31st March 2023

**Title:**

Discovery of novel AMC genes in humans by mining the zebrafish phenotype data

**Aim:**

To identify new candidate genes implicated in AMC phenotypes.

**Methods:**

**Step 1)** Filter zebrafish genes with human orthologs- HT

**Step 2)** Filter genes using AMC descriptive terms- HT

**Step 3)** Filter out genes found in “Structural eye disease PanelAPP” v2.1, considering gene duplication in zebrafish (paralogues)- HT

**Step 4a)** Apply filtered genes to AMC patient datasets, such as our 400+ patient cohort (including Bionano data), Genomics England datasets, and the DDD cap study- in progress LT and Hande)

1. I created the following python scripts to correct and handle filtering of the patient F3 file (original copy located in google drive under Ragge\_WES\_filtered-data folder). There is a total of three scripts (‘**Step0\_Correcting\_F3\_file**’, ‘**Step1\_Patient\_F3\_Filtering**’ and ‘**Step2\_Patient\_F3\_Filtering\_Return\_All\_Variants**’)each one with its own folder and subfolders to organise input and output files.

The hierarchy of the folders and files are outlined below.

**Step0\_Prep**

Input

* + - * + *F3\_1%\_all-probands\_20221013b\_csv.csv*

Output

* + - * + *Corrected\_F3\_with\_unique\_family\_no.csv*
        + *Corrected\_F3\_with\_unique\_family\_no\_nan.csv*
        + *No\_unique\_families\_in\_cohort.txt*

*Step0\_Prep.py*

**Step1\_Patient\_F3\_Filtering**

Anophthalmia

Input

* + - * + *Corrected\_F3\_with\_unique\_family\_no\_nan.csv*
        + *Human\_Anophthalmia\_Synonyms\_SingleColumn.csv*

Output

* + - * + *Corrected\_F3\_with\_unique\_family\_no.csv*
        + *Corrected\_F3\_with\_unique\_family\_no\_nan.csv*
        + *No\_unique\_families\_in\_cohort.txt*

Microphthalmia

Input

* + - * + *Corrected\_F3\_with\_unique\_family\_no\_nan.csv*
        + *Human\_Microphthalmia\_Synonyms\_SingleColumn.csv*

Output

* + - * + *Final\_Filtered\_F3\_data.csv*
        + *Final\_Filtered\_F3\_data\_greyout.xlsx*
        + *No\_unique\_filtered\_families.txt*

Coloboma

Input

* + - * + *Corrected\_F3\_with\_unique\_family\_no\_nan.csv*
        + *Human\_Coloboma\_Synonyms\_SingleColumn.csv*

Output

* + - * + *Final\_Filtered\_F3\_data.csv*
        + *Final\_Filtered\_F3\_data\_greyout.xlsx*
        + *No\_unique\_filtered\_families.txt*

*Step1\_Patient\_F3\_Filtering.py*

**Step2\_Patient\_F3\_Filtering\_Return\_All\_Variants**

Anophthalmia

Input

* + - * + *Corrected\_F3\_ nan.csv*
        + *Final\_Filtered\_F3\_data.csv*
        + *Human\_Anophthalmia\_Synonyms\_SingleColumn.csv*

Output

* + - * + *Anophthalmia\_geneList\_All\_variants.csv*
        + *Anophthalmia\_geneList\_All\_variants.xlsx*
        + *Anophthalmia\_geneList\_All\_variants\_greyout\_nan.xlsx*

Microphthalmia

Input

* + - * + *Corrected\_F3\_nan.csv*
        + *Final\_Filtered\_F3\_data.csv*
        + *Human\_Microphthalmia\_Synonyms\_SingleColumn.csv*

Output

* + - * + *Microphthalmia\_geneList\_All\_variants.csv*
        + *Microphthalmia\_geneList\_All\_variants.xlsx*
        + *Microphthalmia\_geneList\_All\_variants\_greyout\_nan.xlsx*

Coloboma

Input

* + - * + *Corrected\_F3\_nan.csv*
        + *Final\_Filtered\_F3\_data.csv*
        + *Human\_Coloboma\_Synonyms\_SingleColumn.csv*

Output

* + - * + *Coloboma \_geneList\_All\_variants.csv*
        + *Coloboma\_geneList\_All\_variants.xlsx*
        + *Coloboma\_geneList\_All\_variants\_greyout\_nan.xlsx*

*Step2\_Patient\_F3\_Filtering\_Return\_All\_Variants\_v3.py*

OLD

*Step2\_Patient\_F3\_Filtering\_Return\_All\_Variants.py*

*Step2\_Patient\_F3\_Filtering\_Return\_All\_Variants\_v2.py*

# **Summary of python scripts**

## **Step0\_Prep.py:**

Prepares patient F3 file (located in google drive folder *🡪* ‘Ragge\_WES\_filtered-data’) for filtering in subsequent steps.

* Make corrections below columns:

**‘Gene’** (e.g., ‘Mar-02’ to ‘MARC2’)

**‘Gene.refGene2019’** (e.g., ‘Mar-02’ to ‘MARC2’)

**‘Gene.knownGene’** (e.g., ‘37315’ to ‘MARCHF6’)

**‘MutPred\_score’** (e.g., ‘-’ to ‘nan’)

**‘Comments’** (some empty cells are converted to ‘nan’s therefore ‘nan’ values are converted back to blanks (‘’))

**‘Gene.symbol’** (e.g., ‘37316’ to ‘MARC2’)

* Extract unique family numbers (IDs) and store as new column.
* Construct column for lowest reported allele frequency based on columns:

**‘AF\_exome’**

**‘AF\_genome’**

**‘ExAC\_ALL’**

**‘esp6500siv2\_all’**

**‘1000g2015aug\_all’**

* Save file as **‘Corrected\_F3.csv’**
* Replace all instances of ‘.’ with ‘nan’.
* Save file as **‘Corrected\_F3\_nan.csv’**
* Produce text file with the number of families in the F3 file **‘No\_unique\_families\_in\_cohort.txt’** (Note: not the same as individuals with atypical phenotypes).

===============

## **Step1\_Patient\_F3\_Filtering.py:**

Returns .csv and .xslx file of gene variants in genes matching gene of interest lists (individually anophthalmia, microphthalmia and coloboma genes lists). Produces a colour coded excel file for easier interpretation.

* Copy files into input folder
* Filter **‘Corrected\_F3.csv’** for genes found in the anophthalmia, microphthalmia and coloboma gene list.
* Save filtered file as **‘Final\_Filtered\_F3\_data.csv’**.
* Produce text file with the number of families in the F3 file **‘No\_unique\_families\_in\_cohort.txt’** (Note: not the same as individuals with atypical phenotypes).
* Colourise filtered file using masks (in accordance with conditions outlined below):
  + - * + Mask1: Column **‘CLNSIG’** contains terms ‘benign’ or ‘Benign’ 🡪 Grey font.
        + Mask2: Column **‘InterVar\_automated’** contains terms ‘benign’ or ‘Benign’ 🡪 Grey font.
        + Mask3: Column **‘CADD\_phred’** is greater than 15 (values closer to 99 is deleterious) & is not blank (unknown) 🡪 Grey font.
        + Mask4: Unused.
        + Mask5: Column **‘Constructated\_min\_allel\_freqs’** is greater than 0.01 (equivalent of 1% in population) & is not blank (unknown) 🡪 Grey font.
        + Mask6: Column **‘Func.refGene’** is not ‘exonic’ or ‘exonic;splicing’ or ‘ncRNA\_exonic’ or ‘ncRNA\_exonic;splicing’ or ‘splicing’ 🡪 Grey font.
* Save file as **‘Final\_Filtered\_F3\_data\_greyout.xlsx’** (excel file format).

===============

## **Step2\_Patient\_F3\_Filtering\_Return\_All\_Vartiants.py:**

Returns .csv and .xslx files of gene variants in genes matching gene of interest lists (individually anophthalmia, microphthalmia and coloboma genes lists). Also returns other variants found in the same families and colours the excel file to easier interpretation.

* Copy files into input folder
* Filter **‘Corrected\_F3.csv’** for genes found in the anophthalmia, microphthalmia and coloboma gene list and any other variants found in the same families.
* Save filtered file as:
  + - * + **‘Anophthalmia\_geneList\_All\_variants.csv’**
        + **OR ‘Microphthalmia\_geneList\_All\_variants.csv’**
        + **OR ‘Coloboma\_geneList\_All\_variants.csv’**
* Save filtered file as:
  + - * + **‘Anophthalmia\_geneList\_All\_variants.xlsx’**
        + **OR ‘Microphthalmia\_geneList\_All\_variants.xlsx’**
        + **OR ‘Coloboma\_geneList\_All\_variants.xlsx’**
* Produce text file with the number of families in the F3 file **‘No\_unique\_families\_in\_cohort.txt’** (Note: not the same as individuals with atypical phenotypes).
* Colourise filtered file using masks (in accordance with conditions outlined below):
  + - * + Mask1: Column **‘CLNSIG’** contains terms ‘benign’ or ‘Benign’ 🡪 Grey font.
        + Mask2: Column **‘InterVar\_automated’** contains terms ‘benign’ or ‘Benign’ 🡪 Grey font.
        + Mask3: Column **‘CADD\_phred’** is greater than 15 (values closer to 99 is deleterious) & is not blank (unknown) 🡪 Grey font.
        + Mask4: Column **‘gene’** is found in the gene of interest list (anophthalmia, microphthalmia, coloboma) 🡪 **Bold font & dark grey background**.
        + Mask5: Column **‘Constructated\_min\_allel\_freqs’** is greater than 0.01 (equivalent of 1% in population) & is not blank (unknown) 🡪 Grey font.
        + Mask6: Column **‘Func.refGene’** is not ‘exonic’ or ‘exonic;splicing’ or ‘ncRNA\_exonic’ or ‘ncRNA\_exonic;splicing’ or ‘splicing’ 🡪 Grey font.
* Save file as **‘Anophthalmia\_geneList\_All\_variants\_greyout.xlsx’**/ **‘Microphthalmia\_geneList\_All\_variants\_greyout.xlsx’**/ **‘Coloboma\_geneList\_All\_variants\_greyout.xlsx’** (excel file format).