## Supplementary information S1 | Alignment of reference and study datasets

Genotype imputation combines two (or more) datasets together and it is critical that this is done correctly. One issue is that genotypes produced by genotyping technologies and calling algorithms are “expressed” relative to either the + or - strand of the human genome reference. For example, a SNP which has alleles A and G relative to the + strand has alleles T and C relative to the + strand. Thus, it is crucial that each SNP has its genotypes expressed relative to the same strand in all the datasets used in imputation. All of the HapMap panels have had their SNP alleles expressed relative to the + strand of the human genome reference sequence and so genotype data from study samples will also need alleles expressed relative to the + strand. Fortunately, most approaches to imputation help users carryout this task and most genotyping chips have associated annotation files that list the strand of each SNP. It is also important that SNPs base-pair positions for each dataset use the same co-ordinate system i.e. the same build of the human genome. For example, using a version of HapMap in Build 36 and a set of genotype data in Build 35 could cause problems.