Harvard School of Public Health

*Department of Biostatistics*

BST 280: Introductory Genomics & Bioinformatics for Health Research

Fall 2017

**Homework Assignment #1**

Due November 20 – to be submitted through the course drop box or by e-mailing it to   
Conor Foley, cof090@mail.harvard.edu

One of the surprising things, more than fifteen years after the human genome was announced as complete, is that the genome sequence is still evolving. While there is an accepted human genome consensus build, one has to be careful to work with the current version (or fix the version used among collaborators) or risk making unintended errors.

Even though the major data resources such as NCBI, EBI, and UCSC use a consensus standard for the genome, what appears as the annotation is not standardized. While nearly all of of the canonical genes can be found at each site, each has its own version of the genome annotation, and each differs, sometimes dramatically. And at each site, different version of the genome can have different annotation sources or different resources mapped to it.

The goal of this exercise to explore some of that complexity because being aware of it is extremely important if you are going to conduct research using genomic data.

Chronic obstructive pulmonary disease (COPD) is a disease of the lungs, most often caused by tobacco smoking, in which the airways become narrower, leading to shortness of breath. COPD is a major cause of death and disability. Chronic bronchitis and emphysema are both forms of COPD. Like many complex diseases, COPD has both genetic and environmental components. Today we are going to look at a genetic marker associated with the disease, *despite the fact that the greatest risk factor is cigarette smoking rather than genetics.*

The most common genetic markers used today are single nuclear polymorphisms (SNPs), single base changes in DNA that can studied in large populations to determine if its presence is associated with physical traits such as disease.

We’ll start with preliminaries to give you the necessary background, then move on to the assignment.

**Now, the assignment. Questions are in orange and numbered below. Please include a screenshot**

**We’ll start with a quick PubMed Search** to give you background. Go to PubMed and do a Boolean search for COPD and FAM13

1. **What are the three most recent publications associating COPD and FAM13?**

**Next, we’ll look at a resource we haven’t covered, SNPedia (http://www.snpedia.com).** If you search SNPedia for COPD, you will find references to many SNPs, including rs7671167. This SNP has an association with a gene, FAM13A, although not much is known about how it that gene influences COPD (<http://www.snpedia.com/index.php/Rs7671167>). **This SNP will be important later.**

1. **On which chromosome is rs7671167 located? Paste a SNPedia screenshot below.**

**Searching at NCBI and use the Genome Data Viewer (https://www.ncbi.nlm.nih.gov/genome/gdv/)**

Searching is an art. Start by doing a few simple queries to see whether FAM13A’s association with COPD has made it to the Genome Data Viewer.

* **Go to NCBI Genome Data Viewer** ` and find the link for Homo sapiens GRCh38.p11. Note that this genome version is different from the “version 10 patch” (GRCh38.p10) at the other sites, which I do not think affects the rest of this assignment.
* **Enter “lung disease” into the query box and search. Also try “COPD**.” Look in the list of genes and notice that FAM13A doesn’t appear there.
* **Try searching for FAM13A.**
  + Grab a screenshot of the region and find the genomic length of the FAM13A gene by mousing over the gene in the display. (This is an exceptionally long gene.)

1. **How many basepairs in length is the FAM13 gene? (Include a screenshot.)**

* We are interested in SNP **rs7671167.** You will notice that there are way too many SNPs to find it from this display, so try searching for **rs7671167** (see the search box on the left-hand side)

1. **Does this SNP fall within an intron or exon? (Include a screenshot.)**

Now go to **dbSNP** and search for **rs7671167.** You will notice that the “old view” comes up with a link to the new Report Page**. Please use the New Report Page. You will notice that this borrows from the Genome Data Viewer (but uses an older genome patch as the default).**

1. **Find the three most recent publications on the role of this SNP in COPD.**

**Searching using the UCSC genome browser**

**Go to the UCSC genome website**, open the human genome browser, **pick the latest genome build**, and search using the terms “lung disease” and “COPD.”

1. **What genes are associated with lung disease? Is FAM13A on the list?**
2. **What genes are associated with COPD? Is FAM13A on the list?**

**Go to the UCSC genome website**, open the human genome browser, **pick the latest genome build**, and search using the term “FAM13A.”

1. **Look at the results and find the RefSeq gene representation (RefSeq is the old name for the Gene database). Find its genomic location.**

**Go to the UCSC genome website**, open the human genome browser, **pick the latest genome build**, and search for the SNP **rs7671167**

Use the genome coordinates and the **Table Browser** to search in the region; use the Table Brower filters to help narrow your search. There is a video that may help you at <https://genome.ucsc.edu/training/vids/#vid04>.

1. **Use the Table Browser to find the SNPs around rs7671167.**

**Finding Genes using the UCSC genome browser**

The goal of this part of the exercise is to identify **two genes** that you might want to investigate in the region of rs7671167 under the hypothesis that these genes might play a role in COPD. The minimum criteria for these genes is

* 1. They have more than 1 exon (single exon genes are often, not always, pseudogenes)
  2. They are protein coding genes.
  3. There is a mouse homolog

Use the UCSC genome browser as the source for your information. It has more links to other data sources which will help you fill out the details. Using the **Table Broswer** and use of filters to generate a list.

NOTE: If you type in a name in the output file text box, the data will be exported as a text file which can be imported into Excel. Sorting columns in Excel will make it easier to answer the questions at the end of the exercise. You can also use the genome browser and display options for Gencode or RefSeq to filter out non-coding genes.

1. **How many Gencode genes are within a 1 MB region of your reference SNP?**
2. **Of the Gencode genes, how many are protein coding?**
3. **How many RefSeq genes are within a 1 MB region of your reference SNP?**
4. **How many of the RefSeq genes appear to be isoforms of the same gene locus?**
5. **In a 1Mb region around your reference SNP, pick two genes to knock out and provide the following information (in table format):**

* Genode gene ID
* RefSeq gene ID (if available)
* Gene name
* Description
* Mouse homolog
* Number of exons
* Amino acid length
* KEGG pathways in which it is found (if available)
* Presence of non-synonymous coding SNPs (name of SNP and location if applicable)

1. **For each gene, generate a graphic using the UCSC genome browser interface.**

This graphic should include only the genomic region for your gene and about 1000 bases upstream and downstream. Include on the graphic:

* UCSC genes
* RefSeq genes
* Human spliced ESTs
* SNPs (with non-synonymous coding SNPs colored different from the rest)

You may include other tracks and I leave it to your discretion how to display the tracks (pack, dense, full, etc.). Imagine that this will be a figure in a report for your committee so you want to present to them the necessary information without making it too messy.

**Export the graphic as a PDF file and attach to the homework (electronically) or cut and paste it into this assignment.**

**Finding Genes using the Ensembl genome browser**

I also want you to be familiar with Ensembl so you will also to identify **two genes** that you might want to investigate in the region of rs7671167 under the hypothesis that these genes might play a role in COPD. The minimum criteria for these genes is

* 1. They have more than 1 exon (single exon genes are often, not always, pseudogenes)
  2. They are protein coding genes.
  3. There is a mouse homolog

Use the Ensembl genome browser as the source for your information.

First, open the Ensembl browser and search for rs7671167. At the bottom of the page are a set of icons that link to various resources. Click on the “Phenotype Data” link.

1. **What three phenotypic measures is this SNP associated with?**

Display the genomic region using the “Location” tab. Once there, you can explore features of the genomic region.

Select the “Comparative Genomics” menu item, “Region Comparison” from the left-hand navigation menu and create a display between this human region and the same region in mouse.

1. **Which human Ensembl protein-coding genes are within a 1 MB region of your reference SNP?**
2. **Which mouse Ensembl genes are within the same 1MB window on the mouse genome?**