Estimation of genetic relationships between individuals across cohorts: resolving cross-platform issues through imputation.

Iryna O. Fedko[1], Jouke-Jan Hottenga[1,2], Carolina Medina-Gomez[4,6,7], Irene Pappa [4,8], Catharina E.M. van Beijsterveldt[1], Meike Bartels [1,2,3], Erik. A. Ehli [9], Gareth E. Davies[9], Fernando Rivadeneira[4,6,7], Henning Tiemeier[4,5], Morris A. Swertz[10,11], and Dorret I. Boomsma[1,2,3]

[1] Dept Biological Psychology, VU Univ Amsterdam; [2] EMGO Inst Health and Care Research, VU University Medical Center; [3] Neuroscience Campus Amsterdam; [4] Generation R Study Group, Erasmus Medical Center, Rotterdam; [5] Department of Child & Adolescent Psychiatry, Erasmus Medical Center, Sophia Children's Hospital; [6] Department of Epidemiology, Erasmus Medical Center; [7] Department of Internal Medicine, Erasmus Medical Center; [8] School of Pedagogical and Educational Sciences, Erasmus University Rotterdam; [9] Avera Institute for Human Genetics, Sioux Falls; [10] Univ Groningen, University Medical Center Groningen, Dept Genetics, Groningen; [11] Univ of Groningen, Univ Medical Center Groningen, Genomics Coordination Center, Groningen

Introduction: Increased sample sizes are beneficial for the estimation of SNP- heritability due to common SNPs, based on analyzing a Genetic Relationship Matrix (GRM) as can be obtained from the GCTA software (Genome-wide Complex Trait Analysis). Combing SNP data across multiple cohorts should be done at SNP level, where stratification due to different genotyping platforms may be an issue. Here, we analyze SNP data from two cohorts that were genotyped on different platforms. We explore imputation phasing as a tool to perform such combination and test the method on a 'benchmark' phenotype, i.e. height.

METHODS

- **Cohorts:**

1) GENR, 2226 individuals, mean age is 6 (0.4) yrs and height is 119.6 (5.6) cm.
2) NTR, 2072 individuals, mean age is 7.7 (1.4) yrs and height is 129.6 (9.8) cm.

- **QC prior combination:** ethnic outliers removed, sample call rate > 97.5 %, SNP call rate > 95%, MAF > 0.001, HWE P value < 10^{-5}, strand, build, heterozygosity rate, IBD/IBS status, gender mismatch. SNPs that are present in both platforms were selected from the GoNL reference set. SNPs different in frequency between cohorts and reference set ($p < 10^{-5}$) were removed.

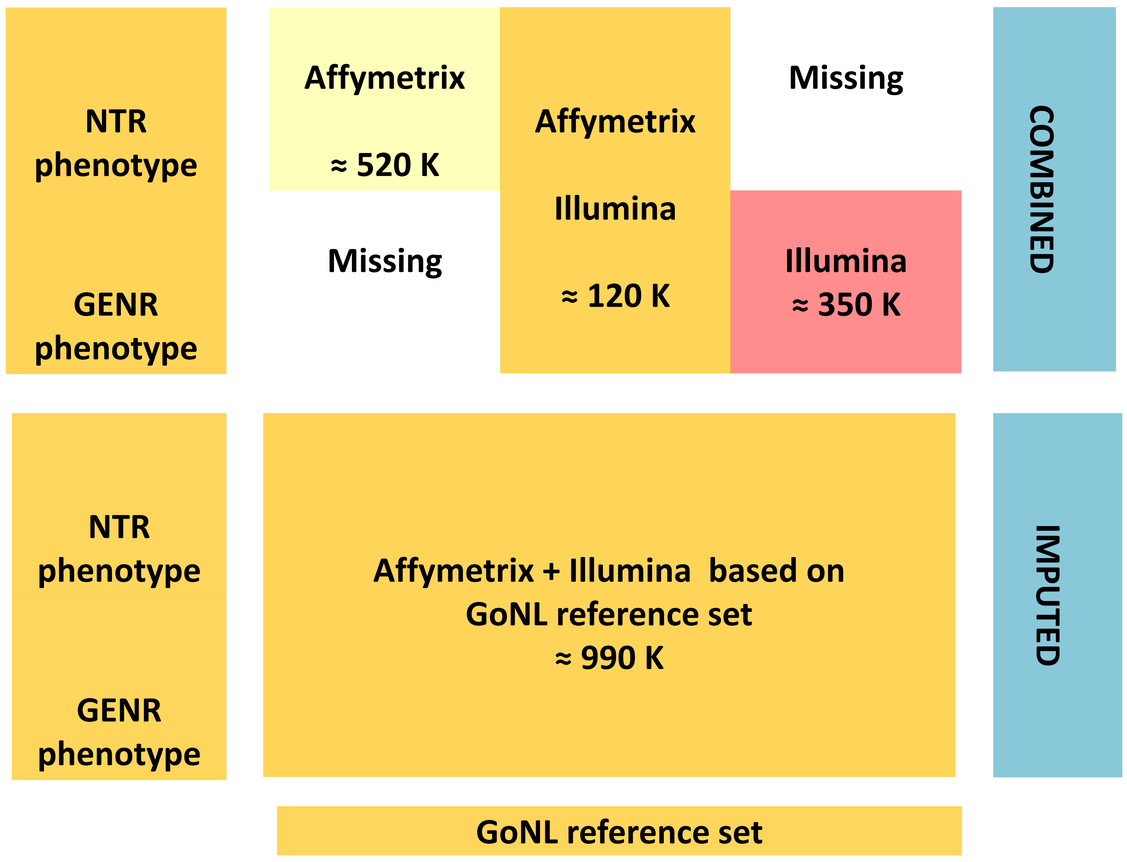
- **Imputation:** both data sets were merged in one set, phased and inherently imputed against GoNL reference set with MaCH-Admix.

- **Post-imputation QC:** R^2, alleles frequencies difference between cohorts, principal components analysis.

- **GRMs:** 1) Based on imputed data, MAF > 0.01 and R^2 > 0.8; 2) Based on combined (i.e. merged) data, MAF > 0.01; 3) Based on each separate cohort.

- **SNP-heritability** of height was estimated using GREML in distantly related individuals adjusted for age and sex

Table 2. Genotyping platforms



RESULTS

Table 1. Height heritability estimates

Data set	h^2	SE	N	Pval
Imputed [a]	0.51	0.10	3124	$1*10^{-7}$
Imputed clean [b]	0.49	0.10	3124	$2.9*10^{-7}$
Combined [c]	0.43	0.10	3124	$2*10^{-6}$
NTR independent [d]	0.49	0.28	1172	0.04
GENR independent [d]	0.59	0.17	1993	$1.8*10^{-4}$

[a] GRM based on data cross-platform imputed SNPs
[b] GRM based on data cross-platform imputed SNPs, excluding SNPs significantly different in frequency
[c] GRM based on the combined SNP data without imputation
[d] GRM based on each genotyped sample separately

DISCUSSION

- Different GRM build strategies have an effect on height heritability estimates (Table 1).
- Heritability estimates based on imputed GRM are closer to the estimates of the individual studies, but with a smaller standard error.
- Underestimation of height heritability in combined GRM is because the matrix is stratified into three groups based on SNPs: NTR AFFY platform SNPs, GENR Illumina SNPs, NTR-GENR AFFY-Illumina SNPs (Table 2).
- Post-Imputation QC shows a little residual stratification between cohorts.
- Strict prior and post imputation QC might remove true population differences between the two cohorts. Therefore when combining cohorts with a different ethnicity other approaches should be used.

CONCLUSION

- Haplotype information of a reference set for phasing and imputation of all SNPs on two different genotyping platforms, allows the combination of cohort data genotyped on both of these platforms.
- When combining genotype data across platform or cohort thorough pre - and post QC is required.
- To control for residual imputation stratification or true cohort population allele frequency differences, cohort should be included as a covariate.
- Cohorts combination strategy depends on the number of overlapping SNPs in relation to the total number of genotyped SNPs for both cohorts, and their ability to tag all the genetic variance related to the specific trait .