTAAGTACCTAAGTACCTTAGATGCCAggtcttccaactccccattcaagacttaattgactctgTAGGTTTAATTTCATCTTCTAATTTAGAATCTctctcgaagttggcaatatctataactcaacctctgcttctgagattctaagtagccaagggtcttccaactccccattcaagacttaattgactctgTAGGTTTAATTTCATCTTCTAATTTAGAATCTctctcgaagttggcaatatctataactcaacctctgcttctgagattctaagtagccaagggtcttccaactccccattcaagacttaattgactctgTAGGTTTAATTTCATCTTCTAATTTAGAATCTctctgaagttggcaatatctataactcaacctctGCTCTGAGATTCTAAGTACCTAGAGTGCCAAGTACgtcttccaactcccattcaagacttaattgactctgTAGGTTTAATTTCAACTTCTAATTTAGAATCTctctgaagttggcaatatctataactcaacctCTGCTTCTGAGATTCTAAGTACCTAGAGTGCCAAGTACATgtcttccaactcccattcaagacttaattgactctgTAGGTTTAATTTCAATTTAGAATCTctctgaagttggcaatatctataactcaacctCTGCTTCTGAGATTCTAAGTACCTTAGATGCCAAGTACATTctccaactcccattcaagacttaattgactctgTAGGTTTAATTCAACTTCAAGTTAGGCAATATCTATAACTCAACCTCGCCTCTGGGCAATTCTAAGTACCTTAGATGCCAAGTACCTTAGATGCCAAGTACATTACTACTAAGT
TAGGTTTAATTTCATCTTCTAATTTAĞ GCCAATCAAGCCTTCTCGAAGTTGGCAATATCTATAACTCAACCTCGCTTCTGAGATTCTAAGTACC cattactatattggtgttatcgggtcttccaactcctcattcaagacttaattgactctg TAGGTTTAATTTCATCTTCTAATTTAG CAATCAAGCCCTCTCGAAGTTGGCAATATCTATAACTCAACCTCGCTTCTGAGATTCTAAGTACC tgttatcgggtcttccaactcctccattcaagacttaattgactctg TAGGTTTAATTTCATCTTCTAATTTAG CAATCAAGCCCTCTCGAAGTTGGCAATATCTATAACTCAACCTCGCTTCTGAGATTCTAAGTACCT TAGGTTTAATTTCATCTTCTAATTTAG GCCCCCCCCGAAGTTGGCAATATCTATAACTCAACCTCGCTTCTGAGATTCTAAGTACCTT GGCTTCAATTTCATCTTCTAATTTAG GGCCATCAAGCCCCTCCGAAGTTGGCAATATCTATAACTCAACCTCGCTTCTGAGATTCTAAGTACCTT GGCTTCCAAATCCCAACCTCCCACTCCCAAGTTGGCAATATCTATAACTCAACCTCGCTTCTGAGATTCTAAGTACCTTAGATGCCA GGCCTTCCAACTCCCCCCCTCCGAAGTTGGCAATATCTATAACTCAACCTCGCTTCTGAGATTCTAAGTACCTTAGATGCCA GGCCTTCCAACTCCCCCCCCCC
TAGGTTTAATTTCATCTTCTAATTTAGCAATCAAGCCCTCTCGAAGTTGGCAATATCTATAACTCAACCTCGCTTCTGAAGTTCTAAGTACCtgttatcgggtcttccaactcctccattcaagacttaatgacttgTAGGTTTAATTTCATCTTCTAATTTAGCAATCAAGCCCTCTCGAAGTTGGCAATATCTATAACTCAACCTCGCTTCTGAGATTCTAAGTACCTTgggtcttccaactcctccattcaagacttaatgacttgTAGGTTTAATTTCATCTTCTAATTTAGgccctctgaagttggcaatatctataactcaacctctgcttctgagattctaagtgccGGTCTTCCAACTCCTCCAACTCCCCATTCAAGACTTAATTGACTCTGTAGGTTTAATTTCATCTTCTAATTTAGAATCCCTCTCGAAGTTGGCAATATCTATAACTCAACCTCTGCTTCTGAGATTCTAAGTACCCAggtcttccaactcctccattcaagacttaattgactctgTAGGTTTAATTTCATCTTCTAATTTAGAATCctctcgaagttggcaatatctataactcaacctcgcttctgagattctaagtgccaagggtcttccaactcctccattcaagacttaattgactcgTAGGTTTAATTTCATCTTCTAATTTAGAATCTctctcgaagttggcaatatctataactcaacctcgcttctgagattctaagtgccaagggtcttccaactcctccattcaagacttaattgactcgTAGGTTTAATTTCATCTTCTAATTTAGAATCTctctcgaagttggcaatatctataactcaacctcgcttctgagattctaagtgccaaggtcttccaactcctccattcaagacttaattgactcgTAGGTTTAATTTCATCTTCTAATTTAGAATCTCCGAAGTTGGCAATATCTATAACTCAACCTCGCTTCTGAGATTCTAAGTACCTTAGATGCCAAGTACgtcttccaactcctccattcaagacttaattgactcgTAGGTTTAATTTCATCTTCTAATTTAGAATCTCGAAGTTGGCAATATCTATAACTCAACCTCGCTTCTGAGATTCTAAGTACCTTAGATGCCAAGTACATTcttccaactcctccattcaagacttaattgactcgTAGGTTTAATTTCATCTTCTAATTTAGAATCTCGAAGTTGGCAATATCTATAACTCAACCTCGCTTCTGAGATTCTAAGTACCTTAGATGCCAAGTACATTcttccaactcctccattcaagacttaattgactcgTAGGTTTAATTCACTCTCTAATTTAGAATCTCGAAGTTGGCAATATCTATAACTCAACCTCGCTTCTGAGATTCTAAGTACCTTAGATGCCAAGTACATTcttccaactcctccattcaagacttaattgactctgTAGGTTTAATTCACTCTCTAATTAGAATCTAAGTTGGCAATATCTATAACTCAACCTCGCTTCTGGAGATTCTAAGTACCTTAGATGCCAAGTACATTACTAAGTACCTTAGATGCCAAGTACATTACTAAGTACCTTAGATGCCAAGTACATTACTAAGTACCTTAGA
TAGGTTTAATTTCATCTTCTAATTTAGCAATCAAGCCCTCTCGAAGTTGGCAATATCTATAACTCAACCTCGCTTCTGAAGTTCTAAGTACCTTgggtcttccaactcctccattcaagacttaatgaccttgTAGGTTTAATTTCATCTTCTAATTTAGgccctctgaagttggcaatatctataactcaacctcgcttctgaagttctaagtaccttagatgccGGTCTTCCAACTCCCCCATTCAAGACTTAAGCTCGTTAGGTTTAATTTCATCTTCTAATTTAGAATCCCTCTCGAAGTTGGCAATATCTATAACTCAACCTCGCTTCTGAGATTCTATAGTGCCAggtcttccaactcctccattcaagacttaattgactctgTAGGTTTAATTTCATCTTCTAATTTAGAATCTctctgaagttggcaatatctataactcaacctctgcttctgaagttctaagtgccaagggtcttccaactcctccattcaagacttaattgactctgTAGGTTTAATTTCATCTTCTAATTTAGAATCTctctgaagttggcaatatctataactcaacctctgcttctgaagttctaagtgccaagggtcttccaactcctccattcaaggacttaattgactctgTAGGTTTAATTTCATCTTCTAATTTAGAATCTCCCAAGTTGGCAATATCTATAACTCAACCTCGCTTCTGAGATTCTAAGTACCTTAGGTGCCAAGTACATGTCTTCCAACTCCTCCATCCAAGCCTCATGACGCCAAGTACCTTGACTTCGCTTCTGAGATTCTAAGTGCCAAGTACCTTAGATGCCAAGTACATTAGGTTTAATTTCATCTTCTAATTTAGAATCTCGAAGTTGGCAATATCTATAACTCAACCTCAGCTTCTGAGATTCTAAGTACCTTAGATGCCAAGTACATgtcttccaactcctccattcaaggacttaattgactctgTAGGTTTAATTCATCTTCTAATTTAGAATCTCGAAGTTGGCAATATCTATAACTCAACCTCGCTTCTGAGATTCTAAGTACCTTAGATGCCAAGTACATTcttccaactcctccattcaaggacttaattgactctgTAGGTTTAATTCATCTTCTAATTTAGAATCTAAGTTGGCAATATCTATAACTCAACCTCGCTTCTGAGATTCTAAGTACCTTAGATGCCAAGTACATTcttccaactcctccattcaaggacttaattgactctgTAGGTTTAATTCATCTTCTAATTTAGAATCTAAGTTGGCAATATCTATAACTCAACCTCGCTTCTGAGATTCTAAGTACCTTAGATGCCAAGTACATTcttccaactcctccattcaaggacttaattgactctgTAGGTTTAATTCATCTCTAATTTAGAATCTCGAAGTTGGCAATATCTATAACTCAACCTCGCTTCTGGAGATTCTAAGTACCTTAGATGCCAAGTACATTACTATAActtccaactccccattccaaggacttaattgactctgTAGGTTTAATTCATCTATCTAATTCAAGTACCTTAGATGCCAAGTACATTAGATCTTAAGTACCTTAGATGCCAAGTACATTAGATGGCAAGTTACAT
TAGGTTTAATTTCATCTTCTAATTTAGgccctctgaagttggcaatatctataactcaacctctgcttctgagattctaagtaccttagatgccGGTCTTCCAACTCCCATTCAAGACTTCAAGACCTTAATTGACTCTGTAGGTTTAATTCATCTTCTAATTTAGAATCCCCTCTGAAGTTGGCAATATCTATAACTCAACCTCGCTTCTGAGATTCTAAGTACCCAggtcttccaactcctccattcaagacttaattgactctgTAGGTTTAATTTCATCTTCTAATTTAGAATCTctctcgaagttggcaatatctataactcaacctctgcttctgagattctaagtaccttagatgccaagggtcttccaactcctccattcaagacttaattgactctgTAGGTTTAATTTCATCTTCTAATTTAGAATCTctctcgaagttggcaatatctataactcaacctctgcttcTGAGATTCTAAGTACCTAGAGTGCCAAGTAGTCTTCCAACTCCACTCCATTCAAGTTAGCACTTGGTAGGTTTAATTTCATCTTCTAATTTAGAATCTCCGAAGTTGGCAATATCTATAACTCAACCTCGCTTCTGAGATTCTAAGTACCTTAGATGCCAAGTACAgtcttccaactcctccattcaagacttaattgactctgTAGGTTTAATTTCATCTTCTAATTTAGAATCTCGAAGTTGGCAATATCTATAACTCAACCTCGCTTCTGAGATTCTAAGTACCTTAGATGCCAAGTACATTcttccaactcctccattccaagacttaattgactctgTAGGTTTAATTCATCTTCTAATTTAGAATCTAAGTTGGCAATATCTATAACTCAACCTCGCTTCTGAGATTCTAAGTACCTTAGATGCCAAGTACATTcttccaactcctccattccaagacttaattgactctgTAGGTTTAATTCATCTCAATTAGAATCTAAGTTGGCAATATCTATAACTCAACCTCGCTTCTGAGATTCTAAGTACCTTAGATGCCAAGTACATTACTTAAActtccaactcctccattccaagacttaattgactctgtaggttttaatttcatcttctaatttagaatcttgccCAATATCTATAACTCAACCTCGCTTCTGAGATTCTAAGTACCTTAGATGCCAAGTACATTACTATAATCTGACCTTCAGAGTCCAAGTACCTTAGATGCCAAGTACATTACTTAAGTACCTTAGATGCCAAGTACATTACTTAAGTACCTTAGATGCCAAGTACATTACTTAACTCAACCTCGCACCTCGCTTCTGAGATTCTAAGTACCTTAGATGCCAAGTACATTACTATAACTCGACcttccaagtacttaattggttatcgggtcttccaacTAGGTTTAATTCACTTCAATTCAAGTACTTGCCAACTATAACTCAACCTCCACCTCCGCTTCTGAGATTCTAAGTACCTTAGATGCCAAGTACATTACTAAGTACCTTAGATGGCCAAGTACATTACTTAAGTACCTTAGATGCCAAGTACATTACTTAAGTGCCAAGTACATTACTGACCTTAGATGGCAAGTACATTACTTGAAGTACCTTAGATGGCCAAGTACATTACTGACACTTGCCAActtccaagtacttaattgatgtcttggtacttggt
TAGGTTTAATTTCATCTTCTAATTTAGAATCT ctctcgaagttggcaatatctataactcaacctctgcttctgagattctaagtaccttagatgccaag ggtcttccaactcctccattcaagacttaattgactctg TAGGTTTAATTTCATCTTCTAATTTAGAATCT CTCGAAGTTGGCAATATCTATAACTCAACCTCGCTTCTGAAGTTCTAAGTGCCAAGTA GTCTTCCAACTCCCCATTCAAGACTTAAATGACCTGG TAGGTTTAATTTCATCTTCTAATTTAGAATCT CGCAAGTTGGCAATATCTATAACTCAACCTCGCTTCTGAAGATTCTAAGTGCCAAGTA gtcttccaactcctccattcaagacttaattgactctg TAGGTTTAATTTCATCTTCTAATTTAGAATCT CGAAGTTGGCAATATCTATAACTCAACCTCGCTTCTGAGATTCTAAGTGCCAAGTACCAAGTACCAT gtcttccaactcctccattcaagacttaattgactctg TAGGTTTAATTTCATCTTCTAATTTAGAATCT CGAAGTTGGCAATATCTATAACTCAACCTCGCTTCTGAGATTCTAAGTGCCAAGTACCATT cttccaactcctccattcaagacttaattgactctg taggtttaatttcatcttctaatttagaatcttgcc CAATATCTATAACTCAACCTCGCTTCTGAGATTCTAAGTACCTTAGATGCCAAGTACATTACATAACT cttccaagtcctccattcaagacttaattgactctg TAGGTTTAATTTCATCTATTAGAATCTTGCCA CAATATCTATAACTCAACCTCGCTTCTGAGATTCTAAGTACCTTAGATGCCAAGTACATTACATAATTGGTG CTTCCAACTCCACTCCACTCATCAAGTACCTTGAGTACCTTAGATGCCAAGTACATTAGATCTTCAAGTCCTCAAGTACCTTAGATGCCAAGTACATTAGTAGCGGC TAGGTTTAATTCATCTCTCAATTTAGAATCTGCCAA CTATAACTCAACCTCACCTCGCTCTGGAGTTCTAAGTACCTTAGATGCCAAGTACATTAGATCGCCAAGTACATTCAAGTACCTTCAAGTACCTTAGATgatccttagatgccaagtacattactaatggtgttatcgggtcttccaac CTCCAACTCAACCTCAACCTCGCTGCTGGCAACCTTGAGTGCCAAGTACCTTAGATGGCCAAGTACCTTAGATGGCCAAGTACCTTCAAGTACCTTAGATGCCAAGTACCTTAGATGCCAAGTACATTAGACCTCGCACGC cttctgagattctaagtagccaagtacattactaatatggtgttatcgggtcttccaac CTCCACTCCAACCCACCCACCCCACCAACCCCACCCCCCC
TAGGTTTAATTTCATCTTCTAATTTAGAATCT CTCGAAGTTGGCAATATCTATAACTCAACCTCTGCTTCTGAAGATCTAAGTACCTTAGATGCCAAGTA GTCTTCCAACTCCCCATTCAAGACCTTAATTGACTCTG TAGGTTTAATTTCATCTTCTAATTTAGAATCT CGAAGTTGGCAATATCTATAACTCAACCTCGCTTCTGAAGATCTAAGTACCTTAGATGCCAAGTACA gtcttccaactcctccattcaagacttaattgactctg TAGGTTTAATTTCATCTTCTAATTTAGAATCT CGAAGTTGGCAATATCTATAACTCAACCTCGCTTCTGAGATTCTAAGTACCTTAGATGCCAAGTACAT gtcttccaactcctccattcaagacttaattgactctg TAGGTTTAATTTCATCTTCTAATTTAGAATCT AAGTTGGCAATATCTATAACTCAACCTCGCTTCTGAGATTCTAAGTGCCAAGTACATT cttccaagacttcattgagacttaattgactctg taggtttaatttcatcttctaatttagaatcttgcc CAATATCTATAACTCAACCTCGCTTCTGAGATTCTAAGTACCTTAGATGCCAAGTACATTAACTAAATTGGTG cttccaactcctccattcaagaacttaattgactcTG TAGGTTTAATTTCATCTTCTAATTTAGAATCTTGCCA CTATAACTCAACCTCACCTCGCTTCTGAGATTCTAAGTACCTTAGATGCCAAGTACATTAATTGGTG CTCCAATCCAACCTCAAGTACCTG TAGGTTTAATTCATCTTCAATTTAGAATCTTGCCAA cttctgagattctaagtagccttagtagatgacattactagatgccaagtacattactagatgccttagggtcttccaac CTCCAATTCAAGTACTGCA taggttttaatttcatcttctaatttagaatcttgccaatcaagcc cttctgagattctaagtagccttagatgccaagtacattactataattggtgttatcgggtcttccaac CTCCAATCAAGTACATTCAAGTACTGTGCA
TAGGTTTAATTTCATCTTCTAATTTAGAATCT CGAAGTTGGCAATATCTATAACTCAACCTCAGCTTCTGAAGTTCTAAGTGCCAAGTACA gtcttccaactcctccattcaagacttaattgacttgf TAGGTTTAATTTCATCTTCTAATTTAGAATCT AAGTTGGCAATATCTATAACTCAACCTCGCTTCTGAAGTTCTAAGTGCCAAGTACATT cttccaactcctccattcaagacttaattgacttgf taggttttaatttcatcttctaatttagaatcttgcc CAATATCTATAACTCAACCTCGCTTCTGAGATTCTAAGTGCCAAGTGCCAAGTACATT cttccaactcctccattcaagacttaattgacttgf TAGGTTTAATTTCATCTTCTAATTTAGAATCT CAATATCTATAACTCAACCTCGCTTCTGAGATTCTAAGTGCCAAGTGCCAAGTACATTACTATAA cttccaactcctccattcaagacttaattgactctgf TAGGTTTAATTTCATCTTCTAATTTAGAATCTTGCCA CTATAACTCAACCTCTGCTTCTGAGATTCTAAGTACCTTAGATGCCAAGTACATTACTATAATTGGTG CTTCCAACTCCATCAACTTAGCTGCGCA TAGgTTTAATTTCATCTTCtaatttagaatcttgccaatcaagcc cttctgagattctaagtagcctaagtagcattaattggtgttatcgggtcttccaac CTCCATCAAGCACTTAATTGACTGG
TAGGTTTAATTTCATCTTCTAATTTAGAATCT AAGTTGGCAATATCTATAACTCAACCTCGCTTCTGAAGTTCTAAGTACCTTAGATGCCAAGTACATT cttccaactcctccattcaagacttaattgacttgt taggtttaatttcatcttctaatttagaatcttgcc CAATATCTATAACTCAACCTCGCTTCTGAGATTCTAAGTACCTTAGATGCCAAGTACATTACTATAAA cttccaactcctccattcaagacttaattgactctgt TAGGTTTAATTTCATCTTCTAATTTAGAATCTTGCCA CTATAACTCAACCTCTGCTTCTGAGATTCTAAGTACCTTAGATGCCAAGTACATTACTATAATTGGATG cttccaactcctccattcaagagtctcaattgactcTGG TAGGTTTAATTTCATCTTCTAATTTAGAATCTTGCCA CTATAACTCAACCTCGCTCTGGAATTCTAAGTACCTGG CTTCCAAGTCACCTCCCATTCAAGACCTTGCCTGG taggtttaatttcatcttctaatttagaatcttgccaatcaagacc cttctgagattctaagtagccaagtacattactataattggtgttatcgggtcttccaac cccatccaagacttaattgactcTGG
taggtttaatttcatcttctaatttagaatcttgcc CAATATCTATAACTCAACCTCGCTTCTGAAGATTCTAAGTACCTTAGATGCCAAGTACATTACTATAA cttccaactcctccattcaagacttaattgactctg TAGGTTTAATTTCATCTTCTAATTTAGAATCTTGCCA CTATAACTCAACCTCGCTCCTGCTGCTGAGGATTCTAAGTACCTTAGATGCCAAGTACATTACTATAATTGGTG CTTCCAACTCCATCCAAGACCTTAATGACCTTG TAGGTTTAATTTCATCTTCTAATTTAGAATCTTGCCA CTATAACTCAACCCCCGCTCCTGGATCCTAAGTACCTTAGATGCCAAGTACATTACTATAATTGGTG CTTCCAACTCCATCCAAGACCTTAATGACCTCG TAGGTTTAATTTCATCTTCTATTTAGAATCTTGCCAA cttctaagtaccttagagtcctaagtgccaagtacattactataattggtgttatcgggtcttccaac CTCCATCAAGACCTTAATTGACCTCG taggtttaatttcatcttctaatttagaatcttgccaatcaagcc cttctgagattctaagtaccttagatgccaagtacattactataattggtgttatcgggtcttccaac tccattcaagacttaattgaccttg
TAĞĞTTTAATTTCATCTTCTATTTTCATCTTAĞAATCTTĞCCA CTATAACTCAACCTCACCTC
TAGGTTTAATTTCATCTTCTAATTTAGAATCTTGCCAA cttctgagattctaagtaccttagatgccaagtacattactataattggtgttatcgggtcttccaac CTCCATTCAAGACTTAATTGACTCTGT taggtttaatttcatcttctaatttgagatcttgccaatcaagcc cttctgagattctaagtaccttagatgccaagtacattactataattggtgttatcgggtcttccaac tccattcaagacttaattgacttgt
taggtttaatttcatcttctaatttagaatcttagccaatcaagcc cttctgagattctaagtaccttagatgccaagtacattactataattggtgttatcgggtcttccaac tccattcaagacttaattgactctg
TAGGITTAATTTCATCTTCATGATCATCATCATCAAGCC CCCCCCCCCC
taggtttaatttcatcttctaatttagaatcttggtgtttaccgggtcttccaagccc ttcgagattctaagtaccttagatgccaagtacattaattggtgttatcgggtcttccaact tccattcaagacttaattgactctg TAGGTTTAATTTCATCTTCTATCTTAGAATCTTGCCAATCAAGCCC tgagattctaagtaccttagatgtccaagtaccattaattggtgttatcgggtcttccaagtacttcaagacttaattgactctg
TAGGTTTAATTTCATCTTCTATTTCAGATCTTGCCAATCAAGCCCTC tgagattctaagtaccttagatgccaagtacattaattggtgttatcgggtcttccaagtacttcaagaacttaattgacttg
TAGGTTTAATTCATCTTCAATTTAGAATCTTGCCAATCAAGCCCTCTCGAAG tgggattctaagattaattgatgccaagtacattaattagtggttatcagggtttacaggattaattgatctg
TAGGTTTAATTTCATCTTCTAATTTTAGAATCTTGCCAATCAAGCCCTCTCGAAG gagattctaagtaccttagtaccttagtaccttagtaccttagtgccagtacattactatattggtgttatcgggtcttccaactgct
ATTTCATCTTCTATTTAGAATCTTGCCAATCAAGCCCTCTCGAAGTTGGCAATATCTATAACTCAAC agattctaagtaccttagtqccaagtacattactataattggtgttatcgggtcttccaactcctcc cttaattgacttg
TTCATCTTCTAAGTAATCTTGCCAATCAAGCCCTCTCGAAGTTGGCAATATCTATAACTCAACCT AGATTCTAAGTACCTTAGATGCCAAGTACATTACTATAATTGGTGTTATCGGGTCTTCCCAACTCCCC attgactctgt
gattctaagtaccttagatgccaagtacattactataattggtgttatcgggtcttccaactccca
gattctaagtaccttagatgccaagtacattactataattggtgttatcgggtcttccaactccca
gattctaattaccttagatgccaagtacattactataattggtgttatcgggtcttcccaactcccca
aagtaccttagatgccaagtacattactataattggtgttatcgggtcttccaactccattcaag
cttccaactcctccattcaagacttaattgactctgt
TTCCAACTCCTCCATTCAAGACTTAATTGACTCTG
TCCAACTCCTCCATTCAAGACTTAATTGACTCGT
caactcctccattcaagacttaattgactctgt
caactcctccattcaagacttaattgactctgt
aactectcattcaagacttaattgactetg
aactcctccattcaagacttaattgactctgt tccattcaagacttaattgactctgt

ccattcaagacttaattgactctg

ccattcaagacttaattgactctg

www.broadinstitute.org/igv/ C

☆ a



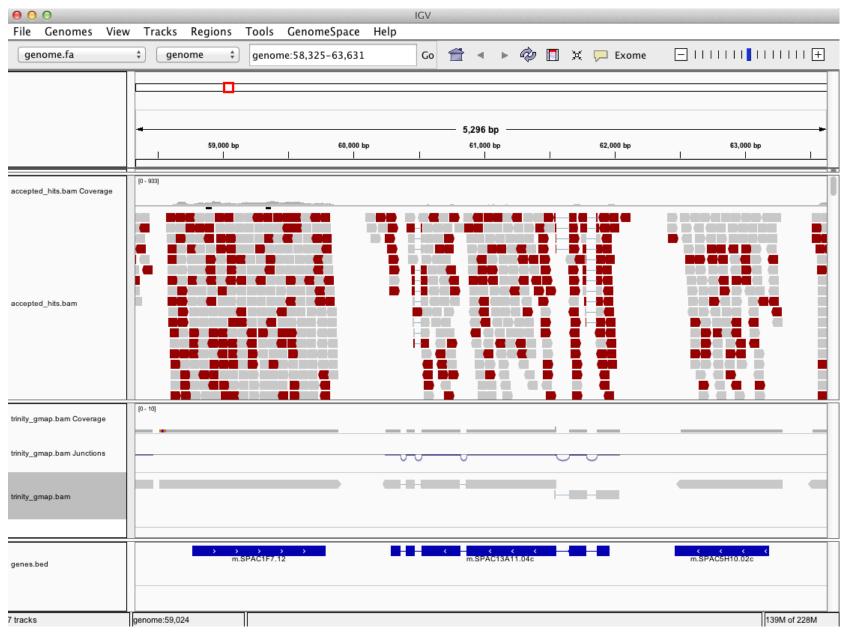
Overview

Can Examine Transcript Read Support Using IGV

🖲 🖸 🔵	Тгэг	🔀 IGV c <u>k</u> s Regions Tools GenomeSpace Help	
Trinity.fasta			
L		■ 254 bp	•
		p 100 bp 200 bp 100 bp	-
GSNO_SRR1582647.bowtie.csoi am Coverage GSNO_SRR1582647.bowtie.csoi am			
GSNO_SRR1582646.bowtie.csoi am Coverage GSNO_SRR1582646.bowtie.csoi am			
GSNO_SRR1582648.bowtie.csoi am Coverage GSNO_SRR1582648.bowtie.csoi am			
wt_SRR1582649.bowtie.csortec Coverage wt_SRR1582649.bowtie.csortec	a.		
wt_SRR1582650.bowtie.csortec Coverage wt_SRR1582650.bowtie.csortec	u.		
wt_SRR1582651.bowtie.csortec Coverage wt_SRR1582651.bowtie.csortec	d.		
		(_DN130_c0_g1_i]	- -

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

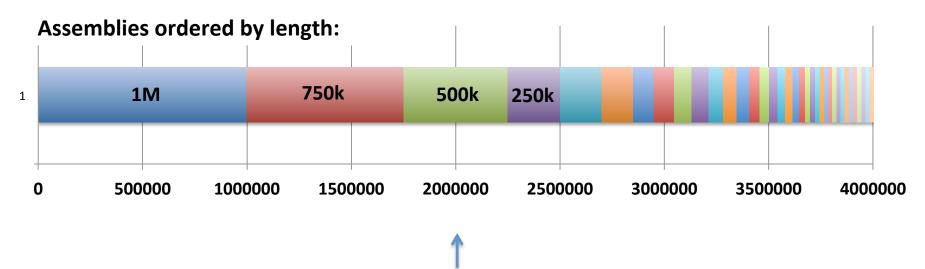
(Trinity transcripts aligned to the genome using GMAP)



The Contig N50 statistic

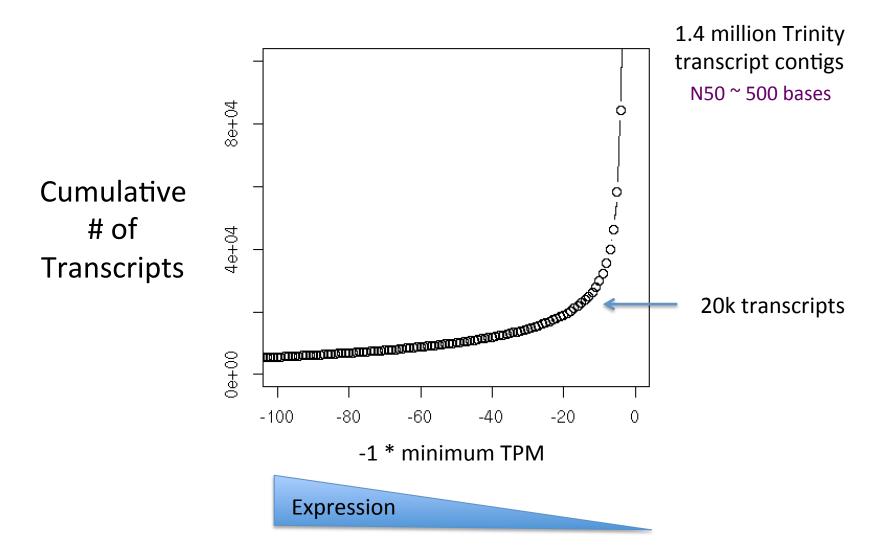
"At least half of assembled bases are in contigs that are at least **N50** bases in length"

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



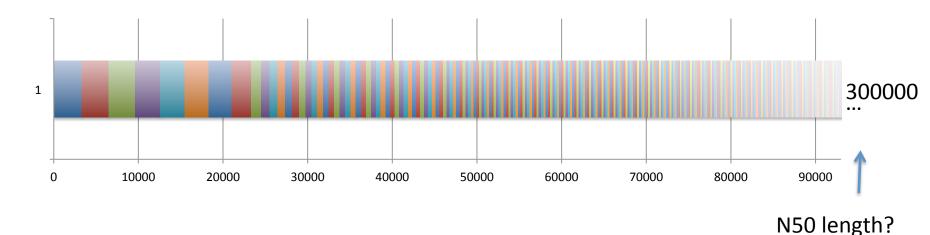
N50 contig length = 500k

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



* Salamander transcriptome

N50 Calculation for *Transcriptome* Assemblies??



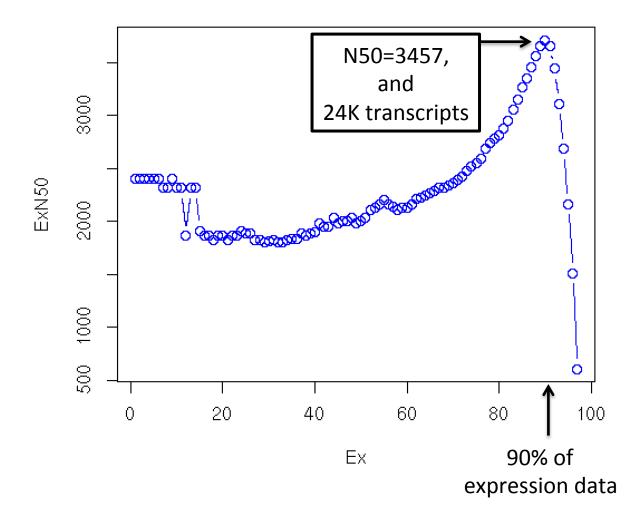
(small)

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

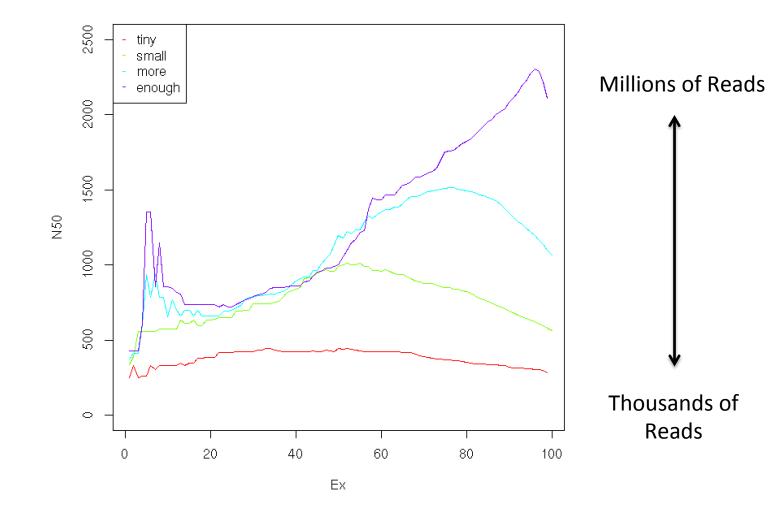
- Overzealous isoform annotation for long transcripts drives higher N50
- Very sensitive reconstruction for short lowly expressed transcripts drives lower N50

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

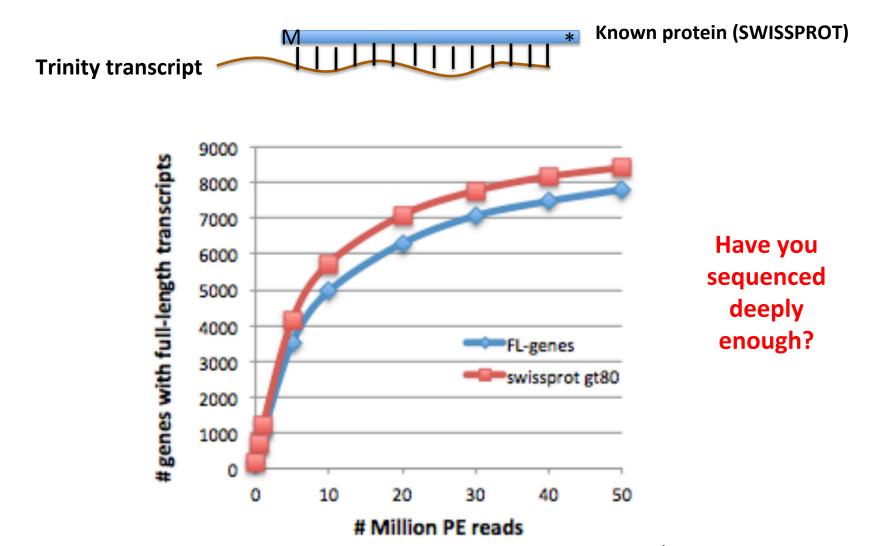


Note shift in ExN50 profiles as you assemble more and more reads.

* Candida transcriptome

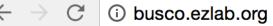
Evaluating the quality of your transcriptome assembly

Full-length Transcript Detection via BLASTX



* Mouse transcriptome

Haas et al. Nat. Protoc. 2013





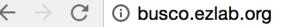
BUSC _{v2}

Assessing genome assembly and annotation completeness with <u>Benchmarking Universal Single-</u> <u>Copy Orthologs</u>

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.



UNIVERSITÉ

DE GENÈVE



☆

BUSC ,

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**
- **Fragmented BUSCOs** 139
- 222 **Missing BUSCOs**
- **Total BUSCO groups searched** 3023

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."

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

"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"

$$\log P(A, D) = \log \int_{\Lambda} P(D|A, \Lambda) P(A|\Lambda) P(\Lambda) d\Lambda$$

$$\approx \underbrace{\log P(D|A, \Lambda_{\text{MLE}})}_{\text{likelihood}} + \underbrace{\log P(A|\Lambda_{\text{MLE}})}_{\text{assembly prior}}$$

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

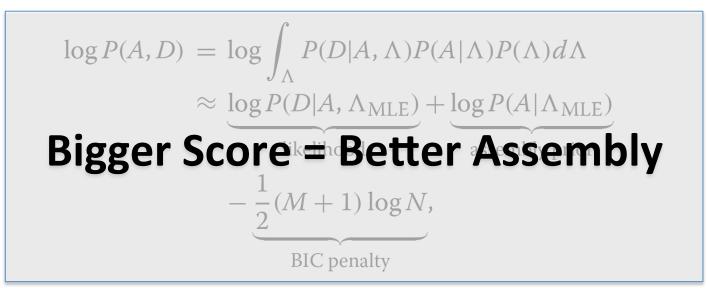
Li et al. Evaluation of de novo transcriptome assemblies from RNA-Seq data, Genome Biology 2014

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."

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

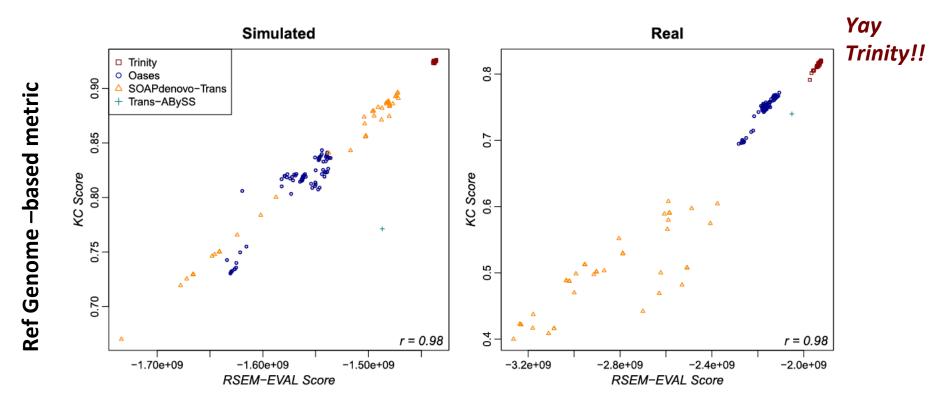
"the RSEM-EVAL score of an assembly is defined as the log joint probability of the assembly A and the reads D used to construct it"



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."



RSEM-EVAL Genome-free metric

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