Imputation quality thresholds for rare and common variants

G Pistis[1,3], C Sidore[1,2,3], A Mulas[1], M Zoledziewska[1], R Berutti[3,4], F Reinier[4], MF Urru[4], A Maschio[1,2], M Marcelli[4], A Angius[1,4], C Jones[4], G Abecasis[2], S Sanna[1], F Cucca[1,3]

[1]Istituto di Ricerca Genetica e Biomedica-CNR, Monserrato (CA), 09042, Italy

[2]University of Michigan, School of Public Health, Ann Arbor, MI, 48109

[3]Università degli Studi di Sassari, Dip Scienze Biomediche, Sassari, 07100, Italy

[4]Center for Advanced Studies, Research and Development in Sardinia – CRS4, Italy

Genotype imputation allows the analysis of variants discovered by sequencing to extend across much larger numbers of individuals. A key issue in these analyses is deciding which variants can be well imputed. Here, we use a large panel of 1,146 sequenced individuals to drive genotype imputation and investigate the behaviour of commonly used imputation quality thresholds. For this analysis, we compared imputed and real genotypes at 47,010 high quality SNPs from the ExomeChip, typed in 6,020 individuals from the SardiNIA study. We observed that for MAF ≥1% (28,267 SNPs), the standard imputation quality cutoff of RSQR >0.3, suggested for the algorithms implemented in MACH and minimac, discards the highest proportion of bad quality SNPs (genotype-dosage correlation <0.5) while keeping the highest proportion of good quality ones (genotype-dosage correlation >0.9). Indeed, with this threshold, we would discard 96.7% of poor quality SNPs and keep 99.7% of those with good quality. On the other hand, if we consider less common variants (0.5%≤ MAF <1%) (4,071 SNPs), the same RSQR threshold will exclude only 89.7% of the low quality imputed markers. Using instead an RSQR >0.55, the percentage of low quality markers excluded increases to 97.2%, while losing only 2.9% of those imputed with high accuracy. Finally for rare variants (0.05%≤ MAF <0.5%) (9,266 SNPs), an RSQR >0.65 can guarantee 96.7% and 4.7%, respectively. Our results thus suggest that more restrictive quality cutoffs might reduce false positive signals in association analyses, particularly for imputed rare variants.