A novel computational and experimental approach for allele-specific expression analysis in high-throughput reporter assays

C. Kalita, C. Brown, X. Wen, R. Pique-Regi, F. Luca

The majority of the human genome is composed of non-coding regions containing regulatory elements such as enhancers, which are crucial for controlling gene expression. Many variants associated with complex traits are in these regions, and may disrupt gene regulatory sequences. Consequently, it is important to not only identify true enhancers but also to test if a variant within an enhancer affects gene regulation. Recently, allele-specific analysis in high-throughput reporter assays, such as massively parallel reporter assays (MPRA) and STARR-seq (self-transcribing active regulatory region sequencing), have been used to functionally validate non-coding variants. However, we are still missing high-quality and robust experimental protocols and data analysis tools for these datasets.

Computationally, current data-analysis methods to measure allele specific expression (ASE) are limited in their ability to account for the uncertainty on the original plasmid proportions, over-dispersion, and sequencing errors. Experimentally, methods derived from the STARR-seq protocol do not have a barcoding strategy to remove PCR duplicates. We have further developed our method QuASAR (quantitative allele-specific analysis of reads), to analyze ASE in barcoded read count data. Using a beta-binomial distribution, QuASAR-MPRA better models the variability present in the allelic imbalance of these synthetic reporters and results in a test that is statistically well calibrated under the null. Additionally, the provided allelic skew estimate and its standard error simplifies meta-analysis of replicate experiments.

To develop QuASAR-MPRA we re-analyzed a recent MPRA study and identified 602 SNPs with allele-specific signal (FDR 10%). Additionally we validated computational annotations of regulatory variants, for subsequent experimental validation. To this end we modified the standard STARR-seq protocol to include the use of a UMI (unique molecular identifier) and used it to test 50,609 computationally and experimentally predicted regulatory variants. We confirmed the regulatory function of 2234 SNPs (FDR 10%). Our study shows that by having the appropriate data analysis tools and experimental approach, we can greatly improve the power to detect allelic effects in high throughput reporter assays. Furthermore, we can accelerate the discovery of causal variants by iteratively improving our computational models that prioritize genetic variants for reporter assay validation.