**MPRA design tool**

An R script for designing probes for massively parallel reporter assays (MPRA), investigating the influence of SNP genotype on enhancer/promoter activity.

The scripts accepts lists of genetic variants, designs probes (REF / ALT alleles and Fwd/ Rev orientations) and outputs a text file of sequences that can be sent directly for oligo synthesis.

MPRA simultaneously assess the enhancer / promoter activity of hundreds or thousands of DNA sequences. MPRA can be used to assess the influence small genetic polymorphisms (SNPs / indels) on enhancer activity for genetic variants linked to trait, for example from a GWAS study. This script accepts a list of SNP/Indels and designs probes to be used in an MRPA experiment such as (<https://www.nature.com/articles/s41596-020-0333-5>). Users can set their own probes length, adapter sequence and other features such as removing/including indels.

A feature of the script is that it will refine user provided risk loci by reference to linkage scores from LDlink (<https://ldlink.nci.nih.gov>). Users must register here for an API token before use, see instructions for further information. Users provide a summary SNP file containing variants tagging regions to be investigated. This is used to extract variants within a user defined genomic space. Within each region loci are recursively split based on an optional R2 threshold into independent loci. Independent loci are then filtered using a P values threshold based on the lead variant in that loci.

The script is designed to work with the output from META (<https://mathgen.stats.ox.ac.uk/genetics_software/meta/meta.html>) a frequently used tool for GWAS meta-analysis.

Functions performed by the script ;

**Part 1** : extract variants within a defined genetic interval around sentinel/leads SNPs from which probe sequences will be created. Sentinel/leads SNPs are defined in **--summary\_SNP\_file** and full SNP list specified by the **--all\_SNP\_file.**

Look up SNPs against a reference database file to ensure mappings and alleles are correct.

Determine reference and alternate alleles.

Optionally, remove SNPs that are known sequencing artifacts.

**Part 2** : split SNPs into linkage based loci by extracting R2 information from NIH LDlink <https://ldlink.nci.nih.gov/?tab=apiaccess>.

Apply a *P*-value filter based on the lead SNP in each loci.

Optionally add proxies that are not in the original SNP list (which may have been filtered eg due to info score).

**Part 3** : Extract the sequences for the REF and ALT alleles.

Add adapters and then optionally alter or filter sequences that have homopolymers.

Generate control sequences.

Perform some final checks (remap sequences to verify consistency).

**Part 4** : Optionally generate figures of the SNP filtering