Analysis plan for Amy-WIP

Contents

[Background: 1](#_Toc172025504)

[Purpose: 2](#_Toc172025505)

[Scope: 2](#_Toc172025506)

[Methods: 2](#_Toc172025507)

[Organization: 3](#_Toc172025508)

[Github: 3](#_Toc172025509)

[Folder Layout: 3](#_Toc172025510)

[Presentation: 5](#_Toc172025511)

[Citations: 5](#_Toc172025512)

# Background:

Variants have been found in filtering has been performed. Additional information about these variants needs to be added and processed to create a human readable document.

# Purpose:

To create documents with variant information from families. This should be automated as much as possible to allow for a feedback iteration loop. The process will be setup in a github repository with instructions on running it.

# Scope:

Information that is relevant to the specific variants with a focus on neurological disorders.

# Methods:

1. Prepare the Excel Spreadsheet

Get variant information from Jim. Ensure your Excel spreadsheet is well-organized. Each row should represent a genetic variant and columns should include relevant information such as:

* Variant ID
* Gene name
* Nucleotide change
* Amino acid change
* Zygosity (homozygous/heterozygous)
* Associated conditions
* Family ID
* Other relevant details that would be searched for

1. Prepare Additional Spreadsheets Containing Additional Gene Information using an automated process:

Each additional spreadsheet should focus on a specific aspects of gene information and should include information about where the information was found. This information can be searched for manually but an automated process should be developed if possible. Common categories of information might include but are not limited to:

* Gene Details: Includes basic information about the gene (Gene symbol, full gene name, chromosomal location, function).
* Gene Variants: Lists known variants of the gene, their clinical significance, and population frequency.
* Literature: References to research articles related to the gene or its variants.
* Functional Studies: Outcomes of functional studies such as expression levels, protein function, and interaction partners.
* Disease Associations: Information on diseases and conditions linked to the gene.

1. Linking Gene Information with Variant Data:

To enrich the documents so they contain all relevant information, link the data from these additional spreadsheets by gene name. This can be achieved by merging data frames or looking up relevant information during document generation.

1. Use an Automated Process to Incorporate Variant and Gene Information into a Document:

Extract information into a document using an automated process. The format can be a Word or other document form which can be opened in Word. This needs to be done in an automated process due to the number of variants being processed. The process can be a python script, LLM or any other method that is approved of by the group.

1. Integrate feedback and Update the Automated Process:

Collect feedback on the documents that are produced. Update the scripts or LLM prompts to apply updates to the documents. Integrating the feedback loop and automating the updates allows the document creation process to be more flexible, and ensures that high-quality, up-to-date reports are consistently produced.

# Organization:

## Github:

Code, LLM prompts and example data will be organized in a specific repository on github. The repository will only contain generic example data. Any rare generic variants will be located on the will be stored on Umich computers only.

## Folder Layout:

The analysis structure will use cookie cutter data science layout.

* [Cookiecutter Data Science (drivendata.org)](https://cookiecutter-data-science.drivendata.org/)

A screenshot of a computer

Description automatically generated

Figure 1: GitHub layout using cookie cutter v2 folders.

The following is the generic folder layout for cookie cutter v2

├── LICENSE <- Open-source license if one is chosen

├── Makefile <- Makefile with convenience commands like `make data` or `make train`

├── README.md <- The top-level README for developers using this project.

├── data

│ ├── external <- Data from third party sources.

│ ├── interim <- Intermediate data that has been transformed.

│ ├── processed <- The final, canonical data sets for modeling.

│ └── raw <- The original, immutable data dump.

│

├── docs <- A default mkdocs project; see mkdocs.org for details

│

├── models <- Trained and serialized models, model predictions, or model summaries

│

├── notebooks <- Jupyter notebooks. Naming convention is a number (for ordering),

│ the creator's initials, and a short `-` delimited description, e.g.

│ `1.0-jqp-initial-data-exploration`.

│

├── pyproject.toml <- Project configuration file with package metadata for gen\_analysis\_module

│ and configuration for tools like black

│

├── references <- Data dictionaries, manuals, and all other explanatory materials.

│

├── reports <- Generated analysis as HTML, PDF, LaTeX, etc.

│ └── figures <- Generated graphics and figures to be used in reporting

│

├── requirements.txt <- The requirements file for reproducing the analysis environment, e.g.

│ generated with `pip freeze > requirements.txt`

│

├── setup.cfg <- Configuration file for flake8

│

└── gen\_analysis\_module <- Source code for use in this project.

│

├── \_\_init\_\_.py <- Makes gen\_analysis\_module a Python module

│

├── data <- Scripts to download or generate data

│ └── make\_dataset.py

│

├── features <- Scripts to turn raw data into features for modeling

│ └── build\_features.py

│

├── models <- Scripts to train models and then use trained models to make

│ │ predictions

│ ├── predict\_model.py

│ └── train\_model.py

│

└── visualization <- Scripts to create exploratory and results oriented visualizations

└── visualize.py

# Presentation:

How data will be presented.

1. Save data in an excel spreadsheet.
2. Transform that data into a document in a specific presentation format based on feedback.

# Citations:

* [Cookiecutter Data Science (drivendata.org)](https://cookiecutter-data-science.drivendata.org/)
* [Annotation & Prediction (ensembl.org)](https://useast.ensembl.org/info/genome/index.html)