Genome alignment & gene annotation

1. Prepare the working directory

1.1 Create a working directory "/workdir/$USER/genome_aln".

Download the data files to your working directory. We will use two DNA sequence files provided by MUMmer for this exercise.

```
mkdir -p /workdir/$USER/genome_aln
cd /workdir/$USER/genome_aln
wget http://mummer.sourceforge.net/examples/data/H_pylori26695_Eslice.fasta
wget http://mummer.sourceforge.net/examples/data/H_pyloriJ99_Eslice.fasta
```

2. MUMmer

2.1 Run genome alignment using nucmer and promer

- The "export PATH ..." command adds the executable files of MUMmer package into the PATH, so that you can execute the tools;
- The "nucmer" commands align two large DNA sequences. "-c" sets the minimum cluster size, "-p" sets the prefix of output file names.
- The "promer" command generates amino acid alignments between two DNA input files (by 6-frame translation).

```
export PATH=/programs/mummer-4.0.0beta2/bin:$PATH
nucmer  -c 100 -p  pylori_100  H_pylori26695_Eslice.fasta
H_pyloriJ99_Eslice.fasta
nucmer  -c 250 -p pylori_250 H_pylori26695_Eslice.fasta H_pyloriJ99_Eslice.fasta
promer -p promer_100 -c 100  H_pylori26695_Eslice.fasta H_pyloriJ99_Eslice.fasta
```

The output are "*.delta" files. The MUMmer software uses the "delta" file format to represent the all-vs-all alignment between the input sequences. You do not need to know the details of the "delta" format, MUMmer provides a "show-coords" tool to show alignment coordinates in the "delta" files.

```
show-coords pylori_100.delta
show-coords pylori_250.delta
show-coords promer_100.delta
```

Press "space" to continue, press "q" to exit. "S1" and "E1" are "start" and "end" positions of the match on the first sequence. "S2" and "E2" are "start" and "end" positions of the match on the 2nd sequence. "LEN" is the length of the match, "% IDY" percent identity of the alignment.

2.2 Generate reports from the delta files.

```
dnadiff   -d  pylori_100.delta  -p  pylori_100
dnadiff   -d  pylori_250.delta  -p  pylori_250
```

These commands will generate quite a few report files for each delta file. "-d" for input file name, "-p" for prefix of the output file names. The dnadiff command does not work for the delta file created from promer.

If your directory has too many files, it might be difficult to tell which files are newly generated by the software. This "ls -lrt" command would order the files based on the time. The new files are at the end of the list.

```
ls -lrt
```

The results are all text files, you can check the contents of the output file with the "less" command.

Output files are
out.report - Summary of alignments, differences and SNPs
out.delta - Standard nucmer alignment output
out.1delta - 1-to-1 alignment from delta-filter -1
out.mdelta - M-to-M alignment from delta-filter -m
out.1coords - 1-to-1 coordinates from show-coords -THrcl .1delta
out.mcoords - M-to-M coordinates from show-coords -THrcl .mdelta
out.snps - SNPs from show-snps -rlTHC .1delta
out.rdiff - Classified ref breakpoints from show-diff -rH .mdelta
out.qdiff - Classified qry breakpoints from show-diff -qH .mdelta
out.unref - Unaligned reference sequence IDs and lengths
out.unqry - Unaligned query sequence IDs and lengths

For more details see https://github.com/marbl/MUMmer3/blob/master/docs/dnadiff.README

2.3 Generate dot plots using delta files

Three dot plots will be created using the delta files from step 2.1.

```
mummerplot   -png -p   pylori_100    pylori_100.delta
mummerplot   -png -p   pylori_250    pylori_250.delta
mummerplot   -png -p   pylori_promer100    promer_100.delta
```

Run "ls -lrt", and you will find three new ".png" files. Use "Filezilla" to download the .png files to you laptop, and double click to open.

Comparing the three plots, you would find:

- "-c" specify the minimum cluster length. Increase this number would decrease the sensitivity, but make the run faster.
- Promer has a much higher sensitivity than nucmer.

3. Minimap2

3.1 genome alignment with minimap2

- "-c": Include CIGAR in the output file. CIGAR is a string to define indels in a sequence match (https://sites.google.com/site/bioinformaticsremarks/bioinfo/sam-bam-format/what-is-a-cigar)
- "-x asm10": Preset code to "asm10", which is "assembly" to "reference genome" mapping, with 1% sequence divergence. "-cx asm10" is equivalent to "-c -x asm10".
- "--cs=long" include sequence alignment strings in the output. This would make output files much larger. For large sequence alignment, you might want to skip this parameter.

```
/programs/minimap2-2.17/minimap2 -cx asm10 --cs=long H_pylori26695_Eslice.fasta
H_pyloriJ99_Eslice.fasta  >minimap2.paf
```

Minimap2 produce a "paf" formatted file. It is a tab-delimited text file. You can use "less" to examine the file "minimap2.paf".

Col	Type	Description
1	string	Query sequence name
2	int	Query sequence length
3	int	Query start coordinate (0-based)
4	int	Query end coordinate (0-based)
5	char	'+' if query/target on the same strand; '-' if opposite
6	string	Target sequence name
7	int	Target sequence length
8	int	Target start coordinate on the original strand
9	int	Target end coordinate on the original strand
10	int	Number of matching bases in the mapping
11	int	Number bases, including gaps, in the mapping
12	int	Mapping quality (0-255 with 255 for missing)

3.2 Use "paftools.js" to analyze the paf file.

"paftools.js" is a tool distributed with Minimap2. It can be used for converting the "paf" to other formats, generating statistics report, et al.

```
export PATH=/programs/minimap2-2.17/misc:$PATH

#print the paftools help menu
paftools.js

#generate statistic report of the alignment
paftools.js stat minimap2.paf

#convert paf to blast file
paftools.js view minimap2.paf > minimap2.blast.txt
less minimap2.blast.txt
```

3.3 Generate a dot plot

Use the R script pafCoordsDotPlotly.R (https://github.com/tpoorten/dotPlotly) to create a dot plot from the minimap2 result.

- -p 5: plot size 5x5 inches;
- -i: input paf file name;
- -o: output file name;
- -m 1: filter out alignments with lengths smaller than cutoff value. Set "-m 1" to turn off this filter.
- -q 1: filter queries with total alignments less than cutoff value. Set "-q 1" to turn off this filter.

```
/shared_data/alignment2020/pafCoordsDotPlotly.R  -p 5 -i minimap2.paf -o
minimap2.png  -m 1  -q 1
```

Use Filezilla to download the minimap2.png to your laptop and open the image file.

4. Gene annotation

Gene Ontology (GO) annotations are used in RNA-seq and Chip-seq data analysis. In this section, we will work on how to retrieve/create GO annotations.

4.1 Download GO annotations from Ensembl BioMart.

The Ensembl database has GO annotations for many species. Its web site provides a BioMart tool to download data. Here we will use "mouse" as an example.

1. On your laptop, use a web browser to open the web site http://ensembl.org/ . (For plant species, the URL is http://plants.ensembl.org/);
2. Click "BioMart", then select "Ensembl Gene xxx" and "Mouse genes";
3. You will need to set "Filters" and "Attributes". "Filters" define the rows of the output table, and "Attributes" define the columns in the output table.
    - Filters: Set "Filters" only if you want to retrieve data for a subset of genes. In this case, we want all genes, so no "filters".
    - Attributes: Click "Attributes" in the left panel. First, expand the "External" tab and check the boxes for "GO term accession" and "Go term name". Then you need to define what type of gene IDs to use in the output file. Expand the "GENE" tab and uncheck all boxes except "Gene stable ID".

(BioMart provides many different types of Gene IDs for you to choose from. Some are under the "GENE" tab. Others, like Refseq accessions, are under the tab "External references")

4. Click "Results" at the top of left panel. You would see top 10 rows of the output table. Check the box "Unique results only" and click "Go" to download the whole file. Some times, if it takes more than 5 minutes to download, you will need to choose the option to get file through Email (From the menu "Export all results to", select "Compressed web file - notify by email"). You can open the downloaded file "mart_export.txt" in Excel. Examine the results.

4.2. Generate GO for a custom genome

Quite often, you cannot find GO annotations for the genome you are working on. In that case, you will have to generate GO annotations by yourself. Two software can be used: InterProScan and Blast2GO. You can run either one of them, or you an run both, and Blast2GO has a function to integrate both results into one file. Integrating InterproScan results into Blast2GO is optional, it would increate the number of annotated gene by a small percentage.

You will not do this part by yourself, as it would not fit into the training computer you are using. The detailed instructions and commands can be found at https://biohpc.cornell.edu/lab/userguide.aspx?a=software&i=73#c .

Here are the steps. Run step 1 & 2 on any BioHPC computer in the categories "medium memory gen2" or "large memory gen2". Run step 3 on the machine "cbsumm10".

Step 1. Run DIAMOND. This step takes less than 20 minutes for 10,000 proteins. The output file is "blastresults.xml".

Step 2. (Optional) Run InterproScan. It takes about 1 hour per 10,000 proteins. The output file is "ipsout.xml".

Step 3. Generate GO annotation with Blast2GO. It takes about 30 minutes for 10,000 proteins.

After the three steps, there are three files you want to save: *.pdf, *.b2g and *.annot. You can find BLAST2GO results for the honey bee genes in the directory /shared_data/alignment2020/project2 .

4.3. Post-process the Blast2GO results

First copy the three Blast2GO result files to your working directory:

```
cp /shared_data/alignment2020/project2/myresult.annot ./
ls -lrt
```

Then you need to check the Blast2GO report "myresult.pdf" file, and find out what percentage of genes CANNOT be annotated. I would like to see the percentage below 30-40%. Download the PDF file to your laptop using Filezilla, open the file, and find these lines:

Annotation summary:

Annotated sequences: 21511
Sequences that could not be annotated: 1960
Assigned Gene Ontology terms: 140105
Assigned enzyme codes: 6815
Sequences with enzyme codes assigned: 5471

In this case, the un-annotated gene percentage is 1960/(21511+1960)=8%. So that is really good.

If too many genes cannot be annotated, you might need to re-run Blast2GO with modified setting. Including:

- Run InterProScan if you have not done so;
- When running DIAMOND, increase the number for --max-target-seqs to 200, and the evalue cutoff to "-e-5".

You might need to re-format the *.annot file for it to work in some software.

- The BLAST2GO software is primarily written for Windows. Its output file has different line terminators from a Linux text file. You need to convert the BLAST2GO output to Linux/Mac style text file. The tool is "dos2unix".

  In the following commands, you will run "file myresult.annot" before and after "dos2unit". "file" is a Linux tool to check file type. Do you see any difference? Before "dos2unix", you would see "with CRLF line terminators". *CRLF* is a Windows line terminator.

```
cp /shared_data/alignment2020/project2/myresult.annot ./

file myresult.annot

dos2unix myresult.annot

file myresult.annot
```

- To use "myresult.annot in" Bioconductor R package "topgo", it needs to be reformatted. Read the software manual to make sure that you are using the right format. Here I am using the tool "toTopGO.pl" to reformats "myresult.annot" into a new file named "topgoAnnot".

```
/shared_data/annotation2019_2/toTopGO.pl myresult.annot

less myresult.annot

less topgoAnnot
```

Check the difference between the two files: myresult.annot and topgoAnnot. You would find that in the original file myresult.annot, a gene has multiple lines. In the topgoAnnot file, each gene has only one line but with multiple GO accessions.