|  |  |  |  |
| --- | --- | --- | --- |
| Input file | Scripts, run in this order, all in one /scripts/ folder | Output file | Description of output file |
| /1\_raw/v1.1\_genes\_TF\_homoeolog\_info.txt | step2\_triads.sh | /2\_triads/triads.txt | List of genes in 1:1:1 triads and their TF families |
| /2\_triads/triads.txt  /gff\_files/Triticum\_aestivum.IWGSC.51.gff3.gz | step2\_b\_genenames.sh  step2\_c\_triad\_coordinates.sh  step2\_d\_triadgene\_coordinates.sh | /2\_triads/Triticum\_aestivum.IWGSC.51.genesintriads2.gff3 | GFF Coordinates & annotation of genes in 1:1:1 triads |
| /He2019/all.GP08\_mm75\_het3\_publication01142019.vcf | sed\_chr\_step2\_f.sh | /He2019/He19\_no\_chr.vcf | VCF file with chromosome column modified to match the gff3 file |
| /He2019/He19\_no\_chr.vcf  /2\_triads/Triticum\_aestivum.IWGSC.51.genesintriads2.gff3 | bedtools\_intersect\_variants\_step2\_e.sh | /2\_variants/variants\_in\_genes\_in\_triads\_He\_2019.vcf | VCF file of SNPs intersecting genes in triads |
| /2\_variants/variants\_in\_genes\_in\_triads\_He\_2019.vcf | vep\_step3\_a.sh | /3\_vep/variant\_effect\_output\_all.txt  /3\_vep/variant\_effect\_output\_all.txt\_summary.html  /3\_vep/variant\_effect\_output\_all.txt\_warnings.txt | Ensembl variant effect predictor (VEP) output for SNPs in genes in triads |
| /3\_vep/variant\_effect\_output\_all.txt | grep\_step3\_b\_all.sh  grep\_coding\_variant\_step3\_c\_all.sh | /3\_filtered/variant\_effect\_coding\_variant\_all.txt | VEP output, only variant effects in the coding sequence of canonical transcripts |
| /He2019/He19\_no\_chr.vcf  /3\_filtered/variant\_effect\_coding\_variant\_all.txt | grep\_vcf\_coding\_variant\_step6\_a.sh  grep\_vcf\_coding\_variant\_step6\_a\_fields10.sh | /4\_analysis/He19F\_coding\_variant\_in\_triad\_genes.vcf  /4\_analysis/He19F\_coding\_variant\_in\_triad\_genes\_fields10\_header.vcf | VCF file of SNPs intersecting genes in triads AND affecting the coding sequence of canonical transcripts |
| /4\_analysis/He19F\_coding\_variant\_in\_triad\_genes.vcf | calculate\_allele\_frequencies\_step6\_a.R | allele\_frequency\_coding\_variant\_2022-01-21.csv | Allele frequencies of SNPs in the coding sequence of canonical transcripts |
| /2\_triads/triads.txt  variant\_effect\_coding\_variant\_all.txt  allele\_frequency\_coding\_variant\_2022-01-21.csv | coding\_variants\_preparation\_step6\_b.R | /4\_analysis/coding\_variant\_tf\_allele\_frequency\_table\_MAFall\_2022-01-10.csv | combined table with variant effects, allele frequencies, and TF families for SNPs in the coding sequence of canonical transcripts. Exclude splice region variants and missense variants without SIFT scores. |
| allsel.bed  extremeload.txt  choulet\_URGI\_tpm.tsv  /4\_analysis/He19F\_coding\_variant\_in\_triad\_genes\_fields10\_header.vcf  /4\_analysis/coding\_variant\_tf\_allele\_frequency\_table\_MAFall\_2022-01-21.csv | bedtools\_intersect\_sweep\_regions\_step6\_g.sh  calculate\_allele\_frequencies\_step6\_a\_2022.R  write\_sites\_to\_exclude.R  filter\_expressed\_genes\_step6\_f\_2022.R | He19\_coding\_variant\_not\_in\_sweep\_regions.vcf  /4\_analysis/coding\_variant\_tf\_allele\_frequency\_table\_MAFall\_expression\_2022-01-30.csv | Add columns:  Expressed: is this gene expressed in at least 1 tissue?  Sweep: is this gene in a selective sweep region?  Synonly: is the synonymous SNP only synonymous, with no other annotations? |
| /4\_analysis/coding\_variant\_tf\_allele\_frequency\_table\_MAFall\_expression\_2022-01-30.csv | coding\_variants\_allele\_frequencies\_step6\_d.R | /4\_analysis/coding\_variant\_tf\_allele\_frequency\_table\_MAF0.01\_2022-02-10.csv | Exclude variants with >25% missing values and filter for MAF>0.01 |
| /4\_analysis/coding\_variant\_tf\_allele\_frequency\_table\_MAF0.01\_2022-02-10.csv | functions\_graph\_chisquared.R  stacked\_plot\_by\_tf\_family\_step7\_a.R | /figures/stacked\_plot\_of\_big\_TF\_families\_MAF0.01\_2022-02-10.pdf | Figure 4B: Stacked bar graph of SNP effects by TF family.  Exclude genes which are not expressed, SNPs which are in selective sweep or introgression regions, and synonymous SNPs with other annotations.  Include TF families with >=10 triads and >=5 SNPs, and non-TFs for comparison. |

NB: R scripts were run through wrapper scripts, e.g. run\_calculate\_allele\_frequencies\_step6\_a\_2022.sh runs calculate\_allele\_frequencies\_step6\_a\_2022.R.

Links

GFF3 annotation file from Ensembl:

<ftp://ftp.ensemblgenomes.org/pub/plants/release-51/gff3/triticum_aestivum>

VCF file of filtered and imputed SNPs “1000 wheat exomes project”:

He, F., Pasam, R., Shi, F., Kant, S., Keeble-Gagnere, G., Kay, P., . . . Akhunov, E. (2019). Exome sequencing highlights the role of wild-relative introgression in shaping the adaptive landscape of the wheat genome. *Nature Genetics, 51*(5), 896-+. doi:10.1038/s41588-019-0382-2

<https://wheat-urgi.versailles.inra.fr/Seq-Repository/Variations>

Available online:

choulet\_URGI\_tpm.tsv

Remaining input files which should be included here:

v1.1\_genes\_TF\_homoeolog\_info.txt

extremeload.txt

allsel.bed