# QC of the genotype sample IMAGEN (Released in July and September 2009)

March 17, 2010

## Preamble

This document describes the work of Ch Lalane, V Frouin JB Poline, A Lourdusamy and G Schumann. A. Lourdusamy and we proposed different approaches for the QC of genotype data. The process described here was chosen following examples from the literature (eg. [MGW+07]) and after several discussions with A Lourdusamy and G Schumann. The corresponding data are released on https://imgen.cea.fr server [Browse Genetic].

## 1 Introduction

The files 23juillet2009.zip et 24 septembre2009.zip were downloaded from https://imagen.cea.fr. From the files Final_ok.txt and XXX_succes.txt a binary file is created with plink. From the original files we got 705 individuals; we left apart 3 individuals with genotyping rate lower than 95%. From the information available to date, 28 individuals out of the 705 have non known sex.

We give in this short report the 3-step process applied to the data:

1. Study of the population homogeneity with Structure [PSD00]
2. Study of potential relatedness between individuals with MDS
3. Study of the reaining outlers with iterative PCA [PPP+07]

Only synthetic results are reported here. A study of the robustness of the parameters of the methods used are reported elsewhere and will be available soon (eg. effect of the number of SNPs or number of clusters used for the Structure step). Finally a few figures are given in Table 1 at the end of this document.

For the rest of the analysis we kept the 705 indivuduals and refer it as the Image Sample IS

## 2 Filtering

**Selection of a set of SNP in linkage equilibrium.** We select such SNP using pairwise correlation estimated with plink to get 7922 SNP.

```
plink parameters : --indep-pairwise 50 10 0.02
```

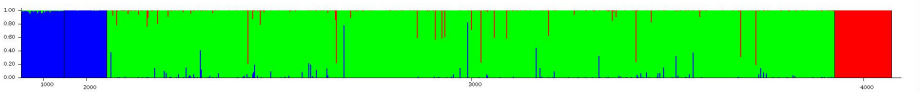


Figure 1: Output plot from code Structure. The bar plot of the posterior probabilities. CEU individuals are in green and outliers appear in non green color bars.

## 2.1 Population homogeneity of the sample

We used these SNP to collect data from IS and Hapmap; from Hapmap we only kept independent individuals from CEU, JPT+BJI and YRI using the hapmap files released with Plink release (Hapmap.r23). Those data were merged into a single file with 705(IS) + 90 (CEU) + 45 (JPT) + 45 (BJ) + 90 (YRI) individuals. We formated a file for *Structure* code with LOCPRIOR set for each ethnica origin (4 different labels). From the 7922 snp yielded by the previous filtering, we got 7060 snp that were in common with between IS the data set from hapmap. The following of the processing considered those 7060 SNPs.

We used the code Structure from Pritchard *et al.* [PSD00]. The main parameters are :

- K (number of class) : 3
- LOCPRIOR : 1. For each individual we set a LOC parameter coding for CEU, JPT, BJI and YRI. The individual from IS are coded CEU.
- Burnin and Iteration numbers : 10.000 and 10.000

The stochastic code was run 5 times with different seeds.

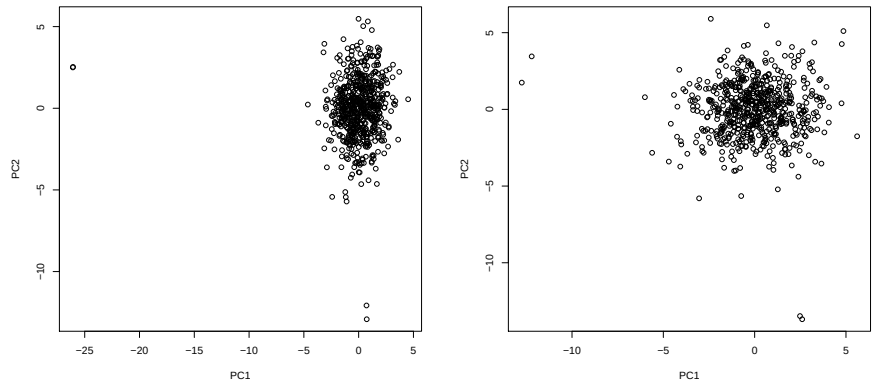
On figure 1, the population a priori labeled as JPT+BJI and YRI appear in green and red in the K=3 clustering result ; their posterior probability is consistent with the knowledge we have on the Hapmap population. Several individuals from IS showed posterior probability with low values. We chose a threshold value of 0.95 and this produced the same set of individual subgroup. We kept this group apart that consists in 74 individuals.

## 2.2 Outlier detection: MDS

We first applied an MDS (5 components, euclidean distance) on the $N = 631$ subjects remaining $(705 - 74)$. Subjects coordinates in the space defined by the first two dimensions are shown in Figure 2 (a). Two pairs of subjects appear to lie at the extreme of the PC1 and PC2 scale and suggest that individuals closely ressemble one to the other within each pair. These are subjects 6210895 and 94180305 (pair1: PC1 coordinate $< -20$), and 97096149 and 35587656 (pair2: PC2 coordinate $< -10$).

The pair 6210895 and 94180305 consists in very related samples. A thorough comparison of the genotypes showed that only 59 of the 582892 SNPs have a different value but the SNPs with missing data were different. We suspect that one same sample was measured twice and this invalidate the genetic data of both 6210895 and 94180305 individuals.

We filtered out only the two individuals from pair1. We removed subjects




(a) (b)

Figure 2: Result for the euclidean MDS.

from pair1 and checked that the resulting MDS looks less distorded (Fig. 2 (b)).

## 2.3 Outlier detection (iterative PCA)

Following [PPP+07] (p. 904), we applied a "principal components analysis to genotype data to infer continuous axes of genetic variation. Intuitively, the axes of variation reduce the data to a small number of dimensions, describing as much variability as possible (...)". We applied an iterative PCA (R, version 2.10) whereby outliers, defined here as those individuals located at more than 4 SD of the mean PCA scores on one of the first 20 dimensions, are removed at each stage, considering 10 iterations. Our results indicated that 9 individuals should be considered as potential outliers following this model. They are highlighted in red in Figure 3. Only one iteration has been required.

Finally, we ended up with a set of $N = 620$ subjects, with the following $74 + 2 + 9 = 85$ subjects removed. The final list is available on the server (file QClist620.txt) and genotype data on the server are subset'ed accordingly.

**74 subjects removed at step1:**

```
000011834339, 000036007181, 000000724314, 000007478069, 000000602062,
000013025821, 000061589239, 000033925730, 000034343813, 000023194232,
000012967847, 000001330914, 000051829237, 000022531572, 000067630794,
000041322973, 000028250707, 000077689671, 000020796903, 000029708729,
000079591035, 000009740318, 000030695561, 000006721723, 000057135861,
000043217738, 000057133727, 000029280369, 000097084811, 000019025504,
000040686493, 000039711869, 000006000160, 000035524006, 000084104893,
000001283761, 000020667836, 000074104786, 000082922411, 000046161596,
000067040028, 000038690448, 000032581973, 000007062184, 000053744289,
000052774607, 000027981638, 000055619168, 000047110302, 000083490451,
000009252997, 000023063758, 000037740824, 000082621571, 000047028244,
```

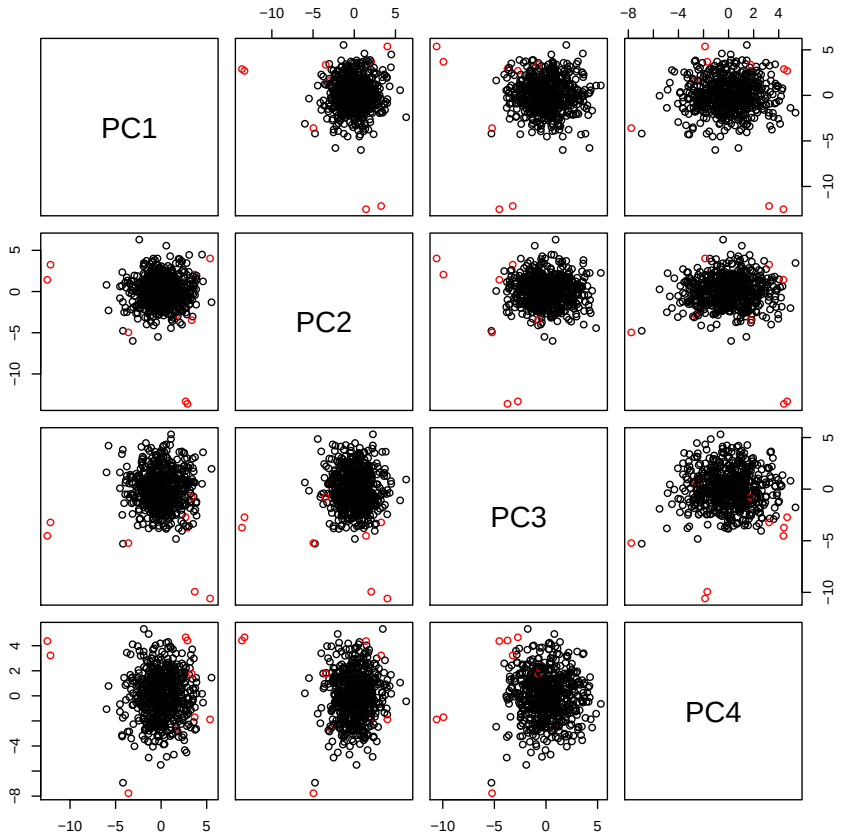



Figure 3: Plot of subjects scores in the subspaces defined by the first four dimensions of the PCA.

000079977505, 000039923280, 000031733662, 000054198873, 000012288410, 000022698017, 000004032825, 000077729708, 000018494678, 000084158838, 000058140414, 000034754250, 000099104307, 000079206781, 000057348284, 000043813881, 000041213271, 000002371789, 000000556983

**2 subjects removed at step2:**

00006210895, 000094180305

**9 subjects removed at step3:**

000084838602, 000099550415, 000065032538, 000099604669, 000034268536, 000056896962, 000097096149, 000035587656, 000065334806

Table 1: Descriptive Statistics by Exclusion

| | N | No | Yes | Test Statistic |
|---|---|---|---|---|
| | | $N = 620$ | $N = 85$ | |
| Gender : female | 677 | 52% (311) | 55% ( 44) | $\chi^2_1 = 0.24,\ P = 0.625$[1] |
| Age | 671 | 14.1 14.4 14.7 | 14.1 14.5 14.8 | $F_{1,669} = 0.05,\ P = 0.823$[2] |
| Father : non-caucasian | 699 | 0% ( 3) | 55% ( 46) | $\chi^2_1 = 338.63,\ P<0.001$[1] |
| Mother : non-caucasian | 699 | 1% ( 9) | 57% ( 47) | $\chi^2_1 = 302.04,\ P<0.001$[1] |
| Centre : London | 694 | 16% (96) | 17% (14) | $\chi^2_7 = 17.23,\ P = 0.016$[1] |
| Nottingham | | 16% (95) | 10% ( 8) | |
| Dublin | | 14% (83) | 7% ( 6) | |
| Berlin | | 12% (74) | 12% (10) | |
| Hamburg | | 13% (82) | 14% (12) | |
| Mannheim | | 12% (72) | 21% (18) | |
| Paris | | 10% (64) | 18% (15) | |
| Dresden | | 7% (44) | 1% ( 1) | |
| language : de | 693 | 45% (272) | 49% ( 41) | $\chi^2_2 = 6.28,\ P = 0.043$[1] |
| en | | 45% (274) | 33% ( 28) | |
| fr | | 10% ( 63) | 18% ( 15) | |

$_a\ b\ _c$ represent the lower quartile $a$, the median $b$, and the upper quartile $c$ for continuous variables.
$N$ is the number of non–missing values.
Numbers after percents are frequencies.
Tests used:
[1]Pearson test; [2]Wilcoxon test

# 3 Lists comparison

The Table 1 sums up descriptive statistics for the excluded and non excluded groups.

# References


[MGW+07] Amanda J Myers, J. Raphael Gibbs, Jennifer A Webster, Kristen Rohrer, Alice Zhao, Lauren Marlowe, Mona Kaleem, Doris Leung, Leslie Bryden, Priti Nath, Victoria L Zismann, Keta Joshipura, Matthew J Huentelman, Diane Hu-Lince, Keith D Coon, David W Craig, John V Pearson, Peter Holmans, Christopher B Heward, Eric M Reiman, Dietrich Stephan, and John Hardy. A survey of genetic human cortical gene expression. *Nat Genet*, 39(12):1494–1499, Dec 2007.

[PPP+07] Alkes L Proce, Nick J Patterson, Robert M Plenge, Michael E Weinblatt, Nancy A Shadick, and David Reich. Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, 38(8):904–909, 2007.

[PSD00] J. K. Pritchard, M. Stephens, and P. Donnelly. Inference of population structure using multilocus genotype data. *Genetics*, 155(2):945–959, Jun 2000.