Imputation reliability on DNA biallelic markers for drug metabolism studies

PhD Day - XXVI Cycle

PhD Student: Vladan Mijatovic
Tutor: Giovanni Malerba
Dept. Of Life and Reproduction Sciences

Background

Imputation is a statistical process used to predict genotypes of loci not directly assayed in a sample of individuals. Our goal is to measure the performance of imputation in predicting the genotype of the best known gene polymorphisms involved in drug metabolism using a common SNP array genotyping platform generally exploited in genome wide association studies.

Methods

Thirtynine (39) individuals were genotyped with both Affymetrix Genome Wide Human SNP 6.0 (AFFY) and Affymetrix DMET Plus (DMET) platforms. AFFY and DMET contain nearly 900000 and 1931 markers respectively. We used a 1000 Genomes Pilot + HapMap 3 reference panel. Imputation was performed using the computer program Impute, version 2. SNPs contained in DMET, but not being imputed, were analyzed studying markers around their chromosome regions. The efficacy of the imputation was measured evaluating the number of successful imputed SNPs (SSNPs, defined as SNPs with 5% of genotype error rate).

MAF	Shared	RPO	NAR
0	36	107	952
0-0.05	29	124	31
0.05-0.10	27	77	4
0.1-0.2	29	98	5
0.2-0.3	26	97	3
0.3-0.4	25	72	2
0.4-0.5	33	79	4
tot	205	654	1001

Table 1: SSNPs in NAR SNP class

MAF: Minor Allele Frequency
Shared: markers genotyped by both DMET and AFFY
RPO (Reference Panel Only): Number of markers that are present in DMET but not in AFFY
NAR (Neither in AFFY nor in Reference Panel): Number of markers not contained in reference panel so these SNPs cannot be directly imputed
For instance, 36 out of 205 shared SNPs, 107 out of 654 RPO SNPs and 952 out of 1001 NAR SNPs resulted to be monomorphic in the sample of 39 individual studied

Results

The imputation predicted the genotypes of 654 SNPs not present in the AFFY array but contained in the DMET array. Approximately 1000 SNPs were not annotated in the reference panel and therefore they could not be directly imputed. After testing three different imputed genotype calling threshold (IGCT) we observed that imputation performs at its best for IGCT value equal to 50% with a rate of SSNPs (MAF>0.05) equal to 85%.

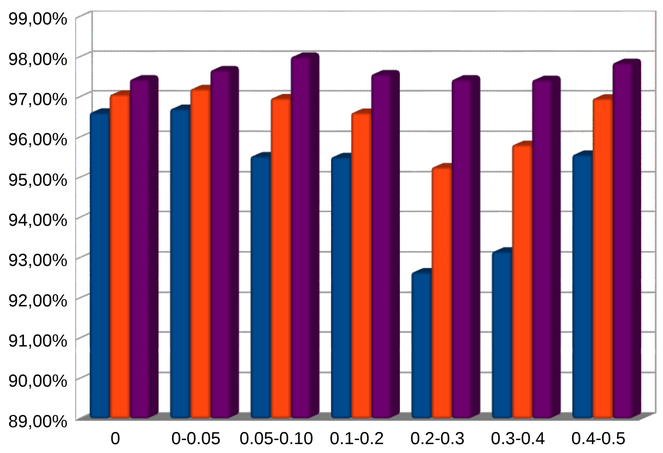

Chart 1: Concordance of IGCT for thresholds of 50%, 70 and 90%

Chart 2: Distribution of Shared (widest circle), RPO (middle circle) and NAR (smallest circle) by MAF

Conclusions

Most of genes involved in drug metabolism can be imputed with high efficacy using standard genome-wide genotyping platforms and imputing procedure.

Future development

We intend to impute chromosome X markers, using the first version of the computer program Impute, that allows to perform such type of analysis. Moreover, In order to validate the obtained results we are exploring alternative imputation software and procedures (i.e. Beagle, Mach).

		IGCT 50%				IGCT 70%				IGCT 90%			
		concordance	discordance	No-call	N. of SSNPs	concordance	discordance	No-call	N. of SSNPs	concordance	discordance	No-call	N. of SSNPs
MAF	#SNP												
0	107	96,62%	3,38%	0,00%	97	97,07%	2,90%	1,13%	92	97,44%	2,49%	2,47%	89
0-0.05	124	96,71%	3,29%	0,02%	103	97,21%	2,75%	1,55%	95	97,67%	2,23%	4,07%	79
0.05-0.10	77	95,53%	4,46%	0,13%	59	96,98%	2,93%	3,06%	55	98,01%	1,83%	8,09%	45
0.1-0.2	98	95,51%	4,47%	0,31%	75	96,61%	3,27%	3,43%	68	97,58%	2,20%	9,37%	56
0.2-0.3	97	92,65%	7,27%	1,11%	66	95,26%	4,39%	7,43%	61	97,44%	2,17%	15,25%	42
0.3-0.4	72	93,16%	6,77%	1,07%	53	95,81%	3,88%	7,26%	46	97,43%	2,21%	14,10%	37
0.4-0.5	79	95,57%	4,41%	0,26%	63	96,97%	2,92%	3,60%	59	97,85%	1,95%	9,22%	47

Table 2: Imputation on RPO SNPs

MAF: Minor Allele Frequency
#SNP: Number of SNPs
IGCT: imputed genotype calling threshold
Concordance: defined as the proportion of genotype calls for which the imputed genotype matched the expermental genotype call, averaged over all SNPs.
Discordance: one minus the concordance, indicates a genotype error rate
No-call: proportion of genotypes whose posterior probability did not reach a pre-specified IGCT
For instance, the first row reports the results of 107 SNPs having a MAF=0 in the study sample for IGCT=50%: the concordance is 96.62%, genotype error rate is 3.38% and there are no-call rate is 0.00%. For the IGCT=90%, the concordance, genotype error rate and no-call are 97.44%, 2.49% and 2.47%, respectively.

Vladan Mijatovic is an IT Ingeneer. He graduated in 2009 at the University of MM.FF.NN in Verona. Afterwards he worked as a research assistant at the faculty of Informatics. From 2011 he attends actively the PhD School in Translational Biomedicine.

References

- Single-sample analysis methodology for the DMET™ Plus Product
- Marchini J, Howie B: **Genotype imputation for genome-wide association studies**. *Nat Rev Genet.* 2010,**11:**499-511
- www.pharmgkb.org
- dbsnp:http://www.ncbi.nlm.nih.gov/projects/SNP/
- Marchini J, Howie B, Myers S, McVean G, Donnelly P: **A new multipoint method for genome-wide association studies by imputation of genotypes**. *Nat Genet* 2007, **39**:906-13.
- Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, et al.: **A second generation human haplotype map of over 3.1 million SNPs**. *Nature* 2009, **449**: 851–861.
- Li N, Stephens M: **Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data.** *Genetics* 2003, **165**:2213-33

The poster is freely downloadable at:
medgen.univr.it/~vlad/