Quality control report for KGPH3abionet_qc

H3Agwas QC Pipeline

Fri Jul 8 09:20:36 SAST 2022

1 Introduction

The input file for this analysis was KGPH3abionet.orig.{bed,bim,fam}. This data includes:

- 2133532 SNPs
- 2504 participants

The input files and md5 sums were

```
KGPH3abionet.bed      167dcda2acb129ab5cd05c42fd7f42c9
KGPH3abionet.bim      d82c762f48366d210b0fa4bfea920be1
KGPH3abionet.fam      3da1e9a1b1ecb2ddad534d93cb241241
```

Note that some statistics are shown twice – on the raw input data and on the final result, since these statistics are needed or different purposed.

Approach

The pipeline takes an incremental approach to QC, trading extra computation time in order to achieve high quality while removing as few data as possible. Rather than applying all cut-offs at once, we incrementally apply cutoffs (for example, removing really badly genotyped SNPs before checking for heterozygosity will result in fewer individuals failing heterozygosity checks).

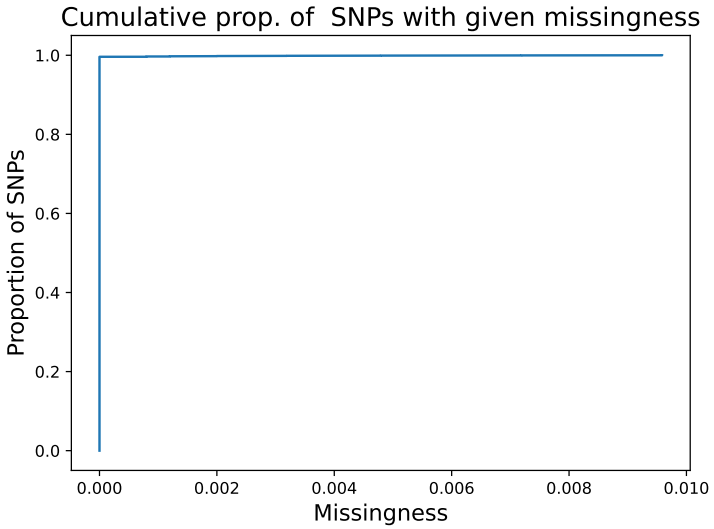
2 QC Phase 0

This phase only removes SNPs which are duplicated (based on SNP name). No other QC is done and so the output of this phase should really be considered as raw data.

1. There were 0 duplicate SNPs. The file with them (if any) is called KGPH3abionet.dups. Note that duplicate SNPs are determined by the names of the SNPs. SNPs which appear at the same position are probably duplicates but may not be. You can control whether you want to detect these using the parameter remove_on_bp. **It is crucial to examine this file to avoid inadvertently removing SNPs. On some chips there are duplicate SNPs at a position – you should select what you what want.**
2. 495 individuals had discordant sex information – the full PLINK report can be found in KGPH3abionet-nd.sexcheck and an extract of the PLINK report showing only the failed reports can be found can be found in KGPH3abionet-nd.badsex, and a more detailed analysis can be found in Section 5.

Figure 1 shows the spread of missingness per SNP across the sample, whereas Figure 2 shows the spread of missingness per individual across the sample. Note that this shows missingness before any filtering or cleaning up of the data.

Figure 1 SNP missingness: For each level of missingness specified on the x axis, the corresponding y-value shows the proportion of SNPs which have missingness *less* than this. [File is `{KGPH3abionet-nd-snpmiss_plot.pdf}`]



Minor allele frequency. Table 1 on page 2 shows the minor allele frequency spectrum for the raw data. The number of monomorphic SNPs is shown in the first row. Note that some of the MAFs with very low MAF are actually monomorphic, with the polymorphisms due to genotyping error. Figure 3 on page 4 shows the cumulative distribution of MAF. This can be used to determine an appropriate MAF cut-off.

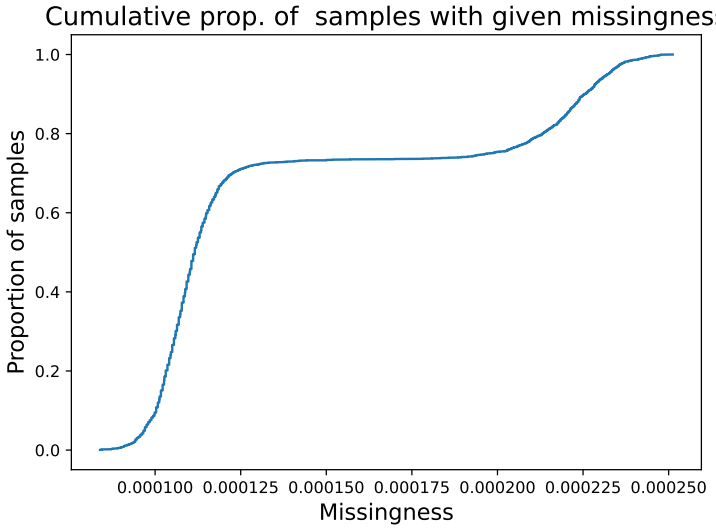
Note that the *minor* allele is determined with respect to the frequency spectrum in this data – 'minor' is not synonym for alternate or non-reference allele, or the allele that has minor frequency in some other data set. Under this definition the MAF is always ≤ 0.5.

Table 1 Minor Allele Frequency spectrum of the raw data. The number of apparently monommorphic SNPs is shown in the row labelled 0; the other rows show the number of SNPs in the bins shown.

MAF bin	Num SNPs
0	170
(0.0, 0.005]	80982
(0.005, 0.01]	95744
(0.01, 0.02]	199806
(0.02, 0.03]	130005
(0.03, 0.04]	88191
(0.04, 0.05]	69375
(0.05, 0.1]	252997
(0.1, 0.15]	206665
(0.15, 0.2]	179802
(0.2, 0.25]	162666
(0.25, 0.3]	147657
(0.3, 0.4]	269344
(0.4, 0.5]	250128

Hardy Weinberg Statistics : Figure 4 shows the cumulative distribution of Hardy-Weinberg p-value for the SNPs in the raw data. This can be used to assess the cost of excluding SNPs with a

Figure 2 Missingness per indvididual: For each level of missingness specified on the x axis, the corresponding y-value shows the proportion of individuals which have missingness *less* than this. [File is `{KGPH3abionet-nd-indmiss_plot.pdf}`]



particular p-cutoff. We expect the curve the fit tightly to the main diagonal, except for a very small p values (and this deviation may not be observable on a linear plot).

The QQ plot for the HWE scores can be found in Figure 5. The region of deviation from the line of expected versus observed p-values will be more observable here. Note that if there are very small observed p-values in relation to expected values, the expected curve may be very flat — pay attention to the x and y axis coordinates. Since we are plotting on a negative log-scale, note that regions of low probabiliy of deviation from HWE (p-value close to 1) are at the left, and regions of high probability (low p-value) are at the right. The tail of the plot where deviation from the diagonal occurs is likely to be a good cut-off to use for QC.

However, care needs to be taken not exclude SNPs. We are using HWE p-value as a proxy for something having gone wrong with the sample or genotyping, and this is a little crude. In a study with participants from different population groups in a recently admixed group, deviation from HWE is expected and does not indicate problems with QC. Moreover, in a disease study, it is likely that those individuals that are affected, those SNPs that are associated with the condition under study will not in be in HWE. Care needs to be taken – it is easier to handle in a pure case/control study. In a population cross-section study with different conditions being considered, it might be advisable to re-run the QC pipeline for HWE for each study. The current version of the pipeline does not support his more complex analysis, though we plan to extend.

Figure 3 Cumulative frequency of SNPs. For a frequency shown on the x-axis, the corresponding y-value shows the proportion of SNPs with frequency *at least this frequency*; that is, it shows the proportion of SNPs which will *remain* if the MAF filter of this x-value is chosen.

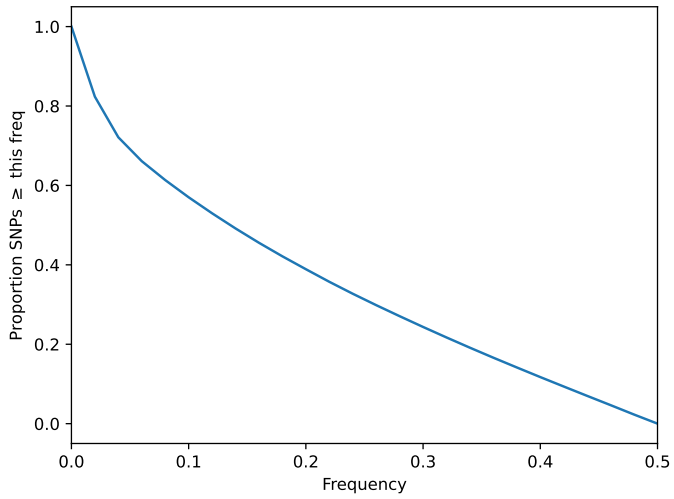


Figure 4 HWE distribution. For an HWE-value shown on the x-axis, the corresponding y-value shows the proportion of SNPs with HWE p-value *at least this frequency*; that is, it shows the proportion of SNPs which will *be removed* if the HWE filter of this x-value is chosen. File is `KGPH3abionet-nd-inithwe.pdf`.

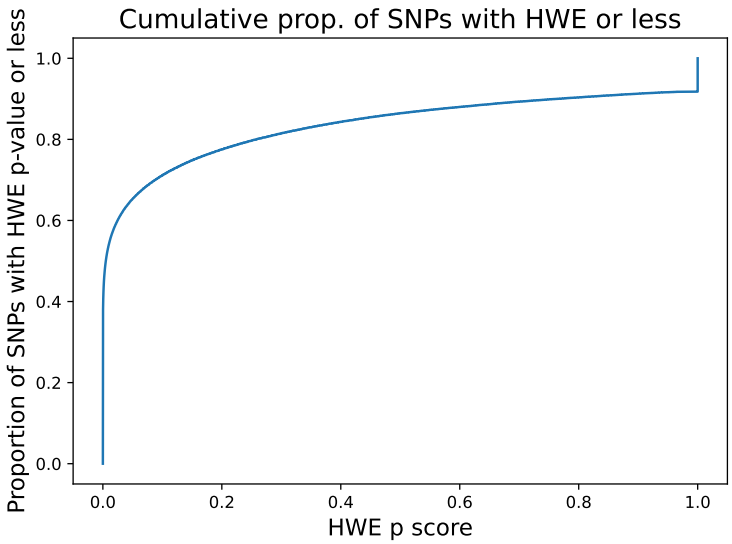
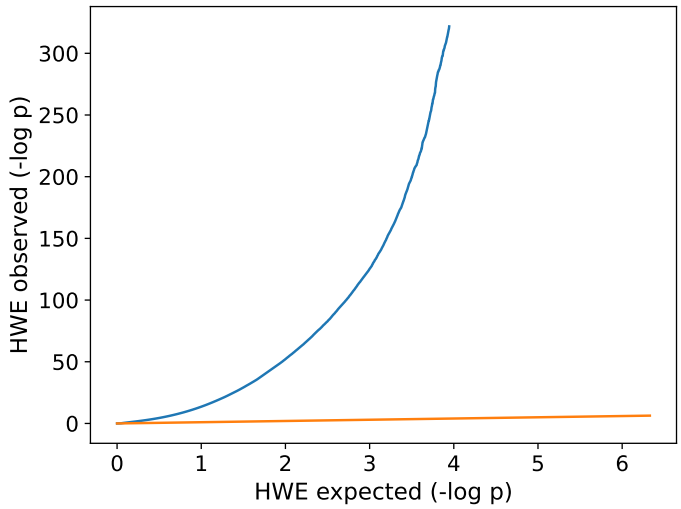


Figure 5 QQ plot for Hardy-Weinberg scores. The graphic can be found in the file `KGPH3abionet-nd-inithwe-qq.pdf`



3 Plate Analysis

No plate analysis can be done because a 10% GenCall score can't be found in the sample sheet or no sample sheet given.

4 QC Phase 1

The details of the final filtering can be found in the Nextflow script. Note that the exact ordering or removal will affect the final results. However, we take a conservative approach.

1. Only autosomal SNPs are included
2. SNPs and individuals that have been very poorly genotyped (missingness exceeding 10 per cent) are removed.
3. Individuals with missingness greater than 0.02 are removed (by filtering out very badly genotyped individuals or SNPs) in the previous steps we *may* save a few individuals in this step;
4. SNP with missingness greater than 0.01 are removed ;
5. minor allele frequency less than 0.001 (and greater than 1-0.001) are removed;
6. HWE adjusted p-value less than 0.0005 are removed

Using this approach,

- 0 SNPs that are non-autosomal were removed;
- 0 SNPs were removed due missing genotype threshold constraints;
- 0 individuals were removed due to missing genotype constraints (the list of missing individuals, if any, can be found in the file `KGPH3abionet-nd-c.irem`);
- 0 SNPs were removed as the MAF was too low.
- 824921 SNPs were removed as they were out of the specified Hardy-Weinberg equilibrium.

5 Batch report

The batch quality results shown here (missingness and sex-check failures) is based on the raw input data (other than duplicate SNPs being deleted) as this gives insight into the quality of the raw data differences between batches and sub-batches. The PC analysis is based on the refined data.

5.1 Overall missingness and sex anomaly report

Table 2 on page 6 shows the error rate as shown by `pheno_1` as found in file `pheno.phe`.

Table 2 For each group shown, we have: the average percentage missingness in that group (100 × the sum the number of missing calls over all individuals divided by the total number of genotype calls over all individuals); the percentage of samples in that group with a genotyping error rate above 4 percent; and the percentage of samples in that group that fail the sex check using the standard PLINK parameters on the raw input data.

pheno_1	Num samples	Missing rate	\% poor	Sex Checkfail
1	2253	0.01	0.00	19.71
2	251	0.01	0.00	19.92

Table 3 on page 7 shows the error rate.

5.2 Detailed Sex Check Analysis

This section shows a detailed analysis of sex check anomalies and/or unusual patterns in the X-chromosome. The purpose of this analysis is to help identify trends between sub-groups, as well as possible labelling and sample handling errors. The term *anomaly* or *error* is used to label individuals where the sex of the individual as described in the manifest does not *stricly* match analysis of the X-chromosome and so for QC should be considerd further.

In this analysis, we use PLINK to analyse the non-recombining regions of the X-chromosome, and in particular its computation of the inbreeding co-efficient of the X-chromosome. If the F statistic is greater than 0.8, PLINK infers that the sample is male; if it is less than 0.2, it infers that the sample is female.

Reminder: the checking of sex on the raw data before any other QC is shown in Section 1. The rest of this section analyses the data after basic QC on genotype has been done and so differences may be seen.

There are two types of apparent anomaly that can happen. *Soft* anomalies are those cases where an individual is slightly above or below the stated threshholds. These may not be sample handling errors – since the F cut-off values are arbitrary, a too strict F-value may be chosen, or there may be uncommon patterns within the individuals studied. How these samples should be treated will require some thought, but these are not a sign of problems of the experimental protocol, and not by itself probably a sign of problems with genotyping or DNA quality errors. We define a soft anomaly as an indvidual having an F value between 0.2 and 0.8, which is possible but unusual.

Hard errors are cases where the sex in the manifest/fam file is markedly different from what the F-statistic predicts (e.g., the F-statistic says 0.996 and we have this as a female). While there may be very unusual cases were this occurs, many such examples in the data are likely to be a sign of sample handling errors. Great care needs to be taken with this.

A summary of the detailed analysis is shown below. In the output, the file `KGPH3abionet-nd-c-prune_missing_and_sexcheck.csv` is a CSV file that has sample ID, per-individual missingness rate in the raw data (`F_MISS` is the individual missingness rate *not* the F-statistic), the `pheno1` status, and whether that sample is hard (H) or soft error (S) for given tolerated per-individual genotyping error rates on the X-chromosome. A '-' indicates that the indivdual is filtered out at that rate of missingness This can be used to assess whether anomalous sex results are due to poor genotyping rates in individuals.

Table 3 For each group shown, we have: the average percentage missingness in that group (100 × the sum the number of missing calls over all individuals divided by the total number of genotype calls over all individuals); the percentage of samples in that group with a genotyping error rate above 4 percent; and the percentage of samples in that group that fail the sex check using the standard PLINK parameters on the raw input data.

all	Num samples	Missing rate	\% poor	Sex Checkfail
1	2504	0.01	0.00	19.73

Table 4 of anomalous sex calls. The results are shown for different values *mind* the maximum per-sample error rate tolerated in the X-chromosome (PLINK parameter). *mind*==1 means all SNPs included. *Tot* shows the number of individuals included with the given *mind* value. *HErr* is the number of hard errors. *SAnm* is the number of soft apparent anomalies or unusual results.

pheno 1	mind=1			mind=0.01			mind=0.03			mind=0.05		
	Tot	SAnm	HErr	Tot	SAnm	HErr	Tot	SAnm	HErr	Tot	SAnm	HErr
overall	2504	389	105	2504	389	105	2504	389	105	2504	389	105
1	2253	351	93	2253	351	93	2253	351	93	2253	351	93
2	251	38	12	251	38	12	251	38	12	251	38	12

5.3 Relatedness

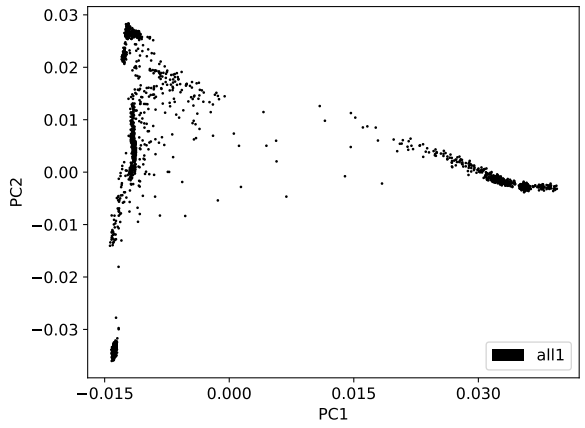
There are no related pairs of individuals ($\widehat{\pi} > 0.8$, the parameter of the pipeline).

5.4 Principal component anlysis

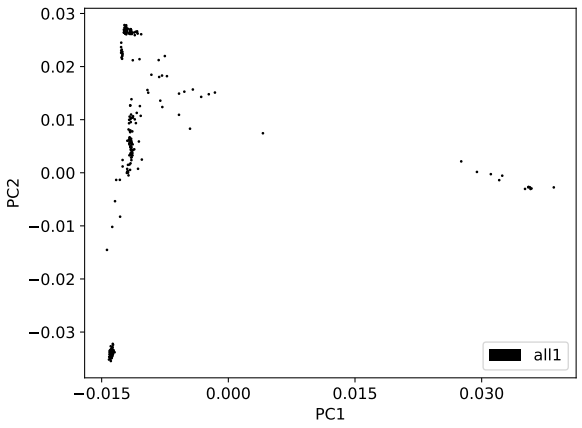
In the PC analysis, the principal components were computed from all samples in the data together, and in the figure(s) we extract the samples for that analysis. The labels and scale of the axes may differ, but the coordinates are the same: the position (0,0) is the same in all sub-figures.

Figure 6 Principal component analysis, using the batch of the samples as the label.

(a) pheno 1 1



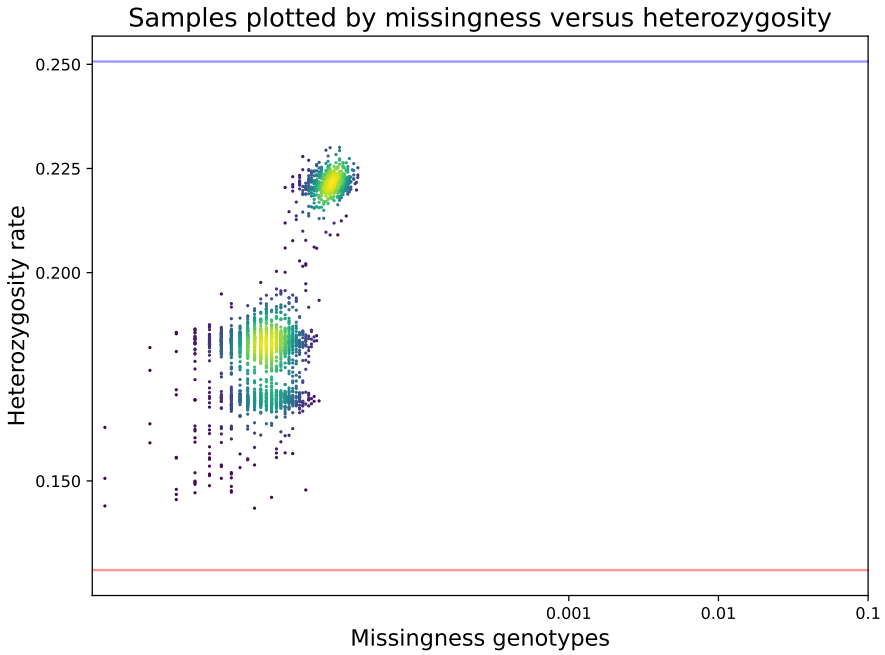
(b) pheno 1 2



6 Heterozygosity check

Levels of heterozygosity were examined in the data filtered in the previous step. Figure 7 shows plots of heterozygosity versus individual missingness (i.e., the number of SNPs missing per individual). Levels of heterozygosity should be between the ranges given in the confguration file – anything higher may indicate that there is sample contamination, lower may indicate inbreeding. However, each set of data must be treated on its own merits and the analyst must apply their mind the problem. Missingness should be low.

Figure 7 Missingness versus heterozygosity: the lines show the mean heterozygosity plus/minus standard deviations (the brighter the colours, the greater the density). If there is zero missingness, only heterozygosity is shown in a violin plot. [File is `{KGPH3abionet-nd-c-imiss-vs-het.pdf}`]



Individuals out of range heterozygosity were removed. Any indviduals with heterozygosity:

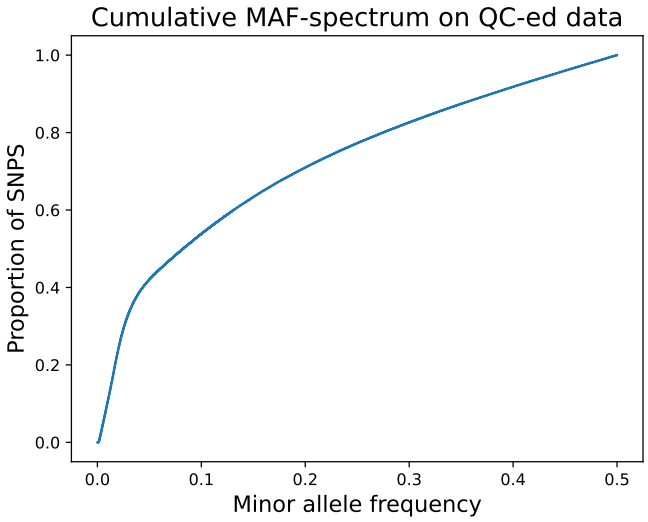
- less than 0.15 are removed. This may indicate inbreeding.
- greater than 0.343 are removed. This may indicate sample contamination.

Overall 16 individuals were removed. These individuals, if any, can be found in the file `KGPH3abionet-nd-c-fail_het.txt`.

7 Minor Allele Frequency Spread

Figure 8 shows the cumulative distribution of minor allele frequency in the data **after** quality control (the figures shown in Section 2 show the MAF before QC. The MAF cut-off should be chosen high enough that one is sure that the variants seen are real (so this would depend on the size of the sample and the quality of the genotyping and whether some of the data is imputed). In this analysis the cut off was 0.001. Again, note that the *minor* allele is determined with respect to the frequency spectrum in this data – 'minor' is not synonym for alternate or non-reference allele, or the allele that has minor frequency in some other data set. Under this definition the MAF is always ≤ 0.5.

Figure 8 Minor allele frequency distribution [File is `{KGPH3abionet_qc-maf_plot.pdf}`]



8 Differences between cases and controls

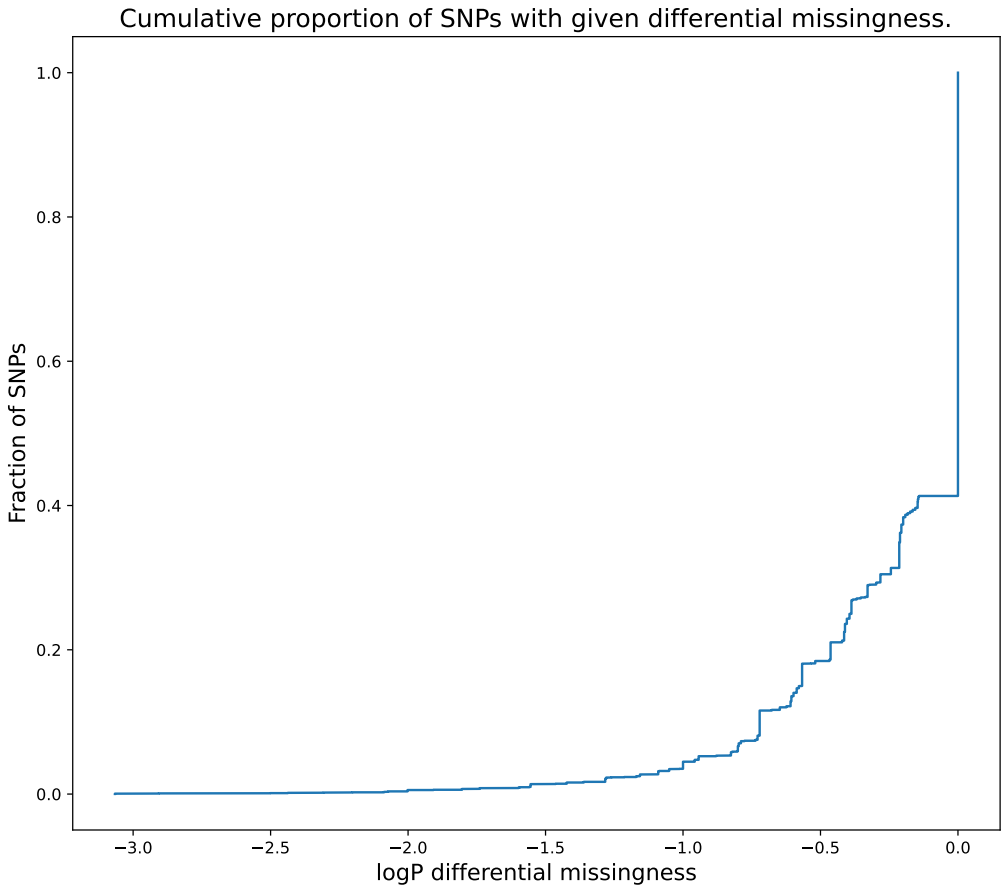
We do not expect there to be large, observable macro-scale differences between cases and controls. Great caution needs to be taken in this case. If the samples are from heterogeneous groups, where the case-control status differs between groups, then there may well statistically significant differences between the cases and controls and a batch analysis should be undertaken.

We compute for each SNP the missingness in the cases, and the missingness in the controls, and the corresponding p-value describing the difference in missingness.

We expect very few SNPs to have highly significant differences. Where many SNPs with very highly significant p-values are found, great care should be taken.

Figure 9 plots the differences between cases and controls, showing the SNP-wise p-value, unadjusted for multiple testing

Figure 9 The plot shows differential missingness for each (log) level of significance, the number of SNPs with p-value of at least this significance. The flatter the better. [File is `KGPH3abionet-c-c-diff-snpmiss_plot.pdf`]



For removal of SNPs, we compute the p-value adjusted for multiple testing, by performing permutation testing (1000 rounds) using the PLINK mperm maxT option. SNPs are removed from the data set if their adjusted (EMP2) differential missingness p-value is less than 0.05. The SNPs that are removed can be found in the file `KGPH3abionet-nd-c-c_missing-failed_diffmiss.snps`

Figure 10 shows a principal component analysis of the data, identifying the cases and controls. Should the cases and controls cluster differ signficantly, something is likely wrong. Moreover should there be any significant clusters or outliers, association testing should take into account stratification. Statistical testing could also be done. Figure 11 shows for each principal component what the corresponding eigenvalue is. The shape of the curve may indicate if there is any structure in the data. Informally, the "broken stick" model sees two processes generating PCs – population structure and

random fluctuation. The eigenvalue is a function of both – the former has a major impact but drops relatively rapidly; the latter has less effect but drops more slowly. If (!) the model is correct then you will see an inflection point in the graph, where the random effects become the leading factor in the change. More formal analysis may be desirable, e.g., using the Tracy-Widom statistic or Velicer's MAP test.

Figure 10 Principal Component Analysis of Cases Versus Controls [File is `{KGPH3abionet-nd-c-prune-pca.pdf}`]

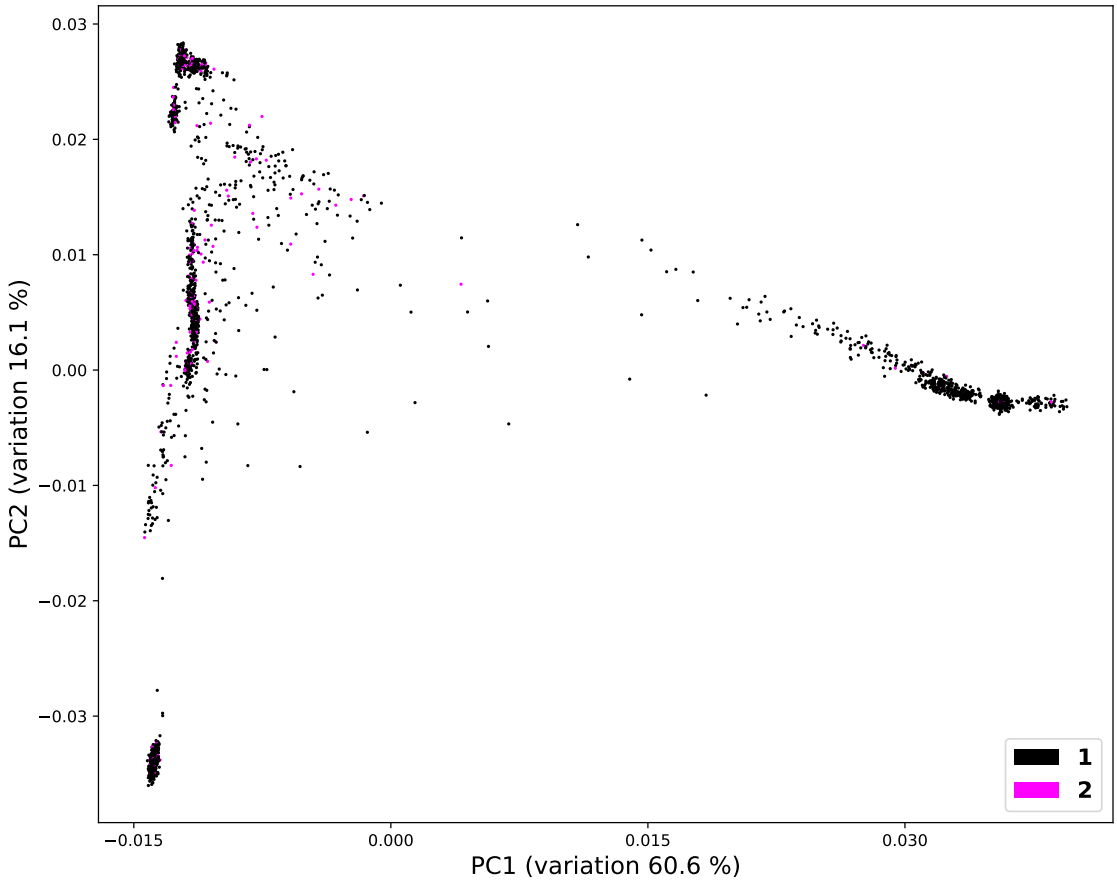
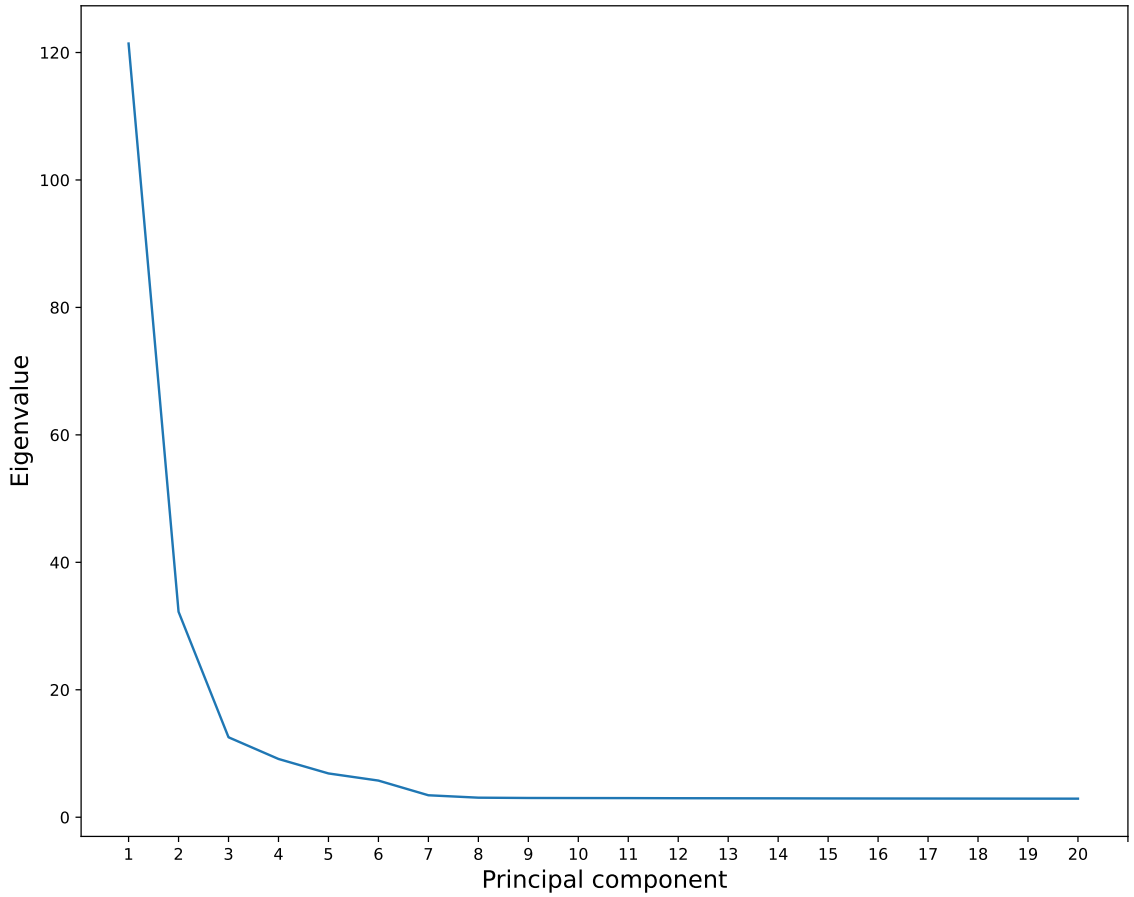


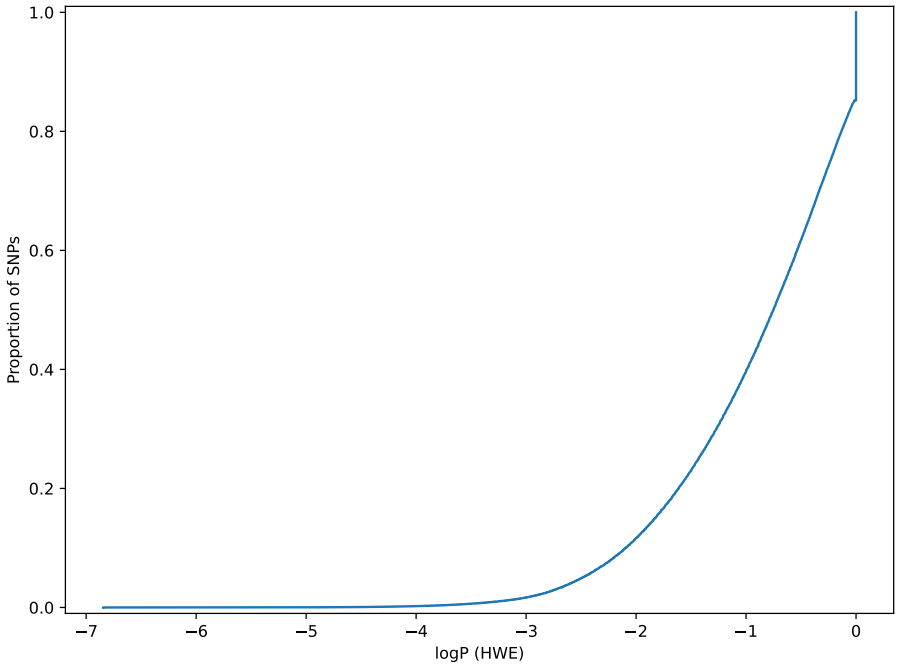
Figure 11 Eigenvalues for each principal component: the shape of the curve gives some indication of how many PCs are important [File is `{eigenvalue.pdf}`]



9 Hardy-Weinberg Equilibrium

Deviation for Hardy-Weinberg Equilibrium (HWE) may indicate sample contamination. However, this need not apply to cases, nor in a situation where there is admixture. For each SNP, we compute the probability of the null hypothesis (that the deviation from HWE is by chance alone). Figure 12 shows a plot of the corresponding p-value versus the frequency of occurrence.

Figure 12 The plot shows for each level of significance, the number of SNPs with H WE p-value [File is `{KGPH3abionet-nd-c-c-unaff-hwe_plot.pdf}`]



10 Final results

The quality control procedures as described below were applied to the data. The final, cleaned result contains:

- 1258341 SNPs
- 2010 participants

The final output files are

- `KGPH3abionet_qc.bed`,
- `KGPH3abionet_qc.bim`, and
- `KGPH3abionet_qc.fam`.

The ouput files' md5 sums are shown below

```
KGPH3abionet_qc.bed 3fd118679aff92dd39d6518b6a90db5a
KGPH3abionet_qc.bim 13a0809309fdb8cb94e78d0d48b9d63a
KGPH3abionet_qc.fam 9ebde08339da52cb6db2def7c6281308
```

Table 5 Docker Images Used

Nextflow process	Docker Image
default	quay.io/h3abionet_org/py3plink

11 Technical details

The analysis and report was produced by the h3aGWAS pipeline (http://github.com/h3abionet/h3agwas) produced by the Pan-African Bioinformatics Network for H3Africa (http://www.h3abionet.org).

The following tools were used:

- PLINK v1.90b6.16 64-bit (17 Feb 2020) [Chang et al 2015]
- 22.04.3 [Di Tommaso et al, 2017]
- A local copy of the workflow was used
- The command line below was called [NB: if the command line is long, the linebreak may break oddly after a hyphen or dash so take care.]

```
nextflow run /home/jeantristan/Travail/git/h3agwas/qc/main.nf -profile
    slurmSingularity -c utils/params_1000G_h3abionet_qc.params
```

- The profile slurmSingularity was used: the docker images used are found in Table 5
- The full configuration can be found in the appendix.

12 References

- Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, and Lee JJ. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience*, 4(1), 1-16. `http://doi.org/10.1186/s13742-015-0047-8`
- Di Tommaso P, Chatzou M, Prieto Barja P, Palumbo P, Notredame C. Nextflow enables reproducible computational workflows. *Nature Biotechnology*, 35, 316-319, 2017. Nextflow can be downloaded from `https://www.nextflow.io/`

A Configuration

A.1 config

```
singularity.cacheDir = "/home/jeantristan/.singularity/"
```

A.2 nextflow

```
plugins {
  id 'nf-azure'
}

py3Image = "quay.io/h3abionet_org/py3plink"
gemmaImage="quay.io/h3abionet_org/py3plink"
latexImage="quay.io/h3abionet_org/h3agwas-texlive"
py2fastImage="quay.io/h3abionet_org/py2fastlmm"

swarmPort = '2376'
queue = 'batch'

manifest {
    homePage = 'http://github.com/h3abionet/h3agwas'
    description = 'GWAS Pipeline for H3Africa'
    mainScript = "main.nf"
}


aws {
accessKey='*******'
secretKey='*******'
    region    ='us-east-1'
}

/*cloud {
    // imageId = "ami-710b9108"      // specify your AMI id here
    instanceType = "m4.xlarge"
    subnetId = "null"
    sharedStorageId   = "null"
    sharedStorageMount = "/mnt/shared"
    bootStorageSize = "20GB"     // Size of disk for images spawned
    //   instanceStorageMount = ""   // Set a common mount point for images
    //   instanceStorageDevice = ""  // Set a common block device for images
    autoscale {
        enabled = true
        maxInstances = 1
        terminateWhenIdle = true
    }

}
*/


params {

    // Directories
    work_dir                = "/$PWD"
    input_dir               = "sample"
    // Can use S3 too
    //input_dir               = "s3://h3abionet/sample"
    input_pat               = "sampleA"

    output_dir              = "${params.work_dir}/output"
    scripts                 = "${params.work_dir}/scripts"
    output                  = "out"

    max_forks            = 95
```

```
    high_ld_regions_fname = ""
    sexinfo_available    = true
    cut_het_high         = 0.343
    cut_het_low           = 0.15
    cut_diff_miss         = "0.05"
    cut_maf                = "0.01"
    cut_mind              = "0.02"
    cut_geno              = 0.01
    cut_hwe               = 0.008
    pi_hat                = 0.11
    super_pi_hat    = 0.7
    f_lo_male             = 0.8 // default for F-sex check -- >= means male
    f_hi_female           = 0.2 //  <= means female
    gc10                 = 0.4  // 10% Gen Call confidence -- 0.4 is very low
    case_control         = "${params.input_dir}/sample.phe"
    case_control_col     = "PHE"


    phenotype = "0"
    pheno_col = ""
    batch = "0"
    batch_col = 0

    idpat                =  0  // or "(\w+)-DNA_(\w+)_.*" or ".*_(.*)"

    plink_mem_req        = "1750MB"
    other_mem_req        = "1750MB"
    big_time             = '12h'
    sharedStorageMount   = "/mnt/shared"
    max_plink_cores      = 4


}



profiles {


    // For execution on a local machine, no containerization. -- Default
    standard {
        process.executor = 'local'
    }

    awsbatch {
         process.executor = "awsbatch"
 aws.region    ='us-east-1'
         aws.uploadStorageClass = 'ONEZONE_IA'
         process.queue = 'h3a-00'
    }

    azurebatch {
        process.executor = 'azurebatch'
    }

    slurm {
       process.executor = 'slurm'
       process.queue = queue
     }



    // Execute pipeline with Docker locally
    docker {
        docker.remove      = true
```

```
        docker.registry    = 'quay.io'
        docker.enabled     = true
        docker.temp        = 'auto'
        docker.fixOwnership= true
docker.runOptions = '-u $(id -u):$(id -g) --rm'
        docker.process.executor = 'local'
    }


    // Execute pipeline with Docker Swarm setup
    dockerSwarm {
        docker.remove      = true
        docker.runOptions  = '--rm'
        docker.registry    = 'quay.io'
        docker.enabled     = true
        docker.temp        = 'auto'
        docker.fixOwnership= true
        docker.process.executor = 'local'
        docker.engineOptions = "-H :$swarmPort"
    }

    // For execution on a PBS scheduler, no containerization.
    pbs {
        process.executor = 'pbs'
        process.queue = queue
    }




    // For execution on a SLURM scheduler, no containerization.
    slurm {
        process.executor = 'slurm'
        process.queue = queue
    }


    slurmSingularity {
        singularity.cacheDir = "${HOME}/.singularity"
        process.executor = 'slurm'
        singularity.autoMounts = true
        singularity.enabled = true
        singularity.runOption = "--cleanenv"
        process.queue = queue

    }




    singularity {
        singularity.cacheDir = "${HOME}/.singularity"
        singularity.autoMounts = true
        singularity.enabled = true
     }
}




process {

    withLabel:bigMem {
```

```
      memory = '8GB'
    }

    container=py3Image

    withLabel: latex {
          container = latexImage
    }

    withLabel: py2fast {
          container = py2fastImage
    }

}

timeline {
    enabled=true
    file = "nextflow_reports/timeline.html"
}

report {
    enabled = true
    file = "nextflow_reports/report.html"
}
```

A.3 params_1000G_h3abionet_qc

```
params {
  // what is input direction of you plink
  input_dir = "KGPH3abionet/geno_all/"
  // what is pattern of you plink file (without extension)
  input_pat = "KGPH3abionet"
  // output directory
  output_dir = "$PWD/KGPH3abionet_qc"
  //output header
  output = "KGPH3abionet_qc"
  // pipeline requrie a case control to test association
  case_control = "KGPH3abionet/simul_pheno/qual_pheno/KGPH3abionet_ql.pheno"
  //header of your case control
  case_control_col = "pheno_1"
  // if you have batch file : check if there is a biais in batch
  // check phenotype
  phenotype = "KGPH3abionet/simul_pheno/qual_pheno/KGPH3abionet_ql.pheno"
  pheno_col = "pheno_1"
  // high ld region : high_ld_regions_fname: this is optional -- it is a list of regions which are in very high LD
  //https://github.com/genepi-freiburg/gwas/blob/master/single-pca/high-LD-regions.txt
  //high_ld_regions_fname = "high_LD_regions.txt"
  // GC10 : score of genomic control (optional)
  // maximum allowable relatedness
  pi_hat = 0.8
  // What is the maximum allowable heterozygosity for individualsl;
  cut_het_high          = 0.343
  // What is the minimum allowable heterozygosity for individualsl;
  cut_het_low           = 0.15
  // allowable differential missingness between cases and controls;
  cut_diff_miss         = "0.05"
  // Minor allele frequencie
  cut_maf                     = "0.001"
  // maximum allowable per-individual missingness
  cut_mind              = "0.02"
  // maximum allowable per-SNP mssingness
  cut_geno              = 0.01
  // minimum allowable per-SNP Hardy-Weinberg Equilibrium p-value
  cut_hwe               = 0.0005
  // f_low_male and f_hi_female. Discordant sex genotype is done on the X-chromosome using the non-recombining part
  f_lo_male            = 0.8 // default for F-sex check -- >= means male
```

```
  f_hi_female           = 0.2 //  <= means female
  // sex info available : do you have sex from another sex (in plink file must be )
  sexinfo_available    = true
  plink_mem_req = "20G"
}
```