Joint calling of Infinium Expanded Multi Ethnic Genotyping Array (MEGA^{EX}*) and QC summary on GCATcore sample

Iván Galván-Femenía[1], David Piñeyro[2], Anna Carreras[1], Laia Ramos[2], Xavi Duran[1], Raquel Pluvionet[2], Susanna Aussó[2], Lauro Sumoy[2] and Rafael de Cid [1]

1 Genomes for Life - GCAT lab Group, Program of Predictive and Personalized Medicine of Cancer (PMPPC), Germans Trias i Pujol Research Institute (IGTP), Crta. de Can Ruti, Camí de les Escoles s/n. Can Ruti Campus, 08916 Badalona, Spain

2 High Content Genomics and Bioinformatics Unit, Program of Predictive and Personalized Medicine of Cancer (PMPPC), Germans Trias i Pujol Research Institute (IGTP), Crta. de Can Ruti, Camí de les escoles s/n. Can Ruti Campus, 08916 Badalona, Spain

* MEGA_Consortium_v2_15070954_A2

Content

Genome Studio Analysis

Loading files

Load the intensity files (.idat) into Genome Studio with the sample sheet and manifest files.

- The .idat files are the raw data files from the MEGA provided by the facility that performed the genotyping.
- The sample sheets are .CSV files that contain sample information, such as plate ID, cell ID, gender and so on.
- Manifest file contains information about each probe, such as genomic position, sequence, strand, etc. It exists in text tab delimited format (human readable) or binary format (to input to GenomeStudio). The manifest file version used was "MEGA_Consortium_v2_15070954_A2".

Genotype clustering and QC

Samples analyzed to create this cluster file: GCATcore (plates 1-60)
Included samples: 5459 GCATcore samples + 177 Hapmap CEPH samples + 60 negative controls

Output> MEGA_FINAL_Cluster_File_QC

All the following QC criteria were automatically applied using own Python and R scripts.

QC on Sample Call Rate

Exclude samples with Call Rate < 0.94.
61 GCAT samples and all Negative controls (60) were excluded.

QC on SNPs located in haploid genome

Male X chromosome:
Zero out SNPs with AB Freq (heterozygous freq) > 0.8.
Manually inspect AB Freq >= 0.28.
Y chromosome:
Manually inspect all SNPs (7216).
MT chromosome:
Manually inspect all SNPs (>800).

QC by GenTrain Score and ClusterSep

Zero out SNPs with GenTrain < 0.67.
Manually inspect from 0.67 to 0.7.

Zero out SNPs with cluster separation <= 0.425.
Manually inspect from 0.4 to 0.45.

QC by Mendelian errors and replicate errors

Zero out any SNP with Mendelian or replicate errors.

QC by other criteria

AB Freq:

Zero out SNPs with AB Freq > 0.575.
Manually inspect from 0.55 to 0.575.

Call Freq:

Zero out all SNPs with Call Freq < 0.5.
Zero out SNPs from a diploid chromosome (excluding X, Y and MT) with Call <= 0.8.
Manually inspect X SNPs from 0.5 to 0.8.

AA T Deviation (AA cluster Theta value deviation).

Zero out >= 0.05.

BB T Dev.

Zero out >= 0.05.

by AA R Mean (AA cluster R mean value).

Zero out <= 0.2.

by AB R Mean.

Zero out <= 0.2.

by BB R Mean.

Zero out <= 0.2.

OUTPUT > excluded SNPs

TOTAL No Call (zeroed) SNPs: 257647.
TOTAL SNPs with GenTrain Score < 0.7: 211794.
TOTAL Edited (zeroed or manually edited) SNPs: 287168 out of 2036060 SNPs.

Post Genome Studio analysis.

Previous to QC PLINK we perform mapping and conversions steps to be sure of the position, the reference and the DNA strand. We conserve both annotation files.

Exporting from GenomeStudio

Manifest file remapping:

From the MEGA EX manifest file we observed 28949 non-mapped probes, 32542 positions (SNPs) interrogated by more than a single probe and 65280 probes interrogating the same genomic position as they were the exact same sequence. For these reasons, we decided to remap the entire set of probes. To perform the remapping process to the hg19 and hg38 human genome assemblies, we used own Python scripts to collect and format probe sequences and bowie2 (version 2.3.2, options: -f -N 1 --end-to-end -k 20) to align probes and define their genomic positions. Finally, we used own Python scripts to collect real genomic positions, as well as discard non-aligning and multi-mapping probes.

Reporting PLUS strand allele for hg19 and hg38 human genome assemblies:

GenomeStudio software is only capable of generating PLINK reports for either TOP or Forward (based on dbSNP) alleles. In order to transform our genotype callings to PLUS strand (used in downstream analysis) and also to correct the genomic positions for each SNP (based on our manifest file remapped), we used an in-house developed Python script. As input, it uses data exported from GenomeStudio using "PLINK input report plug-in" (v. 2.1.3. Illumina, Inc.) reporting TOP strand alleles. Additionally, this script removes all the multi-mapping and non-mapped probes. After processing, final .ped and .map files are produced, containing genotype calls in the PLUS strand and genomic positions corrected. This files are generated for both, hg19 and hg38 genome assemblies. As a result of removing multi-mapping and non-mapped probes, from the original 2036060 probes present in the array, 2000071 and 1987046 SNPs were reported for hg19 and hg38, respectively.

QC0 Mapping Array Data and Convert all SNPs to HG38 plus strand

We import from GS genotype a report data references in **forward strand format based on illumina annotation**.

We generate a *final report* genotype data references based on **GRch38/hg38 plus strand genotypes:**

We generate two set lists

- QC0.1 MAPPED VARIANTS in HG38 plus strand annotated reference
- QC0.2 UNMAPPED VARIANTS

Those file are used as references for ALL further analysis

We convert PLINK exported filed from GS to bed file

> -make bed --out GCAT_pl_1_60 (bedhg38plus)

INPUT >

GS Forward strand:

CUSTOM_01_Pl_1-60_hg19_PLUS.map and .ped
GCAT_pedCEPH.csv

Status: plates 1 to 60
5,696 individuals and 2,036,060 mapped SNPs

OUTPUT>

GCAT_pl_1_60_QC_0.bed

Status: plates 1 to 60
5,575 individuals and 1,987,046 mapped SNPs

Removed SNPs 48,816 multi-aligned SNPs
198 no aligned SNPs
Removed Samples 60 negative controls samples
61 Call rate = 0 samples

GCAT_pl_1_60_Unmapped_positions.txt
Contains list ID of Unmapped positions **(410 SNPs)**

* **GCAT_pl_1_60_Unmapped_positions.csv** with number of Unmapped SNPs
Unmapped positions are updated as chromosome 0

GCAT_pl_1_60_Mapped_positions.txt
Contains list ID of mapped positions, with both position HG38 (GRch38) and HG19 (GRch37)

QC1. Exclude Probes duplicates and SNPs 0 alleles

We check for any SNP duplicate probe (by chromosome position hg38), and we identify and remove lower quality genotypes. Finally, we generate a bed with the BEST UNIQUE and MAPPED HG38 Genome Variants.

QC scoring is performed with the GenTrain score (GT) and the best GT is conserved.

ALL GenTrain scores and duplicated SNPs list are generated by Genome Studio analysis

All SNPs with 0 alleles were excluded

INPUT>

***SNP_Table_GCAT.txt*. Generated by Genome studio**
Contains Gen train Score for All SNPs. Generated by Genome studio. ***2,036,060 SNPs***
***MEGA_Original_repeated.txt*. Generated by Genome studio**
Contains duplicated SNPs Probes **65,280 probes**

OUTPUT>

Filter_out _duplicates_probes_GCAT_pl_1_60_QC_1.txt TXT file with removed duplicated SNPs probes 32,614 SNPs (1.6%)
SNPs_zero_alleles.txt
Contains SNPs with all zero alleles 226,584 SNPs

GCAT_pl_1_60_QC_1.bed
Bed file with best Unique SNPs Probes: 1,728,123

Exclusion file with SNP designs

SNP designs (G>C and A>T) lead to missleading strand definition during Imputation step. Since combination **(G>C and A>T)** we can generate flipped positions (assign and change a T (A, G o C) in the opposite strand, and create artefactual imputed outputs, we will create a txt file to exclude those genomic variants previous to IMPUTATION STEP. Starting from the **first and second alleles columns** from the **bim file**, select the SNPs with the following conditions: the first allele is a C and the second allele is a G, first allele is a G and the second is a C, the first allele is an A and the second a T and finally the first allele is a T and the second an A.

All these variants are excluded from final analysis, and they will be incorporated later on the association analysis from the IMPUTE output file.

Table QC1.1: Percentage of SNPs by type of polymorphism

	A/C	A/G	A/T	C/G	C/T	G/T	D/I
%SNP	9	35	4	6	36	9	1

INPUT>

GCAT_pl_1_60_QC_1.bed

Output >

AT_CG_sites.txt

This file contains the SNP ID of 140,071 (G>C and A>T)

QC2. Exclusion of Bad variants

We perform QC on ALL chromosomes (diploid and haploid), looking for bad cluster scores and failed SNPs (GTscores =0).

We exclude *all variants* in the Array with a defective clustering score, based on the **GenTrain** and **ClusterSep** metrics. Those metrics, generated in the raw data analysis will be used as standards to discard problematic or BAD variants.

(Illumina's genotyping solutions use the GenCall software application to automatically cluster, call genotypes, and assign confidence scores. The GenCall application incorporates a clustering algorithm (GenTrain) and a calling algorithm. Genotyping calls for a specific DNA are made by the calling algorithm, relying on information provided by the GenTrain clustering algorithm.)

- **GenTrain.** Measures the shapes of the clusters and their relative distance to each other with a statistical score (varies from 0–1).
- **Cluster Sep.** Measures the separation between the three genotype clusters in the theta dimension and varies from 0–1. Evaluate individual SNPs for overlapping clusters, starting with those having low Cluster Sep. If clusters are not well separated, or overlapping, should be zeroed.

We retain all SNPs with: Gen Train >0.7 and ClusterSep >0.4

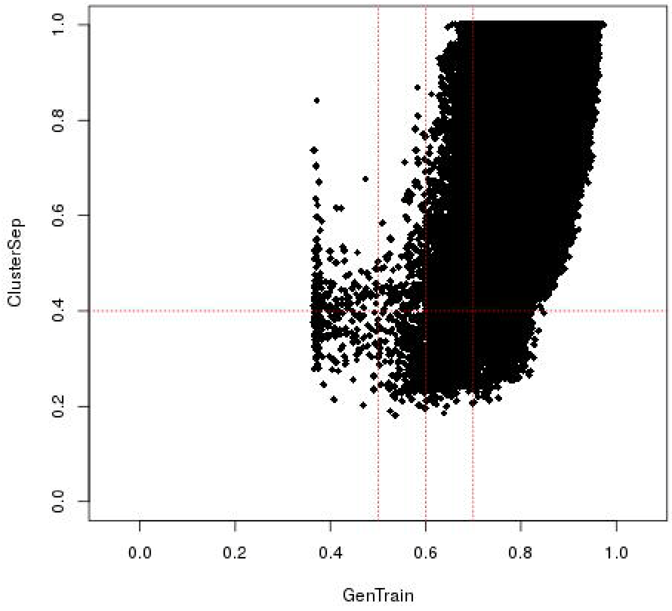


Figure QC2.1: Scatter plot of Gen Train and ClusterSep scores

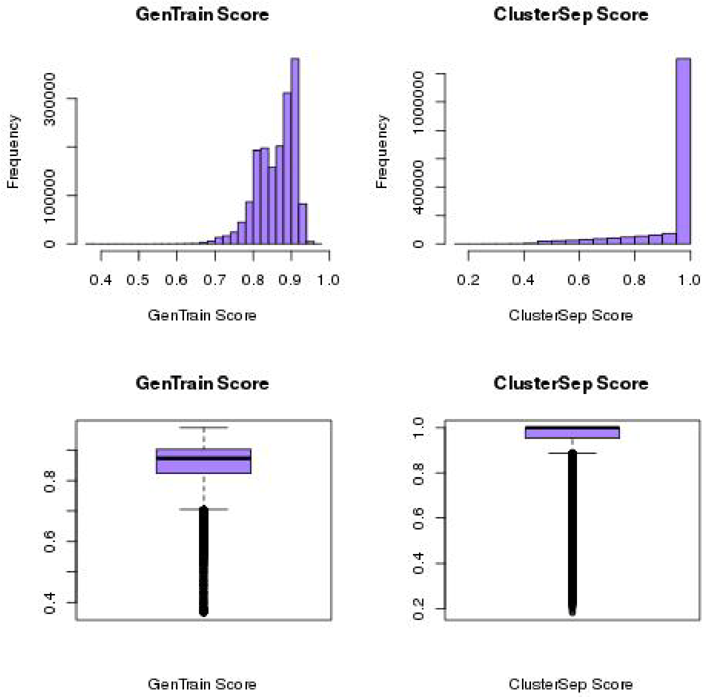


Figure QC2.2: Initial GenTrain and ClusterSep score distribution

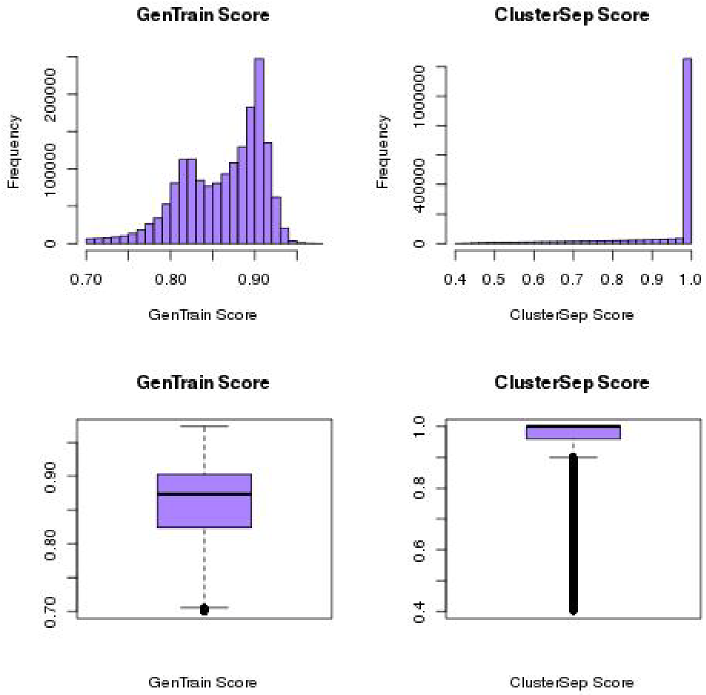


Figure QC2.3. Final GenTrain and ClusterSep score distribution

Table QC2.1: Number of selected SNPs below the threshold for Gen Train and ClusterSep scores

	Threshold: 0.4	Threshold: 0.5	Threshold: 0.6	Threshold: 0.7
GenTrain	153	276	786	11,818
ClusterSep	2,648	25,259	72,396	139,781

INPUT>

GCAT_pl_1_60_QC_1.bed

OUTPUT>

GCAT_pl_1_60_QC_2.bed
Contains all unique and GOOD clustered (QC1 + QC2)

5,575 individuals and 1,715,385 SNPs

Filter_out _BAD_GCAT_pl_1_60_QC_2.txt

List **12,738** SNPs removed by Bad cluster / Failed scores

QC3. Missing Call rates

On **RACES POOLED** we analyze the Call rate to excluded poor sample and SNP calls.

Poor call rate, with high missing calls, for sample or SNPS, could be due to different reasons (SNP design, sequence effect, technical, DNA quality,). If not checked could lead to misleading conclusions in the final analysis. This effect is even worst in the case o rare variation enrichment arrays, as the MEGA array.

To define the cutoff to remove poor Call rates, we calculate the missing call rates for (1) each sample and for (2) each SNP from samples with different missing call rates; >10, >5 % , >2%, >1%

1. **Missing call rate per SNP**, over all samples **(including control samples**)
2. **Missing call rate per samples**, overall all SNPs

Table QC3.1 Number of samples and SNPs with missing call rate >=10, >=5 %, >=2%, >=1%.

	Missing>=1 %	**Missing>=2 %**	**Missing>= 5%**	**Missing>= 10%**
Sample	0	0	0	0
SNPs	15,550	8,150	4,988	2,802

Table QC3.2. Average Missing call rate on selected samples with different missing call rate

	Missing>=1 %	**Missing>=2 %**	**Missing>= 5%**	**Missing>= 10%**
Average Missing call rate SNP	0.9997571	0.9996996	0.9996429	0.9995509
Average Missing call rate sample	0.9992869	-	-	-

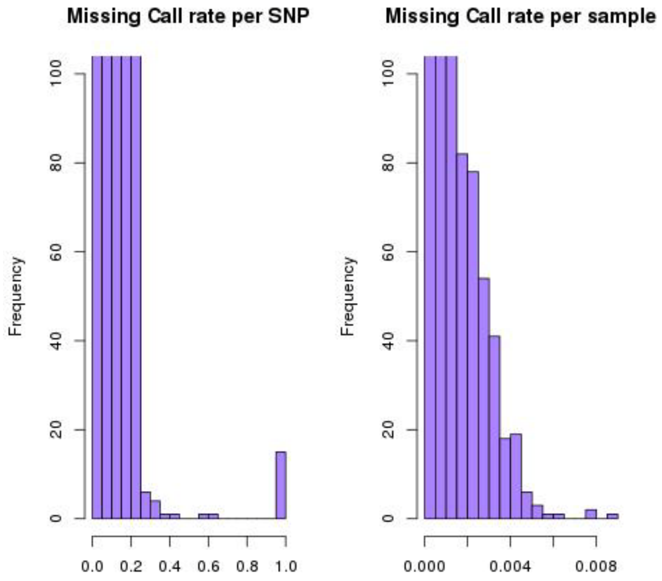


Figure QC3.1: Histogram of the missing call rate per SNP and sample

Based on recommendations and with these figures, we filter out samples with missing call rate >2% and SNPs with missing call rate >2%

INPUT>

GCAT_pl_1_60_QC_2.bed

Samples_CallRate_SexDiscordance_Pl_1-60.txt
This txt file contains a list of ID identifiers and call rate annotated by GenomeStudio (GS) analysis. This file is generated by GS QC Steps.

OUTPUT >

GCAT_pl_1_60_QC_3.bed
Contains all unique and GOOD clustered (QC1 + QC2) with ≤2% missing SNP and Sample (QC3)

Table QC3.2. **Table average missing call rate for classes**

Figure QC3.1 **Histogram of missing call rate for clases and average**

5,575 individuals and 1,707,235 SNPs
0.9997 missing call rate average on the final INCLUDED sample

QC4. Gender mismatch

<u>On RACES POOLED</u>, **<u>and MAF>0.1</u>** we analyze gender mismatch to identify and exclude problematic samples.

To check gender identity we observe X chromosomes *heterozygosis rates* and the *means of the intensities of SNP probes* on the X and Y chromosomes.

The expectation is that males and females fall into distinct clusters that differs markedly in X and Y intensities. This analysis is usually done during the GenomeStudio analysis.

To check for gender mismatch we use the PLINK function filtering for **MAF >0.1 to avoid rare variants effect.Gender information is automatically exported into the PLINK file, provided that it is available in the sample sheet.

```
> plink --maf 0.1 --check-sex --out
```

** --maf 0.1 Avoiding rare variants we eliminate problems with haployd genome and reduced heterogeneity regions thus mismatch gender calling

Samples_CallRate_SexDiscordance_Pl_1-60.txt

This txt file contains a list of ALL unfiltered SNPs with all ID identifiers and gender differences annotated by GenomeStudio (GS) analysis was used to compare. This file is generated by GS QC Steps. (190 samples; 61 GS=0 eliminated now in the QC, then 129 discordances (HapMap and GCAT)

After examination of LOH patterns, 19 samples were excluded (18 females / 1 male) because high inbreeding F values >0.20.

INPUT>

GCAT_pl_1_60_QC_3.bed

OUTPUT>

Filter_out _gender_mismatch_ GCAT_pl_1_60_QC_4.txt

This txt file contains a list of ID identifiers and gender differences annotated for exclusion: **19 individuals**

GCAT_pl_1_60_QC_4.bed

Contains all unique and GOOD clustered (QC1 + QC2) with ≤2% missing SNP and Sample (QC3) and gender mismatched removed (QC4)

<u>5,556 individuals and 1,707,235 SNPs</u>

QC5. Mendelian errors

We analyze non-mendelian inheritance patterns, parent-parent-child (PPC), to excluded problematic SNPs.

We analyze mendelian errors by MAF treshold, >=1% >5% >10%.

Table QC3.1 Number of SNPs with mendelian errors by MAF: >=1%, >=5%, >=10%.

	MAF>=1%	**MAF>=5%***	**MAF>= 10%**
SNPs	127	116	97

*Median GT / Cluster Sep for MAF>5% is 0.85 and 0.92

- We use CEPH reference included controls **GCAT_pedCEPH.csv**
- We exclude ALL SNPs with PPC>1 (but 532 > o =1 error)

INPUT>

GCAT_pl_1_60_QC_4.bed

SNP_Mendelian_Errors_Pl_1-60.txt

This txt file contains a list of ID identifiers with non-mendelian inheritance annotated by GenomeStudio (GS) analysis. This file is generated by GS QC Steps. (**1,398 discordances all SNPs with zero Gen Train Score)**

GCAT_pedCEPH.csv

This txt file contains a list of ID identifiers and Pedigree information for mendelian inheritance analysis.

OUTPUT >

GCAT_pl_1_60_QC_5.bed

Contains all unique and GOOD clustered (QC1 + QC2) mapped SNPs with <5% missing SNP and >5% missing (QC3) Sample, no gender mismatch (QC4) and mendelian SNPs (QC5)

Filter_out _Mendelian_pl_1_60_QC_5.txt

Contain list ID of SNPs excluded: **116 SNPs (MAF>=5%)**

5,556 individuals and 1,707,119 SNPs

QC6. Allele frequencies. Common Polymorphic variants.

On **RACES POOLED , from GOOD genotyped SNPs**, we can now identify frequency statistic and distribution of monomorphic SNPs (MAF=0 in any RACE group) and MAF>0 variants in our dataset

We will generate a figure and table statistics with

- CSVQC6. MAF REFERENCE ALLELE.
  List ID with frequency allele (A1 being the used Reference allele PLUS STRAND)*
  *PLINK reports the A1 columm as the minor allele frequency detected in the sample.

- Figure QC6 MAF STATISTICS
  Figure representation of frequency allele distribution in 6 classes: from 0, 0-0.1%, 0.1-1%, 1-5%, 5-10% , >10%

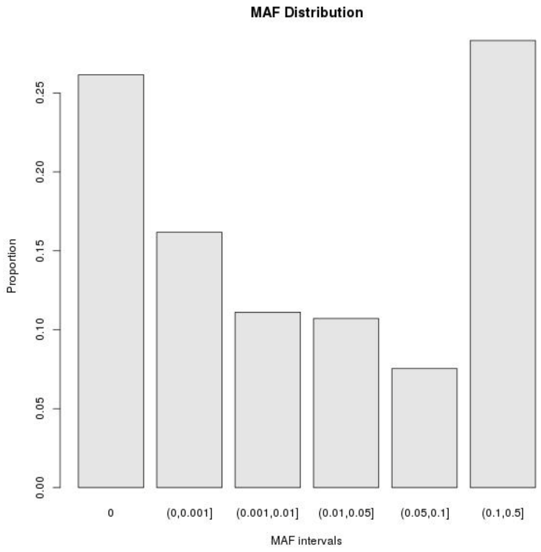


Figure QC6 MAF STATISTICS. Bar plot of the minor allele frequency distribution in 6 classes: from 0, 0-0.1%, 0.1-1%, 1-5%, 5-10% , >10%

- **446,367 SNPs are monomorphic**
- **722,475 SNPs have minor allele frequency shorter than 0.001**

On QC5 step we generate a SUBSET selection with common SNPs with MAF > 10%

INPUT >
GCAT_pl_1_60_QC_5.bed

OUTPUT >
descriptives
CSVQC6. MAF REFERENCE ALLELE
FigureQC6 MAF STATISTICS
GCAT_pl_1_60_QC_5_subset_MAF_10.bed SUBSET selection with common SNPs with MAF > 10%

5,556 individuals and 483,517 SNPs

QC7. Checking for duplicates and relatedness

ON RACE POOLED and AUTOSOMAL, MAF >0.40

HWE < 0.05 and CEPH controls were removed.

We analyze the Genetic relationships to test for duplicates and familiars. GCAT CORE is based on UNRELATED general population; a number of familiars are presents .

We use LOW LD autosomal common variants;> *plink –-remove controls.txt --maf 0.4 --chr 1-22 --hwe 0.05 --indep-pairwise 50 5 0.2 --out {indepSNP}.* Finally **32,228** SNPs are used for the analysis.

We calculate IBD (Method of moments (MoM) with PLINK --*genome*) and check for duplicates and relatedness.

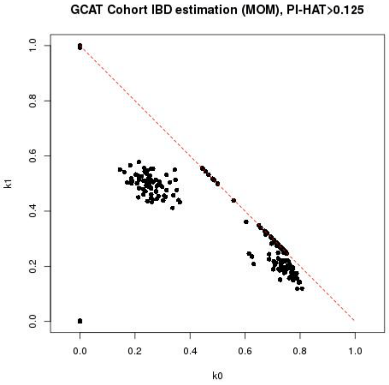


Figure QC7.1: Probability of sharing 0 IBD alleles (k0) vs probability of sharing 1 allele IBD. PI-HAT > 0.125 threshold for 3rd degree was applied

Based on a PIHAT threshold > 0.125 (3rd degree) we suggest 199 related pairs of individuals. These results based on PLINK are not accurate. Thus these relationships are identified but not removed from dataset.

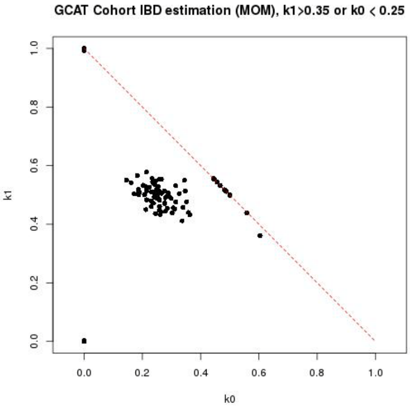


Figure QC7.2: Probability of sharing 0 IBD alleles (k0) vs probability of sharing 1 IBD allele (k1). k1 > 0.35 and k0<0.25 threshold for 1st and 2nd degrees was applied

Generate TXT files containing list of excluded samples

- Duplicates_GCAT_pl_1_60_QC_8.txt (6 re- typed, 1 by duplicate labeling error, same person different ID)
- 2ndegree_GCAT_pl_1_60_QC_8.txt (88 individuals)

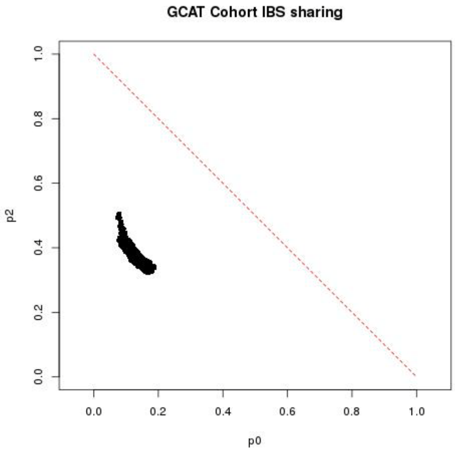


Figure QC7.3: Probability of sharing 0 IBS alleles (p0) vs probability of sharing 2 IBS alleles (p2).

INPUT > GCAT_pl_1_60_QC_5.bed

OUTPUT> GCAT_pl_1_60_QC_7.bed

Filter_out _ Duplicates_GCAT_pl_1_60_QC_7.txt
Contains the Sample list of **8** excluded Samples

Filter_out _2nd_degree_GCAT_pl_1_60_QC_7.txt
Contains the Sample list of **88** excluded Samples

Relatedness_type_pairs_2nd_degree_GCAT_pl_1_60.txt
Contains the type of relationship for the excluded Samples

5,283 individuals (+ 177 CEPH) and 1,707,119 SNPs
*3rd degree relationships identified are not excluded from this analysis

QC8. Stratification and Race mismatch

We analyze the Genetic relationships to test population stratification. CEPH controls were removed.

- GCAT CORE is a RACE POOLED sample;
  81,6% Caucasian, **17%** Hispano/Latino American and **1,4%** others: Black, Asian, Maghrebi Arabic.. etc.
  (code GCAT questionare: 1 caucasians, 2 black, 3 asian, 4 gypsi 5 maghrebi, 6latin, 7 others)

- GCAT CORE is a GEOGRAPHICALLY POOLED sample
  97,3% European, **2,7%** abroad.
  95,5% of the individuals from the GCAT Core were born in Spanish regions

We compute Multidimensional Scaling (MDS) analysis from a subset of SNPs (common SNPs with the 1000G phase III 1000Genomes project). Then we compare GCAT with non-Caucasians, and other European populations.

MDS was computed on GCAT sample after Q1-QC7 exclusions, on autosome chromosomes with independent SNPs LD>0.2 and MAF >0.1.

```
> plink -—remove controls.txt --maf 0.1 –chr 1-22 --indep-pairwise 50 5 0.2 --out {indepSNP} ----cluster --mds-plot 2
```

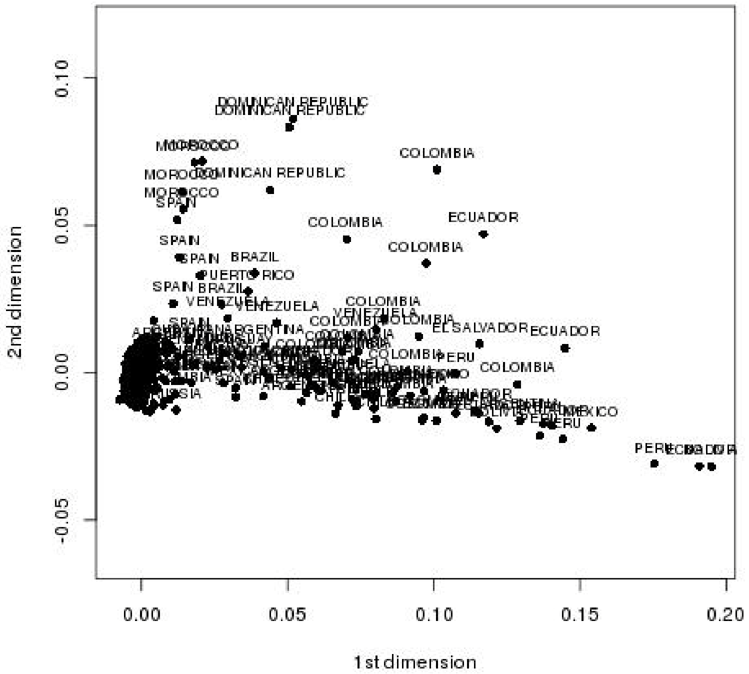


Figure QC8.2: First and second dimensions of MDS analysis of the GCAT Core. The label for each point is the born country self-reported

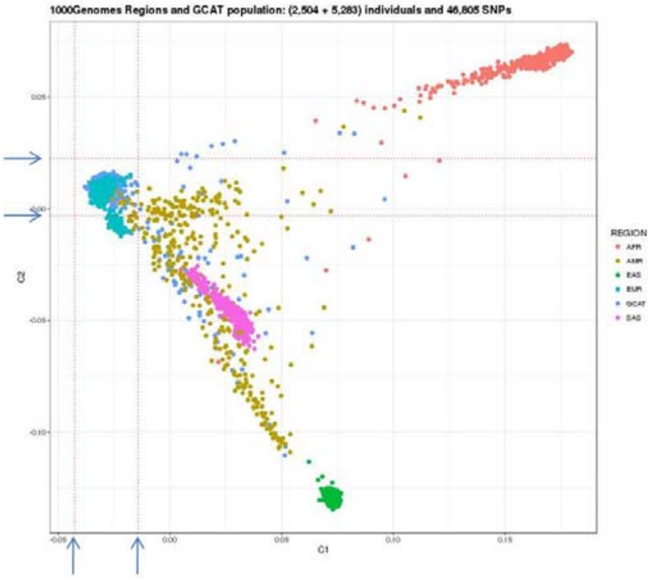


Figure QC8.2: First and second dimensions of MDS analysis of the GCAT Core and 1000Genomes populations. **96 Samples were excluded**

Scored samples were defined as outliers and excluded or considered for a separate analysis, considering Principal Dimension analysis (MDS) results we exclude samples from other populations by considering a threshold = **mean (C1) +- 2*sd (C1) and mean (C2) +- 2*sd (C2)**

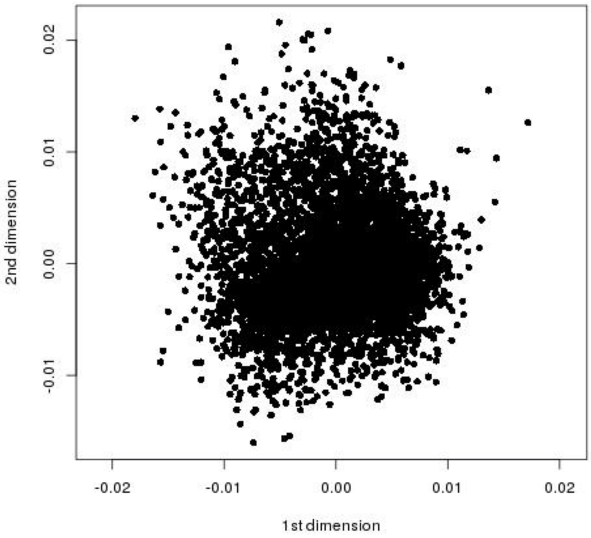


Figure QC8.3: MDS of the GCAT Core after threshold applied.

Table QC8.1 Self-reported ethnicity of the 96 excluded samples

	1	3	6	NA
counts	80	1	12	3

Table QC8.1 Self and father-mother reported ethnicity of the 96 excluded samples

	111	311	611	666	NA11	NANANA
counts	80	1	1	11	1	2

*1 caucasians, 2 black, 3 asian, 4 gypsi 5 maghrebi, 6latin, 7 others; NA not data available

Most of the excluded samples were self-reported as Caucasians (80) africans (1) and latins (12), but were born abroad SPAIN (Figure QC 8.2).

INPUT> **GCAT_pl_1_60_QC_7.bed**

OUTPUT> **GCAT_pl_1_60_QC_8.bed**

Filter_out_MDS_GCAT_pl_1_60_QC_8.txt
Contains the Sample list by homogenous **genetic distance>**
96 excluded

Samples by MDS
Filter_out_MDS_country_of_birth_GCAT_pl_1_60_QC_8.txt
Contains a table with the country of birth of the 96 excluded Samples by MDS

Filter_out_MDS_self_reported_ethnicity_GCAT_pl_1_60_QC_8.txt
Contains a table with the self-reported ethnicity of the excluded 96 Samples by MDS

5,187 individuals (+ 177 CEPH) and 1,707,119 SNPs

QC9. Heterozygosis and inbreeding outliers

On LOW LD AUTOSOMAL Variants and CEPH controls removed.

Heterozygosis excess or loss is an alteration of the Hardy-Weinberg equilibrium (HWE) laws.

Population admixture from two populations with different allele frequencies (natural or artificial) could relate a deficit of Heterozigosity. Non random mating or non random reproduction (another violation of the HWE assumptions) could also introduce biases

We use SNPs present in the autosome and independent SNPs with MAF >0.1

> plink --maf 0.1 --chr 1-22 --indep-pairwise 50 5 0.2 --out {indepSNP}

Heterozigosity rate

This metric was calculated by the formula: (N(NM)-O(Hom))/N(NM) where N(NM) is the number of markers non missing and O(Hom) is the number of observed homozygote markers.

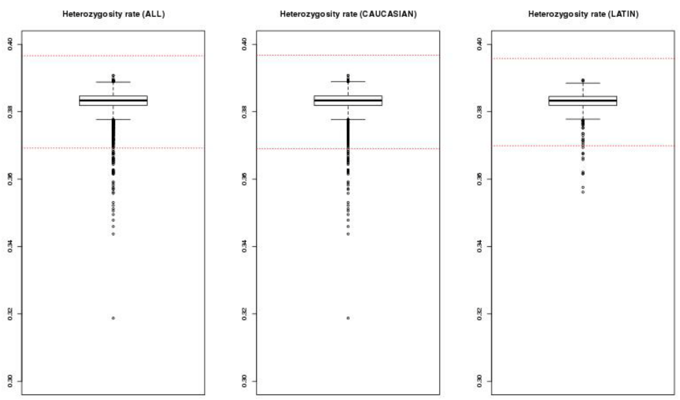


Figure QC9.1. Plot of Heterozigosity rates by RACE. Dashed lines show the mean+-4*SD thresholds

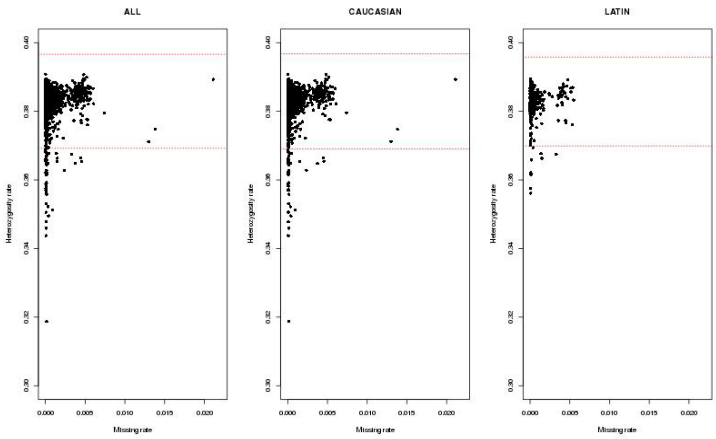


Figure QC9.2. Plot of Heterozigosity rates versus missing call rate by RACE. Dashed lines show the mean+-4*SD thresholds for the heterzygosity rate

We exclude all Samples SNPs with heterozygosity values out of mean +/-4SD:

Estimation of inbreeding coefficient

Deviations from HWE due to population structure are expected result in an excess of homozygote or a positive inbreeding coefficient estimate (1- (observed heterozigotes /expected heterozigotes).

We use SNPs present in the autosome and independent SNPs with MAF >0.1

> plink --maf 0.1 --chr 1-22 --indep-pairwise 50 5 0.2 --out {indepSNP}

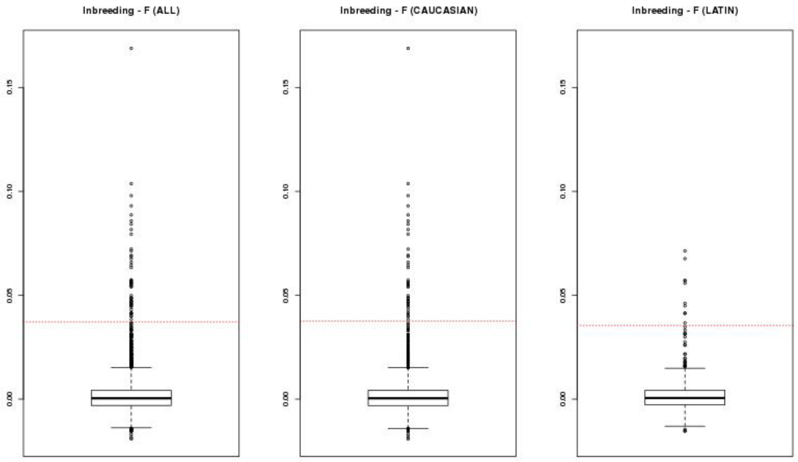


Figure QC9.2. Plot of Inbreeding coefficient F. Dashed lines show the mean+-4*SD thresholds

We exclude all Samples SNPs with Inbreeding coefficient values out of mean +/-4SD

IMPUT >

GCAT_pl_1_60_QC_8.bed

OUTPUT>

GCAT_pl_1_60_QC_9.bed

Incorporates all QC1-QC8 plus QC9 sample exclusions by Het and Inbreeding
Figure QC9.1 and QC9.2.

Filter_outHet_GCAT_pl_1_60_QC_9.txt
Contains the Sample list of **52** excluded Samples

Filter_outF_GCAT_pl_1_60_QC_9.txt
Contains the Sample list of 52 excluded Samples

5,135 individuals (+ 177 CEPH) and 1,707,119 SNPs

QC10. Hardy-Weinberg equilibrium (HWE) outliers

On **RACES POOLED, and ALL chromosomes (autosome and Sexual) and CEPH controls removed.**

We perform the exact HWE test to identify SNPs that deviate from HWE using the –hardy option on PLINK.

Because assumption of HWE, we use (1) unrelated subjects, (2) MAF>0.01 (3) p value= 0.05/ number of SNPs

We separate analysis by chromosomes (Autosomal-PAR, sexual).

We apply conservative GWAs thresholds.

Here we plot the results by frequency (Fig QC10.1), Quantile-Quantile (Fig QQ plot) (QC10.2), and Manhattan (Fig QC10.3) plot for Genome wide distribution in autosomal-Par chromosomes.

AUTOSOMAL-PAR CHROMOSOMES

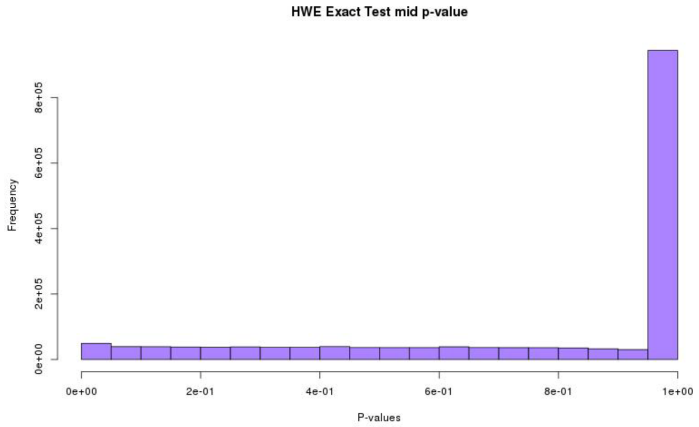


Figure QC10.1 Frequency P-values distribution

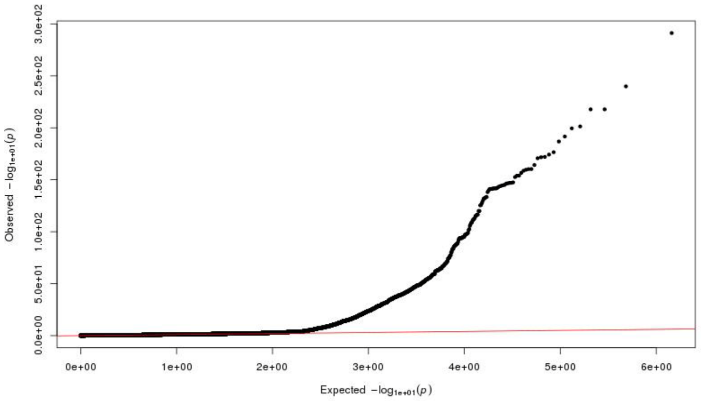


Figure QC10.2 QQ plot of HWE outliers

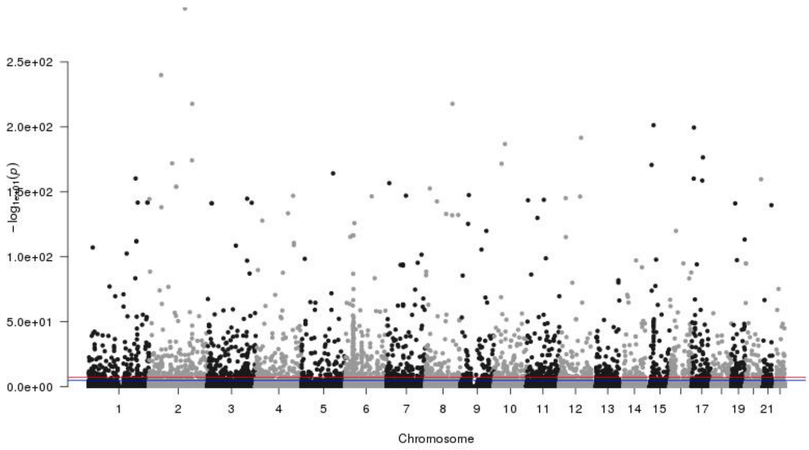


Figure QC10.2 Manhattan plot of HWE outliers

HWE for SNPs from the **PAR chromosome** is calculated with PLINK.

HWE for SNPS from the **X chromosome** is calculated with the R package “HardyWeinberg”

Y chromosome is composed for males and there are only one allele by individual. There are no tests for HWE in the Y chr.

- SNPs that deviate severely from HWE can be identified in the .hwe file

INPUT >

GCAT_pl_1_60_QC_9.bed

OUTPUT>

GCAT_pl_1_60_QC_10.bed

Incorporates all QC1-QC9 plus QC10 SNP exclusions

Filter_out HWE_GCAT_pl_1_60_QC_10.txt
Contains the SNPID list of **2,861** SNPs with Bonferroni correction

5,135 individuals (+ 177 CEPH) and 1,704,532 SNPs
*excluded SNPs deviates severely (Bonferroni 3 x 10^{-8}) from HWE

QC11. Allele frequencies statistics.

FINALLY FROM QC10 SAMPLE we generate a Figure and Table statistics of allelic frequencies in the selected sample

We selected individuals genetically homogenous group (MDS) and excluding NOT country of birth SPAIN (based on questionnaire survey).

Major group was considered born in SPAIN after MDS QC.8 filter applied, it remains **4,988** individuals

Table QC11.1 Self-reported ethnicity of the 4,988 homogeneous samples

	1	2	3	5	6	7	NA
counts	4,152	18	1	1	794	4	18

Table QC8.1 Self and father-mother reported ethnicity of the 4,988 homogeneous samples

	111	666	211	161	NANANA	116	1NANA	661	1NA1	Others
counts	4,100	780	16	14	13	11	8	7	6	33

*1 caucasians, 2 black, 3 asian, 4 gypsi 5 maghrebi, 6latin, 7 others; NA not data available

Even if 780 reported as latins, most of the self-reported as hispanic-latin were genetically classified as Caucasians and born in SPAIN,

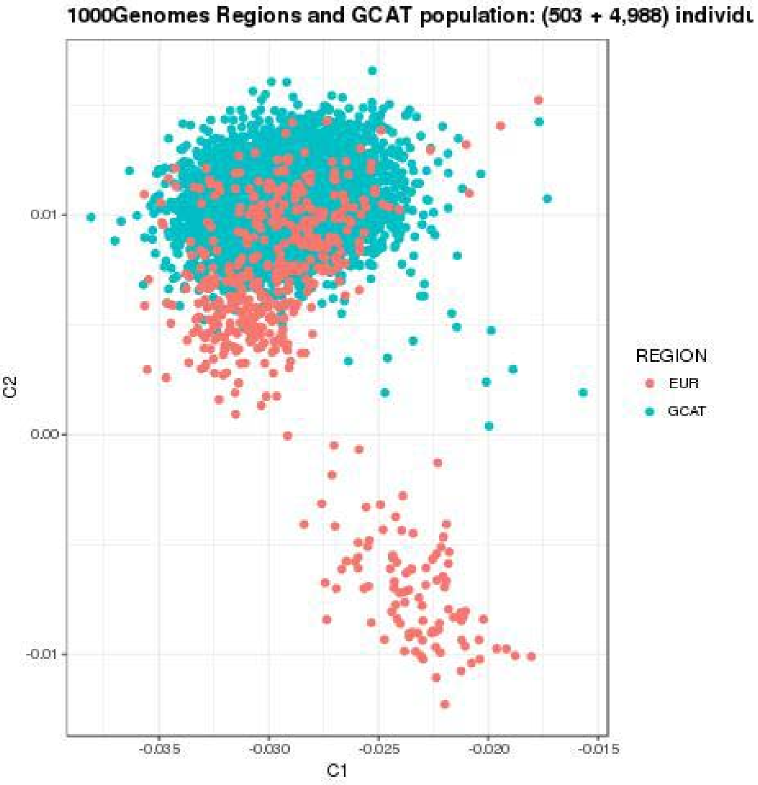


Figure QC11.1: MDS of the GCAT Core for Spanish self-reported after QC.8 filter applied.

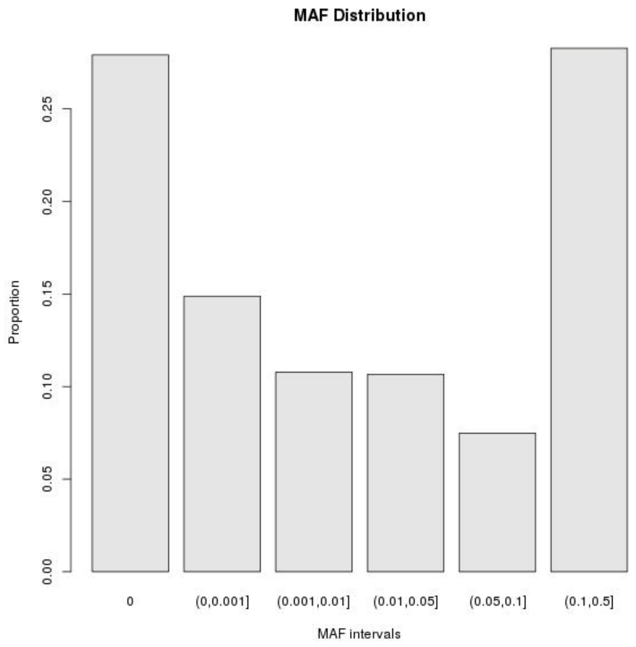


Figure QC11.2 MAF STATISTICS. Barplot of frequency allele distribution in 6 classes: from 0, 0-0.1%, 0.1-1%, 1-5%, 5-10% , >10%

Table QC11.3 MAF STATISTICS. SNP counts by minor allele frequency distribution in 6 classes: from 0, 0-0.1%, 0.1-1%, 1-5%, 5-10% , >10%

MAF	**0**	**(0-0.001]**	**(0.001-0.01]**	**(0.01-0.05]**	**(0.05-0.1]**	**(0.1-0.5]**
SNPs	475,843	253,659	183,776	181,749	127,555	481,950

INPUT > GCAT_pl_1_60_QC_10.bed

MDS plot + Only Spanish self-reported **4,988 individuals and 1,704,532 SNPs**

OUTPUT > descriptives

CSVQC11. MAF REFERENCE ALLELE
FigureQC11MAF STATISTICS

QC12. Genotype Consistency. Concordance and Discordance

Genotyping errors rates could be estimated from duplicate discordances rates.

Concordance should be estimated on MAF classes, since low MAF give more probability to observe a Homozygote and thus concordance than higher MAF SNPs.

1.- Concordance between GCAT controls

We compare all the pairs of duplicates of the **29 GCAT controls**. A total of **717 pairs** of replicates were used. For these pairs, we establish a threshold to eliminate a large fraction of SNPs with high error rates, while retaining a large fraction with low error rates. This threshold is based on the binomial distribution (GWASTools, R Bioconductor, *duplicateDiscordanceProbability*( ) function). With a filter threshold of >5 discordant SNPs, **this retains > 99% SNPs with an error rate < 1e-3.** This threshold eliminates **24, 452 SNPs** of 1,704,532 SNPs from the GCATCore.

discordant calls	error=1e-05	error=1e-04	error=0.001	error=0.01	SNPs
>0	0.0141	0.1324	0.7584	1.0000	51755
>1	0.0001	0.0092	0.4148	1.0000	51755
>2	0.0000	0.0004	0.1710	0.9999	45721
>3	0.0000	0.0000	0.0557	0.9996	45716
>4	0.0000	0.0000	0.0149	0.9984	40986
>5	**0.0000**	**0.0000**	**0.0034**	**0.9951**	**24452**
>6	0.0000	0.0000	0.0007	0.9872	24059
>7	0.0000	0.0000	0.0001	0.9711	23268

Table QC12.1 Probability of observing more than the given number of discordant calls in 717 pairs of duplicate samples.

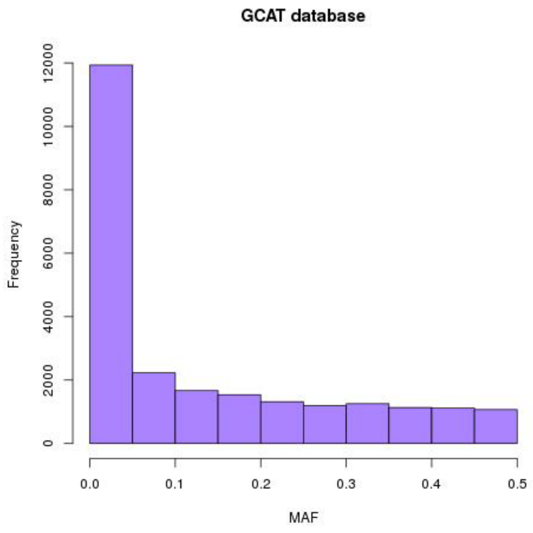


Figure QC12.1 MAF of the 24,452 discordant SNPs using controls in the GCAT database.

2.- Concordance between HAPMAP samples and GCAT controls.

GCAT controls were compared with HAPMAP Phase II + III (4,031,093 SNPs):

(ftp://ftp.ncbi.nlm.nih.gov/hapmap/genotypes/2010-08_phaseII+III/)

HAPMAP data is built at the ncbi_b36 (hg18) assembly. We used liftOver tool from UCSC to change the variant coordinates to hg38 assembly.

http://genome.sph.umich.edu/wiki/LiftOver#Lift_genome_positions

We intersect HAPMAP database and GCATCore by chromosome and position. GCAT share **479,370 SNPs** with HAPMAP database. We compared **27 controls** from GCAT with HAPMAP individuals. 687 duplicated pairs were compared. With a filter threshold of >5 discordant SNPs, **this retains > 99% SNPs with an error rate < 1e-3.** This threshold eliminates **10,884 SNPs**.

discordant calls	error=1e-05	error=1e-04	error=0.001	error=0.01	SNPs
>0	0.0135	0.1272	0.7434	1.0000	26130
>1	0.0001	0.0084	0.3940	1.0000	18303
>2	0.0000	0.0004	0.1565	0.9999	16279
>3	0.0000	0.0000	0.0490	0.9993	12873
>4	0.0000	0.0000	0.0126	0.9976	12441
>5	**0.0000**	**0.0000**	**0.0027**	**0.9927**	**10884**
>6	0.0000	0.0000	0.0005	0.9816	6837
>7	0.0000	0.0000	0.0001	0.9599	6658

Table QC12.2 Probability of observing more than the given number of discordant calls in 687 pairs of duplicate samples.

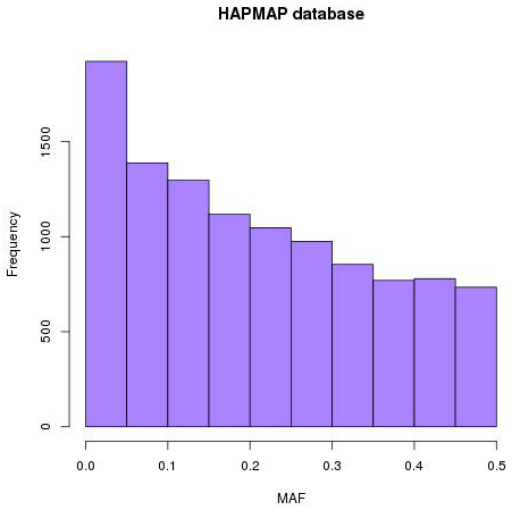


Figure QC12.2 MAF of the 10,884 discordant SNPs using controls in the HAPMAP database.

QC13. Allelic frecuencies with 1000Genomes

We compare minor allele frequency consistence with 1000Genomes project (CEU, IBS, TSI, GRB, FIN pop., n=503) and GCATcore, by the major group (**4,988** spanish individuals). A total of **1,428,126 SNPs** (Autosomal, X and Y chr) were compared by chromosome and position with the 1000Genomes project. We exclude SNPs with absolute diference of the ratio(MAF 1000Genomes /MAF GCATCore)> 2. This threshold eliminates **18,014 SNPs with MAF>0.01**

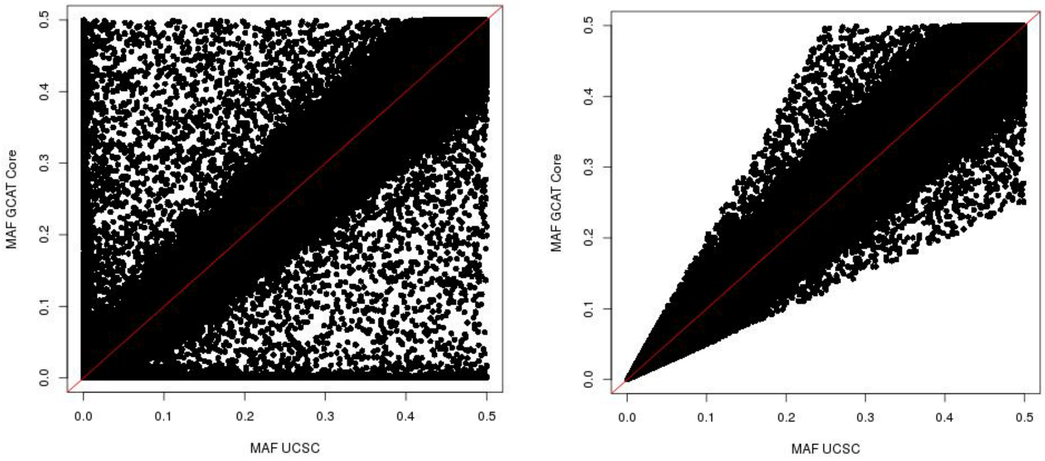


Figure QC13.1 Scatterplot of the MAF from 1000Genomes and the GCATcore. A total of **1,428,126 SNPs** were compared. The red line show the y=x identity line. Left: scatterplot without absolute difference of ratios >2 filter criteria. Right: Scatterplot with filter criteria.

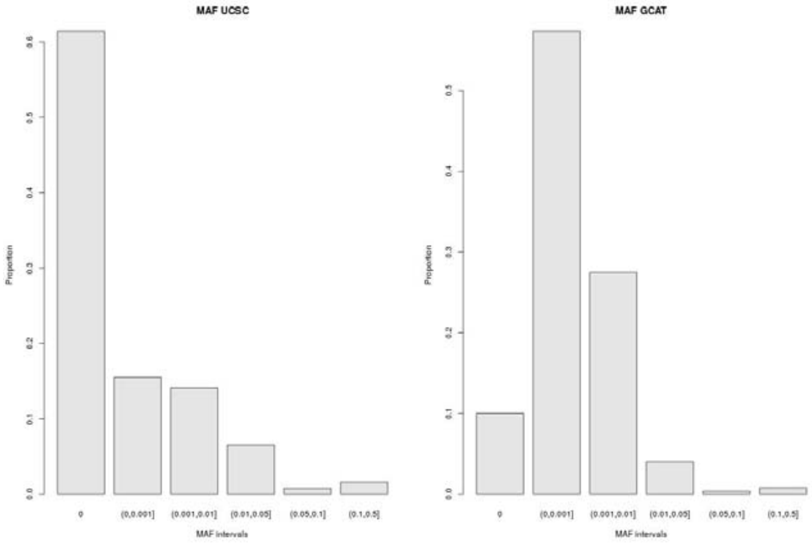


Figure QC13.2 Barplot of MAF intervals for SNPs with ratio(MAF 1000Genomes /MAF GCATCore)> 2.

QC14. Checking for batch effects

Each plate contains 96 samples. We plot the log10 missing call rate of the samples categorized by plate. Despite all the samples have a Call Rate > 98%, the samples from plates 22, 23, 33, 36 and 43 show high CallRate (in median).

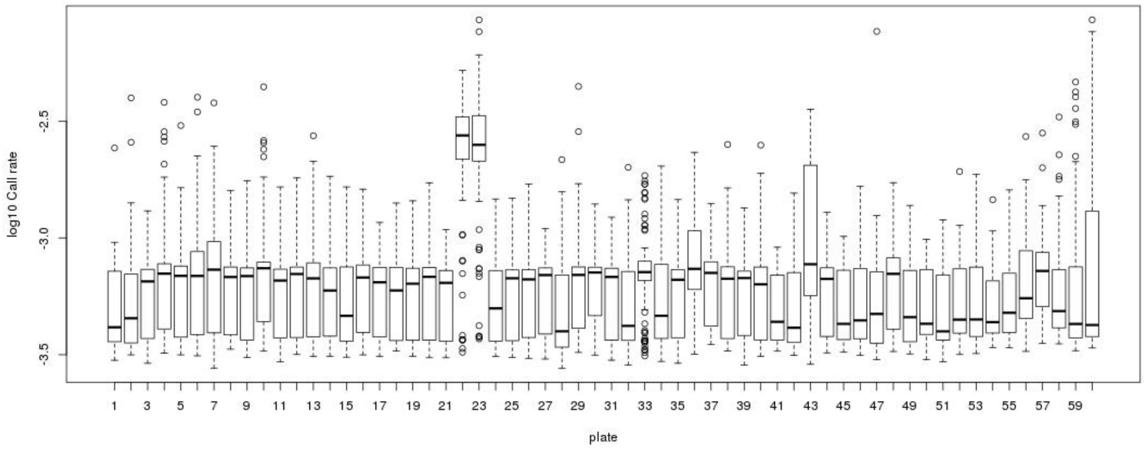


Figure QC14.1 Boxplot of the log10 missing call rate for each sample from QC3 by plate. N= **5,405** samples.

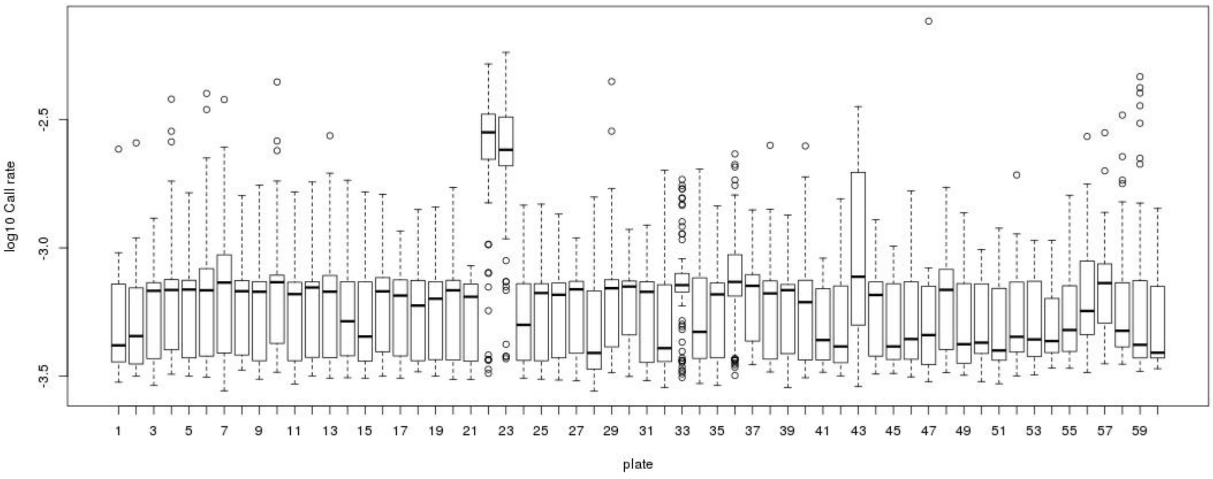


Figure QC14.2 Boxplot of the log10 missing call rate for each sample from QC11 by plate. N= **4,988** samples.

POPRES panel and GCAT

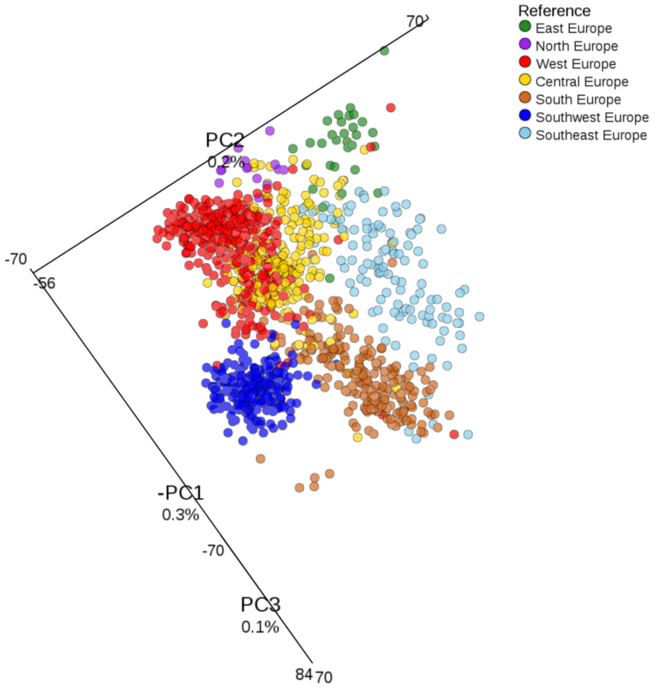
We compare the GCATcore selected (4,988 spanish individuals) with the imputed Population Reference Sample (POPRES) panel of 37 populations, 1,385 samples and 4,212,452 SNPs (Novembre et. al 2008).

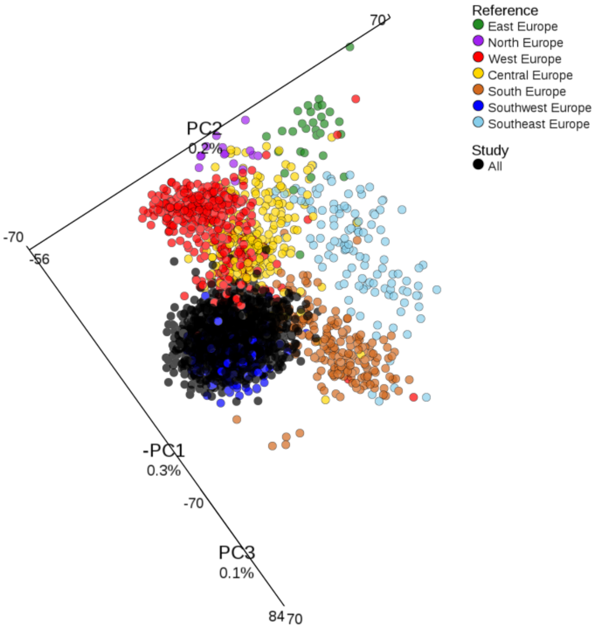
We update 608,850 SNPs at the LASER server (https://laser.sph.umich.edu/start.html#!pages/home)

 (hg19 reference genome) with QC1-10 filters applied and MAF > 0.05.

We intersect **235,406 NPs** with the POPRES panel and inspect the three principal components.

The GCATCore individuals overlapp with Spanish-Portuguese (Southwest Europe, blue color group) from the POPRES panel.





QC Summary

Filter	SNPs lost	SNPs kept
SNP probes		2,036,060
Multi-aligned SNPs	48,816	1,987,244
No aligned	198	1,987,046
Duplicate probes	32,614	1,954,432
Bad variants (GenTrain <0.7 & ClusterSep<0.4)	239,322	1,715,385
Missing call rate >= 2%	8,150	1,707,235
Mendelian errors	116	1,707,119
HWE	2,861	1,704,258
Concordance GCAT controls	24,452	-
Concordance Hapmap controls	10,884	-
Concordance MAF 1000Genomes	18,014	1,652,023
MAF < 0.01	895,250	756,773
MAF < 0.10	297,685	459,088
AT - CG sites	37,330	421,758

Filter	Individuals lost	Individuals kept
GCAT Cohort		5,575
Missing call rate >= 2%	0	5,575
Gender mismatch	19	5,556
Duplicate individuals	8	5,548
Related individuals (until 2nd degree)	88	5,460
MDS out of mean +/- 2sd	96	5,364
Heterozigosity and Inbreeding out mean +/- 4sd	52	5,312
CEPH controls	177	5,135
Non Spanish individuals (self-reported survey)	147	4,988

Requirements

Files

Intensity files (.idat). These files are the raw data files from the MEGA chip. They should be provided by the facility that performed the genotyping.

Sample sheets. The sample sheets are CSV files that contain sample information, such as plate ID, cell ID, gender and so on. The sample sheets should be provided by the genotyping facility.

Manifest file. It contains information about each probe, such as genomic position, sequence, strand, etc. and exists in text tab delimited format (human readable) or binary format (to input to GenomeStudio). The manifest file version used was "MEGA_Consortium_v2_15070954_A2".

Hardware

Windows workstation (required for microarray imaging and GenomeStudio software). Dell Precision T7600. Operating system: Windows Vista Professional (64 Bit. CPU: 2 x Intel Xeon X5667 3.06 GHz (4 cores, 8 threads). Memory: 48 GB, DDR3. Storage: 4 TB, 7,200 RPM..

Linux workstation (required for data analysis). HP Z440 workstation. Operating system: debian 8.8. CPU: Intel Xeon E5-1650 v3 3.5 GHz (6 cores 12 threads). Memory: 32 GB DDR4. Storage: 2TB, 7200 RPM.

Software

GenomeStudio v2011.1 with Genotyping module v1.9.4. GenomeStudio is commercial software developed by Illumina, and it is the only commercial software required for this protocol. It contains many analysis modules and is the only genotyping module required to process Illumina exome chip data. GenomeStudio and the Genotyping module can be downloaded and purchased from https://support.illumina.com/array/array_software/genomestudio/downloads.html

PLINK[22] v1.9. PLINK is a WGA analysis toolset with QC features that are useful for checking the integrity of the exome chip data. PLINK has been built for multiple operating systems; the Linux version is recommended. PLINK can be downloaded freely from ~purcell/plink/

R v3.4.0 64bit. R is a statistical programming language with excellent ability to create figures. R scripts have been provided to perform genotype QC and draw QC figures. R is built for multiple operating systems.

Python v2.7.13. Python is an scripting programming language. In this protocol, some of our in-house scripts were developed in Python and require a working Python installation to run. Python is built for multiple operating systems; the Linux version is recommended because large data sets can more easily be processed in the Linux system than using Windows.

WinSCP 5.9.6. WinSCP is a secure file transfer tool. In this protocol, it can be used to transfer data between Windows and Linux workstations. WinSCP can be downloaded freely at https://winscp.net/en

Other scripts and resources are necessary for the implementation of the protocol. These scripts were developed by the High Content Genomics and Bioinformatics Unit (IGTP) and are available under request.