**Background**

MPRA (Massively Parallel Reporter Assay) is a technique used in molecular biology to study the regulatory elements of DNA and their effects on gene expression. It involves introducing a library of DNA sequences into cells and measuring the activity of these sequences using a reporter gene. By comparing the activity of different DNA sequences, researchers can identify functional elements and study their effects on gene expression.

The expression effects in MPRA refer to the changes in gene expression that are observed when different DNA sequences are tested. By analyzing the activity of different DNA variants, researchers can determine which sequences enhance or suppress gene expression and to what extent. This information provides insights into the regulatory elements and mechanisms that control gene expression.

Allelic effects in MPRA specifically refer to the impact of genetic variations or allelic differences on gene expression. By introducing different alleles of a particular DNA sequence into cells and measuring their activity, researchers can assess how genetic variation influences gene expression levels. This allows them to identify alleles that enhance or diminish gene expression and understand the functional consequences of genetic variation on gene regulation.

MPRA experiments can help uncover important regulatory regions, such as enhancers or promoters, within the genome. They can also shed light on the effects of specific genetic variants on gene expression, which is particularly relevant for understanding the genetic basis of diseases and traits. By combining the information obtained from MPRA with other genomic data, researchers can gain a deeper understanding of gene regulation and its contribution to phenotypic variation.

**Papers**

**Paper1: Functional regulatory variants implicate distinct transcriptional networks in dementia**

**Genome**: hg19

**Region**: regulatory regions (enhancer/promoter)

**Diseases**: Alzheimer’s disease (AD) and progressive supranuclear palsy (PSP)

**Cell lines**: human embryonic kidney 293T (HEK293T) cells

**GWAS/QTL**:

**Finding:** First, we performed two MPRAs to screen both alleles of 5706 non- coding variants in linkage disequilibrium [squared correlation coefficient (R2) > 0.8] with all genome-wide significant polymorphisms from two recent AD GWASs (5, 6) and a PSP GWAS (8). In the first MPRA (MPRA 1), we tested 5223 variants encompassing 14 AD and five PSP GWAS loci. In the second MPRA (MPRA 2), we replicated 326 variants from MPRA 1 to determine assay reproducibility and screened another 483 variants encom- passing 11 additional AD loci and four newly identified suggestive loci for PSP (Fig. 1, B and C; fig. S1; and Table 1) (20).

The majority of frVars (94%) overlapped two or more known functional annotations in human brain tissue or blood, nearly two-thirds of which were predicted to disrupt transcription factor binding, indicating their potential relevance in human disease. Forty-two high- confidence regulatory variants distributed across 15 AD loci and three PSP loci were selected for validation, using either CRISPR droplet sequencing (CROP-seq) or direct CRISPR excision in induced pluripotent stem cell–derived neurons, microglia, and astrocytes, enabling validation of 19 functional variants, implicating 20 risk genes across 11 loci.

**Paper2: Multiple causal variants underlie genetic associations in humans**

**Genome**: hg19

**Region**:

**Diseases**: None

**Cell lines**: Lymphoblastoid cell lines (LCL)

**GWAS/QTL**: We applied an MPRA to systematically char- acterize causal variants underneath multi- ple expression QTL (eQTL) and GWAS loci. We selected independent, common, and top- ranked eQTL across 744 eGenes identified in the CEU cohort (which comprises Utah resi- dents of Northern and Western European an- cestry). Each eQTL had a median of six lead associated variants (range of 1 to 472) in per- fect LD. For each lead variant, we identified all additional variants with a correlation co- efficient (r2) ≥ 0.85 that were associated with the same gene, as well as a set of variants (n = 2114 non-eQTLs) that were not associated with any gene’s expression. Our final library included 30,893 variants, with a median of 50 variants per eQTL (range of 2 to 2824) (Fig. 1A).

**Finding:** To measure regulatory effects from oligo counts, we used negative binomial regression. For each variant, we computed the allele- independent regulatory effects of an oligo (“expression” effects) and the difference in regulatory effects between reference and alternative allele-containing oligos (“allelic” effects). We detected 8502 expression effects and 1264 allelic effects across all tested variants.

By design, a subset of tested variants (n = 782) were previously identified as expression- modulating variants in (8). This overlapping subset was highly enriched for expression and allelic effects (Fig. 1, C and D). Further, we ob- served that 89.6% of allelic MPRA hits in both datasets were directionally concordant (fig. S2A). From these results, we constructed a concordant, high-confidence “MPRA-positive” variant set that contains 250 variants with expression effects and 120 with allelic effects.

**Paper3: Transcriptomic signatures across human tissues identify functional rare genetic variation**

**Genome**: hg38

**Region**: whole genome

**Diseases**: None

**Cell lines**: GM12878 LCL

**GWAS/QTL**: GTEx RV

**Finding:** We further applied both a massively parallel reporter assay (MPRA) and a CRISPR-Cas9 assay to assess the impact of Watershed- prioritized RVs. We experimentally tested the regulatory effects of 52 variants with mod- erate Watershed expression posterior (≥0.5) and 98 variants with low Watershed expres- sion posterior (<0.5) using MPRA. Next, we assessed the functional effects of 20 variants by editing them into inducible-Cas9 293T cell lines. These included 14 rare stop-gained variants and six non-eQTL common variants as negative con- trols.

**Paper4: Genome-wide functional screen of 30 UTR variants uncovers causal variants for human disease and evolution**

**Genome**: hg19

**Region**: 3UTR

**Diseases**: NHGRI-EBI GWAS catalog

**Cell lines:**

|  |
| --- |
| HMEC |
| HEK293FT |
| HEPG2 |
| K562 |
| GM12878 |
| SKNSH |

**GWAS/QTL**: We applied MPRAu to identify functional 30UTR variants asso- ciated with human disease and evolutionary selection, testing 12,173 30 UTR variants. As the causal variant(s) underlying human traits and diseases can be among many variants associated with GWAS tagging (tag) SNPs, we tested 30 UTR SNPs and insertion/ deletions (indels) (minor allele frequency [MAF] S5%) in strong genetic linkage, LD, with tag SNPs (LD threshold: minimum r2 = 0.8) from the NHGRI-EBI GWAS catalog (Welter et al., 2014), totaling 2,153 putative disease-associated variants from 1,556 independent association loci (Table S1). We also incorpo- rated a set of 9,325 30UTR SNPs and indels overlapping regions identified as being under positive selection in humans (Gross- man et al., 2013) (Table S1). We also included a set of 46 rare 30UTR variants (MAF %0.01 in Europeans) that are in genes with outlier expression signatures across tissues in the Geno- type-Tissue Expression (GTEx) project, which are known to have potential deleterious consequences (Li et al., 2017) (Table S1). Lastly, across all tested variants, 2,955 were also tested un- der alternative allelic backgrounds to account for the potential effect of surrounding sequence variants.

**Finding**: Confident in our assay’s ability to assess oligos with regulatory activity, we then identified tamVars altering 30UTR functionality by comparing expression changes between alleles of the same 30UTR (using as a threshold a Benjamini-Hochberg adjusted p value [BH p-adj] <0.1) (Figure 1D). We found 2,368 tamVars in to- tal across all cell types (Table S1). \

**Paper5: Functional testing of thousands of osteoarthritis-associated variants for regulatory activity**

**Genome**: NA

**Region**: noncoding

**Diseases**: **osteoarthritis**

**Cell lines:** Saos-2 cells, an osteosarcoma cell line

**GWAS/QTL**: We compiled a list of 35 lead SNPs associated with OA in Europeans via GWAS, with minor allele frequencies over 5%10–27. Each SNP represents an independent signal with p < 5e-8 (genome-wide significant; n = 20) or p < 5e-5 (genome-wide suggestive; n = 15) (Supple- mentary Table 1). We identified all SNPs in LD with an r2 > 0.8 in Europeans using rAggr (Fig. 1a), resulting in a list of 1,605 candidate SNPs. For the major and minor allele of each SNP, we synthesized 196 nucleotides (nt) of genomic sequence, centered on the SNP and flanked by adaptor sequences, on a microarray

**Finding:** To date, genome-wide association studies have implicated at least 35 loci in osteoarthritis but, due to linkage disequilibrium, the specific variants underlying these associations and the mechanisms by which they contribute to disease risk have yet to be pinpointed. Here, we functionally test 1,605 single nucleotide variants associated with osteoarthritis for regulatory activity using a massively parallel reporter assay. We identify six single nucleotide poly- morphisms (SNPs) with differential regulatory activity between the major and minor alleles. We show that the most significant SNP, rs4730222, exhibits differential nuclear protein binding in electrophoretic mobility shift assays and drives increased expression of an alter- native isoform of HBP1 in a heterozygote chondrosarcoma cell line, in a CRISPR-edited osteosarcoma cell line, and in chondrocytes derived from osteoarthritis patients. This study provides a framework for prioritization of GWAS variants and highlights a role of HBP1 and Wnt signaling in osteoarthritis pathogenesis.

**Paper6: Transcriptional-regulatory convergence across functional MDD risk variants identified by massively parallel reporter assays**

**Genome**: hg19

**Region**: enhancer

**Diseases**: major depressive disorder (MDD)

**Cell lines: N2A**

**GWAS/QTL**: Identifying candidate psychiatric GWAS regulatory variants To prioritize putative regulatory variants from neuropsychiatric disease GWAS regions (predominantly MDD; Fig. 1A), SNPs in linkage disequili- brium (LD) with GWAS tag variants at R2>0.1 were collected and intersected with histone modification, eQTL, Hi-C, and enhancer segmen- tation datasets from human postmortem tissue and neural lineage cell lines (see Supplemental Methods, Fig. 1B). SNPs were manually selected based on diversity and density of annotation overlap within each locus (Supplemental Methods). As a negative control, we identified candidates from one additional locus associated with several anthropomorphic traits [64], in a trait-blinded manner. Altogether, 1453 SNPs were selected. Final LD of selected SNPs was distributed similarly to starting SNPs (Fig. 1D)

**Finding:** Using Massively Parallel Reporter Assays (MPRAs), we functionally screened over 1000 SNPs prioritized from 39 neuropsychiatric trait/disease GWAS loci, selecting SNPs based on overlap with predicted regulatory features—including expression quantitative trait loci (eQTL) and histone marks—from human brains and cell cultures. We identified >100 SNPs with allelic effects on expression in a retinoid-responsive model system. Human genomic sequence (hg19) tiles up to 126bp were taken centered on the 1454 candidate enhancer SNPs.

Using Massively Parallel Reporter Assays (MPRAs), we functionally screened over 1000 SNPs prioritized from 39 neuropsychiatric trait/disease GWAS loci, selecting SNPs based on overlap with predicted regulatory features—including expression quantitative trait loci (eQTL) and histone marks—from human brains and cell cultures. We identified >100 SNPs with allelic effects on expression in a retinoid-responsive model system. Functional SNPs were enriched for binding sequences of retinoic acid-receptive transcription factors (TFs), with additional allelic differences unmasked by treatment with all-trans retinoic acid (ATRA).

We find that principles of the omnigenic model appear to hold true for MDD risk genetics, including the presence of far more functional variants (a total of 277 SNPs with allelic and/or interaction effects of 1178 assessed across the two assays; Supplemental Table 1) than there were GWAS loci (i.e., tag SNPs).

**Data availablity**: <https://bitbucket.org/jdlabteam/n2a_atra_mdd_mpra_paper/src/master/>

**Paper7: Saturation mutagenesis of twenty disease-associated regulatory elements at single base-pair resolution**

**Genome**: hg19/hg38

**Region**: promoter/enhancer

**Diseases**: 20 disease- associated gene promoters and enhancers

**Cell lines:** HepG2 (HB-8065), HEK293T (CRL-11268), HeLa (CCL-2), HaCaT (CRL-2404), Neuro-2a (CCL-131), LNCaP (CRL-1740), and SK-MEL-28 (HTB-72)

**GWAS/QTL**: We perform satura- tion mutagenesis in conjunction with massively parallel reporter assays on 20 disease- associated gene promoters and enhancers, generating functional measurements for over 30,000 single nucleotide substitutions and deletions.

**Finding:** Alto- gether, our MPRAs quantified the regulatory effects of 31,243 potential mutations (min. ten tags) at 9834 unique positions (Supplementary Table 8) and we setup an interactive website for exploring this dataset (https://mpra.gs.washington.edu/ satMutMPRA/). Of the unique mutations, 4830 (15%) were identified as causing significant changes relative to the wild-type promoter or enhancer sequence (p-value of fit <10−5). Of those causing significant changes, 1789 (37%) increased expression (by a median of 20%) and 3041 (63%) decreased expression (by a median of 24%). The majority of significant effects were of small magnitude. If we require a minimum two-fold change, we identify a total of 83 activating and 559 repressing mutations.

**Sever: https://mpra.gs.washington.edu/satMutMPRA/**

**Paper8: Functional dissection of inherited non-coding variation influencing multiple myeloma risk**

**Genome**: hg19

**Region**: 22 risk loci of MM

**Diseases**: multiple myeloma

**Cell lines:** L363 and MOLP8

**GWAS/QTL**: To identify putative causal variants, we first designed an MPRA14,19–21 to screen 1039 variants in high LD (r2 > 0.8) with MM lead variants for transcriptional activity

**Finding:** In L363, 142 variants were significant (FDR <5%), including 33 with strong effects (absolute log2 score >0.2). In MOLP8, 28 were significant, including 21with strong effects (Fig. 2e, f and Supplementary Data 1). The higher number of significant variants in L363, compared to MOLP8, cells was congruent with a higher transfection efficiency (54% for L363 versus 15% for MOLP8) and higher post-transfection viability (90% for L363 versus 65% for MOLP8). In total, 23 variants were significant in both screens, and eight of these showed concordant plasma cell cis-eQTLs, making them putative causal variants that were selected for follow-up (Table 1, Fig. 3, and Supplementary Figs. 4, 6). The other 15 had discordant or no plasma cell cis-eQTLs, either because of technical limitations (e.g., TERC was not in our eQTL data; the JARID2 and RUNX3 variants are rare), or because these alter gene expression in another cell state (e.g., TNFRSF13B is primarily expressed in switch-memory B-cells33 and had a cis-eQTL in total mature B-cells; Supplementary Table 4).

**Paper9: Global discovery of lupus genetic risk variant allelic enhancer activity**

**Genome**: hg19

**Region**: enhancer

**Diseases**: Systemic Lupus Erythematosus (SLE)

**Cell lines:** GM12878, Jurkat T cell lines

**GWAS/QTL**: We first collected all SLE-associated risk loci reaching genome-wide association sig- nificance (p < 5 × 10−8) published through March 2018 (Supple- mentary Data 1). Studies of all ancestral groups were included, and independent risk loci were defined as loci with lead (tag) variants at r2 < 0.2. For each of these 91 risk loci, we performed linkage disequilibrium (LD) expansion (r2 > 0.8) in each ancestry of the initial genetic association(s), to include all possible disease- relevant variants (Supplementary Data 2). In total, this procedure identified 3073 genetic variants. All established alleles of these variants were included, with 36 variants having three or more alleles. We also included 20 additional genetic variants from a previously published study19 as positive and negative controls to assess the library’s performance (Supplementary Data 3).

**Finding:** Transfection into the Epstein-Barr virus-transformed B cell line GM12878 reveals 482 variants with enhancer activity, with 51 variants showing genotype-dependent (allelic) enhancer activity at 27 risk loci. Comparison of MPRA results in GM12878 and Jurkat T cell lines highlights shared and unique allelic transcriptional regulatory mechanisms at SLE risk loci. In-depth analysis of allelic transcription factor (TF) binding at and around allelic variants identifies one class of TFs whose DNA-binding motif tends to be directly altered by the risk variant and a second class of TFs that bind allelically without direct alteration of their motif by the variant. Collectively, our approach provides a blueprint for the discovery of allelic gene regulation at risk loci for any disease and offers insight into the transcriptional regulatory mechanisms underlying SLE.

**Paper10: Prioritization of autoimmune disease-associated genetic variants that perturb regulatory element activity in T cells**

**Genome**: hg19

**Region**: open chromatin

**Diseases**: five autoimmune diseases

**Cell lines:** Jurkat T cells,

**GWAS/QTL**: We collected 578 GWAS index variants (representing 531 dis- tinct GWAS loci) and variants in tight LD (r2>0.8) from the above-cited studies, totaling 18,312 variants, and designed MPRA libraries by centering each variant within 200 bp of its genomic con- text to test for allele-specific effects on reporter expression

**Finding:** Here, we applied massively parallel reporter assays (MPRAs) and accessible chromatin in T cells to prioritize ~18,000 variants associated with five autoimmune diseases, including type 1 dia- betes (T1D), inflammatory bowel disease (IBD) (including ulcer- ative colitis (UC) and Crohn’s disease (CD)), rheumatoid arthritis (RA), psoriasis and multiple sclerosis (MS). Through integrating these methods, we found 60 likely causal variants that enriched up to 57.8-fold for causal variants according to fine-mapping.

We found 7,095 elements that had higher reporter expression than expected from their prevalence in plasmid libraries for at least one variant allele (termed putative cis-regulatory elements (pCREs); We found 313 variants that had statistically significant differences in expression between the reference and alternate alleles, which we term expression-modulating variants (emVars) (Fig. 1b and Supplementary Table 3).

**Paper 11: Massively parallel reporter assays and variant scoring identified functional variants and target genes for melanoma loci and highlighted cell-type specificity**

**Genome**: hg19

**Region**: 54 GWAS risk loci

**Diseases**: melanoma

**Cell lines:** malignant melanoma and normal melanocyte cells

**GWAS/QTL**: 54 GWAS risk locis

**Finding:** Of 1,992 risk-associated variants tested in MPRAs, we identified 285 from 42 loci (78% of the known loci) displaying significant allelic transcriptional activities in either cell type (FDR < 1%). We first focused on the variants displaying allelic tran- scriptional activity in each cell type, identifying 134 (7% of tested variants) in UACC903 melanoma (Figure 1C; Table S6) and 208 (10% of tested variants) in C283T mela- nocyte cell lines (Figure 1D; Table S7) that pass an FDR < 0.01 cutoff (two-sided Wald test with robust sandwich type variance estimate; multiple testing correc- tion by Benjamini and Hochberg27 method; material and methods). We defined these 285 unique variants (FDR < 0.01 in either cell line; 14% of tested variants) as ‘‘MPRA-significant variants.’’ 78% of the melanoma GWAS loci (42 of 54 loci) displayed at least one MPRA-sig- nificant variant. For 83% of these loci (35 of 42 loci), MPRA-significant variants were identified from both cell types, while the rest were from only one cell type (three loci in melanoma and four loci in melanocyte). For eight loci, a single MPRA-significant variant was identified, while 2–36 MPRA-significant variants were identified for 34 loci.

**The cis-regulatory effects of modern human-specific variants**

**Genome**: hg38

**Region**: noncoding

**Diseases**: modern and archaic humans

**Cell lines:** embryonic stem cells, neural progenitor cells, and bone osteoblasts

**GWAS/QTL**:

**Finding:** We used the library of 14,042 pairs of archaic and modern human sequences, together with posi- tive and negative control sequences, to infect each cell type. As positive controls for ESCs and NPCs, we added a set of 199 sequences with known regulatory capacity from previous MPRAs (Supplementary file 1d). To our knowledge, there have not been any MPRAs conducted in osteo- blasts, so we searched the literature for putative regulatory regions in osteoblasts and other bone cell types and used these as putative positive controls (Supplementary file 1d, see ’Materials and methods’). As negative controls, in all cell types, we randomly chose 100 sequences from the library and scrambled the order of their bases, creating a set of GC content matching sequences that had not been previously established to drive expression Here, we used a massively parallel reporter assay in embryonic stem cells, neural progenitor cells, and bone osteoblasts to investigate the regulatory effects of the 14,042 single-nucleotide modern human-specific variants. Overall, 1791 (13%) of sequences containing these variants showed active regulatory activity, and 407 (23%) of these drove differential expression between human groups

**Paper12: Massively parallel reporter assays of melanoma risk variants identify MX2 as a gene promoting melanoma**

**Genome**: hg19

**Region**:

**Diseases**: melanoma

**Cell lines:** Combined UACC903 & HEK293FT

**GWAS/QTL**: GWAS

**Finding:** Genome-wide association studies (GWAS) have identified ~20 melanoma susceptibility loci, most of which are not functionally characterized. Here we report an approach integrating massively-parallel reporter assays (MPRA) with cell-type-specific epigenome and expression quantitative trait loci (eQTL) to identify susceptibility genes/variants from multiple GWAS loci. From 832 high-LD variants, we identify 39 candidate functional variants from 14 loci displaying allelic transcriptional activity, a subset of which corroborates four colocalizing melanocyte cis-eQTL genes.

**Paper13: A screen of 1049 schizophrenia and 30 Alzheimer's-associated variants for regulatory potential**

**Genome**: hg19

**Region**:

**Diseases**: AD and schizophrenia

**Cell lines:** K562 chronic myelogenous leukemia lymphoblasts and SK-SY5Y human neuroblastoma cells.

**GWAS/QTL**:

**Finding:** We used a massively parallel reporter assay to screen, 1,049 SZ and 30 AD variants in 64 and nine loci, respec- tively for allele differences in driving reporter gene expression. A library of synthetic oligonucleotides assaying each allele five times was transfected into K562 chronic myelogenous leukemia lymphoblasts and SK-SY5Y human neuroblastoma cells. One hundred forty eight variants showed allelic differences in K562 and 53 in SK-SY5Y cells, on average 2.6 variants per locus. Nine showed significant differences in both lines, a modest overlap reflecting different regulatory landscapes of these lines that also differ significantly in chromatin marks. Eight of nine were in the same direction.

**Paper14: Systematic Functional Dissection of Common Genetic Variation Affecting Red Blood Cell Traits**

**Genome**: hg19

**Region**:

**Diseases**:

**Cell lines:** erythroid cell line K562

**GWAS/QTL**: We selected 2,756 SNPs or small indels that were in high LD (R2 > 0.8) with 75 previously reported GWAS hits to include in a high- throughput screen (Table S1). These variants were identified in the most comprehensive analysis to date that measured the ef- fects of genetic variation on RBC traits, comprising over 135,000 cases from over 30 individual studies (van der Harst et al., 2012). Positive control variants that disrupt the binding site of the erythroid TF GATA1 in an erythroid enhancer element, resulting in vivo in severe human erythroid disorders, were also included

**Finding:** We employ a massively parallel reporter assay (MPRA) to simulta- neously screen 2,756 variants in strong linkage disequilibrium with 75 sentinel variants associated with red blood cell traits. We show that this assay iden- tifies elements with endogenous erythroid regulatory activity. Across 23 sentinel variants, we conservatively identified 32 MPRA functional variants (MFVs).

**High-throughput identification of human SNPs affecting regulatory element activity**

**Genome**: hg19

**Region**:

**Diseases**:

**Cell lines:** K562 and HepG2

**GWAS/QTL:** These libraries enabled us to test promoter/enhancer activity of both alleles of 5,919,293 SNPs, which include 4,569,323 (57%) of the ~8 million known common SNPs (minor allele frequency (MAF) > 5%) worldwide1.

**Finding:** To systematically annotate SNPs we combined the complete SuRE datasets from the four genomes for each transfected cell line. The sequencing data of the SuRE libraries then enabled us to group, for each SNP, the overlapping genomic DNA fragments by the two alleles (Fig. 1c,d). This allowed us to identify SNPs for which fragments carrying one allele produced significantly dif- ferent SuRE signals compared to those carrying the other allele. Because all of these fragments differ in their start and end coor- dinate, the activity of each allele is tested in a multitude of local sequence contexts, providing not only statistical power but also biological robustness. For each SNP we calculated a P value and we used a random permutation strategy to estimate false discovery rates (FDR) (Supplementary Fig. 1e,f). We also required that the strongest allele showed an average SuRE signal of at least fourfold over background. We refer to the resulting SNPs at FDR < 5% as reporter assay QTLs (raQTLs).

This analysis yielded a total of 19,237 raQTLs in K562 cells and 14,183 in HepG2 cells (Fig. 1e). The average allelic fold change of these SNPs was 4.0-fold (K562) and 7.8-fold (HepG2) (Supplementary Fig. 1g,h). In 72% of cases the SuRE effect could be assigned to a single SNP; when SNPs were spaced less than ~200bp apart, their effects could typically not be resolved (Supplementary Fig. 1i).

**Availability:** lists of raQTLs and a table with SuRE data for all 5.9 million SNPs are available from the Open Science Framework (https://osf.io/w5bzq/wiki/home/?view).

**Paper15: Direct Identification of Hundreds of Expression- Modulating Variants using a Multiplexed Reporter Assay**

**Genome**: hg19

**Region**:

**Diseases**:

**Cell lines:**  lymphoblastoid cell lines

**Finding:** Evaluation of 32,373 variants associated with eQTLs in lymphoblastoid cell lines. 842 variants showed differential gene expression between alleles

**Human 5′ UTR design and variant effect prediction from a massively parallel translation assay**

**Genome**: hg19

**Region**: 5UTR

**Diseases**:

**Cell lines:** HEK293T

**GWAS/QTL**:

**Finding:** The ability to predict the impact of cis-regulatory sequences on gene expression would facilitate discovery in fundamental and applied biology. Here we combine polysome profiling of a library of 280,000 randomized 5′ untranslated regions (UTRs) with deep learning to build a predictive model that relates human 5′ UTR sequence to translation. We test 35,212 truncated human 5′ UTRs and 3,577 naturally occurring variants and show that the model predicts ribosome loading of these sequences. Finally, we provide evidence of 45 single-nucleotide variants (SNVs) associated with human diseases that substantially change ribosome loading and thus may represent a molecular basis for disease.

**Paper16: Direct Identification of Hundreds of Expression- Modulating Variants using a Multiplexed Reporter Assay**

**Genome**: hg19

**Region**:

**Diseases**:

**Cell lines:**  lymphoblastoid cell lines

**Finding:** Evaluation of 32,373 variants associated with eQTLs in lymphoblastoid cell lines. 842 variants showed differential gene expression between alleles

**Paper17: Multi-level functional genomics reveals molecular and cellular oncogenicity of patient-based 30 untranslated region mutations**

**Genome**: hg38

**Region**: 3UTR

**Diseases**: cancer

**Cell lines:**  PC3 cell

**Finding:** 30 untranslated region (30 UTR) somatic mutations represent a largely unexplored avenue of alternative onco- genic gene dysregulation. To determine the significance of 30 UTR mutations in disease, we identify 30 UTR so- matic variants across 185 advanced prostate tumors, discovering 14,497 single-nucleotide mutations enriched in oncogenic pathways and 30 UTR regulatory elements. By developing two complementary massively parallel reporter assays, we measure how thousands of patient-based mutations affect mRNA translation and stability and identify hundreds of functional variants that allow us to define determinants of mutation significance. We demonstrate the clinical relevance of these mutations, observing that CRISPR-Cas9 endogenous editing of distinct variants increases cellular stress resistance and that patients harboring oncogenic 30 UTR mutations have a particularly poor prognosis. This work represents an expansive view of the extent to which disease-rele- vant 30 UTR mutations affect mRNA stability, translation, and cancer progression, uncovering principles of reg- ulatory functionality and potential therapeutic targets in previously unexplored regulatory regions.

Patient-based 30 UTR mutations functionally affect gene-specific translation efficiency To functionally assess how patient mutations change 30 UTR- mediated aspects of post-transcriptional gene regulation, we first built a polysome profiling-based MPRA able to simulta- neously measure how thousands of mutations change transla- tion efficiency (TE) (Figure 2A). Two hundred base pair se- quences of the 30 UTR around each of 6,892 mutations from recurrently mutated genes were inserted downstream of the luciferase CDS in a modified pGL4 plasmid backbone, pLuc2CP-noARE (Table S2). Sequencing of this plasmid library confirmed adequate and even representation of 30 UTR inserts (Figure S2A). **We discover 180 30 UTR point mutations that significantly change TE, with a more than 1,000-fold dynamic range (Fig- ure 2B).**

Patient-based 30 UTR mutations significantly alter oncogenic mRNA stability  
In addition to TE, the 30 UTR is an important mediator of mRNA stability. To determine how mCRPC patient mutations affect mRNA degradation, we designed and implemented a second, complementary MPRA using an RNA sequencing (RNA-seq) time course of in vitro transcribed (IVT) mRNA (Figure 3A). We used the integrated T7 promoter to in vitro transcribe our plasmid library into a fully capped and poly-A tailed mRNA library. We transfected PC3 cells with this mRNA library, then collected li- brary mRNA from the cells over a 24 h time course in 6 biological replicates. **Overall, we find 150 patient-based 30 UTR mutations that signif- icantly change mRNA stability, many of which are in known onco- genes (Figure 3B).**

**Paper18: Systematic investigation of allelic regulatory activity of schizophrenia-associated common variants**

**Genome**: hg19

**Region**:

**Diseases**: Schizophrenia

**Cell lines:**  HEK293s / HNPS

**Finding:** Genome-wide association studies (GWASs) have successfully identified 145 genomic regions that contribute to schizophrenia risk, but linkage disequilibrium makes it challenging to discern causal variants. We per- formed a massively parallel reporter assay (MPRA) on 5,173 fine-mapped schizophrenia GWAS variants in primary human neural progenitors and identified 439 variants with allelic regulatory effects (MPRA-positive variants). Transcription factor binding had modest predictive power, while fine-map posterior probability, enhancer overlap, and evolutionary conservation failed to predict MPRA-positive variants. Furthermore, 64% of MPRA-positive variants did not exhibit expressive quantitative trait loci signature, suggesting that MPRA could identify yet unexplored variants with regulatory potentials. To predict the combinatorial effect of MPRA-positive variants on gene regulation, we propose an accessibility-by-contact model that combines MPRA-measured allelic activity with neuronal chromatin architecture.

As a result, we identified 439 MPRA-positive variants that show allelic regulatory activity at a false discovery rate (FDR) threshold of 0.1 (Figures 1C and 1D; Table S1). We found that 102 out of 143 GWS loci contained at least one MPRA-positive variant (Figure 1D). Out of 102 GWS loci that harbor regulatory variants, index variants (variants with the strongest GWAS association statistics at given loci) of 12 loci showed regulatory activity (Figure 1E), suggesting that the most significant GWAS association cannot accurately predict functional variants.

We defined MPRA-positive variants as variants that show statistical RNA count difference between reference and alternative allele at FDR <0.1, while defining MPRA-negative variants as variants with no significant allelic regulatory activity at nominal p > 0.1.

**Paper:**

**Genome**: hg19

**Region**:

**Diseases**:

**Cell lines:**

**Finding:**