Spit For Science Genetic Data

Quality analysis by sample

v0.6.0
11 July 2023

Table of genotype array quality statistics per sample. This is a tab-delimited text file.

Fields

1. FID = family ID
2. IID = individual ID
3. ClinSex = self-reported sex from the clinical database
4. Xhet = proportion of heterozygous genotypes on the X chromosome
5. YcallRate = proportion of non-missing genotypes on the Y chromosome
6. GeneticSex = estimated sex based on Xhet and YcallRate; 1=male, 2=female, NA=inconclusive
7. SexChr = inferred sex chromosome count
8. CallRate = proportion of non-missing genotypes across the genome
9. AutoHet = proportion of heterozygous genotypes on the autosomes
10. ClosestAncestry = approximate ancestry
11. SexProblems = Is GeneticSex inconsistent with ClinSex? 1=yes, 0=no
12. QCfail = did the sample fail quality control for reasons other than SexProblems? 1=yes, 0=no
13. Reasons = if QCfail=1, reason(s) for failure

Details

FID

Family ID provided by the genotyping lab. May or may not have any significance.

IID

Individual ID ("TAG" ID). All letters have been converted to uppercase as necessary.

ClinSex

Self-reported biological sex from the clinical database. 1=male, 2=female, 3=other, 4=prefer not to answer, 999=unknown/don't know.

Xhet

Proportion of heterozygous genotypes on the X chromosome. Calculated using bcftools 1.9.

YcallRate

Proportion of non-missing genotypes on the Y chromosome. Calculated using PLINK 1.9b6.21.

GeneticSex

Inferred genetic sex based on Xhet and YcallRate. Inferred to be male (GeneticSex=1) if Xhet < 0.05 and YcallRate > 0.75. Inferred to be female (GeneticSex=2) if 0.1 < Xhet < 0.3 and YcallRate < 0.4. Otherwise, inconclusive (GeneticSex=NA).

SexChr

Inferred sex chromosome count:

- XY if Xhet < 0.05 and YcallRate > 0.75
- XX if 0.1 < Xhet < 0.3 and YcallRate < 0.4
- XXY if 0.1 < Xhet < 0.3 and YcallRate > 0.75
- unknown otherwise.

CallRate

Proportion of non-missing genotypes across the genome, after excluding

- SNPs with call rate <97%
- X chromosome SNPs with call rate <97% in males after setting heterozygous calls to missing.

Calculated using PLINK 1.9b6.21. Individuals were considered to have failed the CallRate step if CallRate < 97%.

AutoHet

Proportion of heterozygous genotypes on the autosomes. Calculated using bcftools 1.9. Individuals were considered to have failed the AutoHet step if AutoHet > 6 times the interquartile range (IQR) based on all samples.

ClosestAncestry

Approximate ancestry based on principal components analysis with 1000Genomes data as a reference. Performed using PLINK 1.9b6.21. Individuals were assigned to have the same continental ancestry as the 1000 Genomes sample "closest" to them, where "closest" was defined as Euclidean distance based on the first 3 principal components. Distances were calculated using R version 3.6.1.

SexProblems

Was ClinSex inconsistent with GeneticSex? 1=yes, 0=no. Inconsistent if

- GeneticSex is inconclusive
- ClinSex not missing and ClinSex not the same as GeneticSex.

QCfail

Did the individual fail quality control for non-sex reasons? 1=yes, 0=no. Failed if

- CallRate <97% (IMISS)
- AutoHet >6 IQR (HET).

Reasons

If QCfail = 1, reason(s) for failure. NA if QCfail = 0.

Resources

1000 Genomes data: https://tcag.ca/tools/1000genomes.html (independent samples bundle)
bcftools: https://samtools.github.io/bcftools/bcftools.html
PLINK: https://www.cog-genomics.org/plink/
R: https://cran.r-project.org/