European Journal of Human Genetics (2017) 25, 267–270

www.nature.com/ejhg

SHORT REPORT

A method to customize population-specific arrays for genome-wide association testing

Erik A Ehli*,[1], Abdel Abdellaoui[2], Iryna O Fedko[2], Charlie Grieser[3], Sahar Nohzadeh-Malakshah[3], Gonneke Willemsen[2], Eco JC de Geus[2,4], Dorret I Boomsma[1,2,4], Gareth E Davies[1,2] and Jouke J Hottenga[2,4]

As an example of optimizing population-specific genotyping assays using a whole-genome sequence reference set, we detail the approach that followed to design the Axiom-NL array which is characterized by an improved imputation backbone based on the Genome of the Netherlands (GoNL) reference sequence and, compared with earlier arrays, a more comprehensive inclusion of SNPs on chromosomes X, Y, and the mitochondria. Common variants on the array were selected to be compatible with the Illumina Psych Array and the Affymetrix UK Biobank Axiom array. About 3.5% of the array (23 977 markers) represents SNPs from the GWAS catalog, including SNPs at FTO, APOE, Ion-channels, killer-cell immunoglobulin-like receptors, and HLA. Around 26 000 markers associated with common psychiatric disorders are included, as well as 6705 markers suggested to be associated with fertility and twinning. The platform can thus be used for risk profiling, detection of new variants, as well as ancestry determination. Results of coverage tests in 249 unrelated subjects with GoNL-based sequence data show that after imputation with 1000G as a reference, the median concordance between original and imputed genotypes is above 98%. The median imputation quality R^2 for MAF thresholds of 0.001, 0.01, 0.05, and >0.05 are 0.05, 0.28, 0.80, 0.99, respectively, for the 1000G imputed SNPs, with a similar quality for the autosomes and X chromosome, showing a good genome-wide coverage for association studies after imputation.

European Journal of Human Genetics (2017) **25,** 267–270; doi:10.1038/ejhg.2016.152; published online 23 November 2016

INTRODUCTION

Genome-wide association studies (GWAS) in large population samples have been the key method to identify genetic variants involved in complex human traits.[1,2] Multiple successful GWAS studies have been reported ranging from body size, metabolomics, and medically relevant traits (reviewed),[3,4] to hormones,[5] personality,[6] educational attainment,[7] and lifestyle characteristics.[8–10]

The major technology behind these successes is the relatively cheap genotyping, in comparison with full genome sequencing, of DNA samples on genotyping arrays with 300 K–5 M single-nucleotide polymorphisms (SNPs), followed by imputation of the unmeasured SNPs. Initially, the contents of these arrays were determined by the manufacturers, but recently companies also allow researchers to select the variants on an array. Here, we focus on the Axiom array, a genotyping solution from Affymetrix, Inc., which provides a high throughput platform for high-density SNP genotyping on a diverse range of sample types. This array has been used for several large population-wide genome-screening projects including the UK biobank[11] and the GERA cohorts.[12] We describe a similar custom-made Axiom array for the Netherlands population, the Axiom-NL that allows good imputation and enhances association, and risk score analysis on DNA samples collected within Dutch Biobanks, such as the large number of Biobanks collaborating in BBMRI-NL.[13] Notwithstanding the application to this specific population, the SNP selection procedures and coverage testing can provide a general guideline for customizing the Axiom array in other populations for which valid reference sequence genomes are available.

MATERIALS AND METHODS

SNP selection for the Axiom-NL array

An overview of the SNP selection is provided in Figure 1, a stepwise procedure of the selection is given in the Supplementary materials. The core of the array was optimized for genome-wide coverage using genotype imputation. The Affymetrix SNP 6.0 array formed the starting point for selection of markers.[14] SNPs were selected that passed quality control (see supplementary methods), including if the replicate genotype error rate in control samples was $<1\%$ in prior experiments, and provided the most tagging information. We then selected up to 10 additional markers per mega base in areas that had a low imputation quality (mean $R^2<0.35$) based on imputations with these SNPs alone. Here, SNPs were selected only if they were present in the Dutch population based on the GoNL reference sequence data.[13] We prioritized them based on the following criteria: $R^2<0.30$, MAF>0.07, and preferably high LD $r^2>0.5$–0.9 with other weakly imputed SNPs in the 1 MB region. Selected SNPs in high LD with each other were removed (PLINK 1.07 –indep 200 10 1.5).[15] Applying a MAF 0.07 threshold here is crucial, because the LD pruning step otherwise selects only rare independent SNPs. The same SNP selection approach was used for chromosome X, to achieve similar coverage as the autosomes.

In addition, laboratory validated SNPs on commercially available microarrays (eg, Affymetrix UK Biobank Axiom Array, Affymetrix Axiom Biobank Array used in the Million Veteran Program, and the Illumina (San Diego, CA, USA) Infinium Psych Array) were added as they are informative for several traits and disease studies. An important consideration for selection of SNPs from these platforms was to focus on selecting common SNPs (MAF>0.01), with the

[1]Avera Institute for Human Genetics, Sioux Falls, SD, USA; [2]The Netherlands Twin Register, Vrije Universiteit (VU), Amsterdam, The Netherlands; [3]Affymetrix Inc., Santa Clara, CA, USA; [4]EMGO institute for Health and Care Research, VU and VU University Medical Center, Amsterdam, The Netherlands
*Correspondence: Dr EA Ehli, Scientific Director, Avera Institute for Human Genetics, 3720W. 69th Street, Sioux Falls, SD, 57108 USA. Tel: +1 605 322 5976; Fax: +1 605 322 3051; E-mail: erik.ehli@avera.org

Received 16 February 2016; revised 5 October 2016; accepted 7 October 2016; published online 23 November 2016

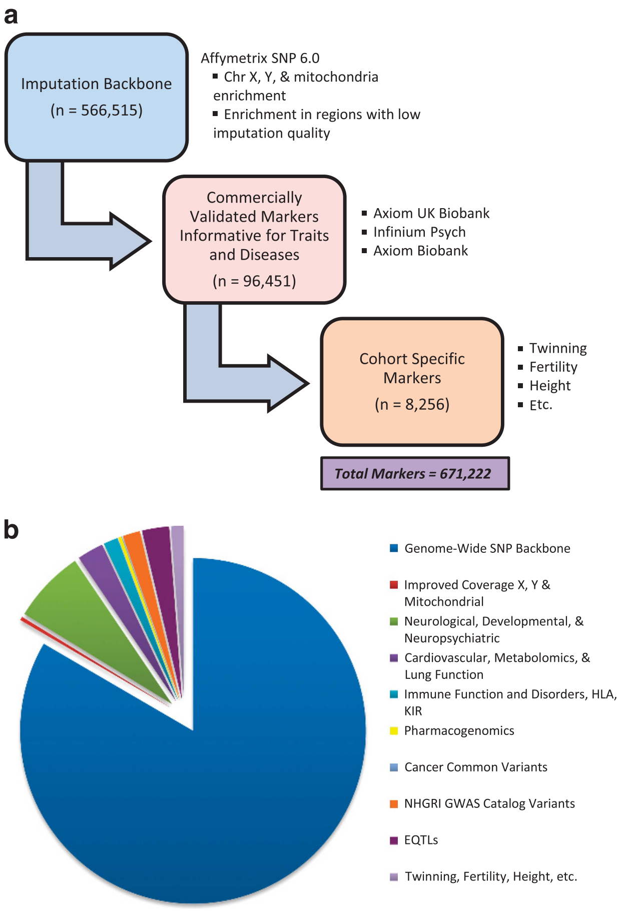
a
Imputation Backbone
(n = 566,515)
Affymetrix SNP 6.0
▪ Chr X, Y, & mitochondria enrichment
▪ Enrichment in regions with low imputation quality
Commercially Validated Markers Informative for Traits and Diseases
(n = 96,451)
▪ Axiom UK Biobank
▪ Infinium Psych
▪ Axiom Biobank
Cohort Specific Markers
(n = 8,256)
▪ Twinning
▪ Fertility
▪ Height
▪ Etc.
Total Markers = 671,222
b
Genome-Wide SNP Backbone
Improved Coverage X, Y & Mitochondrial
Neurological, Developmental, & Neuropsychiatric
Cardiovascular, Metabolomics, & Lung Function
Immune Function and Disorders, HLA, KIR
Pharmacogenomics
Cancer Common Variants
NHGRI GWAS Catalog Variants
EQTLs
Twinning, Fertility, Height, etc.

Figure 1 Axiom-NL marker selection and content. (**a**) Simplified flow chart depicting the strategy for selecting markers for Axiom-NL array. (**b**) A pie graph representing the broad breakdown of marker categories included in Axiom-NL.

exclusion of important rare SNPs known to have associations with complex traits of interest (Supplementary Figure 1). Furthermore, chromosome Y markers were selected to be the same as the Axiom UK Biobank. Mitochondrial markers were selected from Axiom UK Biobank and the Human Mitochondrial Genome Database to represent the most frequent Dutch haplotypes. The final annotation file of the selected 671 222 SNPs is available for download at www.avera.org/axiom.

Estimating genome-wide concordance

To estimate the coverage of the Axiom-NL design, we used a two genome-wide reference approach, where SNPs from the GoNL reference were re-imputed with the 1000 genomes reference based on only the selected Axiom-NL SNPs. The GoNL reference data are 769 individuals, spread across the Netherlands that were sequenced, aligned to genome build 37, variant-called, and phased for imputation for chromosomes 1–22 and X.[16] The 1000 Genomes reference panel are 2504 sequenced individuals from several populations worldwide.[17]

From the GoNL sequence data, 249 unrelated women were identified and their genotype data for chromosomes 1–22 and X were extracted (N = 22 932 747 SNPs). Data from only women were selected to facilitate that chromosome X imputes similarly to the autosomes. From these data, the 671 222 SNPs that were present on the Axiom-NL array were extracted as input data for the 1000G imputation. These SNPs were filtered with the following criteria: minor allele frequency (MAF) <0.01, call rate <0.95, Hardy–Weinberg Equilibrium test P-value $<10^{-5}$, and SNPs having the same alleles as in 1000G, leaving 618 889 SNPs. Strands were checked (SHAPEIT 2.7r790) and flipped (PLINK 1.07) if required before phasing. This set thus mimics a quality controlled pre-imputation genotyped Axiom-NL data set. For comparison the same procedure was applied for the SNP annotation lists from two similarly sized commercially available arrays, the Affymetrix Axiom Biobanking array and the Illumina Infinium OmniExpress-24 BeadChip. Subsequently, the three data sets were phased with SHAPEIT 2.7r790 and imputed against the 1000G all reference panel Phase 3 (October 2014) for the autosomes, and 1000G All Phase 1 interim (June 2011) for the X chromosome. Note that we could not use the GoNL or HRC as an imputation reference panel here, since the subjects from our to be imputed GoNL data set are present in these reference panels. As such they will be imputed back perfectly as their haplotypes would match 100% (previously tested). This would therefore not tell us anything about the ability to impute the genome from just the Axiom-NL SNP list. The imputations were done with IMPUTE2.3.1 using standard protocols.[18] From the imputed data, best-guess genotypes were calculated for all 82 943 231 SNPs (Plink 1.90). For 12 205 845 overlapping SNPs between GoNL and 1000G with MAF > 0 in the imputed data of all three sets and both references, the concordance between the GoNL sequence and the 1000G imputations was calculated with PLINK 1.90. For each SNP, polymorphic in all sets, the median, average, and SD was calculated for the imputation quality R^2 (Quicktest 0.95) in MAF bins >0–0.001, >0.001–0.01, >0.01–0.05, and >0.05 (SPSS 22). This was done for the full 1000 genomes imputation and for the SNPs overlapping with GoNL.

RESULTS

For the imputation quality R^2, where zero is extremely sub-optimal and one is excellent, the results from the three platforms are presented in Table 1. For all 1000 genomes imputed SNPs, autosomes, and chromosome X, the median R^2 values for the Axiom-NL platform are 0.0496 for MAF 0–0.001, 0.281 for MAF >0.001–0.01, 0.805 for MAF >0.01–0.05, and 0.991 for MAF >0.05 (Table 1). For chromosome X alone these values are 0.120, 0.555, 0.838, and 0.993, respectively, indicating that the rare chromosome X SNPs are imputed slightly better than the autosomes and the common SNPs equally well. With these results, the Axiom-NL platform is just in between the Affymetrix Axiom Biobanking array and the Illumina Infinium OmniExpress-24 BeadChip as shown in Table 1. The differences in imputation quality are, however, extremely small, between the three platforms for all MAFs and all chromosomes. When selecting SNPs that are present in the GoNL and the 1000G reference data, the true variants in the Dutch population, the results show an even better imputation quality. The main reason behind this is the large number of rare SNPs, which are likely absent in the Dutch population are now excluded, which improves the median and mean scores.

The concordance rates of the genotyped GoNL SNPs that were re-imputed with a 1000 Genomes imputation were generally high for most SNPs in the genome (Table 2). For the 12 205 845 markers with MAF > 0, being polymorphic in the imputed data for all three platforms in the 249 women, up to 59.6% can be re-imputed at high quality. With a lower level of quality (down to 80% concordance), only 2.8% of the genome is not covered well. For the concordance measurements, the Axiom-NL array is again in between the Affymetrix Axiom Biobanking array and the Illumina Infinium OmniExpress-24 BeadChip, where the Illumina chip performs slightly better and the Axiom Biobanking array slightly worse. Genome wide, there are no large differences between the chips imputation quality.

DISCUSSION

The Axiom-NL Array was developed with a custom backbone to provide optimal imputation for the Dutch population with an improvement of coverage for chromosome X. The design incorporates a significant clinical relevance focus by including the common variants from two large consortia (Psychiatric Genomics Consortium and

Table 1 Imputation metrics comparing Axiom-NL, Axiom Biobanking, and the Infinium HumanOmniExpress arrays

SNP set	MAF range[a]	N SNPs	AXIOM_NL Median R^2	AXIOM_BB Median R^2	INF_OMNI Median R^2	AXIOM_NL Mean R^2 (SD)	AXIOM_BB Mean R^2 (SD)	INF_OMNI Mean R^2 (SD)
1000G	>0–0.001	13 193 788	0.0496	0.0489	0.0484	0.188 (0.275)	0.186 (0.273)	0.187 (0.278)
	>0.001–0.01	6 396 705	0.281	0.284	0.270	0.395 (0.361)	0.394 (0.356)	0.396 (0.368)
	>0.01–0.05	3 185 136	0.805	0.783	0.828	0.621 (0.390)	0.616 (0.374)	0.624 (0.398)
	>0.05	7 942 323	0.991	0.979	0.994	0.928 (0.168)	0.924 (0.158)	0.938 (0.167)
1000G-GONL	>0–0.001	1 425 657	0.345	0.337	0.369	0.427 (0.362)	0.423 (0.362)	0.438 (0.365)
	>0.001–0.01	2 500 511	0.768	0.763	0.783	0.688 (0.277)	0.678 (0.277)	0.700 (0.275)
	>0.01–0.05	1 777 554	0.952	0.924	0.963	0.860 (0.215)	0.837 (0.224)	0.872 (0.209)
	>0.05	6 502 123	0.994	0.983	0.995	0.956 (0.106)	0.950 (0.103)	0.967 (0.098)

[a] Since the imputation of minor alleles is dependent on the platform data, we choose the minor allele frequency of the 1000G population to be the value determining the MAF category. This makes the comparison consistent as the same SNPs remain in one category, and do not change over minor allele frequency bin based on their imputed MAF in a particular platform. All SNPs monomorphic in 1000G as well as in any of the imputed platform data were excluded (MAF>0 in 1000G, AXIOM-NL, AXIOM_BB, and INF_OMNI). Subsequently, only the SNPs were selected that were present in 1000G and GONL.

Table 2 Concordance metrics for Axiom-NL, Axiom Biobanking, and Infinium HumanOmniExpress arrays

Genotype Concordance	AXIOM_NL N SNPs(%)	AXIOM_BB N SNPs(%)	INF_OMNI N SNPs (%)
>99%	7 277 252 (59.6%)	6 322 752 (51.8%)	7 724 224 (63.3%)
>95–99%	3 582 818 (29.4%)	4 554 885 (37.4%)	3 404 295 (27.9%)
>80–95%	999 240 (8.2%)	1 054 715 (8.6%)	794 725 (6.5%)
>50–80%	285 103 (2.3%)	221 354 (1.8%)	226 040 (1.9%)
⩽50%	61 432 (0.5%)	52 139 (0.4%)	56 561 (0.5%)

Total number of 1000G re-imputed SNPs, polymorphic on all platforms, and present in GONL = 12 205 845.

UK Biobank). Over 60 000 markers are included from the UK BioBank array including known GWAS hits from the NHGRI GWAS catalog, with additional modules including apoE, HLA, cardiometabolic, and mitochondrial SNPs. For projects of interest to twin registers, 6705 additional candidate SNPs implicated in fertility and twinning were selected.

With a general reference set, and markers selected for the Axiom-NL array we can re-impute with high confidence, keeping in mind the exception that rare alleles (MAF < 0.001) are never imputed well.[19] Also the methods we utilized, having only used the sequence of 249 samples, the imputation and presence of minor alleles with MAF < 0.01 was likely not optimal. However, for comparison tests this should not matter and the imputation is of similar quality to other commercially available chips namely the Affymetrix Axiom Biobanking array and the Illumina Infinium OmniExpress-24 BeadChip. Finally, our method tested the coverage using two reference data sets and the concordance between genotyped SNPs, and re-imputed SNPs inherently assumes that the SNPs need to be present in both reference data sets. As such we thus assume that population-specific SNPs, for example, present only in GoNL are covered and imputed just as well.

Knowledge generation in genetic epidemiology depends increasingly on the use of SNP-array based GWA studies, including (bivariate) GCTA and polygenic risk score analysis, or the combination of summary statistic information for multiple traits as in LD-score regression and Mendelian Randomisation.[20] We here show that customized population-specific arrays for imputation-based GWA testing can be a valuable tool to generate high quality GWA results.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ACKNOWLEDGEMENTS

This work was supported by Biobanking and Biomolecular Resources Research Infrastructure (BBMRI-NL) [184.021.007], the National Institutes of Health (NIH) [RC2MH08995], ENGAGE (NIHHEALTHF4–2007–201413), the Royal Netherlands Academy of Science Professor Award (PAH/6635) to DIB, the EMGO+ Institute for Health and Care Research, Neuroscience Campus Amsterdam, and the Avera Institute for Human Genetics. We would like to thank and acknowledge all the individuals and scientists who participated in the Genome of the Netherlands Project (GoNL). Charlie Grieser and Sahar Nohzadeh-Malakshah are employees of Affymetrix, Inc.

1 Stranger BE, Stahl EA, Raj T: Progress and promise of genome-wide association studies for human complex trait genetics. *Genetics* 2011; **187**: 367–383.
2 Visscher PM, Brown MA, McCarthy MI, Yang J: Five years of GWAS discovery. *Am J Hum Genet* 2012; **90**: 7–24.
3 Geschwind DH, Flint J: Genetics and genomics of psychiatric disease. *Science* 2015; **349**: 1489–1494.
4 Chang CQ, Yesupriya A, Rowell JL *et al*: A systematic review of cancer GWAS and candidate gene meta-analyses reveals limited overlap but similar effect sizes. *Eur J Hum Genet* 2014; **22**: 402–408.
5 Ruth KS, Campbell PJ, Chew S *et al*: Genome-wide association study with 1000 genomes imputation identifies signals for nine sex hormone-related phenotypes. *Eur J Hum Genet* 2016; **24**: 284–290.
6 Genetics of Personality Consortium, de Moor MH, van den Berg SM *et al*: Meta-analysis of genome-wide association studies for neuroticism, and the polygenic association with major depressive disorder. *JAMA Psychiatry* 2015; **72**: 642–650.
7 Rietveld CA, Medland SE, Derringer J *et al*: GWAS of 126,559 individuals identifies genetic variants associated with educational attainment. *Science* 2013; **340**: 1467–1471.
8 Tobacco Genetics Consortium: Genome-wide meta-analyses identify multiple loci associated with smoking behavior. *Nat Genet* 2010; **42**: 441–447.

9 Agrawal A, Lynskey MT, Hinrichs A *et al*: A genome-wide association study of DSM-IV cannabis dependence. *Addict Biol* 2011; **16**: 514–518.

10 Coffee Caffeine Genetics Consortium, Cornelis MC, Byrne EM *et al*: Genome-wide meta-analysis identifies six novel loci associated with habitual coffee consumption. *Mol Psychiatry* 2015; **20**: 647–656.

11 Hagenaars SP, Harris SE, Davies G *et al*: Shared genetic aetiology between cognitive functions and physical and mental health in UK Biobank (N = 112 151) and 24 GWAS consortia. *Mol Psychiatry* 2016; **21**: 1624–1632.

12 Kvale MN, Hesselson S, Hoffmann TJ *et al*: Genotyping informatics and quality control for 100,000 subjects in the Genetic Epidemiology Research on Adult Health and Aging (GERA) cohort. *Genetics* 2015; **200**: 1051–1060.

13 Boomsma DI, Wijmenga C, Slagboom EP *et al*: The genome of the Netherlands: design, and project goals. *Eur J Hum Genet* 2014; **22**: 221–227.

14 Scheet P, Ehli EA, Xiao X *et al*: Twins, tissue, and time: an assessment of SNPs and CNVs. *Twin Res Hum Genet* 2012; **15**: 737–745.

15 Purcell S, Neale B, Todd-Brown K *et al*: PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 2007; **81**: 559–575.

16 Genome of the Netherlands Consortium: Whole-genome sequence variation, population structure and demographic history of the Dutch population. *Nat Genet* 2014; **46**: 818–825.

17 The 1000 Genomes Project Consortium, Auton A, Brooks LD *et al*: A global reference for human genetic variation. *Nature* 2015; **526**: 68–74.

18 van Leeuwen EM, Kanterakis A, Deelen P *et al*: Population-specific genotype imputations using minimac or IMPUTE2. *Nat Protoc* 2015; **10**: 1285–1296.

19 Zheng HF, Rong JJ, Liu M *et al*: Performance of genotype imputation for low frequency and rare variants from the 1000 genomes. *PLoS One* 2015; **10**: e0116487.

20 Bulik-Sullivan B, Finucane HK, Anttila V *et al*: An atlas of genetic correlations across human diseases and traits. *Nat Genet* 2015; **47**: 1236–1241.

Supplementary Information accompanies this paper on European Journal of Human Genetics website (http://www.nature.com/ejhg)