**Annotating Genomic Regions of Interest and Enrichment Analysis**

Mackenzie R. Gavery

University of Washington

School of Aquatic and Fishery Sciences

mackenzie.gavery@noaa.gov

ConGen lecture/worksheet, 12 September 2018

**Introduction**

We’re entering an era when reference genomes are becoming available for many non-model organisms. Many questions being addressed in conservation genetics/genomics studies can now utilize annotated reference genomes to dive deeper into the biological processes associated with genomic regions of interest identified through genetic, transcriptomic and epigenetic studies.

Often, one of the challenges in making biological interpretations of genomics data is the ability to annotate regions of interest or to functionally summarize genes of interest at a genome-wide scale. Luckily, there are many tools available to address these challenges. The objective of the following exercises is to introduce steps commonly used to annotate regions of interest at a genome-wide scale. This includes the very important (and often overlooked) ability to easily manipulate very large text files.

Here we will first perform an exercise to functionally annotate differentially methylated regions (DMRs) identified in an epigenetics study comparing two cell-types (red blood cells (RBCs) and sperm) in steelhead (*O. mykiss*). We will accomplish this by performing ‘genome arithmetic’ on widely-used genomic file formats (i.e. BED and GFF). Next, we will use the online tool DAVID to identify enriched biological processes associated with our gene list. Finally, we will use a genome browser to visualize our regions of interest relative to genes.

**Useful Definitions**

* *GFF file*: General Feature Format is a file format used for describing genes and other regions of DNA (<https://uswest.ensembl.org/info/website/upload/gff.html>)
* *BED file*: Browser Extensible Data or BED format is similar to GFF, but only requires 4 fields (<https://www.ensembl.org/info/website/upload/bed.html>)
* *Gene Ontology (GO) Enrichment analysis*: gene ontology (GO) terms are commonly used to perform enrichment analysis on gene sets. For example, given a set of genes that are differentially regulated or methylated under certain conditions, an enrichment analysis will find which GO terms are over-represented using annotations for that gene set.

**Software**

* *Bedtools*: the self-proclaimed ‘swiss army knife of genome arithmetic’ <https://github.com/arq5x/bedtools2>
* *DAVID*: Database for Annotation, Visualization and Integrated Discovery (useful background: <https://david.ncifcrf.gov/content.jsp?file=functional_annotation.html>)
* *Integrated Genome Viewer (IGV)*: a high-performance visualization tool for interactive exploration of large, integrated genomic datasets <https://software.broadinstitute.org/software/igv/>

**Exercises**

**Annotate regions of interest**

We will annotate DMRs in 3 steps. First, we will filter and reformat the output of methylKit (R package to identify differentially methylated regions) to make a BED file of DMRs. Next, we will annotate genes associated with the DMRs using bedtools. Finally, we will merge the gene numbers to the UniProt gene information by joining the two tables based on a common field.

Insider information: The most common task I perform in terms of bioinformatics is reformatting and merging very large tables. I like using R Studio for these types of tasks because I can easily save, run and modify scripts that I used frequently. Today we will use R studio to reformat, filter and join tables. Bedtools can be run from the command line, but today we will run it from our R script using the ‘system’ command. This is a great way to keep commonly used pipelines all in one script. Details of each step are in the script (Congen\_Gavery\_2018.R, which you should open now), but more information about each step is below:

1. **Preparing a BED file**

We need to reformat the DMR output file to a genomic coordinate format called a BED (Browser Extensible Data). We will use the command line to reformat the output to contain the 3 required fields: ‘chrom’, ‘chromStart’ and ‘chromEnd’ and the optional 4th field ‘name’. BED files typically need to be sorted by chromosome then start position before they can be used for downstream applications like bedtools or visualized in genome browsers.

1. **Using genomic arithmetic to annotate regions of interest**

We will use the software bedtools to identify genes in close proximity to the DMRs. Bedtools has many incredibly useful utilities for performing ‘genomic arithmetic’. For this exercise we will use the *closest* command to find the gene closest to each DMR in our BED file. The file containing the position and annotation of genes is a GFF (General Feature Format) file. The *-D* option says we want to report the distance in base pairs relative to the gene file (e.g. a negative distance will indicate the DMR is upstream of the gene).

1. **Joining tables and filtering based on distance and quality of annotation**

We will need to filter based on distance (+/- 10kb) and high-quality annotations (i.e. e-value). The quality of the annotation is derived from the BLAST output. This is a very important aspect of gene annotation that will not be discussed today. An overview and exercises for using BLAST to annotate a transcriptome can be found here: <http://sfg.stanford.edu/BLAST.html>

**Enrichment analysis**

We will use the publicly available, online tool DAVID for enrichment analysis. One of the benefits of DAVID is that it can perform GO annotation based on the gene ID. Annotating to the GO database yourself takes up a lot of space and time, so this is great. The other benefit of DAVID that is so useful for non-model species is that it will accept a mixed-species gene list.

1. Browser to DAVID: <https://david.ncifcrf.gov/>
2. Select ‘Start Analysis’ from Toolbar
   1. Step 1: ‘Enter Gene List: Paste your gene list (DMR\_UNIPROT\_ID) into the space provided
   2. Step2: ‘Select Identifier’: Select ‘UNIPROT\_ID’ from the pull-down
   3. Step 3: ‘List Type’: Select ‘Gene List’
   4. Click ‘Submit List’
   5. Select ‘OK’ after reading the note: *“Please note that multiple species have been detected in your gene list […] As a default, all species in your list will be used for analysis. Also note that you may need to select an appropriate background under the "BACKGROUNDS" tab in the manager to the left. By default, the background corresponding to the first species in the list will be selected...”*
3. Explore the output:



* 1. With ‘Check Defaults’ selected, explore the ‘Combined View’ charts and tables
     1. Functional Annotation Clustering: <https://david.ncifcrf.gov/helps/functional_annotation.html#E4>
     2. Functional Annotation Chart: <https://david.ncifcrf.gov/helps/functional_annotation.html#E3>
     3. Functional Annotation Table: gene-centric view (no stats)
  2. Alternatively, if you were just interested in the enriched GO Biological Processes, click the ‘+’ sign next to Gene\_Ontology and click on the Chart for GOTERM\_BP\_DIRECT
     1. Explore this output by clicking the ‘+’ sign and changing some of the options
     2. Click on the list of genes that are ‘not in the output’ to see which genes are not included in the enriched processes

**Visualizing data using a genome browser**

In the world of big data, it’s nice to be able *look* at data. For this exercise, we’ll be using IGV to visualize DMRs relative to genes and CG islands. We will use the command line to open IGV then use the java-based software to load in the genome, genes, CG islands and DMRs. Note: I have created subset of the *O. mykiss* genome and annotation files limited to those that are in our DMRs due to size of files, working with whole genomes can take a little longer.

Navigate to the IGV software folder then Execute:

java -Xmx750m -jar igv.jar

Now we can load in our files.

1. Load the genome
   1. click, on ‘Genomes’ from tool bar then select ‘Load Genome From File’ and select the genome file (“Omy\_scaffolds\_subset.fa)

(NOTE: IGV will automatically create an index file for the genome (.fai) in the same folder)

1. Load the ‘tracks’ of annotations: genes, CG islands and DMRs
   1. click on ‘File’ from tool bar then select ‘Load from File’ and select the following 3 files: Omy\_annotations\_subset.gff3, Omy\_CGislands.gff, DMRs\_hypo.sorted.bed
2. View a single scaffold by selecting ‘scaffold\_244’ from the pull-down list then look in more detail
   1. Right click on the gene track to change the view to ‘Expanded’ to see exons
   2. Zoom in on the region containing the DMRs by right-clicking the ruler region and selecting a small region
3. Load a SAM file to view the mapped bisulfite reads
   1. click on ‘File’ from tool bar then select ‘Load from File’ and select the following SAM file ‘RBC\_bisulfite\_subset.sam’
   2. Click ‘OK’ to have IGV create an index file
   3. Zoom in on a region containing a DMR to see individual reads
   4. IGV infers the SAM is bisulfite data and will colors the alignment in ‘bisulfite mode’. This view highlights all the CG dinucleotides and shows methylated CGs in red. Note a few things in this view:
      1. The coverage track is useful to see areas of deep coverage. It summarizes read depth and proportion methylated cytosines
      2. The bisulfite data was generated using a method called RRBS. Notice how a majority of the reads overlap with CG islands. This is by design to enrich for regions that might be regulated by DNA methylation.