

## **FUMA: Functional mapping and annotation of genetic associations**

Kyoko Watanabe<sup>1</sup>, Erdogan Taskesen<sup>1,2</sup>, Arjen van Bochoven<sup>3</sup>, Danielle Posthuma<sup>1,4\*</sup>

### **Affiliations:**

<sup>1</sup>Department of Complex Trait Genetics, Center for Neurogenomics and Cognitive Research, VU University Amsterdam, Amsterdam, the Netherlands

<sup>2</sup>VU University Medical Center (VUMC), Alzheimercentrum, Amsterdam, the Netherlands

<sup>3</sup>Faculty of Science, VU University Amsterdam, Amsterdam, the Netherlands

<sup>4</sup>Department of Clinical Genetics, VU University Medical Center, Amsterdam Neuroscience, Amsterdam, the Netherlands

K.W.: [k.watanabe@vu.nl](mailto:k.watanabe@vu.nl), E.T.: [e.taskesen@vu.nl](mailto:e.taskesen@vu.nl), A.v.B.: [a.e.van.bochoven@vu.nl](mailto:a.e.van.bochoven@vu.nl), D.P.: [danielle.posthuma@vu.nl](mailto:danielle.posthuma@vu.nl)

\* Corresponding author

**Word count:** abstract: 68; main text: 1652

**Display items:** 3 (Figures 1-3)

Includes **Online Methods, Supplementary Materials** (incl. Results S1-S2, Fig. S1-S11) and **Supplementary Tables** in excel S1-S26

## ABSTRACT

1 A main challenge in genome-wide association studies (GWAS) is to prioritize genetic variants and  
2 identify potential causal mechanisms of human diseases. Although multiple bioinformatics  
3 resources are available for functional annotation and prioritization, a standard, integrative approach  
4 is lacking. We developed FUMA: a web-based platform to facilitate functional annotation of  
5 GWAS results, prioritization of genes and interactive visualization of annotated results by  
6 incorporating information from multiple state-of-the-art biological databases.

7

8 **Subject terms:** genome-wide association study (GWAS), prioritization, functional annotation,  
9 visualization, expression quantitative trait loci (eQTLs), gene mapping

10

## 11 MAIN TEXT

12 In the past decade, more than 2,500 genome-wide association studies (GWAS) have identified  
13 thousands of genetic loci for hundreds of traits<sup>1</sup>. The past three years have seen an explosive  
14 increase in GWAS sample sizes<sup>2-4</sup>, and these are expected to increase even further to 0.5-1 million  
15 in the next year and beyond<sup>5</sup>. These well-powered GWAS will not only lead to more reliable results  
16 but also to an increase in the number of detected disease-associated genetic loci. To benefit from  
17 these results, it is crucial to translate genetic loci into actionable variants that can guide functional  
18 genomics experimentation and drug target testing<sup>6</sup>. However, since the majority of GWAS hits are  
19 located in non-coding or intergenic regions<sup>7</sup>, direct inference from significantly associated single  
20 nucleotide polymorphisms (SNPs) rarely yields functional variants. More commonly, GWAS hits  
21 span a genomic region ('GWAS risk loci') that is characterized by multiple correlated SNPs, and  
22 may cover multiple closely located genes. Some of these genes may be relevant to the disease,  
23 while others are not, yet due to the correlated nature of closely located genetic variants,  
24 distinguishing relevant from non-relevant genes is often not possible based on association P-values  
25 alone. Pinpointing the most likely relevant, causal genes and variants requires integrating available  
26 information about regional linkage disequilibrium (LD) patterns and functional consequences of  
27 correlated SNPs. Ideally, functional inferences obtained from different repositories are integrated,  
28 and annotated SNP effects are interpreted in the broader context of genes and molecular pathways.  
29 For example, consider a genomic risk locus with one lead SNP associated with an increased risk for  
30 a disease, and several dozen other SNPs in LD with the lead SNP that also show a low association P  
31 value, spanning multiple genes. If none of these tested SNPs and none of the other (not tested but  
32 known) SNPs in LD with the lead SNP are known to have a functional consequence (i.e. altering  
33 expression of a gene, affecting a binding site or violating the protein structure), no causal gene can  
34 be indicated. However, if one or several of the SNPs are known to affect the function of one of the  
35 genes in the area, but not the other genes, then that single gene has a higher probability of being  
36 functionally related to the disease.

37 In practice, the extraction and interpretation of the relevant biological information from available  
38 repositories is not always straightforward, and can be time-consuming as well as error-prone. We  
39 have, therefore, developed FUMA, which functionally annotates GWAS findings and prioritizes the  
40 most likely causal SNPs and genes using information from 14 biological data repositories and tools  
41 (Supplementary Table 1). Results are visualized to facilitate quick insight into the implicated  
42 molecular functions. FUMA is available as an online tool at <http://fuma.ctglab.nl>, where users can  
43 set several parameters to filter SNPs or specify specific tissues to be used for annotation based on  
44 expression data (Supplementary Table 2 and Supplementary Fig. 1). As input, FUMA takes  
45 summary statistics from GWAS.

46 The core function of FUMA is the *SNP2GENE* process (Fig.1; Online Methods), in which SNPs are  
47 annotated with their biological functionality and mapped to genes based on positional and  
48 functional information of SNPs. First, conditional on the provided summary statistics, independent  
49 lead SNPs and their surrounding genomic loci are identified depending on LD structure. Lead SNPs  
50 and SNPs which are in LD with the lead SNPs are then annotated for functional consequences on  
51 gene functions (based on Ensembl genes (build 85) using ANNOVAR<sup>8</sup>), deleteriousness score  
52 (CADD score<sup>9</sup>), potential regulatory functions (RegulomeDB score<sup>10</sup> and 15-core chromatin state  
53 predicted by ChromHMM<sup>11</sup> for 127 tissue/cell types<sup>12,13</sup>) and effects on gene expression using  
54 expression quantitative trait loci (eQTLs) of various tissue types (see Online Methods). At this stage,  
55 lead SNPs and correlated SNPs are also linked to the GWAS catalog<sup>1</sup> to provide insight into  
56 previously reported associations with a variety of phenotypes. Functionally annotated SNPs are  
57 subsequently mapped to genes based on functional consequences on genes annotated by  
58 ANNOVAR (positional mapping) and/or eQTLs of user defined tissue types (eQTL mapping).  
59 Gene mapping can be controlled by setting several parameters (Supplementary Table 2) that allow  
60 to in- or exclude specific functional categories of SNPs. For example, positional mapping may  
61 optionally use only *coding* SNPs for gene mapping. For eQTL mapping, specific tissues can be  
62 selected to only include SNPs that influence the expression of genes in the selected tissue(s) (Online

63 Methods and Supplementary Table 2). By combining positional mapping of deleterious coding  
64 SNPs and eQTL mapping across (relevant) tissue types (i.e. functional mapping; Online Methods),  
65 FUMA enables to prioritize genes that are highly likely involved in the trait of interest. Due to the  
66 use of eQTL information, the prioritized genes – although influenced by SNPs within a disease-  
67 associated locus - are not necessarily themselves located inside that locus.

68 To obtain insight into putative causal mechanisms, the *GENE2FUNC* process annotates the  
69 prioritized genes in biological context (Fig. 1; Online Methods). Specifically, biological  
70 information of each input gene is provided to gain insight into previously associated diseases as  
71 well as drug targets by mapping OMIM<sup>14</sup> ID and DrugBank<sup>15</sup> ID. Tissue specific expression  
72 patterns for each gene are visualized as an interactive heatmap, and provide information on whether  
73 a gene is expressed in a certain tissue. Overrepresentation in sets of differentially expressed genes  
74 (DEG; sets of genes which are more (or less) expressed in a specific tissue compared to other tissue  
75 types) for each of 53 tissue types (Supplementary Table 3) based on GTEx v6 RNA-seq data<sup>16</sup> is  
76 also provided to identify tissue specificity of prioritized genes (Online Methods; Supplementary  
77 Table 3). Enrichment in biological pathways and functional categories is tested using the  
78 hypergeometric test against gene sets obtained from MsigDB<sup>17</sup> and WikiPathways<sup>18</sup>.

79 To validate the utility of FUMA, we applied it to summary statistics of the most recent GWAS for  
80 Body Mass Index (BMI)<sup>19</sup>(see Online Methods). FUMA identified 95 lead SNPs (from 223  
81 independent significant SNPs) across 77 genomic risk loci (Fig. 2 and Supplementary Table 4-6), in  
82 accordance with the original study. Functional mapping prioritized 151 unique genes; 23 genes with  
83 deleterious coding SNPs (positional mapping), 128 genes with eQTLs that potentially alter  
84 expression of these genes (eQTL mapping), and 16 genes that had both deleterious coding SNPs  
85 and eQTLs (Supplementary Table 7). The 151 genes include 55 genes that were also reported in the  
86 original study<sup>19</sup> and 96 novel genes implicated by FUMA (Fig. 2). These novel candidates have  
87 shared biological functions with the 55 previously known candidate genes such as ‘metabolism of  
88 carbohydrate’, ‘metabolism of lipid and lipoprotein’, ‘immune system’ and ‘calcium signalling’

89 (Supplementary Table 8). In addition, the FUMA results showed that, although several genomic  
90 loci for BMI included multiple prioritized genes, a single gene was prioritized in 22 loci, suggesting  
91 that these 22 genes have a high probability of being the causal gene in that region. The 22 ‘highly  
92 likely causal genes’ include several well-known genes for BMI such as *NEGR1*, *TOMM40* and  
93 *TMEM18* (Supplementary Fig.2 and Supplementary Table 7). The strongest GWAS association  
94 signal for BMI was on 16q.12.2 where 3 genes were prioritized; *FTO*, *RBL2* and *IRX3* (Fig. 3).  
95 These three genes were only prioritized by eQTL mapping as the positional mapping showed no  
96 deleterious coding SNPs located in these genes. The original study<sup>19</sup> only mentioned *FTO*, because  
97 the associated SNPs were located in this gene, however none of the associated SNPs have a  
98 potential direct affect such as coding SNPs on *FTO*. Two of the genes prioritized by FUMA (*RBL2*  
99 and *IRX3*) are physically located outside the genomic locus and are missed when using  
100 conventional approaches that prioritize genes located in the locus of interest based on LD around  
101 the top SNP. Although the *IRX3* gene was not reported in the original study<sup>19</sup>, recent functional  
102 work has indeed validated this as the causal gene whose expression is affected by SNPs in the  
103 16q.12.2 locus<sup>20</sup>. To assess whether the prioritized genes converge on biological shared functions or  
104 pathways, FUMA tested for enrichment in GO terms, and canonical pathways. 15 significantly  
105 enriched GO terms were detected, including known and novel pathways, e.g. ‘Zinc ion homeostasis’  
106 and ‘Glutathione related biological processes’ (Supplementary Table 10). Thus, using BMI  
107 summary statistics, FUMA confirmed known genes but also prioritized novel genes, including  
108 potential causal genes located *outside* the GWAS risk loci of BMI, which were missed in the  
109 original study.

110 To further illustrate its utility, we applied FUMA to the summary statistics of two other traits:  
111 Crohn’s disease<sup>21</sup> (CD) and Schizophrenia<sup>3</sup> (SCZ) (see Online Methods), where we obtained similar  
112 results: FUMA confirmed several genes that were reported in the original study, yet also prioritized  
113 genes that had not previously been reported (see Supplementary Results for details). For every  
114 prioritized gene, FUMA provides the reason for pinpointing this gene, such as for example when

115 the expression of the prioritized gene is altered by a SNP that is in LD with or associated with the  
116 disease of interest. Interactive regional plots (Supplementary Fig. 5-7, 10-11) show which genes in  
117 a genomic risk locus are prioritized and which genes are not, and the annotated SNPs in the  
118 prioritized genes facilitate the generation of hypotheses for functional validation experiments. For  
119 example, if a gene is prioritized because of an associated loss-of-function SNP, follow-up validation  
120 experiments focusing on a knock-out of this gene may provide disease relevant functional  
121 information. On the other hand, if a gene is prioritized because a risk associated allele of a SNP  
122 increases expression of this gene in brain, then an overexpression experiment of this gene in  
123 neuronal cell cultures would be a more relevant experiment.

124 In summary, FUMA provides an easy-to-use tool to functionally annotate, visualize, and interpret  
125 results from genetic association studies and to quickly gain insight into the directional biological  
126 implications of significant genetic associations. FUMA combines information of state-of-the-art  
127 biological data sources in a single platform to facilitate the generation of hypotheses for functional  
128 follow-up analysis aimed at proving causal relations between genetic variants and diseases.

129

130 **ACKNOWLEDGEMENTS**

131 This work was funded by The Netherlands Organization for Scientific Research (NWO VICI 453-  
132 14-005) and Ingrosyl. We thank the GIANT consortium, WTCCC and PGC for providing GWAS  
133 summary statistics and GTEx Portal for RNA-seq and eQTL data.

134

135 **AUTHOR CONTRIBUTIONS**

136 D.P. conceived the study. K.W. and A.v.B. developed the web application. K.W. performed  
137 analyses and drafted the manuscript. K.W., E.T., and D.P. participated in the discussions,  
138 interpretation of the results, and editing of the manuscript. All authors provided relevant input at  
139 different stages of the project and approved the final manuscript.

140

141 **COMPETING FINANCIAL INTERESTS**

The authors disclose no potential conflicts of interest.

142

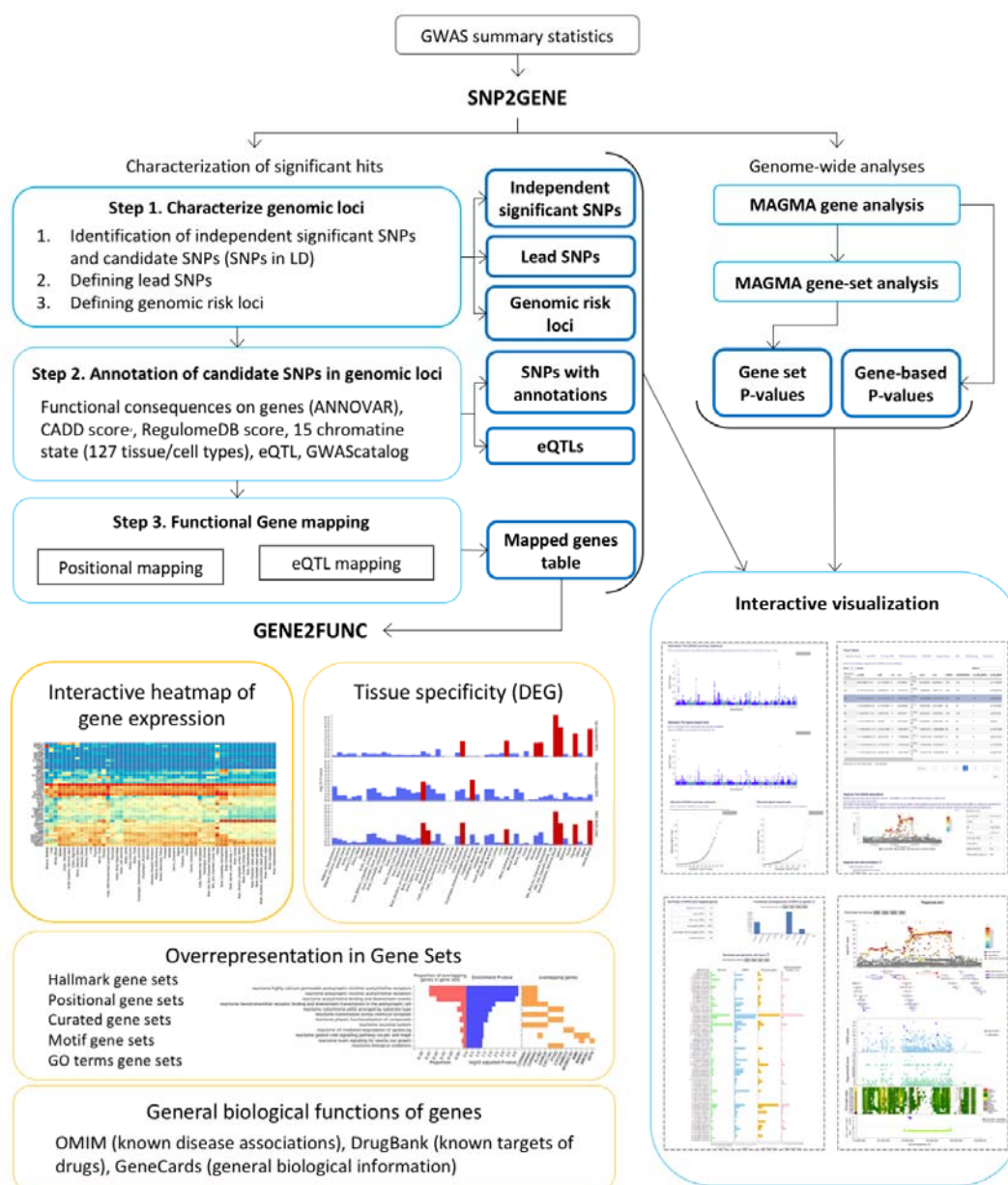


## 143 **References**

- 144 1. Welter, D. *et al. Nucleic Acids Res.* **42**, D1001–D1006 (2014).
- 145 2. Wood, A. R. *et al. Nat. Genet.* **46**, 1173–86 (2014).
- 146 3. Ripke, S. *et al. Nature* **511**, 421–427 (2014).
- 147 4. Okbay, A. *et al. Nature* **533**, 539–542 (2016).
- 148 5. Sudlow, C. *et al. PLoS Med.* **12**, 1–10 (2015).
- 149 6. Breen, G. *et al. Nat. Neurosci.* **19**, 1392–1396 (2016).
- 150 7. Maurano, M. T. *et al. Science (80-. ).* **337**, 1190–1195 (2012).
- 151 8. Wang, K., Li, M. & Hakonarson, H. *Nucleic Acids Res.* **38**, e164 (2010).
- 152 9. Kircher, M. *et al. Nat. Genet.* **46**, 310–315 (2014).
- 153 10. Boyle, A. P. *et al. Genome Res.* **22**, 1790–1797 (2012).
- 154 11. Ernst, J. & Kellis, M. *Nat. Methods* **9**, 215–216 (2012).
- 155 12. Roadmap Epigenomics Consortium. *Nature* **518**, 317–330 (2015).
- 156 13. The ENCODE Project Consortium. *Nature* **489**, 57–74 (2012).
- 157 14. Amberger, J. S., Bocchini, C. A., Schiettecatte, F., Scott, A. F. & Hamosh, A. *Nucleic Acids*  
158 *Res.* **43**, D789–D798 (2015).
- 159 15. Wishart, D. S. *et al. Nucleic Acids Res.* **34**, D668–D672 (2006).
- 160 16. The GTEx Consortium. *Science (80-. ).* **348**, 648–660 (2015).
- 161 17. Liberzon, A. *et al. Bioinformatics* **27**, 1739–1740 (2011).
- 162 18. Kutmon, M. *et al. Nucleic Acids Res.* **44**, D488–D494 (2016).
- 163 19. Locke, A. E. *et al. Nature* **518**, 197–206 (2015).
- 164 20. Claussnitzer, M. *et al. N. Engl. J. Med.* **373**, 895–907 (2015).

165 21. Franke, A. *et al. Nat. Genet.* **42**, 1118–25 (2010).

166



167

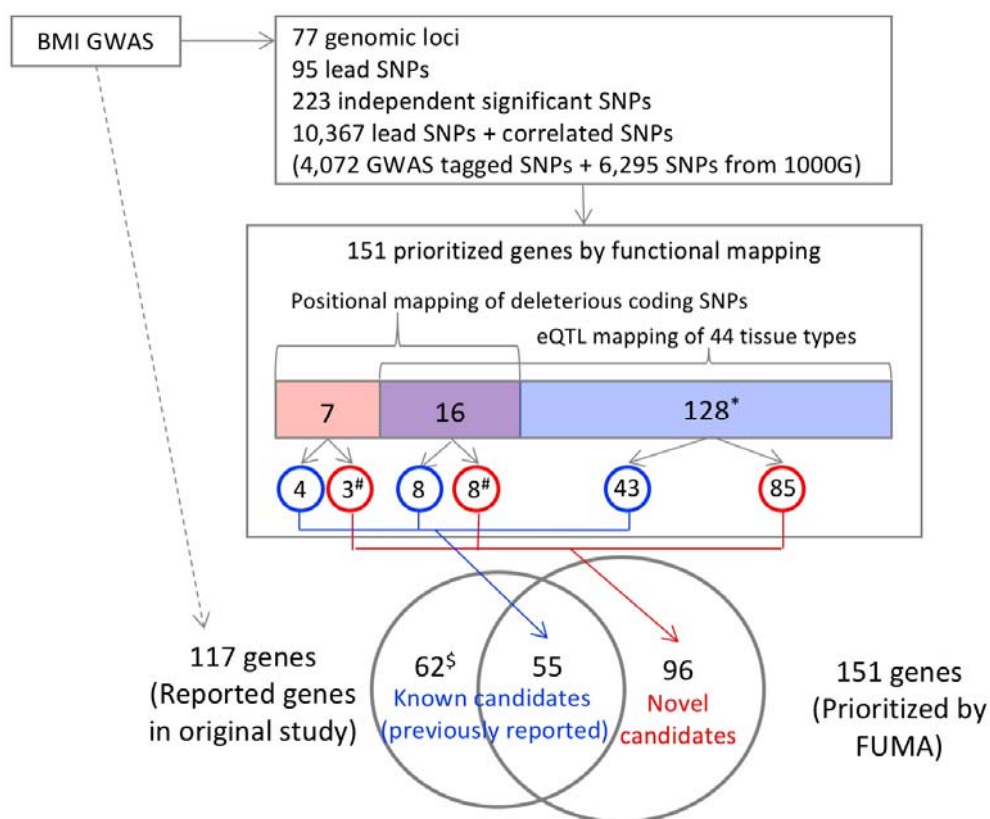
168 **Figure 1. Overview of FUMA**

169 FUMA includes two core processes, *SNP2GENE* and *GENE2FUNC*. The input is GWAS summary  
 170 statistics. *SNP2GENE* prioritizes functional SNPs and genes, outputs tables (blue boxes), and  
 171 creates manhattan, quantile-quantile (QQ) and interactive regional plots (box at right bottom).  
 172 *GENE2FUNC* provides four outputs; a gene expression heatmap, enrichment of differentially  
 173 expressed gene (DEG) sets in a certain tissue compared to all other tissue types, overrepresentation

174 of gene sets, and links to external biological information of input genes. All results are

175 downloadable as text files or high-resolution images.

176



177

178 **Figure 2. Overview of prioritized genes from BMI GWAS by FUMA**

179 Starting from the BMI GWAS summary statistics, boxes represent results of the *SNP2GENE*

180 process. The annotated SNPs include all independent lead SNPs and SNPs which are in LD with

181 these lead SNPs. Prioritized genes are divided into three categories; genes that are implicated by

182 deleterious coding SNPs (colored in pink), by eQTLs for these genes (colored in blue), or genes

183 implicated by both strategies (colored in purple). The prioritized genes are further categorized into

184 previously reported genes (blue circles) and novel genes (red circles) prioritized genes by FUMA.

185 \*50 of the 128 genes are located outside of GWAS risk loci. #These genes are located within the

186 GWAS risk loci (since they have coding SNPs) but were not reported in the original study because

187 of the following reasons: 1) FUMA considers all independent significant SNPs while only top SNPs

188 were considered in the original study, or 2) FUMA incorporates non-GWAS tagged SNPs which are

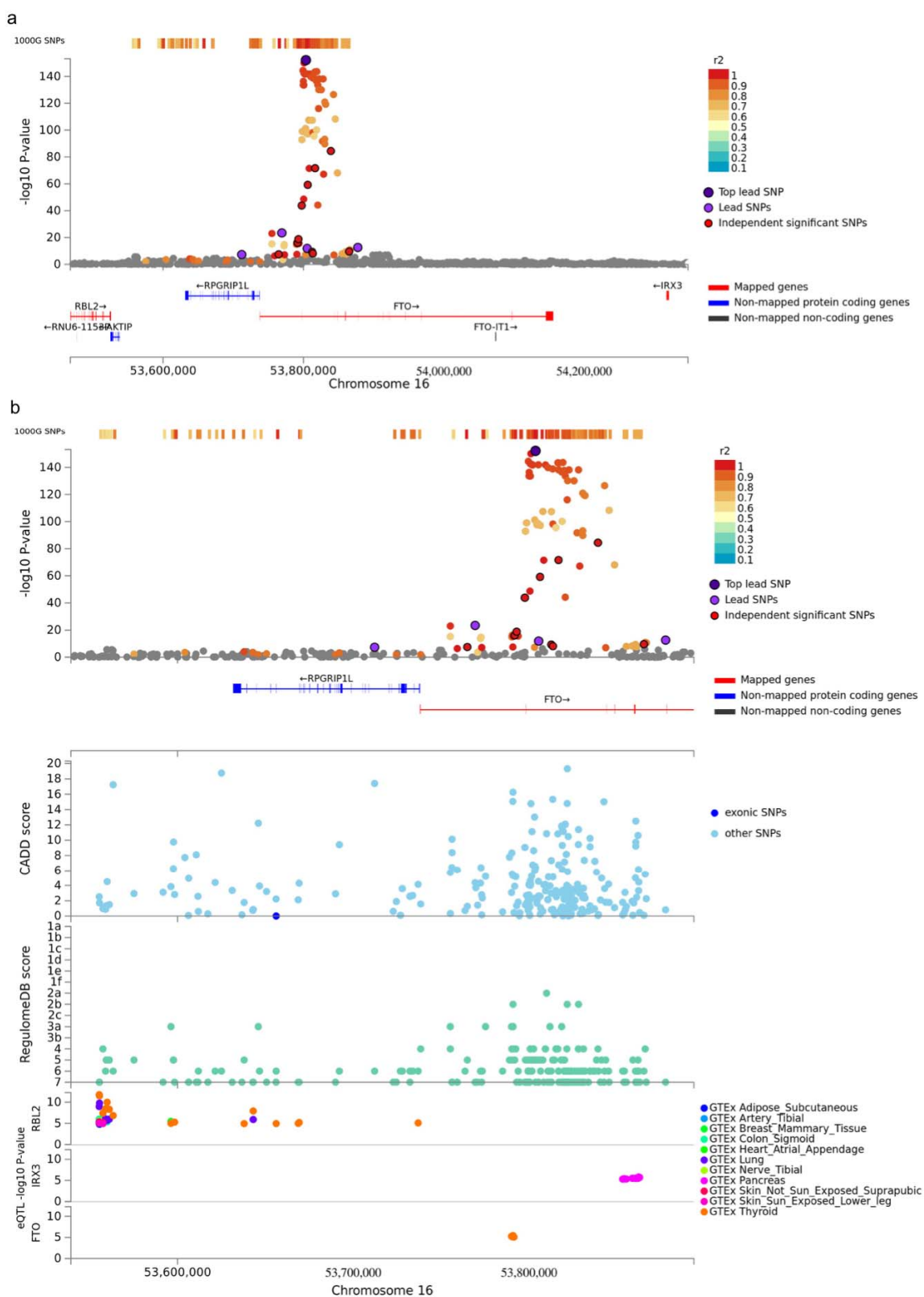
189 in LD of independent significant SNPs, or 3) reported genes do not necessary include all genes that

190 are located within GWAS risk loci because the authors only choose to highlight a subset. \$These

191 genes were not prioritized by functional mapping since they do not have either deleterious coding

192 SNPs or eQTLs, although they are located within GWAS risk loci.

193



194

195 **Figure 3. Regional plot of the locus 16q.12.2 of BMI GWAS and prioritized genes.**

196 (a) Extended region of the FTO locus, which includes prioritized genes *RBL2* and *IRX3*. Genes

197 prioritized by FUMA are highlighted in red. (b) Zoomed in regional plot of FTO locus with, from  
198 the top, GWAS P-value (SNPs are colored based on  $r^2$ ), CADD score, RegulomeDB score and  
199 eQTL P-value. Non-GWAS-tagged SNPs are shown in the top of the plot as rectangles since they  
200 do not have a P-value from the GWAS, but they are in LD with the lead SNP. eQTLs are plotted  
201 per gene and colored based on tissue types. From these results, it can be seen e.g. that SNPs that  
202 were not originally included in the GWAS, but are known to be in LD with the lead SNP using the  
203 1000 genomes reference panel, influence expression of *RBL2* in several different tissues. In addition,  
204 GWAS SNPs with a significant BMI association P-value and which are located in the FTO gene act  
205 as eQTL for expression of *IRX3* in the pancreas. The web-based version of this plot is interactive  
206 and allows zooming in or out as well as obtaining specific details about single SNPs.

207



## 208 **ONLINE METHODS**

### 209 **Data Sources and Pre-processes**

210 Data repositories and tools used in FUMA are available in Supplementary Table 1. All genetic data  
211 sets used in this study are based on the hg19 human assembly and rsIDs were mapped to dbSNP  
212 build 146 if necessary. To compute minor allele frequencies and LD structure, we used the data  
213 from the 1000 Genomes Project<sup>22</sup> phase3. Minor allele frequency and  $r^2$  of pairwise SNPs (up to  
214 1Mb apart) were pre-computed using PLINK<sup>23</sup> for each of available populations (AFR, AMR, EAS,  
215 EUR and SAS). Functional annotations of SNPs were obtained from the following three  
216 repositories; CADD<sup>9</sup>, RegulomeDB<sup>10</sup> and core 15-state model of chromatin<sup>11-13</sup>. Cis-eQTL  
217 information was obtained from the following 4 different data repositories; GTEx portal v6<sup>16</sup>, Blood  
218 eQTL browser<sup>24</sup>, BIOS QTL Browser<sup>25</sup> and BRAINEAC<sup>26</sup> and genes were mapped to ensemble  
219 gene ID if necessary. Genomic coordinate of GWAS catalog<sup>1</sup> reported SNPs was lifted down using  
220 liftOver software from hg38 to hg19. Normalized gene expression data (RPKM, Read Per Kilo base  
221 per Million) from GTEx portal v6<sup>16</sup> for 53 tissue types were processed for different purposes. The  
222 details are described in ‘GTEx Gene Expression Data Set’ section. Curated pathways and gene sets  
223 from MsigDB v5.2<sup>17</sup> and WikiPathways<sup>18</sup> which are assigned entrez ID.

224

### 225 **Characterization of genomic risk loci based on association summary statistics (step 1 in**

#### 226 ***SNP2GENE*)**

227 To define genomic loci of interest to the trait based on provided GWAS summary statistics, pre-  
228 calculated LD structure based on 1000G of the relevant reference population (EUR for BMI, CD  
229 and SCZ) is used. First of all, independent significant SNPs which have the genome-wide  
230 significant P-value ( $\leq 5e-8$ ) and independent from each other at  $r^2$  0.6. For each independent  
231 significant SNP, all known (i.e. regardless of being available in the GWAS input) SNPs that have  $r^2$   
232  $\geq 0.6$  with one of the independent significant SNPs are included for further annotation (candidate

233 SNPs). These SNPs may thus include SNPs that were not available in the GWAS input, but are  
234 available in the 1000G reference panel and are in LD with an independent significant SNP.  
235 Candidate SNPs can be filtered based on a user defined minor allele frequency ( $MAF \geq 0.01$ ).  
236 Based on the identified independent significant SNPs, lead SNPs were defined if they are  
237 independent from each other at  $r^2 \geq 0.1$ . Additionally, if LD blocks of independent significant SNPs  
238 are closely located to each other (less than 250kb, distance if based on the most right and left SNPs  
239 from each LD block), they are merged into one genomic locus. Each genomic locus can thus  
240 contain multiple independent significant SNPs and lead SNPs.  
241 Besides using FUMA to determine lead SNPs based on GWAS summary statistics, users can  
242 provide a list of pre-defined lead SNPs. In addition, users can provide a list of pre-defined genomic  
243 regions to limit all annotations carried out in FUMA to those regions.

244

#### 245 **Annotation of candidate SNPs in genomic risk loci (step 2 in *SNP2GENE*)**

246 Functional consequences of SNPs on genes are obtained by performing ANNOVAR<sup>8</sup> (“gene based  
247 annotation”) using Ensembl genes (build 85). Note that SNPs can be annotated to more than one  
248 gene in case of intergenic SNPs which are annotated to the two closest up- and down-stream genes.  
249 CADD, RegulomeDB score and 15-core chromatin state are annotated to all SNPs in 1000G phase  
250 3 by matching chromosome, position, reference and alternative alleles. eQTLs are also extracted by  
251 matching chromosome, position and alleles for each user selected tissue types, wherein SNPs can  
252 have multiple eQTLs for distinct genes and tissue types. Information on previously known SNP-  
253 trait associations reported in the GWAS catalog is also retrieved for all SNPs of interest by  
254 matching chromosome and position.

255

#### 256 **Gene Mapping (step 3 in *SNP2GENE*)**

257 Gene annotation is based on Ensembl genes (build 85). To match external gene IDs, we mapped  
258 ENSG ID to entrez ID yielding 35,808 genes which consist of 19,436 protein-coding genes, 9,249  
259 non-coding RNA and other 7,123 genes (e.g. pseudogenes, processed transcripts, immunoglobulin  
260 genes and T cell receptor genes).

261 Positional mapping is performed based on annotations obtained from ANNOVAR<sup>8</sup> for which we  
262 provide two options; maximum distance from SNPs to genes and functional consequences of SNPs  
263 on gene. When the former option is defined, FUMA maps SNPs to genes based on ANNOVAR  
264 annotation and a user defined maximum distance is applied for intergenic SNPs. Note that  
265 ANNOVAR prioritize an annotation of SNPs which are located in a genomic region where multiple  
266 genes are overlapped. For these SNPs, they are mapped to the annotated gene by ANNOVAR.  
267 When the latter option is provided, FUMA maps only SNPs which have selected annotations  
268 annotated by ANNOVAR.

269 For eQTL mapping, all independent significant SNPs and SNPs in LD of them are mapped to  
270 eQTLs in user defined tissue types. By default, only significant SNP-gene pairs ( $FDR \leq 0.05$ ) are  
271 used. Optionally, eQTLs can be filtered based on a user defined P-value. eQTL mapping maps  
272 SNPs to genes up to 1Mb apart (cis-eQTLs).

273 Optional filtering of SNPs based on functional annotations obtained in step 2 of *SNP2GENE* (i.e.  
274 CADD score, RegulomeDB score, 15-core chromatin state) can be performed for positional and  
275 eQTL mappings separately.

276

### 277 **Functional Mapping: identification of potential causal genes from functional SNPs**

278 We refer to “functional mapping” as the combination of positional mapping of deleterious coding  
279 SNPs, and tissue specific eQTL mapping. With functional mapping, we aim to further identify  
280 candidate causal genes based on biological function of SNPs. We include deleterious coding SNPs,

281 either being exonic or splicing with CADD score  $\geq 12.37$  (defined by Kircher *et al.*<sup>9</sup>), and eQTLs of  
282 defined tissue types (FDR  $\leq 0.05$ ).

283

#### 284 **MAGMA: Gene Analysis and Gene set Analysis**

285 In FUMA, input GWAS summary statistics is used to compute gene-based P-values (gene analysis)  
286 and gene set P-value (gene set analysis) by MAGMA<sup>27</sup> to provide a genome-wide distribution of  
287 genetic associations. For gene analysis, the gene-based P-value was computed for protein-coding  
288 genes by mapping SNPs to genes if SNPs are located within the genes. For gene set analysis, the  
289 gene set P-value was computed using gene-based P-value for 4,728 curated gene sets (including  
290 canonical pathways) and 6,166 GO terms obtained from MsigDB v5.2. For both analyses, the  
291 default setting (SNP-wise model for gene analysis and competitive model for gene set analysis)  
292 were used, and the Bonferroni correction (gene) or False Discovery Rate (gene-set) was used to  
293 correct for multiple testing.

294

#### 295 **GTEEx Gene Expression Data Set**

296 Normalized gene expressions (Reads Per Kilo base per Million, RPKM) of 53 tissue types were  
297 obtained from GTEEx (Supplementary Table 3). A total of 56,320 genes was available in GTEEx,  
298 which we filtered on an average RPKM per tissue greater or equal to 1 in at least one tissue type.  
299 This resulted in transcripts of 28,520 genes, of which 22,146 were mapped to entrez ID (see ‘Gene  
300 Mapping’ section for details). In the *GENE2FUNC*, the heatmap of prioritized genes displays two  
301 optional expression values; *i.* the average  $\log_2(\text{RPKM}+1)$  per tissue per gene, wherein RPKM was  
302 winsorized at 50, which allows comparison of expression level across genes and tissue types and *ii.*  
303 the average of the normalized expression (zero mean of  $\log_2(\text{RPKM}+1)$ ) per tissue per gene which  
304 allows comparison of expression level across tissue types within a gene.

305 To obtain differentially expressed gene (DEG; genes which are significantly more or less expressed  
306 in a given tissue compared to others) sets for each of 53 tissue type, the normalized expression (zero  
307 mean of  $\log_2(\text{RPKM}+1)$ ) was used. Two-sided Student's t-tests were performed per gene per tissue  
308 against all other tissues. After the Bonferroni correction, genes with corrected p-value  $\leq 0.05$  and  
309 absolute log fold change  $\geq 0.58$  were defined as a DEG set in a given tissue, i.e. for these gene  
310 expression in the given tissue had the largest discrepancy with expression in all other tissues. In  
311 addition, we distinguished between genes that were up- and down-regulated in a specific tissue  
312 compared to other tissues, by taking the sign of t-score into account. In *GENE2FUNC*, genes are  
313 tested against those DEG sets by hypergeometric tests to evaluate if the prioritized genes (or a list  
314 of genes of interest) are overrepresented in DEG sets in specific tissue types.

315

### 316 **Gene Set Enrichment Test**

317 To test for overrepresentation of biological functions of prioritized genes, the prioritized genes (or a  
318 list of genes of interest) are tested against gene sets obtained from MsigDB (i.e. hallmark gene sets,  
319 positional gene sets, curated gene sets, motif gene sets, computational gene sets, GO gene sets,  
320 oncogenic signatures and immunologic signatures) and WikiPathways, using hypergeometric tests.  
321 The set of background genes (i.e. the genes against which the set of prioritized genes are tested  
322 against) is 19,264 protein-coding genes. Background genes can also be selected from gene types as  
323 described in 'Gene Mapping' section. Custom sets of background genes can also be provided by the  
324 users. Multiple testing correction (i.e. Benjamini-Hochberg by default) is performed per data source  
325 of tested gene sets (e.g. canonical pathways, GO biological processes, hallmark genes). FUMA  
326 reports gene sets with adjusted P-value  $\leq 0.05$  and the number of genes that overlap with the gene  
327 set  $> 1$  by default.

328

### 329 **Validation with BMI GWAS**

330 GWAS summary statistics for the BMI GWAS were obtained from  
331 [http://portals.broadinstitute.org/collaboration/giant/index.php/GIANT\\_consortium\\_data\\_files](http://portals.broadinstitute.org/collaboration/giant/index.php/GIANT_consortium_data_files) and  
332 were used as input for FUMA. Parameters were set as described in the ‘Functional mapping’  
333 section and we used eQTLs in 44 tissue types from GTEx. Indels were excluded. rsID was mapped  
334 to dbSNP build 146 and chromosome and positions were extracted based on human genome hg19  
335 reference. Only protein-coding genes were used in gene mapping and enrichment of DEG in 53  
336 tissue types, Canonical Pathways and GO terms were tested.

337

### 338 **Application to CD GWAS**

339 GWAS summary statistics of CD was obtained from  
340 <ftp://ftp.sanger.ac.uk/pub/consortia/ibdgenetics/>. We set parameters as described in the ‘Functional  
341 Mapping’ section and we used eQTLs in 5 tissue types from GTEx which are relevant to CD, i.e.  
342 Small Intestine, Colon Sigmoid, Colon Transverse, Stomach and Whole Blood. The MHC region  
343 was excluded from the analysis. Since the input GWAS summary statistics only contained results  
344 from the discovery phase, we manually submitted the 71 reported lead SNPs to FUMA in addition  
345 to the independent lead SNPs that were identified as described above (Supplementary Table 11).  
346 Only protein-coding genes were used in mappings and enrichment of DEG in 53 tissue types,  
347 Canonical Pathways and GO terms were tested.

348

### 349 **Application to SCZ GWAS**

350 GWAS summary statistics were obtained from <http://www.med.unc.edu/pgc/results-and-downloads>.  
351 Parameters were set as described in the ‘Functional mapping’ section, and eQTLs in 10 brain  
352 tissues from GTEx. The extended MHC region (25Mb – 34Mb), Chromosome X and indels were  
353 excluded from this analysis. The input GWAS summary statistics are based on the discovery phase  
354 and not all reported lead SNPs from the combined results of discovery and replication phases

355 reached genome-wide significance. To include all reported lead SNPs, 111 non-indel lead SNPs  
356 were provided to FUMA and additional independent lead SNPs were identified at  $P \leq 5e-8$   
357 (Supplementary Table 19). Only protein-coding genes were used in mappings and enrichment of  
358 DEG in 53 tissue types, Canonical Pathways and GO terms were tested.

359

360 **References**

- 361 22. Auton, A. *et al. Nature* **526**, 68–74 (2015).
- 362 23. Purcell, S. *et al. Am. J. Hum. Genet.* **81**, 559–575 (2007).
- 363 24. Westra, H.-J. *et al. Nat. Genet.* **45**, 1238–43 (2013).
- 364 25. Zhernakova, D. V *et al. Nat. Genet.* **49**, 139–145 (2016).
- 365 26. Ramasamy, A. *et al. Nat. Neurosci.* **17**, 1418–28 (2014).
- 366 27. de Leeuw, C. A., Mooij, J. M., Heskes, T. & Posthuma, D. *PLoS Comput. Biol.* **11**, 1–19
- 367 (2015).
- 368