FUMA Practical Boulder 2023

FUMA GWAS uses information from publicly available annotation and mapping datasets (see https://fuma.ctglab.nl/tutorial#celltype – Data sets)

We will work with results from published GWAS studies on either Neuroticism or Schizophrenia – you can choose which trait to work on.

- Schizophrenia (SCZ; Ripke et al., 2014)
- Neuroticism (NEU; Nagel et al., 2018)

Finding your data

- Go to: http://fuma.ctglab.nl/
- Click 'Tutorial' (top of the page) to see a schematic overview of what you can do with FUMA. If something's unclear during the practical, first consult the tutorial, it's probably explained in there!
- We will use previously published GWAS results that have already been submitted to FUMA by Kyoko Watanabe. Click 'Browse Public Results', and scroll down to the trait of your choice (Schizophrenia Ripke et al. or Neuroticism Nagel et al.). E.g., if you chose schizophrenia, you should click this line:

3	Example results of Schizophrenia	Kyoko Watanabe	k.watanabe@vu.nl	Schizophrenia	PMID: 29184056	https://www.med.unc.edu/pgc/results-and-downloads/downloads	PMID: 25056061	Predefined lead SNPs in Supplementary Data 14 was used.	2017-05-19	2017-06-24

- For Neuroticism, choose 'Example results of Neuroticism'.

Section: Summary of results

The questions below concern the 'Summary of results' section in FUMA.

Due to privacy concerns the publicly available data (as used in this practical) is sometimes slightly different from the data on which published results are based. E.g., the personal genomics company 23andMe shares data, but researchers are not allowed to include the 23andMe participants in any data they make publicly available.

Schizophrenia (if the question is not related to one specific trait, (part of) the answer is printed in red font too)

Neuroticism

- As you saw in the lecture, SNPs located close to each other often show similar association signal (e.g., if SNP A is strongly associated then SNP B that lies right next to it will likely also show fairly strong association, i.e. low *P*-value). This has to do with the concept of linkage disequilibrium (LD).
  Q1: What LD threshold does FUMA apply to define independent significant SNPs? (see Watanabe et al. (2017) or tutorial).
  Independent significant SNPs are (A) genome-wide significant and (B) independent from each other at $r^2 < 0.6$. In other words, these SNPs are statistically associated to the trait of interest, and not in (strong) LD.

- SNPs in exonic regions potentially affect the expression of the gene or the function of the gene product (protein).
  Q2: How many SNPs are annotated to exonic regions in your example results (schizophrenia or neuroticism)?
  When hovering over the bar plot in the top right corner, you see that FUMA reports 157 exonic SNPs.
  When hovering over the bar plot in the top right corner, you see that FUMA reports 135 exonic SNPs.
- FUMA established genomic risk loci, based on the independent associations. If you scroll down you see four bar plots, showing information for each genomic locus.
  Q3: What is the largest genomic locus? (Copy-paste the y-axis label)
  18:52747689-53804156. This locus spans >1,000 kb (or: > 1Mb).
  6:26903585-28833101 This locus spans >1929 kb (or: > 1Mb).
  Q4: Does this locus also contain the most SNPs? If not, which locus does?
  No, this locus contains 245 SNPs, whereas locus 1:73275828-74077588 contains 956 SNPs.
  No, this locus contains 477 SNPs, whereas locus 17:43460181-44874453 contains 2417 SNPs.
  Q5: How come the number of genes physically located within a locus is not always equal to the number of genes that is mapped to the locus?
  FUMA uses several strategies to map SNPs to genes. Only for positional mapping the SNP is located within (or very close) to the gene. Other strategies, eQTL and chromatin interaction mapping, can map SNPs to genes that are further away!

Section: Genome-wide plots

The questions below concern the 'Genome-wide plots' section in FUMA.

- Below the SNP-level Manhattan plot you find the gene Manhattan plot. This figure shows the gene associations.
  Q6: Why is the threshold for significance different from that in the SNP Manhattan plot?
  The (genome-wide) significance threshold in the SNP Manhattan plot is based on the assumption that we test approximately 1,000,000 independent associations ($0.05/1{,}000{,}000 = 0.05 \times 10^{-8}$). In the gene-based test, the multiple testing correction is based on the number of (protein coding) genes, which is 'only' ~18,950 in the current data.
- Scrolling down you see a histogram, showing whether 53 types of tissue are enriched for the genes associated to the trait of interest. In other words; this informs you on whether the identified genes are primarily expressed in a specific tissue type.
  Q7: Is there a tissue type that shows significant enrichment? If so, which one(s)? Is this what you would expect, given the trait of interest?
  Yes. All brain tissues appear to show significant enrichment for the schizophrenia related genes. Since schizophrenia is generally viewed as a brain disorder, this is not surprising.

Yes. All brain tissues appear to show significant enrichment for the neuroticism related genes. Since neuroticism is generally viewed as a brain-related trait, this is not surprising.

Section: Results

The questions below concern the 'Results' section in FUMA (note that this section has several tabs).

The 'Genomic risk loci' tab lists all genomic loci, and provides information on the location, the number of SNPs, the lowest SNP *P*-value etc. Moreover, if you click a locus, an interactive regional plot appears below the table (you can, for example, click on a SNP for info on that particular SNP). Essentially this is a zoomed in version of the Manhattan plot, with each dot representing a SNP.

Depending on your trait of interest, create a regional plot for genomic locus number:

- Schizophrenia: 97
- Neuroticism: 73

Q8: What is the rsID of the top lead SNP? Which chromosome is it on? How many SNPs are in LD with this lead SNP? What is the minor allele frequency?

rs8082590
285 SNPs in LD (from table on the right side of the regional plot)
MAF = 0.3767 (hover the mouse of the lead SNP)
Chromosome 17
rs11066591
311 SNPs in LD (from table on the right side of the regional plot)
MAF = 0.319 (hover the mouse of the lead SNP)
Chrosome 12

Now click the 'Plot' button on the bottom to open a new window, showing the same regional plot with additional more information. Try zooming, scrolling and clicking a bit to see what happens.

Q9: What is the nearest mapped gene to the lead SNP?

GID4
MYO1H

- The plot below shows the CADD score. Since GWAS often identifies many SNPs, we need a criterion to prioritize which ones to study further.

  Q10: Why is the CADD score helpful for this? Which SNP has the highest CADD score in this locus? Is this below or above the threshold mentioned in Watanabe et al. (2017)?

  CADD stands for 'Combined Annotation Dependent Depletion'. The CADD score is a measure of the deleteriousness of a variant. Read more here: https://cadd.gs.washington.edu/

  SCZ: rs4584886 has the highest CADD score (33), exceeding the threshold of 12.37.

  NEU: rs7298565 has the highest CADD score (22.7), exceeding the threshold of 12.37.

  Q11: Was the SNP with the highest CADD score genome-wide significant?

  No, *P* = 1.601e-7

  Yes, *P* = 2.301e-9

  Q12: How many exonic SNPs are in the locus?

3

7

Q13: Based on these results, which gene would you recommend to study in more detail? Explain why.

First of all, before setting up expensive lab experiments into the function of a specific gene, more evidence is required. However, given the information available here, *LRRC48*, is an interesting candidate. This gene is located closely to the exonic SNP with a high CADD score, suggesting that it might negatively influence the function of this gene.

First of all, before setting up expensive lab experiments into the function of a specific gene, more evidence is required. However, given the information available here, *UBE3B* is an interesting candidate. This gene is located closely to the exonic SNP with a high CADD score, suggesting that it might negatively influence the function of this gene.

There exist many collections of genes that have something in common; gene sets. Sometimes these are composed by experts, e.g. listing all genes that are involved in a specific biological process. In FUMA, you can find those results in the 'Gene sets' section. We will have a look at the 'GWAS catalog reported genes' (second-last), gene sets based on association with a large variety of traits (listed in the GWAS catalog).

Q14: What is the top 3 gene sets reported for your trait? Does this surprise you?

Schizophrenia, Bipolar disorder & Intelligence. Schizophrenia and bipolar disorder are known to be genetically quite similar. Intelligence correlates (negatively) with schizophrenia, so in that sense you would expect some genes to influence both. However, it's still unknown what the exact role of these genes in both schizophrenia and intelligence is.

Autism spectrum disorder or schizophrenia, neuroticism and schizophrenia. Indeed neuroticism and autism spectrum disorder are known to share genetic risk factors, and neuroticism is a risk factor for schizophrenia.

Chromatin interactions

- In the 'Results' section, click 'Chromatin interactions'. If you scroll down you find circos plots, showing eQTL and chromatin interactions per chromosome (read the info text on the circos plots!). As mentioned earlier, these techniques allow SNPs to be mapped to genes that are further apart.

  Q15: In the circos plots, do you find examples of these long-range interactions, where genes are implied through interactions with (physically) distant regions? What do the different layers and colors mean?

  Manhattan plot: The most outer layer. Only SNPs with P < 0.05 are displayed. SNPs in genomic risk loci are color-coded as a function of their maximum r^2 to the one of the independent significant SNPs in the locus, as follows: red ($r^2 > 0.8$), orange ($r^2 > 0.6$), green

($r^2 > 0.4$) and blue ($r^2 > 0.2$). SNPs that are not in LD with any of the independent significant SNPs (with $r^2 \leq 0.2$) are grey.
The rsID of the top SNPs in each risk locus are displayed in the most outer layer. Y-axis are ranges between 0 to the maximum -log10(P-value) of the SNPs.

Chromosome ring: The second layer. Genomic risk loci are highlighted in blue.

Mapped genes by chromatin interactions or eQTLs: Only mapped genes by either chromatin interaction and/or eQTLs (conditional on user defined parameters) are displayed. If the gene is mapped only by chromatin interactions or only by eQTLs, it is colored orange or green, respectively. When the gene is mapped by both, it is colored red.

Chromosome ring: The third layer. This is the same as second layer but without coordinates to make it easy to align position of genes with genomic coordinate.

Chromatin interaction links: Links colored orange are chromatin interactions.

eQTL links: Links colored green are eQTLs.

Further reading

FUMA:

- Watanabe (2017) - Functional mapping and annotation of genetic associations with FUMA

Studies from which data was used:

- Ripke et al. (2014) - Biological insights from 108 schizophrenia-associated genetic loci
- Nagel et al., 2018 - Meta-analysis of genome-wide association studies for neuroticism in 449,484 individuals identifies novel genetic loci and pathways