GWAS Catalog pipeline

Authors:

Geraldine San Ramon

Jessie Chan

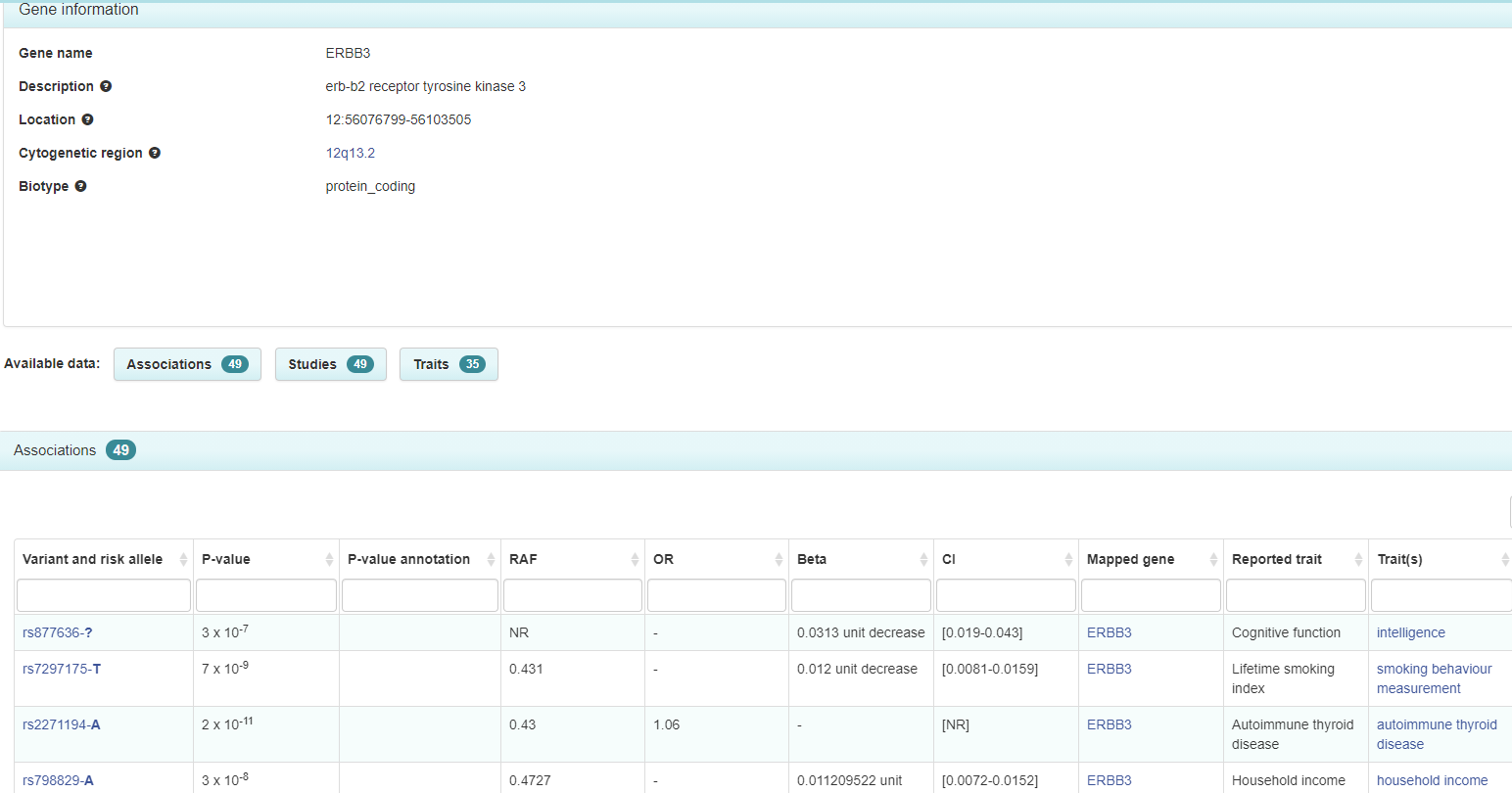
Zain Anwar

Created: 3/16/2021

Last Update: 3/17/2021

With widespread use of GWAS, thousands of SNPs were identified to associate to phenotypes, but it is not clear which genes has causal effect on phenotypes. In a previous study “A gene-based association method for mapping traits using reference transcriptome data” by Dr. Wheeler and her team, they tried to use a gene-based association method called PrediXcan to detect genes associated with disease traits by directly testing the molecular mechanism of the gene that affect the phenotype.

Now that they have identified these genes of interest, they want to find out if these genes were already known or studied to be associated to the phenotype they found. This is done by manually searching the GWAS catalog for the gene and looking at the list of SNPs and check if the trait matches the phenotype found to be associated to the said gene. See example of a GWAS catalog search of gene ERBB3 below:



Gene of interest

List of SNPs identified in the gene

Phenotype to search through to find match

Thus, the goal of this project is to automate this process of GWAS Catalog queries of PrediXcan genes to find SNPs mapped to this gene with the phenotype found to be associated to the gene.

Associated to

Given: Gene phenotype (or trait)

Associated to

Find: SNPs phenotypes that match

From the catalog search of the SNPs, they want to get all the phenotypes and filter to the matching phenotype they’re looking for and separate the ones that don’t match but still output this non-match to see if the phenotype is just entered differently.

Proposed solution:

Given:

1. Gene (accession ID)

2. phenotype

Search GWAS catalog using the gene accession ID

SNP

Trait match given phenotype?

List of SNP (with summary statistics) that match phenotype

List of SNP (with summary statistics) that don’t match phenotype

YES

NO

1

2

3

STEPS:

1. Automate searching the gene from a list to the GWAS catalog.
   1. Search GWAS using the gene accession id from a list (in csv format). There’s another rawer format that needs to be parsed to get it into a list form but for now we’ll focus on the list.
      1. Using GWAS Catalog API (Application Programming Interface) <https://www.ebi.ac.uk/gwas/docs/api>
      2. Gencode annotations (may be needed to map Ensembl gene ID’s to gene names) https://www.gencodegenes.org/
2. Pull the SNPs from the search query.
   1. gwascat or gwasrapidd or write a pipeline in bash/python.
3. Match the given phenotype to the traits column of the SNP search result.

Steps

1. Zain
2. Jessie
3. Geraldine

Week 1 – Figure out each step, API and suggested software

Week 2 – Coding and see we’re we’ll have some issues. I expect the third step will be tricky.

Week 3 – Check each other’s assigned task and see if we can connect the workflow.

Week 4 – Start debugging