Testing the Regulatory Consequences of 1,049 Schizophrenia Associated Variants With a Massively Parallel Reporter Assay

Leslie Myint[1], Ruihua Wang[2], Leandros Boukas[2], Kasper D. Hansen[1,2], Loyal A. Goff[2,3] Dimitrios Avramopoulos[2,4,*]

1 Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health

2 McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins School of Medicine

3 Department of Neuroscience, Johns Hopkins School of Medicine

4 Department of Psychiatry, Johns Hopkins School of Medicine

*Corresponding author

Mailing Address: Johns Hopkins University, 733 North Broadway, MRB Room 507, Baltimore MD 21205.

adimitr1@jhmi.edu

Short tittle: A MPRA for 1049 schizophrenia associated SNPs

Abstract

Recent genome-wide association studies (GWAS) have identified numerous schizophrenia (SZ) associated loci. As is common across disorders, many of the SZ associated variants are located outside protein-coding regions and are hypothesized to differentially affect the transcription of nearby genes. To systematically identify which variants affect potential regulatory activity, we assayed 1,049 variants in high Linkage Disequilibrium (LD) within 64 reported SZ-associated loci and an additional 30 SNPs from 9 loci associated with Alzheimer's disease (AD) using a massively parallel reporter assay (MPRA). Each variant was assayed on the center of a 95bp synthetic oligonucleotide. The resulting library was transfected 3 independent times into K562 chronic myelogenous leukemia lymphoblasts and another 6 times into SK-SY5Y human neuroblastoma cells. We identified 148 SNPs with significant allelic differences in their effect on expression of the reporter gene in the K562 cells and 53 in the SK-SY5Y cells with an average of 2.6 such SNPs per locus and a median of 1. The overlap between cell lines was modest with 9 SNPs having significant allelic differences in both lines, 8 of these 9 in the same direction. We do not observe a direction preference (increased or decreased enhancer activity) for risk vs. non risk alleles. We find that large LD blocks have a greater density of functional SNPs supporting instances of combinatorial SNP effects that may lead to selection at the haplotype level. Our results help determine driver GWAS variant(s), guide the functional follow up of disease associated loci and enhance our understanding of the genomic dynamics of gene regulation.

Author summary

To characterize the functional significance of disease associated variation we performed a massively parallel reporter assay targeting 1,079 disease associated DNA variants, 1,049 with Schizophrenia and 30 with Alzheimer's disease. This assay tests in an artificial system whether DNA variants have the potential to influence nearby gene expression. We found that 18% of these disease variants did, suggesting they might also do so in the brain. Many of the variants overlapped with functional locations identified by different experiments supporting the validity of the results. Further, in genomic regions where many variants are inherited together in the population, we identified relatively more functional variants which supports prior observations that combinatorial effects of multiple nearby variants might be common, possibly maintained by selection. These results will be a significant spring board for future research on the genetics of these disorders and the complex role of gene regulation in disease risk.

Introduction

Over the last decade collaborative genome wide association studies (GWAS) have identified thousands of DNA variants that show robust associations with an array of complex phenotypes [1]. For example, over 100 independent loci have been identified contributing to the risk for schizophrenia (SZ) [2,3] and based on recent reports from the SZ working group of the psychiatric genomics consortium (PGC, World Congress of Psychiatric Genetics 2017, Orlando, FL) over twice that number of loci will soon be published. In Alzheimer's disease (AD) a large meta-analysis in 2013 [4] reported 19 risk loci, of which 11 were novel.

Identifying reliable associations of genetic variants with disease has been a significant first step towards understanding the causes of neuropsychiatric disorders such as SZ and AD. Understanding their functional consequences and linking them to specific affected genes and biological processes is the next step to unlock their translational potential. Across disorders GWAS have taught us some important lessons that help with this goal. For one, variants are most often located in non-coding sequences, and 40% of the time their haplotype blocks do not include coding exons [5–7]. Further, they concentrate in regions of regulatory DNA marked by deoxyribonuclease I (DNase I) hypersensitive sites (DHSs) [8], and Quantitative Trait Loci (eQTL) studies suggest that they are often regulatory [5,9,10]. It must be noted that such studies likely underestimate how often disease variants are regulatory since variation in the time, place or circumstances under which a regulatory sequence is activated can lead to false negatives both for regulatory DNA marks and eQTLs.

Typically each GWAS locus contains multiple variants in high linkage disequilibrium (LD), all of which show strong evidence of association. This restricts efforts to identify the one or few variants that drive the association signal at a given locus and to evaluate their functional consequences. For each locus that is associated with disease risk, there must be at least one strongly correlated functional variant explaining the observed association, while it is possible that in some loci there are more than one functional variants in LD with each other along with other "passenger" variants. Such loci might contain specific haplotypes under selective pressure because of their distinct functional profiles, as recently reported for Mendelian disease mutations [11]. Episomal reporter assays have been commonly used to assess the potential of DNA sequences to drive transcription and to identify functional variants among the many in LD. In these assays, a plasmid is constructed carrying the candidate enhancer sequence inserted next to a minimal promoter driving the expression of a reporter gene. The plasmid is transfected into cells without genome integration and the expression of the reporter gene is measured and compared across different sequences. More recently, to meet the needs of genome level analysis, these types of assays have been redesigned to achieve higher throughput. Examples of such high throughput assays include massively parallel reporter assays (MPRA) [12] that examine thousands of candidate sequences in parallel and self-transcribing active regulatory region sequencing (STARR-seq) which screen entire genomes for enhancers based on their activity [13].

Because of our interest in SZ and the large number of reported robust associations, we chose to survey the regulatory landscape of GWAS loci for this complex disorder using a custom MPRA assay, and interrogate the differential regulatory activity of the disease-

associated sequences and the differences between risk and non-risk alleles. In our list of variants of interest and because of available space on the synthesis array, we included a few of the variants reported to be associated with AD, another interest of our laboratory. As each of these robust associations is bound to be driven by variants of some biological importance, identifying variants that alter enhancer activity in-vitro can be a means to screen for the subset of variants among the many in LD that are of functional significance.

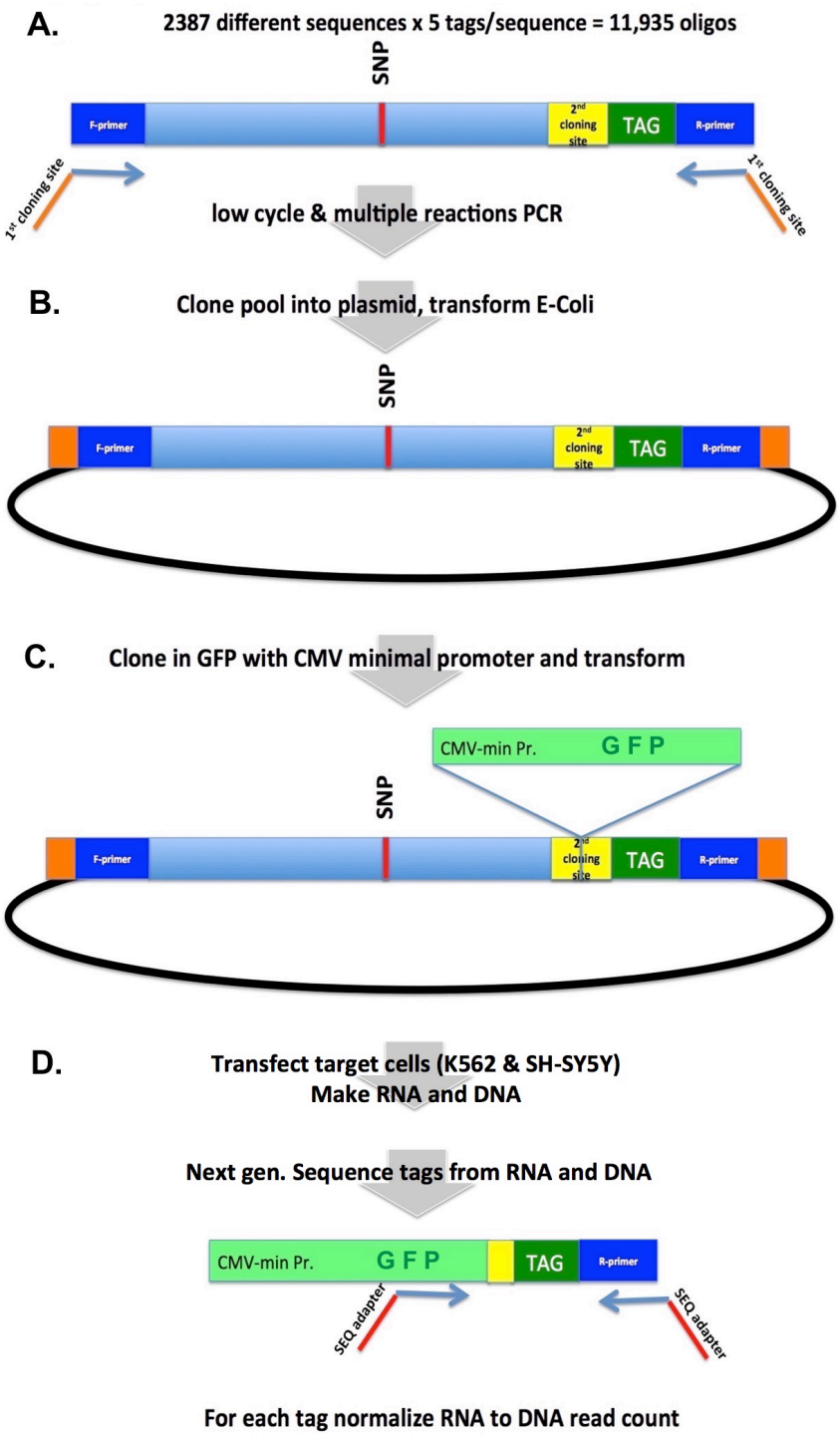

Figure 1: Experimental protocol for MPRA (See text for details).

Results

MPRA design

As described in detail in the methods our final MPRA design included oligonucleotides of 150 bp including 95 bp centered on each. Each allele of 1,053 SZ and 30 AD SNPs was tagged by 5 different tags. A positive control consisted of tiled oligos from the promoter of the human *EEF1A1* gene. A negative control was similarly tiled segments of a 465 bp sequence from a pseudogene intron, overlapping no open chromatin signals in the encode data (chr1:14,992-15,456, hg38). The final pool consisted of 11,935 oligonucleotides. The experimental process is summarized in Figure 1. We performed 3 independent transfections in the K562 cell line and 2 batches of 3 independent transfections in SH-SY5Y.

MPRA quality checks

Barcode representation

	K562	SH-SY5Y (batch 1)	SH-SY5Y (batch 2)	SH-SY5Y (both batches)
Barcode representation: DNA	94.7-95.6%	93.7-96.9%	87.3-92.1%	87.3-96.9%
Barcode representation: RNA	84.9-85.5%	78.5-81.0%	57.1-81.7%	57.1-81.7%
Barcodes always absent: DNA	2.50%	1.90%	5.10%	1.30%
Barcodes always absent: RNA	7.90%	13.10%	15.30%	9.10%
Elements always absent: DNA	0.2% (5)	0.3% (6)	0.5% (11)	0.2% (5)
Elements always absent: RNA	0.4% (10)	1% (24)	1.2% (29)	0.5% (12)

Table 1. Barcode and element representation across cell lines. Sample size for rows 1 to 4 is 10,106 barcodes. Sample size for rows 5 and 6 is 2,383 elements. The ranges in rows 1 and 2 indicate a range over the 3 or 6 samples represented in a column.

Of the initial 11,935 designed barcodes, 1,829 were never present in DNA or RNA. We consider the remaining pool of 10,106 barcodes (corresponding to 2,383 elements of the

original 2,387) to be the total pool of possible barcodes for assessing experimental quality. Barcode representation (fraction of barcodes with non-zero counts) was high in both DNA and RNA in the two cell lines tested (Table 1). For DNA, approximately 94% of barcodes were represented in K562 samples (where a sample is an independent transfection experiment) and SH-SY5Y batch 1 samples, and about 90% in SH-SY5Y batch 2 samples. For RNA, approximately 85% of barcodes were represented in K562 samples, 80% in SH-SY5Y batch 1 samples, and 60% in SH-SY5Y batch 2 samples. While we cannot be certain about the completeness of the oligonucleotide synthesis, the less than 100% representation is likely due to loss of oligonucleotides during the amplification and cloning steps rather than the transfection (as suggested by earlier quality control experiments - not shown here). The lower representation in RNA is likely due to low/undetectable expression of some transfected oligonucleotides. Some of the 10,106 barcodes were unrepresented in all transfections (Table 1, rows 3 and 4). After summing counts across barcodes for each element, only a small number of elements still had missing counts (Table 1, rows 5 and 6).

Correlation of count and activity measures between samples

To summarize count information for each MPRA element, we sum counts over barcodes to obtain one count per element per sample. These aggregated counts are used to compute activity measures

$$\alpha = log2\left(\frac{aggregated\ RNA + 1}{aggregated\ DNA + 1}\right)$$

This estimator is expected to have lower bias than an estimator that uses the mean of barcode-specific activity measures [14].

Aggregated DNA counts show high between-sample correlations (Figures 2 and 3). Pairwise correlations of the log2-transformed aggregated counts ranged from 0.981 to 0.998. Correlations between samples within a single cell line were about the same as correlations between samples in different cell lines. For RNA, pairwise correlations of the log2-transformed aggregated counts ranged from 0.862 to 0.978 (Figure 2). Correlations between samples within a single cell line were somewhat higher than correlations between samples in different cell lines. The latter correlations ranged from 0.827 to 0.928.

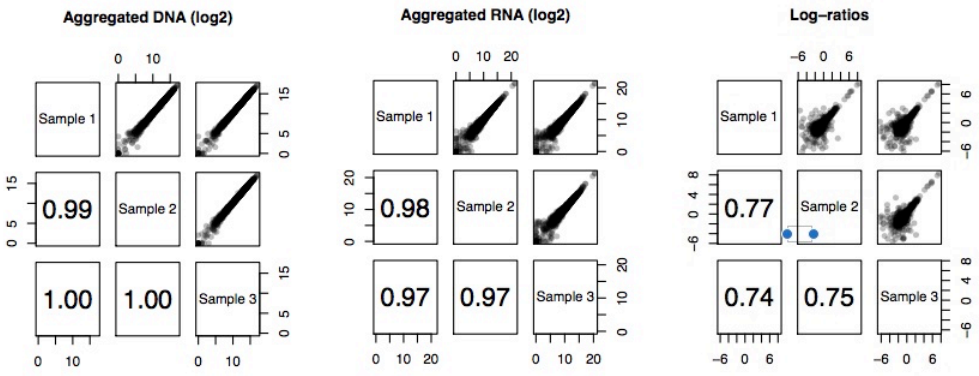

Figure 2. Correlations in the K562 cell line. The upper panels show scatterplots of MPRA quantities between samples. The lower panels indicate the correlation between a

pair of samples.

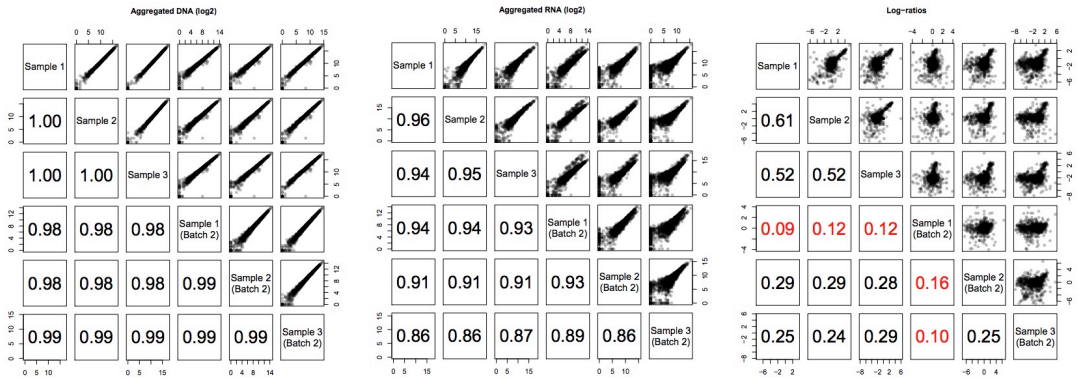

Figure 3. Correlations in the SH-SY5Y cell line. The upper panels show scatterplots of MPRA quantities between samples. The lower panels indicate the correlation between a pair of samples. Correlations are displayed in red if less than 0.2.

Activity measures show between-sample correlations ranging from 0.742 to 0.774 in the K562 cell line (Figure 2). In the first batch of the SH-SY5Y cell line, activity measures show between-sample correlations ranging from 0.519 to 0.607 (Figure 3). In the second batch, between-sample correlations were low, ranging from 0.104 to 0.251. Generally, correlations between samples within a single cell line are higher than correlations between samples in different cell lines. Figure 4). The correlations for the second batch of SH-SY5Y cells were low and we considered removing them from analysis, we found however that including it did increase our power for detecting activity differences. We note that the lower correlations between activity measures in SH-SY5Y are driven by an increased abundance of inactive elements, whose activity measures are simply noise. We do see higher correlation between samples for the active elements.

All SNPs and their p-values for differential activity in either cell line are listed in supplementary table 1. The same file includes a tab listing SNPs that only reached significance with the inclusion of batch 2 . Based on previous experience of our laboratory, we believe that the lower performance of the SH-SY5Y cell line is largely due to the nature of these cells. In our experience, these cells are more difficult to transfect and show lower efficiencies compared to K562.

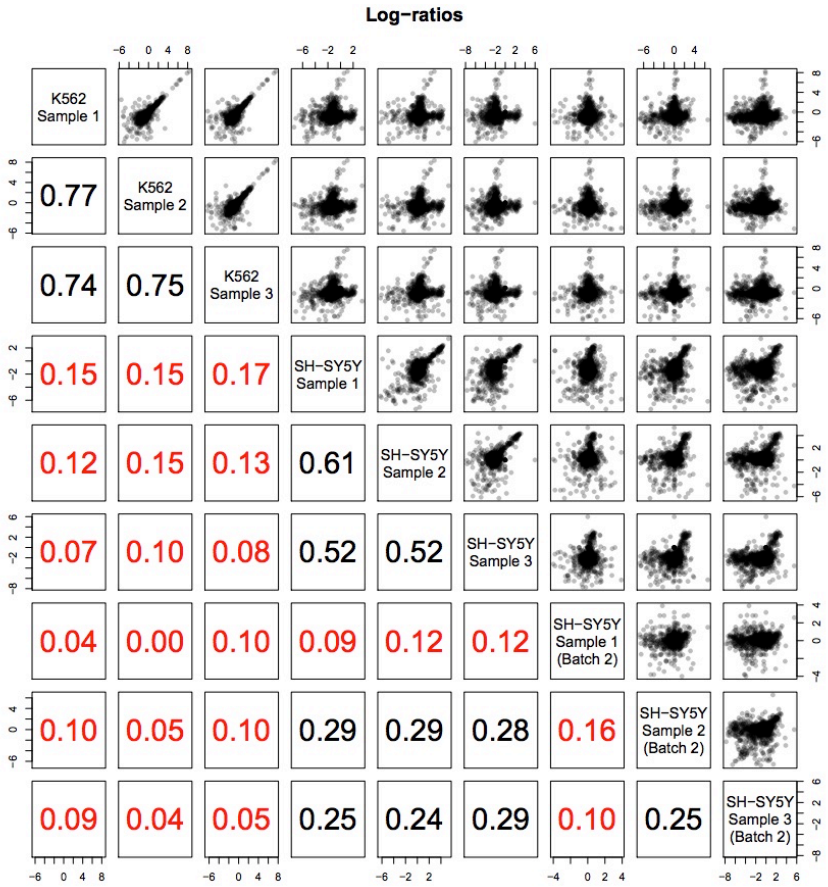

Figure 4. Correlations of log-ratio activity measures in both the K562 and SH-SY5Y cell lines. Correlations are displayed in red if less than 0.2. Generally, correlations within cell lines are higher than across cell lines.

Activity of positive and negative control sequences

In the K562 cell line, the *EEF1A1* promoter sequences (positive control) showed significantly higher activity than the negative control sequences (95% CI for difference in log-ratio activity measures: 0.50-1.12). We should note that the negative control was chosen for the absence of evidence of being an enhancer, but we have no evidence that definitely exclude such function.

In the SH-SY5Y cell line, the *EEF1A1* promoter sequences when examined all together did not show significantly higher activity measures than the negative control sequences in either of the two batches (Batch 1 95% CI for difference in log-ratio activity measures: -0.25-0.33. Batch 2 95% CI for difference in log-ratio activity measures: -0.31-0.21). We noticed however that not all the *EEF1A1* promoter sequences which were tiled in overlapping segments show activity. This is illustrated in Figure 5 where counts and log-ratio activity measures are shown as a function of tile location. Oligos that show clear elevated activity are more pronounced in the K562 cell line. The same oligos however show elevated activity in the SH-SY5Y cell line albeit to a lesser extent. The overall lack of significantly increased signal over the negative controls in the neural cell line appears to be attributable to the inclusion of a large fraction of inactive sequences and the increased variability of count measures in the neural cell line.

When we select sub-sets of *EEF1A1* oligos at different activity levels in the K562 cell line and compare to the negative control oligos, we see that above a certain threshold these subsets of positive controls also show significantly higher activity than the negative controls in both batches of the neuronal line (Figure 6). Not surprisingly, batch 1 shows

significant increases in positive control activity over a wider range of thresholds than batch 2.

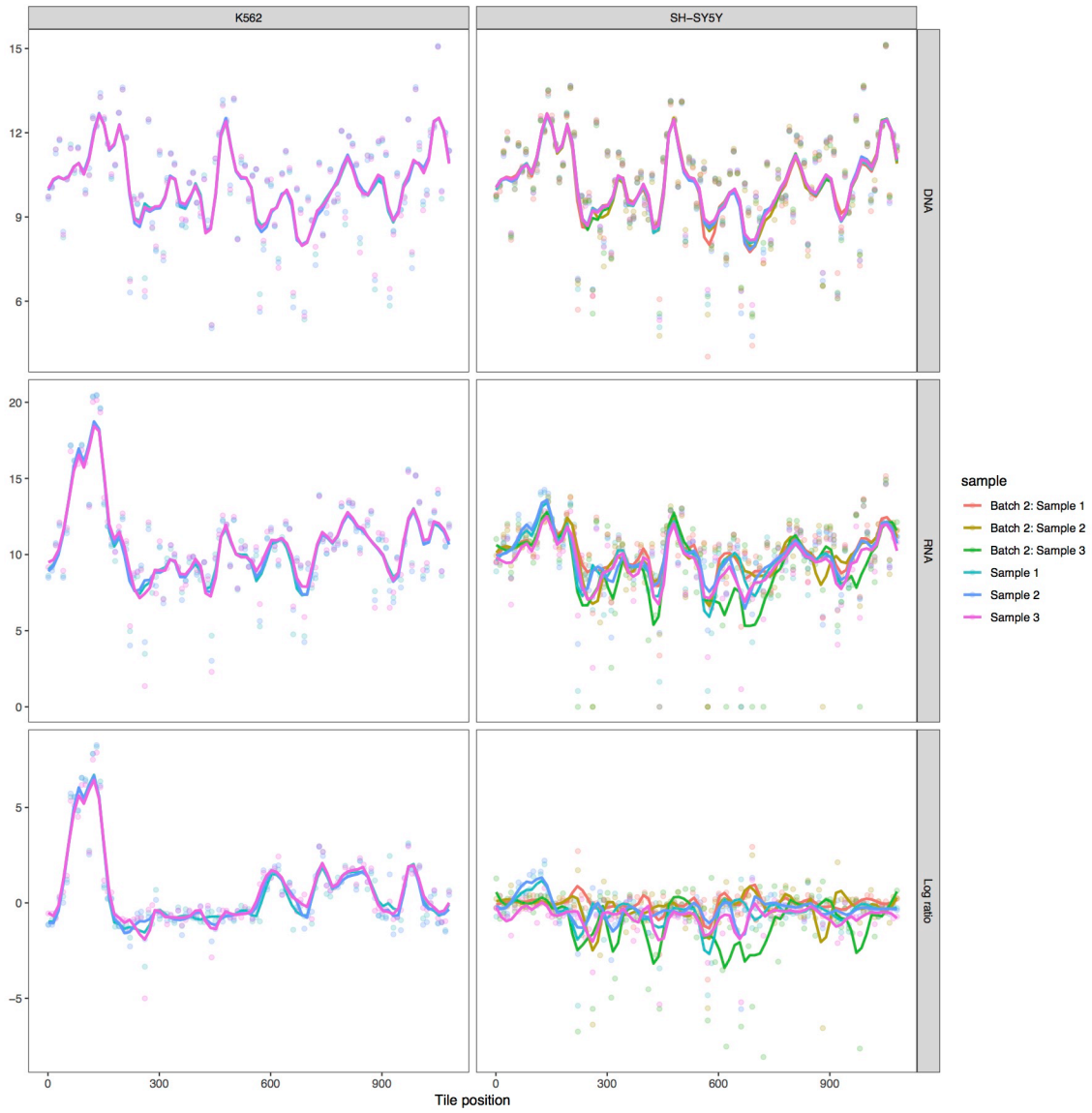

Figure 5. MPRA measures for *EEF1A1* positive control sequences. Activity is shown on the Y axis for every tile, ordered on the Y axis, for each replicate (sample), batch and cell line.

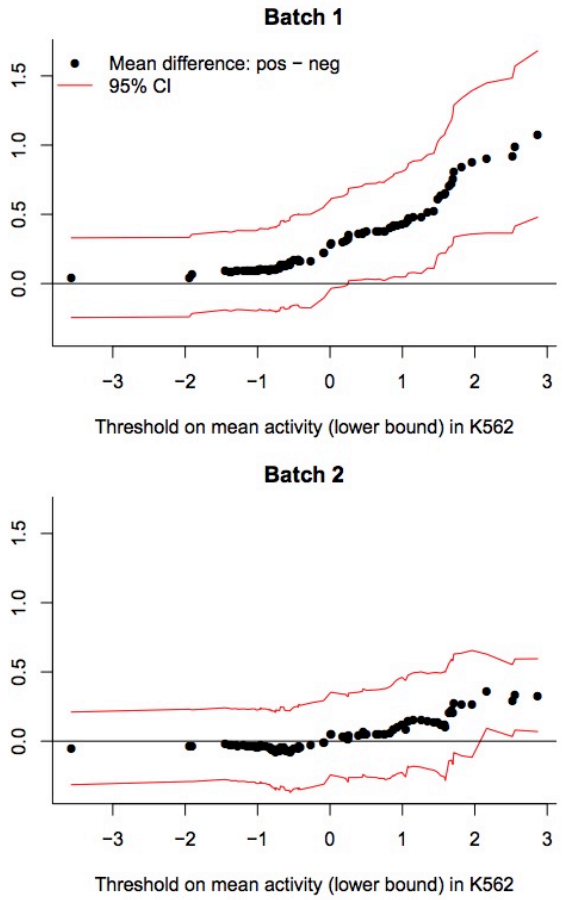

Figure 6. Comparison of activity levels in subsets of *EEF1A1* positive control and negative control sequences: neuronal cell line. The x-axis shows different activity thresholds for defining the set of positive controls, using data in the K562 cell line. The set of negative controls is fixed. For each set of positive controls we show the difference with the negative controls (Y axis) and the confidence intervals for the difference. In batch 1, moderately active positive control oligos (K562 mean activity greater than ~ 0) show significantly higher activity than the negative controls. In batch 2, only the most highly active positive control oligos show significantly higher activity than negative controls.

Evaluation of variants that show differential enhancer activity

Location of SNPs with differential enhancer activity

We successfully assayed 1,079 SNPs and found 144 SZ and 4 AD SNPs that show differential enhancer activity between alleles (significant allelic skew) in the K562 cell line and 50 and 3 in the SH-SY5Y cell line (FDR < 0.05). These SNPs, which we call significant SNPs for the purposes of this paper, are located on several different chromosomes (Figure 7). Nine SNPs showed allelic differences in both K562 and SH-SY5Y: rs73036086, rs2439202, rs6801235, rs134873, rs13250438, rs2605039, rs7582536, rs1658810, rs8061552. Of those all except rs2439202 were in the same direction (binomial p=0.02). Six of the remaining 8 were also eQTLs for one or more gene in the dorsolateral prefrontal cortex (DLPFC) in the CommonMind consortium data (http://www.nimhgenetics.org/available_data/commonmind).

On average, there are 2.6 significant SNPs per locus (median: 1 significant SNP per locus), and on average, 18% of SNPs tested were significant (median: 13%). Out of the 73 succesfully tested loci, 19 do not have any significant SNPs, and 5 of these 19 are single-SNP loci. Of the 54 GWAS loci with SNPs showing significant allelic differences, in 31 there were eQTLs for nearby genes among the SNPs (66%), however the rate was also high for 16 of the loci not showing significant allelic differences (74%).

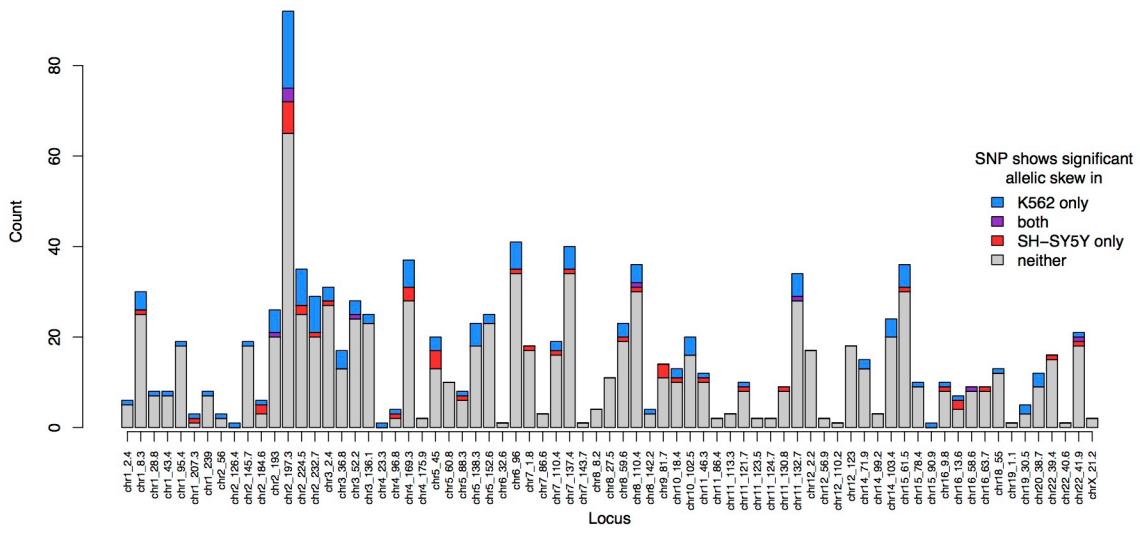

Figure 7. Genomic locations of assayed SNPs and SNPs showing allele specific activity. The SNPs assayed in this MPRA come from a variety of linkage disequilibrium (LD) blocks across several chromosomes. SNPs that show differential regulatory activity between alleles are also present in a wide variety of loci. Loci are labeled by chromosome and location on hg19.

There is no preferred direction of effect on enhancer/repressor activity for risk versus non-risk alleles

If a SNP exerts its influence in SZ or AD risk by affecting an enhancer for a nearby gene, is it more likely to decrease or increase enhancer activity? To explore this question, we overlaid our MPRA results with GWAS information on effect directions. Because of the larger number of SNPs and to avoid exploring different diseases together we did the analysis only for SZ SNPs.

Of the 144 significant SZ SNPs in K562, 81 (56%, 95% binomial CI: 47%-64%) are SNPs for which the risk allele from GWAS shows lower MPRA enhancer activity than the non-risk allele. Of the 50 significant SZ SNPs in SH-SY5Y 25 (50%, 95% binomial CI: 35%-65%) are SNPs for which the risk allele shows lower MPRA enhancer activity. Overall, these results suggest that there is no specific direction of effect on expression characterizing disease risk variants. It must be noted however that this analysis examines the effect of each individual SNP on the disease-associated haplotype in isolation. It is likely that selection has favored LD of alleles with opposing effects, balancing each other towards a favorable combined effect. It is also likely that there are interactions between such regulatory sequences that are not captured in our experiment.

MPRA activity levels show concordance with chromatin accessibility measures

We find that MPRA activity levels in the two cell lines show good overlap with cell line-specific digital genomic footprinting (DGF) tracks from UCSC (see methods). In particular, highly active SNPs in the K562 cell line show greater enrichment for K562 DGF sites than highly active SNPs in the SH-SY-5Y cell line. Inversely, highly active SNPs in the SH-SY-5Y cell line show greater enrichment for SkNSHRA (a retinoic acid-treated cell line of which SH-SY5Y are a sub clone) DGF sites than highly active SNPs in the K562 cell line (Figure 8).

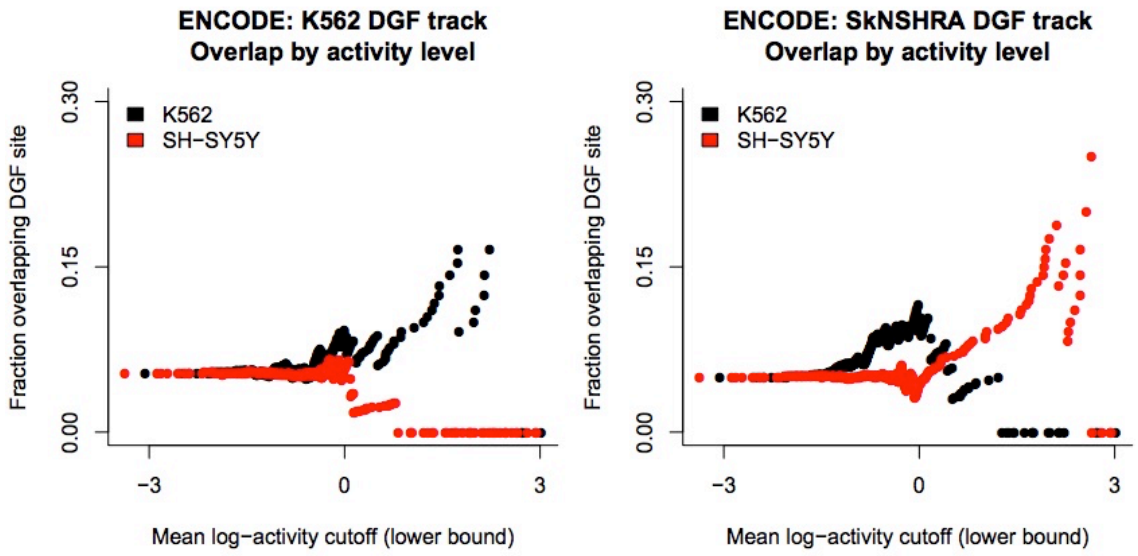

Figure 8. DGF enrichment by activity level. Every point corresponds to a set of SNPs. This set is defined by a cutoff on the mean MPRA activity. For these sets, the fraction of SNPs overlapping a DGF site is shown on the y-axis for the K562 cell line (left) and for the SkNSHRA cell line (right).

Identification of disrupted features contributing to differential activity

We searched for transcription factor (TF) binding motifs overlapping SNP positions and compared binding scores between alleles to assess potential disruptions in TF binding. We used the union of the JASPAR2016 and ENCODE databases for a total of 2450 TF motifs. The main match metric that we use is the relative score, which is the ratio of the observed position weight matrix (PWM) score at a given position divided by the maximum possible score for the PWM. Relative scores fall between 0 and 1. Due to the large number of motifs examined, most SNPs in our MPRA assay showed "convincing" binding (high relative score) for at least one TF motif. In fact, all MPRA SNPs have at least one TF motif match with a relative score of at least 0.963 (Figure 9).

Figure 9. Distribution of PWM relative scores. The y-axis shows the fraction of MPRA SNPs that have at least one TF motif match with a relative score greater than or equal to the value given on the x-axis. The dashed vertical line shows the highest relative score (0.963) at which all SNPs still have a motif match.

Given that a PWM relative score gives a measure of how strongly a particular TF might bind to a region, we found it reasonable to ask if a higher score was associated with higher enhancer activity as measured by our MPRA. For each match between a TF motif and an MPRA oligo, we obtained a PWM relative score for each allelic version of the oligo. For each MPRA oligo, we could also compute the mean activity level for each allele from MPRA data. We determined whether the PWM relative scores for the different alleles were positively or negatively correlated with the mean activity levels for the different alleles. We did not observe a correlation of PWM with MPRA activity levels in either the K562 or the SH-SY5Y cell line. This is probably expected given the noise in examining such a wide variety of motifs across so many sequences.

Support for combinatorial SNP effects leading to large LD blocks

Ours [15] and others' work [16] suggests that within large LD blocks there are often multiple regulatory variants that presumably combine their effects in regulating their target gene. In such genomic regions selective pressure is likely to favor specific arrangements of alleles to optimally combine positive and negative regulators. By doing so selection may drive the formation of larger haplotypes containing more functional SNPs along with some passenger SNPs in LD. In such regions the fraction of functional SNPs $\left(\frac{\text{Functional}}{\text{Functional+Passengers}}\right)$ would be greater than in genomic regions where only one SNP is functional and all others are passengers. Our data that include a mix of smaller and larger LD groups provides an opportunity to test this hypothesis. We did so on our K562 cell line data, as it had the most positive results. We did not include positives from SH-SY5Y to avoid confounding from differences in regulation between cell lines. We first tested whether blocks that include more SNPs have a higher fraction of positives $\left(\frac{\text{Significant}}{\text{all SNPs in LD}}\right)$ in our assay than smaller blocks. The upper bar graph in Figure 10 shows the fraction of significant SNPs at different LD block sizes.

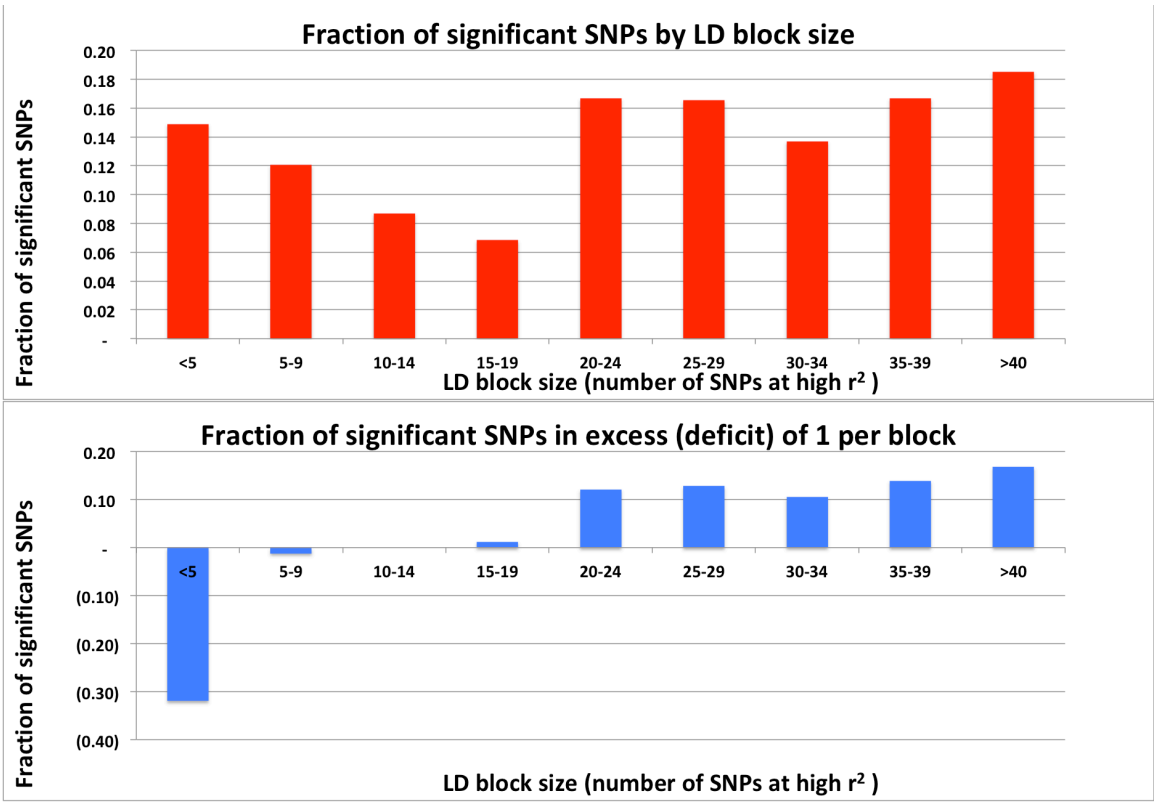

Figure 10. Fraction of significant SNPs by LD block size. The fraction of tested SNPs that are significant is shown for different bins of block size in terms of number of SNPs in LD (red bars). It is observed that smaller blocks have lower density. A density increase at very small sizes is also evident, likely due to a constant enumerator (1 expected significant SNP) and a quickly decreasing denominator (all SNPs in the block). After converting total number of significant SNPs to excess significant SNPs (more than the minimum expected one per block) the difference becomes more pronounced (lower graph with blue bars). The number of observations in each bin from left to right is 47, 83, 115, 161, 132, 133, 95, 144 and 173.

We observed that blocks of fewer than 20 SNPs showed a smaller fraction of significant SNPs, a difference that was highly significant (Chi-square p = 7.5e-4, top of Figure 10)

even if we were to test all 8 possible cut offs (corrected p=0.006). There was also a reverse trend from small to very small blocks. This reversal may be because every block is expected to contain at least one functional SNP, but for very small blocks the fraction $\left(\frac{\text{Significant}}{\text{all SNPs in LD}}\right)$ will also have a small denominator. If we remove the expected 1 significant SNP for every block in each size bin and graph the excess significant SNPs for every bin we see a much more pronounced difference between larger and smaller blocks (blue bars at bottom graph in Figure 10). We stress, however that this correction introduces a bias and this graph is only meant to illustrate the point, not to assess statistical significance of this effect.

Discussion

We report the results of an MPRA for SNPs associated with SZ and AD on two different cell lines, K562 and SH-SY5Y. The purpose of this work is to add to the data available to us and other investigators in the effort to sort among the multiple SNPs showing statistical associations with disease and identify those likely to have a functional consequence. We report a total of 192 SNPs among the total of 1079 tested (18%) that show statistically significant allelic differences (FDR = 0.05) in driving reporter gene expression. While we recognize the many caveats inherent to such experiments as discussed below, we consider these variant sequences strong candidates for underlying the observed disease associations and an important resource for SZ and AD research.

There are a number of observations that support the validity of our results. The differences between our positive and negative controls along with the concordance of

direction of signals that emerged as significant in both cell types are all consistent with true signals. The significant and cell type specific overlap with open chromatin marks is also an expected feature of true signals. Whether a higher overlap should be expected between the two cell lines is less clear. One might also argue that it would be expected to observe more eQTLs among the loci harboring significant SNPs. The numbers of eQTLs we observe are already high at ~70% of tested loci. Whether we would see a enrichment in the loci showing allelic differences in our MPRA depends on many factors including the differences between the cell lines we used and the DLPFC bulk tissue that was used to identify eQTLs.

Despite the evidence for the validity of the data there remain important caveats to keep into consideration. First, the results from the SH-SY5Y cells were significantly weaker than those of the K562 cells. We had expected that SH-SY5Y being a neuroblastoma cell line would be more informative for studying brain disorders than the K562 chronic myelogenous leukemia cell line. Unfortunately, our experience has been that SH-SY5Y cells are much less efficient for transfection both in these and in other experiments. This explains both the lower oligo representation and the lower yield of positive results. We should note that, in fact, K562 might be a very relevant cell type for neuropsychiatric phenotypes. For one, K562 is enriched for genes expressed in immunity-related tissues reported by the PGC for the SZ GWAS signals. We have also observed that the overlap of chromatin marks and SZ SNPs is higher for K562 cells than SK-N-SH-RA that SH-SY5Y are derived from (5.4% of all SZ SNPs we assayed overlap K562 DGF hotspots compared to 5.1% overlapping SK-N-SH-RA DGF hotspots). Further, the overlapping significant SNPs between the two cell types, while highly consistent in direction, are

relatively few. While it would be encouraging to observe more overlaping SNPs, this is not necessarily of concern. The activity of an episomally-introduced sequence can be influenced by numerous factors, including the combinations of transcription factors present in each cell type and the likely necessity of combinatorial effects with other enhancers that may be absent in the 95 bp DNA fragments that we test here. Finally, it must be noted that all the caveats affecting classic reporter assays, an artificial expression system, also apply here. The risk of false negatives is high, while false positives can also not be ruled out in the case that a tested variant resides in closed chromatin in the disease-relevant tissues and therefore while it has a functional potential is not important for the disease. Having tested two cell types reduces false negatives while it should have little effect on this type of false positive rate. Nevertheless, our results are only meant as a screen for functional elements. More labor intensive methods would need to be employed for each of these variants to confirm their function. Such methods, as for example specific genomic editing of variant bases and examination of the consequences in disease relevant brain cell types, can provide a definite answer regarding their role in disease.

Conclusion

In summary, we have screened 1,079 disease associated variants and found that 192 show significant allelic differences in driving reporter gene expression. These were located on 54 of 73 tested disease loci or almost three quarters. Many of the 54 loci contained more than one regulatory variant in LD and we observed a relative increase of signal density in larger LD blocks, which might indicate complex regulation and selection of haplotypes that combine specific regulatory alleles as previously suggested by us and others [15,16].

The results of our large screen of disease associated variation will be a significant spring board to facilitate future research on the genetics of these disorders and the complex role of gene regulation.

Materials and methods

Selection of variants

Variants were selected from two large GWAS, one for schizophrenia [3] and one for Alzheimer's disease [4]. The schizophrenia PGC2 data was downloaded from https://www.med.unc.edu/pgc/ which provides data on the association between schizophrenia and 9,444,231 imputed SNPs across the genome. We first determined the size of the groups of SNPs that were in LD with each other and therefore represented the same association signal. To maximize the number of independent loci investigated in our assay while comprehensively examining each tested locus, we followed specific selection rules. First, we identified the lead SNPs at each locus and identified all SNPs in the same locus with p-values up to 15 times larger. To maximize efficiency, we excluded loci where more than 45 SNPs fit this criterion (the only exception was a block on chromosome 2 with 92 SNPs). This resulted in 1,198 SNPs in 64 schizophrenia loci moving forward to oligonucleotide design. Next we added SNPs from the AD GWAS [4]. Here we first used the lead SNPs reported by Lambert to identify all SNPs in strong LD ($r^2 > 0.9$) using the bioinformatics tool SNAP [17]. The largest LD groups were then removed resulting in the addition of 30 SNPs across 9 AD GWAS loci to our assay, for a total of 1,228 SNPs selected to be included in the MPRA. All successfully assayed SNPs and the results of our assay are in supplementary Table 1.

MPRA design and methods

Oligonucleotides of 150 bp including 95 bp centered on each SNP and 45 flanking based for amplification and cloning purposes (see below) were synthesized for each sequence of interest, as described in [18] and [19]. Target sequences were flanked with a multiple cloning site, a unique tag and primers for PCR amplification (Figure 1). Each tested sequence (each allele in the case of SNPs) was tagged by 5 different tags. We successfully designed oligonucleotides for 1083 of our selected SNPs (1053 for SZ and 30 for AD), including 27 that had 3 alleles and 5 had 4 alleles. In addition, we synthesized 545 oligonucleotides covering 1,179 bp from the promoter of the human *EEF1A1* gene (chr6:73,520,057-73,521,235, hg38) in 95 bp segments with 10 bp overlap as a positive control. A negative control was similarly tiled segments of a 465 bp sequence from a pseudogene intron, overlapping no open chromatin signals in the encode data (chr1:14,992-15,456, hg38). The final pool of 11,935 oligonucleotides was ordered from the Broad Technology Labs (Broad Institute, Cambridge MA) This included 5 oligonucleotides, each with a different barcode, for each SNP allele for a total of 10 oligonucleotides per bi-allelic SNP.

The library we received from the BROAD institute was amplified, cloned and transfected as described [18,19] with slight modifications. The experimental design is summarized in Figure 1. First, using the primer sequences built into the synthetic oligos the library was amplified through a low-cycle (25 cycles) PCR reaction in 7 separate wells to preserve oligo representation and minimize PCR bias (Figure 1A). The 7 PCR products were gel-purified and pooled together for subsequent steps. The primers used for PCR introduced

two distinct SfiI sites as described [19]. These sites were used for the subsequent directional cloning (1st cloning site in Figure 1A) into the pMPRA1 plasmid (Addgene ID# 49349, Figure 1B). The plasmids were used to transform E.Coli by electroporation. The synthesized oligos contained a second directional cloning site (restriction enzymes KpnII, XbaI) which was used in the next step to insert a CMV minimal promoter driving GFP between the putative regulatory sequence and the corresponding tag sequence within the synthesized oligonucleotide (2nd cloning site, Figure 1C). After these two rounds of cloning - E.Coli transformation, we used the resulting plasmid library to perform three independent transfections of K562 chronic myelogenous leukemia cells and six independent transfections in SH-SY5Y human neuroblastoma cells in two batches by lipofectamine3000 (Thermofisher Scientific cat. no L3000008). We then extracted DNA and RNA and performed 50 bp non-paired end sequencing reading from the 3'end, through the unique tag and into the GFP transcript on an ILLUMINA MiSeq DNA Sequencer (Illumina, Inc. San Diego, CA). For each triplicate transfection we acquired 10 - 27 million reads of DNA and RNA.

Differential analysis

Before statistical analysis, we normalize both the RNA and DNA counts using total count normalization. That is, we scale the counts in each sample so that all samples have the same library size. For differential analysis of MPRA activity levels between alleles, we use the mpralm method [14]. This approach uses linear models to directly model activity measures and uses a combination of observation-level weighting and empirical Bayes methods to improve estimation of element-specific variances. In our study, comparisons

of activity measures between alleles are paired in the sense that each biological replicate measures the activity of all (2 or more) alleles of a SNP. The assaying of all alleles simultaneously in a given sample leads to correlation between their measurement, and for this reason, we use the mixed model approach of mpralm for making comparisons.
Our data for the SH-SY5Y cell line come from two separate batches, so before differential analysis we use the ComBat method [20] to correct the log-ratio activity measures. This method estimates location and scale parameters of batch effects at the element level and moderates these estimates using an empirical Bayes technique that pools information on these estimates across all elements. With these parameter estimates, batch-corrected outcome measures can be computed. From here, we proceed as detailed above with differential analysis.

Transcription factor binding analysis

To search for transcription factor (TF) motifs that can be affected by the SNPs in our study, we used the TFBSTools package available on Bioconductor [21]. We used the union of the JASPAR2016 and ENCODE TF databases for a total of 2450 motifs. We only considered matches in which the motif overlapped the SNP position within the oligo and in which the position weight matrix (PWM) score was at least 70% of the maximum score possible for the PWM (i.e. a relative score of at least 70%) for at least one allele of a SNP.

Comparisons with open chromatin measures

To assess whether MPRA active elements where overlapping with cell line-specific digital genomic footprinting (DGF) tracks we downloaded from UCSC the following files: hgdownload.cse.ucsc.edu/goldenpath/hg19/encodeDCC/wgEncodeUwDgf, file names wgEncodeUwDgfK562Hotspots.broadPeak.gz and wgEncodeUwDgfSknshraHotspots.broadPeak.gz. These correspond to data from K562 cells and retinoic acid treated SK-N-SH cells. SH-SY5Y are a sub-clone of SK-N-SH which is the closest cell line for which we could access DGF data.

Acknowledgements

This work was supported in part by NIMH grants R56MH113215 to DA and P50MH094268 (project 1 PI: DA).

References

1. Visscher PM, Wray NR, Zhang Q, Sklar P, McCarthy MI, Brown MA, et al. 10 Years of GWAS Discovery: Biology, Function, and Translation. Am J Hum Genet. 2017;101: 5–22. doi:10.1016/j.ajhg.2017.06.005

2. Pardiñas AF, Holmans P, Pocklington AJ, Escott-Price V, Ripke S, Carrera N, et al. Common schizophrenia alleles are enriched in mutation-intolerant genes and in regions under strong background selection. Nat Genet. 2018;50: 381–389. doi:10.1038/s41588-018-0059-2

3. Schizophrenia Working Group of the Psychiatric Genomics Consortium. Biological insights from 108 schizophrenia-associated genetic loci. Nature. 2014;511: 421–427. doi:10.1038/nature13595

4. Lambert JC, Ibrahim-Verbaas CA, Harold D, Naj AC, Sims R, Bellenguez C, et al. Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. Nat Genet. 2013;45: 1452–1458. doi:10.1038/ng.2802

5. Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, et al. Potential etiologic and functional implications of genome-wide association loci for

human diseases and traits. Proc Natl Acad Sci U S A. 2009;106: 9362–9367. doi:10.1073/pnas.0903103106

6. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, et al. Finding the missing heritability of complex diseases. Nature. 2009;461: 747–753. doi:10.1038/nature08494

7. Visel A, Rubin EM, Pennacchio LA. Genomic views of distant-acting enhancers. Nature. 2009;461: 199–205. doi:10.1038/nature08451

8. Maurano MT, Humbert R, Rynes E, Thurman RE, Haugen E, Wang H, et al. Systematic localization of common disease-associated variation in regulatory DNA. Science. 2012;337: 1190–1195. doi:10.1126/science.1222794

9. Cookson W, Liang L, Abecasis G, Moffatt M, Lathrop M. Mapping complex disease traits with global gene expression. Nat Rev Genet. 2009;10: 184–194. doi:10.1038/nrg2537

10. Schaub MA, Boyle AP, Kundaje A, Batzoglou S, Snyder M. Linking disease associations with regulatory information in the human genome. Genome Res. 2012;22: 1748–1759. doi:10.1101/gr.136127.111

11. Castel SE, Cervera A, Mohammadi P, Aguet F, Reverter F, Wolman A, et al. Modified penetrance of coding variants by cis-regulatory variation contributes to disease risk. Nat Genet. 2018;50: 1327–1334. doi:10.1038/s41588-018-0192-y

12. Inoue F, Ahituv N. Decoding enhancers using massively parallel reporter assays. Genomics. 2015;106: 159–164. doi:10.1016/j.ygeno.2015.06.005

13. Muerdter F, Boryń ŁM, Arnold CD. STARR-seq - principles and applications. Genomics. 2015;106: 145–150. doi:10.1016/j.ygeno.2015.06.001

14. Myint L, Avramopoulos DG, Goff LA, Hansen KD. Linear models enable powerful differential activity analysis in massively parallel reporter assays. bioRxiv. 2018; doi:10.1101/196394

15. Eckart N, Song Q, Yang R, Wang R, Zhu H, McCallion AS, et al. Functional Characterization of Schizophrenia-Associated Variation in CACNA1C. PloS One. 2016;11: e0157086. doi:10.1371/journal.pone.0157086

16. Corradin O, Saiakhova A, Akhtar-Zaidi B, Myeroff L, Willis J, Cowper-Sal lari R, et al. Combinatorial effects of multiple enhancer variants in linkage disequilibrium dictate levels of gene expression to confer susceptibility to common traits. Genome Res. 2014;24: 1–13. doi:10.1101/gr.164079.113

17. Johnson AD, Handsaker RE, Pulit SL, Nizzari MM, O'Donnell CJ, de Bakker PIW. SNAP: a web-based tool for identification and annotation of proxy SNPs using

HapMap. Bioinforma Oxf Engl. 2008;24: 2938–2939. doi:10.1093/bioinformatics/btn564

18. Kheradpour P, Ernst J, Melnikov A, Rogov P, Wang L, Zhang X, et al. Systematic dissection of regulatory motifs in 2000 predicted human enhancers using a massively parallel reporter assay. Genome Res. 2013;23: 800–811. doi:10.1101/gr.144899.112

19. Melnikov A, Murugan A, Zhang X, Tesileanu T, Wang L, Rogov P, et al. Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. Nat Biotechnol. 2012;30: 271–277. doi:10.1038/nbt.2137

20. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostat Oxf Engl. 2007;8: 118–127. doi:10.1093/biostatistics/kxj037

21. Tan G, Lenhard B. TFBSTools: an R/bioconductor package for transcription factor binding site analysis. Bioinforma Oxf Engl. 2016;32: 1555–1556. doi:10.1093/bioinformatics/btw024