**Instructions**

**What it does**

The script has three parts;

**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.

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 the 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

**Before running**

The script will connect to the NIH website to extract linkage information. You will need to go to this link (<https://ldlink.nci.nih.gov/?tab=apiaccess>) and register for an API token. In the script options please enter you token before running, using **-T**, **--API\_token**.

**Quick start**

When running the script ensure you have at least 40Gb of RAM as the script reads into memory both genome and reference SNP files.

To execute in its most simple form

> Rscript /path\_to\_script/design\_library\_v0.2.R \

-i / path\_to/summary\_SNPs.txt \

-a /path\_to/all\_SNPs.txt \

-d /path\_to/SNPdb.vcf.gz \

-g /path\_to/reference\_genome.fa.gz \

-o /path\_to/outdir \

-T {API\_token}

**Libraries**

The script has the following dependencies, they should install automatically on first run, if they don’t please install manually.

From CRAN- stringr, optparse, LDlinkR, data.table, dplyr, remotes, BiocManager, taRifx, RColorBrewer.

From BiocManager –rtracklayer, VariantAnnotation, Biostrings.

From Github– sarlacc, SparseSummarizedExperiment.

**Run environment information**

R version 4.0.3 (2020-10-10)

Platform: x86\_64-conda-linux-gnu (64-bit)

Running under: CentOS Linux 8 (Core)

Matrix products: default

BLAS/LAPACK: /opt/software/applications/anaconda/3/envs/r-essentials4.0/lib/libopenblasp-r0.3.12.so

attached base packages:

parallel, stats4,stats, graphics, grDevices, utils, datasets,methods, base

other attached packages:

SparseSummarizedExperiment\_0.0.0.9021, sarlacc\_1.0.0, VariantAnnotation\_1.36.0, Rsamtools\_2.6.0, Biostrings\_2.58.0, XVector\_0.30.0, SummarizedExperiment\_1.20.0 Biobase\_2.50.0, MatrixGenerics\_1.2.1, matrixStats\_0.58.0, rtracklayer\_1.50.0 GenomicRanges\_1.42.0, GenomeInfoDb\_1.26.7, IRanges\_2.24.1, S4Vectors\_0.28.1 BiocGenerics\_0.36.1, RColorBrewer\_1.1-2, taRifx\_1.0.6.2, BiocManager\_1.30.15 remotes\_2.3.0, dplyr\_1.0.5, data.table\_1.14.0, LDlinkR\_1.1.2, optparse\_1.7.1, stringr\_1.4.0

loaded via a namespace (and not attached):

httr\_1.4.2, bit64\_4.0.5, assertthat\_0.2.1, askpass\_1.1, BiocFileCache\_1.14.0, latticeExtra\_0.6-29, blob\_1.2.1, BSgenome\_1.58.0, GenomeInfoDbData\_1.2.4, progress\_1.2.2, pillar\_1.6.0, RSQLite\_2.2.3, lattice\_0.20-41, glue\_1.4.2, Matrix\_1.3-2, plyr\_1.8.6, XML\_3.99-0.6, pkgconfig\_2.0.3, ShortRead\_1.48.0, biomaRt\_2.46.3, zlibbioc\_1.36.0, purrr\_0.3.4, jpeg\_0.1-8.1, getopt\_1.20.3, BiocParallel\_1.24.1, tibble\_3.1.1, openssl\_1.4.3, generics\_0.1.0, ellipsis\_0.3.1, cachem\_1.0.4, GenomicFeatures\_1.42.3, magrittr\_2.0.1, crayon\_1.4.1, memoise\_2.0.0, fansi\_0.4.2, hwriter\_1.3.2, xml2\_1.3.2, tools\_4.0.3, prettyunits\_1.1.1, hms\_0.5.3, lifecycle\_1.0.0, DelayedArray\_0.16.3, AnnotationDbi\_1.52.0, compiler\_4.0.3, rlang\_0.4.10, grid\_4.0.3, RCurl\_1.98-1.3, rstudioapi\_0.13, rappdirs\_0.3.3, bitops\_1.0-7, curl\_4.3, DBI\_1.1.0, reshape2\_1.4.4, R6\_2.5.0, GenomicAlignments\_1.26.0 fastmap\_1.0.1, bit\_4.0.4, utf8\_1.1.4, stringi\_1.5.3, Rcpp\_1.0.7, png\_0.1-7, vctrs\_0.3.6, dbplyr\_2.0.0, tidyselect\_1.1.0

**INPUT FILES**

For both the summary\_SNP\_file and main\_SNP\_file;

chr/CHR notation should **be consistent** in both files.

P\_value/P notation should be exponential format eg 2.3755e-64

**Required inputs**

1. **-i, --summary\_SNP\_file** : "/path\_to/summary\_SNPs.txt"

Textfile with header containing a list of lead SNPs used to define loci to design oligos within. Tab delimited (can be gzipped)

eg

CHR POS RSID P

1 22503282 rs2807367 3.27E-08

Must contain the columns “**CHR”**, “**POS”** and “**RSID”,** can contain others.

1. **-a, --all\_SNP\_file** : "/path\_to/meta\_results\_header.txt"

Text file with header containing the full list of genetic variants. May be the output from META. Tab delimited (can be gzipped)

Must contain the columns “**chr**”, “**pos**”, “**rsid**”, "**P\_value**",”**allele\_A**”,” **allele\_B**”**;**

optional headers;

**“RAF”** (relative allele freq)if present variants are filtered to remove those with a RAF below a value specified by the min\_RAF option.

“**P\_heterogeneity**” if present and option filter\_phet=T variants with significant heterogeneity in a region where the lead SNP doesn’t exhibit heterogeneity are filtered. (intended for studies from multiple populations)

Can contain others columns.

1. **-d, --snp\_db\_file** : ”/path\_to /00-common\_all.vcf.gz"

Reference data baseSNP file for the species and genomic build used for generating SNP lists/GWAS.

Must be vcf file format.

Human dbSNP reference file (Hg37) can be download from [https://ftp.ncbi.nih.gov/snp/organisms/](https://ftp.ncbi.nih.gov/snp/organisms/human_9606_b151_GRCh37p13/VCF/)

1. **-g, --genome\_file** : ”/path\_to /hg19.fa.gz"

**F**asta file of the relevant species and genome build, must be the same as the SNPdb file.

Must be fasta file format.

**Optional inputs**

1. **-n, --negative\_control\_file** : "/path\_to /H3K27me3.broadPeak.gz"

A bed file use to design control probes. Enriched regions are refined by the **‘--ChIP\_enrichment**’ option. Regions with an enrichment score below this value are removed. Retained regions are mapped to known variants in the snp\_db\_file. Overlapping variants are randomly selected and used for probe design.

Must be headerless bed file with the columns arranged as per;

**"chr","start","stop","rank","score","strand","enrichment","p-val","q-val"**

Eg <https://egg2.wustl.edu/roadmap/data/byFileType/peaks/consolidated/broadPeak/>

1. **-u, --unfiltered\_file\_path** : "/path\_to/unfiltered\_rsid\_{CHROM}.txt.gz"

Directory containing the per chr output from meta, without any P value filtering. If the --**add\_proxies** option is true, these files are used to prevent importing variants which were not in -- **all\_SNP\_file** due to P-value filtering but are linked to risk variants. Chromosome names are inferred from the SNPs in the --**summary\_SNP\_file**. These are then substituted for “CHROM” in the path to specify the relevant file path. Eg for variants on chromosome 1 the script will attempt to import unfiltered\_rsid\_1.txt.gz. The “CHROM” in the path must be present to allow substitution with chr names. The remainder of the path must match the relevant file name. File format same as --**all\_SNP\_file.**

1. **-b, --black\_list\_snp\_file** : "/path\_to/b151\_rs\_without\_GRCh38\_mapping.bcp"

Text file.

Must be headerless file with the columns arranged as per;

**"chr",”pos”,"rsid”**

Other columns can after these can exist. Any consistent separator can be used.

Some SNPs in build hg37 are artefacts and have been removed in hg38 dbSNP builds they are listed here # <https://gatk.broadinstitute.org/hc/en-us/articles/360035890951-Human-genome-reference-builds-GRCh38-or-hg38-b37-hg19>

**OPTIONS**

1. **-o, --outdir :** ”/path\_to/output\_dir”

Path to output dir where output files are saved.

1. **-s, --suffix :** string.

Suffix added to output file names.

1. **-T**, **--API\_token** : string

Token to access LDlink database see 'https://ldlink.nci.nih.gov/?tab=apiaccess' to obtain.

1. **-F, --Forward\_adapter** : character string, default="AGGACCGGATCAACT".

Sequence added to the 5' end of each oligo.

1. **-R, --Reverse\_adapter** : character string, default="CATTGCGTGAACCGA".

Sequence added to the 3' end of each oligo.

1. **-w,--loci\_window** : positive integer, default=250000.

Variants less than this number of base pairs from a lead SNP will be used for designing probes.

1. **-z, --insert\_size** : positive integer, default=199 eg 99bp/SNP/99bp.

Desired length of probes/cis-regulatory sequences containing a SNP; if indels are retained probe length will differ from selected value.

1. **-c, --n\_controls** : positive integer, default=100.

The number of control sequences to design. One control = 4 oligos; Fwd/Rev : Allele1/Allele 2.

1. **-e, --ChIP\_enrichment** : positive integer, default=4.

If providing a control region file only design control probes/CRSs in regions with an enrichment > than this value.

1. **--prop\_control** : numeric, default=0.5 , range 0-1.

If providing a control region file what proportion of controls sequences should be taken from this. If less than 1 remainder are randomly generated.

1. **-m, --minor\_allele\_frequency** : numeric, default=0.01, range 0-1.

Variants with a MAF/RAF lower than this value are removed.

1. **-P, --P\_value\_cuttoff** : numeric, default=5e-8, format must be scientific notation Xe-y.

Hard P-value cut-off. Variants with a P-value above this value are removed.

1. **-f, --P\_cuttoff\_factor** : numeric, default=0.7, range 0-1, higher the number the more variants filtered.

Used to generate locus specific P\_value thresholds. Calculated as -log(lead\_Pval,10)\*P\_cuttoff\_factor. Variants within a locus with a P value above this are removed. lead\_Pval = P value of the lead SNP within a defined loci.

1. **-S, --SNVs\_only** : logical, default=TRUE, values TRUE|FALSE.

If true retain only single nucleotide variants.

1. **-L, --max\_indel\_length** : positive integer, default=10.

Indels with an alt allele length > this value are removed; irrelevent if -S is TRUE.

1. **-U,--keep\_unmapped\_variants :** logical, default=TRUE, values TRUE|FALSE.

If TRUE user provided variants not present in reference SNP file (**snp\_db\_file**) are retained.

1. **--update\_alleles** : logical, default=FALSE, values TRUE|FALSE.

If TRUE user provided variants not matching the reference SNP file (**snp\_db\_file**) alleles are reverted to the ref db alleles.

1. **--output\_figures** : logical, default=TRUE, values TRUE|FALSE.

Generate figures showing filtered and retained variants.

1. **--split\_by\_r2** : logical, default=TRUE, values TRUE|FALSE.

If TRUE variants in a region surrounding a sentinel SNP will be split into separate loci based on linkage. The SNP with the lowest p-value in a new loci will be the new lead and used for p-value filtering.

1. **-r, --r2\_cuttoff** : numeric, default=0.3, range 0-1.

When splitting loci variants with an r2, to the lead SNP, below this value are split into separate loci.

1. **-M, --mutate\_homopolyers** : logical, default=TRUE, values TRUE|FALSE.

Probes with homopolymers (specified by **--max\_hp\_len**) will be mutated to remove homopolymers.

1. **--filter\_homopolymers** : , logical, default=TRUE, values TRUE|FALSE.

Probes with homopolymers (specified by **--max\_hp\_len**) will be filtered. If **--mutate\_homopolyers** is TRUE successfully mutated homopolymers are retained.

1. **-l, --min\_dist\_mut2snp** : positive integer, default=5.

Bases within a homoploymer < this number of base pairs from the reference or alternative allele are not mutated.

1. **-p, --populations** : character, default="EUR", values 'EUR|SAS|EAS|AMR|AFR'.

Populations used for r2 extraction. Multiple populations separate by a comma.

1. **-x, --add\_proxies** : logical, default=FALSE, values TRUE|FALSE.

Import variants linked to a lead SNP but absent from user provided variants. If TRUE necessary to specify **--unfiltered\_file\_path.**

1. **--proxy\_maf** : numeric, default=0.01, value 0-1

If **--add\_proxies** is TRUE, proxies with a MAF less than this value are filtered.

1. **--proxy\_r2** : numeric, default=0.7, range 0-1

Filter proxies with an r2, to loci lead SNP, less than this value.

1. **--filter\_phet** : logical, default=FALSE, values TRUE|FALSE.

Where lead SNPs has phet > value specified in **--P\_het\_cuttoff** option (ie not heterogenous) filter variants within that loci with phet < **P\_het\_cuttoff**. If FALSE no filtering is performed.

1. **--P\_het\_cuttoff** : numeric, default=0.01, value 0-1.

If **--filter\_phet** = TRUE. This value defines the cut off.

1. **-C, --cycle\_limit** : Positive integer, default=5.

Risk loci are split into loci containing only linked variants, this value sets the upper limit on the number of cycles used to split loci.

**OUTPUT FILES**

**MPRA\_library\_sequences.txt** - contains the final probe set names and sequences that can be submitted directly to an oligo synthesis company for manufacture.

**MPRA\_summary.txt** - details the number of oligos, SNPs, and loci in the final probe set.

**MPRA\_filtered\_SNPs.txt** - contains all the filtered SNPs and the filter which was applied to them.

**MPRA\_summary\_filtered.txt** – contains summary information about the numbers of SNPs filtered due to various constraints.

**MPRA\_summary\_stats.txt** – describes each loci after filtering, provides information on the number of SNPs pre and post filtering and the P-value threshold with was applied to each.

**MPRA\_library\_all\_variants.txt** – contains additional information on all test SNPs in the final probe set. Including, lead SNP, original probe sequence and loci P-value cut-off.

**MPRA\_library\_included\_variants\_and\_cotrols.txt** – contains additional information on all, test and control SNPs, in the final probe set. Including, lead SNP, original probe sequence and loci P-value cut-off.

**MPRA\_prefilter\_sequences.txt –** contains all variants extracted for probe design before filtering.

**#** in all output files unless otherwise specified RSID names are those updated based on the supplied SNP reference file.