European Journal of Human Genetics (2010) 18, 1322–1326

www.nature.com/ejhg

ARTICLE

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

Monica Leu*[,1,2], Keith Humphreys[1], Ida Surakka[2,3], Emil Rehnberg[1], Juha Muilu[2], Päivi Rosenström[2], Peter Almgren[4], Juha Jääskeläinen[5], Richard P Lifton[6], Kirsten Ohm Kyvik[7], Jaakko Kaprio[2,8,9], Nancy L Pedersen[1], Aarno Palotie[2,10,11], Per Hall[1], Henrik Grönberg[1], Leif Groop[4], Leena Peltonen[2,3,10,11], Juni Palmgren[1,12] and Samuli Ripatti*[,2,3]

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 through the web portal (http://www.nordicdb.org). The current version of NordicDB pools together high-density genome-wide SNP information from ~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 (MDS) analysis and pairwise allele frequency differences between the studies. The plot of the first two MDS 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.
European Journal of Human Genetics (2010) **18,** 1322–1326; doi:10.1038/ejhg.2010.112; published online 28 July 2010

Keywords: 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 ~5000 healthy individuals. At present, 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 >5% of data missing, as well as individuals with sex inconsistencies between the genotype data and the indicated sex. First- or second-degree relatives were filtered out on the basis of IBD values >0.2. On the basis of QC, on average, <3% of markers and <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 (Molgenis; http://molgenis.sourceforge.net/).[10] 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

[1]Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden; [2]Institute for Molecular Medicine, Finland, FIMM, University of Helsinki, 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, Chevy Chase, MD, USA; [7]Department of Epidemiology, Institute of Public Health, University of Southern Denmark, Odense Area, 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, Helsinki, Finland; [10]The Broad Institute of Harvard and MIT, Cambridge, MA, USA; [11]Department of Human Genetics, Wellcome Trust Sanger Institute, Hinxton, Cambridge, United Kingdom; [12]Department of Mathematical Statistics, Stockholm University, Stockholm, Sweden
*Correspondence: Dr M Leu, Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, PO Box 281, Nobels vag 12a, SE-17177, Stockholm, Sweden. Tel: +46 8 524 82 337; Fax: +46 8 31 49 75; E-mail: monica.leu@ki.se and Dr S Ripatti, FIMM, PO Box 20, FIN-00014 University of Helsinki, Helsinki, Finland. Tel: +358 20 610 8159; Fax: +358 9 47448480; E-mail: samuli.ripatti@thl.fi
Received 29 October 2009; revised 24 March 2010; accepted 4 June 2010; published online 28 July 2010

Table 1 GWAS studies contributing controls to the Nordic Control Database

Study	*Number of samples*	*Number of SNPs*	*Genotyping platform*	*Sampling location*	*Sample characteristics*
Cancer Prostate in Sweden (CAPS) 1[3]	502	492 555	Affymetrix 550K	Central and Northern Sweden	Males; mean age: 66.4 ± 7.1 years
Cancer Prostate in Sweden (CAPS) 2[3]	519	440 092	Affymetrix 5.0	Central and Northern Sweden	Males; mean age: 66.4 ± 7.1 years
Cancer And Hormonal Replacement in Sweden (CAHRES)[4]	764	561 274	Illumina HumanHap-550-v3	Sweden	Females; mean age: 63 ± 6.5 years
Diabetes Genetics Initiative (DGI)[5]	1467	496 963	Affymetrix 550K	Southern Sweden and Western Finland (Botnia)	654 males/701 females; mean age: 58.3 ± 6.5 years (SWE); 59 ± 10 years (FIN)
SGENE and MS[6]	241	318 212// 314 691	Illumina HumanHap-300-v2.0/v1.0//	Helsinki region	148 males/93 females; mean age 43.1 ± 11 years (SGENE)
Aneurysm study[7]	697	341 389	Illumina HumanCNV-370-v1.0	Kupio and Helsinki	304 males/393 females; mean age 58.1 ± 18.66 years
GenomEUtwin-DK[8,9]	173	318 212	Illumina HumanHap-300-v2.0	Denmark	Females; age range 20–80 years
GenomEUtwin-SWE[8,9]	302	318 212	Illumina HumanHap-300-v2.0	Sweden	Females; age range 20–80 years
GenomEUtwin-FIN[8,9]	157	318 212	Illumina HumanHap-300-v2.0	Finland	13 males/144 females; age range 20–80 years

their physical positions are on build 36. Individual level data can be accessed through 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 (http://www.nordicdb.org/drb) consisting of the PIs of the studies in NordicDB. At the time of preparing this paper, the Data Review Board members were affiliated to the 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 with or sold to 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. The FIMM Technology Center (FTC) will provide information regarding which investigators should be acknowledged.
- An annual report on research progress and publications, in which control data have been used, will be submitted to the FTC.

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

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

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 sub-populations 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.[11,12] Therefore, population structure must be considered carefully when pooling controls that originate from different populations.[13] Recent studies have shown that, even for small isolated populations or for populations within restricted areas, stratification should be evaluated and accounted for when assessing genetic association.[6,14]

As the NordicDB samples were collected from different Nordic countries, we investigated potential layers of stratification through the multidimensional scaling (MDS) analysis in PLINK.[15] Before performing the MDS analysis, we removed non-autosomal SNPs, SNPs in known inverted regions,[16] 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 available at http://nordicdb.org/database/Data.html). The MDS analysis was based on SNPs which were common across platforms ($\sim$45k SNPs). From the restricted SNP set, only SNPs and individuals with $<5\%$ missingness were included and only SNPs with low LD.[13] To prune SNPs in LD, the pairwise genotypic correlation was calculated between all SNPs within windows of 20 SNPs and 1 SNP was excluded from each pair if the LD was found >0.1. A forward shift of five SNPs was assumed between windows. For the purpose of the MDS assessment, a Finnish reference data set 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[17,18]). SNPs from the Illumina Human 1M-Duo chip (Illumina, San Diego, CA, USA) and the Affymetrix Genome-Wide Human SNP Array 6.0 chip (Affymetrix, Santa Clara, CA, USA) were

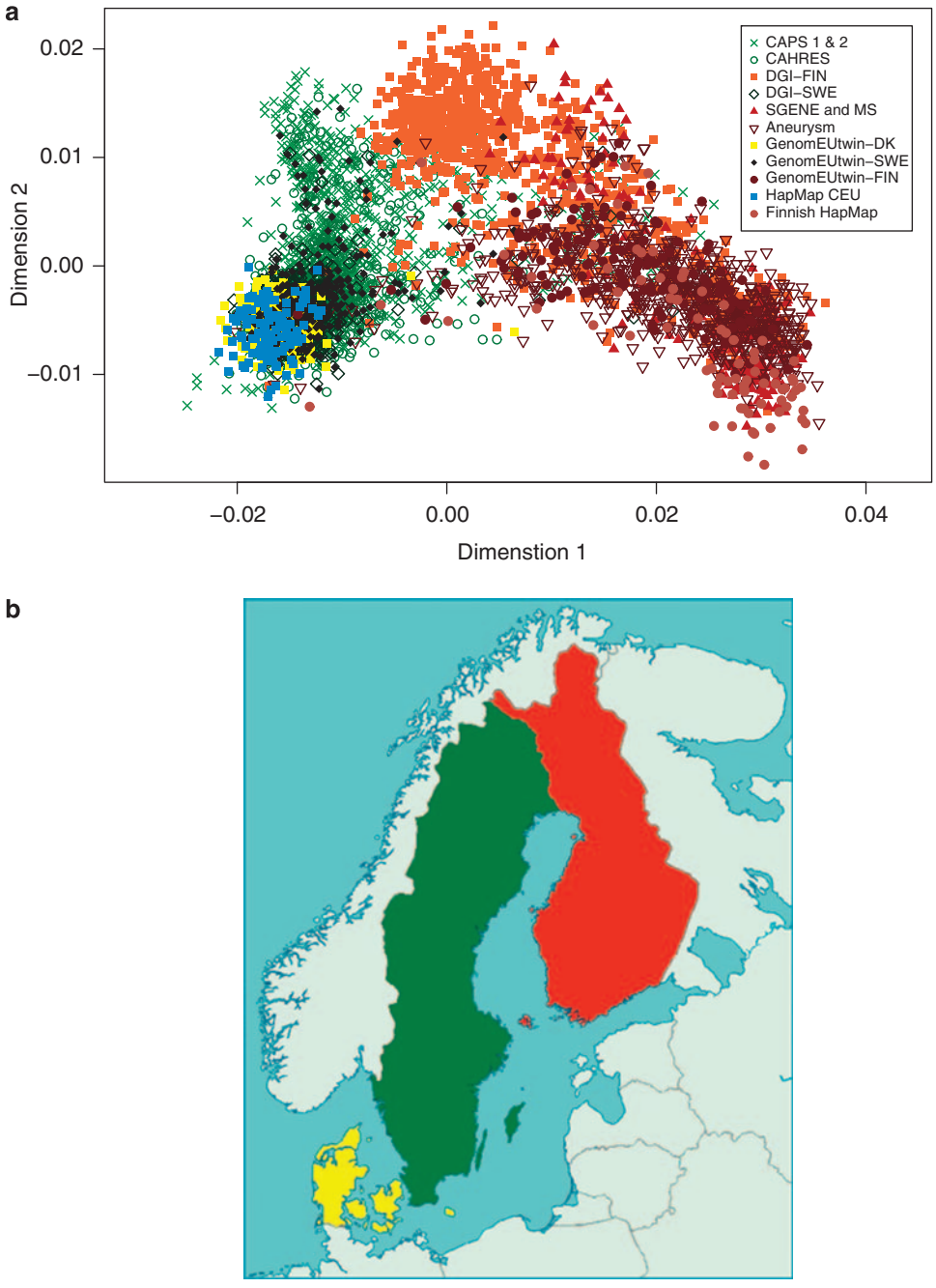

Figure 1 (**a**) Top axes of genetic variation in the Nordic Control Database, NordicDB (4620 samples) contrasted with the HapMap CEU (108 samples) and a Finnish HapMap reference population (81 samples). The MDS analysis was performed on ~45 000 SNPs that were common between genotyping platforms. The controls are part of the following studies: Cancer Prostate in Sweden (CAPS) 1 and 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). (**b**) Geographical map of Scandinavia with three countries highlighted to show the origin of the samples in panel **a**: Finland (red), Sweden (green) and Denmark (yellow).

genotyped, resulting in 1 163 280 SNPs after applying QC. The haplotypes in this data set were phased similarly to the HapMap 3 CEU samples (individuals with NW European ancestry) and Tuscany in Italy (TSI). Figure 1a shows the first two 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[19]

Table 2 Pairwise F_{ST} values for data sets in the Nordic Control Database

Study[a]	CAPS 1 and 2	CAHRES	DGI-FIN	DGI-SWE	SGENE and MS	Aneurysm	Genom EUtwin-DK	Genom EUtwin-SWE	Genom EUtwin-FIN	CEU HapMap 3[b]	Finnish reference
CAPS 1 and 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											—

To easily distinguish F_{ST} values between countries, the following text formatting was used: Finland–Finland: roman, Sweden–Sweden: bold, Sweden–Finland: italics, Sweden–Denmark: underline, Denmark–Finland: bold italics.
Calculations were based on ~2500 SNPs, chosen with a low LD between each other (pairwise LD values were calculated within windows of 50 SNPs and 1 SNP was excluded from each pair if LD was found >0.006. A forward shift of five SNPs was used between windows.
[a]More complete names of the studies are provided in Table 1.
[b]All the values in this column pertain to general European reference population, thus neither Finnish, Swedish or Danish.

(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 Illumina (http://www.illumina.com) and Affymetrix (http://affymetrix.com/index.affx) platforms, which differ in terms of genomic coverage, call rate and accuracy, array processing time and ease of use, typically have an SNP overlap of ~10%. Thus, to provide a harmonized SNP set, imputation of non-overlapping SNPs is required. We use Impute software (University of Oxford, Oxford, UK; https://mathgen.stats.ox.ac.uk/impute/impute.html)[20] to impute genotypes of the individuals in NordicDB against a common reference set. Choice of the reference population was based on comparing accuracy of imputing in three data sets (namely 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 with 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, among 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

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

where K is the number of SNPs in the subset of imputed SNPs and N 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). Therefore, we used this reference population to impute all data sets in the database. The imputation procedure is described in more detail on the portal, (http://www.nordicdb.org) in which 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 ~99% (Table 4, last column).

Table 3 Comparison of imputation accuracy for three reference populations[a]

Study/reference population	CEU HapMap 2	CEU HapMap 3	CEU+TSI HapMap 3
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

[a]Mean prediction error (RMSEP values) over SNPs and individuals. Calculations are based on overlapping SNPs between platforms.

DISCUSSION

We have described an open resource (NordicDB) that pools GWAS samples from the Nordic countries. With population substructure present across the Nordic populations,[6,14] 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 variation[21] 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 QC 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.

Table 4 Imputation accuracy for the Nordic Control Database[a]

Study	*Number of imputed genotypes*[b]	*% Called at 0.9 threshold*	*% Concordance*[c]
Cancer Prostate in Sweden (CAPS) 1 and 2[3]	9 311 088	89.46	98.41
Cancer And Hormonal Replacement in Sweden (CAHRES)[4]	11 134 120	92.76	99.02
Diabetes Genetics Initiative (DGI)[5]	12 787 169	89.45	98.35
SGENE and MS[6]	1 721 751	84.99	98.04
Aneurysm study[7]	5 925 910	86.17	98.15
GenomEUtwin-DK[8,9]	1 316 505	86.53	98.46
GenomEUtwin-SWE[8,9]	2 453 897	86.35	98.43
GenomEUtwin-FIN[8,9]	1 245 029	85.47	98.11

[a]Calculations are based on chromosome 15.
[b]Calculated as the number of typed SNPs (in at least 90% of the individuals) multiplied by the number of individuals in the data set.
[c]The concordance is based on the SNPs in the second column that were called after the imputation using a 0.9 threshold.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

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 and 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).

1 Wellcome Trust Case Control Consortium: Genome-wide association study of 14 000 cases of seven common diseases and 3000 shared controls. *Nature* 2007; **447**: 661–678.
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**: 905–908.
3 Zheng SL, Sun J, Wiklund F *et al*: Cumulative association of five genetic variants with prostate cancer. *N Engl J Med* 2008; **358**: 910–919.
4 Einarsdöttir 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 ResearchSaxena R, Voight BF, Lyssenko V *et al*: Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. *Science* 2007; **316**: 1331–1336.
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**: 787–794.
7 Bilguvar K, Yasuno K, Niemelä M *et al*: Susceptibility loci for intracranial aneurysm in European and Japanese populations. *Nat Genet* 2008; **40**: 1472–1477.
8 McEvoy BP, Montgomery GW, McRae AF *et al*: Geographical structure and differential natural selection among North European populations. *Genome Res* 2009; **19**: 804–814.
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**: 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**: 2075–2083.
11 Freedman ML, Reich D, Penney KL *et al*: Assessing the impact of population stratification on genetic association studies. *Nat Genet* 2004; **36**: 388–393.
12 Tian C, Gregersen PK, Seldin MF: Accounting for ancestry: population substructure and genome-wide association studies. *Hum Molec Genet* 2008; **17**: R143–R150.
13 Yu K, Wang Z, Li Q *et al*: Population substructure and control selection in genome-wide association studies. *PLoS ONE* 2008; **3**: e2551.
14 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**: e3519.
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 Price AL, Weale ME, Patterson N *et al*: Long-range LD can confound genome scans in admixed populations. *Am J Hum Genet* 2008; **83**: 132–135.
17 Nevanlinna HR: The Finnish population structure. A genetic and genealogical study. *Hereditas* 1972; **71**: 195–235.
18 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**: 604–612.
19 Guillot G, Santos F, Estoup A: Inference in population genetics with Geneland: a sensitivity analysis to spatial sampling scheme, null alleles and isolation by distance, 2009. Submitted.
20 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**: 906–913.
21 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**: 904–909.