



HAL
open science

NordicDB: A Nordic pool and portal for genome-wide control data

Monica Leu, Dr. Humphreys, Ida Surakka, Emil Rehnberg, Juha Muilu, Päivi Rosenström, Peter Almgren, Juha Jääskeläinen, Richard Lifton, Kirsten Ohm Kyvik, et al.

► To cite this version:

Monica Leu, Dr. Humphreys, Ida Surakka, Emil Rehnberg, Juha Muilu, et al.. NordicDB: A Nordic pool and portal for genome-wide control data. *European Journal of Human Genetics*, 2010, 10.1038/ejhg.2010.112 . hal-00560306

HAL Id: hal-00560306

<https://hal.science/hal-00560306>

Submitted on 28 Jan 2011

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

NordicDB: A Nordic pool and portal for genome-wide control data

Monica Leu^{1,2}, Keith Humphreys¹, Ida Surakka^{2,3}, Emil Rehnberg¹, Juha Muilu², Päivi Rosenström², Peter Almgren⁴, Juha Jääskeläinen⁵, Richard P. Lifton⁶, Kirsten Ohm Kyvik⁷, Jaakko Kaprio^{2,8,9}, Nancy L. Pedersen¹, Aarno Palotie^{2,10,11}, Per Hall¹, Henrik Grönberg¹, Leif Groop⁴, Leena Peltonen^{2,3,10,11}, Juni Palmgren^{1,12}, Samuli Ripatti^{2,3}

¹ Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Sweden

² Institute for Molecular Medicine, Finland, FIMM, University of Helsinki, Finland

³ Public Health Genomics Unit, National Institute for Health and Welfare, Helsinki, Finland

⁴ Department of Clinical Sciences, Diabetes and Endocrinology, Lund University Diabetes Centre, Malmö, Sweden

⁵ Department of Neurosurgery, Kuopio University Hospital, Kuopio, Finland

⁶ Department of Genetics, Howard Hughes Medical Institute, Yale University, USA

⁷ Department of Epidemiology, Institute of Public Health, University of Southern Denmark, Denmark

⁸ Mental Health Problems and Substance Abuse Services Unit, National Institute for Health and Welfare, Helsinki, Finland

⁹ Department of Public Health, University of Helsinki, Finland

¹⁰ The Broad Institute of Harvard and MIT, Cambridge, Massachusetts, USA

¹¹ Wellcome Trust Sanger Institute, Hinxton, Cambridge, United Kingdom

¹² Department of Mathematical Statistics, Stockholm University, Sweden

Corresponding authors:

Monica Leu, PhD

MEB

Karolinska Institutet

PO Box 281

SE-17177, Stockholm, Sweden

Email: monica.leu@ki.se

And

Samuli Ripatti

FIMM

PO Box 20

FIN-00014 University of Helsinki

Finland

Email: samuli.ripatti@thl.fi

ABSTRACT

A cost efficient way to increase power in a genetic association study is to pool controls from different sources. The genotyping effort can then be directed to large case series. The Nordic Control database, NordicDB has been set up as a unique resource in the Nordic area and the data are available for authorized users via the web-portal (<http://www.nordicdb.org>). The current version of NordicDB pools together high-density genome-wide SNP information from approximately 5000 controls originating from Finnish, Swedish and Danish studies and shows country-specific allele frequencies for SNP markers. The genetic homogeneity of the samples was investigated using multidimensional scaling analysis and pairwise allele frequency differences between the studies. The plot of the first two multidimensional scaling components showed excellent resemblance to the geographical placement of the samples, with a clear NW-SE gradient. We advise researchers to assess the impact of population structure when incorporating NordicDB controls in association studies. This harmonized Nordic database presents a unique genome-wide resource for future genetic association studies in the Nordic countries.

KEY WORDS: common controls, genome-wide data, Nordic Control Database, population stratification

Introduction

Genetic association studies aim to identify variants that predict disease susceptibility, prognosis or therapy response. Many association studies use geographically matched cases and controls, with controls selected and genotyped for each study. Recent successes in reusing existing controls for newly genotyped cases^{1,2} indicate possibilities for designing more cost-effective designs of the next generation of studies. Pooling controls from different studies can be a cost efficient way to increase the power to detect or verify loci of modest effect size.

The Nordic Center of Excellence in Disease Genetics (<http://www.ncoedg.org>), formed by the Joint Committee of the Nordic Medical Research Councils, the Nordic Council of Ministers and the Nordic Research Board, announces the release of the Nordic Control database, ***NordicDB***, providing high-density genome-wide SNP information for approximately 5000 healthy individuals. Currently, NordicDB contains randomly ascertained samples from Finland, Sweden and Denmark. The portal (<http://www.nordicdb.org>), which is under continual development, provides population statistics and web-based tools for efficient use of this resource. Thus, for example, the portal describes quality control (QC) and imputation methods and provides imputed genotype probabilities (HapMap 3 SNPs). This paper introduces the NordicDB and its first release of the imputed data.

Materials and Methods

The Nordic Control Database, ***NordicDB***

NordicDB pools together samples from Finnish, Swedish and Danish studies. The selection of studies came from PIs at NCoEDG sites. These

samples are individuals chosen to be controls in the original case-control studies. Table 1 presents the contributing studies with number of samples and genotyped SNPs, genotyping platform, sample characteristics, sampling location and reference to papers describing the respective studies in more detail.

When constructing NordicDB, each data set was individually subjected to unified genotype QC measures. Briefly, SNPs were aligned to top strand and updated to build 36. We removed markers with ambiguous allele coding, and individuals and markers with more than 5% of data missing, as well as individuals with sex inconsistencies between the genotype data and the indicated sex. 1st or 2nd degree relatives were filtered out based on IBD values greater than 0.2. Based on quality control, on average less than 3% of markers and less than 4% of individuals were excluded from the data sets.

Database and portal

The relational database and the web-based data management application were built using the MOLGENIS application generator^{10,11}. The database contains information and statistics on samples, markers, genotype data releases and sampling location. The sample identifiers were anonymized for the purpose of this database and cannot be linked to the original study identifiers. All SNPs are on top strand alignment and their physical positions are on build 36. Individual level data can be accessed via an application process using the application form available on the portal (<http://www.nordicdb.org/database/Access.html>). Applications will be reviewed by the Nordic Center of Excellence Data Review Board (www.nordicdb.org/drb) consisting of the PIs of the

studies in NordicDB. At the time of preparing this manuscript, the Data Review Board members are affiliated with Lund University (Sweden), Karolinska Institutet (Sweden), Sanger Institute (UK) and the University of Tartu (Estonia). The potential user has to specify the data set(s) that he would be requesting and a brief description of the proposed research use of the requested data. The user must also offer the following assurances that:

- the data will only be used only for approved research, as follows
 - As control data for case – control study design or as population set for population genetics analyses
 - As example data for software algorithm development:
 1. Addressing challenges associated with the analysis of sets of genotypic data.
 2. Detecting differences in allele frequency based on phenotypic data.
 3. Development of advanced analysis tools for the genetic community.
- data confidentiality will be strictly protected
- all applicable laws, local institutional policies, and terms and procedures specific to the study's data access policy for handling anonymized population control data will be followed
- no attempts will be made to identify individual study participants from whom genotype data were obtained using genotype data or by trying to combine genotype data with any other information
- no information regarding the obtained control data set will be shared or sold with third parties
- the contributing investigator(s) who conducted the original study

and the funding organizations involved in supporting the original study will be acknowledged in publications resulting from the analysis of those data. FIMM Technology Center (FTC) will provide information which investigators should be acknowledged.

- an annual report on research progress and publications, where control data has been used, will be submitted to FTC

Finally, the control data use agreement must be co-signed by a group/department/institute leader, who represents the institution for which applicant works. Since data access policies are still being developed, these requirements and policies may change from what is described here without notice. Some data sets will require the original contributing investigator to be contacted and getting his approval in addition to application approval by FTC.

The data can be also accessed via the European Genotype Archive (www.ebi.ac.uk/ega). Specifically, for each sample, researchers will be able to obtain genotype data and an indication of the study from which genotypes originate. Because samples from different studies have been genotyped with different technology and SNP locations, we also provide imputed genotypes (see Section *Imputed data*).

Population structure in the Nordic Control database

Population structure can be measured in terms of differences in allele frequencies and linkage disequilibrium (LD) patterns between subpopulations due to systematic ancestry differences. In genetic association studies, when there are differences in allele frequencies between individuals with different disease/trait status due to population

structure sampling differences by disease status, the false positive error rate is inflated^{12,13}. Population structure must therefore be considered carefully when pooling controls that originate from different populations¹⁴. Recent studies have showed that, even for small isolated populations or for populations within restricted areas, stratification should be evaluated and accounted for when assessing genetic association^{15,6}.

Since the NordicDB samples were collected from different Nordic countries, we investigated potential layers of stratification through the multidimensional scaling (MDS) analysis in PLINK¹⁶. Before performing the MDS analysis we removed non-autosomal SNPs, SNPs in known inverted regions¹⁷, SNPs with MAF < 0.01 and SNPs that failed the Hardy-Weinberg equilibrium test at the significance threshold of 1e-06. Individuals identified as outliers based on the inbreeding coefficient were also excluded (Supplementary material). The MDS analysis was based on SNPs which were common across platforms (approximately 45k SNPs). From the restricted SNP set, only SNPs and individuals with less than 5% missingness were included and only SNPs with low LD¹⁴. In order to prune SNPs in LD, the pairwise genotypic correlation was calculated between all SNPs within windows of 20 SNPs and one SNP was excluded from each pair if the LD was found greater than 0.1. A forward shift of 5 SNPs was assumed between windows. For the purpose of the MDS assessment, a Finnish reference dataset was included. This consists of 81 individuals, 40 individuals collected from the capital area, representing genetically general population and 41 individuals from a Finnish isolate, late-settlement area (LSFIN, described elsewhere^{18,19}). SNPs from the Illumina Human 1M-Duo chip and the Affymetrix Genome-Wide Human SNP Array 6.0 chip were genotyped, resulting in

1,163,280 SNPs after applying QC. The haplotypes in this dataset were phased similarly to the HapMap 3 CEU samples (individuals with NW European ancestry) and Tuscany in Italy (TSI). Figure 1A shows the first 2 axes of genetic variation in NordicDB, CEU HapMap 3 data and the Finnish reference set. The analysis was based on 4809 samples: 2458 Swedish, 2082 Finnish, 161 Danish and 108 from CEU. The plot of the first two MDS components shows excellent resemblance to the geographical placement of the samples (Figure 1B), with a clear NW-SE gradient. To validate the SNP set used in the MDS analysis we compared patterns of variation based on all available SNPs and on the restricted set, using two studies genotyped on the same chip (CAPS and DGI). The results were similar (data not shown).

Table 2 shows summary statistics for allele frequency differences and similarities between study populations. We calculated pairwise F_{ST} values using Weir and Cockerham's approach implemented in the R package *Geneland*²⁰ (see <http://www.nordicdb.org>). The largest differences were those between Finnish and Swedish studies, with magnitude varying according to the location of the Finnish study.

Imputed data

The limited overlap of SNPs across genotyping platforms and chips is a key issue for NordicDB to address. The Illumina²¹ and Affymetrix²² platforms, which differ in terms of genomic coverage, call rate and accuracy, array processing time and ease of use, typically have a SNP overlap of approximately 10%. Thus, in order to provide a harmonized SNP set, imputation of non-overlapping SNPs is required. We use IMPUTE software^{23,24} to impute genotypes of the individuals in NordicDB against a common reference set. Choice of reference population was

based on comparing accuracy of imputing in three data sets (CAPS1, CAPS2 and CAHRES) using different populations, CEU HapMap 2, CEU Hapmap 3, and the combined HapMap 3 European populations CEU and TSI, in a subset of SNPs from chromosomes 21 and 22. Genotypes of directly typed SNPs were compared to their calls after imputing. The subset of SNPs was chosen by first selecting all SNPs that were common to the genotyping platforms that were used in the three studies (see Table 1) and then removing a minimum number of them such that the maximum pairwise r^2 value was 0.2, amongst the remaining SNPs. Genotypes for SNPs in the selected subset were imputed using genotypes of all other typed SNPs on chromosomes 21 and 22. To assess imputation accuracy we calculated the root mean square error of prediction (**RMSEP**) over SNPs and individuals. Writing y_{ki} to denote the observed genotype for SNP k of individual i , and p_{jki} to denote the posterior probability of genotype $j \in \{0,1,2\}$, obtained from IMPUTE), for SNP k , individual i , RMSEP was calculated as

$$RMSEP = \frac{1}{N * K} \sum_{i=1}^N \sum_{k=1}^K \sum_{j=0}^2 p_{jki} (y_{ki} - j)^2 \quad (1)$$

where K is the number of SNPs in the subset of imputed SNPs and N is the number of individuals in the data set. Accurate imputation results are reflected by low RMSEP values. Without exception, lowest RMSEP values were achieved for the CEU and TSI populations combined (Table 3). We therefore used this reference population to impute all datasets in the database. The imputation procedure is described in more detail on the portal, (www.nordicdb.org) where information on how to download imputed data is also provided.

Table 4 presents a summary of imputation accuracy for the Nordic Control database, based on those SNPs that were genotyped in at least

90% of the individuals in the originating study. For chromosome 15, a minimum of 85% of SNPs were called after the imputation at a threshold of 0.9, with a concordance rate of approximately 99% (Table 4, last column).

Discussion

We have described an open resource (NordicDB) that pools GWAS samples from the Nordic countries. With population substructure present across the Nordic populations^{15,6}, there is an obvious need to assess its impact when using NordicDB with a new study population of cases. In dealing with substructure, one should consider adjustment for the main axes of genetic variation²⁵ or selecting a subset of controls that are ancestrally compatible with the cases. An obvious limitation of the Nordic DB is that it includes no environmental variables and therefore users will not be able to adjust for environmental confounders in performing their own association analyses.

The samples in NordicDB were genotyped with different technologies. This called for harmonizing the quality control measures and for imputing the non-overlapping markers using the publicly available LD data from HapMap 3. This allows scientists interested in studying Nordic populations to use their preferred platform to genotype new cases and use NordicDB to pick readily genotyped controls for their studies.

Acknowledgements

Ilkka Lappalainen from EBI is thanked for discussions over the project. NordicDB is financially supported by the Nordic Center of Excellence in Disease Genetics, Wallenberg Foundation, FP6 coordinated action PHOEBE (Promoting Harmonization of Epidemiological Biobanks in

Europe), the Wallenberg Consortium North, Sweden, Center of Excellence for Complex Disease Genetics of the Academy of Finland (grants 213506, 129680) the Biocentrum Helsinki Foundation, The Nordic Centre of Excellence (NCoE) Programme in Molecular Medicine. KH acknowledges support from the Swedish Research Council (grant number 523-2006-972).

Conflict of interest

The authors declare no conflict of interest.

References

- [1] Wellcome Trust Case Control Consortium: Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* 2007; **447(7145)**: 661-78.
- [2] Wrensch M, Jenkins RB, Chang JS, et al: Variants in the CDKN2B and RTEL1 regions are associated with high-grade glioma susceptibility. *Nat Genet* 2009; **41(8)**: 905-8.
- [3] Zheng SL, Sun J, Wiklund F et al: Cumulative association of five genetic variants with prostate cancer. *N Engl J Med* 2008; **358(9)**: 910-9.
- [4] Einarsdóttir K, Humphreys K, Bonnard C et al: Linkage disequilibrium mapping of CHEK2: common variation and breast cancer risk. *PLoS Med* 2006; **3**:e168.
- [5] Diabetes Genetics Initiative of Broad Institute of Harvard and MIT, Lund University, and Novartis Institutes of BioMedical Research et al: Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. *Science* 2007; **316(1)**: 1331-6.
- [6] Jakkula E, Rehnström K, Varilo T et al: The Genome-wide Patterns of Variation Expose Significant Substructure in a Founder Population. *Am J Hum Genet* 2008; **83(6)**: 787-94.

- [7] Bilguvar K, Yasuno K, Niemelä M et al: Susceptibility loci for intracranial aneurysm in European and Japanese populations. *Nat Genet* 2008; **40(12)**: 1472-7.
- [8] McEvoy BP, Montgomery GW, McRae AF et al: Geographical structure and differential natural selection among North European populations. *Genome Res* 2009; **19(5)**: 804-14.
- [9] Aulchenko YS, Ripatti S, Lindqvist I et al: Loci influencing lipid levels and coronary heart disease risk in 16 European population cohorts. *Nat Genet* 2009; **41(1)**: 47-55.
- [10] Swertz MA, de Brock EO, van Hijum S et al: Molecular Genetics Information System (MOLGENIS): alternatives in developing local experimental genomics databases. *Bioinformatics* 2004; **20(13)**: 2075-83.
- [11] MOLGENIS web page: <http://molgenis.sourceforge.net/>
- [12] Freedman ML, Reich D, Penney KL et al: Assessing the impact of population stratification on genetic association studies. *Nat Genet* 2004; **36(4)**: 388-93.
- [13] Tian C, Gregersen PK, Seldin MF: Accounting for ancestry: population substructure and genome-wide association studies. *Hum Molec Genet* 2008; **17(R2)**: R143-50.
- [14] Yu K, Wang Z, Li Q, Wacholder S, Hunter DJ et al: Population substructure and control selection in genome-wide association studies. *PLoS ONE* 2008; **3(7)**: e2551. doi:10.1371/journal.pone.0002551
- [15] Salmela E, Lappalainen T, Fransson I et al: Genome-Wide Analysis of Single Nucleotide Polymorphisms Uncovers Population Structure in Northern Europe. *PLoS ONE* 2008; **3(10)**: e3519. doi:10.1371/journal.pone.0003519.
- [16] 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(3)**: 559-75.

[17] Price AL, Weale ME, Patterson N et al: Long-Range LD Can Confound Genome Scans in Admixed Populations. *Am J Hum Genet* 2008; **83(1)**: 132-5

[18] Nevanlinna HR: The Finnish population structure. A genetic and genealogical study. *Hereditas* 1972; **71(2)**: 195-235.

[19] Varilo T, Laan M, Hovatta I et al: Linkage disequilibrium in isolated populations: Finland and a young sub-population of Kuusamo. *Eur J Hum Genet* 2008; **8(8)**: 604-12.

[20] Guillot G, Santos F, Estoup A (2009) Inference in population genetics with Geneland: a sensitivity analysis to spatial sampling scheme, null alleles and isolation by distance. Submitted.

[21] <http://www.illumina.com>

[22] <http://affymetrix.com/index.affx>

[23] Webpage for IMPUTE download:

<https://mathgen.stats.ox.ac.uk/impute/impute.html>

[24] Marchini J, Howie B, Myers S, McVean G, Donnelly P: A new multipoint method for genome-wide association studies via imputation of genotypes. *Nat Genet* 2007; **39(7)**: 906-13.

[25] Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D: Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* 2006; **38(8)**: 904-9.

Titles and legends to figures

Figure 1A: Top axes of genetic variation in the Nordic Control Database, NordicDB (4620 samples) contrasted with the HapMap CEU (108 samples) and a Finnish HapMap reference population (81 samples). The MDS analysis was performed on approximately 45000 SNPs that were common between genotyping platforms. The controls are part of the following studies: Cancer Prostate in Sweden (**CAPS**) 1&2, Cancer and Hormonal Replacement in Sweden (**CAHRES**), Diabetes Genetics Initiative in Western Finland and Southern Sweden (**DGI-FIN** and **DGI-SWE**), **SGENE** and **MS** in the Helsinki region, **Aneurysm** study in the Helsinki region, GenomEUtwin Denmark (**GenomEUtwin-DK**), GenomEUtwin Sweden (**GenomEUtwin-SWE**) and GenomEUtwin Finland (**GenomEUtwin-FIN**).

Figure 1B: Geographical map of Scandinavia with three countries highlighted to show the origin of the samples in *Figure 1A*: Finland (red), Sweden (green) and Denmark (yellow).

TABLE 1

GWAS studies contributing controls to the Nordic Control Database					
<i>Study</i>	<i>Number of samples</i>	<i>Number of SNPs</i>	<i>Genotyping platform</i>	<i>Sampling location</i>	<i>Sample characteristics</i>
Cancer Prostate in Sweden (CAPS) ^{1 3}	502	492555	Affymetrix 550K	Central and Northern Sweden	Males; Mean age: 66.4 ± 7.1 yrs
Cancer Prostate in Sweden (CAPS) ^{2 3}	519	440092	Affymetrix 5.0	Central and Northern Sweden	Males; Mean age: 66.4 ± 7.1 yrs
Cancer And Hormonal Replacement in Sweden (CAHRES) ⁴	764	561274	Illumina HumanHap-550-v3	Sweden	Females; Mean age: 63 ± 6.5 yrs
Diabetes Genetics Initiative (DGI) ⁵	1467	496963	Affymetrix 550K	Southern Sweden & Western Finland (<u>Botnia</u>)	654 males/ 701 females; Mean age: 58.3 ± 6.5 yrs (SWE); 59±10 yrs (FIN)
SGENE and MS ⁶	241	318212//314691	Illumina HumanHap-300-v2.0/v1.0//	Helsinki region	148 males/ 93 females; Mean age 43.1 ± 11 yrs (SGENE)
Aneurysm study ⁷	697	341389	Illumina HumanCNV-370-v1.0	<u>Kupio and Helsinki</u>	304 males/ 393 females; Mean age 58.1 ± 18.66 yrs
GenomEUtwin-DK ^{8,9}	173	318212	Illumina HumanHap-300-v2.0	Denmark	Females; Age range 20-80 yrs
GenomEUtwin-SWE ^{8,9}	302	318212	Illumina HumanHap-300-v2.0	Sweden	Females; Age range 20-80 yrs
GenomEUtwin-FIN ^{8,9}	157	318212	Illumina HumanHap-300-v2.0	Finland	13males/ 144 females; Age range 20-80 yrs

TABLE 2

Pairwise F_{ST} values for datasets in the Nordic Control Database											
<i>Study</i> *	CAPS 1&2	CAHRES	DGI-FIN	DGI-SWE	SGENE and MS	Aneurysm	Genom EUtwin-DK	Genom EUtwin-SWE	Genom EUtwin-FIN	CEU HapMap 3	Finnish reference
CAPS 1&2	-	0	0.001	0	0.004	0.004	0	0	0.003	0.001	0.006
CAHRES		-	0.001	0	0.004	0.004	0	0	0.004	0	0.006
DGI-FIN			-	0.001	0.002	0.002	0.001	0.001	0.001	0.002	0.004
DGI-SWE				-	0.006	0.005	0	0	0.004	0.001	0.005
SGENE and MS					-	0.001	0.007	0.005	0.001	0.007	0.001
Aneurysm						-	0.005	0.004	0	0.006	0.001
Genom EUtwin-DK							-	0.001	0.004	0.001	0.006
Genom EUtwin-SWE								-	0.003	0.001	0.005
Genom EUtwin-FIN									-	0.005	0.001
CEU HapMap 3										-	0.007
Finnish reference											-

*More complete names of the studies are provided in Table 1.

In order to easily distinguish F_{ST} values between countries, the following color-coding was used: Finland-Finland: **red**, Sweden-Sweden: **green**, Sweden-Finland: **blue**, Sweden-Denmark: **orange**, Denmark-Finland: **brown**.

Calculations were based on approximately 2500 SNPs, chosen with a low LD between each other (pairwise LD values were calculated within windows of 50 SNPs and one SNP was excluded from each pair if LD was found greater than 0.006. A forward shift of 5 SNPs was used between windows.

TABLE 3

<i>Comparison of imputation accuracy for three reference populations*</i>			
<i>Study/ reference population</i>	<i>CEU HapMap 2</i>	<i>CEU HapMap 3</i>	<i>CEU+TSI HapMap 3</i>
Cancer Prostate in Sweden (CAPS) 1			
Chr 21	0.159	0.151	0.144
Chr 22	0.179	0.180	0.175
Cancer Prostate in Sweden (CAPS) 2			
Chr 21	0.164	0.151	0.144
Chr 22	0.194	0.189	0.187
Cancer And Hormonal Replacement in Sweden (CAHRES)			
Chr 21	0.079	0.069	0.064
Chr 22	0.081	0.073	0.068
*Mean prediction error (RMSEP values) over SNPs and individuals. Calculations are based on overlapping SNPs between platforms.			

TABLE 4

<i>Imputation accuracy for the Nordic Control Database*</i>			
<i>Study</i>	<i>Number of imputed genotypes[#]</i>	<i>% called at 0.9 threshold</i>	<i>% concordance^{##}</i>
Cancer Prostate in Sweden (CAPS) 1&2 ³	9311088	89.46	98.41
Cancer And Hormonal Replacement in Sweden (CAHRES) ⁴	11134120	92.76	99.02
Diabetes Genetics Initiative (DGI) ⁵	12787169	89.45	98.35
SGENE and MS ⁶	1721751	84.99	98.04
Aneurysm study ⁷	5925910	86.17	98.15
GenomEUtwin-DK ^{8,9}	1316505	86.53	98.46
GenomEUtwin-SWE ^{8,9}	2453897	86.35	98.43
GenomEUtwin-FIN ^{8,9}	1245029	85.47	98.11
<p>* Calculations are based on chromosome 15 [#] Calculated as the number of typed SNPs (in at least 90% of the individuals) multiplied by the number of individuals in the data set. ^{##} The concordance is based on the SNPs in the 2nd column that were called after the imputation using a 0.9 threshold.</p>			



