CSE 291 - PERSONAL GENOMICS FOR BIOINFORMATICIANS

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 genetic mapping.
- 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

Exercises

1. **(2 points)** For a single patient (Sample ID XX), how many candidate genes have at least one loss of function mutation (nonsense 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 (*MLL2*) 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 *MLL2* rank?
4. **(2 points)** How many *MLL2* loss of function mutations are present in ExAC that are indistinguishable from disease mutations? What are their allele frequencies?
5. **(2 points)** Using the provided sequence of the canonical *MLL2* transcript and the given per-base mutation probabilities, calculate the expected incidence of Kabuki Syndrome. Does it match known incidence (1 in 32,000)?

Acknowledgements

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