Effects of filtration on imputation in clusterised variants

C. M. Charon, R. Allodji, J.F. Deleuze

Effects of filtrations on imputation in clusterised variants

Charon C[1*], Allodji R[2], Deleuze JF [1]
* Presenting author
1 CNRGH, Institut François Jacob, CEA Paris-Saclay, 2 rue Gaston Crémieux, Evry, France
2 Institut Gustave Roussy, Biostatistics and Epidemiology, Paris, France

BACKGROUND

Many unidentified loci are responsible for complexe diseases :
Same criteria for quality controls are mainly used from GWAS to imputations => Compare direct effects on imputed variants in presence and absence of QC pre-filtration and with different post-filtrations conditions using the same seed.
We also compared the results with our curated NCBI dbSNP (ie. 1,089 ID) harboring similar ethnicities and numbers to the individuals studied (1,031 ID)

METHODS

Empirical data for 1,031 individuals on Chr 20, 2Mb with 1,031 individuals [1]
ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/working/20130606_sample_info.xlsx
Shapeit2 [1] Impute2 [2] with re-phased reference 1000 Genome Build 37 of 1,089 ID [3]
ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase1/analysis_results/shapeit2_haplotypes
Statistical R package : 3.2.4 [4]
Create SQL Ncbi dbSNP Build 137 (GRCh37.4) followed by our curation

RESULTS

1 EFFECTS ON VARIANTS

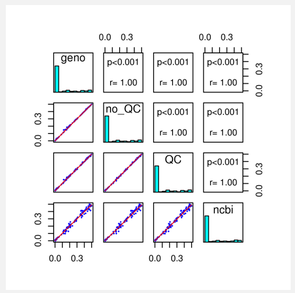

Fig 1. 17.5% of SNP genotypes didn't pass QC (MAF <0.01, HWE< 1[e]-06, success rate < 0.99) however their maf imputed in absence of filtration correlated with their maf imputed after QC pre-filtration and NCBI's maf.
Their info-impute 2 scores was max = 1

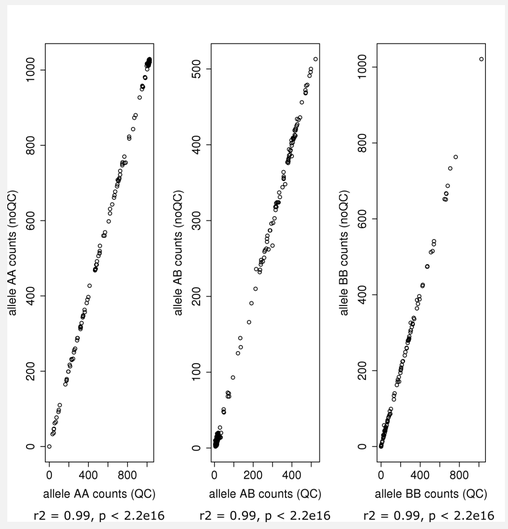

Fig 2. Alleles counts of 17.5% SNP

Table 1. Wilcoxson paired test of maf prior and after QC on all markers imputed

	maf noQC	maf QC	wilcoxson test		maf noQC	maf QC	wilcoxson test
maf categories	mean	mean	p-value	maf categories	mean	mean	p-value
0-0	0.00	0.00	< 2.20 E-16 ****	5E-03-1E-02	7.1 E -03	7.00 E-03	2.14 E-01
1E-04-5E-04	3.00 E-04	2.90 E-03	5.97 E-01	1E-02-5E-02	2.3 E-02	2.29 E-02	4.55 E-01
5E-04-1E-03	7.30 E-04	7.30 E-04	1.49 E-11 ****	5E-02-1E-01	7.3 E-02	7.30 E-02	9.20 E-01
1E-03-5E-03	2.4 E-03	2.40 E-03	8.99 E-11 ****	1E-01-5E-01	5 E-01	5 E-01	5.12 E-01

maf : minor alleles frequenies - QC : quality control. 2-sided levels of significance **** < 0.0001, *** < 0.001, ** < 0.01, * < 0.05
no significant differences were obtained in the total number (n) of variants prior (no QC) and after QC with the wilcoxson test (pvalue = 9.45 E-01).
****Differences were small and equal to the mean differences of 1.85 E-05 prior and after QC

2 NCBI and IMPUTATIONS

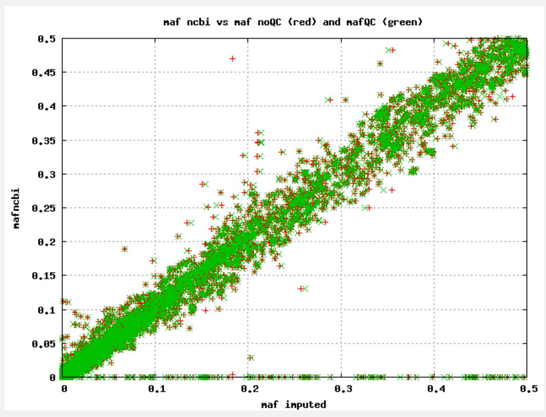

Fig 3. ncbi unreliable records : imputed variants QC/noQC vs NCBI with similar profiles
showing 145 null alleles records in dbSNP were in majority reliably imputed with or without QC

IMPUTATIONS QC vs no QC

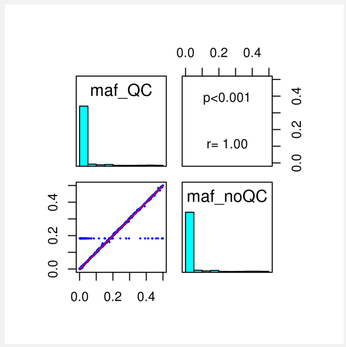

Fig 4. discordant mafs prior and after QC for only 44 SNP (0.15%) :
31 records with positions and no ncbi repositary names + 12 records with null alleles in ncbi

3 COMPARING DATA AND DATABASES

Differences : 145 variants : 84% with info > 0.8 were all structurals (Fig 3) + 9.5% variants with info score between 0.3-0.8 – The 44 showed discordant SV between QC/noQC (Fig 4) - 18 variants with repository names neither imputed nor in ncbi (null alleles) - 18 dual imputations with repository names, ie. indel + SNP = 36 variants - 13 SNP type 3 ie. not on the reference genome

4 EFFECTS ON QUALITY

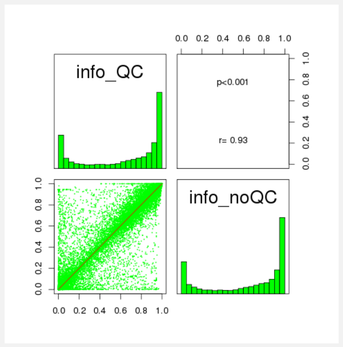

Fig 5. The same imputed variant will not necessarily produce the same quality score in presence or absence of pre-filtration score

Table 2. Wilcoxson right tail test comparing information-impute 2 before and after QC

maf categories	info p-value	pseudomedians	maf categories	info p-value	pseudomedians
0-0	1.00	-0.00995	5E-03-1E-02	< 2.20 E-16 ****	0.00740
1E-04-5E-04	3.30 E-04 ***	0.00360	1E-02-5E-02	< 2.20 E-16 ****	0.00198
5E-04-1E-03	3.48 E-16 ****	0.01350	5E-02-1E-01	9.50 E-03 *	0.00580
1E-03-5E-03	< 2.20 E-16 ****	0.00650	1E-01-5E-01	< 2.20E-16 ****	0.00969
0-5E-01	< 2.20E-16 ****	0.00206	0-5E-01	< 2.20E-16 ****	0.00206

info : information imputation score, maf : minor allele frequency, QC : quality control
right tail levels of significance **** < 0.00005, *** < 0.0005, ** < 0.005, * < 0.025 (paired test)
Differences of the distributions prior and after QC were at least equal to their pseudomedians

5 POST-FILTRATIONS

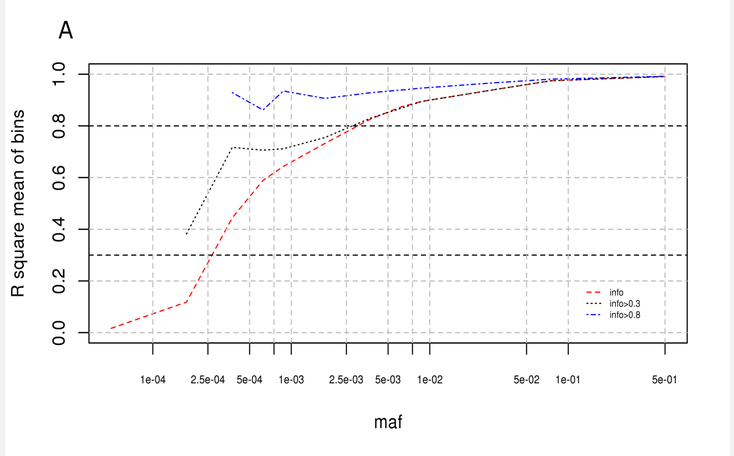

Fig 6. increasing filtration from 0.3 to 0.8 was drastic in eliminating very rare variants < 0.001 (decreased by 2.5 fold) and furthermore rare SNV < 0.01 (decreased by 1.8 fold). Average maf > 0.01 showed a mean quality score > 0.8

CONCLUSIONS

- In this data set better imputation quality scores were obtained in absence of QC pre-filtration compared with QC pre-filtration
- Pre-QC filtrations vs no QC pre-filtration had a significant effect on the maf although of small magnitude (1.85[E]-05) in one of each category of the very rare and rare variants
- The choice of post-filtration at quality score of 0.3 or 0.8 depends on whether to keep very rare and rare variants with less stringency (0.3) or to maintain less of those with confidence at more conservative threshold (0.8)
- Public curated database can be used to control the reliability of the imputed variants and vice et versa

THIS PRESENTED PAPER IS UNDER REVIEW

References : [1] Delaneau et al (2012) : Nat Methods, Vol 9 : 179-181. [2] Impute2 : Howie et al (2011) : G3 , Vol 1 : 457-470 - [3] Delaneau & Marchini (2014) : Nat Commun, Vol 5 : 3934. [4] R core Team : Bioconductor 3.2. R v 3.2.4 in R studio 0.99.115 2016-03-10. **Acknowledgements** : CNRGH initially CNG, Bioinformatics, Evry , France for building the db SNP SQL