# ELGAN imputation documentation

###### Weifang Liu

###### Mar 2020

This file documents what have been done for the imputation of ELGAN GWAS data, including pre-imputation quality control, imputation, and post-imputation quality control.

## Pre-imputation Quality Control

## SNP-level Quality Control

There are 731,442 markers before SNP-level QC.

We removed SNPs which meet any of the following criteria: 1) call rate < 90%; or (2) MAF < 1%. Details on SNP-level filtering are listed in Table 1. We did not use deviation from Hardy-Weinberg Equilibrium as an exclusion criterion since this is a mixed population.

**Table 1. Summary of excluded SNPs**

|  |  |  |
| --- | --- | --- |
| **Criteria** | **# of SNPs** | **Percent (%)** |
| call rate < 90% | 13,376 | 1.83 |
| MAF < 1% | 17,893 | 2.45 |

We end up with 700,845 SNPs after SNP-level QC.

## Sample-level Quality Control:

There are 733 individuals before sample-level QC.

Among them, 4 individuals have sample-level missing rate > 10% using plink program v.1.90. Table 2 summarizes these samples.

**Table 2. Summary of excluded samples**

|  |  |  |  |
| --- | --- | --- | --- |
| **ID** | **#of missing SNPs** | **# of non-obligatory missing genotypes** | **% of missing SNPs** |
| 8803676149\_R06C02 | 141682 | 729763 | 19.41% |
| 8803713055\_R03C02 | 131632 | 729763 | 18.04% |
| 8803676149\_R05C02 | 105645 | 729763 | 14.48% |
| 8803713013\_R03C01 | 81484 | 731442 | 11.14% |

We are left with 729 samples after sample QC.

**Note:** There are potential sex or data quality issues with the data as suggested by the number of heterozygous genotypes and nonmissing nonmale Y chromosome genotypes, but since we are unclear what causes the issues, we will focus on autosomes for imputation. We did not further remove samples based on sex check. More details about the sex check results generated by plink can be found in the Appendix. I removed the four samples failed at sample QC post-imputation rather than pre-imputation, but this should have minimal effects on the results.

## Imputation

## Strand flipping

We first performed strand flipping according to our reference panel (TOPMed Freeze 5) to improve imputation accuracy. For non-ambiguous SNPs (that is, SNPs that are not A/T or C/G), the alleles of our cohort were flipped if they were chosen from different strand from the reference panel (for example, the alleles in our cohort are A/G, while they are T/C or C/T in the reference panel). We dropped the ambiguous SNPs and SNPs with zero positions.

## Phasing and imputation

We used the Michigan Imputation Server for phasing (eagle) and imputation (minimac4), using the TOPMed Freeze 5 as the reference panel. Genotype data was lifted over to hg38.

**Note:** the imputed vcfs from the Michigan Imputation Server do not contain typed only variants, which are not in the reference panel, but these are about only 1%-2% of the SNPs in the original genotyped data. There are a few duplicates in the SNPs with same genotypes. The results keep only one of them. An example of duplicates is shown below: same chromosome, same position, different rsIDs, same alleles.

chr position rsID ref allele alt allele

15 74720646 rs1799814 G T

15 74720646 VG15S12103 G T

## Post-imputation QC

## Number of well-imputed variants (R squared > 0.8)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **chr** | **maf>0.5% count (Rsq>0.8)** | **maf>1% count (Rsq>0.8)** | **maf>5% count (Rsq>0.8)** | **total count Rsq>0.8** | **total count without any filters** |
| **1** | 1178225 | 660360 | 189165 | 2672938 | 18682637 |
| **2** | 929956 | 365955 | 456961 | 2940712 | 20192008 |
| **3** | 515894 | 651595 | 361504 | 2442939 | 16501050 |
| **4** | 1266308 | 514914 | 201448 | 2405326 | 15927657 |
| **5** | 998980 | 287791 | 401917 | 2241488 | 14962700 |
| **6** | 551947 | 749577 | 318031 | 2121886 | 13945760 |
| **7** | 1080000 | 597437 | 178487 | 1940461 | 13299033 |
| **8** | 855971 | 340127 | 429453 | 1891082 | 12722217 |
| **9** | 477406 | 746985 | 341612 | 1487197 | 10244523 |
| **10** | 1075179 | 592017 | 192847 | 1685052 | 11265600 |
| **11** | 856243 | 331432 | 325314 | 1675703 | 11405448 |
| **12** | 485392 | 718033 | 261014 | 1615025 | 10917306 |
| **13** | 982893 | 571267 | 149153 | 1240386 | 8229011 |
| **14** | 773212 | 318596 | 341469 | 1077517 | 7301851 |
| **15** | 428402 | 550215 | 270510 | 972738 | 6780354 |
| **16** | 952613 | 437830 | 151337 | 1044137 | 7668915 |
| **17** | 761433 | 249813 | 205442 | 923128 | 6680061 |
| **18** | 433888 | 478447 | 164773 | 963726 | 6454890 |
| **19** | 864615 | 381452 | 95394 | 709087 | 5121859 |
| **20** | 686969 | 213073 | 203380 | 767277 | 5305586 |
| **21** | 384004 | 430438 | 161578 | 447913 | 3044173 |
| **22** | 832703 | 340697 | 90110 | 450226 | 3297817 |
| **sum** | 17372233 | 10528051 | 5490899 | 33715944 | 229950456 |

**Table 3. Number of well-imputed variants**

Table 3 tabulates the number of well-imputed variants by chromosome with different minor allele frequency (MAF) thresholds. The fourth column shows the number of well-imputed variants without any MAF filters. Note that a lot of the imputed variants are rare and have low MAF, and many of the rare variants do not have high R squares.

## Appendix

**Table A.1 Summary of sex check problems detected by plink**

|  |  |  |  |
| --- | --- | --- | --- |
| ID | Sex in pedigree file  (1=male, 2=female) | Sex as determined by chrX (0=unknown) | chrX inbreeding estimate (F) |
| 7930631002\_R01C02 | 2 | 0 | 0.2284 |
| 8803713091\_R05C02 | 2 | 0 | 0.2448 |
| 8803684058\_R01C02 | 2 | 0 | 0.2934 |
| 8803713018\_R05C02 | 2 | 0 | 0.2491 |
| 8803713028\_R03C02 | 2 | 0 | 0.3745 |
| 8803713034\_R03C01 | 2 | 0 | 0.2152 |
| 9236426092\_R03C02 | 2 | 0 | 0.2364 |
| 9236426092\_R04C02 | 2 | 0 | 0.2494 |
| 9236426109\_R03C01 | 2 | 0 | 0.2358 |
| 9236440023\_R01C02 | 2 | 0 | 0.2121 |
| 9236440041\_R04C01 | 2 | 0 | 0.2342 |

\*We used the –check-sex option in plink, which uses X chromosome data to determine sex (i.e. based on heterozygosity rates) and flags individuals for whom the reported sex in the PED file does not match the estimated sex (given genomic data). A male call is made if F is more than 0.8; a female call is made if F is less than 0.2.