

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



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 Rautainen, 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 Kerpeljiev, Melanie Kirsch, Mikhail Kolmogorov, Jonas Korlach, Milinn Kremitzki, Heng Li, Valerie V. Maduro, Tobias Marschall, Ann M. McCarty, 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. Sedlacek, 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?].



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)
      --nодc                  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

```
[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
  <m1>, <m2>, <r> are sequences themselves, not files
  -c          skip the first <int> reads/pairs in the input (none)
  -s/--skip <int> stop after first <int> reads/pairs (no limit)
  -u/--upto <int> trim <int> bases from 5'/left end of reads (0)
  -5/--trim5 <int> trim <int> bases from 3'/right end of reads (0)
  -3/--trim3 <int> qualities are Phred+33 (default)
  --phred33   qualities are Phred+64
  --phred64   qualities encoded as space-delimited integers

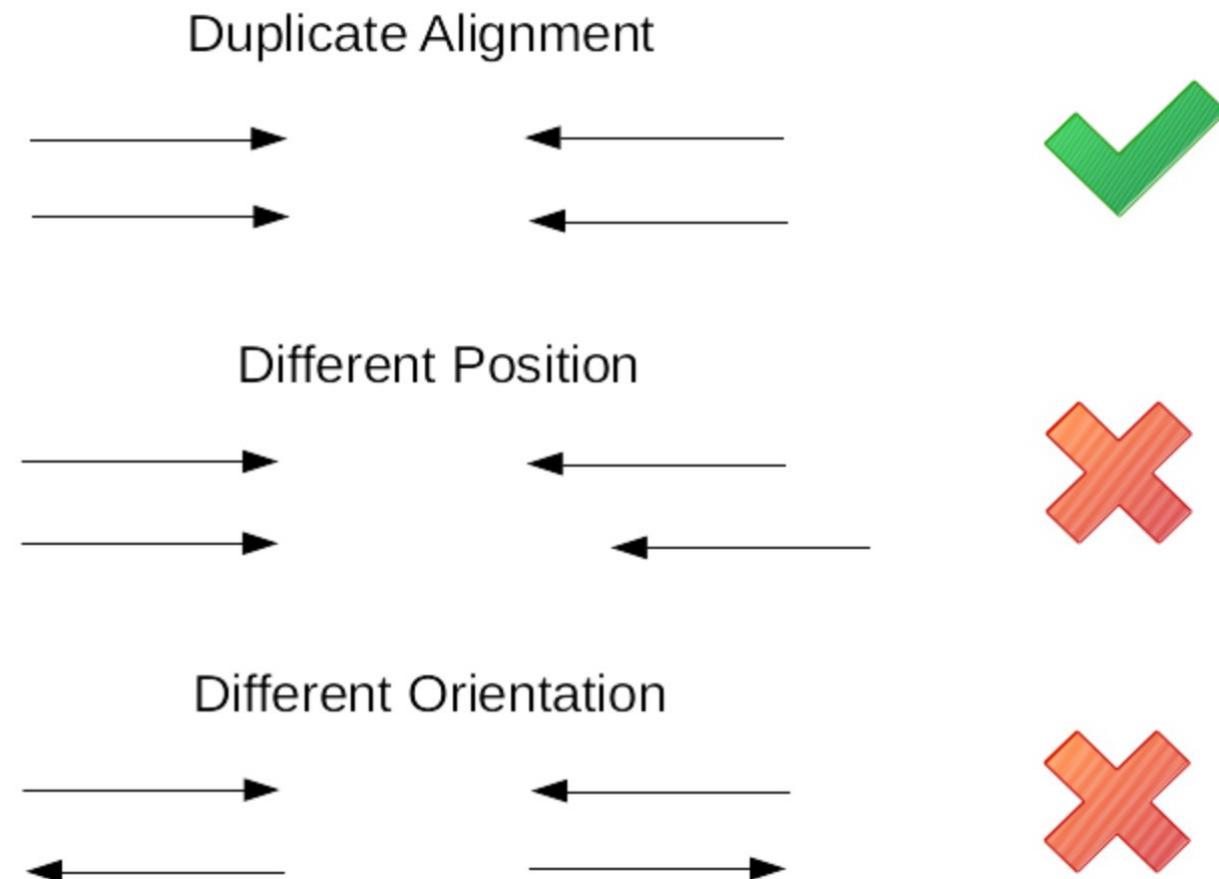
Presets:           Same as:
  For --end-to-end:
  --very-fast    -D 5 -R 1 -N 0 -L 22 -i S,0,2.50
  --fast          -D 10 -R 2 -N 0 -L 22 -i S,0,2.50
  --sensitive    -D 15 -R 2 -N 0 -L 22 -i S,1,1.15 (default)
  --very-sensitive -D 20 -R 3 -N 0 -L 20 -i S,1,0.50

  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



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

```
asouthard-smith@genomics:~$ samtools

Program: samtools (Tools for alignments in the SAM format)
Version: 1.7 (using htslib 1.7-2)

Usage: samtools <command> [options]

Commands:
  -- Indexing
    dict          create a sequence dictionary file
    faidx        index/extract FASTA
    index         index alignment

  -- Editing
    calmd        recalculate MD/NM tags and '=' bases
    fixmate      fix mate information
    reheader     replace BAM header
    targetcut   cut fosmid regions (for fosmid pool only)
    addreplacerg adds or replaces RG tags
    markdup     mark duplicates

  -- File operations
    collate      shuffle and group alignments by name
    cat          concatenate BAMs
    merge        merge sorted alignments
    mpileup     multi-way pileup
    sort         sort alignment file
    split        splits a file by read group
    quickcheck  quickly check if SAM/BAM/CRAM file appears intact
    fastq       converts a BAM to a FASTQ
    fasta        converts a BAM to a FASTA

  -- Statistics
    bedcov      read depth per BED region
    depth       compute the depth
    flagstat    simple stats
    idxstats   BAM index stats
    phase       phase heterozygotes
    stats       generate stats (former bamcheck)

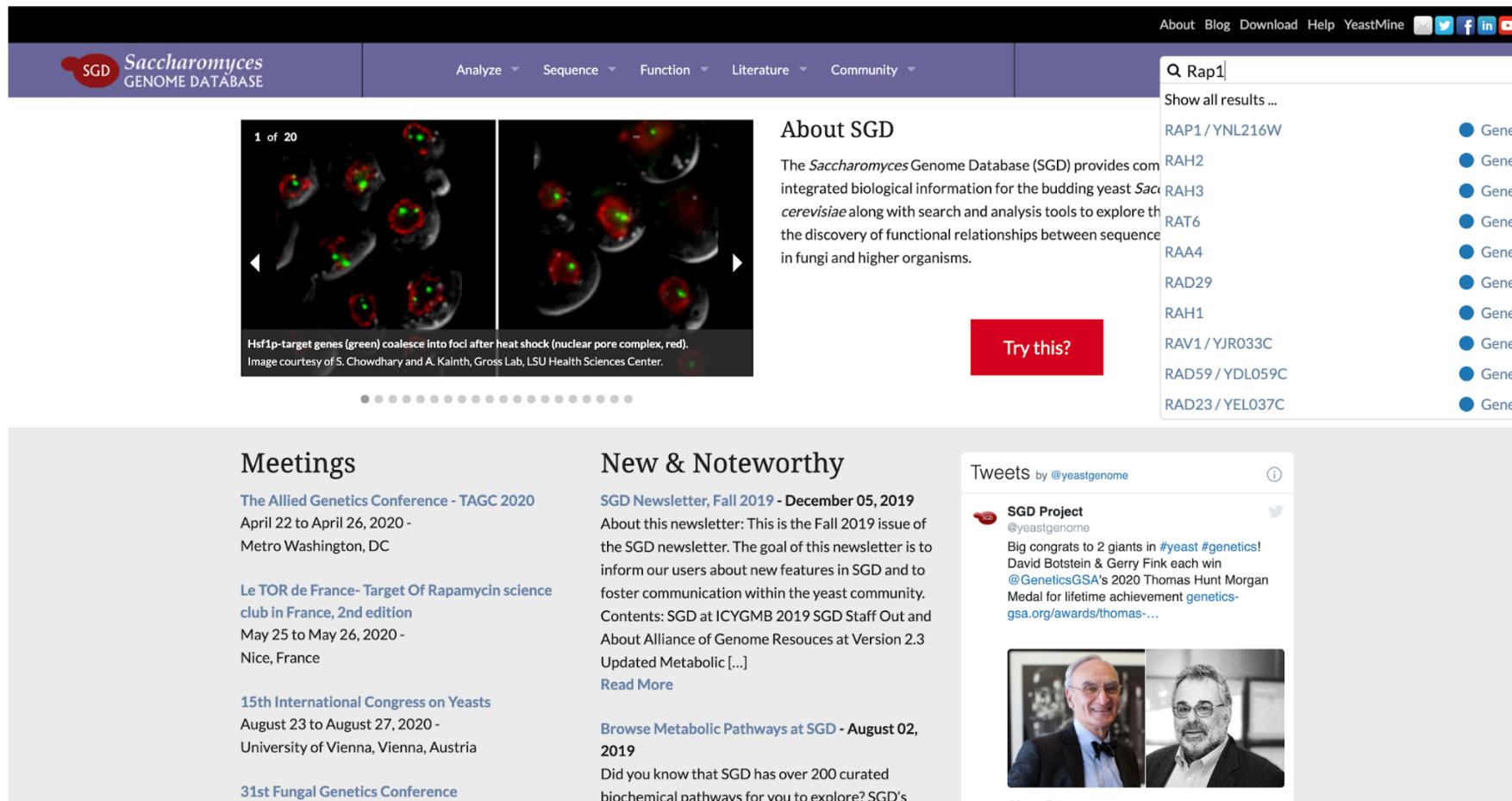
  -- Viewing
    flags        explain BAM flags
    tview       text alignment viewer
    view        SAM<->BAM<->CRAM conversion
    depad      convert padded BAM to unpadded BAM

asouthard-smith@genomics:~$
```

Part 3: Running BLAST

Step 1: Obtain Gene Sequence

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



The SGD website interface is shown. The top navigation bar includes links for About, Blog, Download, Help, YeastMine, and social media icons. A search bar on the right contains the query 'Rap1'. Below the search bar, a list of gene results is displayed, each preceded by a blue circular icon and the text 'Gene'. The results include RAP1/YNL216W, RAH2, RAH3, RAT6, RAA4, RAD29, RAH1, RAV1/YJR033C, RAD59/YDL059C, and RAD23/YEL037C. The main content area features a banner image of yeast cells with green and red fluorescence, and a text box explaining Hsf1p-target genes coalesce into foci after heat shock. Below this are sections for 'Meetings', 'New & Noteworthy', and a 'Tweets' feed.

SGD Saccharomyces GENOME DATÁBASE

Analyze Sequence Function Literature Community

Q Rap1

About SGD

The *Saccharomyces* Genome Database (SGD) provides com integrated biological information for the budding yeast *Sac cerevisiae* along with search and analysis tools to explore th the discovery of functional relationships between sequence in fungi and higher organisms.

Try this?

1 of 20

Hsf1p-target genes (green) coalesce into foci after heat shock (nuclear pore complex, red). Image courtesy of S. Chowdhary and A. Kainth, Gross Lab, LSU Health Sciences Center.

Meetings

The Allied Genetics Conference - TAGC 2020
April 22 to April 26, 2020 -
Metro Washington, DC

Le TOR de France- Target Of Rapamycin science club in France, 2nd edition
May 25 to May 26, 2020 -
Nice, France

15th International Congress on Yeasts
August 23 to August 27, 2020 -
University of Vienna, Vienna, Austria

31st Fungal Genetics Conference

New & Noteworthy

SGD Newsletter, Fall 2019 - December 05, 2019
About this newsletter: This is the Fall 2019 issue of the SGD newsletter. The goal of this newsletter is to inform our users about new features in SGD and to foster communication within the yeast community. Contents: SGD at ICGM 2019 SGD Staff Out and About Alliance of Genome Resources at Version 2.3 Updated Metabolic [...] [Read More](#)

Browse Metabolic Pathways at SGD - August 02, 2019
Did you know that SGD has over 200 curated biochemical pathways for you to explore? SGD's

Tweets by @yeastgenome

SGD Project
@yeastgenome
Big congrats to 2 giants in #yeast #genetics! David Botstein & Gerry Fink each win @GeneticsGSA's 2020 Thomas Hunt Morgan Medal for lifetime achievement genetics-gsa.org/awards/thomas-...

Step 1: Obtain Gene Sequence

SGD *Saccharomyces*
GENOME DATABASE

Analyze ▾ Sequence ▾ Function ▾ Literature ▾ Community ▾

search: actin, kinase, glucose

Summary Sequence Protein Gene Ontology Phenotype Interactions Regulation Expression Literature

RAP1 / YNL216W Overview

Standard Name: RAP1¹

Systematic Name: YNL216W

SGD ID: SGD:S000005160

Aliases: GRF1¹⁷, TBA1, TUF1

Feature Type: ORF, Verified

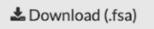
Description: Essential DNA-binding transcription regulator that binds many loci; involved in transcription activation, repression, chromatin silencing, telomere length maintenance; relocates to cytosol under hypoxia; conserved protein with N-terminal BRCT domain, central region with homology to Myb DNA binding domain, and C-terminal Rap1-specific protein-interaction domain (RCT domain); recruits Sir complex to telomeric DNA; present in quiescent cell telomere hyperclusters^{2 3 4 5 6 7 8 9}

Name Description: Repressor/Activator site binding Protein¹

Comparative Info: Integrated model organism details available at the [Alliance of Genome Resources](#) website

Sequence

Sequence Details 

  View in: JBrowse

Genomic DNA  Chromosome XIV 241689..244172

Genomic DNA +/- 1kb

Coding DNA  RAP1

Protein  IES2 PEX17 VID27

Custom Retrieval  RRG9

Genetic Position: -148 cM

Step 2: Run BLASTx Web Tool

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

The screenshot shows the NCBI BLAST homepage. At the top, there are links for NIH, U.S. National Library of Medicine, NCBI National Center for Biotechnology Information, and a 'Sign in to NCBI' button. The main navigation bar includes 'Home', 'Recent Results', 'Saved Strategies', and 'Help'. A 'NEWS' sidebar on the left highlights a new version of IgBLAST (1.15.0) released on Wednesday, 22 Jan 2020, 16:00:00 EST. The 'Basic Local Alignment Search Tool' section describes BLAST's function of finding regions of similarity between biological sequences. Below this, the 'Web BLAST' section features three search tools: 'Nucleotide BLAST' (nucleotide to nucleotide), 'blastx' (translated nucleotide to protein, highlighted with a red arrow), and 'tbstn' (protein to translated nucleotide). To the right is 'Protein BLAST' (protein to protein). At the bottom, the 'BLAST Genomes' section includes a search bar and links for 'Human', 'Mouse', 'Rat', and 'Microbes'.

Step 2: Run BLASTx Web Tool

NIH U.S. National Library of Medicine NCBI National Center for Biotechnology Information Sign in to NCBI

BLAST® > blastx Home Recent Results Saved Strategies Help

Translated BLAST: blastx

blastn blastp blastx tbblastn tbblastx

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s) BLASTX search protein databases using a translated nucleotide query. [more...](#)

Query subrange

Or, upload file No file chosen

Genetic code

Job Title

Enter a descriptive title for your BLAST search

Align two or more sequences

BLAST results will be displayed in a new format by default. You can always switch back to the Traditional Results page. 

Choose Search Set

Database Non-redundant protein sequences (nr)

Organism exclude Enter organism name or id—completions will be suggested

Optional Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown.

Exclude Models (XM/XP) Non-redundant RefSeq proteins (WP) Uncultured/environmental sample sequences

Optional

BLAST

Search database nr using Blastx (search protein databases using a translated nucleotide query) Show results in a new window

Algorithm parameters

Step 2: Run BLASTx Web Tool

[Algorithm parameters](#) [Restore default search parameters](#)

General Parameters

Max target sequences: 100 

Select the maximum number of aligned sequences to display 

Expect threshold: 10 

Word size: 6 

Max matches in a query range: 0 

Scoring Parameters

Matrix: BLOSUM62 

Gap Costs: Existence: 11 Extension: 1 

Compositional adjustments: Conditional compositional score matrix adjustment 

Filters and Masking

Filter: Low complexity regions 

Mask: Mask for lookup table only 
 Mask lower case letters 

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.