Bioinformatics, 34(15), 2018, 2682–2683
doi: 10.1093/bioinformatics/bty150
Advance Access Publication Date: 21 March 2018
Applications Note

OXFORD

Gene expression

Design tools for MPRA experiments

Andrew R. Ghazi[1], Edward S. Chen[2], David M. Henke[2], Namrata Madan[3], Leonard C. Edelstein[3] and Chad A. Shaw[2,*]

[1]Department of Quantitative and Computational Biosciences and [2]Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX 77030, USA and [3]Cardeza Foundation for Hematologic Research and Department of Medicine, Thomas Jefferson University, Philadelphia, PA 19107, USA

*To whom correspondence should be addressed.

Associate Editor: Janet Kelso

Received on April 10, 2017; revised on February 28, 2018; editorial decision on March 7, 2018; accepted on March 19, 2018

Abstract

Motivation: Genetic reporter assays are a convenient, relatively inexpensive method for studying the regulation of gene expression. Massively Parallel Reporter Assays (MPRA) are high-throughput functionalization assays that interrogate the transcriptional activity of many genetic variants at once using a library of synthetic barcoded constructs. Despite growing interest in this area, there are few computational tools to design and execute MPRA studies.

Results: We designed an online web-tool and R package that allows for interactive MPRA experimental design encompassing both power analysis and design of constructs. Our tool is tuned using data from real MPRA studies. Users can adjust experimental parameters to examine the predicted effect on assay power as well as upload VCFs for automated construct sequence generation.

Availability and implementation: The MPRA Design Tools web application is available here: https://andrewghazi.shinyapps.io/designmpra/, https://github.com/andrewGhazi/designMPRA and https://github.com/andrewGhazi/mpradesigntools.

Contact: cashaw@bcm.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Genome wide expression quantitative trait loci (eQTL) studies have identified large numbers of genetic variants that show statistically significant associations with expression of many genes. Despite growing availability and research interest, associations revealed by eQTLs give incomplete insight into the functional mechanisms of these variants because eQTL peaks represent association signals and not causative mechanisms that drive differential expression. To address this gap between correlation and causation, an experimental approach of Massively Parallel Reporter Assays (MPRA) (Melnikov *et al.*, 2012) was developed. This method can rapidly assess the transcriptional activity of many variants at once. Briefly, in MPRA a library of barcoded oligonucleotides containing reference and alternate alleles of variants of interest along with surrounding genomic sequence are cloned into reporter gene plasmids and transfected into cells. The cells use the barcoded sequences as DNA templates for transcription and sequencing is used to count the barcodes in the plasmid library and RNA output from the cells. Transcription preserves the barcodes. Thus, the log ratio of mRNA counts to DNA counts of a single barcode gives a measurement of the 'transcriptional activity' of the barcode's corresponding variant sequence:

$$\text{Activity} = \log\left(\frac{\text{mRNA count}}{\text{DNA count}}\right)$$

$$\text{Transcriptional Shift} = \text{Act}_{\text{mut}} - \text{Act}_{\text{ref}}.$$

Comparing the mean activity difference between barcode replicates of the alternate and reference alleles gives insight into the transcriptional shift of a variant (see Supplementary Data S1). MPRA Design Tools is implemented in R and Shiny, a web application framework for R.

Many factors can reduce the sensitivity of MPRAs. Variants prioritized by GWAS association signal may represent functional classes that cannot operate in an episomal setting such as MPRA; for instance variants that only function in certain cell types or conditions may be undetectable by MPRA. Moreover, variants tested by MPRA may be nonfunctional yet emerge as testable candidates from GWAS due to linkage disequilbrium. In addition to these and potentially other factors, noise

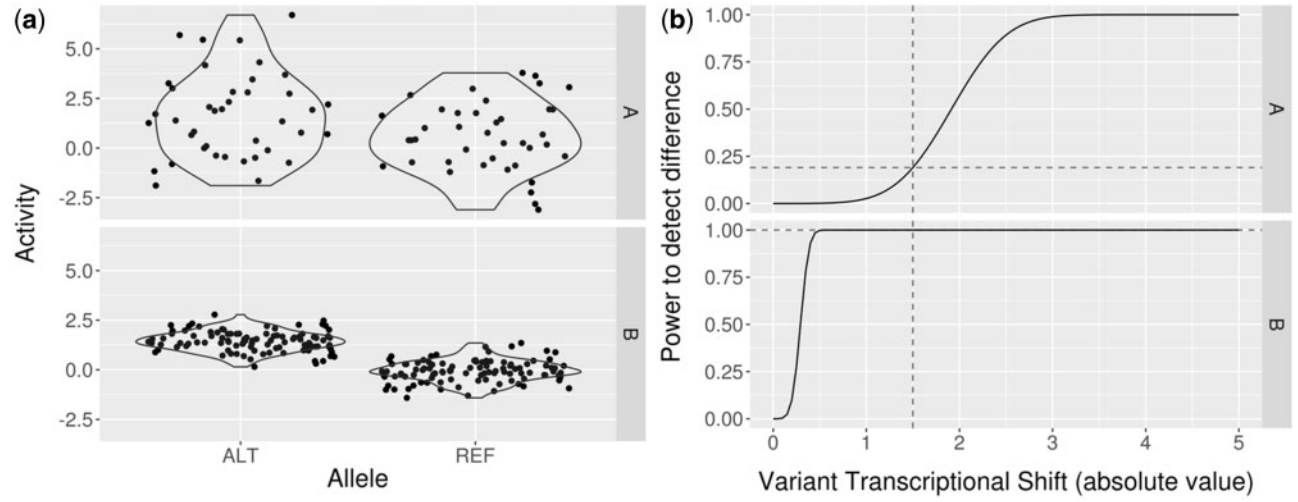

Fig. 1. (**a**) Two synthetic example MPRA outputs for hypothetical variants (see Supplementary Data S2) in scenarios A and B with transcriptional shifts of 1.5. Variant A uses 40 barcodes per allele and exhibits a standard deviation of 2 while B uses 100 barcodes per allele and exhibits a standard deviation of .5. Variances are based on observed data from Ulirsch *et al.* (2016) and Tewhey *et al.* (2016). (**b**) Line plots showing power to detect changes using the given multiplicities of barcodes and standard deviations in a MPRA assay with 1000 variants. In scenario A the statistical power to detect the shift is less than 20% while in scenario B the power is effectively 1. This example shows the necessity of selecting design parameters that maximize experimental efficiency

within the assay reduces the statistical power to detect variant function because the mRNA output per DNA input can be highly variable. Our analysis of published MPRA studies (Tewhey *et al.*, 2016; Ulirsch *et al.*, 2016) has shown that the standard deviation of activity across barcoded replicates of a given oligonucleotide is on average around .95, corresponding to a 2.6-fold difference in mRNA per input DNA molecule (see Supplementary Data S2). This level of noise suggests the assay cannot sensitively detect functional variants with small effect sizes. This can be addressed in part by increasing the number of barcodes per construct as well as through repeated transfections.

Here we show MPRA Design Tools for interactive design of MPRA studies. Users can adjust parameters such as barcodes per allele and activity variance to examine the estimated effect on statistical power, as shown by a synthetic demonstrative example in Figure 1. After selecting parameters, a tab allows users to upload VCFs of their variants to obtain MPRA construct sequences based on the hg38 genome. A companion R package provides more customizable sequence generation features.

Our tool differs from existing software such as MPRAnator (Georgakopoulos-Soares *et al.*, 2016) in both ease-of-use and interactivity. The MPRA Design Tools guides the user to choose experimental parameters that best fit goals while the sequence generation can be done either with VCF upload or using the R package. Our tool acquires genomic context from the hg38 reference genome, rather than requiring input by the user.

2 MPRA statistical power

Variant activity measurements are noisy; therefore, a set of multiple barcodes are used for each allele. These activity measurements are approximately normally distributed (see Supplementary Data S3), so our tool uses this assumption and published data (Tewhey *et al.*, 2016; Ulirsch *et al.*, 2016; see Supplementary Data S2 and S3) to estimate statistical power of detecting true shifts at a range of effect sizes (see 'Power' tab; see Supplementary Data S4 for modelling details). Changes to input parameters automatically update the plot output. This interface allows users to design experiments that optimize statistical power.

3 MPRA sequence generation

MPRA experiments require thousands of uniquely barcoded sequences that need to meet specific design parameters. The barcodes must be unique, must not generate restriction sites that would cause the constructs to be degraded in the plasmid library, and must be otherwise transcriptionally inert (see Supplementary Data S5). Additional parameters such as sequence context range or restriction enzymes may be adjusted.

We added functionality to allow for automatic generation of MPRA sequences based on variants input by the user. After inputting the number of barcodes per allele, the range of sequence context and other design parameters, users provide a VCF containing variants to receive a tab-separated file containing the MPRA sequences for their experiment.

4 Conclusions

MPRA Design Tools allows users to rapidly and interactively design MPRA experiments. The tool is available by web and R source.

Funding

This study was supported by United States National Institutes of Health Grant R01HL128234.

Conflict of Interest: none declared.

References

Georgakopoulos-Soares,I. *et al.* (2016) MPRAnator: a web-based tool for the design of Massively Parallel Reporter Assay experiments. *Bioinformatics (Oxford, England)*, **33**, 1–2.

Melnikov,A. *et al.* (2012) Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. *Nat. Biotechnol.*, **30**, 271–277.

Tewhey,R. *et al.* (2016) Direct identification of hundreds of expression-modulating variants using a multiplexed reporter assay. *Cell*, **165**, 1519–1529.

Ulirsch,J.C. *et al.* (2016) Systematic functional dissection of common genetic variation affecting red blood cell traits. *Cell*, **165**, 1530–1545.