# qc-rotterdam1-report

*Øyvind Helgeland*

*May 2, 2018*

## Module 0 : Data conversion

Data conversion from GenomeStudio to PLINK format was carried out prior to data delivery.

## Module 1 Data preparation

Module 1 prepared data for QC procedures. In the exported data from GenomeStudio, samples were coded using project specific retrieval IDs and non-informative family IDs. Also, no sex information was available.

Information about declared sample sex and pedigree structure provided by MoBa were used to update the dataset.

Markers with poor cluster separation, low 10% GC score and high AA theta deviation were removed. Clustering metrics were provided by the SNP table exported from GenomeStudio.

The array contains some duplicated/triplicated markers. Duplicates/triplicates were removed to avoid potential problems downstream.

Illumina provides a conversion list for converting marker IDs used on the array to rsIDs. In the provided list some markers are named ‘.’. To avoid duplicate/non-informative IDs, the original chip ID was used in situations were no rsID was available.

Illumina technical markers (assigned to chromosome 0) were excluded and markers with poor clustering were eliminated using metrics available from the variant table exported from GenomeStudio. Problematic markers reported by other consortia were subsequently excluded.

NOTE: All markers in the PAR region was correctly assigned to chr 24, no update necessary.
NOTE: The lab in Rotterdam removed 7710 markers showing poor performance prior to delivery (markers not included in the dataset)

**Updating sample IDs:**
Samples in: 17949
Samples updated: 17949
Samples not updated: 0

**Updating parental information:**
Samples in: 17949
Samples assigned one or more parents: 5984
Samples not updated: 11965

**Updating sex:**
Samples in: 17949
Samples where sex was updated: 17949
Samples not updated for sex: 0

**Remove markers by cluster separation < 0.4:**
Markers in: 692367
Markers removed: 18154
Markers remaining: 674213

**Remove markers by 10% GC score:**
Markers in: 674213
Markers removed: 19915
Markers remaining: 654298

**Remove markerst by AA theta dev:**
Markers in: 654298
Markers removed: 4123
Markers remaining: 650175

**Remove duplicated markers:**
Markers in: 650175
Markers removed: 480
Markers remaining: 649695

**Update SNPs to rsID:**
Markers in: 649695
Markers updated: 338017
Markers not updated: 311678

**Removed technical markers (chr0):**
Markers in: 649695
Markers removed: 1910
Markers remaining: 647785

## Module 2 Identify core samples and infere pedigree

In module 2 an ethnically homogeneous core set of samples were identifed for use in module 3 (marker cleaning). Marker cleaning requires an ethnically homogeneous sample set in order to facilitate marker cleaning sensitive to ethnic outliers. Initially markers with MAF > 10% were temporarily removed. Samples with missingness rate > 5% were permanently removed. Markers with missingness > 2% were temporarily removed. The resulting dataset was then used to assess and update the sex of samples where reported and genetic sex did not match.

Further, strand-ambiguous markers, non-autosomal markers, and markers with HWE p < 1e-4 were temporarily removed before IBD estimation.

the dataset was subsequently split into a parent dataset and offspring dataset, the latter including only one child in case of siblings (selected at random). The resulting datasets were cleaned separately in the modules described in the following modules and merged in module 7.

**Markers and samples in**
Markers in start: 647785
Samples in start: 17949

**Temporary removal of markers with MAF < 10%**
Markers in: 647785
Markers removed: 415378
Markers remaining: 232407

**Permanent removal of samples with missingness > 5%**
Samples in: 17949
Samples removed: 207
Samples remaining 17742

**Temporary removal of markers with missingness > 2%**
Markers removed: 232407

Markers removed: 3879
Markers remaining: 228528

**Temporary removal of non-autosomal markers**
Markers in: 228528
Markers removed: 8521
Markers remaining: 220007

**Temporary removal of markers with HWE p < 1e-4**
Markers in: 220007
Markers removed: 842
Markers remaining: 219165

**Temporary removal of strand ambiguous markers**
Markers removed: 219165
Markers removed: 1205
Markers remaining: 217960

**Temporary remove markers in high LD**
Markers in: 217960
Markers removed: 6338
Markers remaining: 211622

**Prune set og markers using –indep-pairwise 200 100 0.1**
Markers in: 211622
Markers removed: 171386
Markers remaining: 40236

JONAS STUFF…

**Removal of pedigree inconsistent samples**
Samples in: 17742
Samples for removal: 64
Samples removed: 64
Samples remaining: 17678

**PCA after merge with HapMap**
Markers after HapMap merge (used for PCA): 20308

**Sample selection post PCA**
Samples in: 17678
Samples removed after PCA: 785
Samples remaining after PCA: 16893

**Split dataset into founders and offspring**
Samples in: 16893
Founders: 11316
Offspring: 5577

Founders

**IBD estimation** Samples in: 11316 Markers in: 647785

**Remove samples with excess accumulated PIHAT:** Samples in: 11316 Samples for removal: 14 Samples removed: 14 Samples remaining: 11302

**Remove one in a pair of samples with PIHAT > 0.1:** Samples in: 11302 Samples for removal: 478 Samples removed: 461 Samples remaining: 10841 </div>

<div class="column-right"> Offspring

**IBD estimation** Samples in: 5577 Markers in: 647785

**Remove samples with excess accumulated PIHAT:** Samples in: 5577 Samples for removal: 11 Samples removed: 11 Samples remaining: 5566

**Remove one in a pair of samples with PIHAT > 0.1:** Samples in: 5566 Samples for removal: 90 Samples removed: 87 Samples remaining: 5479

## Module 3 Identify good markers

## Founders

**Number of markers and samples at start of cleaning:**
Samples in start: 10841
Markers in start: 647785

**Remove markers with missingness > 10%:**
Markers in: 647785
Markers removed: 288
Markers remaining: 647497

**Remove individuals with missingsness > 5%:**
Samples in: 10841
Samples removed: 1
Samples remaining: 10840

**Remove markers with missingness > 5%:**
Markers in: 647497
Markers removed: 1417
Markers remaining: 646080

**Remove individuals with missingess > 3%:**
Samples in: 10840
Samples removed: 19
Samples remaining: 10821

**Remove markers with missingness > 2%:**
Markers in: 646080
Markers removed: 8322
Markers remaining: 637758

**Remove individuals with missingness > 2%:**
Samples in: 10821
Samples removed: 18
Samples remaining: 10803

**Remove autosomal markers with HWE p < 1e-7:**
Markers in: 637758
Markers removed: 1659
Markers remaining: 636099

**Remove samples with HET excess > 4SD using common autosomal markers (MAF > 0.01)**
Samples in: 10803
Samples removed: 1
Samples remaining: 10802

**Remove autosomal markers with HWE p < 1e-6**
Markers in: 636099

Markers removed: 208
Markers remaining: 635891

**Remove samples with HET excess > 4SD using rare autosomal markers (MAF > 0.01)**
Samples in: 10802
Samples removed: 61
Samples remaining: 10741

**Remove markers with missingness > 2%:**
Markers in: 635891
Markers removed: 3
Markers remaining: 635888

**Temporarily remove samples failing sex check (F: 0.2, 0.8):**
Samples in: 10741
Samples for removal: 5
Samples removed: 5
Samples out: 10736

**Markers into sex clean:**
X markers in: 15099
Y markers in: 712
PAR markers in: 564
MT markers in: 126

**Remove chrX and PAR markers with HWE p < 1e-6 (only female):**
Markers (X + PAR) in: 15663
Markers removed: 29
Markers remaining: 15634

**Remove chrX marker if any male has at least one heterozygote genotype:**
Markers removed: 694
Markers remaining 635165

**Markers after sex clean:**
Autosomes markers out: 619387
X markers out: 14404
Y markers out: 712
PAR markers out: 536
MT markers out: 126
TOTAL: 635165

## Offspring

**Number of markers and samples at start of cleaning:**
Samples in start: 5479
Markers in start: 647785

**Remove markers with missingness > 10%:**
Markers in: 647785
Markers removed: 246
Markers remaining: 647539

**Remove individuals with missingsness > 5%:**
Samples in: 5479
Samples removed: 2
Samples remaining: 5477

**Remove markers with missingness > 5%:**
Markers in: 647539
Markers removed: 1223
Markers remaining: 646316

**Remove individuals with missingess > 3%:**
Samples in: 5477
Samples removed: 19
Samples remaining: 5458

**Remove markers with missingness > 2%:**
Markers in: 646316
Markers removed: 7548
Markers remaining: 638768

**Remove individuals with missingness > 2%:**
Samples in: 5458
Samples removed: 9
Samples remaining: 5449

**Remove autosomal markers with HWE p < 1e-7:**
Markers in: 638768
Markers removed: 974
Markers remaining: 637794

**Remove samples with HET excess > 4SD using common autosomal markers (MAF > 0.01)**
Samples in: 5449
Samples removed: 4
Samples remaining: 5445

**Remove autosomal markers with HWE p < 1e-6**
Markers in: 637794
Markers removed: 123
Markers remaining: 637671

**Remove samples with HET excess > 4SD using rare autosomal markers (MAF > 0.01)**
Samples in: 5445
Samples removed: 34
Samples remaining: 5411

**Remove markers with missingness > 2%:**
Markers in: 637671
Markers removed: 5
Markers remaining: 637666

**Temporarily remove samples failing sex check (F: 0.2, 0.8):**
Samples in: 5411
Samples for removal: 2
Samples removed: 2
Samples out: 5409

**Markers into sex clean:**
X markers in: 15251
Y markers in: 712
PAR markers in: 564
MT markers in: 126

**Remove chrX and PAR markers with HWE p < 1e-6 (only female):**
Markers (X + PAR) in: 15815

Markers removed: 25
Markers remaining: 15790

**Remove chrX marker if any male has at least one heterozygote genotype:**
Markers removed: 477
Markers remaining 637164

**Markers after sex clean:**
Autosomes markers out: 621013
X markers out: 14774
Y markers out: 712
PAR markers out: 539
MT markers out: 126
TOTAL: 637164

## Module 4 : Individuals for analyses

### Founders

**Markers and samples at beginning of module:**
Markers start: 635165
Samples start: 11316

**Remove markers not surviving QC in both parents and offspring:**
Markers in: 635165
Markers removed: 427
Markers remaining: 634738

**Remove samples with missingness rate > 2%:**
Samples in: 11316
Samples removed: 42
Samples remaining: 11274

**Remove samples with HET excess > 4SD using common autosomal markers (MAF > 0.01):**
Samples in: 11274
Samples removed: 1
Samples remaining: 11273

**Remove samples with HET excess > 4SD using rare autosomal markers (MAF < 0.01):**
Samples in: 11273
Samples removed: 65
Samples remaining: 11208

**Remove samples with excess accumulated PIHAT:**
Samples in: removed 11208
Samples removed: 11
Samples remaining: 11197

**Remove one in a pair of samples with PI_HAT > 0.1:**
Samples in: 11197
Samples removed: 457
Samples remaining: 10740

**Offspring**

**Markers and samples at beginning of module:**
Markers start: 637164
Samples start: 5577

**Remove markers not surviving QC in both parents and offspring:**
Markers in: 637164
Markers removed: 2426
Markers remaining: 634738

**Remove samples with missingness rate > 2%:**
Samples in: 5577
Samples removed: 33
Samples remaining: 5544

**Remove samples with HET excess > 4SD using common autosomal markers (MAF > 0.01):**
Samples in: 5544
Samples removed: 4
Samples remaining: 5540

**Remove samples with HET excess > 4SD using rare autosomal markers (MAF < 0.01):**
Samples in: 5540
Samples removed: 38
Samples remaining: 5502

**Remove samples with excess accumulated PIHAT:**
Samples in: removed 5502
Samples removed: 7
Samples remaining: 5495

**Remove one in a pair of samples with PI_HAT > 0.1:**
Samples in: 5495
Samples removed: 86
Samples remaining: 5409

## Module 5: Preparation for phasing and imputation

**Samples and markers into module:**
Samples in: 17742
Markers in: 647785

**Remove markers not passing QC for both offspring and founders:**
Markers in: 647785
Markers shared: 634738
Markers removed: 13047
Markers remaining: 634738

**Remove markers above chr 23:**
Markers in: 634738
Markers removed: 1374
Markers remaining: 633364

**Set mendelian errors to missing:**
Mendelian errors zeroed: 186590

**HRC harmonizing:**
Markers in: 633364

Marker chromosomes changed: 0
Marker positions changed: 0
Marker strand flips: 37109
Marker allele flips: 575573
Markers excluded (not in HRC): 65089
Markers after exclusion: 568275

**Number of markers per chromosome sent to phasing:**

| Chromosome | Number |
|---|---|
| 1 | 45268 |
| 2 | 46161 |
| 3 | 38636 |
| 4 | 35311 |
| 5 | 33424 |
| 6 | 39844 |
| 7 | 31182 |
| 8 | 29286 |
| 9 | 24176 |
| 10 | 28150 |
| 11 | 28293 |
| 12 | 27051 |
| 13 | 19613 |
| 14 | 18290 |

15

17219

16

18758

17

17044

18

16219

19

13131

20

13854

21

7705

22

8191

X

11469