Design and implementation of a customized genotyping array for imputation-based genome-wide association studies: the Avera-NTR Global Screening Array

Jeffrey J. Beck[1], Erik. A. Ehli[1], Hamdi Mbarek[2], Noah Kallsen[1], Shanna Peyton[1], Abdel Abdellaoui[2], Iryna O. Fedko[2], Gonneke Willemsen[2], Eco J.C. de Geus[2], Dorret I. Boomsma[1,2], Gareth E. Davies[1,2], Jouke J. Hottenga[2]

1 Avera Institute for Human Genetics, Avera McKennan Hospital and University Health Center, Sioux Falls, South Dakota,
2 Department of Biological Psychology, Vrije Universiteit, Amsterdam, Netherlands

Basic Strategy

GONL reference 249 sequenced female individuals 22.1 million SNPs

→ Illumina GSA Extraction + QC
- MAF > 0.01
- M Ind > 0.90
- M SNP > 0.95
- HWE > 10^{-5}
- In 1000G reference

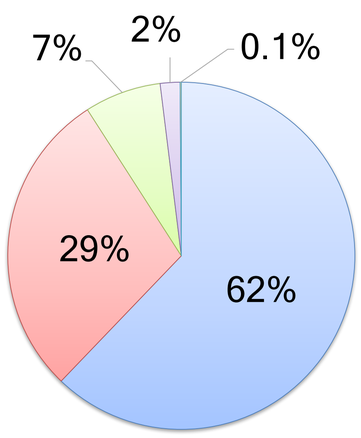
→ Like a genotyped GSA array with 616,961 SNPs Chr 1-22

↓ Phasing (Shapeit) + Imputation (Impute2) against 1000G phase 3 reference

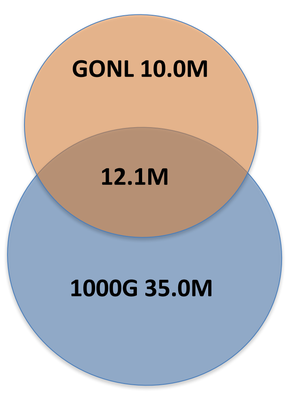
→ 47.1 million imputed SNPs 1000G

↑ Concordance? Coverage? Imputation Quality R^2?

Concordance Results



	N SNPs
>99%	7506660
>95%-99%	3470030
>80%-95%	861745
>50%-80%	221831
<50%	14204



Estimates for the Netherlands:
62.2% of the genome fully covered
35.8% of the genome covered with imputation
2% of the genome not covered

Avera-NTR GSA Array Content

Marker type	**#SNPs (N=697,486)**
GSA Core Backbone	**~660,000**
Sex Chromosomes	19,660 X; 1,691 Y; 547 PAR
ADME Genes/Exons	6,668; 2,787
ClinVar	17,020
MHC	9797
Ancestry Informative	3212
Fine Mapping Content (candidate genes and imputation)	**~30,000**
Custom Markers	**~8,000**
Fertility and Twinning	
Body Stature (height, BMI) and Sports and Exercise Behavior	
Mental State and Health (happiness, depression, schizophrenia)	
Chromosome X (imputation)	
Educational Attainment	
Pharmacogenomics	

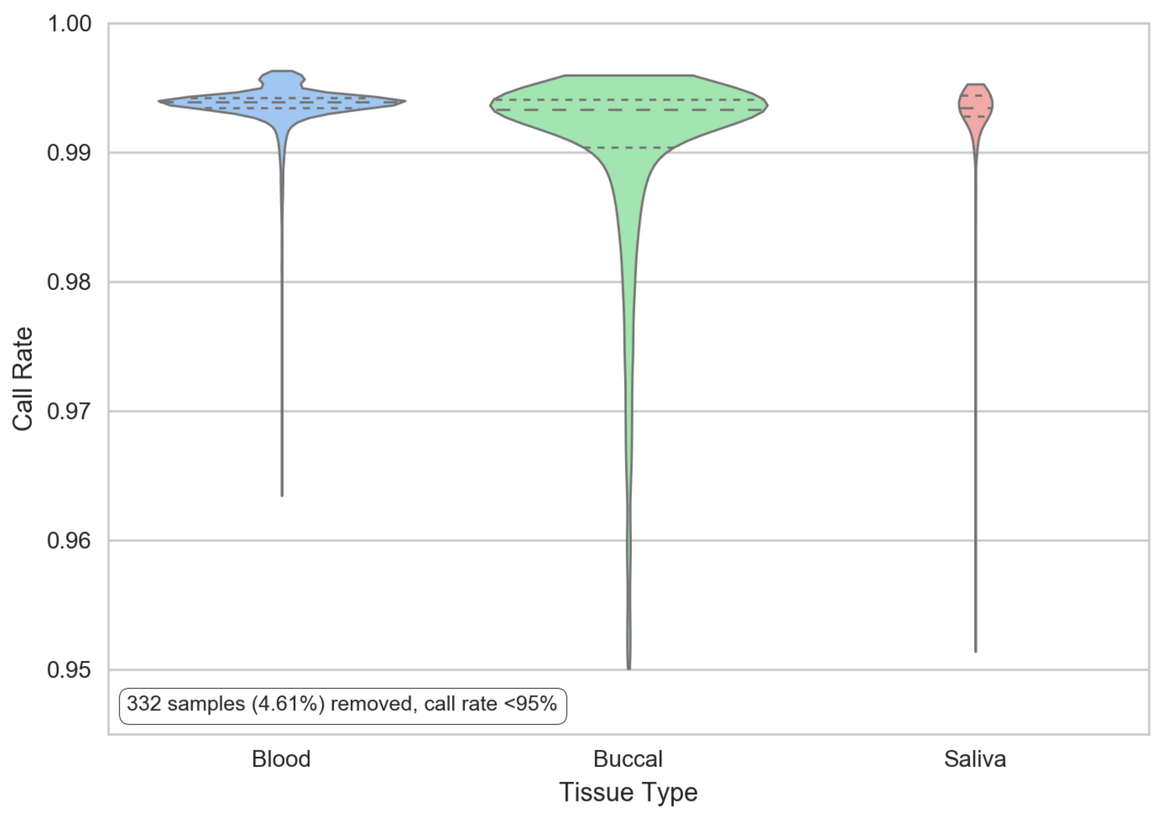
Imputation Quality R^2 Results

Table 1: Imputation quality measure per minor allele frequency bin

Selected SNPs	**Chr**	**MAF**	**N SNPs**	**Median R^2**	**Mean R^2**	**SD**
1000G All SNPs*	1-22	>0.000-0.001	21,373,838	0.02	0.05	0.08
Full 1000G imputation with		>0.001-0.01	6,853,643	0.69	0.64	0.28
Asian/African/Other SNPs		>0.01-0.05	2,863,052	0.97	0.91	0.13
not present in the Dutch population		>0.05	6,974,825	0.99	0.96	0.08
GONL and 1000G*	1-22	>0.000-0.001	1,003,022	0.04	0.08	0.10
Overlapping SNPs between		>0.001-0.01	2,736,096	0.80	0.74	0.24
GONL and 1000G		>0.01-0.05	2,461,024	0.97	0.92	0.12
**Polymorphic SNPs*		>0.05	5,874,328	0.99	0.97	0.07

Genotyping Performance Per Tissue

Kernel density estimates for 6,871 samples [N=7,203] with CR>95%: 3,040 blood; 3,420 buccal; 411 saliva



Considerations/Limitations

- Coverage, concordance, and imputation were assessed with the GSA backbone
- Genotyping performance data presented here are derived from preliminary annotation and cluster files

Conclusions

- Backbone of the array was designed for the global population
- The Dutch genome can be imputed very well
- Allows for the rapid genotyping of a large number of highly relevant SNPs
  - GWAS loci
  - Common variant risk profiling
  - Zygosity and IBD determination