ROC Incidentalome App

1. Extend your analysis of variants in the two genes MYBPC/MYH7 in 1000G and ExAC to all HCM-related genes tested on our collaborator's (the LMM) panel, available here: [http://personalizedmedicine.partners.org/laboratory-for-molecular-medicine/tests/cardiomyopathy/default.aspx](https://urldefense.proofpoint.com/v2/url?u=http-3A__personalizedmedicine.partners.org_laboratory-2Dfor-2Dmolecular-2Dmedicine_tests_cardiomyopathy_default.aspx&d=CwMFaQ&c=-dg2m7zWuuDZ0MUcV7Sdqw&r=TFO3U2IN65ALPevhA0baNJ973d9SgAubepCjWZrUKdI&m=bL_tYGbVFmf3fqR0T8hLrEifLdBlbnubZUfZenkaerw&s=-fIEqTvxrQR60Ogwiw-ahbRDptW7BOSCEzgkbkqeRP0&e=) [you want the 20 genes on the HCM panel]. The approach is up to you but I would suggest scripting this task [it will take some time, but this will be very useful for the ACMG task below] - you should have enough tools to do this now.

(a) For 1000G, please read about the VCF format on the website, and check out this page on accessing individual-level data for a specific population: [http://www.1000genomes.org/faq/can-i-get-genotypes-specific-individualpopulation-your-vcf-files/](https://urldefense.proofpoint.com/v2/url?u=http-3A__www.1000genomes.org_faq_can-2Di-2Dget-2Dgenotypes-2Dspecific-2Dindividualpopulation-2Dyour-2Dvcf-2Dfiles_&d=CwMFaQ&c=-dg2m7zWuuDZ0MUcV7Sdqw&r=TFO3U2IN65ALPevhA0baNJ973d9SgAubepCjWZrUKdI&m=bL_tYGbVFmf3fqR0T8hLrEifLdBlbnubZUfZenkaerw&s=BTnnCWTsFsxio7eeOH0tYUG19jkgySOtEtZRtKFLpkw&e=). You should retrieve and store individual-level variant data for each HCM gene, for each 1000G population [this will take some time].

(b) For ExAC, there are many ways to do this. For example, you could download the large VCF of all variation or use scrapeR for individual genes - choose an approach, justify why you chose that approach, and proceed. You should retrieve variant-level allele frequent data for each ExAC population, for each HCM gene.

(c) Are you able to reconcile the findings between (a) and (b) on a per gene level?

2. Compute an HCM allele frequency threshold for the 'general population' assuming an HCM prevalence of 1:500, penetrance cutoff of 80%, and allelic heterogeneity of 0.001. Identify all individuals in 1000G who have such an HCM variant in any HCM gene based after removing variants above your computed allele frequency threshold in ExAC. Make basic plots by population/gene to visualize what fraction of variation/individuals are being captured/removed by the filtering.

3. Vary the threshold in 2 as a function of three parameters: (a) the assumed HCM prevalence in 1000G and ExAC [1/10000 to 1/1], and (b) the assumed allelic heterogeneity [1/10000 to 1], (c) the penetrance cutoff of 80% [0 to 100%]. Using the population in 2. as the 'gold standard', extend your R/Shiny app to create an reactive ROC curve that visualizes how the classifier's performance moves in ROC space as a function of these three parameters. Additionally, plot gene-level individual-level variant data (e.g. a hist for each gene for each population) in the R/Shiny App, as well as ExAC allele frequency data.

4. Take and send a few representative screenshots as well as a screencast of your app.

5. Commit all changes to GitHub.

ACMG 46 Findings

1. Read the ACMG paper closely [entitled "ACMG recommendations for reporting of incidental findings in clinical exome and genome sequencing" by Green et al.].

2 Based upon your reading, scripting tools created thus far, and literature review (I assume you have access to PubMed?), create a SQL DB with columns:

variant\_name | gene | chrom | pos | ref | alt | disease | 1000g\_freq | ExAC\_freq | disease\_prev | citation

for each variant in ExAC residing in one of the ACMG recommended genes. Don't worry about filtering for KP and EP for now - just include all variants.

“KP = known pathogenic, sequence variation is previously reported and is a recognized cause of the disorder; EP = expected pathogenic, sequence variation is previously unreported and is of the type which is expected to cause the disorder. Note: The recommendation to not report expected pathogenic variants for some genes is due to the recognition that truncating variants, the primary type of expected pathogenic variants, are not an established cause of some diseases on the list.”

3. Plot the summed allele frequency vs. the disease prevalence for each distinct disease.