DeepWAS: Directly integrating regulatory information into GWAS using deep learning supports master regulator MEF2C as risk factor for major depressive disorder

Gökcen Eraslan[*,1], Janine Arloth[*,1,2], Jade Martins[2], Stella Iurato[2], Darina Czamara[2], Elisabeth B. Binder[2,4], Fabian J. Theis[1,3], Nikola S. Mueller[1,#]

[1] Institute of Computational Biology, Helmholtz Zentrum München Neuherberg 85764, Germany

[2] Department of Translational Research in Psychiatry, Max Planck Institute of Psychiatry, Munich 80804, Germany

[3] Department of Mathematics, Technische Universität München Garching 85748, Germany

[4] Department of Psychiatry and Behavioral Sciences, Emory University School of Medicine, Atlanta GA 30322, USA

[*] These authors contributed equally to this work.

[#]To whom the correspondence should be addressed: nikola.mueller@helmholtz-muenchen.de

Abstract

Background: Genome-wide association studies (GWAS) identify genetic variants predictive of common diseases but this does not directly inform on molecular mechanisms. The recently developed deep learning-based method DeepSEA uses DNA sequences to predict regulatory effects for up to 1000 functional units, namely regulatory elements and chromatin features in specific cell-types from the ENCODE project.

Results: We here describe “DeepWAS”, a conceptually new GWAS approach that integrates these predictions to identify SNP sets per functional units prior to association analysis based on multiple regression. To test the power of this approach, we use

genotype data from a major depressive disorder (MDD) case/control sample (total N=1,537). DeepWAS identified 177 regulatory SNPs moderating 122 functional units. MDD regulatory SNPs were located mostly in promoters, intronic and distal intergenic regions and validated with public data. Blood regulatory SNPs were experimentally annotated with methylation quantitative trait loci (QTLs), expression quantitative trait methylation loci and expression QTLs and replicated in an independent cohort. Joint integrative analysis of regulatory SNPs and the independently identified annotations were connected through transcription factors MEF2A, MEF2C and ATF2, regulating a network of transcripts previously linked to other psychiatric disorders. In the latest GWAS for MDD, the *MEF2C* gene itself is within the top genome-wide significant locus. **Conclusions:** DeepWAS is a novel concept with the power to directly identify individual regulatory SNPs from genotypes. In a proof of concept study, MEF2C was identified as a master-regulator in major depression, a finding complementary to recent depression GWAS data, underlining the power of DeepWAS.

Keywords: GWAS concept, functional data, chromatin features, cell-type specificity, ENCODE, deep learning, multiple regression, major depressive disorder, MEF2, MEF2C

Background

Genome-wide association studies (GWAS) have been highly successful in identifying genetic variants associated with risk for common diseases[1]. However, going from pure

47 association to mechanistic insight has been a much more challenging task. The
48 identification of the true functional/causal variants within association signals is
49 hampered by the linkage disequilibrium (LD) block structure of the genome. Due to this
50 structure, the true functional variants can most often not be discerned within associated
51 blocks, which can span several gene loci. One approach to better infer functional
52 variants has been the annotation of association signals by their correlation with more
53 proximal molecular read-outs such as gene expression or DNA methylation in the form
54 of expression and methylation quantitative trait locus (eQTL and meQTL)
55 approaches[2,3]. Very recent approaches such as binding QTL (bQTL) studies for
56 transcription factor (TF) binding now add additional regulatory information[4]. While
57 these approaches can indicate regulatory effects of associated SNPs, they can also not
58 pinpoint single functional variants within an LD block.
59
60 To further close the gap of missing functional annotation, additional post-processing
61 approaches, so called functional GWAS[5], have been used. The majority of the
62 previously published functional GWAS studies are based on the overlap of SNPs with
63 cis-regulatory elements such as promoters and enhancers (see Tak and Farnham[6] for
64 a comprehensive review). These functional GWAS indicate that for common disease,
65 the majority of associated SNPs reside in non-coding, regulatory regions[5,7,8]. One
66 drawback of these methods is that the actual impact of the variant on regulatory
67 elements is not assessed and thus not taken into account for the annotation, which
68 does not go beyond positional overlap. For example, two SNPs that localize to the same
69 chromatin immunoprecipitation with massively parallel sequencing[9] (ChIP-seq) peak

of a TF might have opposing effects or no effects at all. To try to resolve this, *in silico* approaches predicting disruption of transcription factor binding motifs[10,11] have been used, however, our understanding of actual binding based on known motifs is still limited. Given the fact that classical GWAS thus test many variants that are highly unlikely to be functionally relevant, methods integrating functional knowledge of SNP into GWAS could not only allow prioritize relevant variants but also to increase power for such association studies that now often need tens of thousands of cases for robust signals[1].

Recent advances in systems genetics, harnessing the predictive power of deep learning, might have the capacity to enhance the performance of functional SNP prioritization methods. Zhou J and Troyanskaya OG have developed a deep learning method called “DeepSEA” uses only DNA sequence information to predict effects on regulatory chromatin features such as histone marks, TF binding or the presence of open chromatin[12]. For this experimental, publicly available data from the ENCODE project[13] and the Roadmap Epigenomics Project[14] for cell type-specific TF binding, histone modifications and chromatin states was used. This type of functional sequence annotation is superior to pure overlap methods as it computes allele-specific differences in the effects on these regulatory elements and thus discerns SNPs with functional impact, at least in the given cell lines, from those just residing within the annotated element. Furthermore, it allows incorporating cell-type specific regulatory effects of such variants, adding another critical layer to understand disease mechanisms that are often tissue-specific.

93

94 In this manuscript we present a conceptually new approach fusing classical and
95 functional GWAS. We obtain regulatory information on SNPs by generating sets of
96 SNPs in "functional units" using *deep* learning and then performing functional unit-*W*ide
97 *A*ssociation *S*tudies (deepWAS). First, we extracted DNA sequences centred on a SNP
98 to predict close to 1000 allele-specific regulatory effects of chromatin features in various
99 cell types (a pair of one chromatin feature in one cell type is further called "functional
100 unit") using DeepSEA[12]. Second, the resulting significant regulatory SNPs were then
101 used to identify sets of SNPs characterized by their joint moderation of a functional unit.
102 Finally, we identify regulatory SNPs, short "deepSNPs", each coupled to a functional
103 unit by associating each set of predicted regulatory SNPs with a trait or disease. To that
104 end we use a multiple regression model with SNP selection using L1 ("LASSO")
105 penalization[15]. By testing regulatory SNPs within each of the confined functional units,
106 we controlled not only for correlations induced by LD but also for possible joint, thereby
107 again correlated, SNP effects on two or more chromatin features or cell types for which
108 LASSO would select one representative SNP. This optimized variant selection improves
109 our power to identify SNPs that may play a role in the etiology of the disease.

110

111 For a proof of concept, we used data from a published GWAS for major depressive
112 disorder (MDD)[16]. Heritability for this disorder has been reported to be up to 40%[17]
113 and thus comparable with other common diseases, GWAS have, however, only been
114 successful very recently when including over 110,000 cases and 300,000 controls
115 identifying 17 independent SNPs associated with MDD at genome-wide significant

116 level[18,19]. Nonetheless, the required sample size for each discovered disease locus is
117 much higher than for other psychiatric disorders such as schizophrenia, in which 128
118 independent SNPs have been identified with only 35,000 cases[20]. This difference may
119 be attributed to the possibly higher biological heterogeneity not covered with current
120 phenotyping methods and the large relevance of the environment as a risk factor for
121 MDD. A recent study from our group has reported the importance of genetic variants in
122 specific enhancer regions conferring risk for major depression, likely by gene x stress
123 interactions[21]. DeepWAS, with its focus on variants altering functional regulatory
124 elements, could thus help to promote our understanding of the genetic mechanisms
125 underlying depression risk.
126

127 Results

128 DeepWAS approach

129 To identify SNP sets that potentially affect common regulatory functions in a cell type
130 specific manner, we need to investigate their genomic location and sequence alterations
131 within so called “functional units”. We denote a “functional unit” as a gene-regulatory
132 state of a specific cell type characterized by the mapping of one chromatin feature onto
133 the genome in the absence or presence of additional stimulating/treatment conditions,
134 see (**Figure 1A**). For example, the functional unit “p53-HeLa” was defined by all
135 genomic regions covered by ChIP-seq peaks of the TF p53 within baseline conditions in
136 the HeLa cell line. For this manuscript, the 919 DeepSEA features [12] are considered

137 as functional units representing data from the ENCODE project [13] covering
138 combinations of 201 different experimental annotations of epigenetically relevant
139 information. This includes the mapping of different TFs, various DNase hypersensitive
140 (DHS) regions and histone marks across 31 cell lines and 17 treatment conditions.
141 Next, one kilobase DNA sequences were generated for each SNP for both reference
142 and alternative alleles. We used the pre-trained DeepSEA model[12] to predict the
143 probabilities of a SNP allele to moderate a functional unit. DeepSEA is a method that
144 uses deep neural networks and was trained to predict membership of 919 functional
145 units given a one kilobase genomic sequence. To select SNPs moderating a functional
146 unit, we used the metric provided by DeepSEA that estimates the impact of a variant on
147 the functional read out by comparing the probabilities of allele-specific regulation[12].
148 Therefore, the authors used one million random SNPs from the 1,000 Genomes project
149 [22] as a background distribution to calculate e-values for each functional unit by
150 assessing the proportion of random variants with a bigger effect than that of the
151 observed variants. SNPs with significant e-values, thus potentially having allele-specific
152 effects on a functional unit, were defined as a functional-unit specific SNP set. In other
153 words, only SNPs with a predicted and significant difference in allele-specific regulatory
154 effect for a given genomic feature in a cell type (functional unit) were retained for the
155 subsequent association analysis. When overlapping the available genotypes (measured
156 and imputed) in the case/control samples with the functional-unit specific SNP sets, 919
157 sets of genotypes filtered by regulatory effect in a functional unit remained for the
158 analysis (**Figure 1B**).
159

160 For the final step in the DeepWAS, we used the functional genotypes of the case/control
161 sample and possible confounders to associate SNPs set to a disease or trait (**Figure**
162 **1C**). The number of tested SNPs is massively reduced using the above-described
163 functional annotation based on ENCODE data and this variant filtering now enables us
164 to employ regularized regression methods with multiple SNPs. Therefore, each set of
165 regulatory SNPs in one functional unit was subjected to a logistic regression model with
166 L1 regularization (LASSO). In other words, we fitted 919 functional unit-based models to
167 again select regulatory SNPs associated to a trait or disease. For models with at least
168 one non-zero coefficient for a SNP, we further implemented a permutation-based
169 significance test with controlled false positive rate. From all models that withstood this
170 permutation-based multiple testing correction, we finally identified the deepSNPs that
171 are defined as significant non-zero regulatory SNPs associated to the trait or disease
172 and moderating a functional unit. The advantage of DeepWAS over traditional GWAS
173 (**Figure 1D**) is thus twofold, 1) it includes putative regulatory mechanisms in the GWAS
174 analysis from the start and 2) it controls false discovery error by reducing multiple
175 testing.
176

177 DeepWAS of major depressive disorder

178 We used post-quality control genotype data from 1,537 individuals (739 controls and
179 798 cases) of a cohort recruited for recurrent major depressive disorder (MDD). The
180 data were imputed to the 1000 Genomes Phase 1 reference panel [22] using SHAPEIT2
181 [23] and IMPUTE2 [24]. The resulting data set contained more than 1.7 million variants

182 with MAFs of at least 5% and non-missing genotypes. As expected, a classical GWAS
183 approach failed to identify genome-wide significant associations [16] (**Additional file 1:**
184 **Figure S1**).
185
186 For the DeepWAS approach, a total, 6,685 SNPs and 50,704 SNP-functional unit pairs
187 were retained after filtered for allele-specific effects in a given functional unit at a
188 significance level of $5 \cdot 10^{-5}$. These SNPs were then tested for association with
189 depression within 919 LASSO models for each functional unit (see above) including age
190 and sex as additional covariates (number of SNP predictors in 919 LASSO models:
191 min=17, median=51, max=213. See **Additional file 1: Figure S2**). Of these models,
192 193 had at least one non-zero coefficient for a SNP (number of SNPs with non-zero
193 coefficients in 193 models: min=1, median=2 and max=22). Permutation-based
194 significance test identified 122 out of these 193 models to be below the FDR-adjusted p-
195 value threshold of 0.1 (**Additional file 1: Figure S2** and **Additional file 2 and 3**). In
196 other words, deepWAS identified 122 functional units associated with MDD containing
197 in total 177 deepSNPs moderating 74 chromatin features in 31 cell lines.
198
199 We first analysed, whether the MDD deepSNPs were enriched for specific chromatin
200 features or tissue (**Figure 2A**). Not surprisingly, cell lines with a large number of
201 investigated chromatin features such as GM12878, K562 and HepG2 exhibited more
202 significant models since more chromatin features are covered in the data generated
203 from these cell lines. Interestingly, TFs such as c-Myc and EBF1 that are more relevant
204 to cell proliferation did not show significant associations. CTCF on the other hand, an

205 transcriptional regulator relevant for chromatin conformations important for the function
206 of long-range enhancers showed consistent effects across many cell lines. In fact,
207 genetic variants in long range enhancer have been shown to be highly relevant for
208 common disorders [14] and also major depression [21]. Furthermore, NF-κB and
209 glucocorticoid response element related deepSNPs only become apparent in the
210 stimulation condition, underlining the importance of going beyond pure baseline
211 characterization. Both systems and transcription factors, NF-κB – a major immune
212 regulator and the glucocorticoid receptor – central for the stress response, have been
213 previously implicated in the pathogenesis of MDD [21,25,26] (**Figures 2A and**
214 **Additional file 1: Figure S2**).
215
216 We next investigated the genomic regions in which these deepSNPs reside. Using the
217 UCSC knownGene annotations [27], we observed that vast majority of deepSNPs are in
218 promoters, intronic regions and distal intergenic regions that might indicate enhancers
219 (**Figure 2B**). We then examined the enrichment of deepSNPs in tissue-specific *cis*-
220 regulatory elements using genome annotations of the 15-state ChromHMM model[14].
221 The promoters are significantly enriched for deepSNPs in 126 out of 127 epigenomes
222 (99%) spanning all tissue groups, whereas the enhancers are only enriched in
223 “Astrocytes Primary Cells - E125” (FDR-adjusted p-value: 0.093) and in “A549 EtOH
224 0.02pct Lung Carcinoma Cell Line - E114” (FDR-adjusted p-value: 0.006) epigenomes
225 (**Additional file 1: Figure S4**).
226

227 To test whether the deepSNPs could be validated for their predicted allele-specific
228 effects in experimental data, we evaluated the effect of cell line-matched deepSNPs on
229 TF binding using previously published binding QTLs (bQTL) [4]. The bQTLs were
230 identified in GM12878 (Lymphoblastic cell line (LCL) from B-Lymphocytes) using a
231 pooled ChIP-seq approach of 60 Lymphoblastic Yoruban cell lines to identify SNPs
232 significantly altering binding of TFs by comparing the allele frequencies in sequencing
233 data before and after the ChIP experiment. Out of 26 deepSNPs in GM12878, we
234 identified eight SNPs (31 %) that were also experimentally validated bQTLs for JunD,
235 Pu.1, NF-κB and POU2f1 (**Additional file 4**).
236

237 Blood deepSNPs overlap with methylation and expression QTLs

238 Allele-specific effects on chromatin features likely associate with altered transcriptional
239 activation and this is reflected in differences in DNA methylation levels and gene
240 transcription. To test, whether the deepSNPs are indeed preferentially transcriptionally
241 relevant variants, we used DNA methylation and mRNA expression data from peripheral
242 blood in the recMDD samples (see Methods) as well as a second independent sample
243 (MPIP, see Methods). We focused on the detailed analysis of 26 GM12878 deepSNPs
244 (**Figure 3**), as GM12878 is among the best-characterized cell lines, derived from blood
245 cells, B-lymphocytes and amongst the top regulatory cell line identified by the MDD
246 deepWAS.
247

248 First, we conducted a methylation quantitative trait locus (meQTL) analysis using whole
249 blood methylation levels of a subset of the recMDD cohort (n=257 individuals). SNP-
250 CpG associations were carried out in a 3Mb *cis*-window around the deepSNPs to
251 assess the effect of nearby CpG sites. From the 26 deepSNPs, we identified 17
252 significant meQTLs containing seven unique deepSNPs (27 %) and 17 CpGs after
253 Bonferroni-correction for all tested CpGs per SNP (**Table 1**). The majority (>71%) of the
254 meQTL SNPs were associated with more than one DNA methylation site (one to four
255 CpG sites per SNP, **Table 1**). Notably, 25 deepSNPs (96 %) were nominally associated
256 with differential methylation levels ($p<0.05$). The most significant meQTL is located in an
257 enhancer region and all four identified CpGs in regions epigenetically associate with
258 gene silencing (**Figure 4A**). The genotypes at the intergenic SNP rs9293528 were
259 significantly associated with the methylation levels of four moderately correlated CpGs
260 located approximately 30-55 kb away from the SNP: cg1193710, cg14617041,
261 cg14727987 and cg26653990 (**Figures 4B** and **4C**). Sixteen out of the 17 meQTLs
262 were validated in the independent cohort (MPIP cohort, n=229 samples, see Methods),
263 i.e., showed a significant meta-analysis p-value (see Methods) showing equally or more
264 significant associations than in the discovery cohort alone (**Figure 4D** and **Table 1**).
265
266 Interestingly, deepSNPs are also associated with MDD-specific methylation changes. In
267 total we found five deepSNP-CpG pairs with significant differences in the allele-specific
268 methylation profiles between MDD cases and controls in recMDD cohort (see Methods).
269 Two of these MDD specific meQTLs are replicated in the independent MPIP cohort (see
270 **Additional file 1: Table S1** and **Additional file 1: Figure S5**).

271

272 Second, we investigated the effects of the identified 17 CpG methylation sites
273 associated to MDD deepSNPs on matched gene expression levels, if CpG and
274 transcripts were within 1.5Mb of genomic distance. This was investigated in the MPIP
275 cohort, as gene expression data were only available in this sample. Among these, we
276 identified three expression quantitative methylation loci (eQTM) comprising two CpGs
277 and three gene expression probes after genome-wide Bonferroni-correction for the
278 number of tested transcripts per CpG (**Additional file 1: Table S2**) and 11 eQTMs (8
279 CpGs and 9 transcripts) at nominal level. These are: cg19235974-*CACNA2D4 (Calcium*
280 *voltage-gated channel auxiliary subunit alpha2delta 4*), cg19235974-*CACNA1C (Calcium*
281 *voltage-gated channel subunit alpha1 C)* and cg21290162-*MSRA (Methionine sulfoxide*
282 *reductase A*). *CACNA1C* has been associated with schizophrenia, bipolar disorder and
283 major depression[28], *CACNA2D4* with late onset bipolar disorder[29] and *MRSA* with
284 both bipolar disorder and schizophrenia [30,31].

285

286 Finally, we associated the 26 deepSNPs to gene expression directly, by conducting an
287 expression quantitative trait locus (eQTL) analysis using the MPIP cohort data (n=289
288 samples). We examined transcripts within a *cis*-window of 1.5Mb upstream and
289 downstream of the deepSNPs for an association with whole blood gene expression
290 profiles. We identified three deepSNPs with significant eQTL: rs12541159-*MSRA*,
291 rs1868881-*TIMM10* and rs4646797-*ALDH3A2* (**Additional file 1: Table S3**).
292 Interestingly, at nominal significance level we found 27 eQTLs (12 deepSNPs and 27
293 transcripts).

294

295 Overall, the investigation of the functional consequences of a hematopoietic tissue-
296 relevant set of deepSNPs in experimental data indicates that these indeed preferentially
297 tag meQTLs and eQTLs and are thus transcriptionally relevant.

298

299 Predicted regulatory effects of deepSNPs in MDD

300 Finally, we integrated the analysis results of meQTLs, eQTLs, eQTMs in peripheral
301 blood with our GM12878 deeSNPs to explore power of the deepWAS approach for
302 identifying novel, putative disease mechanisms. We combined all pairwise links of
303 meQTL (SNP-CpG), eQTL (SNP-gene), eQTM (CpG-gene) and deepSNP-chromatin
304 features and illustrated them in a network (**Figure 5A**).

305

306 One of six deepSNPs, rs12541159, had a meQTL and eQTL and at the same time
307 harboured an eQTM, in which the CpG site was affecting the transcriptional levels of the
308 same gene (**Figure 5B**). By regulating the methylation and expression profiles of
309 multiple genes via common TFs, deepSNPs may represent putative master regulators
310 of in a given disease. The intergenic deepSNP rs12541159 was predicted to affect
311 binding of the MEF2C TF, correlated with blood methylation levels of two intergenic
312 CpGs (cg21290162 cg11269159) and one exonic CpG cg27411982 with blood gene
313 expression of *MSRA* resigning 517 kb (**Figure 5B,** box 1 and **5C**). The CpGs
314 themselves showed no correlation to each other (**Figure 5D).** The transcriptional levels
315 of *MSRA* were associated with increased cg21290162 methylation (**Figure 5E**). This,

316 motivated us to infer causality and we investigated whether the CpG site could fully
317 explain the observed association between the SNP and *MSRA* expression using causal
318 inference testing[32]. We found partial mediation of the effect of rs12541159 on *MSRA*
319 expression by DNA methylation status of CpG site cg21290162 (**Figure 5E**, p = 0.043).
320
321 Strikingly, the deepWAS results connected the independently identified six SNPs of the
322 nine molecular QTLs (see Figure 5A) through a family of TFs, namely MEF2A (*myocyte*
323 *enhancer factor 2A*), MEF2C (*myocyte enhancer factor 2C*) and ATF2 (*activating*
324 *transcription factor 2*). Predicted regulatory effect to identical SNPs suggest TF co-
325 localization, which has been not been reported so far. We thus utilized ReMap
326 annotation tool [33] to study co-localization and identified a significantly overlap of TF
327 ChIP-seq peaks for these three TFs (**Additional file 1: Figure S6**). Interestingly, a
328 SNP in the locus encompassing the MEF2C gene is the top locus of the largest GWAS
329 for MDD so far, with a $p < 10^{-16}$[19]. Similarly, we identify a regulatory effect of a second
330 TF, P300, of which the respective EB300 gene was a GWAS locus in the discovery
331 cohort[19].
332

333 Discussion

334 In classical GWAS all SNPs are tested independently and in a fully genome-wide
335 manner, thereby implicitly assuming that any SNP can moderate the function of any cell
336 state at any time. It is now clear that disease associations especially to common
337 disorders are driven by SNPs altering the function of regulatory elements. Hence it is

338 likely not necessary to test all SNPs in GWAS but instead useful to focus on functional
339 annotation to help prioritize putative risk variants[14].
340
341 In this manuscript, we address this idea by directly integrating the functional data into
342 the GWAS approach itself. To that end, we employed the powerful DeepSEA method
343 [12] that uses raw DNA sequence data to predict regulatory effects of 919 chromatin
344 features (from the ENCODE project) in various cell types, termed “functional unit” here.
345 These predictions enable determining allele specific effects of single SNPs by
346 comparing predictions of reference and alternative allele sequences (**Figure 1**). As
347 deep learning refers to the idea of learning multiple levels of representations and
348 relevant features from the raw input, it eliminates the need for manual feature
349 extraction, such as extracting k-mer frequencies for DNA sequence analysis [34]. To our
350 best knowledge, our study is the first to use deep learning-based predictors to identify
351 regulatory SNPs in GWAS. All identified deepSNPs were predicted to have a specific
352 regulatory effect of a chromatin feature in a defined cell type (functional unit), thereby
353 allowing having directly testable hypotheses of regulatory mechanisms in a trait or
354 disease.
355
356 Our deepWAS approach is superior to simply overlapping peaks or tracks of chromatin
357 features to GWAS signals, even though it is computationally more expensive due to the
358 deep learning-based predictions, cross-validated LASSO models and permutation-
359 based model selection. In contrast to annotation-based methods, data-driven methods,
360 such as DeepSEA, have a high predictive power on a single base resolution (by

361 comparing predictions of reference and alternative allele sequences) and not only for
362 larger stretches of sequence underlying one or several chromatin features. It thus
363 enables us to systematically assess the effects of sub-threshold SNPs that are missed
364 in classical approaches. DeepWAS does not identify genomic loci containing many
365 possible regulatory SNPs, but identifies single deepSNPs with predicted allele-specific
366 regulatory effect in a function unit, a cell-type and chromatin feature.
367
368 The second major methodological advantage of deepWAS is the implementation of
369 multiple (multi-SNP) regression models with L1 regularization (called LASSO
370 regression) selecting only a few SNPs, so-called deepSNPs. The deepSNPs of one
371 functional unit are only selected when they jointly associate with the disease or trait.
372 Multiple regression models are relatively new in the field of GWAS, nonetheless already
373 show promising result [35,36]. A related multi-SNP approach used pre-clustering of LD
374 blocks prior to LASSO modelling [37]. Models with millions of SNPs are less powerful to
375 associate SNPs, most probably due to the high correlation structure induced by LD. The
376 power of a pre-selection of regulatory SNP sets is underlined by the fact that when all
377 predicted regulatory SNPs (neglecting the information on functional units) were used as
378 input to a single LASSO regression model only one SNP (rs8180478 in osteoblast-
379 H2A.Z functional unit) showed significant association with MDD. This is in contrast to
380 the 177 deepSNPs identified using the pre-selection models.
381
382 Currently deepWAS has two main limitations. First, the proposed study was designed
383 for best-guess genotype data (i.e. 0-1-2 encoding). Since LASSO models require full

384 genotype information for all samples, the numbers of SNPs entering the analysis had to
385 be reduced due to quality control procedures. To address this issue, we are planning to
386 extend the method for dosage data (i.e. probabilistic representation of genotypes) to
387 increase the range of predicted regulatory effects. This will also necessitate as re-
388 evaluation of the LASSO model as many more SNPs in high LD will enter the analysis.
389 The second limitation is the current comprehensiveness of regulatory element
390 catalogues like ENCODE and Roadmap. ENCODE lacks for example a number of
391 relevant disease-specific stimulation conditions as well as disease-related tissues, in
392 our case brain tissues, which are important for psychiatric disease. Our previous and
393 related studies indicated the importance to test SNPs in stimulated conditions, like
394 TNFalpha and glucocorticoids receptor agonist dexamethasone conditions activating
395 NF-κB and GR, respectively [21]. Data from cell lines or bulk tissues will miss variants
396 with effects only in rare cell types or cell type specific effects in native tissue. It is
397 therefore important to be able to retrain the DeepSEA neural network with additional
398 publicly available chromatin features and as well as newly generated experimental data.
399 This will be possible using the deepWAS code publicly available at DOI:
400 10.5281/zenodo.59282. Additional data, from efforts such as the PsychENCODE will
401 make deepWAS even more powerful in the future.
402
403 The power of DeepWAS could be demonstrated in our small GWAS sample for
404 MDD[16]. Only using a subset of SNPs in functionally confined, regulatory units, we
405 identify risk variants for MDD even though the classical GWAS in this small sample was
406 negative (**Additional file 1: Figure S1**). The results in this small MDD sample illustrate

407 the power of deepWAS. 1) DeepSNPs were identified in cell types and enhancers
408 previously shown to be relevant for depression. Furthermore, using the stimulation
409 conditions, the importance of both immune system regulator NF-κB and the
410 glucocorticoid receptor system for depression was also observed in this study. Both
411 systems have been implicated by a large number of previous publications [21,25,26]. 2)
412 deepSNPs for MDD were highly likely to be functional in experimental data and impact
413 both DNA methylation and gene expression (**Figure 4** and **Table 1, Additional file 1:**
414 **Table S2 and Table S3**) as shown by their effects meQTL, eQTL and eQTM analyses.
415 Interestingly, for some of these, differential genotypic effects by major depression status
416 were observed (**Additional file 1: Figure S5** and **Table S1**). 3) The integrated analysis
417 of eQTL, meQTL and eQTM results with the GM12878 deepSNPs, identified the MEF2
418 TF family, including MEF2C, as important risk factors in MDD (**Figure 5**). The MEF2 TF
419 family was already identified to play a major role in neuronal plasticity, which is an
420 important component in disease development of stress-related disorders, like MDD and
421 other psychiatric disorders. Chen and colleagues[38] identified the TF MEF2 as a
422 master regulator of developmental metaplasticity, which is important to guide
423 developmental structural and functional neuronal plasticity. Additional evidence was
424 found by Barbosa et al [39]. Relating MEF2 to activity-dependent dendritic spine growth
425 and suggesting that this TF may suppress memory formation [39,40]. Most relevant for
426 this study, SNPs in the locus encoding MEF2C are the top signal in the latest meta-
427 analysis for major depression with over 130,000 MDD cases and 310,000 controls [19]
428 and minimum p-value of $9.99*10^{-16}$. Our study now identified SNPs altering the binding
429 of this TF to target transcript as relevant for MDD. This implies that not only SNPs in the

430 MEF2C locus itself, as seen in the meta-analysis for MDD [19] but also SNPs altering
431 it's binding in target transcripts, as observed in our deepWAS, are associated with major
432 depression. This shows that our approach focused on regulatory SNPs was able to
433 generate complementary functional data to the largest classical GWAS for MDD to date
434 suggesting that not only transcriptional regulation of MEF2C itself but also of its
435 downstream targets is relevant for MDD.
436
437 Our results supports the transcription factor MEF2C as a master regulators in MDD. In
438 fact we observed that transcripts affected by differential MEF regulation included a
439 network of genes previously implicated in psychiatric disorders. This includes
440 *CACNA1C*, a gene encoding a voltage gated calcium channel that has reported
441 genome-wide significant associations for schizophrenia as well as associations for
442 bipolar disorder and major depression in previous studies[28] as well as another
443 voltage-gated calcium channel, *CACNA2D4* that has been linked to late onset bipolar
444 disorder [29]. *MRSA* has been associated with both bipolar disorder and schizophrenia
445 [30,31] .
446

447 Conclusions

448 Our data indicate that deepWAS, a method combining classical GWAs with deep
449 learning based functional SNP annotation is a powerful tool to uncover disease
450 mechanisms for common disorders, including relevant cell types. It also allows to
451 directly identify functional SNPs by having a single base resolution and not being limited

by the LD structure of the locus. With ever increasing amounts of available functional data the deepWAS approach will become even more valuable and allow to integrate both publicly as well as unpublished data generated by individuals labs. DeepWAS is a versatile tool, publicly available and together with available code from DeepSEA applicable to any GWAS data.

Figures

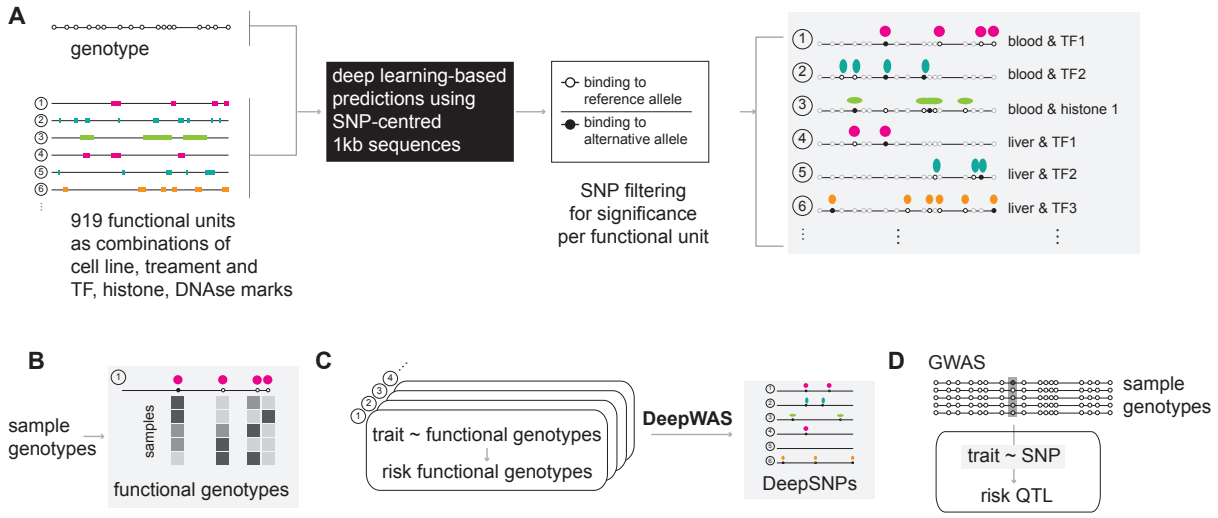


Figure 1

Figure 1: **DeepWAS approach. (A)** The first step of the workflow corresponds to a regulatory variant filtering procedure where only the SNPs with significant regulatory effect are retained in the analysis. Prediction of significant regulatory variants is performed using a pretrained deep neural network, namely DeepSEA, and e-values are calculated by comparing the regulatory effects of observed SNPs to randomly selected

466 SNPs. **(B)** We define the functional genotype as the intersection of all SNPs assessed
467 (genotyped and imputed) in a sample and the preselected SNPs from DeepSEA. **(C)**
468 The association between disease and retained regulatory variants are interrogated via
469 regularized logistic regression. A regression model is fitted for each functional unit that
470 is define by tested TF/histone mark/DNase effects, the cell line and the specific
471 treatment e.g. NF-κB -GM12878-TNFa) where only the SNPs with a regulatory effect in
472 the specific cell type context are included as covariates. **(D)** The classical GWAS design
473 focuses on allele frequency differences and does not include information on the
474 regulatory impact of the tested variants.
475

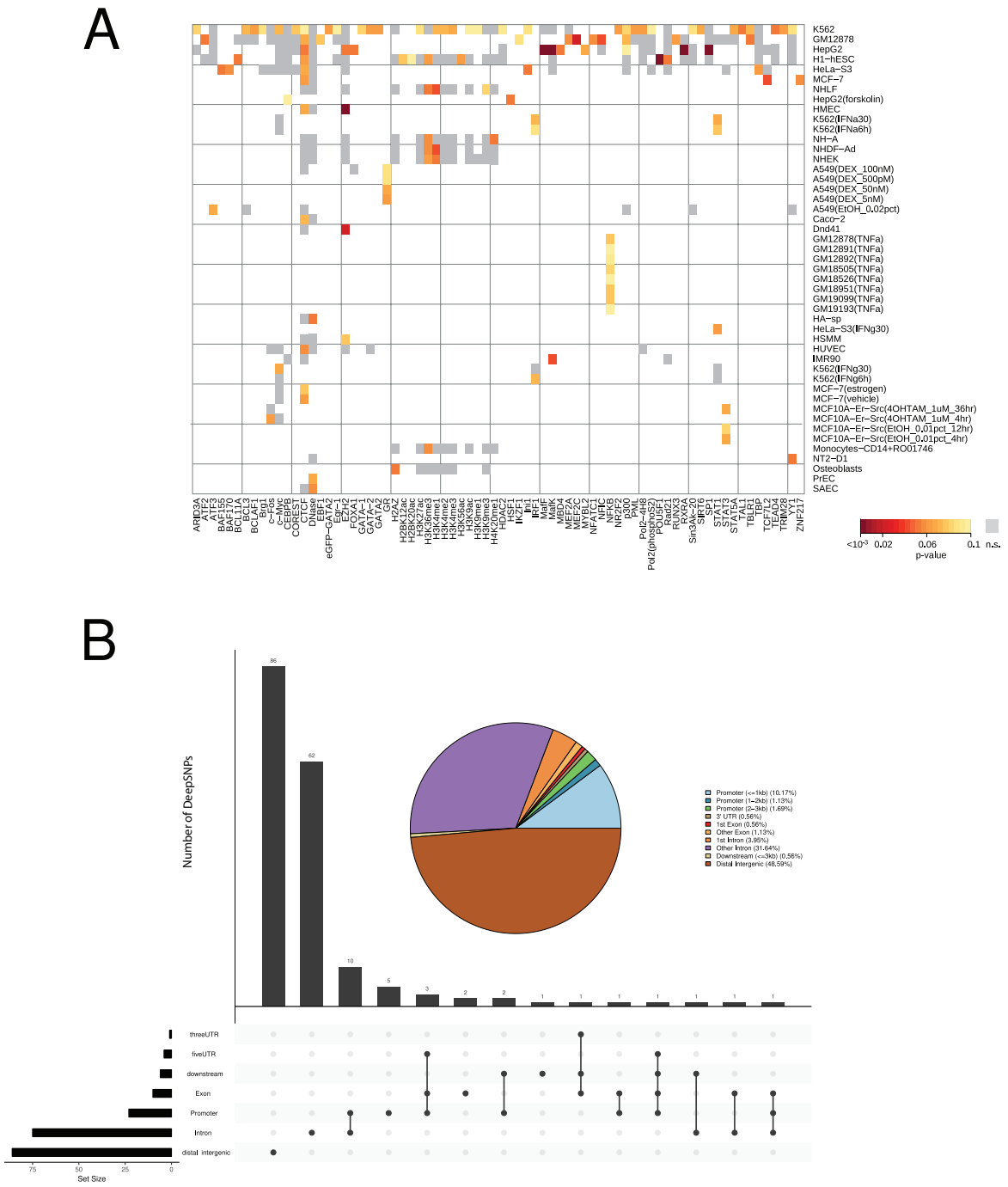
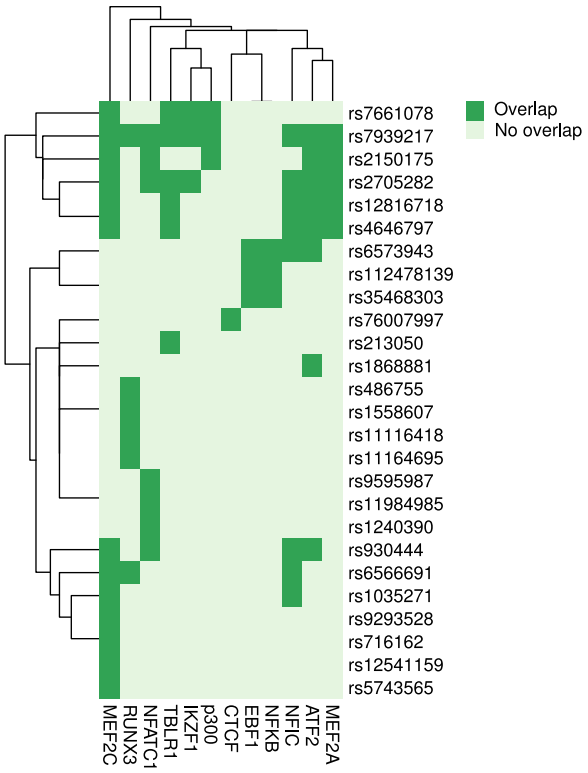


Figure 2: **Significance of DeepWAS models and overlap of deepSNPs with functional annotations (A)** Results of the significance test is given in heatmap. Every functional unit is represented by the p-value, which shows the FDR-adjusted significance of the corresponding LASSO model. Missing values, represented in white,

482 show functional units for which no data were available. The models of the functional
483 units with LASSO coefficients with p > 0.1 and those without any non-zero coefficients
484 are shown in grey and those with p < 0.1 in shades from yellow to red, with red
485 representing the most significant models. **(B)** Overlap of deepSNPs with genomic
486 elements is performed using UCSC knownGene annotations. Vast majority of
487 deepSNPs overlap with non-coding genomic regions.



488

489 **Figure 3: Transcription factors harbouring deepSNPs in GM12878 cell line.** In data
490 from GM12878 26 deepSNPs are predicted to affect the binding to 12 TFs. Group of
491 SNPs that are overlapping exclusively with a group of TFs are highlighted

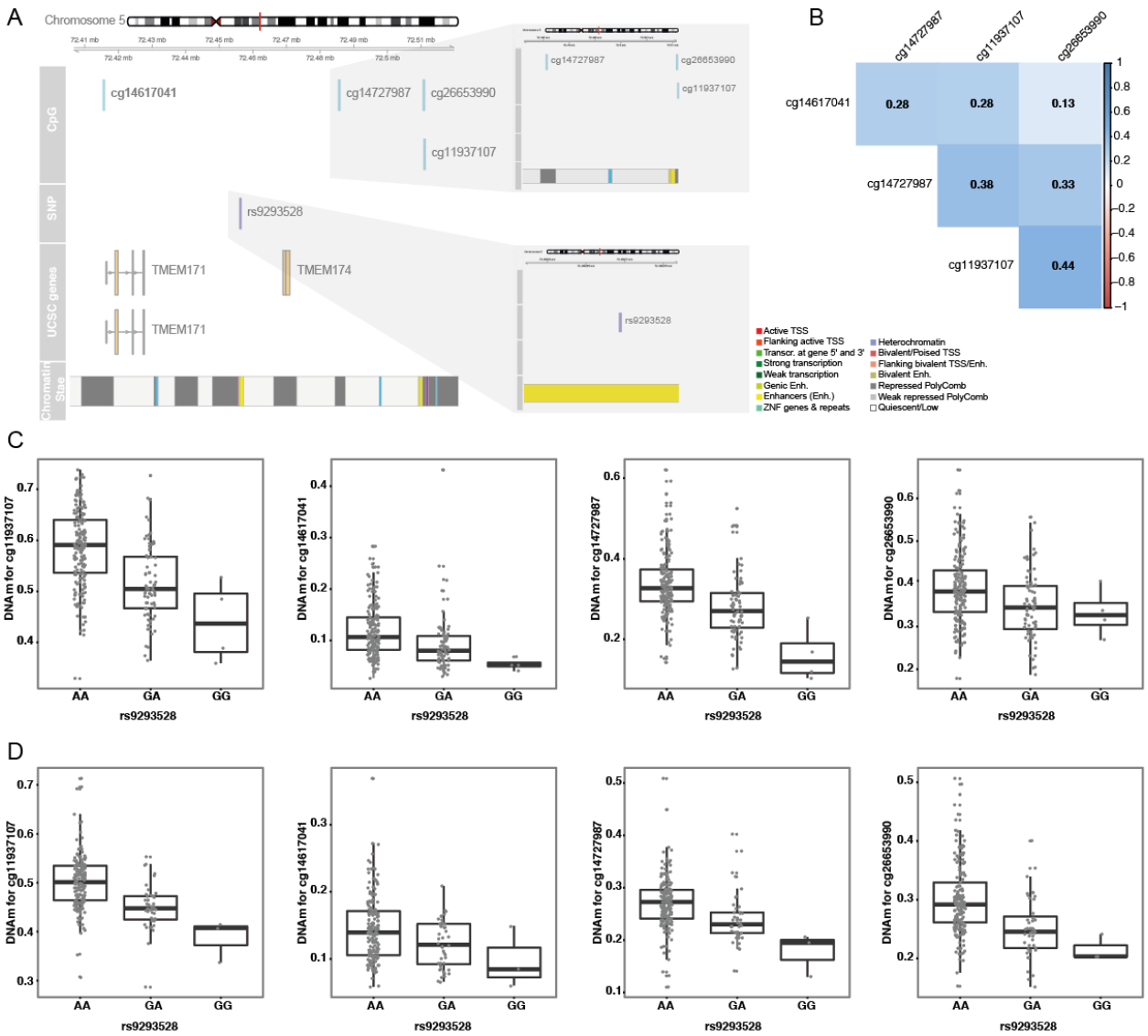
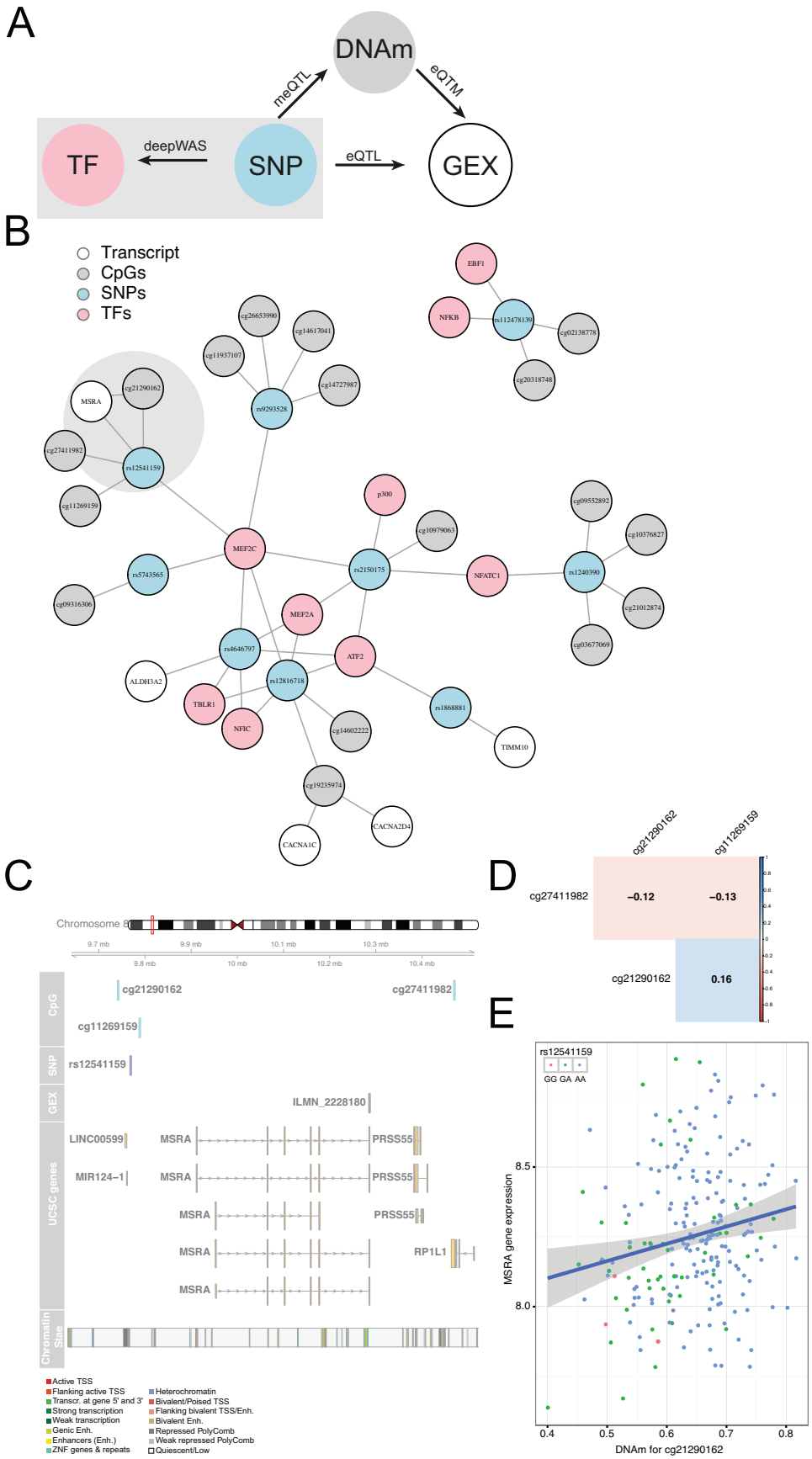


Figure 4: **meQTL results for the GM12878 deepSNP rs9293528. (A)** Overview of the rs9293528 meQTL locus on chromosome 5. Top panel, ideogram for chromosome 5. A red box indicates the region shown enlarged in the bottom panels. Bottom panels: Location of CpGs (n=4) that are significantly associated with rs9293528 genotypes, SNP position, genes nearby (based on USCS knownGene annotation) and annotation of chromatin states (based on GM12878 15-state core ChromHMM model). **(B)** Matrix of Pearson's correlation coefficients between the DNAm levels (beta values) of cg1193710, cg14617041, cg14727987 and cg26653990. **(C)** meQTLs between

501 rs9293528 and cg1193710, cg14617041, cg14727987 and cg26653990 within the
502 recMDD cohort. **(D)** Replication of those meQTLs in the MPIP cohort. Y-axis in C and D
503 displays the DNAm level of a particular CpG site and the x-axis indicates rs9293528
504 genotype.
505


A
DNAm
meQTL
eQTM
TF
deepWAS
SNP
eQTL
GEX
B
Transcript
CpGs
SNPs
TFs
EBF1
NFKB
rs11278139
cg02138778
cg20318748
cg26653990
cg14617041
cg11937107
cg14727987
rs9293528
MSRA
cg21290162
cg27411982
rs12541159
cg11269159
p300
cg09552892
cg10979063
cg10376827
MEF2C
rs2150175
NFATC1
rs1240390
rs5743565
cg09316306
MEF2A
cg21012874
cg03677069
rs4646797
ATF2
ALDH3A2
rs12816718
TBLR1
NFIC
cg14602222
rs1868881
TIMM10
cg19235974
CACNA2D4
CACNA1C
C
Chromosome 8
9.7 mb
9.8 mb
9.9 mb
10 mb
10.1 mb
10.2 mb
10.3 mb
10.4 mb
CpG
cg21290162
cg27411982
cg11269159
SNP
rs12541159
GEX
ILMN_2228180
UCSC genes
LINC00599
MSRA
PRSS55
MIR124–1
MSRA
PRSS55
MSRA
PRSS55
MSRA
RP1L1
MSRA
Chromatin State
Active TSS
Flanking active TSS
Transcr. at gene 5' and 3'
Strong transcription
Weak transcription
Genic Enh.
Enhancers (Enh.)
ZNF genes & repeats
Heterochromatin
Bivalent/Poised TSS
Flanking bivalent TSS/Enh.
Bivalent Enh.
Repressed PolyComb
Weak repressed PolyComb
Quiescent/Low
D
cg21290162
cg11269159
cg27411982
−0.12
−0.13
cg21290162
0.16
E
rs12541159
GG GA AA
8.5
8.0
MSRA gene expression
0.4
0.5
0.6
0.7
0.8
DNAm for cg21290162

506

507 **Figure 5: Joint analysis of all significant meQTL, eQTL, eQTM and deepWAS**
508 **results: (A)** Schematic illustration of our joint analysis. Grey box highlighting the MDD
509 disease effect. **(B)** Network visualization of the joint analysis. Edges represent the
510 association relation of deepSNPs, CpGs, TFs and transcripts. The grey circle highlights
511 the deepSNP rs12541159, which coincides all types of QTLs. **(C)** Regional plot for the
512 deepSNP rs12541159 locus: with top panel illustrating the ideogram for chromosome 8.
513 Bottom panels the location of CpGs (n=3) that are significantly associated with
514 rs12541159 genotypes, SNP position, gene expression probe position sig. association
515 with rs12541159, genes nearby (based on USCS knownGene annotation) and
516 annotation of chromatin states (based on GM12878 ChromHMM). **(D)** Matrix of
517 Pearson's correlation coefficients between the DNAm levels (beta values) of
518 cg27411982, cg11269159 and cg21290162. **(E)** Relationship between cg21290162
519 methylation, MSRA expression, and deepSNP rs12541159 genotype in MPIP samples,
520 including mediation effect rs12541159 → cg21290162 → MSRA expression.
521

522 # Tables

CpG	SNP	Chr	SNP Position	SNP Location	Genes near SNP	CpG Position	CpG Location	Genes near CpG	P-value	Validation in MPIP
cg02138778	rs112478139	20	25110965	intergenic	VSX1,LOC284798	25172805	intergenic	LOC284798,ENTPD6	6.72e-06	yes
cg20318748	rs112478139	20	25110965	intergenic	VSX1,LOC284798	25605228	upstream	NANP	0.00306	yes
cg03677069	rs1240390	10	88725984	intergenic	SNCG,ADIRF	88718366	UTR5	SNCG	0.0148	yes
cg09552892	rs1240390	10	88725984	intergenic	SNCG,ADIRF	88718324	UTR5	SNCG	0.0229	yes

cg10376827	rs1240390	10	88725984	intergenic	SNCG,ADIRF	88730374	exonic	ADIRF	0.0142	yes
cg21012874	rs1240390	10	88725984	intergenic	SNCG,ADIRF	88718443	UTR5	SNCG	0.000177	yes
cg11269159	rs12541159	8	9769098	intergenic	MIR124-1,MSRA	9788795	intergenic	MIR124-1,MSRA	0.00265	yes
cg21290162	rs12541159	8	9769098	intergenic	MIR124-1,MSRA	9741960	intergenic	TNKS,LINC00599	0.000191	yes
cg27411982	rs12541159	8	9769098	intergenic	MIR124-1,MSRA	10470102	exonic	RP1L1	0.0105	no
cg14602222	rs12816718	12	969800	intronic	WNK1	1025664	exonic	RAD52	0.00226	yes
cg19235974	rs12816718	12	969800	intronic	WNK1	1063197	intergenic	RAD52,ERC1	0.0386	yes
cg10979063	rs2150175	6	97269213	intronic	GPR63	97285872	upstream	GPR63	0.000352	yes
cg09316306	rs5743565	4	38805983	UTR5	TLR1	38807387	upstream	TLR1	4.9e-05	yes
cg11937107	rs9293528	5	72456267	intergenic	TMEM171,TMEM174	72511061	intergenic	TMEM174,FOXD1	9.81e-10	yes
cg14617041	rs9293528	5	72456267	intergenic	TMEM171,TMEM174	72415688	upstream	TMEM171	0.00791	yes
cg14727987	rs9293528	5	72456267	intergenic	TMEM171,TMEM174	72485684	intergenic	TMEM174,FOXD1	6.44e-07	yes
cg26653990	rs9293528	5	72456267	intergenic	TMEM171,TMEM174	72510846	intergenic	TMEM174,FOXD1	0.0104	yes

Table 1: Significant GM12878 deepSNPs representing a meQTL (GSK cohort). P-values were corrected according to Bonferroni.

Additional files

Additional file 1: Supplementary Figures S1-S6 and Tables S1-S3 (PDF)

Additional file 2: List of MDD deepWAS results. Table lists deepWAS results for MDD. Triplet of the first three columns together form a functional unit. (XLS)

- Cell Line: Name of cell line of the functional unit
- Chromatin Feature: Name of chromatin feature of the functional unit
- Treatment: Name of Treatment of the cell line of the functional unit
- SNP: The deepSNPs jointly associated to MDD using deepWAS

535 **Additional file 3:** Significance of LASSO models with number of predictors in each
536 model (XLS)

537 • Cell line: Name of cell line of the functional unit
538 • Chromatin feature: Name of chromatin feature of the functional unit
539 • Treatment: Name of Treatment of the cell line of the functional unit
540 • pval.adj: FDR-adjusted p-values which represent the significance of a regression
541 model
542 • pval: p-values which represent the significance of a regression model
543 • devratio: Deviance ratio used as a test statistic for quantify the significance of
544 regression model
545 • n.snps: Number of SNPs that are used as predictors in regression models
546 • n.nonzero.snps: Number of predictors selected by LASSO approach (i.e.
547 covariates with non-zero coefficients)

548 **Additional file 4:** Overlap of bQTLs and MDD deepSNPs. Overlap of MDD deepSNPs
549 in cell line GM12878 with public bQTL study in same cells. (XLS)

550 • Chr: Chromosome number
551 • Pos.h19: Genomic position of SNP in hg19 assembly
552 • Rsid: SNP identifier
553 • bQTL.ref: Reference allele base extracted from bQTL study
554 • bQTL.alt: Alternative allele base extracted from bQTL study
555 • bQTL.pval: p-value extracted from bQTL study
556 • Higher Binding Allele: Base of higher binding allele extracted from bQTL study
557 • bQTL.TF: Transcription factor extracted from bQTL study
558 • deepwas_TF: Transcription factor identified using MDD deepWAS and rsid
559

560

Methods

Clinical Samples

recMDD

The sample consists of 1,537 Caucasian individuals (with 84%, n=1294 of German origin) recruited at the Max-Planck Institute of Psychiatry (MPIP) in Munich, Germany and two satellite hospitals in the Munich metropolitan area (BKH Augsburg and Klinikum Ingolstadt): 739 controls (500 females, 239 males) and 798 cases diagnosed with recurrent major depression (526 females, 272 males). Please see Muglia et al.[16] for more detail in sample recruitment and characterization. All subjects are independent from the MPIP collection below.

MPIP

The group of subjects consists of 289 Caucasian individuals of the Max-Planck Institute of Psychiatry (MPIP), 93 women and 196 men. Recruitment strategies and further characterization of the MPIP cohort have been described previously[21]. One hundred sixty of them were healthy (115 men, 45 women), 129 (81 men, 48 women) had a depressive disorder treated at the MPIP's hospital in Munich.

Genotype data and Imputation (GSK & MPIP)

Human DNA of the GSK and MPIP cohort samples was isolated from EDTA blood samples using the Gentra Puregene Blood Kit (Qiagen) with standardized protocols. Genome-wide SNP genotyping was performed using Illumina HumanHap550 Quad

581 (GSK), Illumina Human610-Quad (MPIP) and OmniExpress (MPIP) genotyping
582 BeadChips according to the manufacturer's standard protocols. Quality control (QC) of
583 genotyped data was conducted in PLINK 1.90b3s [20] separately for each cohort and
584 genotyping BeadChip. QC steps on samples included removal of individuals with
585 genotyping rate < 2 %, cryptic relatives (PI-HAT > 0.05), and genetic outliers (distance
586 in first two MDS components from mean > 2 SD). QC steps on variants included
587 removal of variants with call rate < 2 %, MAF < 5% and HWE test p-value ≤ 10-6.
588 Furthermore, variants on non-autosomal chromosomes were excluded, resulting in
589 GSK: 481,178, MPIP: 481,762 (Human610-Quad) and 544,908 (OmniExpress) SNPs.
590 These sets of SNPs comprised the input for imputation, which was performed
591 separately for each cohort and genotyping BeadChip using IMPUTE v2[24] with the
592 following parameters: 1000G phase I reference panel (released in June 2014, ALL
593 samples), SHAPEIT [23] phasing. QC of imputed probabilities was conducted in
594 QCTOOL 1.4 (http://www.well.ox.ac.uk/~gav/qctool/). Imputed SNPs were excluded if
595 the HWE test p-value ≤ 10^{-6} and the info metric < 0.8. Called genotypes ("best guess
596 genotypes'" with a probability of 70%) were used for further analysis. SNP sets of the
597 MPIP cohort were merged together. GSK and merged MPIP SNPs were further
598 processed in PLINK and variants were excluded if their MAF < 5% and best guess
599 genotypes are not called (call rate = 100%). This yielded a total of 1.78 Mio GSK SNPs
600 and 3 Mio MPIP SNPs. To annotate SNPs for the closest genes, we used Annovar
601 version Februar 2016 [41] with the UCSC knownGene annotations. SNP coordinates are
602 given according to hg19.

603 ## Prediction of regulatory effects

604 In this study, we employed DeepSEA[12] to determine the SNPs that might play an
605 important role in MDD by acting through the alteration of regulatory elements.
606 Pretrained DeepSEA network (v0.94) was downloaded from
607 http://deepsea.princeton.edu/help/ and the predictions of 1.7M SNPs were generated
608 using NVIDIA GeForce GTX TITAN X GPU (Maxwell) in 10 hours. Next step, filtering of
609 regulatory SNPs, was performed using generated so-called "e-value" files that represent
610 the significance of the regulatory effect of SNPs. We applied an e-value cutoff of
611 $5 \cdot 10^{-5}$ where we only take the SNPs that are associated to at least one functional unit
612 into consideration e.g. (rs1035271, GM12878, MEF2C). Set of SNPs that have an
613 impact on functional unit k is depicted here as set S_k.
614
615 We then employed a simple probabilistic genotype encoding where allele-specific
616 regulatory probability of the reference or alternative alleles is used if genotype is
617 reference or alternative homozygous, respectively. If the genotype of the individual is
618 heterozygous then the mean of two allele probabilities are used. Here, we refer to the
619 genotype matrix (1,537 x 1.7M) as G where the rows are individuals and columns are
620 SNPs. Genotypes in the G matrix is encoded such that $G_{ij} \in \{0, 1, 2\}$ where we simply
621 count an arbitrarily chosen allele. We then define D as a tensor representing the allele-
622 specific probabilities for each SNP, functional unit and allele. For example $D_{jk}^{(ref)}$ refers
623 to the probability of reference allele of SNP_j in the context of functional unit k. Therefore,
624 the SNP encoding scheme can be described in terms of G and D matrices:
625

$$X_{ijk} = \begin{cases} D_{jk}^{(ref)} \ if \ G_{ij} = 0 \\ D_{jk}^{(alt)} if \ G_{ij} = 2 \\ \frac{D_{jk}^{(ref)} + D_{jk}^{(alt)}}{2} \ if \ G_{ij} = 1 \end{cases} \quad G_{ij} \in \{0, 1, 2\}$$

or equivalently

$$X_{ijk} = G_{ij} \left(\frac{D_{jk}^{(alt)} - D_{jk}^{(ref)}}{2} \right) + D_{jk}^{(ref)}$$

Here the subscripts i, j and k represent individual, SNP, and functional unit indices, respectively. Resulting encoded matrix is 1,537 x 6,143,515 where we store all SNP-functional unit pairs (919 x 6,685) for all individuals using “bigmemory” R package[32] in order to limit the memory consumption. During model fitting, however, only the list of SNPs in sets S_k are used.

DeepWAS

Penalized regression models

Compared to the classical GWAS approaches where the trait of interest is regressed separately on each SNP, regularized regression approaches provide an alternative way to handle high dimensional data and to identify SNPs associated with the trait of interest using variable selection. Here we utilize L1-regularized logistic regression (LASSO) for variable selection. 919 different LASSO models are fitted for each functional unit in

order to estimate the statistical association between the disease status (y=0 for controls, and 1 for cases) and SNPs in the context of a specific cell line:

$$logit\ P(y_i = 1) = \beta_{0k} + \sum_{j \in S_k} \beta_{jk} X_{ijk} + \beta_{sex,k} sex_i + \beta_{age,k} age_i$$

where k is used as a model index. Note also that the regression coefficients are model specific and therefore indexed with k. In each LASSO model, only the SNPs that are significantly affecting the binding of a specific TF in a specific cell line are included as covariates. This is represented in the equation by the summation over the elements of S_k.

Note that the simple probabilistic encoding does not affect the variable selection, because the effect goes away during the standardization of the design matrix except for possible sign flips. Yet, it leads to the effect sizes that are scaled based on the allele-specific regulatory probabilities reported by DeepSEA since the effect sizes in logistic regression are reported in the scale of original covariates.

The parameters (β) of the logistic regression are optimized with L1 regularization:

$$min_\beta \sum_{i=1}^{M} \left[y_i - \beta_{0k} - \sum_{j \in S_k} \beta_{jk} X_{ijk} + \beta_{sex,k} sex_i + \beta_{age,k} age_i \right]^2 + \lambda \|\beta\|_1$$

where M is the number of individuals, λ parameter represents the strength of the regularization and β vector represents model parameters. We fitted L1-regularized logistic regression model (LASSO) using glmnet R package[42]. Regularization parameter λ is determined by 100-fold cross validation after which we used "lambda.1se" value to determine non-zero parameters.

Permutation-based significance test

We used a permutation-based approach to select significant LASSO models with non-zero coefficients and to control false discovery rate. We generated 1000 random permutations of the response variable and fitted LASSO models on each. λ parameter in each model is determined by 10-fold cross validation in order to reduce the computational cost. After fitting LASSO models on the permuted data, deviance ratio is used as a test statistic.

$$deviance\ ratio\ = 1 - \frac{D}{D_{null}} \text{ where } D = 2\ (L_S\ -\ L_P) \text{ and } D_{null} = 2\ (L_S\ -\ L_N)$$

Here L_S, L_P and L_N are the log likelihoods of saturated model (model with a free parameter per individual), proposed model and null model (intercept model), respectively. Empirical p-values are calculated as the number of permuted models with a deviance ratio greater than or equal to the proposed model divided by 1000. After FDR[34] multiple testing correction with threshold of 10%, models with lower adjusted p-values are retained.

685 ## Functional annotation of deepSNPs

686 ### Enrichment of Roadmap cis-regulatory elements

687 In order to quantify the enrichment of deepSNPs in *cis*-regulatory elements identified by
688 Roadmap Epigenomics Project[14], we first downloaded the segmentation files of core
689 15-state model for 127 epigenomes were downloaded from Roadmap epigenomics web
690 portal
691 (http://egg2.wustl.edu/roadmap/data/byFileType/chromhmmSegmentations/ChmmModel
692 s/coreMarks/jointModel/final/all.mnemonics.bedFiles.tgz) in BED format. Then we
693 sampled random genomic ranges with the same size distribution as promoter (State 1)
694 and enhancer (State 7) regions from the background distribution using GAT[44]. The
695 background distribution was defined as the intersection of mappable regions of the
696 genome
697 (https://www.cgat.org/downloads/public/gat/datasets/hg19/mapability_36.filtered.bed.gz)
698 and genomic segments that belong to the first 13 chromatin states of ChromHMM
699 model (State 14: Weak repressed Polycomb and State 15: Quiescent are excluded). In
700 order to limit the computation time, 1000 samples were drawn and p-values smaller
701 than 10^{-3} were extrapolated from normal distribution using "--pvalue-method=norm"
702 command line option.

703 ### bQTL overlap

704 bQTL results were downloaded from "Supplementary Table 2" of Tehranchi et al.[4] and
705 intersection with GM12878 deepSNPs is computed in R version 3.3.0.

706 Transcription factor co-localization

707 ReMap project[33] provides a consolidated dataset for ChIP-seq peaks of several
708 transcription factors including the ENCODE peaks as well as various public datasets. In
709 order to quantify TF-TF colocalization patterns, we downloaded non-redundant peaks
710 for 11 individual TFs in BED format from the web interface (http://tagc.univ-
711 mrs.fr/remap/index.php?page=download) and then ran the annotation interface
712 (http://tagc.univ-mrs.fr/remap/index.php?page=annotation) for each BED file separately.
713 Downloaded TSV files with –log10 transformed e-values are combined and plotted with
714 corrplot R package[43]. Because very high –log10-transformed e-values obscure
715 visualization, e-values are binarized with a cutoff of 10^{-3}. Out of 12 TFs affected by 26
716 GM12878 deepSNPs 2 TFs are not shown in **Additional file 1: Figure S6** since p300
717 peaks are not available in ReMap dataset and CTCF did not show any significant
718 association with other TFs.

719 DNA methylation (GSK & MPIP)

720 For a subset of GSK (n=257 with 53% MDD cases) and MPIP cohort (n=229 with 52%
721 MDD cases) genomic DNA was extracted from whole blood using the Gentra Puregene
722 Blood Kit (QIAGEN). DNA quality and quantity of both was assessed by NanoDrop 2000
723 Spectrophotometer (Thermo Scientific) and Quant-iT Picogreen (Invitrogen). Genomic
724 DNA was bisulfite converted using the Zymo EZ-96 DNA Methylation Kit (Zymo
725 Research) and DNA methylation levels were assessed for >480,000 CpG sites using
726 the Illumina HumanMethylation450 BeadChip array. Hybridization and processing was
727 performed according to manufacturer's instructions. Quality control of methylation data,

728 including intensity read outs, filtering (detection p-value > 0.01 in at least 75% of the
729 samples), cellular composition estimation according to[44], as well as beta calculation
730 was done using the minfi Bioconductor R package version 1.18.2. Filtered beta values
731 were reduced by eliminating any CpG sites on sex chromosomes, as well as probes
732 found to have SNPs at the CpG site itself or in the single-base extension site with a
733 MAF ≥ 1% in the 1,000 Genomes Project European population and/or non-specific
734 binding probes according to[45]. Additionally, we performed a re-alignment of the array
735 probe sequences using Bismark[45,46]. This yielded a total of around 425,000 GSK and
736 MPIP CpG sites for further analysis. The data were then normalized with functional
737 normalization[47]. Technical batch effects were identified by inspecting the association
738 of the first principal components of the methylation levels with plate and plate position
739 for the GSK sample as well as processing (experiment) date for the MPIP sample as
740 technical batch. The data were then adjusted using ComBat[48]. To annotate CpGs for
741 the closest genes, we used Annovar version February 2016[41] with the UCSCknown
742 gene annotation. CpG coordinates are given according to hg19.

743 Gene expression (MPIP)

744 Whole blood RNA of the MPIP sample (n=289 individuals) was collected using
745 PAXgene Blood RNA Tubes (PreAnalytiX) and processed as described previously[21].
746 The RNA was then hybridized to Illumina HT-12 v3 and v4 expression Bead Chips
747 (Illumina, San Diego, CA). Raw probe intensities were exported using Illumina's
748 GenomeStudio and further statistical processing was carried out using R version 3.3.0.
749 All 29,075 probes present on both BeadChips (v3 vs. v4), excluding X and Y
750 chromosomes as well as cross-hybridizing probes identified by using the Re-Annotator

751 pipeline [21,49] were first filtered with a detection p-value of 0.05 in at least 50% of the
752 samples, leaving 11,994 autosomal expression array probes. Subsequently, each probe
753 was transformed and normalized through variance stabilization and normalization
754 (VSN)[50]. Technical batch effects were identified by inspecting the association of the
755 first principal components of the expression levels for all known batch effects and then
756 adjusted using ComBat with slide, amplification round, array version, and amplification
757 plate column as fixed effects. The position of the gene expression probe and gene
758 symbols were annotated using the Re-Annotator pipeline based on hg19. Blood cell
759 counts were estimated according to CellCODE [51].

760 Statistical analysis of gene expression and methylation data

761 Methylation QTL (meQTL) analysis

762 For each of the 26 deepSNP CpG sites within a 3Mb window around the SNP position
763 were selected for the meQTL analysis. Due to low frequency of minor allele
764 homozygotes for many SNPs, all SNP genotypes were coded and evaluated as a
765 dominant model. Linear regression was used to measure the relationship between the
766 DNA methylation (beta values) and the number of minor alleles (coded 0 and 1),
767 including covariates for age, disease-state, gender and blood cell counts. So we
768 defined the model for a single meQTL mapping as follows: $CpG = \beta_0 + \beta_{sex} sex +$
769 $\beta_{age} age + \beta_{CD8T} CD8T + \beta_{CD4T} CD4T + \beta_{NKcell} NKcell + \beta_{Bcell} Bcell + \beta_{Mono} Mono +$
770 $\beta_{Gran} Gran + \beta_{disesase\ status} disease\ status + \beta_{SNP} SNP + e$, where e is the general
771 error term for any residual variation not explained by the rest of the model.
772

773 In addition a differential meQTL analysis by the disease status was performed.
774 Therefore an additive model adjusting for age, gender and blood cell counts was
775 applied in controls and cases separately:

776 $CpG = \beta_0 + \beta_{sex} sex + \beta_{age} age + \beta_{CD8T} CD8T + \beta_{CD4T} CD4T + \beta_{NKcell} NKcell +$
777 $\beta_{Bcell} Bcell + \beta_{Mono} Mono + \beta_{SNP} SNP + e.$

778 P-values were adjusted for the total number of CpG sites that were within the tested
779 region surrounding the SNP. Case specific meQTLs were defined as meQTLs, which
780 show a significant P-value in cases but not in controls (nominal p-value> 0.05). GSK
781 and MPIP data were analysed independently and a validation of meQTL results was
782 carried out with a sample size-weighted Z-score meta-analysis [52]. This method
783 accounts for different sizes and suggests the robustness the meQTLs.
784 Top hits were plotted with easyGgplot2 (https://github.com/kassambara/easyGgplot2),
785 corrplot R package [43] and Gviz 1.16.1 bioconductor package [53].
786

787 Expression QTL (eQTL) analysis

788 For each of deepSNP all transcripts beginning or ending within 1Mb up- or downstream
789 of the SNP were determined. Associations between genotype (coded 0, 1 and 2) and
790 expression levels (normalized and batch corrected log2 gene expression values) were
791 determined in the MPIP data set by linear regression, using sex, age, disease status
792 and blood cell estimates as covariates:

793 $GEX =$
794 $\beta_0 + \beta_{sex} sex + \beta_{age} age + \beta_{Neutrophil} Neutrophil + \beta_{Mono} Mono + \beta_{Bcell} Bcell +$

795 $\beta_{Tcell}Tcell + \beta_{NKcell}NKcell + \beta_{Plasmacell}Plasmacell + \beta_{DendriticCell}DendriticCell +$

796 $\beta_{diseas\ status}disease\ status + \beta_{SNP}SNP + e.$

797 To correct for multiple testing, p-values were corrected for the number of transcripts per

798 SNP.

799 Expression quantitative trait methylation (eQTM) analysis

800 For each of CpG site of the significant DeepQTL SNP- CpG pairs all transcripts

801 beginning or ending within 1Mb up- or downstream of the CpG were determined.

802 Associations between CpG and expression levels were determined in the MPIP data set

803 by linear regression, using sex, age, disease status and blood cell estimates as

804 covariates:

805 $CpG = \beta_0 + \beta_{sex}sex + \beta_{age}age + \beta_{CD8T}CD8T + \beta_{CD4T}CD4T + \beta_{NKcell}NKcell +$

806 $\beta_{Bcell}Bcell + \beta_{Mono}Mono + \beta_{Gran}Gran + \beta_{disesase\ status}disease\ status + \beta_{GEX}GEX +$

807 $e.$

808 To correct for multiple testing, p-values were first corrected for the number of transcripts

809 per SNP window.

810 Joint analysis

811 For the joint analysis we first measured the overlap of deepSNPs with eQTL, meQTL

812 data and secondly for all overlapping pairs we calculated the overlap of meQTL CpGs

813 and eQTM CpGs. For the deepSNPs, CpG, transcripts triplets we then assessed

814 possible causal relationships by using a causal inference test [32].

815 The network was plotted using the R package igraph [54]. The triplet was plotted using

816 ggplot2[55], corrplot R package [43] and Gviz 1.16.1 Bioconductor package [53].

Declarations

Availability of data and material

Data from MPIP gene expression experiment are deposited at the GEO repository under GEO: GSE64930 and methylation under: GSE74414. DeepWAS source code is available at DOI: 10.5281/zenodo.59282 through Zenodo.

Competing interests

The authors declare that they have no competing interests.

Funding

This work was funded by the German Federal Ministry of Education and Research through the Research Consortium Integrated Network IntegraMent (grant 01ZX1314H, to GE), the LiSyM Verbundprojekt Pillar II/III (grant 031L0047, to NSM) under the auspices of the e:Med Programme as well as the European Research Council (grant 281338 to EBB)

Ethics approval and consent to participate

Both studies (recMDD and MPIP) were approved by the local ethics committee and all individuals gave written informed consent. All experimental methods comply with the Helsinki Declaration.

Authors' contributions

NSM conceived and designed the study. GE and JA performed the research, analysed the data, developed the computational methods and created figures. GE, JA, EBB, FJT and NSM wrote the paper. JM and DC performed imputation of recMDD and MPIP genotypes. SI performed the pre-processing of the recMDD methylation data. EBB and FJT contributed to the study design.

Acknowledgements

We thank Michael Laimighofer for feedback on the statistical approach. We thank Matthias Heinig for valuable comments on the manuscript. We would like to thank Martin Preusse for useful discussions on the approach.

References

1. Visscher PM, Brown MA, McCarthy MI, Yang J. Five Years of GWAS Discovery. Am. J. Hum. Genet. 2012;90:7–24.

2. Gamazon ER, Wheeler HE, Shah KP, Mozaffari SV, Aquino-Michaels K, Carroll RJ, et al. A gene-based association method for mapping traits using reference transcriptome data. Nat. Genet. Nature Research; 2015;47:1091–.

3. Gamazon ER, Badner JA, Cheng L, Zhang C, Zhang D, Cox NJ, et al. Enrichment of cis-regulatory gene expression SNPs and methylation quantitative trait loci among bipolar disorder susceptibility variants. Mol. Psychiatry. Nature Publishing Group; 2013;18:340–6.

4. Tehranchi AK, Myrthil M, Martin T, Hie BL, Golan D, Fraser HB. Pooled ChIP-Seq Links Variation in Transcription Factor Binding to Complex Disease Risk. Cell. 2016;165:730–41.

865 5. Edwards SL, Beesley J, French JD, Dunning AM. Beyond GWASs: Illuminating the
866 Dark Road from Association to Function. The American Journal of Human Genetics.
867 2013;93:779–97.

868 6. Tak YG, Farnham PJ. Making sense of GWAS: using epigenomics and genome
869 engineering to understand the functional relevance of SNPs in non-coding regions of
870 the human genome. Epigenetics & Chromatin 2015 8:1. BioMed Central; 2015;8:1.

871 7. Maurano MT, Humbert R, Rynes E, Thurman RE, Haugen E, Wang H, et al.
872 Systematic Localization of Common Disease-Associated Variation in Regulatory DNA.
873 Science. American Association for the Advancement of Science; 2012;337:1190–5.

874 8. Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, et al.
875 Potential etiologic and functional implications of genome-wide association loci for
876 human diseases and traits. Proc. Natl. Acad. Sci. U.S.A. National Acad Sciences;
877 2009;106:9362–7.

878 9. Robertson G, Hirst M, Bainbridge M, Bilenky M, Zhao Y, Zeng T, et al. Genome-wide
879 profiles of STAT1 DNA association using chromatin immunoprecipitation and massively
880 parallel sequencing. Nature Methods. Nature Publishing Group; 2007;4:651–7.

881 10. Coetzee SG, Coetzee GA, Hazelett DJ. motifbreakR: an R/Bioconductor package
882 for predicting variant effects at transcription factor binding sites. Bioinformatics. Oxford
883 University Press; 2015;31:btv470–3849.

884 11. Thomas-Chollier M, Hufton A, Heinig M, O'Keeffe S, Masri NE, Roider HG, et al.
885 Transcription factor binding predictions using TRAP for the analysis of ChIP-seq data
886 and regulatory SNPs. Nature Protocols. Nature Research; 2011;6:1860–9.

887 12. Zhou J, Troyanskaya OG. Predicting effects of noncoding variants with deep
888 learning-based sequence model. Nature Methods. Nature Research; 2015;12:931–4.

889 13. Consortium TEP. An integrated encyclopedia of DNA elements in the human
890 genome. Nature. Nature Research; 2012;489:57–74.

891 14. Consortium RE, Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, et al.
892 Integrative analysis of 111 reference human epigenomes. Nature. Nature Research;
893 2015;518:317–30.

894 15. Tibshirani R. Regression shrinkage and selection via the Lasso. Journal of the
895 Royal Statistical Society Series B-Methodological. 1996;58:267–88.

896 16. Muglia P, Tozzi F, Galwey NW, Francks C, Upmanyu R, Kong XQ, et al. Genome-
897 wide association study of recurrent major depressive disorder in two European case-
898 control cohorts. Mol. Psychiatry. Nature Publishing Group; 2010;15:589–601.

899 17. Kendler KS, Gatz M, Gardner CO, Pedersen NL. A Swedish national twin study of
900 lifetime major depression. Am J Psychiatry. American Psychiatric Publishing;

901 2006;163:109–14.

902 18. Levinson DF, Mostafavi S, Milaneschi Y, Rivera M, Ripke S, Wray NR, et al. Genetic
903 Studies of Major Depressive Disorder: Why Are There No Genome-wide Association
904 Study Findings and What Can We Do About It? Biol. Psychiatry. 2014;76:510–2.

905 19. Hyde CL, Nagle MW, Tian C, Chen X, Paciga SA, Wendland JR, et al. Identification
906 of 15 genetic loci associated with risk of major depression in individuals of European
907 descent. Nat. Genet. Nature Research; 2016.

908 20. Consortium SWGOTPG, Ripke S, Neale BM, Corvin A, Walters JT, Farh K-H, et al.
909 Biological Insights From 108 Schizophrenia-Associated Genetic Loci. Nature. Europe
910 PMC Funders; 2014;511:421–7.

911 21. Arloth J, Bogdan R, Weber P, Frishman G, Menke A, Wagner KV, et al. Genetic
912 Differences in the Immediate Transcriptome Response to Stress Predict Risk-Related
913 Brain Function and Psychiatric Disorders. Neuron. 2015;86:1189–202.

914 22. Consortium T1GP. A map of human genome variation from population-scale
915 sequencing. Nature. Nature Research; 2011;473:544–4.

916 23. Delaneau O, Zagury J-F, Marchini J. Improved whole-chromosome phasing for
917 disease and population genetic studies. Nature Methods. Nature Research; 2013;10:5–
918 6.

919 24. Howie BN, Donnelly P, Marchini J. A Flexible and Accurate Genotype Imputation
920 Method for the Next Generation of Genome-Wide Association Studies. Schork NJ,
921 editor. PLOS Genet. Public Library of Science; 2009;5:e1000529.

922 25. Wohleb ES, Franklin T, Iwata M, Duman RS. Integrating neuroimmune systems in
923 the neurobiology of depression. Nat. Rev. Neurosci. Nature Research; 2016;17:497–
924 511.

925 26. Mostafavi S, Battle A, Zhu X, Potash JB, Weissman MM, Shi J, et al. Type I
926 interferon signaling genes in recurrent major depression: increased expression detected
927 by whole-blood RNA sequencing. Mol. Psychiatry. Nature Publishing Group;
928 2014;19:1267–74.

929 27. Hsu F, Kent WJ, Clawson H, Kuhn RM, Diekhans M, Haussler D. The UCSC Known
930 Genes. Bioinformatics. Oxford University Press; 2006;22:1036–46.

931 28. Bhat S, Dao DT, Terrillion CE, Arad M, Smith RJ, Soldatov NM, et al. CACNA1C
932 (Cav1.2) in the pathophysiology of psychiatric disease. Prog. Neurobiol. 2012;99:1–14.

933 29. Van Den Bossche MJ, Strazisar M, De Bruyne S, Bervoets C, Lenaerts A-S, De
934 Zutter S, et al. Identification of a CACNA2D4 deletion in late onset bipolar disorder
935 patients and implications for the involvement of voltage-dependent calcium channels in
936 psychiatric disorders. Am. J. Med. Genet. B Neuropsychiatr. Genet. Wiley Subscription

937 Services, Inc., A Wiley Company; 2012;159B:465–75.

938 30. Ni P, Ma X, Lin Y, Lao G, Hao X, Guan L, et al. Methionine sulfoxide reductase A
939 (MsrA) associated with bipolar I disorder and executive functions in A Han Chinese
940 population. J Affect Disord. 2015;184:235–8.

941 31. Walss-Bass C, Soto-Bernardini MC, Johnson-Pais T, Leach RJ, Ontiveros A,
942 Nicolini H, et al. Methionine sulfoxide reductase: a novel schizophrenia candidate gene.
943 Am. J. Med. Genet. B Neuropsychiatr. Genet. Wiley Subscription Services, Inc., A Wiley
944 Company; 2009;150B:219–25.

945 32. Millstein J, Zhang B, Zhu J, Schadt EE. Disentangling molecular relationships with a
946 causal inference test. BMC Genet. BioMed Central; 2009;10.

947 33. Griffon A, Barbier Q, Dalino J, van Helden J, Spicuglia S, Ballester B. Integrative
948 analysis of public ChIP-seq experiments reveals a complex multi-cell regulatory
949 landscape. Nucl. Acids Res. Oxford University Press; 2015;43:e27–7.

950 34. Lee D, Karchin R, Beer MA. Discriminative prediction of mammalian enhancers from
951 DNA sequence. Genome Res. Cold Spring Harbor Lab; 2011;21:2167–80.

952 35. Waldmann P, Mészáros G, Gredler B, Fuerst C, Sölkner J. Evaluation of the lasso
953 and the elastic net in genome-wide association studies. Frontiers in Genetics. Frontiers;
954 2013;4.

955 36. Ayers KL, Cordell HJ. SNP Selection in genome-wide and candidate gene studies
956 via penalized logistic regression. Genetic Epidemiology. Wiley Subscription Services,
957 Inc., A Wiley Company; 2010;34:879–91.

958 37. Dehman A, Ambroise C, Neuvial P. Performance of a blockwise approach in
959 variable selection using linkage disequilibrium information. BMC Bioinformatics. BioMed
960 Central; 2015;16.

961 38. Chen SX, Cherry A, Tari PK, Podgorski K, Kwong YKK, Haas K. The Transcription
962 Factor MEF2 Directs Developmental Visually Driven Functional and Structural
963 Metaplasticity. Cell. 2012;151:41–55.

964 39. Barbosa AC, Kim M-S, Ertunc M, Adachi M, Nelson ED, McAnally J, et al. MEF2C, a
965 transcription factor that facilitates learning and memory by negative regulation of
966 synapse numbers and function. Proc. Natl. Acad. Sci. U.S.A. National Acad Sciences;
967 2008;105:9391–6.

968 40. Pfeiffer BE, Zang T, Wilkerson JR, Taniguchi M, Maksimova MA, Smith LN, et al.
969 Fragile X Mental Retardation Protein Is Required for Synapse Elimination by the
970 Activity-Dependent Transcription Factor MEF2. Neuron. 2010;66:191–7.

971 41. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants
972 from high-throughput sequencing data. Nucl. Acids Res. Oxford University Press;

973 2010;38:–e164.

974 42. Friedman J, Hastie T, Tibshirani R. glmnet: Lasso and elastic-net regularized
975 generalized linear models. R package version; 2009.

976 43. Wie T. Corrplot: Visualization of a correlation matrix-R package version 0.71.
977 Retrieved August; 2013.

978 44. Houseman E, Accomando WP, Koestler DC, Christensen BC, Marsit CJ, Nelson
979 HH, et al. DNA methylation arrays as surrogate measures of cell mixture distribution.
980 BMC Bioinformatics. BioMed Central; 2012;13:86.

981 45. Chen Y-A, Lemire M, Choufani S, Butcher DT, Grafodatskaya D, Zanke BW, et al.
982 Discovery of cross-reactive probes and polymorphic CpGs in the Illumina Infinium
983 HumanMethylation450 microarray. Epigenetics. Taylor & Francis; 2014;8:203–9.

984 46. Krueger F, Andrews SR. Bismark: a flexible aligner and methylation caller for
985 Bisulfite-Seq applications. Bioinformatics. Oxford University Press; 2011;27:1571–2.

986 47. Fortin J-P, Labbe A, Lemire M, Zanke BW, Hudson TJ, Fertig EJ, et al. Functional
987 normalization of 450k methylation array data improves replication in large cancer
988 studies. Genome Biol. BioMed Central; 2014;15:1.

989 48. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression
990 data using empirical Bayes methods. Biostatistics. Oxford University Press;
991 2006;8:118–27.

992 49. Arloth J, Bader DM, Röh S, Altmann A. Re-Annotator: Annotation Pipeline for
993 Microarray Probe Sequences. Xue Y, editor. PLOS ONE. Public Library of Science;
994 2015;10:e0139516.

995 50. Lin SM, Du P, Huber W, Kibbe WA. Model-based variance-stabilizing transformation
996 for Illumina microarray data. Nucl. Acids Res. Oxford University Press; 2007;36:e11–1.

997 51. Chikina M, Zaslavsky E, Sealfon SC. CellCODE: A robust latent variable approach
998 to differential expression analysis for heterogeneous cell populations. Bioinformatics.
999 Oxford University Press; 2015;31:1584–91.

1000 52. Evangelou E, Ioannidis JPA. Meta-analysis methods for genome-wide association
1001 studies and beyond. Nature Reviews Genetics. Nature Research; 2013;14:379–89.

1002 53. Hahne F, Ivanek R. Visualizing Genomic Data Using Gviz and Bioconductor.
1003 Methods Mol. Biol. New York, NY: Springer New York; 2016;1418:335–51.

1004 54. Csardi G, Nepusz T. The igraph software package for complex network research.
1005 InterJournal. 2006.

1006 55. Wickham H. ggplot2. Cham: Springer; 2016.

1007