Lab Assignment 3: Sequence Comparison

2024 02 02

Obtained from 2024 MTEs

Part 1: Finding the difference between hg38 and CHM13

Part 1: Compare hg38 to CHM13

- hg38 is a chimeric reference from a few individuals
  - Maintained and developed by the Genome Research Consortium
  - Many gaps and satellite contigs
- CHM13 is from the CHM13 cell line that expresses human telomerase reverse transcriptase (hTERT) (retains 46,XX karyotype)
  - Developed by the T2T consortium

Cold Spring Harbor Laboratory

bioRxiv

THE PREPRINT SERVER FOR BIOLOGY

bioRxiv posts many COVID19-related papers. A reminder: they have not been formally peer-reviewed and should not guide health-related behavior or be reported in the press as conclusive.

New Results

Follow this preprint

The complete sequence of a human genome

Sergey Nurk, Sergey Koren, Arang Rhie, Mikko Rautiainen, Andrey V. Bzikadze, Alla Mikheenko, Mitchell R. Vollger, Nicolas Altemose, Lev Uralsky, Ariel Gershman, Sergey Aganezov, Savannah J. Hoyt, Mark Diekhans, Glennis A. Logsdon, Michael Alonge, Stylianos E. Antonarakis, Matthew Borchers, Gerard G. Bouffard, Shelise Y. Brooks, Gina V. Caldas, Haoyu Cheng, Chen-Shan Chin, William Chow, Leonardo G. de Lima, Philip C. Dishuck, Richard Durbin, Tatiana Dvorkina, Ian T. Fiddes, Giulio Formenti, Robert S. Fulton, Arkarachai Fungtammasan, Erik Garrison, Patrick G.S. Grady, Tina A. Graves-Lindsay, Ira M. Hall, Nancy F. Hansen, Gabrielle A. Hartley, Marina Haukness, Kerstin Howe, Michael W. Hunkapiller, Chirag Jain, Miten Jain, Erich D. Jarvis, Peter Kerpedjiev, Melanie Kirsche, Mikhail Kolmogorov, Jonas Korlach, Milinn Kremitzki, Heng Li, Valerie V. Maduro, Tobias Marschall, Ann M. McCartney, Jennifer McDaniel, Danny E. Miller, James C. Mullikin, Eugene W. Myers, Nathan D. Olson, Benedict Paten, Paul Peluso, Pavel A. Pevzner, David Porubsky, Tamara Potapova, Evgeny I. Rogaev, Jeffrey A. Rosenfeld, Steven L. Salzberg, Valerie A. Schneider, Fritz J. Sedlazeck, Kishwar Shafin, Colin J. Shew, Alaina Shumate, Yumi Sims, Arian F. A. Smit, Daniela C. Soto, Ivan Sović, Jessica M. Storer, Aaron Streets, Beth A. Sullivan, Françoise Thibaud-Nissen, James Torrance, Justin Wagner, Brian P. Walenz, Aaron Wenger, Jonathan M. D. Wood, Chunlin Xiao, Stephanie M. Yan, Alice C. Young, Samantha Zarate, Urvashi Surti, Rajiv C. McCoy, Megan Y. Dennis, Ivan A. Alexandrov, Jennifer L. Gerton, Rachel J. O'Neill, Winston Timp, Justin M. Zook, Michael C. Schatz, Evan E. Eichler, Karen H. Miga, Adam M. Phillippy

doi: https://doi.org/10.1101/2021.05.26.445798

This article is a preprint and has not been certified by peer review [what does this mean?].

0 0 0 0 85 0 3418

Abstract | Full Text | Info/History | Metrics | Preview PDF

Part 1: Compare hg38 to CHM13

- Copy the finding_gaps_template.py script to your directory
  - /storage1/fs1/workshops/Active/BIO5488/assignments/assignment3/finding_gaps_template.py

- Complete the script to do the following:
  - Count nucleotide frequencies
  - Count dinucleotide frequencies
  - Count and gaps and gap sizes
  - Plot gap size distribution

Part 2: Align Reads to Chr22 and mark Duplicates

Step 1: Create Indices for hg38 and CHM13

- Create an index for both assemblies of chr22
- Example command in assignment

```
asouthard-smith@genomics:~$ bowtie2-build
No input sequence or sequence file specified!
Bowtie 2 version 2.3.4.1 by Ben Langmead (langmea@cs.jhu.edu, www.cs.jhu.edu/~langmea)
Usage: bowtie2-build [options]* <reference_in> <bt2_index_base>
    reference_in            comma-separated list of files with ref sequences
    bt2_index_base          write bt2 data to files with this dir/basename
*** Bowtie 2 indexes work only with v2 (not v1).  Likewise for v1 indexes. ***
Options:
    -f                      reference files are Fasta (default)
    -c                      reference sequences given on cmd line (as
                            <reference_in>)
    --large-index           force generated index to be 'large', even if ref
                            has fewer than 4 billion nucleotides
    -a/--noauto             disable automatic -p/--bmax/--dcv memory-fitting
    -p/--packed             use packed strings internally; slower, less memory
    --bmax <int>            max bucket sz for blockwise suffix-array builder
    --bmaxdivn <int>        max bucket sz as divisor of ref len (default: 4)
    --dcv <int>             diff-cover period for blockwise (default: 1024)
    --nodc                  disable diff-cover (algorithm becomes quadratic)
    -r/--noref              don't build .3/.4 index files
    -3/--justref            just build .3/.4 index files
    -o/--offrate <int>      SA is sampled every 2^<int> BWT chars (default: 5)
    -t/--ftabchars <int>    # of chars consumed in initial lookup (default: 10)
    --threads <int>         # of threads
    --seed <int>            seed for random number generator
    -q/--quiet              verbose output (for debugging)
    -h/--help               print detailed description of tool and its options
    --usage                 print this usage message
    --version               print version information and quit
asouthard-smith@genomics:~$
```

Step 2: Align with bowtie2

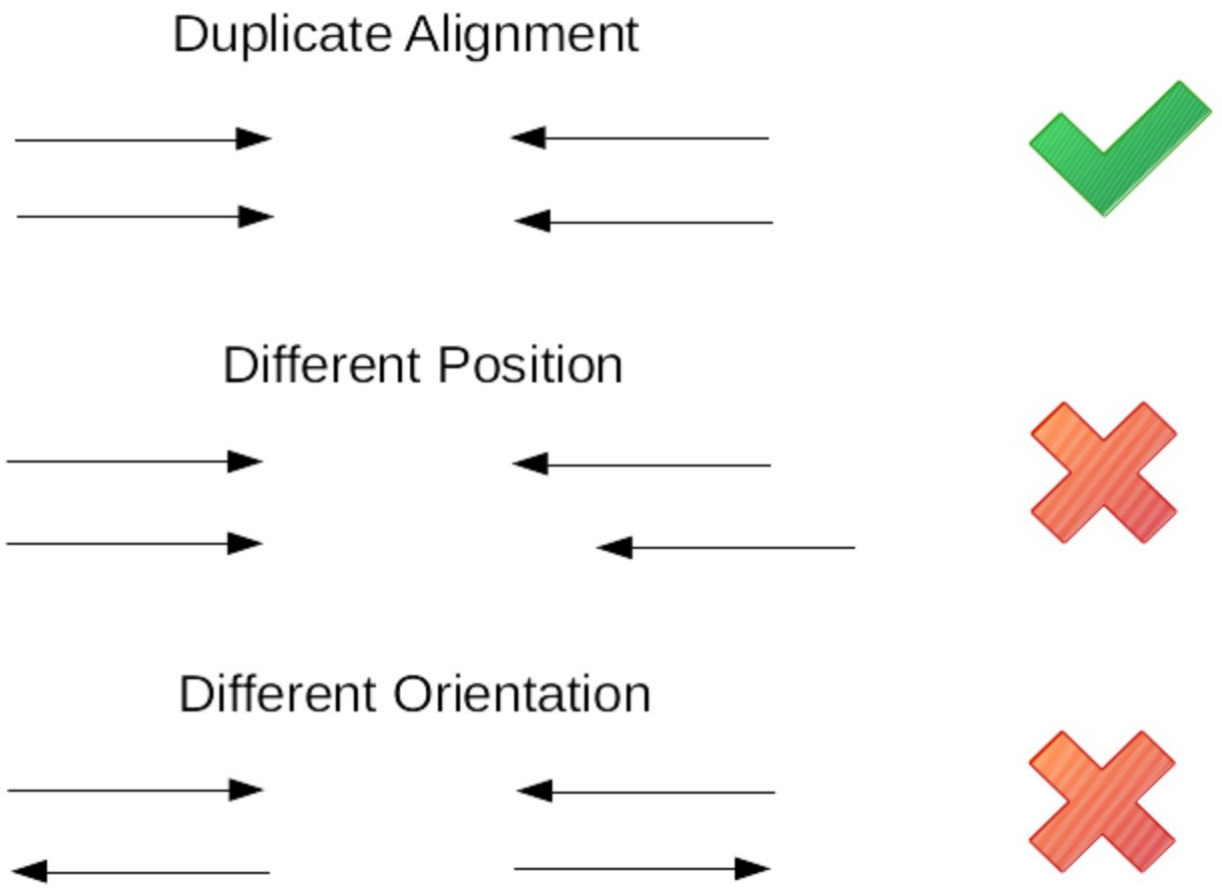
- Reads are in:
  /storage1/fs1/workshops/Active/BIO5488/assignments/assignment3/test-500k.fq
- The fastq file is unpaired
- Save both the alignment and report output files

http://bowtie-bio.sourceforge.net/bowtie2/manual.shtml

```
asouthard-smith@genomics:~$ bowtie2
No index, query, or output file specified!
Bowtie 2 version 2.3.4.1 by Ben Langmead (langmea@cs.jhu.edu, www.cs.jhu.edu/~langmea)
Usage:
  bowtie2 [options]* -x <bt2-idx> {-1 <m1> -2 <m2> | -U <r> | --interleaved <i>} [-S <sam>]

  <bt2-idx>  Index filename prefix (minus trailing .X.bt2).
             NOTE: Bowtie 1 and Bowtie 2 indexes are not compatible.
  <m1>       Files with #1 mates, paired with files in <m2>.
             Could be gzip'ed (extension: .gz) or bzip2'ed (extension: .bz2).
  <m2>       Files with #2 mates, paired with files in <m1>.
             Could be gzip'ed (extension: .gz) or bzip2'ed (extension: .bz2).
  <r>        Files with unpaired reads.
             Could be gzip'ed (extension: .gz) or bzip2'ed (extension: .bz2).
  <i>        Files with interleaved paired-end FASTQ reads
             Could be gzip'ed (extension: .gz) or bzip2'ed (extension: .bz2).
  <sam>      File for SAM output (default: stdout)

  <m1>, <m2>, <r> can be comma-separated lists (no whitespace) and can be
  specified many times.  E.g. '-U file1.fq,file2.fq -U file3.fq'.

Options (defaults in parentheses):

 Input:
  -q                 query input files are FASTQ .fq/.fastq (default)
  --tab5             query input files are TAB5 .tab5
  --tab6             query input files are TAB6 .tab6
  --qseq             query input files are in Illumina's qseq format
  -f                 query input files are (multi-)FASTA .fa/.mfa
  -r                 query input files are raw one-sequence-per-line
  -F k:<int>,i:<int> query input files are continuous FASTA where reads
                     are substrings (k-mers) extracted from a FASTA file <s>
                     and aligned at offsets 1, 1+i, 1+2i ... end of reference
  -c                 <m1>, <m2>, <r> are sequences themselves, not files
  -s/--skip <int>    skip the first <int> reads/pairs in the input (none)
  -u/--upto <int>    stop after first <int> reads/pairs (no limit)
  -5/--trim5 <int>   trim <int> bases from 5'/left end of reads (0)
  -3/--trim3 <int>   trim <int> bases from 3'/right end of reads (0)
  --phred33          qualities are Phred+33 (default)
  --phred64          qualities are Phred+64
  --int-quals        qualities encoded as space-delimited integers

 Presets:                 Same as:
  For --end-to-end:
   --very-fast           -D 5 -R 1 -N 0 -L 22 -i S,0,2.50
   --fast                -D 10 -R 2 -N 0 -L 22 -i S,0,2.50
   --sensitive           -D 15 -R 2 -N 0 -L 22 -i S,1,1.15 (default)
   --very-sensitive      -D 20 -R 3 -N 0 -L 20 -i S,1,0.50

  For --local:
   --very-fast-local     -D 5 -R 1 -N 0 -L 25 -i S,1,2.00
   --fast-local          -D 10 -R 2 -N 0 -L 22 -i S,1,1.75
   --sensitive-local     -D 15 -R 2 -N 0 -L 20 -i S,1,0.75 (default)
   --very-sensitive-local -D 20 -R 3 -N 0 -L 20 -i S,1,0.50

 Alignment:
```

Step 3: Remove Duplicates with Samtools

- Duplicate reads are measured from the position of the 5’ end of the read
- Duplicates are the result of non-uniform amplification and hybridization rates during library construction & sequencing

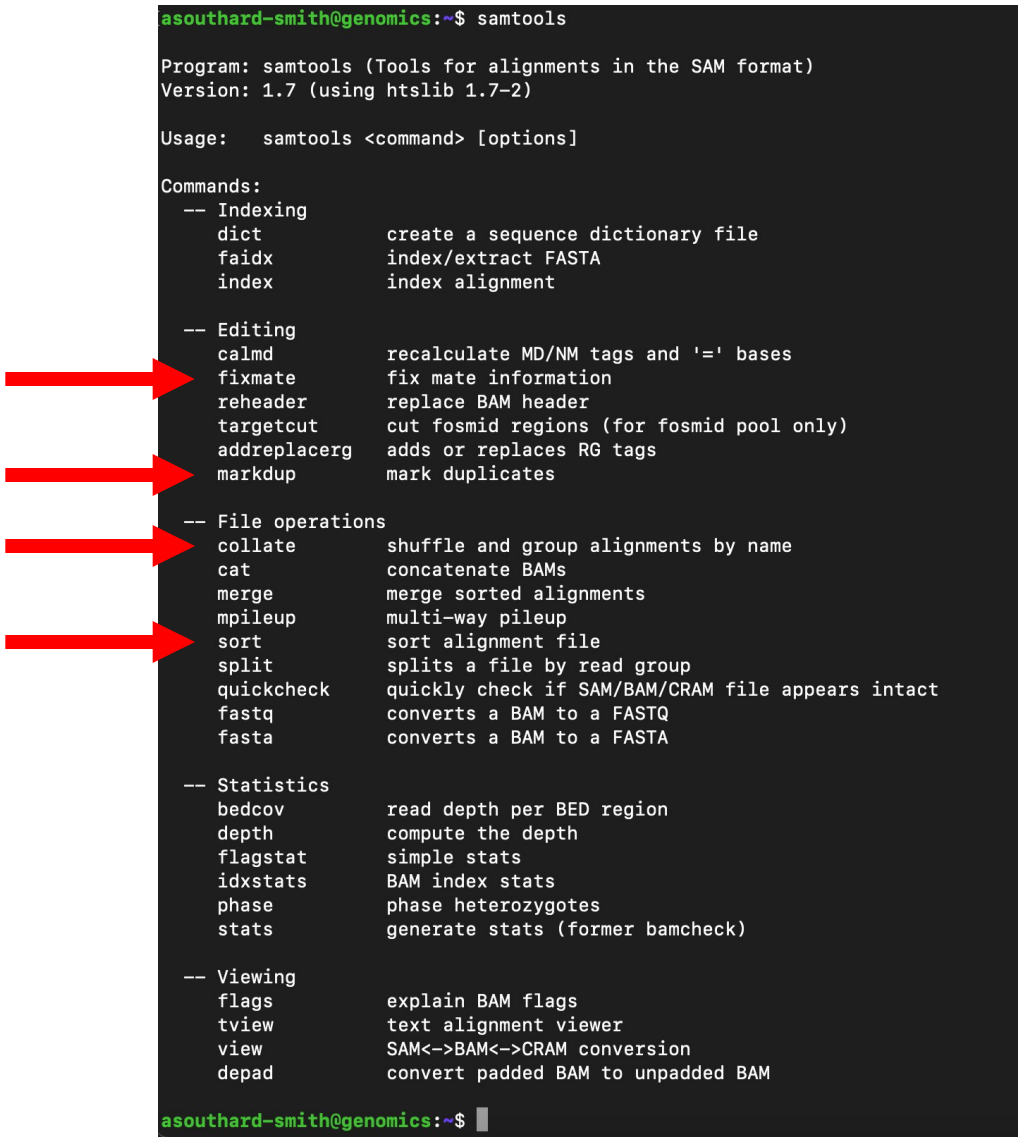


https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-016-1097-3

http://www.htsli b.org/algorithms/duplicate.html

Step 3: Remove Duplicates with Samtools

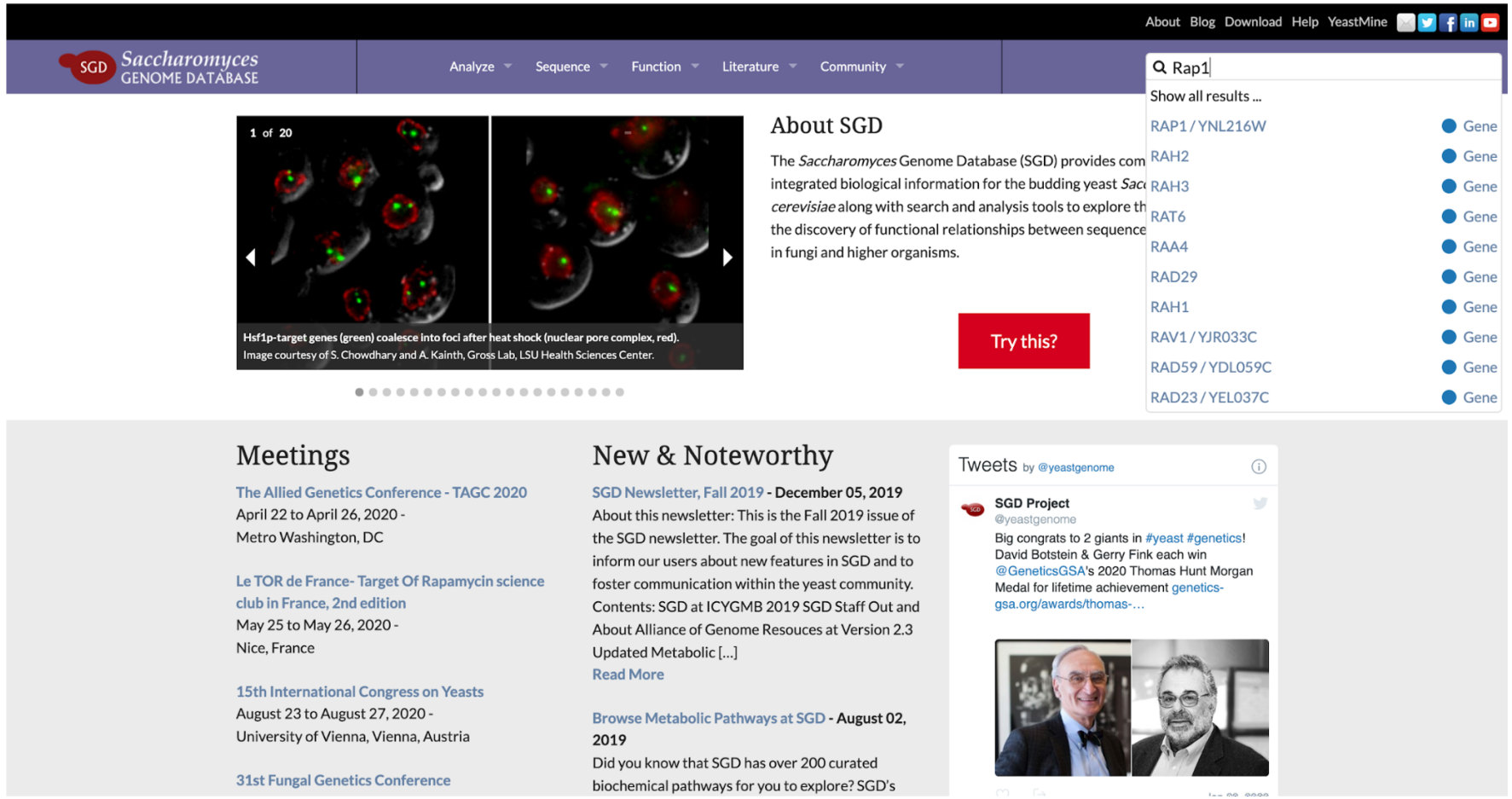
- 4 stages:
  - Collate
  - Fixmate
  - Sort
  - MarkDuplicates
- markdup output should be saved to a file
- markdup report will be output to terminal



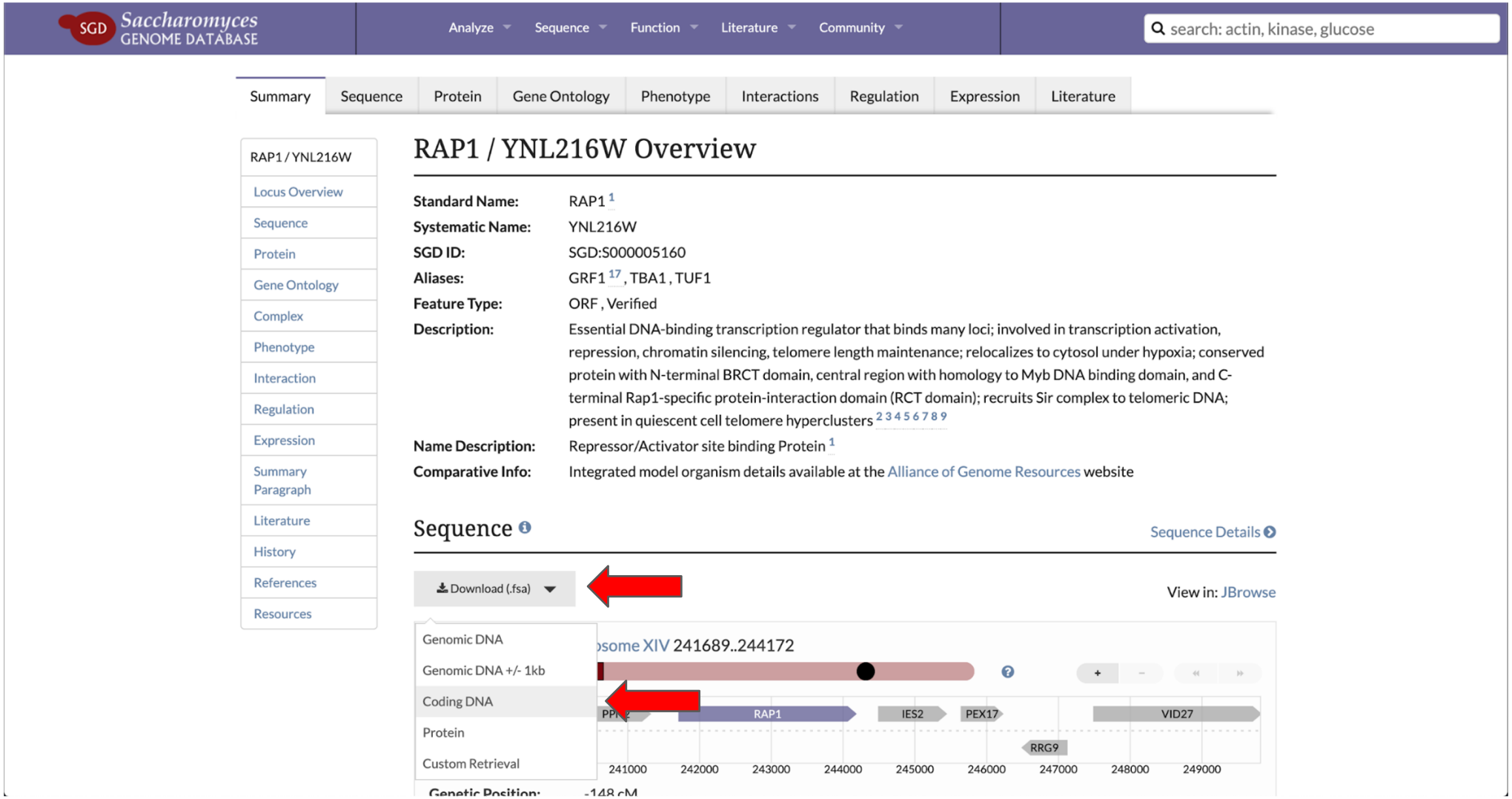
Part 3: Running BLAST

Step 1: Obtain Gene Sequence

- Obtain **coding** sequence for *RAP1* from yeastgenome.org

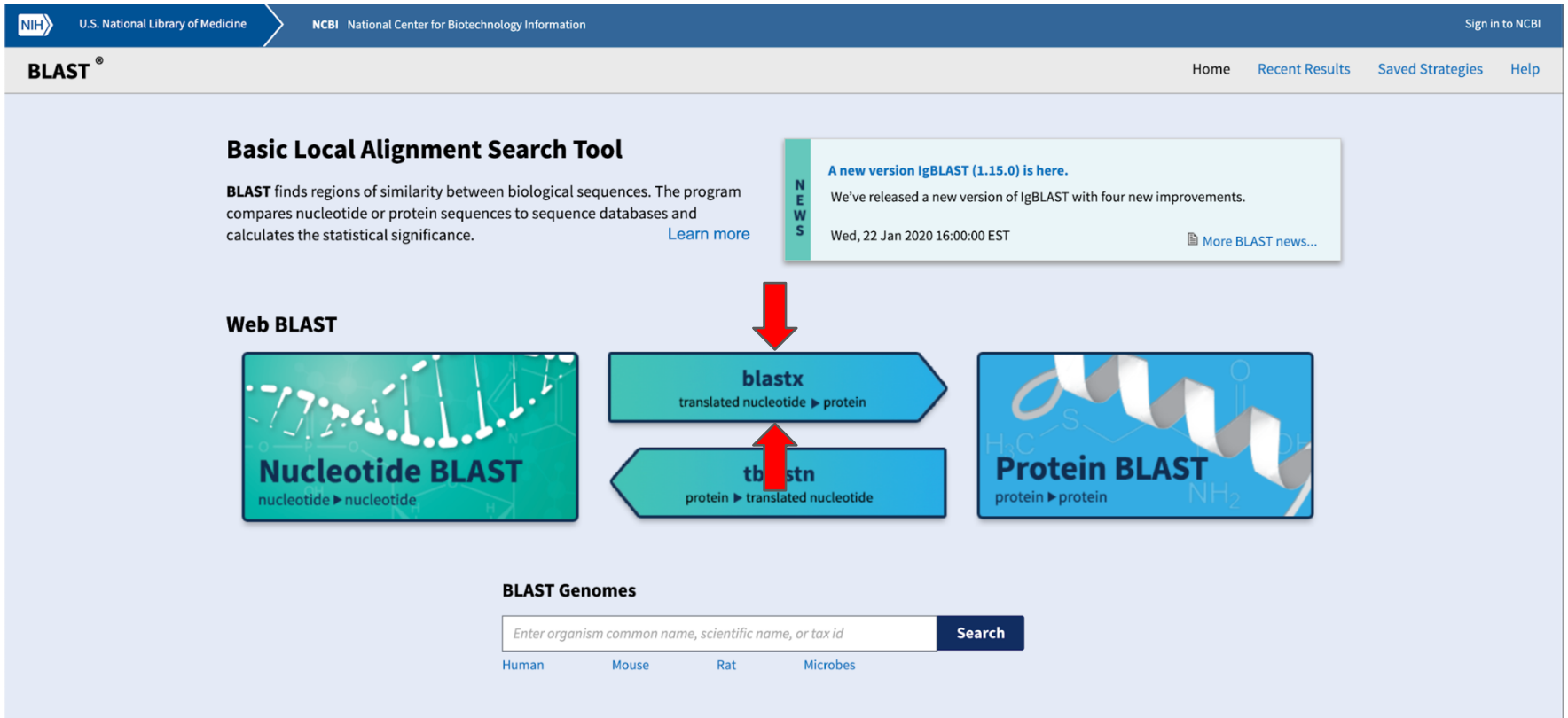


Step 1: Obtain Gene Sequence

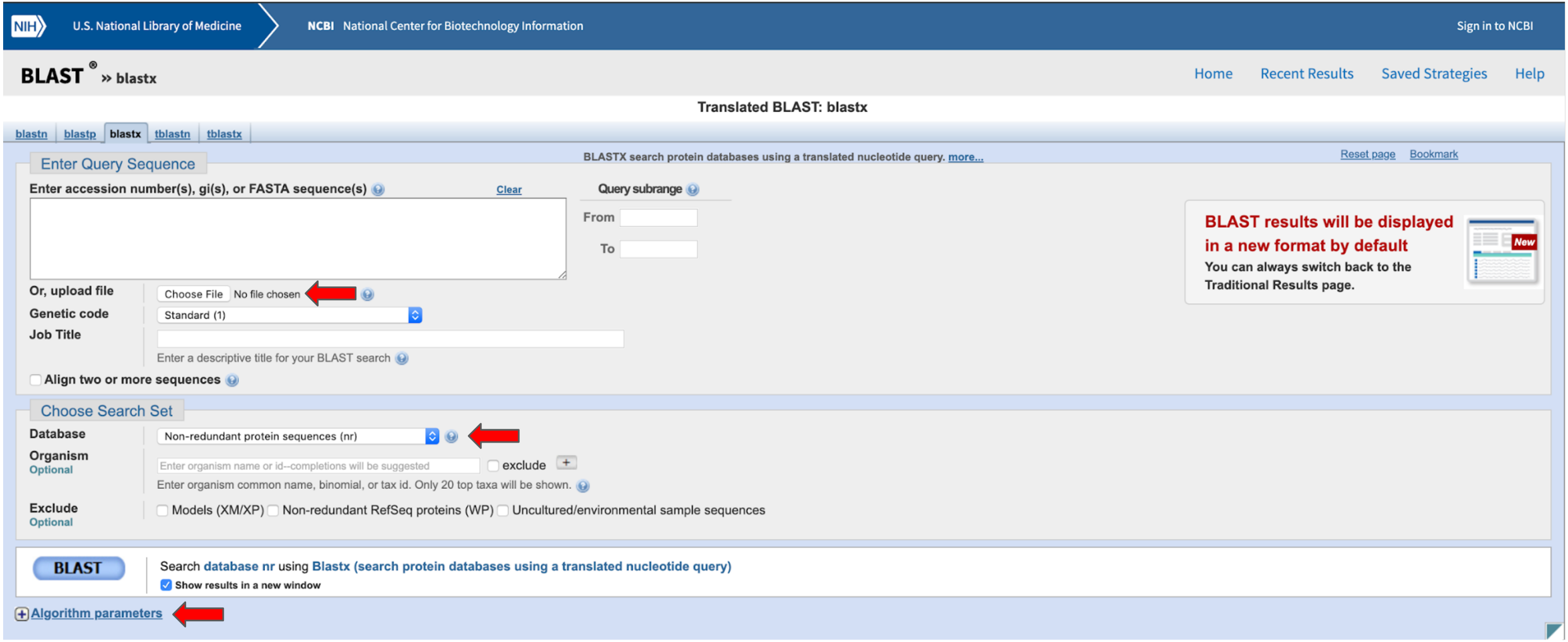


Step 2: Run BLASTx Web Tool

- Compare BLOSUM62 vs. BLOSUM80
- Compare default existence (11), extension (1) penalties vs existence of 7, extension of 2



Step 2: Run BLASTx Web Tool



Step 2: Run BLASTx Web Tool

Things to turn in:

- README.txt
- finding_gaps.py script
- hg38-chr22_gap_distribution.png
- Bowtie2 Alignment files
- Bowtie2 Report files
- Samtools duplicates removed sam file
- Extra credit items:
  - Additional gap size distribution plot
  - Python script for tallying duplicates from the bowtie2 aligned sam file.