# Chapter 4 Allelic variation Analysis pipeline

## Phase 1: Shell scripts run on Bio4 server

Green = Key Files used in final pipeline

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Input file** | **Script**  /scripts | **Output folder** | **Output file** | **Description** | **# SNP Sites kept** | **# SNP**  **Sites failed** |
| Exome\_Capture\_All.vcf.gz  NAC\_target\_genes\_pts.bed | bedtools\_intersect\_target\_genes.sh | 2\_variants | variants\_NAC.vcf | Variants (both SNPs and Indels) overlapping 11 NAC target genes (genomic sequence).  705 sites in 286 individuals. (82 lines header). Takes 35-70 mins on server. | 705/705 |  |
| variants\_NAC.vcf | vcftools\_quality\_stats.sh | 3\_quality\_stats | variants\_NAC\_stats.frq  .lmiss  .imiss  .ldepth.mean  .idepth  .INFO  .qc.INFO  .singletons  .log | Multiple files of SNP-level and individual-level quality stats | 705/705 |  |
| variants\_NAC.vcf | vcftools\_quality\_filters.sh | 4\_filtered | NAC.DP.removed.sites | Read coverage average between 5 and 60, per-sample between 2 and 120 | 703/705 | 2/705 |
|  | vcftools\_quality\_filters.sh | 4\_filtered | NAC.lmiss.removed.sites | <20% missingness | 694/703 | 9/703  11/705 |
|  | vcftools\_quality\_filters.sh | 4\_filtered | NAC.qc.recode.vcf  NAC.biallelic.removed.sites | Biallelic sites (2 possible alleles) | 666/694 | 28/694  30/705 |
| NAC.qc.recode.vcf | vcftools\_quality\_filters.sh | 4\_filtered | NAC.qc.nohet.vcf | No heterozygous calls (optional filter) | 195/666  118/499 | 471/666  381/499 |
| NAC.qc.recode.vcf  NAC.qc.nohet.vcf | vcftools\_quality\_filters.sh | 4\_filtered | NAC.qc.filter.vcf  NAC.qc.nohet.filter.vcf  NAC.qc.filter.removed.sites | quality control  --exclude “QD <2 || FS >60.0 || MQRankSum<-12.5 || ReadPosRankSum<-8.0 || SOR >3.0 || MQ <40”. (Scott et al. 2021) | 499/666  118/195 | 167/666  77/195 |
| ().vcf  NAC.qc.filter.vcf | vcftools\_quality\_filters.sh | 4\_filtered | ().nosingletons.recode.vcf  NAC.qc.filter.nosingletons.recode.vcf | No singletons (found in only 1 variety) | 128/705  131/666  99/499  25/195  23/118 | 577/705  535/666  400/499  170/195  95/118 |
| NAC.qc.filter.nosingletons.recode.vcf | bedtools\_flank\_sequences.sh | 5\_flanking\_regions | NAC.qc.filter.nosingletons.coordinates.bed  ().locations.txt  ().leftflank.txt  ().rightflank.txt | 100bp flanking sequences from CS RefSeq v1.0 |  |  |
| NAC.qc.filter.nosingletons.recode.vcf | vcftools\_ld.sh | 6\_linkage | NAC.qc.filter.nosingletons.geno.ld | Pairwise linkage disequilibrium between SNPs. R2=1 means SNPs are always found together. | 99 |  |

## Phase 2: R scripts and other tools

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Input file** | **Script**  /2022-07-01 RAGT Exome Capture | **Output folder** | **Output file** | **Description** | **# SNP Sites kept** | **# SNP**  **Sites failed** |
| NAC.qc.filter.nosingletons.coordinates.bed  ().locations.txt  ().leftflank.txt  ().rightflank.txt  variants\_NAC\_stats.frq  161010\_Chinese\_Spring\_v1.0\_pseudomolecules\_parts\_to\_chr.xlsx | PolyMarker\_input\_prep\_2022-07-01.R | 2022-07-01 RAGT Exome Capture | Summary\_NAC.qc.filter.nosingletons.csv  Polymarker\_input\_().csv  VeP\_input\_().txt | Summary table including RefSeq v1.0 coordinates  Format for input into PolyMarker (KASP marker building tool)  Format for input into Ensembl VeP (Variant Effect Predictor) | 99 |  |
| Polymarker\_input\_().csv | PolyMarker (Ramirez-Gonzalez, 2015)  [PolyMarker](http://www.polymarker.info/) | 2022-07-01 RAGT Exome Capture | Polymarker\_output\_().csv | KASP markers to target SNPs, mapped to Chinese Spring RefSeq v1.0 | 58/99 | 41/99 |
| Polymarker\_output\_().csv | John ran an RAGT in-house quality control script | 2022-07-01 RAGT Exome Capture | marker\_annotations\_from\_John\_2022-07-14.csv  markers\_analysis\_polymarker\_NAC.nosingletons\_2022-07-11.xlsx | KASP markers categorised as excellent, good and poor | 26/58 | 32/58 |
| VeP\_input\_().txt | Ensembl VeP  (McLaren, 2010)  [Variant Effect Predictor - Triticum\_aestivum - Ensembl Genomes 54](http://plants.ensembl.org/Triticum_aestivum/Tools/VEP) | 2022-07-01 RAGT Exome Capture | VeP\_output\_().txt | Effect of variants on the gene’s RNA and protein sequence | 99 |  |
| Summary\_NAC.qc.filter.nosingletons.csv  NAC\_target\_genes.txt  VeP\_output\_().txt | VeP\_Table\_building\_RAGTExome\_2022-07-06.R | 2022-07-01 RAGT Exome Capture | variant\_summary\_full\_RAGT\_Capture.csv  variant\_summary\_brief\_RAGT\_Capture.csv | **TABLES WITH RAGT DATA ONLY**  summary\_full:  1 row per SNP\*Transcript,  full info  summary\_brief:  1 row per SNP,  brief info | 99 |  |
| summary\_SNP\_Transcript\_Variantsource\_2022-07-06.csv  (From #2 Public Datasets, found in /Haplotypes) | VeP\_Table\_building\_RAGTExome\_2022-07-06.R | 2022-07-01 RAGT Exome Capture | summary\_SNP\_Transcript\_full\_+\_RAGT\_Capture.csv  summary\_SNP\_Transcript\_intermediate\_+\_RAGT\_Capture.csv  summary\_SNP\_Transcript\_brief\_+\_RAGT\_Capture.csv  summary\_SNP\_Transcript\_Variantsource\_+\_RAGT\_Capture.csv | **TABLES WITH RAGT DATA PLUS ALL 5 PUBLISHED DATASETS**  summary\_full:  1 row per SNP\*Transcript,  full info  summary\_ intermediate:  1 row per SNP\*Transcript,  intermediate info  summary\_brief:  1 row per SNP,  brief info  summary\_Variantsource:  1 row per SNP\*Transcript\*Variant source,  full info | 398 |  |
| NAC.qc.filter.nosingletons.recode.vcf  variant\_summary\_full\_RAGT\_Capture\_2022-07-07.csv  markers\_analysis\_polymarker\_NAC.nosingletons\_2022-07-11.xlsx | haplotype\_run\_ragt\_exome\_2022-07-13  haplotype\_functions\_202-07-13.R | 2022-07-14 Haplotype keys  2022-07-13  Figures | haplotype\_key\_NAM-A1.csv  haplotype\_key\_imputed\_NAM-A1.csv  etc. | Haplotype tables for every target gene with  Only non-missing data  Missing data  + Figures | 99 |  |
| 2022-07-14 Haplotype keys  summary\_SNP\_Transcript\_full\_+\_RAGT\_Capture.csv  markers\_analysis\_polymarker\_NAC.nosingletons\_2022-07-11.xlsx | Manual curation | 2022-07-01 RAGT Exome Capture | Haplotype Keys with Het 2022-07-14.xlsx  Haplotype Keys Imputed 2022-07-14.xlsx  Haplotype slides.pptx | **BRING EVERYTHING TOGETHER**  Haplotype tables for every target gene with  Number of RAGT varieties  Heterozygous SNPs  SNP Effect  KASP marker quality  Pretty colours  Conclusions | 99 |  |