BINF6308: Lecture 8

Instructor - Andrej Savol (a.savol@northeastern.edu)

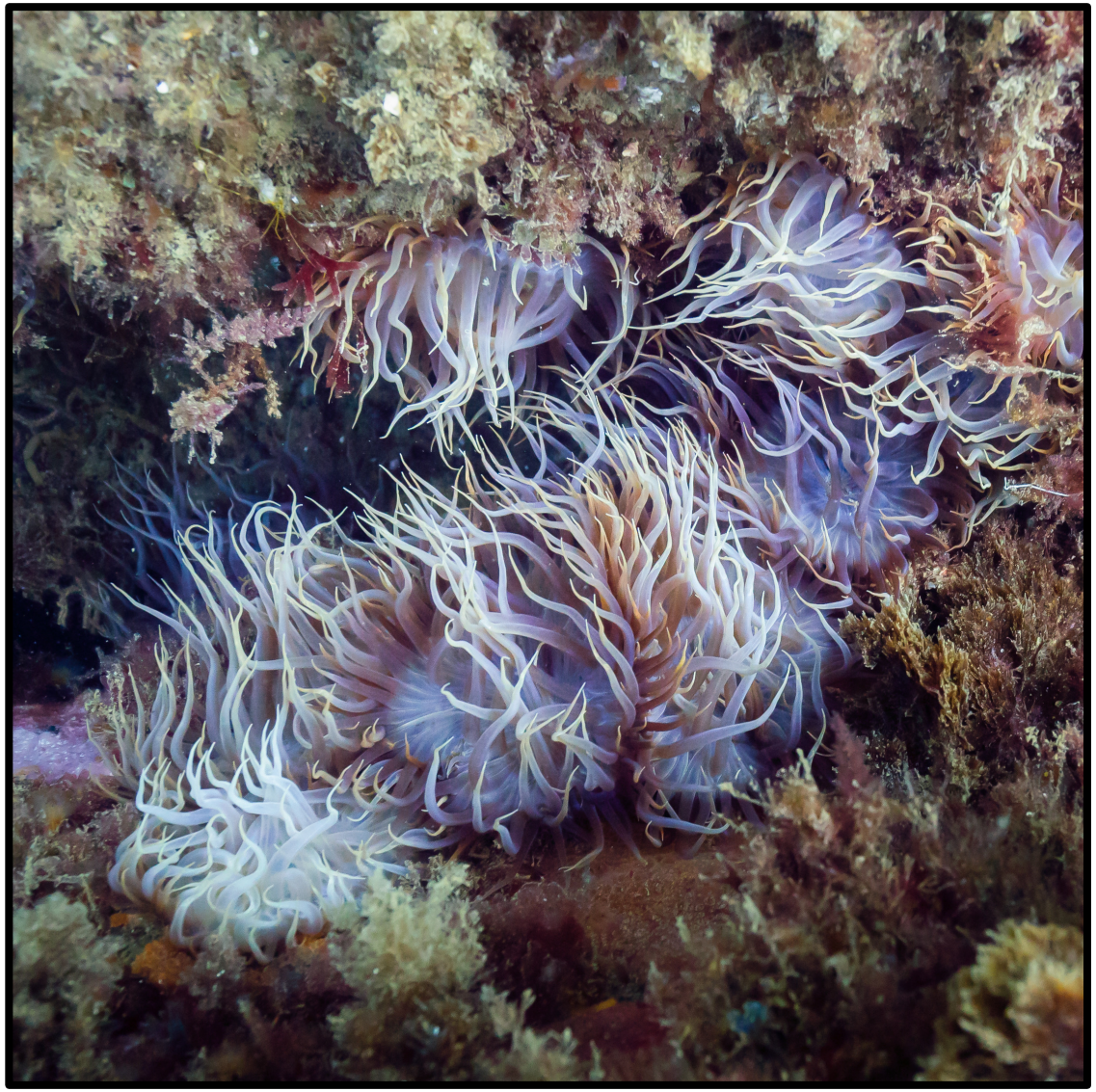
TA Section 12579 (in-person): Lindsey Alexanian (alexanian.l@northeastern.edu)
TA Section 12580 (online): Yicheng Zhang (zhang.yicheng@northeastern.edu)

Meeting times:
Section 1 (in-person): Monday 5:30-8:30pm ISEC 140
Section 2(online): Asynchronous

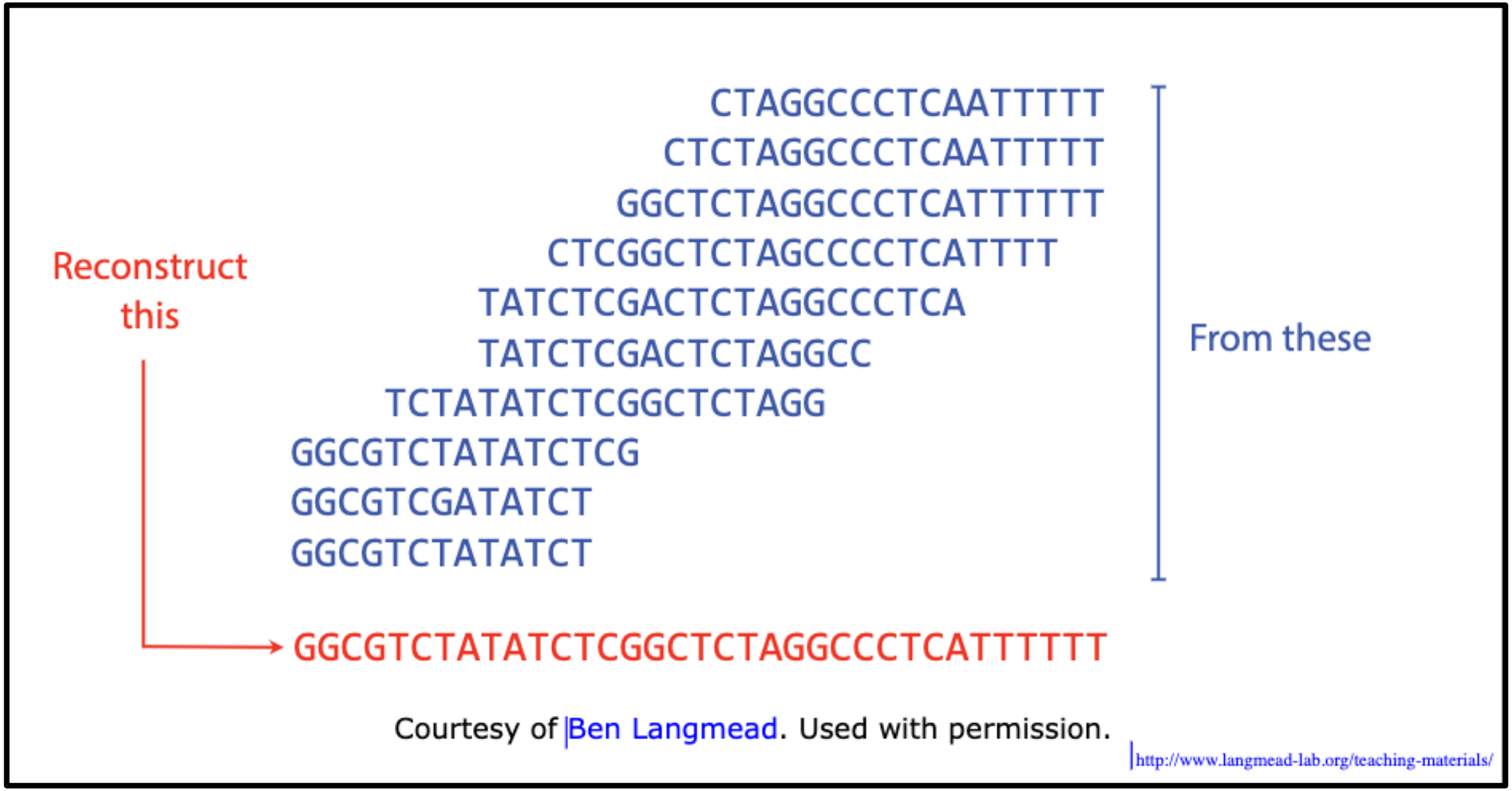
Lecture 8: Outline

- Transcriptome Assembly
  - Assembly v. alignment
  - Reference-guided v. *de novo*
  - Index-based alignment
  - Using Trinity (demo)
- Links in Unix
- Using IGV

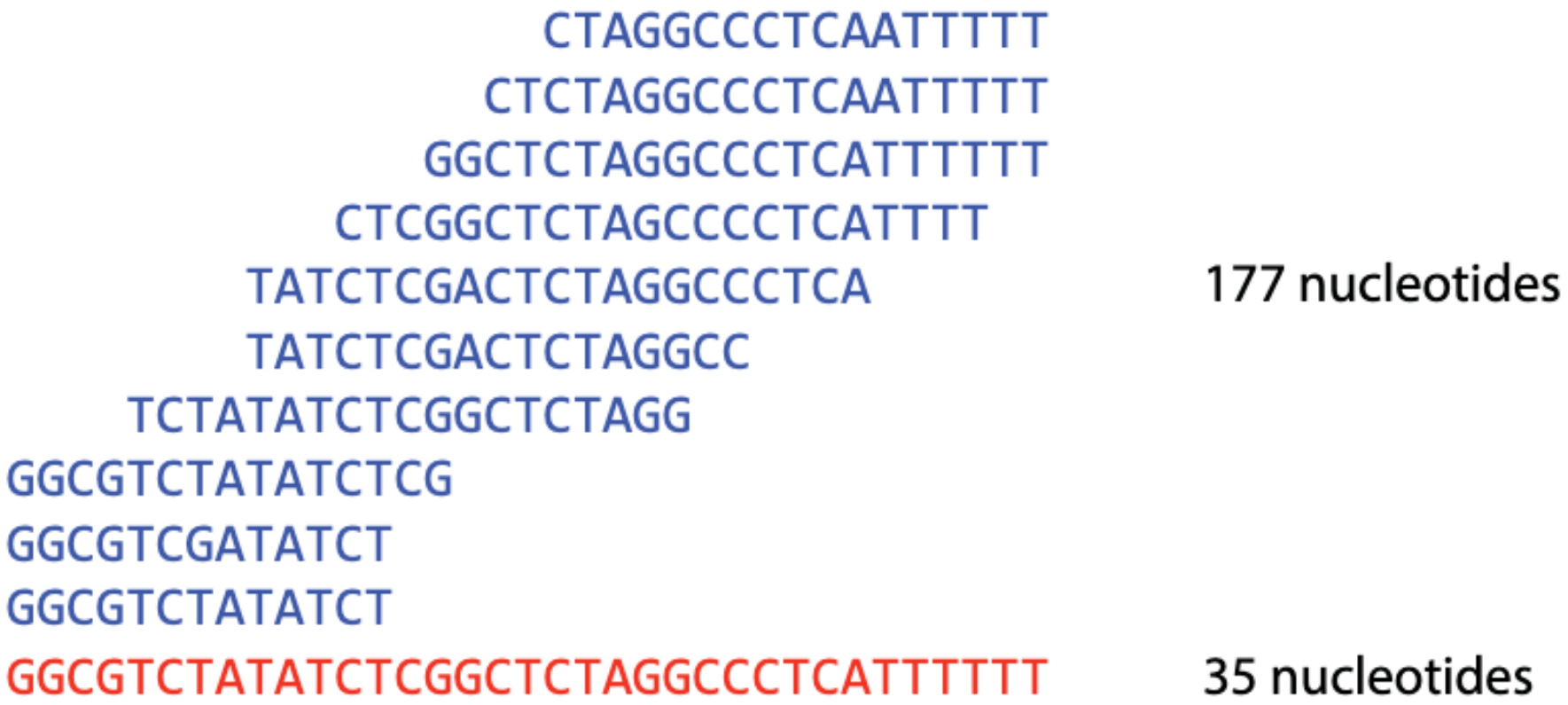
Aiptasia



Using Trinity for transcriptome assembly

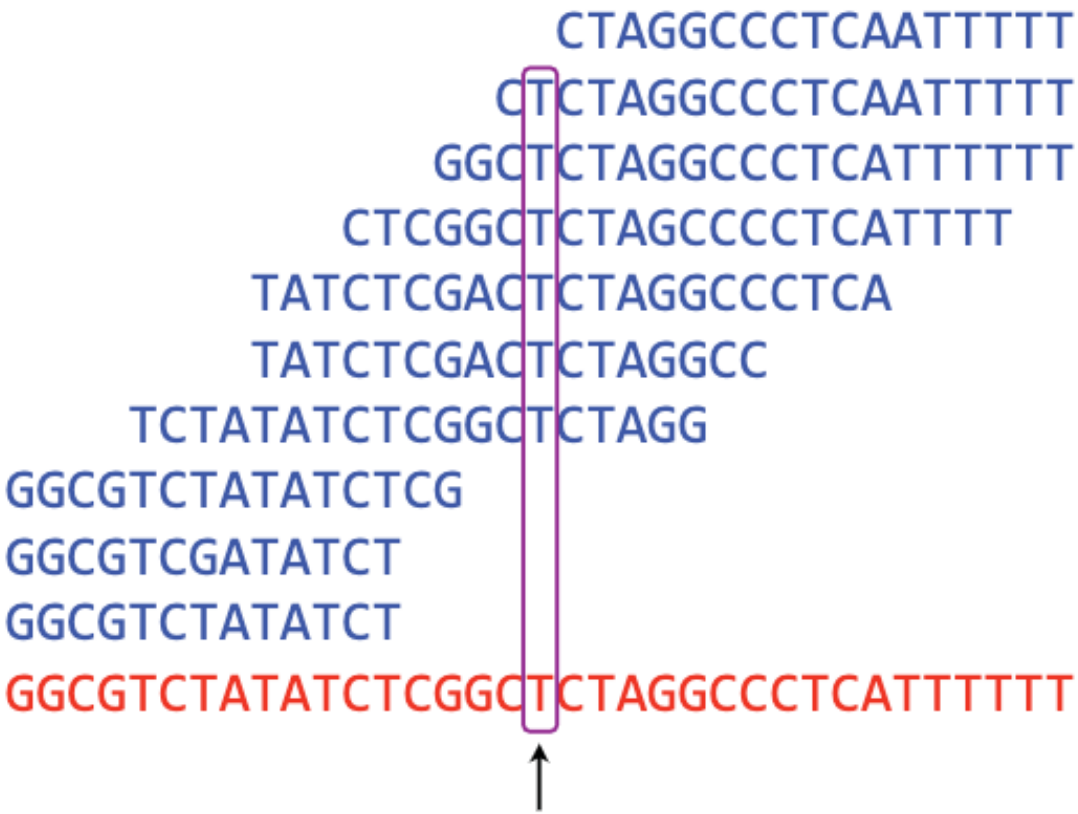


Understanding base coverage
(I.e. average coverage: the average number of reads covering a position in the genome)



What is the approximate coverage here?

Understanding base coverage
(I.e. average coverage: the average number of reads covering a position in the genome)



Coverage at this position = 6

Courtesy of Ben Langmead. Used with permission. http://www.langmead-lab.org/teaching-materials/

What is the approximate coverage here?

Here are standard alignment steps; which steps are different in Assembly?

Pre-alignment QC ("alignment-free" QC)	Base-quality score summaries (FastQC). Base distributions
Pre-processing	Trimming. Fastq interleaving. Demultiplex ("demux")
Alignment	BWA, STAR, HISAT2, GSNAP, bowtie, GMAP
Sorting	Sort bam file by position or read name for easy access.
Indexing	Generate accessory file ("bam index") for fast read access.
Post-alignment QC	% aligned. Intronic/exonic content. Insert size.

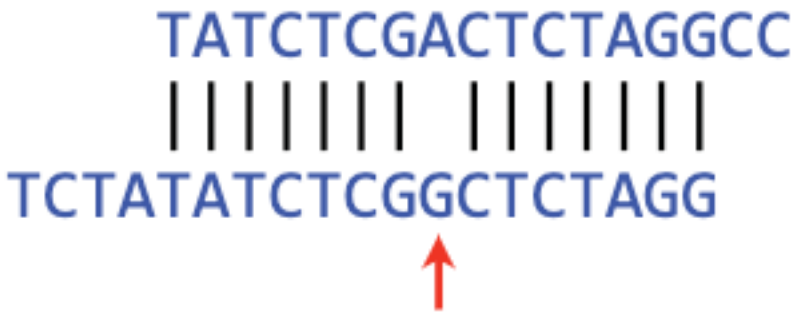
Cheap/Fast

Reasonable/Medium

Expensive/Slow

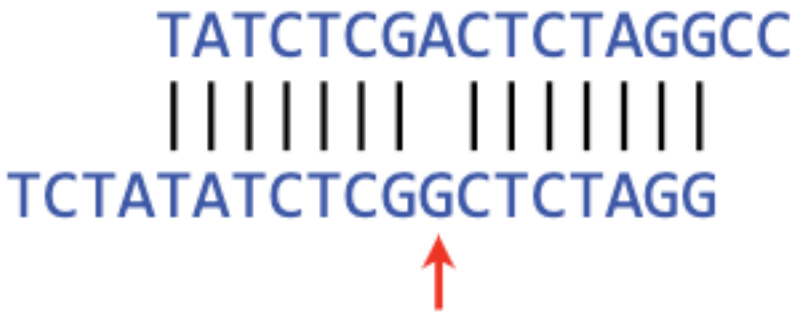
Reference-guided assembly: why assemble a genome if a reference already exists?

Say two reads truly originate from overlapping stretches of the genome. Why might there be differences?



Reference-guided assembly: why assemble a genome if a reference already exists?

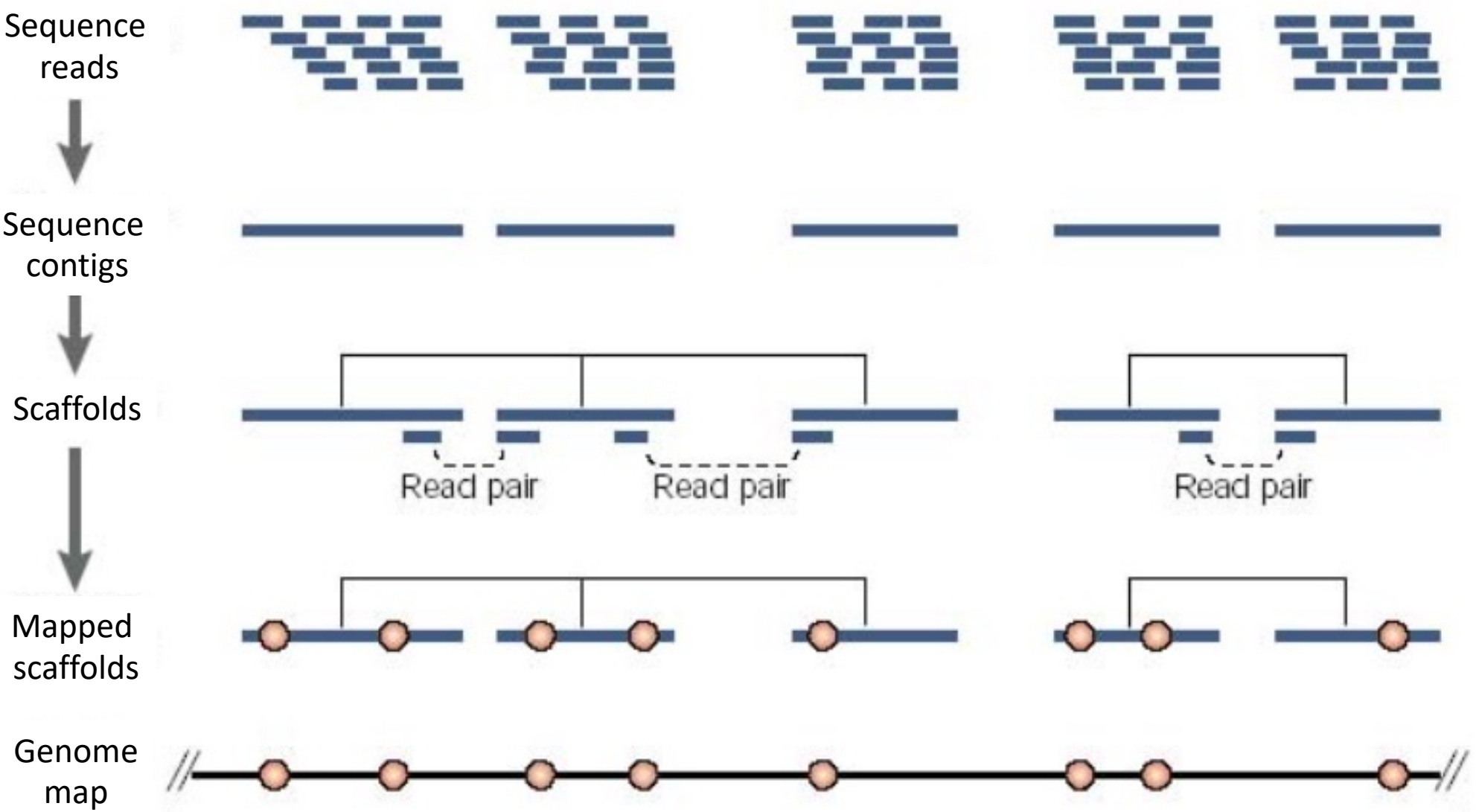
Say two reads truly originate from overlapping stretches of the genome. Why might there be differences?



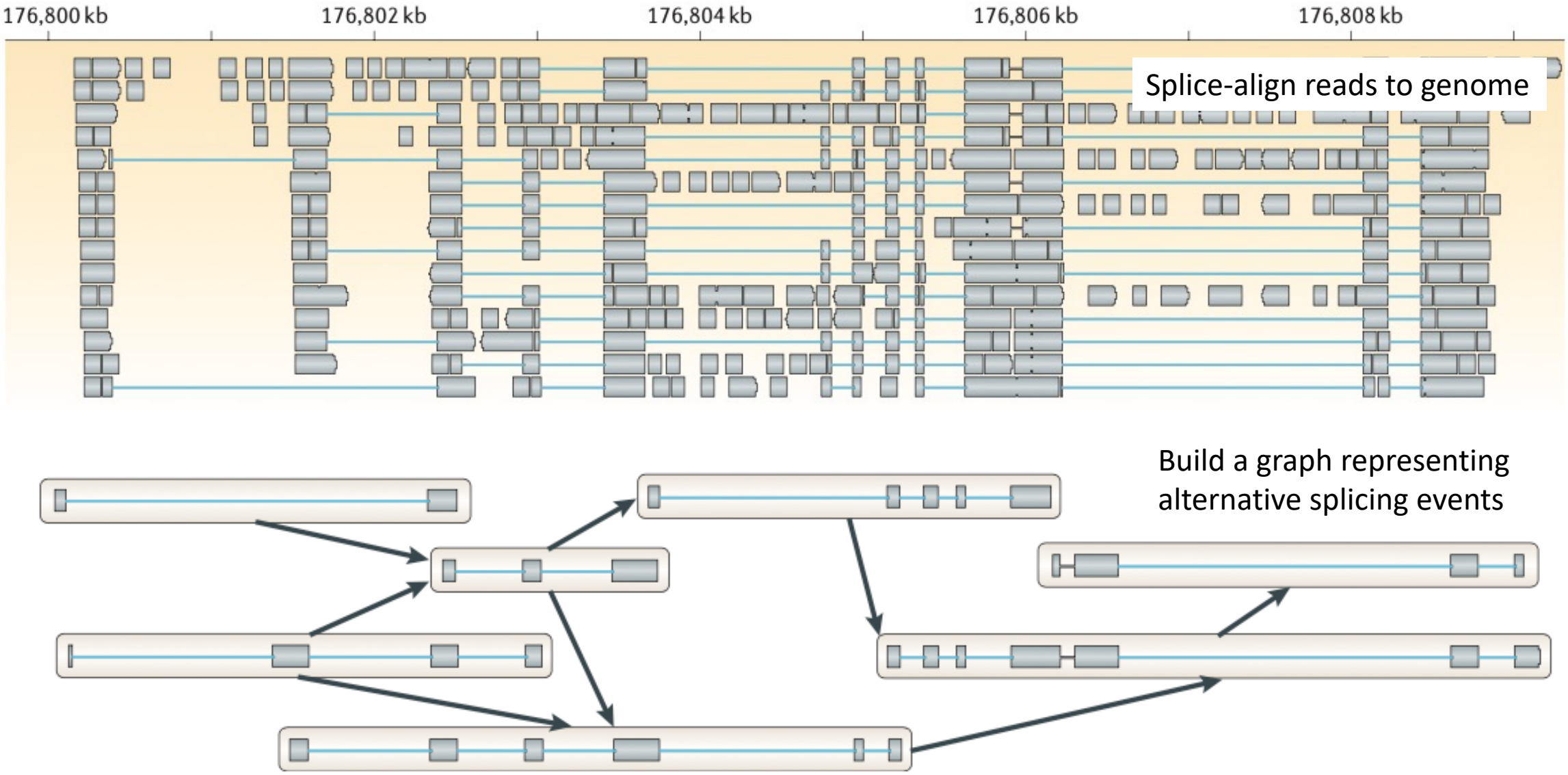
1. Sequencing error
2. Difference between inhereted *copies* of a chromosome

   E.g. humans are diploid; we have two copies of each chromosome, one from mother, one from father. The copies can differ:

de novo whole genome assembly

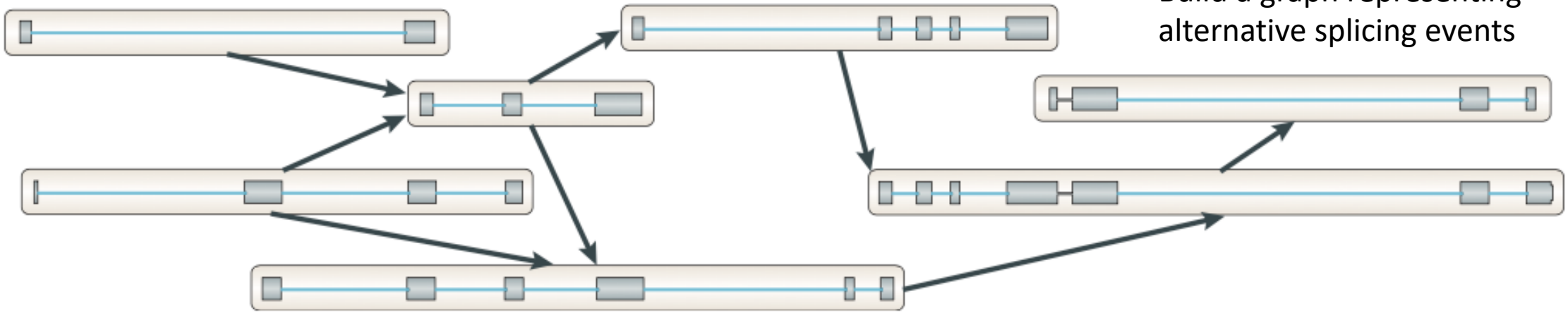


Transcriptome assembly

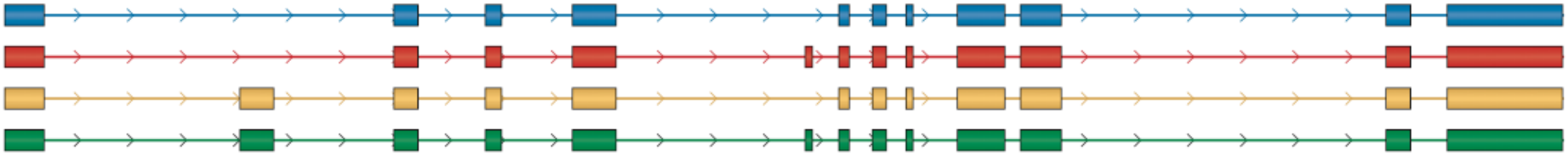


Martin, J., Wang, Z. Next-generation transcriptome assembly. *Nat Rev Genet* **12,** 671–682 (2011). https://doi.org/10.1038/nrg3068

Transcriptome assembly



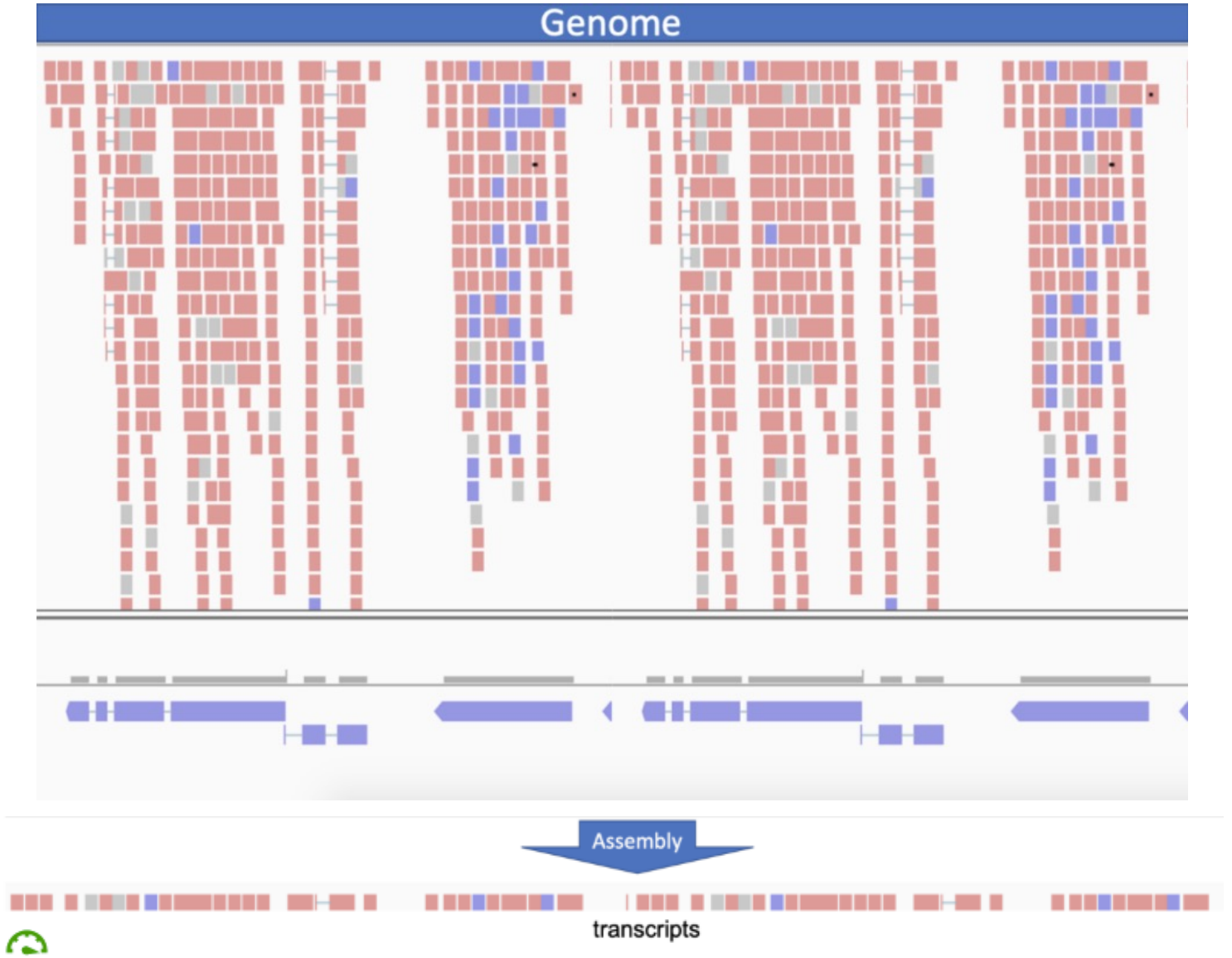
Build a graph representing alternative splicing events



Assembled isoforms

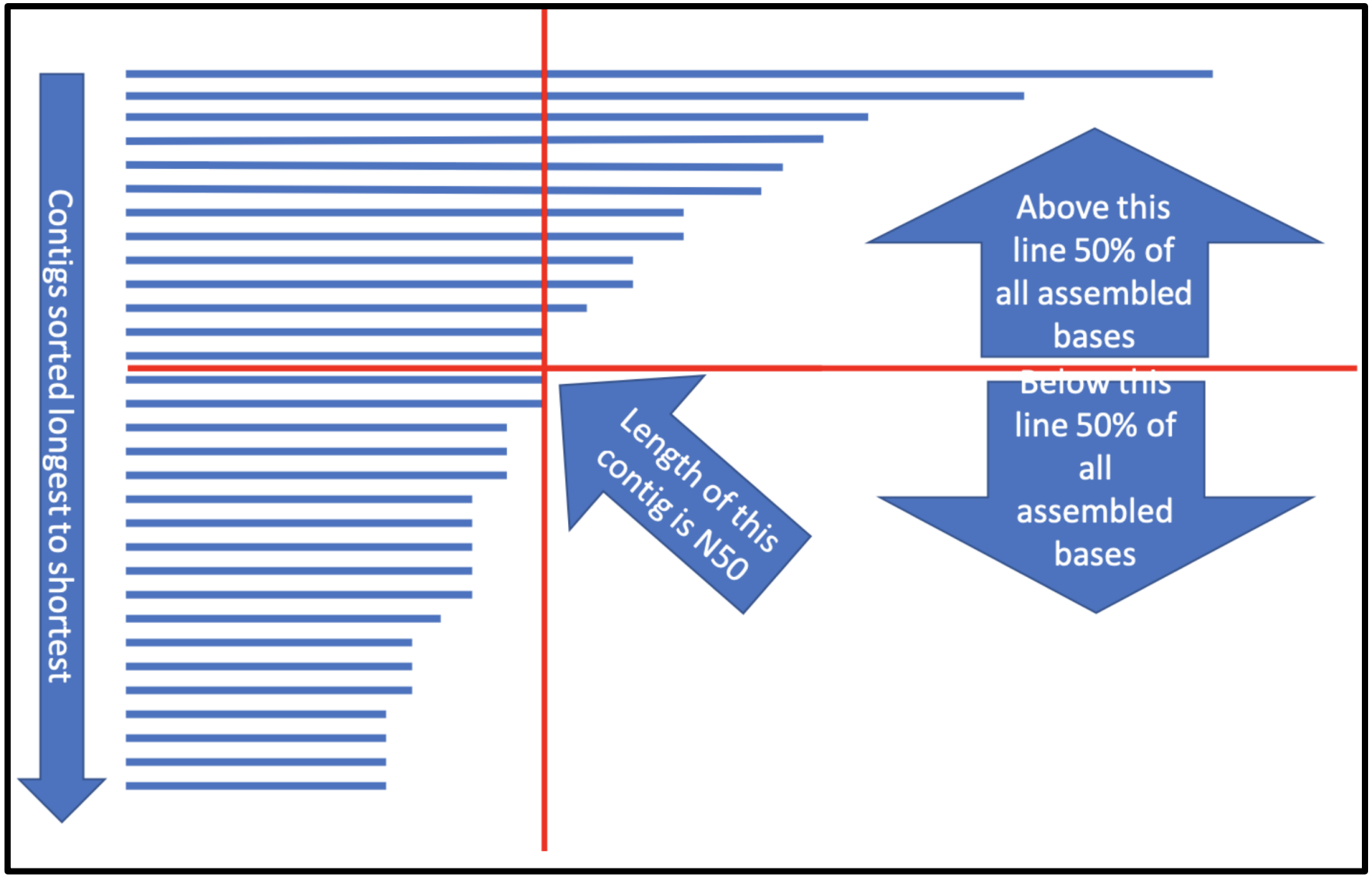
Martin, J., Wang, Z. Next-generation transcriptome assembly. Nat Rev Genet 12, 671–682 (2011). https://doi.org/10.1038/nrg3068

Genome assembly -> transcriptome assembly



How big are the assembled contigs?

N50 indicates the minimal length among the contigs that contain 50 of all assembled bases.



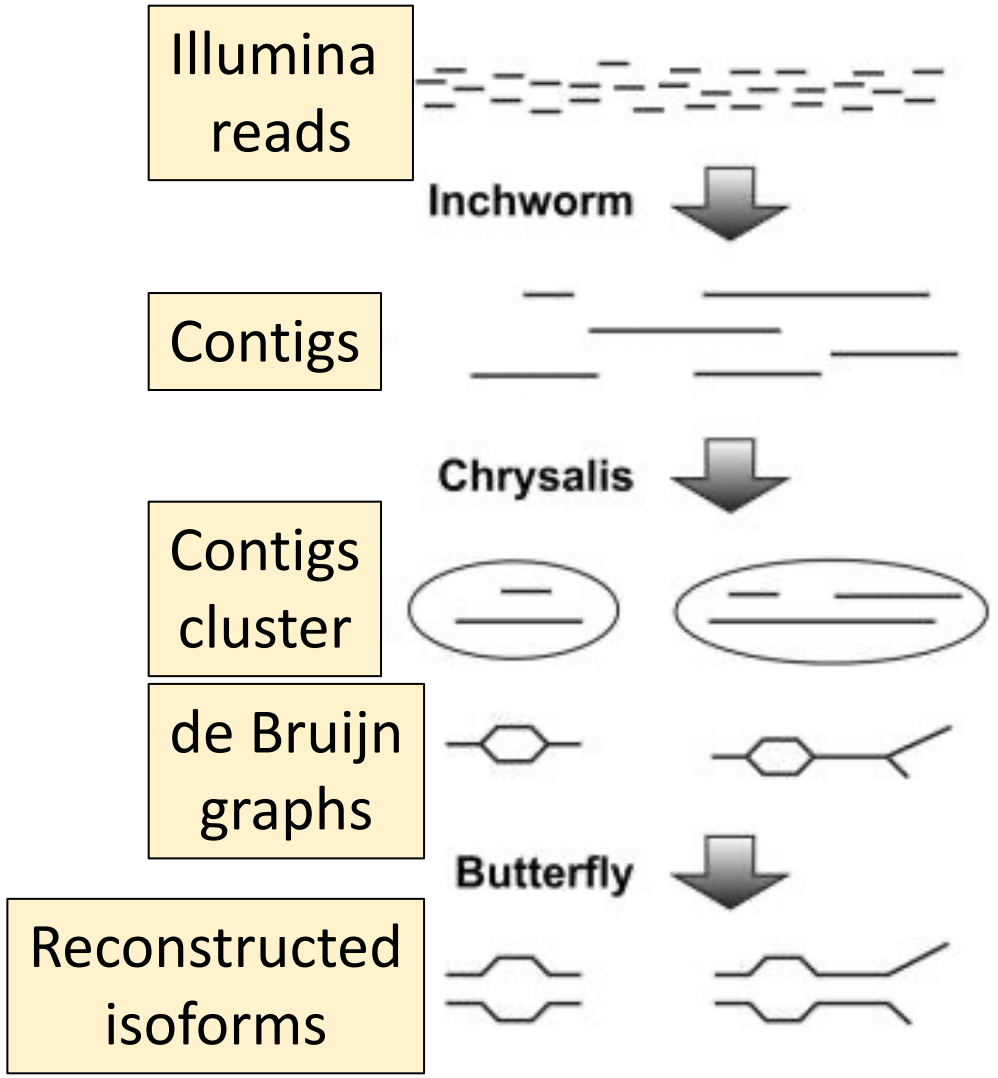
Will N40, N30 be bigger or smaller than N50?

Using Trinity for transcriptome assembly

```
[a.savol@defiance[RNA-Seq]# /usr/local/programs/trinityrnaseq-Trinity-v2.8.4/Trinity
```

```
#######################################################################
#
#  *Note, a typical Trinity command might be:
#
#        Trinity --seqType fq --max_memory 50G --left reads_1.fq  --right reads_2.fq --CPU 6
#
#            (if you have multiple samples, use --samples_file ... see above for details)
#
#    and for Genome-guided Trinity, provide a coordinate-sorted bam:
#
#        Trinity --genome_guided_bam rnaseq_alignments.csorted.bam --max_memory 50G
#                --genome_guided_max_intron 10000 --CPU 6
#
#     see: /usr/local/programs/trinityrnaseq-Trinity-v2.8.4/sample_data/test_Trinity_Assembly/
#          for sample data and 'runMe.sh' for example Trinity execution
#
#     For more details, visit: http://trinityrnaseq.github.io
#
#######################################################################
```

Using Trinity for transcriptome assembly



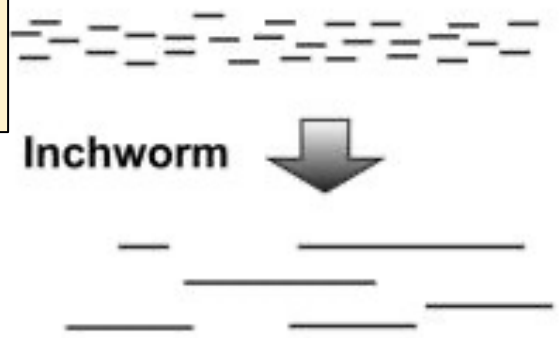
Haas, Brian J et al. "De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis." Nature protocols vol. 8,8 (2013): 1494-512. doi:10.1038/nprot.2013.084

Trinity also uses kmers for faster contig assembly

```
a.savol@defiance[trinity_out_dir]# head -n 10 jellyfish.kmers.fa
>526
AAAAAAAAAAAAAAAAAAAAAAAAA
>1
AGACTGGCCTCTGGTTCAACCCTGC
>1
TGTCTTATTCTAAGCCACACGCAGA
>105
TGTGGGCTGCGCATTCTCTTCCGGA
>1
AAGACCCCGTGTGTTGATGAAGCAG
```

Using Trinity for transcriptome assembly

Illumina reads



Contigs

```
[a.savol@defiance[backup.trinity_out_dir]# vim inchworm.K25.L25.fa
127 >a3;3 total_counts: 76 Seed: 5 K: 25 length: 46
128 AACAGGCTGCCGGGCTAAGGCCCCTAGGATTTAGCCCACAGGAGCC
129 >a4;2 total_counts: 938 Seed: 2 K: 25 length: 562
130 TGGTTTTCAGTCCTCTCCTGAGATAATGCCCTGGGTGTGATGACTTTTAGTAGATGCAGA
131 GAGATTGCTCTGCCACCCTCCAGGCTCCACAGCGTGGGAACTCTTGCATGTTTGTTCACT
132 AACACAGCCAGTCCCAGCACCCTTAGACAGTGAGTCACTGGAGAGGCGTGAGTAGAGAGG
133 CTGAGACTGGCTGCCTACAATCTAATCTACCTCGATAACTTAGGACTAACTCTTCCTAGA
134 ATGTTTGGTTCTTAAGATAAGGCAGCTCATATAAGTAAGCTATCACACAGGCCCAGCTCT
135 GCTCAGAATGGGGGAAGGTCAGCACTCTGACCCCTTCTGCTCTCGAGGGCCTGTCTTTCC
136 TCCTCGGCTCTTACTCCTTTGAGACTGCACCCCCTAACTTCTGTCTCATCTTCATTTATA
137 GGCCTGGACTCTTTGAAGTCTGATATTGATTACAGTGGGATGTCAATTGTCTGCTTGCGA
138 CAGGGCACCTGGAACCTATTCAGATAAGCAAGAGGTTTCCACTGTGAACCCAACGAGCTG
139 CGAGCTGTGACCTTGCTCTCAG
140 >a5;2 total_counts: 79 Seed: 2 K: 25 length: 61
141 TCTTGGGCCCCCAGACACCCCCCTGCCCCCTACACGCCCTGACAGTGTGTCATCCAGCTA
142 C
143 >a6;2 total_counts: 156 Seed: 3 K: 25 length: 89
144 GCCCAATGTTGCCTTGGGAACATTATCTAACTTTATTTGAGCAGTTTTAGGACAATGTCT
145 ACAACTTAAGAGGTAAAGCACCCTTCTAC
146 >a7;2 total_counts: 38 Seed: 2 K: 25 length: 48
147 GGAGGGGCCGCCAAGAGCCCATCCCAAAAGAGGGGTGTGCACCCTTCT
```

Haas, Brian J et al. “De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis.” Nature protocols vol. 8,8 (2013): 1494-512. doi:10.1038/nprot.2013.084

Assessing Trinity accuracy

```
a.savol@defiance[test_Trinity_Assembly]# ${TRINITY_HOME}/util/misc/illustrate_ref_comparison.pl
__indiv_ex_sample_derived/refSeqs.fa trinity_out_dir/Trinity.fasta 90 > assembly_comparison.txt
```

```
Loaded 197879 letters in 84 sequences
Searched 29910 bases in 11 sequences
              1-2919  [============================================================]  mm9chr5-NM_172722;mm9chr5-231713
              1-2919+ ------------------------------------------------------------>  TRINITY_DN6_c0_g1_i3 24-2942:5399 (54.07% aln, 100.00% ID)
              1-1728+ ----------------------------------->                           TRINITY_DN6_c0_g1_i2 24-1751:5528 (31.26% aln, 100.00% ID)
           1729-2919+                                     ------------------------>  TRINITY_DN6_c0_g1_i2 1881-3071:5528 (21.54% aln, 100.00% ID)
           2649-2919+                                                          ------>  TRINITY_DN6_c0_g1_i1 77-347:2804 (9.66% aln, 100.00% ID)

              1-1536  [============================================================]  mm9chr6-NM_001083315;mm9chr6-64213
              1-1536+ ------------------------------------------------------------>  TRINITY_DN0_c0_g1_i1 567-2102:2364 (64.97% aln, 100.00% ID)
              1-1056+ ---------------------------------------->                     TRINITY_DN0_c0_g1_i5 567-1622:2213 (47.72% aln, 100.00% ID)
               1-511+ ------------------->                                          TRINITY_DN0_c0_g1_i2 567-1077:2433 (21.00% aln, 100.00% ID)
             23-1536+ ------------------------------------------------------------>  TRINITY_DN0_c0_g1_i3 425-1938:2200 (68.82% aln, 100.00% ID)
              23-511+ ------------------->                                          TRINITY_DN0_c0_g1_i4 425-913:2269 (21.55% aln, 100.00% ID)
            512-1536+                     ------------------------------------->  TRINITY_DN0_c0_g1_i4 983-2007:2269 (45.17% aln, 100.00% ID)
            512-1536+                     ------------------------------------->  TRINITY_DN0_c0_g1_i2 1147-2171:2433 (42.13% aln, 100.00% ID)
           1208-1536+                                              ------------->  TRINITY_DN0_c0_g1_i5 1623-1951:2213 (14.87% aln, 100.00% ID)

              1-1605  [============================================================]  mm9chr6-NM_022332;mm9chr6-64213
              1-1605+ ------------------------------------------------------------>  TRINITY_DN0_c0_g1_i2 567-2171:2433 (65.97% aln, 100.00% ID)
               1-512+ ------------------->                                          TRINITY_DN0_c0_g1_i1 567-1078:2364 (21.66% aln, 100.00% ID)
               1-512+ ------------------->                                          TRINITY_DN0_c0_g1_i5 567-1078:2213 (23.14% aln, 100.00% ID)
             23-1605+ ------------------------------------------------------------>  TRINITY_DN0_c0_g1_i4 425-2007:2269 (69.77% aln, 100.00% ID)
              23-512+ ------------------->                                          TRINITY_DN0_c0_g1_i3 425-914:2200 (22.27% aln, 100.00% ID)
            582-1605+                       ------------------------------------->  TRINITY_DN0_c0_g1_i3 915-1938:2200 (46.55% aln, 100.00% ID)
            582-1605+                       ------------------------------------->  TRINITY_DN0_c0_g1_i1 1079-2102:2364 (43.32% aln, 100.00% ID)
            582-1125+                       ---------------------->                 TRINITY_DN0_c0_g1_i5 1079-1622:2213 (24.58% aln, 100.00% ID)
           1277-1605+                                               ------------->  TRINITY_DN0_c0_g1_i5 1623-1951:2213 (14.87% aln, 100.00% ID)

              1-1200  [============================================================]  mm9chr9-NM_026942;mm9chr9-69106
               1-240+ ------------>                                                 TRINITY_DN3_c0_g1_i4 66-305:769 (31.21% aln, 100.00% ID)
            133-1200+       ------------------------------------------------------>  TRINITY_DN3_c0_g1_i1 871-1938:2619 (40.78% aln, 100.00% ID)
             133-240+       ----->                                                  TRINITY_DN3_c0_g1_i5 871-978:2899 (3.73% aln, 100.00% ID)
             133-240+       ----->                                                  TRINITY_DN3_c0_g1_i3 871-978:1442 (7.49% aln, 100.00% ID)
            239-1200+            ------------------------------------------------>  TRINITY_DN3_c0_g1_i2 664-1625:2306 (41.72% aln, 100.00% ID)
            241-1200+            ------------------------------------------------>  TRINITY_DN3_c0_g1_i5 1259-2218:2899 (33.11% aln, 100.00% ID)

              1-1104  [============================================================]  mm9chr4-NM_010598;mm9chr4-16498
              1-1104+ ------------------------------------------------------------>  TRINITY_DN2_c0_g1_i3 324-1427:3739 (29.53% aln, 100.00% ID)
                1-77+ --->                                                          TRINITY_DN2_c0_g1_i2 324-400:3697 (2.08% aln, 100.00% ID)
                1-77+ --->                                                          TRINITY_DN2_c0_g1_i1 175-251:3548 (2.17% aln, 100.00% ID)
            120-1104+       ------------------------------------------------------>  TRINITY_DN2_c0_g1_i2 401-1385:3697 (26.64% aln, 100.00% ID)
            120-1104+       ------------------------------------------------------>  TRINITY_DN2_c0_g1_i1 252-1236:3548 (27.76% aln, 100.00% ID)
             501-540+                           -->                                 TRINITY_DN32_c0_g1_i1 217-256:256 (15.62% aln, 100.00% ID)
```

Assessing Trinity accuracy

```
1 Loaded 197879 letters in 84 sequences
2 Searched 29910 bases in 11 sequences
3               1-2919  [============================================================]  mm9chr5-NM_172722;mm9chr5-231713
4               1-2919+ ------------------------------------------------------------>  TRINITY_DN6_c0_g1_i3 24-2942:5399 (54.07% aln, 100.00% ID)
5               1-1728+ ----------------------------------->                           TRINITY_DN6_c0_g1_i2 24-1751:5528 (31.26% aln, 100.00% ID)
6            1729-2919+                                      ------------------------>  TRINITY_DN6_c0_g1_i2 1881-3071:5528 (21.54% aln, 100.00% ID)
7            2649-2919+                                                          ------>  TRINITY_DN6_c0_g1_i1 77-347:2804 (9.66% aln, 100.00% ID)
```

Contig TRINITY_DN6_c0_g1_i3
(length=5399)

```
7 >TRINITY_DN6_c0_g1_i3 len=5399 path=[0:0-1751 3:1752-2670 4:2671-5398]
8 ACTGAGGTTCTGCGCTTGTCATCATGGCGACGCGGGGCCATGTGCAGGACCCTAACGACCGGCGCCTCCGGCCCATTTACGATTATCTTGATAATGGAAATAATAAAATGGCAATTCAGCAAGCTGAT
  AAACTGTTGAAGAAACACAAAGATCTTCACTGTGCAAAGGTTCTCAAGGCCATTGGTTTACAGAGAACTGGCAAGCAGGAAGAAGCTTTCACCTTGGCTCAGGAGGTGGCTGCCCTTGAGCCCACAGA
  TGACAACTCCCTGCAGGCCCTGACGATTCTCTACCGAGAGATGCACCGGCCGGAGTTAGTCACAAAGCTCTATGAGGCGGCTGTGAAGAAAGTTCCCAACAGCGAGGAGTACCACTCTCACCTCTTCA
  TGGCCTATGCCAGAGTGGGCGAGTACAAGAAGATGCAGCAGGCTGGCATGGCTCTGTATAAGATTGTCCCCAAAAACCCTTACTACTTTTGGTCAGTGATGAGCTTGATCATGCAATCTATATCCGCA
  CGGGATGAGAACCTCTCCAAAACCATGTTTCTGCCCCTGGCGGAGAGAATGGTAGAGAAGATGGTGAAGGAGGACAAGATAGAGGCTGAAGCTGAGGTTGAACTTTATTACATGATTCTGGAGCGCTT
  GGGAAAGTACCAGGAAGCCCTGGATGTCATTAGAGGAAAATTAGGAGAGAAGTTGACAAGTGAAATTCAGAGTCGGGAAAACAAATGCATGGCCATGTACAAGAAGCTGAGCAAGTGGCCGGAGTGCA
  ACGCCCTTTCCCGACGCCTCCTTCTGAAAAACTCAGATGACTGGCAGTTCTATCTGACTTATTTCGATTCTGTCTTTCGACTGATTGAGGAAGCCTGGACGCCGCCTGCTGAGGGGGAACACTCCTTA
  GAAGGGGAAGTACACTGCTCTGCAGAAGATGCTGTGAAGTTTATAGAGGACAGGATAACAGAGGCATCACAAAGCTCTCGACATGTCCGAGGCCCACACCTAGCTAAGCTGGAGCTCATCAGGCGCTT
  GCGGAGCCAAGGCTGTAATGACGAATACAAGCTGGGTGACCCTGAAGAGTTAATGTTCCAGTATTTCAAGAAGTTCGGGGATAAGCCATGCTGCTTCACAGACCTGAAGGTGTTCGTTGACCTCTTGC
  CTGCCGCCCAGTGCACACAATTCATTAACCAGCTCCTTGGAGTTGTTCCATTGTCGACGCCGACAGAGGATAAGCTTGCCCTGCCGGCCGACATCCGAGGCCTGCAGCAGCATCTGTGTGTTGTGCAG
  CTGACACGGCTCCTTGGCTTGTACCACAGCATGGATAAGAGTCAGAAGCTGGATGTGGTCAAAGAACTGATGTTAAGGTACCAGCACGGACTGGAGTTTGGGCGCTCTTGTTTGAAGACAGAGTTGCA
  GTTCTCAGACTATTACTGCCTGCTCGCTGTTCATGTGCTCATTGATGTATGGAGAGAAGCAGGTGAGGAGACTGCGGTCTGGCAGGCCCTGACTCTGCTGGAAGAAGGCTTAACCCACAGCCCATCCA
  ATGCTCAGTTCAAATTGCTACTTGTTCGGATCTACTGCGTGCTGGGTGCATTTGAACCAGTGGTGGATCTCTACTCCAGTCTCGATGCTAAGCACATCCAGCACGACACCATTGGCTATCTTCTGACC
  CGATATGCTGCATCCCTGGGTCAGTATGCTGCTGCCTCCCAGTCCTGTAACTTTGCACTCAGGTTTTTCCACTCCAACCAGAAAGATACCTCAGAATACATTATTCAAGCTTACAAATATGGTGCATT
  TGAGAAGATCCCAGAGTTTATCGCTTTTAGGAACAGGCTGAATAATTCTCTCCATTTTGCACAAGTACGCACTGAGCGGATGCTGTTAGACCTTCTACTTGAAGCAAACATATCAATTAGCTTGGCAG
  AAAGTATCAAATCAATGAATCTGAGGCCAGAGGAAGATGATGTTCCATGGGAAGATTTGCGAGACAACAGAGACTTGGATGTTTTCTTCAGTTGGGATCCAAAAGACAGGAATGTTTCTGAAGAGCAT
  AAGAAACTCTCCTTGGAGGAAGAAACCATGTGGTTAAGAATTCGCTCCTTAACGTTGAGGCTGATCAGTGGACTTCCAAGTCTTACCCACCCTGTGGAGCCCAAGAACTCAGAGAAGATGTCAGAGAA
  CGGTGTGTCCTCCCGGATCGACATCCTTCGTTTGCTCCTCCAACAGCTAGAGGTGGCCGTGGAGACGGGGAAGCGGTTTATTGAGAAGGAAATTCAGTACCCCTTCCTTGGGCCTGTTCCCACCAGGA
  TGGGCAGGTTCTTCAGTTCTGGCTGCTGTCAGTGCCAAGTCCAGTCCTTCCATCTGGTCAGCGACATGTATGAGCTTGACACCAGTGGCTTAGAGGGTACAGTGGATATACAGGAACGAATAGAAAAT
  AGTCTTGCATCTTTACTAGAACTATTGAAAGGTGTCTTCAGCACATGTAAAGGGGACCTCTTAGAAGTTACAGATGGCAATGTGAAGACCCAGCCTGCAGTCCTAGAGAATCTGGTCTTCTTCGTTGA
  GACTATTTCTGTCATCCTTTGGGTGTCCAGTTACTGTGAAAGTGTCTTGCGACCATACAAATTAAATATCCAGAAAAAGAAAAAGAAGAAAAAAGAAACGAGCATCATCATGCCTCCAATCTTCACCA
  GTTTCCAAGACTACGTTACTGGACTTCAGACTGTAATTTCCAATGCTGTGGATCATATTAAAGGACTTGAAGCACATCTGATTGCACTTAGACTTGAGGAGCTCACGTTAGAGGAAACTTCTATCTCT
  ACGGAAGAAAGAAAATTTCAAAGACTGTGCAGGGGAAAGTCCAGAGCAGTTACCTGCACTCACTTCTGGAGACAGGGGAGCTGCTGAGAAAAAGACTTGAGACTACAAAGAAACTAAAAATTTAAGG
  AATTCTGAACCATAAACACCGAAGGTTCCACAGCAGAAGATGACACACCACCTTCCCCAGCAAAAACAACATCTGGCTGGCATTGTTCTAATCCAGAACTTCCTCACAAATGCATGAAGGACTTTGAT
  CGTGAATTAATTTTTGAGGTATTAAATATATACAACAAATTAAATTATTGTATATAAACCAGAAGCAGGACCCTCATCTGGGTTTACCACTCTTTATACCTGTCCATACATACAGGACAGCAACAAGA
  GCACCCCACCTCACTCCCAGCTCCAGGAATGGCTGGGGTGGACATTTCTATACAAGCAGCAATAGAGAATAAGTATAGCATCGACCTCAGAGCAGGTGGGTTGCTTTTTCATCATGCAGTGGGAGTTT
  GGGTTGGGTAGCCCATACTGTAGCCGTATGAGGCAAAAGCAGGTGTAGGAGACCCAGGAGCAGGACCCTTCCCTCAGTGGCAGGTTCCTTGTGCCATCCCACAGAACCGGGATGGACCCACAGAGGCT
  CAAGGGAGATGTGGCTGAGTACGTGCCACCACAGGGCACCACACATGACCACAGGTGCTCAGTTTTGGCTTCACAGCCCCAGCAGCCAAGACATCCCCAGCCCCGCTCCTGGGAGAGCTCAGCCTCAG
  GGCAGTGAGGAAGTCCAAGTTCTGCCCACTTTGAGGAAGAAATCTGAGCACAGCCCACCGTGAGCTTCGTAAACTGCTGAGACTGAGCCCCGCTGGTCTTCTGGGAAAGGTGGAAAAGCTGTTGGAAG
  AGGTTGTACGTGGTGTGTGCAGGCGAGTCTGATATTTGTGTCGTCATTAAGTCACCTTGTGCAGCTGATTGATGAGTCACACACAGATACACATATGCATAAATATGCTTGGAAACAGATTTACTACT
  CTAGTGCTTTTAAGGGGGGCTTTGGGATGGAAGGATTAAATATGGGTTTAACCTAGGTTTTTAAGTTTTAGGTTGAAAAAGGTATGTTTTATAATGGGCCTGATGTTGGTGTCCTGTTGACACTTAAG
  AACTGCATCAGGGTGTATTCATGTTGTTCACCTAAGGTTGGTGGCCTGTTCCCTAAGCCTCCAGCTTAAGAGGCTGGTCTGCGAGAAGATGAAAACTGCTCTTGATTATGTCATGCCGAGCTCCGACA
  GCGCTTTCCGTGACTGTTTTGTGTCTGAATCTCCTCTCTCACTTCTGTTCTTGTTCCCTGTCACTTTATTCCTGAGAGTAACTGTAAGGTGAAAAGATGACTTGGTGAGGACTGGTCTTTCCCATGGA
  CTGCTCCAGCAGCACACCATTGCATTCTTTATCGTAACTGTCTGGGATTCACAAATCAGCATAGTGGTTAACTAATTTACTTGTAAGGAAATGTGTGCAAACTACTTTCTTACTTGTAGGATGTGCTG
  GAGCCTACATCATGGTTGCATCAGAGTATCACGTGCTCCATCCCCAAAAAAAGGGCCTTTTTGAATAAGATGGGTTGAAGACCCTGATTTTGACTTGCCTTGAACACCACCTTGTGGTACTGTTCTGT
  AGTTGTGGCCTGTTTCCCATGGGACTGGGTGGGCGTTTGCTTGCACTGCCCTGGCTCCTGGAATAATGTTTCTTGGGATATTCAAATTGTGCTGTGTTGTTCACGGTCACAGATGCTCTAAATCAAGCT
  CTGTCAGCTGACCATGGTCCGTATGGACTGCATGGGCTGGCTATGGCCCTCAGGCATAGCTCCTTCCTCTAATGGTCTATCCTGTCAGTCTCCTGCTCTAACATGTAGGGCCTCCACTGGTCTTGTTG
  ACAAATAATGTCATGGTGGGGTTACACTTCCTCTGTTACTTGCTGATTTTCTTACTAAATGTAGCATTTCAATCAGATCTAATTCTTACATTTTGAGATAACTGTAAAAAGAGTGAAATTATGAAATG
  GCCAATATCTTTTATCAGTCTATTCTTTTGGAAGCTGTCATGCACTATACATTGTGTACAGTTAAAAGTATATATATATATATTCTTACTGAGTGAACGCCTCCTCTCCCCACGCCTGTATGTCAC
  TAGCATCTAAGGAGAATGCTCAAGGCCCAGTGCTGCTGCTGCTGTGGTTTATATGGGTTTTGTTCTGTTTTGTTTTTGTGTGGTAAATTGATATTTAAAAACAACAAAAACCACGACTACTGTTTACA
  GACTGAAAAAAAACAATCACTGCTTTTTATACTACTGAGATCCTAAGTCAAGACTTTGCAAAGCAGGAATCGGGTTCAAGTTACTTCTTTGCTGTGGACGGATAGTCCTCTGTAGTATCTCCACATGA
  TGGAGAGTGCACAAACCTAGGTGTGCTGCCATCAATTTTGTATATTTTCATAATTTTAATTGTTCGAAATTGCATTATATTTTGCAATCACCACATTCAATCTGTATATGTCTTTCATTTCAACTTTT
  TCAATACAAAAAGGGATTCAATT
```

Using Trinity for transcriptome assembly

```
a.savol@defiance[test_Trinity_Assembly]# ./runMe.sh
```

Group coding challenge

```
1 #!/usr/bin/env python
2
3 # define a function is_palindrome that takes a string as argument and returns a boolean
4 # that is True if the string is a palindrome, and False if it is not
5
6 def is_palindrome():
7     pass
```

```
In [2]: is_palindrome('deified')
Out[2]: True

In [3]: is_palindrome('CTCC')
Out[3]: False

In [4]: is_palindrome('C')
Out[4]: True
```

Group coding challenge: 3 possible solutions

```
1 def is_palindrome(s):
2     return(s == s[::-1])  # via slicing (simplest)
```

Group coding challenge: 3 possible solutions

```
1 def is_palindrome(s):
2     return(s == s[::-1])  # via slicing (simplest)
```

```
4 def is_palindrome(s):
5     all_matches = True
6     for i,c in enumerate(s,1):
7         if c != s[-i]:  # walk backward through the string with negative indices
8             all_matches = False
9
10    return(all_matches)
```

Group coding challenge: 3 possible solutions

```
1 def is_palindrome(s):
2     return(s == s[::-1])  # via slicing (simplest)
```

```
4 def is_palindrome(s):
5     all_matches = True
6     for i,c in enumerate(s,1):
7         if c != s[-i]:  # walk backward through the string with negative indices
8             all_matches = False
9
10    return(all_matches)
```

```
12 def is_palindrome(s):
13    pal = True
14    for fwd,rev in zip(s,s[::-1]):
15        # will evaluate pal irreversibly to False if mismatch
16        pal = pal and fwd == rev
17        print('%s==%s: pal=%r' % (fwd,rev,pal))
18
19    return(pal)
```

Using symlinks

A symbolic link (aka *symlink* or *soft link*) is a special type of file that points to another file or directory.

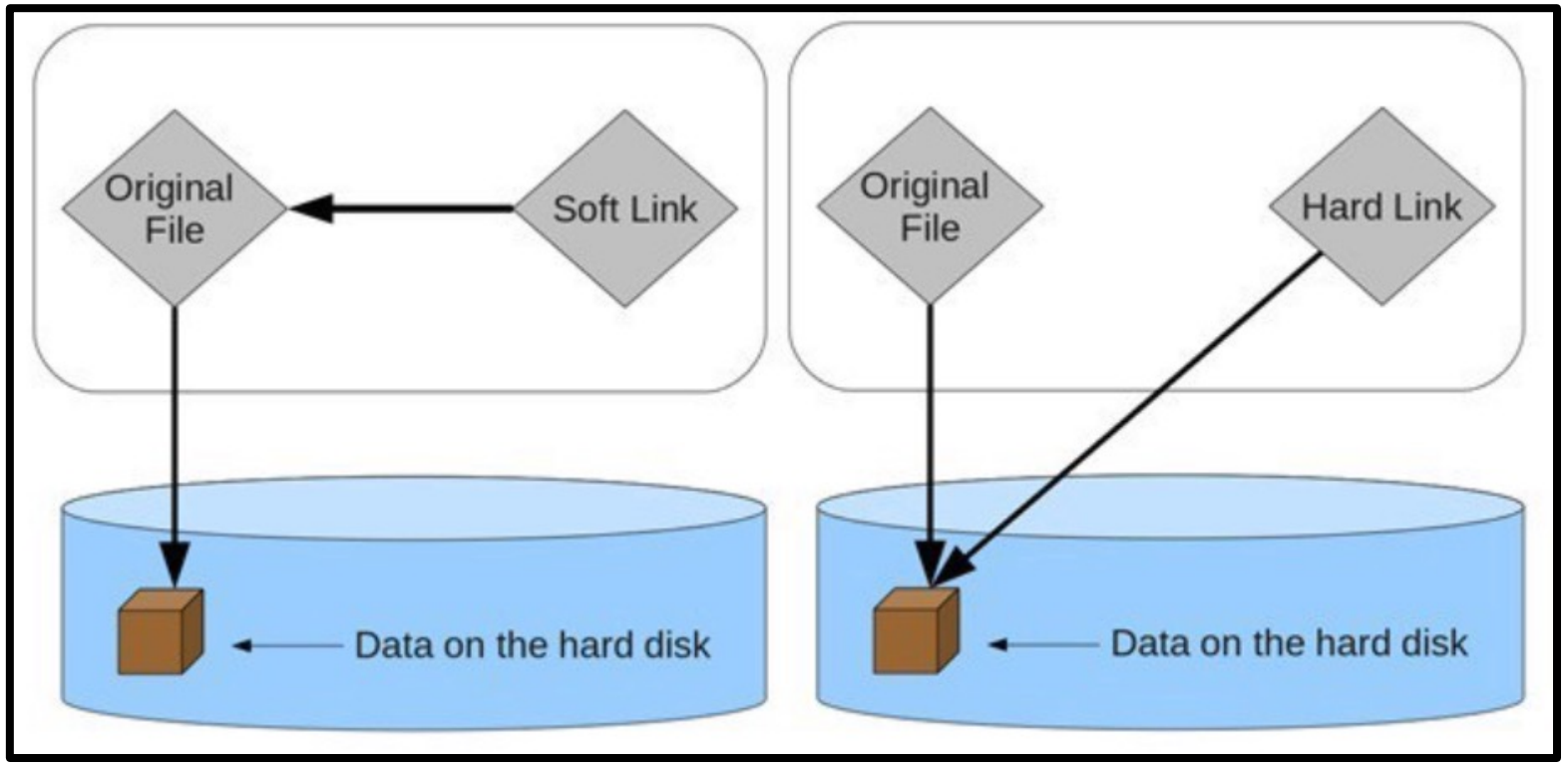
Symlinks enable multiple file names to point to the same file

Symlinks help avoid unnecessary file copying.

```
a.savol@defiance[RNA-Seq]# ln -s ../../module-06-a-savol/RNA-Seq/Paired Paired
a.savol@defiance[RNA-Seq]# ll
total 0
drwxr-xr-x. 2 a.savol BioInfo 74 Mar 20 17:36 .
drwxr-xr-x. 4 a.savol BioInfo 84 Mar 20 17:18 ..
lrwxrwxrwx. 1 a.savol BioInfo 34 Mar 20 17:20 bam -> ../../module06-a-savol/RNA-Seq/bam
lrwxrwxrwx. 1 a.savol BioInfo 38 Mar 20 17:36 Paired -> ../../module-06-a-savol/RNA-Seq/Paired
lrwxrwxrwx. 1 a.savol BioInfo 34 Mar 20 17:20 sam -> ../../module06-a-savol/RNA-Seq/sam
lrwxrwxrwx. 1 a.savol BioInfo 39 Mar 20 17:20 Unpaired -> ../../module06-a-savol/RNA-Seq/Unpaired
```

https://linuxize.com/post/how-to-create-symbolic-links-in-linux-using-the-ln-command/

	Soft link (symbolic link)	Hard link
Command	ln –s <TARGET> <LINK_NAME>	ln <TARGET> <LINK_NAME>
Use case	Access a file from multiple locations (or even multiple computers)	Access same data and be robust to name changes
Consequences of renamed/moved/deleted Original file	Broken link	Nothing
Consequences of deleted link	None	None



https://medium.com/@eitanmayer/hard-vs-symbolic-links-f0584a5d1db5

Merging bam files

```
1 #!/usr/bin/env bash
2 # mergeAll.sh
3 # Usage: bash scripts/mergeAll.sh 1>results/logs/mergeAll.log 2>results/logs/mergeAll.err
4
5 # List the files with ls and redirect (>) the output to bamIn.txt
6 ls data/bam/Aip*.sorted.bam > data/bam/bamIn.txt
7
8 # Merge the files using bamIn.txt with the -b option
9 samtools merge -b data/bam/bamIn.txt data/bam/AipAll.bam
```

Running Transcriptome Assembly with Trinity

```
#!/usr/bin/env bash
# runTrinity.sh
# Usage: bash scripts/runTrinity.sh 1>results/logs/trinity_guided.log 2>results/logs/trinity_guided.err

Trinity \
--genome_guided_bam data/bam/AipAll.bam \
--genome_guided_max_intron 10000 \
--max_memory 10G --CPU 4 \
--output results/trinity_guided
```

Creating a pipeline

```
#!/usr/bin/bash
#SBATCH --partition=short               # choose from debug, express, or short
#SBATCH --job-name=trinity
#SBATCH --time=20:00:00                 # the code pieces should run in far less than 1 hour
#SBATCH -N 1                            # nodes requested
#SBATCH -n 4                            # task per node requested
#SBATCH --mem=10Gb
#SBATCH --exclusive
#SBATCH --output="batch-%x-%j.output"   # where to direct standard output; will be batch-jobname-jobID.output
#SBATCH --mail-type=ALL
#SBATCH --mail-user=<your user ID>@northeastern.edu # Update to your user name!

# Usage: sbatch sbatch_transcriptome.sh
# Assumes input data is in /home/$USER/AiptasiaRNASeq/data/

echo "Starting our analysis $(date)"

echo "Loading our BINF6308 Anaconda environment, which includes Trinity."
module load anaconda3/2021.11
source activate BINF-12-2021
echo "Loading samtools."
module load samtools/1.10

echo "Make directory for data files"
mkdir -p data/

# part of a bigger sbatch script (e.g., #sbatch lines above)
echo "Moving trimmed FASTQ data to the working directory"
cp -r /home/$USER/AiptasiaRNASeq/data/trimmed data/trimmed
echo "Moving alignment files to working directory"
cp -r /home/$USER/AiptasiaRNASeq/data/bam data/bam
cp -r /home/$USER/AiptasiaRNASeq/data/sam data/sam

echo "Make directory for log files"
mkdir -p results/logs/
```

Creating a pipeline

```
37 echo "Starting Guided Assembly $(date)"
38 echo "Merge all BAM alignment files $(date)"
39 bash scripts/mergeAll.sh 1>results/logs/$SLURM_JOB_NAME-$SLURM_JOB_ID-mergeAll.log 2>results/logs/$SLURM_JOB_NAME-$SLURM_JOB_ID-mergeAll
   .err
40
41 echo "Assemble the Guided Transcriptome $(date)"
42 bash scripts/runTrinity.sh 1>results/logs/$SLURM_JOB_NAME-$SLURM_JOB_ID-runTrinity.log 2>results/logs/$SLURM_JOB_NAME-$SLURM_JOB_ID-runT
   rinity.err
43
44 echo "Analyze the Guided Transcriptome $(date)"
45 bash scripts/analyzeTrinity.sh 1>results/$SLURM_JOB_NAME-$SLURM_JOB_ID-trinity_guided_stats.txt 2>results/logs/$SLURM_JOB_NAME-$SLURM_JO
   B_ID-analyzeTrinity.err
46
47 echo "Guided Assembly complete $(date)"
48
49
50 echo "Starting De Novo Assembly $(date)"
51 echo "Assemble the De Novo Transcriptome $(date)"
52 bash scripts/trinityDeNovo.sh 1>results/logs/$SLURM_JOB_NAME-$SLURM_JOB_ID-trinityDeNovo.log 2>results/logs/$SLURM_JOB_NAME-$SLURM_JOB_I
   D-trinityDeNovo.err
53
54 echo "Analyze the De Novo Transcriptome $(date)"
55 bash scripts/analyzeTrinityDeNovo.sh 1>results/$SLURM_JOB_NAME-$SLURM_JOB_ID-trinity_de_novo_stats.txt 2>results/logs/$SLURM_JOB_NAME-$S
   LURM_JOB_ID-analyzeTrinityDeNovo.err
56
57 echo "De Novo Assembly complete $(date)"
58
59 echo "Moving key files back to /home"
60 cp -r results/trinity_guided /home/$USER/AiptasiaRNASeq/data/trinity_guided
61 cp -r results/trinity_de_novo /home/$USER/AiptasiaRNASeq/data/trinity_de_novo
62 cp results/trinity*stats.txt /home/$USER/AiptasiaRNASeq/data/
63
64 echo "Assemblies complete $(date)"
```

Assessing Trinity output

```
##
##
################################
## ## Counts of transcripts, etc.
################################
## Total trinity 'genes':   26028
## Total trinity transcripts:   27595
## Percent GC: 38.49
##
########################################
## Stats based on ALL transcript contigs:
########################################
##
##  Contig N10: 1917
##  Contig N20: 1292
##  Contig N30: 952
##  Contig N40: 727
##  Contig N50: 572
##
##  Median contig length: 356
##  Average contig: 503.55
##  Total assembled bases: 13895363
##
##
#######################################################
## ## Stats based on ONLY LONGEST ISOFORM per 'GENE':
#######################################################
##
##  Contig N10: 1599
##  Contig N20: 1103
##  Contig N30: 820
##  Contig N40: 639
##  Contig N50: 516
##
##  Median contig length: 347
##  Average contig: 470.82
##  Total assembled bases: 12254396
```

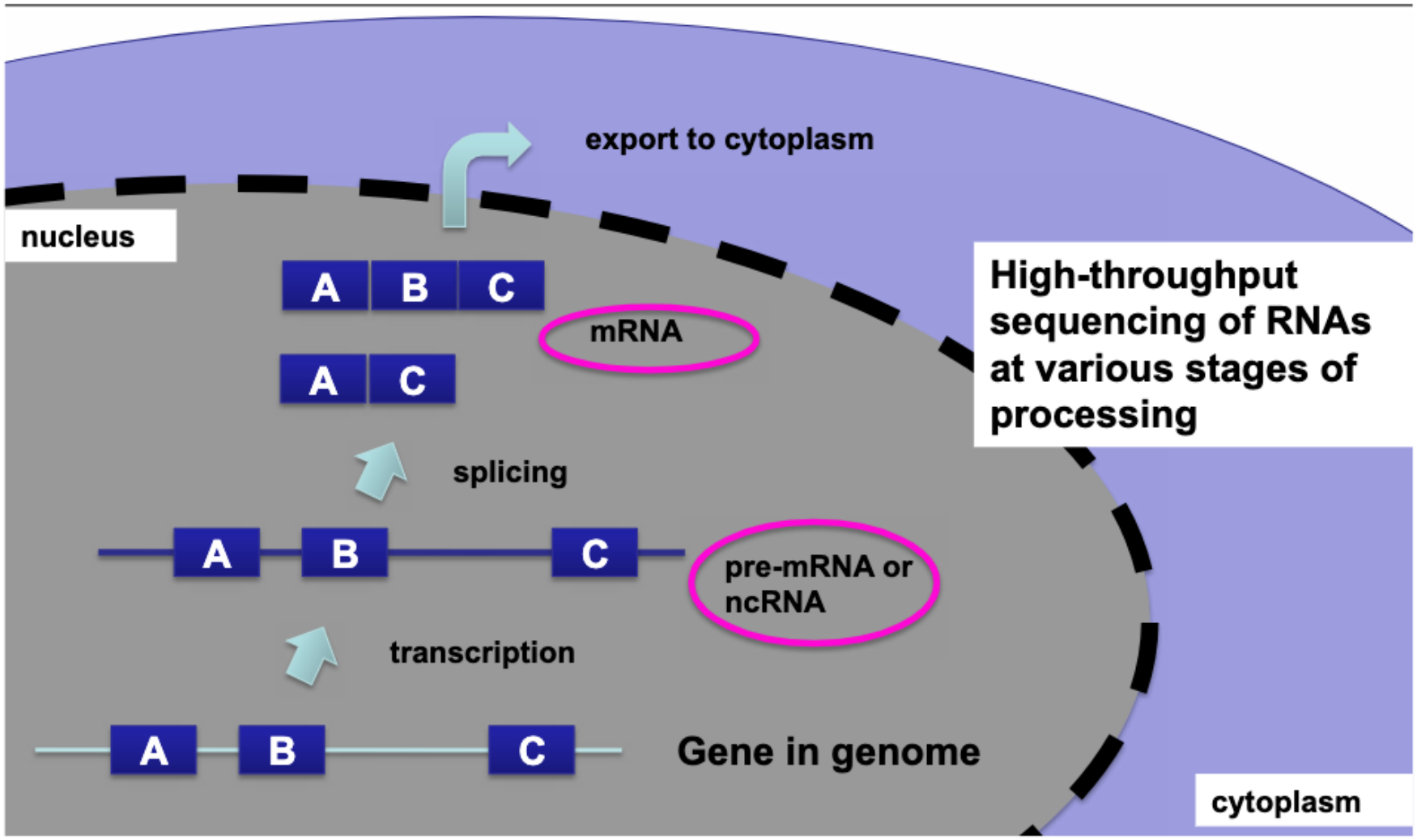
```
#!/usr/bin/env bash
# analyzeTrinity.sh
# Usage: bash scripts/analyzeTrinity.sh 1>results/trinity_guided_stats.txt 2>results/logs/analyzeTrinity.err

TrinityStats.pl results/trinity_guided/Trinity-GG.fasta
```

Lecture 7: Outline

- Transcriptome Assembly
  - Assembly v. alignment
  - Reference-guided v. *de novo*
  - Index-based alignment
  - Using Trinity (demo)
- Links in Unix
- Using IGV (round 1)

Review of gene expression



Courtesy of Cole Trapnell. Used with permission.

Slide courtesy Cole Trapnel

How is RNA-seq different from DNA-seq? What are we looking for?

The message:

‘It was the best of times, it was the vorst of times, it was the age of wisdom, it was the age of foolishness, it was the epoch of belief, it was the epoch of incredulity, it was the season of Light, it was the season of Darkness …’

DNA

RNA

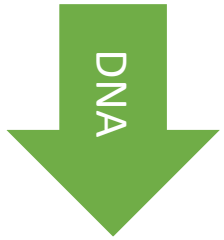
The features:

worst>vorst

1 belief	**2 age**
1 best	**2 epoch**
1 darkness	**2 season**
1 foolishness	**2 times**
1 incredulity	**8 it**
1 light	**8 of**
1 wisdom	**8 the**
1 vorst	**8 was**

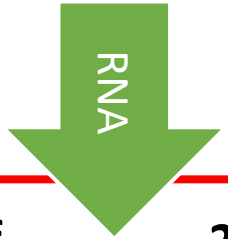
“expression”

How is RNA-seq different from DNA-seq? What are we looking for?



The features:

worst>vorst



1 belief	2 age
1 best	2 epoch
1 darkness	2 season
1 foolishness	2 times
1 incredulity	8 it
1 light	8 of
1 wisdom	8 the
1 vorst	8 was

"expression"

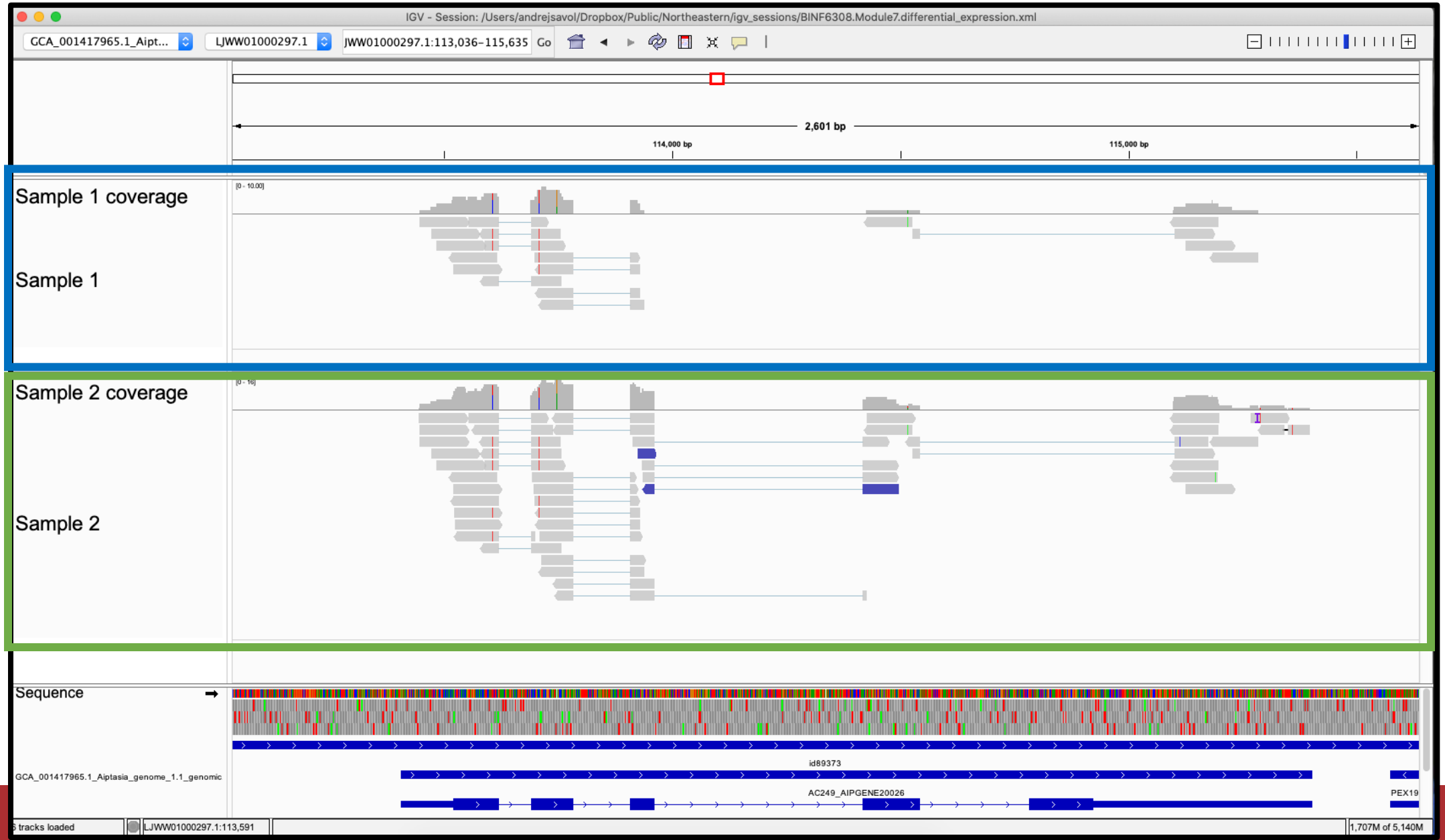
The question:

Which words are misspelled?

Which words occur more often than expected?

Calculating gene expression: a visual introduction

RNA-seq data from two samples

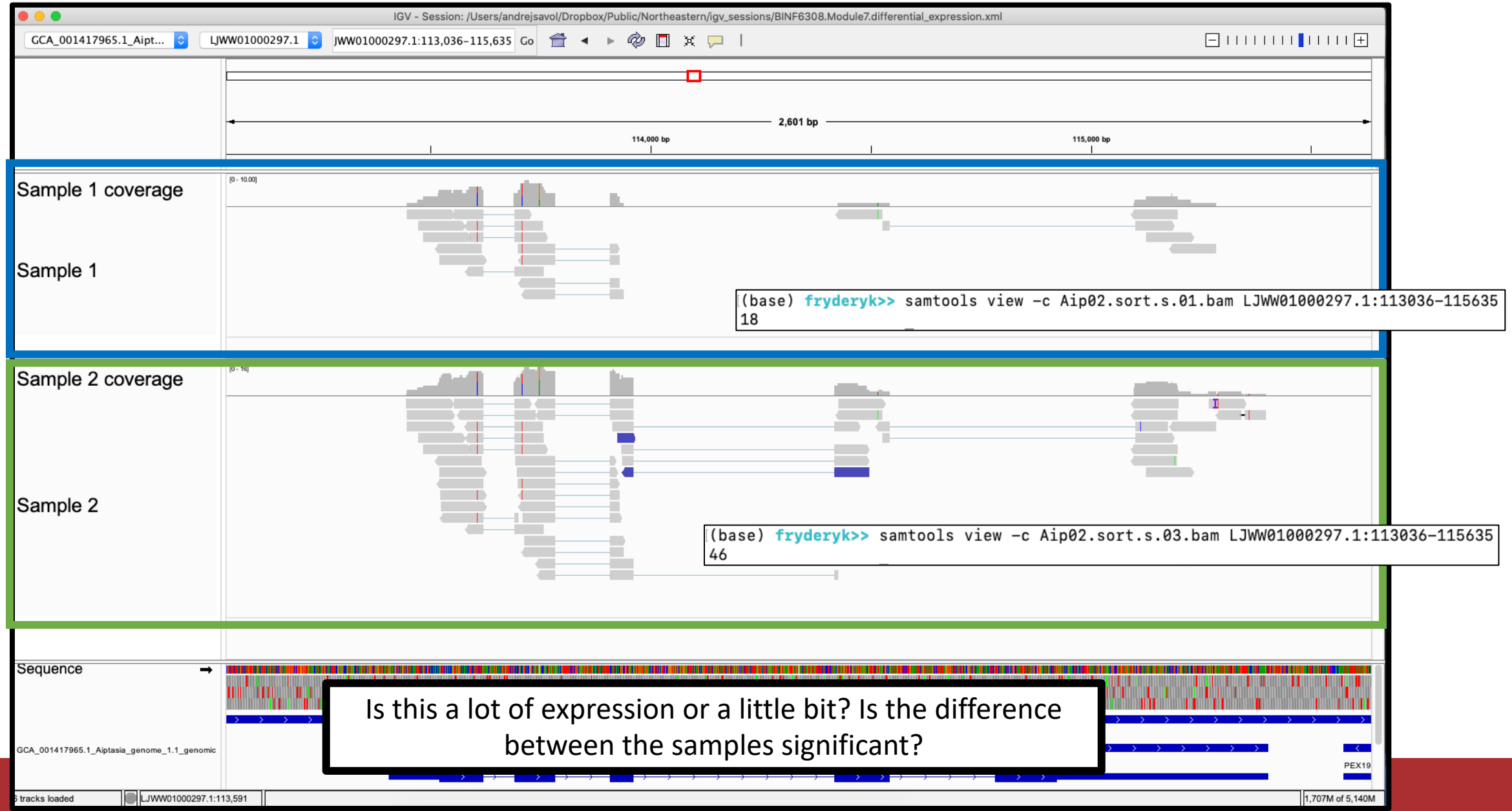


Calculating gene expression: a visual introduction

RNA-seq data from two samples

Calculating gene expression: a visual introduction

RNA-seq data from two samples



Why not just count reads per gene?

Some requirements for a unit of gene expression?

Normalized by gene length (so we can compare between genes)
Normalized by sequencing depth (so we can compare between samples)

RPKM of a gene: $\mathrm{RPKM} = \frac{\text{Number of reads mapped to gene} \times 10^3 \times 10^6}{\text{Total number of mapped reads} \times \text{Gene length in bp}}$

FPKM of a gene: $\mathrm{FPKM} = \frac{\text{Number of fragments mapped to gene} \times 10^3 \times 10^6}{\text{Total number of mapped reads} \times \text{Gene length in bp}}$

TPM of a gene: $\mathrm{TPM} = A \times \frac{1}{\sum(A)} \times 10^6$ Where $\mathrm{A} = \frac{\text{Total reads mapped to gene} \times 10^3}{\text{Gene length in bp}}$

https://shanguangyu.com/articles/differences-between-RPKM-FPKM-and-TPM/

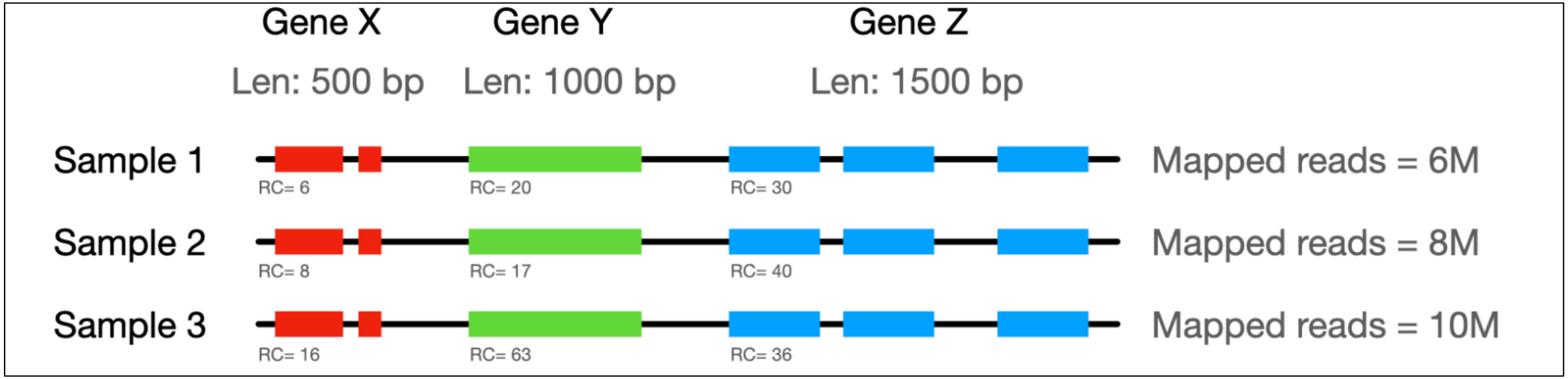
Toy example to compare RPKM and TPM

RPKM of a gene: $\mathrm{RPKM} = \frac{\text{Number of reads mapped to gene} \times 10^3 \times 10^6}{\text{Total number of mapped reads} \times \text{Gene length in bp}}$

$$\frac{\text{Reads}}{\text{Total mapped (in millions)} \times \text{Gene length In kilobases}}$$

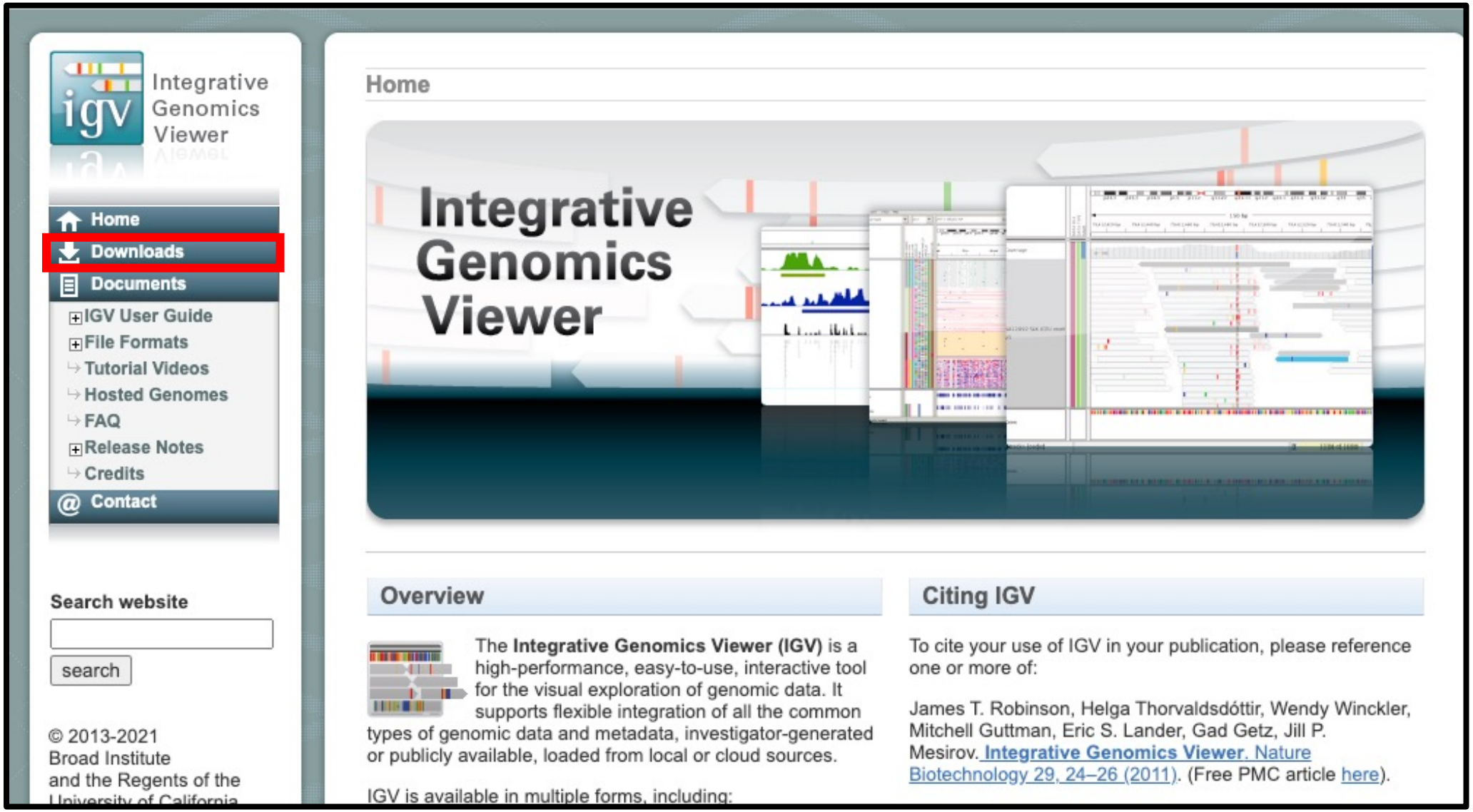
TPM of a gene: $\mathrm{TPM} = A \times \frac{1}{\sum(A)} \times 10^6$ Where $\mathrm{A} = \frac{\text{Total reads mapped to gene} \times 10^3}{\text{Gene length in bp}}$

$$A = \frac{\text{Reads}}{\text{Gene length In kilobases}}$$



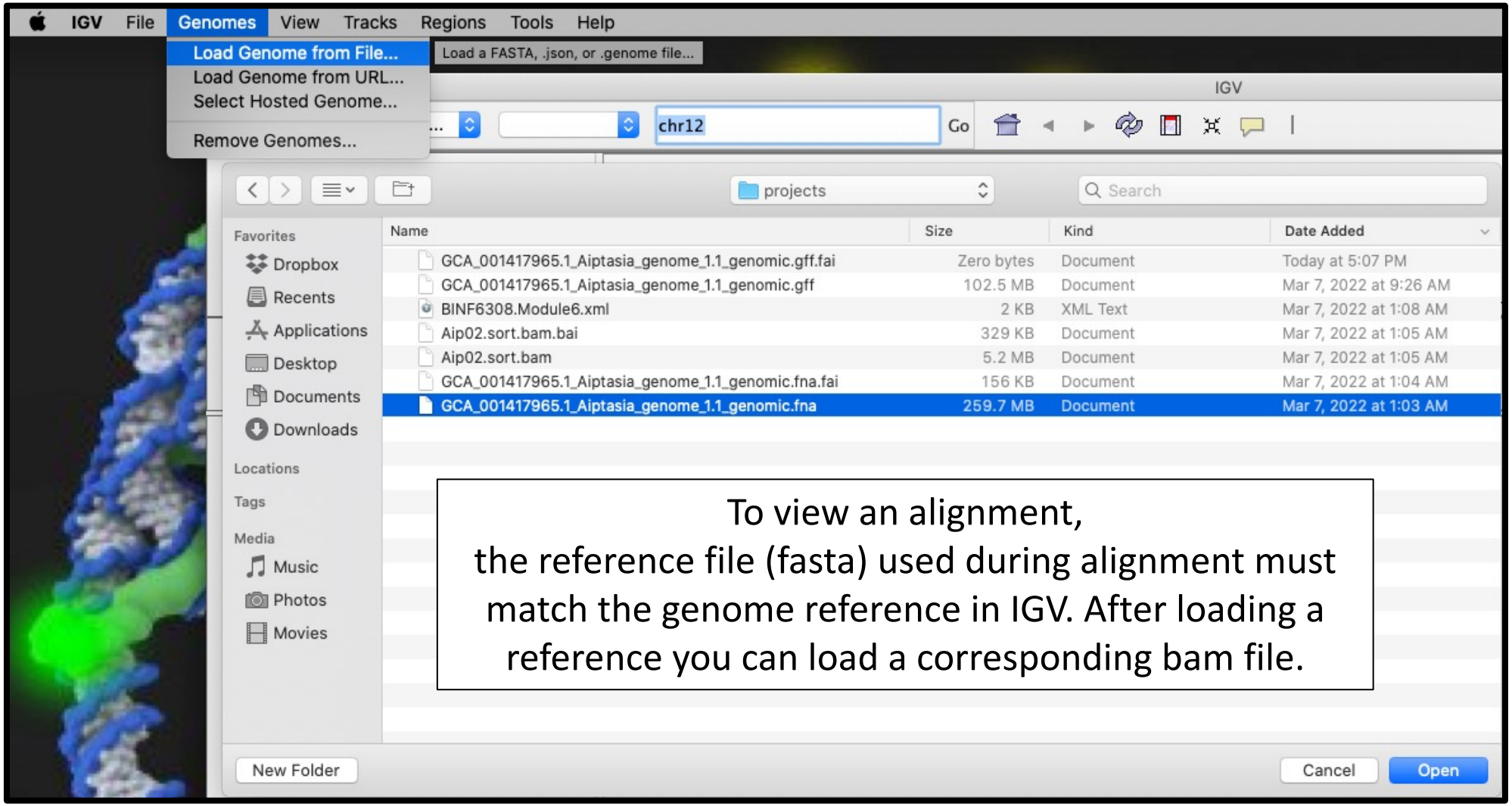
https://shanguangyu.com/articles/differences-between-RPKM-FPKM-and-TPM/

Introduction to IGV (Integrated Genomics Viewer)



https://software.broadinstitute.org/software/igv/

Generating a *.genome file for IGV



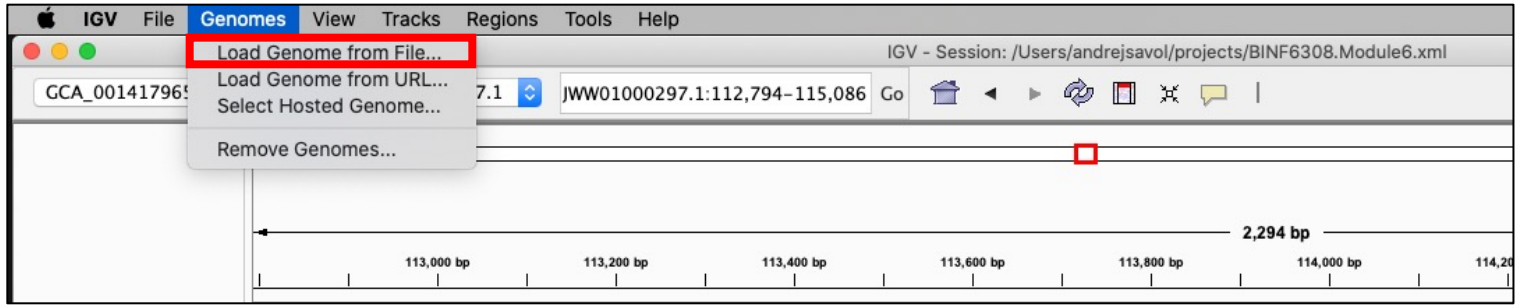
To view an alignment,
the reference file (fasta) used during alignment must match the genome reference in IGV. After loading a reference you can load a corresponding bam file.

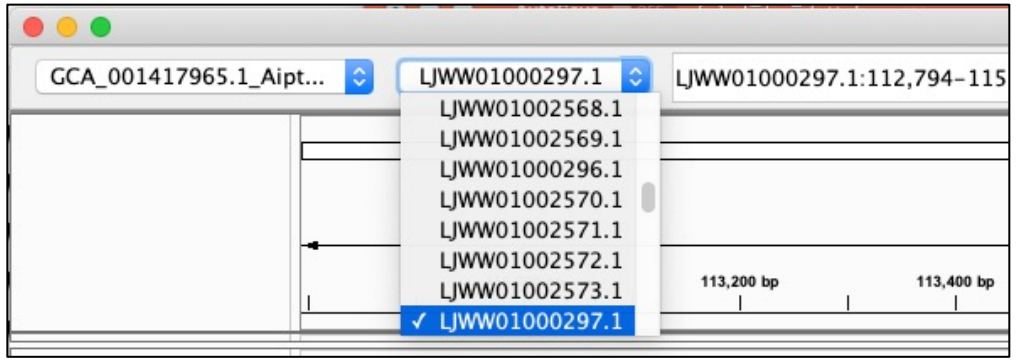
To view an alignment in IGV:

1. Load your reference genome as a fasta file:
   GCA_001417965.1_Aiptasia_genome_1.1_genomic.fna

2. Load your new bam file:
   Aip02.sort.bam

3. Navigate to a chromosome that has a decent number of reads (via the dropdown menu):
   LJWW01000297.1





To view an alignment in IGV:

Coverage track

Individual reads

Individual reads

External genome annotations

Which track is only typical in RNA (versus DNA)?