MAKING WHOLE GENOME ALIGNMENTS USABLE FOR BIOLOGISTS.

EXAMPLES AND SAMPLE ANALYSES.

Table of Contents

Examples:

Introduction to Examples

While these examples can be followed using any regions of interest, the genomic regions covered by the *ABCB1 ATP-binding cassette, sub-family B (MDR/TAP), member 1* gene (commonly known as *MDR1*) will be used within this text. To import these coordinates into Galaxy, use the *UCSC Main table browser* tool.

1. Within the UCSC table browser interface, set
    a. *clade* to 'Mammal',
    b. *genome* to 'Human',
    c. *assembly* to 'Mar. 2006',
    d. *group* to 'Genes and Gene Prediction Tracks',
    e. *track* to 'RefSeq Genes',
    f. *table* to 'refGene' and
    g. *region* to 'genome'.
2. Click the *paste list* button associated with the *identifiers (names/accessions)* setting.
    a. Type 'ABCB1' (without the quotes – this is the gene's official symbol) into the text box that appears and

b. click submit; the output will now only include regions annotated as 'ABCB1'.

3. Make sure *Send output to Galaxy* is checked and that
4. the *output format* is set to BED and then
5. click *get output*.
6. On the new page that is displayed ensure that **one BED record per** *Whole Gene* is selected and then
7. click *Send query to Galaxy*.

A new dataset, *UCSC Main on Human: refGene (genome)*, appears in the history panel and will turn green when all data has been retrieved. A preview of the dataset can be expanded by clicking on the name or the entire dataset can be viewed by clicking on the eye icon. Doing either of these actions reveals a single genomic interval (in the BED12 format) that has 29 exons.

Example 1: Alignment Extractors

Example 1.1: Obtaining a subset of an alignment using the *Extract MAF blocks* tool.

Alignment blocks, from the 28-way alignment, which overlap the human *MDR1* gene are obtained.

1. Start with a set of genomic intervals, such as the coordinates of the *MDR1* gene.
2. In the tools menu, on the left hand side, click the heading *Fetch Alignments* to expand the tool section housing the multiple alignment tools.
3. The interface to the alignment extractor is accessed by clicking on the tool titled *Extract MAF blocks given a set of genomic intervals*, which appears within the newly expanded tool section.
4. Ensure that
    a. the genomic interval dataset (*UCSC Main on Human: refGene (genome)*) is selected under *Choose intervals* and
    b. *MAF Source* is set to 'Locally Cached Alignments'.
5. Under *Choose alignments*, select the 28-way multiZ alignment.
6. After the screen refreshes, the *Choose species* selector will have 28 different genome builds listed; this allows the optional exclusion of undesired species from the extracted MAF blocks.

Either use the *Select All* button to select all species or select no species (in this tool, not selecting any species is the same as selecting them all) to cause the extracted MAF blocks to have all the species that appear in the original source blocks.

7. Clicking *Execute* adds a new dataset to the history, which contains 2,084 MAF blocks that fall within the genomic coordinates of the human *MDR1* gene.

Example 1.2: Viewing alignments graphically

An interactive java-based multiple sequence alignment viewer, Gmaj, has been made available as a tool under the *Graph/Display Data* tool section. To view the extracted alignment blocks in context with the *MDR1* exons, use the *GMAJ Multiple Alignment Viewer* tool.

1. Under *Alignment File*, select the extracted MAF blocks and then click *Add new Annotations*.
2. New configuration settings will appear:
    a. set *Annotation Style* to 'Galaxy',
    b. *Species* to 'hg18' and
    c. select the *MDR1* genomic intervals dataset.
3. A new history item will appear; clicking the dataset's name will reveal a button *Launch GMAJ*. Clicking on this button will cause the Gmaj applet to load with the exon annotated alignment blocks.

Example 2: Format Converters

Example 2.1: Creating a FASTA alignment from a MAF alignment

As an example, a single FASTA alignment block will be generated from the multiple alignments that were obtained in the extract MAF blocks example (*Example 1.1*). All format converters are located under a single tool heading, *Convert Formats*, this is a different location than the majority of the MAF toolset.

1. Clicking *Convert Formats* on the left hand side will reveal the tool titled *MAF to FASTA Converts a MAF formated file to FASTA format*. Access this tool by clicking on its name.
2. Make sure that

a. *MAF file to convert* is set to the MAF blocks obtained in the previous example and

b. set *Type of FASTA output* to 'One Sequence per Species.'

c. Click the *Select All* button under *Species to extract* in order to include all 28 species in the output FASTA block and then

d. click *Execute*.

3. A new dataset, having a single FASTA alignment block with 28 sequences will appear. This file is ready to be used in downstream analyses such as building a phylogenetic tree.

Example 2.2: Building a phylogenetic tree

A phylogenetic tree is built from the alignment of the human *MDR1* gene region using HyPhy.

1. To build a neighbor joining tree with HyPhy, select the tool section entitled *Evolution* and select the tool named *Neighbor Joining Tree Builder*.
2. Select the newly created FASTA alignment under *Fasta file* and choose a desired distance model from the appropriate dropdown menu.
3. Clicking execute on this tool creates two new datasets in the history:

    a. one is a PDF with a graphical representation of the tree and

    b. the other is a text file with the tree in the Newick format.

Example 3: MAF Stitchers

Example 3.1: Stitching MAF blocks by genes

An example of this can be best demonstrated by following the same steps used in the convert MAF to FASTA example above, but this time using the *Stitch Gene blocks given a set of coding exon intervals* tool. The stitching tools are located under the *Fetch Alignments* tool section. This tool requires the input intervals to be of the BED12 format, as exon and coding region information is encoded in additional columns not required of all interval files. All of the intervals returned from the refGene table of the UCSC table browser, of which the *MDR1* gene intervals retrieved earlier belong to, contain this information. Within the *Stitch Gene blocks* tool,

1. set *Gene BED file* to the gene interval,

2. set *MAF Source* to 'Alignments in Your History', and then
3. select the alignment dataset which was created in the extract MAF blocks example under *MAF file*.
4. Click *Select All* under *Choose species* and then
5. click *Execute*.

Alternatively, it is possible to directly use the locally cached 28-way alignment, just as was done in the extract MAF blocks example; the resultant stitched datasets will contain equivalent data. The FASTA dataset generated by this example can now be used in the HyPhy neighbor joining tree tool and the results can be compared to those obtained from following *Example 2.2*.

Sample Analysis

Sample Analysis: Investigating alignments corresponding to human codons containing synonymous SNPs.

In this example, multiple-species alignments of human regions containing synonymous protein coding single nucleotide polymorphisms (SNPs) will be examined after filtering based upon the change in codon usage frequency. This analysis requires three different datasets: the genomic intervals and observed nucleotide bases for human SNPs, the genomic intervals (including coding exon information) of human genes, and also a table containing the nucleotide sequence and human codon usage frequency for each amino acid.

1. The first two datasets can be retrieved using the *UCSC main table browser* tool.
    a. A dataset containing all human genes can be retrieved by following the steps used in the *Introduction to Examples* to obtain the genomic interval for *MDR1*, but this time do not provide any *identifiers (names/accessions)*; be sure to *Clear* any identifiers which may already be set.
    b. To obtain a dataset containing all known human SNPs, access the UCSC table browser interface and
        i. set *clade* to 'Mammal',
        ii. *genome* to 'Human',
        iii. *assembly* to 'Mar. 2006',

iv. *group* to 'Variation and Repeats',

v. *track* to 'SNPs (130)',

vi. *table* to 'snp130',

vii. *region* to 'genome' and

viii. *output format* to 'all fields from selected table'.

ix. Make sure *Send output to Galaxy* is checked and then click *get output*.

x. On the new page that is displayed click *Send query to Galaxy*. A new dataset, *UCSC Main on Human: snp130 (genome)*, appears in the history panel and will turn green when all data has been retrieved.

xi. UCSC informs Galaxy that this dataset is plain tabular data, but this dataset contains all the information required of genomic intervals. This dataset can be changed to genomic intervals by

1. clicking on the pencil icon and
    a. setting *New Type* under **Change data type** to 'interval' and
    b. clicking *Save*.
2. Make sure column assignment metadata is properly set (columns are 2, 3, 4, 7 and 5 for chrom, start, end, strand, and name columns, respectively).

c. The last dataset needed, codon usage frequencies, can be obtained from the Galaxy dataset library *Codon Usage Frequencies* by

i. clicking on the *Data Library* link in the masthead of the Galaxy interface and selecting the library by name.

ii. Import the library item named *Human Codon Usage Frequencies* by filling in the checkbox next to the name and clicking the *Go* button next to *Import into your current history*.

iii. Information about this file, including the original source, can be accessed by clicking on the library item's name before importing.

2. Once all the datasets have been loaded into the history, it is time to start the analysis. The first step is to filter out any SNPs that are not exactly one base long. This is done using the *Filter*

data on any column using simple expressions tool found under the tool section *Filter and Sort*:

   a. set *Filter* to the SNP dataset and

   b. enter 'c3 + 1 == c4' (without the quotes) for *With following condition*; only genomic intervals with start position plus one equaling the end position will be included.

3. While the history has the genomic intervals for all human genes, it is really the genomic intervals for each codon that is needed. The codons can be obtained from the genomic intervals by using the tool *Gene BED To Exon/Intron/Codon BED expander* found in the *Extract Features* tool section.

   a. Set *Extract* to 'Codons' and

   b. *from* to the human genes dataset. The extended information of the BED12 format is used to generate a new dataset that contains all codons composed of nucleotides that appear adjacently within the human genome (codons crossing splice points cannot be represented with a single start and stop position and are discarded by this tool).

4. Next, the *Extract Genomic DNA using coordinates from assembled/unassembled genomes* tool, found in the *Fetch Sequences* tool section, is used to obtain the coding sequences for the human codons.

   a. Set *Fetch sequences corresponding to Query* to the codon-containing genomic interval dataset and

   b. select 'interval' as the *Output data type*.

   c. The result is a new set of genomic intervals, exactly like the one for codons, except that a new column containing the genomic DNA sequence for each codon has been appended.

5. Now it is time to determine the codon usage frequencies for each of the codons, this is done by joining the codon sequence containing genomic intervals with the codon usage frequency table obtained earlier from the Galaxy library.

   a. Before this join can occur, it will be necessary to ensure that all the sequence characters extracted in the last step are uppercase. This is done using the *Compute an expression on every row* tool found in the *Text Manipulation* tool section.

      i. Set *Add expression* to 'c7.upper()' (without the quotes, c7 is the column which

contains the sequence information),

ii. *as a new column to* to the dataset created in the previous step, and

iii. leave *Round result?* as 'NO'.

b. Now the genomic interval data is ready to be joined to the codon usage frequency data. This is accomplished using the *Join two Queries side by side on a specified field* tool, which is found in the *Join, Subtract and Group* section.

i. Set *Join* to the genomic intervals containing all uppercase codon sequences,

ii. *using column* to 'c8' (the uppercased extracted sequences),

iii. *with* to the codon usage frequency table,

iv. *and column* to 'c2' (the column containing the nucleotide sequence for each codon) and

v. leave *Keep lines of first input that do not join with second input*, *Keep lines of first input that are incomplete* and *Fill empty columns* all set to 'No'.

vi. The resultant dataset now contains the genomic intervals of codons coupled with sequence and codon usage frequency information. A keen eye will notice that a handful of codons were lost in this process, this is due to the presence of ambiguous nucleotides (i.e. 'n's) in the human genome reference sequence (hg18 in this case).

6. Now that there is a dataset that contains the genomic intervals, sequences, and usage frequencies for human codons in the history, it is time to join the codons to overlapping SNPs. This is done using a different version of the join tool used earlier, named *Join the intervals of two queries side-by-side* and found under the *Operate on Genomic Intervals* tool section. This tool works by joining on overlapping genomic regions (using three columns: chromosome, start and end), whereas the previous join worked by comparing the contents of a single user-specified identifier column.

a. Set *Join* to the dataset that was created in the previous step,

b. *with* to the dataset containing human SNPs with a length of one,

c. *with min overlap* to '1' and

d. *Return* to 'Only records that are joined (INNER JOIN)'. The new dataset will have only

those codons which overlap with a provided SNP.

7. The next step is to determine the sequence of the SNP containing codon. This is done using the *Mutate Codons with SNPs* tool found in the *Evolution* tool section. This tool will create a new dataset which contains the genomic intervals having different codons than were obtained from the reference genome and which are based upon the nucleotides provided by the *observed* column of the SNP data.
    a. Set *Interval file with joined SNPs* to the dataset created in the last step,
    b. *Codon Sequence column* to 'c8',
    c. *SNP chromosome column* to 'c17',
    d. *SNP start column* to 'c18',
    e. *SNP end column* to 'c19',
    f. *SNP strand column* to 'c22' and
    g. *SNP observed column* to 'c25'.
8. Now it is possible to join the SNP codons with the codon frequency table, as was done in step 6, but this time use 'c34' instead of 'c8'. After this join is complete, the *Filter data* tool is used to remove any non-synonymous codons. In the filter tool, set *With following condition* to 'c9 == c35' (without quotes), this will cause the new dataset to only have genomic intervals where the amino acid for the human reference genome sequence matches the amino acid of the SNP substituted sequence (removes non-synonymous codons).
9. The *Compute* tool is used to determine the ratio of the two codon usage frequencies by setting *Add expression* to 'c15 / c41'. After using the *Count* tool to calculate the counts of each occurrence of the ratios, the cutoffs for the upper and lower approximate 5% of the data was determined (the ratios are naturally binned into discrete values so obtaining exact 5% cutoffs was not possible). The upper and lower cutoff values are used in the *Filter data* tool by setting *With following condition* to 'c42 <= 0.360575241042 or c42 >= 3.06607911917' (without quotes).
10. Due to multiple SNPs overlapping individual codons, the filtered output should be collapsed to only include each codon once. This can be accomplished using the *Count occurrences of each record* tool by selecting the columns 'c1', 'c2', 'c3' and 'c6' (the chromosome, start, end and

strand columns, respectively); the resultant dataset will have 5 fields: the number of times the codon was found, the chromosome, the start position, the end position and the strand. The output of this tool is tabular data, so it is necessary to change the datatype to 'interval' (as was done for the original SNP data retrieved from UCSC) and set the column assignment metadata to the proper values.

11. Now the *Extract MAF blocks* tool is used on this new dataset.
    a. Set *MAF Source* to 'Locally Cached Alignments' and
    b. *Choose Alignments* to the 44-way alignment.
    c. Limit the species in the extracted MAF blocks by selecting only 'hg18', 'panTro2', 'rheMac2', 'mm9' and 'canFam2' under *Choose species*.
12. The *Join MAF blocks by Species* tool is now used to rejoin MAF blocks which may have been disrupted by rearrangements local to an excluded species. Use the *Select All* button to include each of the five species requested in the last step.
13. MAF blocks are again extracted, using the same genomic intervals as in step 11, but this time the joined-by-species MAF alignment set (created in step 12) is used as the MAF source.
14. The last three steps result in an alignment set which is not broken up by rearrangements occurring in species that are not of interest to this analysis and the majority of the blocks will have a length of 3 (matching the human codon locations). The Convert *MAF to Interval* tool is now used to create a set of genomic intervals from the extracted-joined-extracted alignment set. Only the genomic intervals for human are needed, so
    a. leave *Select additional species* blank,
    b. set *Excluded blocks which have missing species* to 'include blocks with missing species' and
    c. *Remove Gap characters from sequences* to 'keep gaps'.
15. The genomic intervals join tool is now used to join the SNP-codon interval dataset created during step 9, where the ratio of codon usage frequencies was calculated *with* the interval file created in the last step (from converting MAF blocks to intervals).
16. The *Filter Data* tool is used to make sure that the codon interval positions exactly match the intervals obtained from the extracted MAF blocks (rearrangements between the included

species can cause these position to not match). Filter the joined SNP-codon-MAF dataset by using the expression 'c1 == c43 and c2 == c44 and c3 == c45 and c6 == c46' (without the quotes; this is checking that the chromosomes, starts, ends and strands are equal).

The resultant datasets contains the genomic interval information and aligned sequences for each of the codons containing synonymous SNPs that have a condon usage ratio which met the specified filtering criteria. Further steps that can be undertaken on this dataset, using only the already available data and the tools discussed, include determining whether the sequences of aligned species could encode synonymous or non-synonymous codons.

Secondary Sample Analysis: Computing Lineage-specific mutation rates

Demonstrating research reproducibility is not only important for any technique or methodology, but also a primary goal of the Galaxy project. We include this sample analysis to demonstrate the ability of this toolset reproduce published results. We will determine the genomic locations and aligned sequences corresponding to ancestral repeats. Using the information provided by a multiple-species whole genome alignment, we will determine which repeats are purportedly ancestral and also calculate lineage-specific substitution rates between dog, human and mouse. The results generated here agree favorably with those generated in Lindblad-Toh, et al. Genome sequence, comparative analysis and haplotype structure of the domestic dog. Nature. 2005. 438:803-819.

1.) The first step is to obtain locations of repeats. A common repeat caller, RepeatMasker, is a part of the UCSC pipeline and tracks are available for most genome builds found at the UCSC table browser. Within the UCSC table browser interface, set

   a. *clade* to 'Mammal',

   b. *genome* to 'Human',

   c. *assembly* to 'Mar. 2006',

   d. *group* to 'Variation and Repeats',

e. *track* to 'RepeatMasker',

f. *table* to 'rmsk' and

g. *region* to 'genome' or a region of interest; for the case of brevity we will restrict this example to chromosome 1.

2.) Alignment blocks, from the 28-way alignment, which overlap the human repeats are obtained.

a. Start with the genomic intervals corresponding to human repeats.

b. In the tools menu, on the left hand side, click the heading *Fetch Alignments* to expand the tool section housing the multiple alignment tools.

c. The interface to the alignment extractor is accessed by clicking on the tool titled *Extract MAF blocks given a set of genomic intervals*, which appears within the newly expanded tool section.

d. Ensure that

i. the genomic interval dataset (*UCSC Main on Human: rmsk (chr21:1-46944323)*) is selected under *Choose intervals* and

ii. *MAF Source* is set to 'Locally Cached Alignments'.

e. Under *Choose alignments*, select the 28-way multiZ alignment.

f. After the screen refreshes, the *Choose species* selector will have 28 different genome builds listed; this allows the optional exclusion of undesired species from the extracted MAF blocks. Select the species that will be used to determine the ancestral status of the repeats

g. Clicking *Execute* adds a new dataset to the history, which contains MAF blocks that fall within the genomic coordinates of the human annotated repeats.

3.) The *Filter MAF blocks by Species* tool is now used to remove MAF blocks which lack a desired species and are therefore not considered ancestral. One possible set is human, mouse, dog and chicken (hg18, mm8 and canFam2).

4.) The filtered MAF blocks are then converted to the genomic interval format for the aligning human repeat regions using the *MAF to Interval* tool.

5.) The *Coverage of a set of intervals on a second set of intervals* tool is used to determine the

portion of the original repeat genomic intervals (step 1, retrieved from UCSC) that are covered by the filtered ancestral repeat regions.

6.) Using, the filter tool, we only keep the repeats that have 50% coverage or better (c8>=0.5).

7.) Using the coverage-filtered repeats we extract MAF blocks for human, dog and mouse from the 28-way alignment.

8.) These MAF blocks are then converted to the FASTA format using the concatenate option.

9.) The *Branch Lengths Estimation* tool is used on the FASTA file containing the purported ancestral repeat alignment data. These results for chromosome 1 are presented in the table here and agree favorably with those presented in Figure 3 in Linblad-Toh et al. 2005 where sliding windows across known ancestral repeats were used.

Branch	**Length**	**LowerBound**	**UpperBound**
canFam2	0.227208	0.22686	0.227549
hg18	0.112318	0.112036	0.112601
mm8	0.392244	0.391749	0.392743
Total Tree	0.73177	0.731149	0.732371

Lineage-specific mutation rates based upon ancestral repeats found on human chromosome 1