Problem Set 5 - Mutation hunting

This homework should be sent to mgymrek@ucsd.edu with subject line [CSE291 PS5:LASTNAME] by the beginning of class on **Thursday, March 9**. The assignment is worth 10 points total.

As in previous problem sets, template code is provided for some problems. Using the template code is optional, it is simply there to guide you.

For all plots, please include the plot in the writeup. For code, please paste the relevant snippets that you wrote into the writeup.

Objectives

- Gain experience with mutation hunting.
- Learn how to use methods for filtering and prioritizing variants in medical genetics studies.

A description of data files for this problem set and several setup steps can be found at PS5 resources.

Overview

Kabuki disease is a rare multiple malformation disorder that was first reported in the 1970s in Japan with an incidence of 1 in 100,000. The first causal gene for this disease was mapped 40 years later in 2010 with the availability of exome sequencing. This took so long since there is very little chance of the disease being inherited, thus it is not possible to do traditional linkage analysis. Discovering the gene was complicated by the fact that there are several different genes that can cause Kabuki Syndrome.

In this problem set, we'll implement a simple filtering pipeline to identify candidate disease causing genes and assess the impact of the number of patients, mutation type, and use of large control datasets on the ability to narrow down to the true causal gene.

In the ps5 data directory, you'll find the following files:

- VEP-annotated VCF file for 99 patients of European ancestry.

  `1kg_phase3_exome_ceu_all_vep.vcf.gz`

  You'll notice this file contains a very long field in the INFO, labeled "CSQ". This gives the annotations from the variant effect predictor (VEP). To make things easier, this file has been preprocessed as described below.

- VEP-annotated VCF file for 99 patients of European ancestry.

  `1kg_phase3_exome_ceu_all_vep.vcf.tab`

  This file contains the same information as the VCF in a tab delimited format, with columns for the chromosome, start, SNP rsid, gene name, consequence, ExAC minor allele frequency, and genotypes.

You'll also find the following scripts and templates in the templates directory:

- Helpful commands to get you started with each question:

  `run_ps5_template.sh`

- A utility script to count how many columns in a given range have values greater than 0

  `count_greater_than_zero.py`

Additionally, if you're interested in how the data was preprocessed, see `preprocess_exom_file.sh` and preprocess_vep.sh.

Note, while the Kabuki Syndrome paper (Ng *et al.*) uses *MLL2* to describe the causal gene, we'll use *KMT2D* to refer to the gene, since this nomenclature is more widely used.

Exercises

1. **(2 points)** For a single patient (Sample ID NA06984), how many candidate genes have at least one loss of function mutation (to simplify, you can look only at things labeled as "stop_gained" or "frameshift".

2. **(2 points)** Rank each gene by (i) the number of patients with a loss of function mutation (nonsense or frameshift) and alternatively by (ii) the number of patients with a missense mutation. Report the top 10 genes and number of each type of mutation. Where does the disease gene (*KMT2D*) rank on your list?

3. **(2 points)** Repeat the above, now filtering out any mutations that are seen in the ExAC dataset of 65,000 control exomes. Report the top 10 genes and number of each type of mutation. Where does *KMT2D* rank? Is it at the top? If not, what is? Explain why you think that is.

4. **(2 points)** How many *KMT2D* loss of function mutations are present in ExAC that are indistinguishable from disease mutations? What are their allele frequencies? Do they tend to fall in a certain area of the gene?

5. **(2 points)** Using the provided ExAC frequencies, what is the average minor allele frequency for variants labeled as "stop_gained", "synonymous_variant", and "missense_variant"? Restrict to variants seen in ExAC. What is your expectation for the relative allele frequencies of each class and why? Do the results match your expectation?

Acknowledgements

This problem set is based on a set of exercises designed by Vikas Bansal.