ORIGINAL ARTICLE

Subpopulation difference scanning: a strategy for exclusion mapping of susceptibility genes

E Salmela, O Taskinen, J K Seppänen, P Sistonen, M J Daly, P Lahermo, M-L Savontaus, J Kere

J Med Genet 2006;**43**:590–597. doi: 10.1136/jmg.2005.038414

See end of article for authors' affiliations

Correspondence to: Dr J Kere, Karolinska Institutet, Department of Biosciences at Novum, CBT, Seventh Floor, 14157 Huddinge, Sweden; juha.kere@biosci.ki.se

Received 11 September 2005
Revised version received 20 December 2005
Accepted for publication 21 December 2005
Published Online First 27 January 2006

Background: Association mapping is a common strategy for finding disease-related genes in complex disorders. Different association study designs exist, such as case-control studies or admixture mapping.
Methods: We propose a strategy, subpopulation difference scanning (SDS), to exclude large fractions of the genome as locations of genes for complex disorders. This strategy is applicable to genes explaining disease incidence differences within founder populations, for example, in cardiovascular diseases in Finland.
Results: The strategy consists of genotyping a set of markers from unrelated individuals sampled from subpopulations with differing disease incidence but otherwise as similar as possible. When comparing allele or haplotype frequencies between the subpopulations, the genomic areas with little difference can be excluded as possible locations for genes causing the difference in incidence, and other areas therefore targeted with case-control studies. As tests of this strategy, we use real and simulated data to show that under realistic assumptions of population history and disease risk parameters, the strategy saves efforts of sampling and genotyping and most efficiently detects genes of low risk—that is, those most difficult to find with other strategies.
Conclusion: In contrast to admixture mapping that uses the mixing of two different populations, the SDS strategy takes advantage of drift within highly related subpopulations.

A common strategy for association mapping of disease genes is to compare allele differences between a group of cases who have the disease and a group of controls who do not. This approach has been used for multifactorial diseases with their complex genotype-phenotype relationships.[1 2] There, the situation can be brought closer to the monogenic one (that is, the genetic homogeneity can be maximised) by applying strict phenotypic criteria in the selection of subjects, by considering subphenotypes, and possibly by studying genetic isolates. However, there are inherent problems in defining disease status due to low penetrance, high phenocopy rates, and often also late disease onset.

We propose here a strategy that we call subpopulation difference scanning (SDS). To avoid the tedious determination of disease status, it compares not cases and controls but instead individuals randomly sampled from two subpopulations that have differing disease incidence. The strategy is outlined in more detail below, and we also use simulations and real genotype and incidence data to study its benefits and limitations.

THE SUBPOPULATION DIFFERENCE SCANNING STRATEGY

Outline

If regional differences in disease incidence (fig 1A, B) are partly determined genetically, it is intuitive that a genetic variant with similar variation trends can explain them (fig 1C, D, respectively), whereas one with a different trend (fig 1E) or one with minor variation (fig 1F) cannot. Thus, the genomic location of the incidence difference causing variant could be narrowed down by sampling the unselected population from regions of high and low incidence, rather than cases and controls, and genotyping the samples with a set of markers across the genome. The genomic areas where no frequency difference between the regional samples exceeds a given threshold could be excluded from subsequent analyses.

The method thus avoids case-control sampling and phenotype ascertainment by using subpopulation pairs where, during population history, cases have been enriched in one subpopulation and controls in the other. Homogeneity of non-genetic factors between the subpopulations is essential to ensure that the observed incidence differences are predominantly of genetic origin.

The procedure will save in cost and labour by eliminating the need for an initial careful phenotype ascertainment and by reducing the potentially interesting genome area. The major determinants of cost are then the sample sizes and number of loci needed to achieve a sufficiently reliable exclusion. As we show by the simulations detailed below, genotyping samples of as few as 100–200 random individuals from two subpopulations for loci at 1 cM distance may suffice to exclude 90% of the genome for acute myocardial infarction (AMI) susceptibility genes.

Population background

The proposed strategy is based on allele and haplotype frequency differences between subpopulations. One possible origin of such differences is shown (fig 2A), in which an initial population splits into several subpopulations that first grow in relative isolation. Founder effects and genetic drift cause shifts in the subpopulation allele frequencies. These shifts will mostly remain visibly large even though later subpopulation contacts can somewhat smooth them. Because the founder effects and drift are random, the size of the shifts they produce varies between loci. If such a shift happens in a

Abbreviations: AMI, acute myocardial infarction; MCMC, Markov chain Monte Carlo; SDS, subpopulation difference scanning

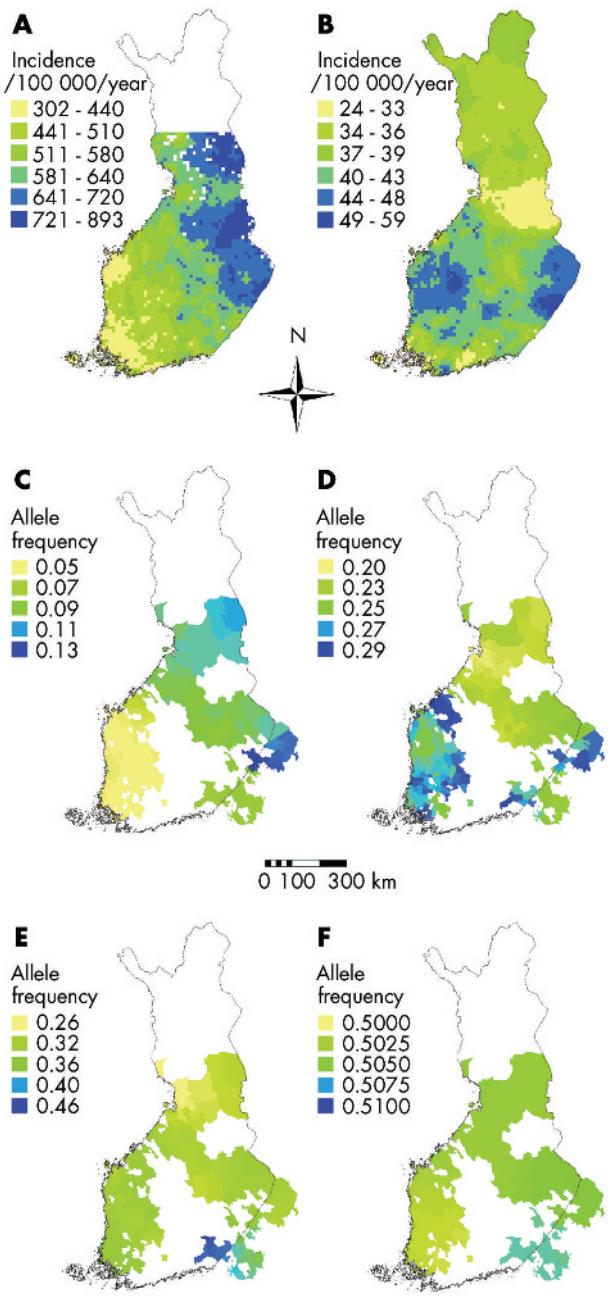

Figure 1 (A) Age standardised incidence of acute myocardial infarction among 35–74 year old men. The map is based on cross sectional years 1983, 1988, and 1993. (B) Age standardised incidence of type 1 diabetes mellitus among children younger than 15 years in 1987–1996. Frequency of one allele of markers: (C) D9S1677; (D) D20S196; (E) D5S641; and (F) D8S277.

disease causing variant, the disease incidence will also differ between the subpopulations; conversely, only the loci that show a certain amount of difference are possible genetic causes of incidence variation of a disease. A disease causing variant with a given difference is thus easiest to locate when average frequency differences in loci are small—that is, when the subpopulations of study are closely related.

The way such frequency differences are reflected in the incidence of a disease is shown in more detail (fig 2B); two populations are depicted, both consisting of carriers and non-carriers of a disease variant. In each population, penetrance determines the fraction of diseased individuals among

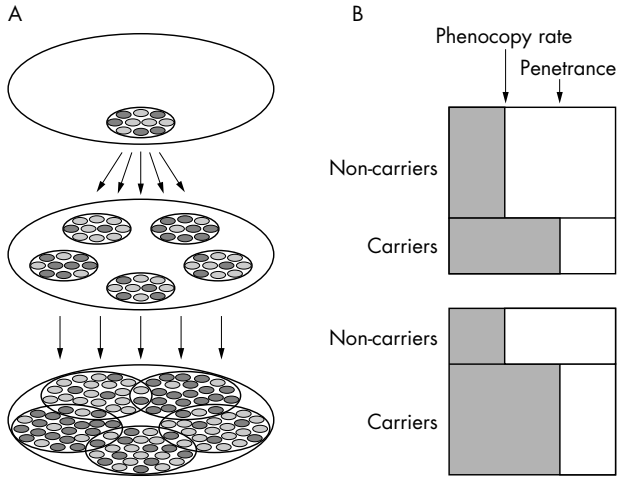

Figure 2 (A) Origin of allele and haplotype frequency differences between subpopulations through founder effects and drift. In the first step, a small population divides into several isolated subpopulations with distinct founder effects on allele frequencies. In the second step, drift modifies these differences further, resulting in regional allele frequency differences within the population. (B) Incidence variation of a disease. Two subpopulations are presented (large squares; individuals with a complex disease are shown in grey, healthy in white). Penetrances of the susceptibility gene are equal in the subpopulations, as are phenocopy rates, but the upper subpopulation has fewer carriers of the susceptibility gene than the lower, and as a result their disease incidences also differ.

carriers, and phenocopy rate the corresponding fraction among non-carriers. Although the populations are similar with respect to penetrance and phenocopy rate, if their variant frequencies (and thus their carrier frequencies) differ, their total proportions of diseased individuals will also differ.

The fact that many key features of the proposed origin of differences (fig 2A) characterise Finnish population history makes Finland a promising test ground for the SDS strategy. Indeed, the small numbers of founders and immigration waves, gradual habitation of the inland regions, relative isolation both within the country and from neighbouring populations, and a fairly recent population growth have led to genetic differences within the country.[3–7] Several diseases also vary in their incidence: in addition to AMI and type 1 diabetes (fig 1A, B) and many monogenic diseases,[8] examples include cerebrovascular diseases, dementia, and malignant neoplasms of colon, prostate ,and ovary, all with between province mortality differences of more than 2.2 fold (StatFin Online Service, Statistics Finland). Some of these differences undoubtedly result from variation in non-genetic risk factors, but, for example, the east-west incidence difference in coronary heart disease seems to be partly genetic, because non-genetic factors do not fully explain it and it has persisted despite a general decline in incidences.[9 10]

Evaluating the strategy

Intuitiveness alone is no proof to the efficiency of the SDS strategy. For instance, the observed incidence differences of a disease of interest might be produced by a variant whose frequency difference is not very large relative to other loci, and thus the exclusion of genome areas would be inefficient. In addition, if the detection of even a large frequency difference would require genotyping a vast number of samples or markers, the strategy would save little in cost or effort.

We have studied the determinants of the efficiency of the strategy in the following sections. Firstly, we tested how different population history parameters affect the frequency

difference distribution in simulations, and compared the result to the distribution in a real dataset from the Finnish population. We then assessed how sample size, genotyping density, and population history influence the efficiency of detecting these differences. Finally, we investigated how the frequency difference of a disease causing variant relates to the incidence difference that the disease will show, both theoretically and in the case of AMI.

MATERIALS AND METHODS

Simulated data

We simulated a population history resembling the one shown (fig 2A). In the simulation, an initial population split into three, and the daughter populations grew exponentially for a given number of generations to a given size in the absence of selection, mutation, or migration. Assuming no migration is realistic for geographically distant regions in a sparsely inhabited country. Generations were non-overlapping, and mating was random with the exception that sibling mating was prevented. In the final generation, non-sibling individuals were randomly sampled from two of the daughter populations.

Further details of the simulation, including a rationale for parameter values, are described in appendix A.

Genotype data

We chose to use microsatellites as the demonstration data for their high information content and sufficient stability to record subpopulation differences within the time scale relevant for the Finnish population. Alternatively, as presented for the simulations, haplotype tagging SNPs would yield similar information.

We genotyped 30 autosomal microsatellite markers in 465 Finnish subjects. All subjects were anonymous, unrelated male blood donors aged 40–55 and had given an informed consent. As the population samples are unidentifiable, no ethics approval was required for their use. The birthplaces of each subject's grandparents lay near each other. The markers (D2S117, D2S2382, D3S1278, D3S1566, D3S1569, D4S391, D4S405, D4S413, D5S641, D5S647, D6S264, D6S462, D7S510, D7S640, D8S277, D9S158, D9S273, D9S286, D9S1677, D10S185, D11S925, D12S345, D13S153, D13S171, D13S285, D13S1265, D15S131, D20S100, D20S107, D20S196) were amplified from 10 ng of DNA in 5 µl volume using fluorescently labelled commercial primers (from Linkage Mapping Set MD-10; Applied Biosystems, Foster City, CA, USA) and Ampli*Taq* Gold DNA polymerase (Applied Biosystems), detected with MegaBACE 1000 capillary electrophoresis instrument (Molecular Dynamics/Amersham Biosciences, Sunnyvale, CA, USA), and analysed with Genetic Profiler allele calling software (version 1.1;Molecular Dynamics/Amersham Biosciences) according to the manufacturers' instructions.

Subpopulation differences

In the real genotype data, frequencies of the observed alleles were calculated in the provinces best corresponding to the areas of high and low AMI incidence (fig 1A): central east (111 samples), central west (46), and southwest (38) Finland. Absolute allele frequency differences were then calculated in east versus west and east versus southwest, and the maximum difference per marker was recorded in each comparison.

In the simulated data, the frequency differences of the number of initially equifrequent haplotypes per locus were calculated between the two sampled daughter populations. Only the largest absolute difference was recorded for each locus. The distribution of the differences was used to determine an effective exclusion threshold and the risk of excluding, with the threshold, a disease gene with a given frequency difference (details in appendix A). Both the threshold and risk were studied from several simulation settings of population history factors, initial LD, number of haplotypes, sample size, and genotyping density.

To examine the relationship of the frequency differences and incidence differences, we constructed a simple model that consists of two subpopulations, each in Hardy-Weinberg equilibrium, and a disease with high incidence in one subpopulation and low in the other. The whole incidence difference was assumed to be genetic and due to a single predisposing allele that was dominant with incomplete penetrance; other risk factors, both genetic and non-genetic, were assumed to be identical in both subpopulations and their distribution in individuals independent of the presence or absence of the predisposing allele. (Note that even when the observed incidence difference of the disease of interest is attributable to several factors, the model is still valid if the power calculations use, instead of the whole incidence difference, the estimated difference due to a single allele—that is, the difference that would result if the assumption of identical distribution of all other factors would hold.) Appendix A gives further model details and estimates of relevant AMI incidences.

Incidence and allele frequency maps

Frequency maps of four microsatellite alleles and incidence maps of AMI and type 1 diabetes (fig 1) were estimated using a Bayesian spatial conditional autoregressive model.[11 12] The allele frequencies were modelled in municipalities, and the incidences on a 10×10 km grid. Model details are given in appendix B.

RESULTS

Exclusion threshold

When differences in disease incidence (resulting from differences in disease gene frequency) are caused by population history factors such as founder effects and drift, the same factors have also caused frequency differences in non-disease causing loci. The efficiency of the SDS strategy to locate interesting genome areas therefore essentially depends on frequency differences in average loci.

Five distributions are shown (fig 3; curves A–E) of simulated haplotype frequency differences between samples of 100 individuals, genotyped at 1 cM intervals, from subpopulations with different histories (see figure legend). From the 90th percentile of the curves, we can estimate that a frequency difference exclusion threshold between 0.12 and 0.25, depending on subpopulation history, should on average reduce the genomic area of interest to approximately 10% of the original.

Fig 3 also shows two distributions (triangles and circles) of allele frequency difference between areas of high and low AMI incidence within Finland based on 30 microsatellites. The median frequencies of the microsatellite alleles depicted in the picture, 0.18 and 0.19, correspond well to the haplotype frequency of 0.20 in the simulations. Although the sample sizes differ slightly more, the overall distributions of real data agree well with the simulations that show low to moderate frequency differences. Thus, the exclusion threshold relevant to AMI is likely a frequency difference of approximately 0.13–0.15.

We also tested several other simulation parameter combinations; table 1 lists some of the resulting exclusion thresholds. In summary, large thresholds resulted from small initial sizes, slow growth rates, large initial haplotype frequencies and small sample sizes. High LD and low genotyping density increased variation between single simulations, but neither affected the average threshold from a set of simulations.

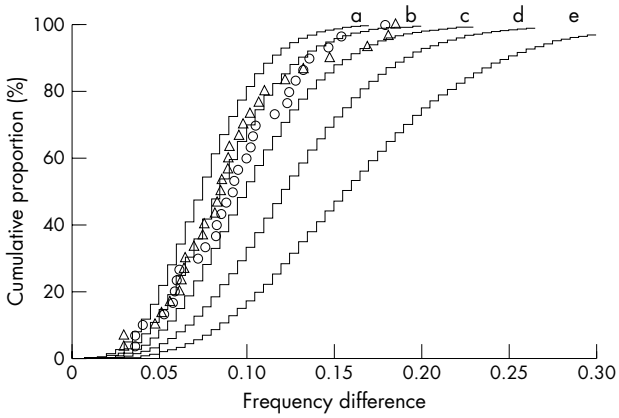

Figure 3 Cumulative distributions of absolute frequency differences from simulations (curves) and from real data (triangles and circles). The curves show maximal haplotype frequency differences of 101 loci at 1 cM intervals in samples of 100 individuals from two subpopulations that have grown into 100 000 individuals in 20 generations from an initial size of 100 (D) or 800 individuals (A), or in 40 generations from 100 E), 400 (C), or 800 individuals (B); initial haplotype frequencies in all simulations are 0.2. The curves are averages of 100 simulations. The circles show maximum allele frequency differences of 30 microsatellite loci between areas of high and low incidence of acute myocardial infarction (AMI): east (111 samples) versus west (46 samples; triangles) or southwest (38 samples; circles) Finland.

Exclusion risk

The exclusion of genome areas is based on sample frequency differences in the genotyped markers, not on true differences of the disease gene. Thus sampling errors or low LD, for instance, can create a risk of excluding the gene.

Table 1 lists, from various simulation settings, the minimal frequency difference that a gene should show between the whole subpopulations in order to have at most a 20% risk of being excluded along with 90% of the genome. For the simulations closest resembling AMI in fig 3, this minimal difference is approximately 0.2 or lower.

When based on 100 simulations, the minimal differences in table 1 are only approximate, owing to the large variation between simulations. General trends are nevertheless visible; large minimal differences result from large exclusion thresholds, low genotyping density, and low LD. Similar or slightly elevated minimal differences were also reachable at a detection power of 0.95 with up to fivefold increases in sample size (data not shown).

Incidence differences

In SDS, only genetic variants with sufficiently large frequency differences can reliably avoid exclusion. How can we recognise diseases caused by such variants—that is, the optimal target diseases for the strategy?

According to fig 2(B) and the analysis in Appendix A, incidence difference Δi is the product of disease risk difference Δp (penetrance minus phenocopy rate) and carrier frequency difference Δc (related to gene frequency difference Δq). In genotyping, we try to detect Δq (or to be precise, its reflection in a nearby locus from the genotyped sample) and therefore the strategy will work best for diseases with large Δi; with small Δi, Δp becomes very limited, if Δq is to remain detectably large. In more detail, the size of Δq that various combinations of Δi and Δp will produce is shown (fig 4) (note that the case shown is the most stringent one where the variant is completely absent from one of the subpopulations; with larger frequencies, a given combination of Δp and Δq will produce a smaller Δi, and thus yield a wider range of potential target diseases.)

What would these restrictions mean in practice, for instance in the case of AMI in Finland? If the genetic lifetime incidence difference is ca. 0.04 (see appendix A),

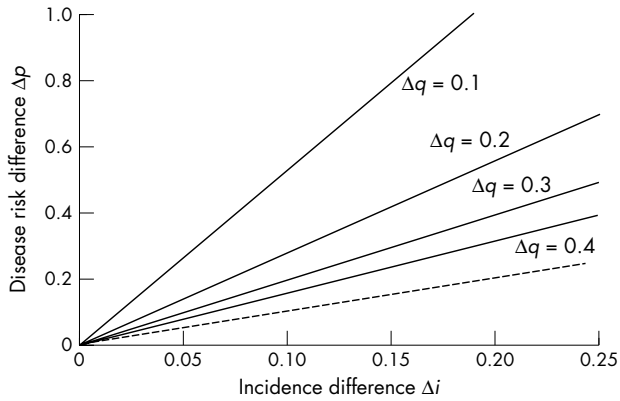

Figure 4 The relation of disease risk differences Δp (penetrance minus phenocopy rate) and incidence differences Δi and disease allele frequency differences Δq between two subpopulations (see fig 2B). Depicted is the most stringent case where the disease allele is absent from one subpopulation. The dashed line indicates the minimal possible disease risk difference $\Delta p = \Delta i$. Diseases with combinations of Δi and Δp that produce a Δq exceeding the corresponding critical gene frequency difference (GFD) listed in table 1 are potential targets for the SDS approach.

Table 1 Haplotype frequency difference thresholds that yield a 90% genome exclusion, and corresponding sizes of gene frequency difference (GFD) that will produce a gene detection power of 0.80 (that is, an exclusion risk of 20%) in various population history simulations with five haplotypes.

Initial size	No. of generations	D'	No. of simulations	Sample size	HFD threshold	GFD
100	20	0.5	100	100	0.19	0.27
100	20	0.7	100	100	0.19	0.23
100	40	0.5	100	100	0.25	0.33
100	40	0.7	100	100	0.25	0.29
400	40	0.5	1000	100	0.16	0.21
400	40	0.7	1000	100	0.16	0.19
800	20	0.5	100	200	0.10	0.13
800	20	0.7	100	200	0.09	0.12
800	40	0.5	100	200	0.12	0.15
800	40	0.7	100	200	0.12	0.14

D', the degree of average linkage disequilibrium at 1 cM distances in the initial population; HFD, haplotype frequency difference; GFD, gene frequency difference.

gene frequency differences of 0.2 or more (as in the section on exclusion risk) would require disease risk differences of 0.12 or less (fig 4). For example, if the phenocopy rate was approximately 0.05, the method would thus work for penetrances of approximately 0.17 or lower.

In the analysis above, the disease related allele was assumed to be dominant and predisposing, but the approach will work identically in finding alleles with a protective effect, and at least equally well for recessive alleles with carrier frequencies <0.5.

DISCUSSION

We have proposed and tested here a strategy for exclusion mapping of complex disease susceptibility loci. We found that the SDS strategy works best for diseases with large incidence difference and with disease alleles that have a small effect on disease risk. The latter might first seem counterintuitive, especially as such alleles tend to be the most difficult for other gene localisation strategies; at a closer look, however, it is obvious that when a given incidence difference results from a low effect allele, the allele frequency difference will be large and thus easier to detect (fig 2B). In more detail, our example of AMI suggested that a 90% exclusion of genome area with a 20% risk of missing the true locus could be accomplished by genotyping loci at 1 cM intervals from 100–200 individuals from a high incidence and 100–200 individuals from a low incidence subpopulation.

The proposed SDS strategy clearly has both limitations and advantages. The main advantage is that subpopulation sampling is easier (and thus cheaper) than the sampling of cases and controls, where the quest for a genetically homogeneous sample makes careful phenotype ascertainment crucial, and is often complicated by low penetrance and high phenocopy rate.

Another advantage of sampling subpopulation individuals regardless of their phenotype is that the results of a one time sampling and genotyping will be valid for any phenotype that has sufficiently large incidence difference between the sampled subpopulations due to a sufficiently low effect variant. In settings that compare cases and controls, the sampling is inevitably phenotype specific.

A third advantage of the phenotype insensitive sampling is that the approach only uses information on allele frequencies in the subpopulations, not on the distribution of those alleles into individuals. Consequently, the correlation of allele frequency differences between the gene of interest and its nearby markers, and thus the genotyping density needed to detect the gene frequency difference, depends on the LD that prevailed when the frequency differences formed through drift. As most drift took place in a small population (fig 2A) that had probably been of relatively constant size, this crucial LD was probably stronger than that observable in the expanded present day populations, for example, in Finland.[13–16] An immediate drawback from this advantage is that knowledge of present day LD patterns, however detailed,[17] will be of limited use in the choosing of loci to be genotyped, and the effectiveness of the SDS approach ultimately needs to be tested in practice.

An obvious limitation of the strategy is that its use is confined to diseases with sufficiently large incidence differences. On the other hand, in Finland several diseases show incidence differences that are of general interest. The strategy is also applicable to diseases where the incidence difference is only partly genetic, as long as the power calculations are based on an estimate of the genetic part.

Another limitation is the array of suitable target populations. It appears (table 1) that the strategy will work best when the differences between subpopulations are moderate; large differences radically limit possible disease parameter values, whereas small differences can be reliably detected only with large samples and dense genotyping.

The performance of the method could be further improved by comparing several high and low incidence subpopulations when such exist. Provided that the subpopulations have separate histories of genetic drift but that their disease causing variant is the same, further investigations could be limited to the fraction of genome where the frequency difference pattern matches the incidence pattern of all the subpopulations.

When the disease of interest has known genetic risk factors, their correlation to the observed incidence differences is naturally worth investigating before proceeding to a whole genome scan. For example, the variation in HLA antigen DR3 frequencies in Finland;[18] (fig 5) could perhaps partly explain the incidence differences in type 1 diabetes (fig 1B); in contrast, the frequency variation of apolipoprotein E allele e4 does not explain the incidence variation observed in coronary heart disease in Finland.[19]

At first glance, our approach might resemble that of admixture mapping (Patterson *et al*[20], and references therein). The differences, however, are also obvious. Admixture mapping focuses on populations that have a long history of separation before a recent admixture event, and finds among loci that are known to differ radically between the ancestral populations those loci that show disproportionate ancestry in the diseased individuals of the admixed population. SDS, in turn, uses subpopulations with long common history and detects the few differences between them. Both approaches use LD to reduce the amount of genotyping needed; in admixture mapping, this LD results from the recent admixture, in SDS from older events of population history (as discussed above; in this sense SDS could be seen as an extension of the ''drift mapping'' approach,[13] and could in fact overcome some of its difficulties of sampling a small population). Both approaches are limited to a subset of diseases in populations with suitable history. Of course, they might be complementary when applied to different populations.

As research tools such as high resolution tagging marker sets and microarrays to assess variation at a multitude of markers in parallel become available, the proposed subpopulation difference scanning strategy could prove a useful addition to the repertoire of complex disease gene localisation strategies. It also shows that population stratification, a

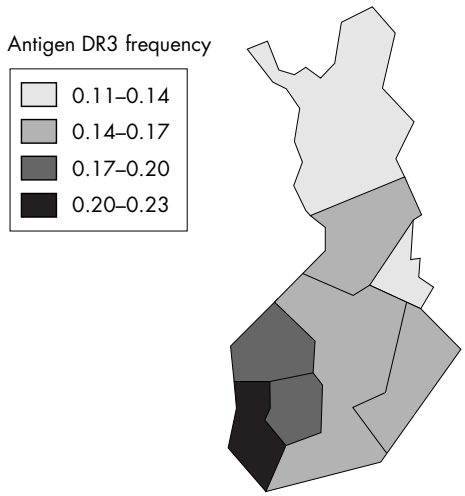

Figure 5 Variation of HLA antigen DR3 frequencies in Finland (data from Sirén *et al*[18]).

ELECTRONIC DATABASE INFORMATION

The URL for data presented herein is as follows:
StatFin -Online service, Statistics Finland, http://statfin.stat.fi

nuisance in case-control analyses, could actually be turned into an advantage.

ACKNOWLEDGEMENTS

We thank M Karvonen and E Moltchanova for discussions on the subject, M Karvonen for providing the AMI and diabetes incidence maps, and T Miettinen for technical help with MatLab and illustrations. This work was supported by Sigrid Jusélius Foundation, Emil Aaltonen Foundation, Graduate School in Computational Biology, Bioinformatics, and Biometry (ComBi) and Academy of Finland. J Kere is a member of Biocentrum Helsinki and Center of Excellence for Disease Genetics at University of Helsinki.

Authors' affiliations

E Salmela, P Lahermo, Finnish Genome Center, University of Helsinki, Finland
O Taskinen, Department of Epidemiology and Health Promotion, Diabetes and Genetic Epidemiology Unit, National Public Health Institute, Helsinki, Finland
J K Seppänen, HIIT Basic Research Unit and Laboratory of Computer and Information Science, Helsinki University of Technology, Helsinki, Finland
P Sistonen, Red Cross Finland Blood Service, Helsinki, Finland
M J Daly, Broad Institute, Cambridge, MA
M-L Savontaus, Department of Genetics, University of Turku, Turku, Finland
M-L Savontaus, Department of Medical Genetics, University of Turku, Turku, Finland
J Kere, Department of Biosciences at Novum and Clinical Research Centre, Karolinska Institutet, Stockholm, Sweden
J Kere, Department of Medical Genetics, University of Helsinki, Helsinki, Finland

Competing interests: there are no competing interests

REFERENCES

1 **Risch NJ**. Searching for genetic determinants in the new millennium. *Nature* 2000;**405**:847–56.
2 **Botstein D**, Risch N. Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. *Nat Genet* 2003;**33**(suppl):228–37.
3 **Nevanlinna HR**. The Finnish population structure. A genetic and genealogical study. *Hereditas* 1972;**71**:195–236.
4 **de la Chapelle A**, Wright FA. Linkage disequilibrium mapping in isolated populations: The example of Finland revisited. *Proc Natl Acad Sci USA* 1998;**95**:12416–23.
5 **Kere J**. Human population genetics: lessons from Finland. *Annu Rev Genomics Hum Genet* 2001;**2**:103–28.
6 **Norio R**. Finnish Disease Heritage I: characteristics, causes, background. *Hum Genet* 2003;**112**:441–56.
7 **Norio R**. Finnish Disease Heritage II: population prehistory and genetic roots of Finns. *Hum Genet* 2003;**112**:457–69.
8 **Norio R**. The Finnish Disease Heritage III: the individual diseases. *Hum Genet* 2003;**112**:470–526.
9 **Pyörälä K**, Salonen JT, Valkonen T. Trends in coronary heart disease mortality and morbidity and related factors in Finland. *Cardiology* 1985;**72**:35–51.
10 **Jousilahti P**, Vartiainen E, Tuomilehto J, Pekkanen J, Puska P. Role of known risk factors in explaining the difference in the risk of coronary heart disease between eastern and southwestern Finland. *Ann Med* 1998;**30**:481–7.
11 **Besag J**, York J, Mollié A. Bayesian image restoration, with two applications in spatial statistics. *Ann Inst Stat Math* 1991;**43**:1–59.
12 **Ranta J**, Penttinen A. Probabilistic small area risk assessment using GIS-based data: a case study on Finnish childhood diabetes. *Stat Med* 2000;**19**:2345–59.
13 **Terwilliger JD**, Zöllner S, Laan M, Pääbo S. Mapping genes through the use of linkage disequilibrium generated by genetic drift: 'drift mapping' in small populations with no demographic expansion. *Hum Hered* 1998;**48**:138–54.
14 **Varilo T**, Laan M, Hovatta I, Wiebe V, Terwilliger JD, Peltonen L. Linkage disequilibrium in isolated populations: Finland and a young sub-population of Kuusamo. *Eur J Hum Genet* 2000;**8**:604–12.
15 **Varilo T**, Paunio T, Parker A, Perola M, Meyer J, Terwilliger JD, Peltonen L. The interval of linkage disequilibrium (LD) detected with microsatellite and SNP markers in chromosomes of Finnish populations with different histories. *Hum Mol Genet* 2003;**12**:51–9.
16 **Kaessmann H**, Zöllner S, Gustafsson AC, Wiebe V, Laan M, Lundeberg J, Uhlén M, Pääbo S. Extensive linkage disequilibrium in small human populations in Eurasia. *Am J Hum Genet* 2002;**70**:673–85.
17 **The International HapMap Consortium**. The International HapMap Project. *Nature* 2003;**426**:789–96.
18 **Sirén MK**, Sareneva H, Lokki ML, Koskimies S. Unique HLA antigen frequencies in the Finnish population. *Tissue Antigens* 1996;**48**:703–7.
19 **Lehtimäki T**, Moilanen T, Viikari J, Åkerblom HK, Ehnholm C, Rönnemaa T, Marniemi J, Dahlen G, Nikkari T. Apolipoprotein E phenotypes in Finnish youths: a cross-sectional and 6-year follow-up study. *J Lipid Res* 1990;**31**:487–95.
20 **Patterson N**, Hattangadi N, Lane B, Lohmueller KE, Hafler DA, Oksenberg JR, Hauser SL, Smith MW, O'Brien SJ, Altshuler D, Daly MJ, Reich D. Methods for high-density admixture mapping of disease genes. *Am J Hum Genet* 2004;**74**:979–1000.
21 **Pitkänen K**. Suomen väestön historialliset kehityslinjat. In: Koskinen S, Martelin T, Notkola IL, Notkola V, Pitkänen K, eds. *Suomen väestö*. Hämeenlinna, Finland: Gaudeamus, 1994:19–63.
22 **Hedrick PW**. Gametic disequilibrium measures: proceed with caution. *Genetics* 1987;**117**:331–41.
23 **Galassi M**, Davies J, Theiler J, Gough B, Jungman G, Booth M, Rossi F. *GNU scientific library reference manual*, 2nd ed. Bristol: Network Theory Ltd, 2003.
24 **Tuomilehto J**, Arstila M, Kaarsalo E, Kankaanpää J, Ketonen M, Kuulasmaa K, Lehto S, Miettinen H, Mustaniemi H, Palomäki P, Puska P, Pyörälä K, Salomaa V, Torppa J, Vuorenmaa T. Acute myocardial infarction (AMI) in Finland – baseline data from the FINMONICA AMI register in 1983–1985. *Eur Heart J* 1992;**13**:577–87.
25 **Rytkönen M**, Ranta J, Tuomilehto J, Karvonen M. Bayesian analysis of geographical variation in the incidence of Type I diabetes in Finland. *Diabetologia* 2001;**44**(suppl):B37–44.
26 **Karvonen M**, Moltchanova E, Viik-Kajander M, Moltchanov V, Rytkönen M, Kousa A, Tuomilehto J. Regional inequality in the risk of acute myocardial infarction in Finland: a case study of 35- to 74-year-old men. *Heart Drug* 2002;**2**:51–60.
27 **Spiegelhalter DJ**, Thomas A, Best NG, Lunn D. *WinBUGS version 1.4 user manual*. Cambridge: MRC Biostatistics Unit, 2003.

APPENDIX A

PARAMETER VALUES USED IN POPULATION SIMULATIONS

The simulations had initial daughter population sizes of 100, 400, or 800 individuals, encompassing the estimates of breeding unit size in a Finnish rural population in the 19th century.[3] Growth times were either 20 or 40 generations, corresponding to the time scale relevant to Finnish population history, with southwestern coastal areas inhabited more than 1000 and eastern inland areas less than 500 years ago.[21] The final size of the subpopulations was 100 000 and chosen mostly for technical convenience; obviously, the initial sizes are more crucial because in large populations the effects of drift will be negligible. Linkage disequilibrium (LD) was measured in terms of the multiallelic extension of Lewontin's normalized LD measure D',[22] and $D' \geqslant 0.5$ at 1 cM distances was mostly used.

MATING AND REPRODUCTION

All initial individuals were assumed non-siblings. In each generation, individuals formed random non-sibling pairs from which the individuals of the next generation randomly ''chose'' their parent pair. Thus, virtually the whole population was reproducing, apart from individuals who did not find a non-sibling pair, or pairs who did not happen to produce children.

RECOMBINATION

Each individual had one pair of 100 cM long chromosomes. In the meioses that produced the child chromosomes from the parent chromosomes, chiasma numbers were Poisson distributed, and the distribution of chiasma locations on the chromosomes was uniform.

SAMPLING AND GENOTYPING

The descent of the initial chromosomes through the population was monitored for 1001 loci located at 0.1 cM intervals. In the final generation, the number of copies of each initial chromosome in each locus was recorded from two of the total daughter populations as well as from independent random samples of 50, 100, 200, 500, and 1000 non-sibling individuals that were drawn without replacement from the daughter populations. Evenly spaced subsets of the 1001 loci were used to mimic the results of genotyping, which were assumed to contain no errors or missing data.

HAPLOTYPES AND INITIAL LD

The initial chromosomes were assigned to carry at each simulated locus one of n equifrequent haplotypes of nonrecombining SNPs. Each of these haplotypes was assumed to be uniquely recognisable from one tagging SNP. Linkage disequilibrium (LD) between haplotypes of adjacent loci in the initial population was created by letting the haplotypes on a given chromosome depend on each other: In the first locus, the haplotypes were assigned to the chromosomes randomly. In each following locus, a given number of randomly chosen chromosomes received haplotypes identical to the previous locus; in the rest of the chromosomes, the haplotypes of the previous locus were randomly shuffled into the new one. Thus, in the two extreme cases, either all loci along a chromosome have the same haplotype ($D' = 1$) or all haplotypes are assigned anew in all loci ($D' = 0$). For intermediate LD values, the appropriate number of chromosomes to be shuffled was determined based on average LD produced in 1000–5000 pairs of loci located 1 cM (that is, 10 reshufflings) apart; for example $D' = 0.5$ is produced by shuffling ca. 6.7% (depending somewhat on haplotype number n) of the chromosomes at each 0.1 cM interval.

IMPLEMENTATION

The simulations of chromosome descent were implemented in C using the GNU Scientific Library (version 1.3)[23] for random number generation, and the haplotype assignment and further calculations were performed in Matlab (release 12.1; MathWorks Inc., Natick, MA, USA). The initial chromosomes were labelled uniquely in the descent simulations, and simulation effort could therefore be minimised by using the same set of descent simulations in several combinations of n and D'; this introduced no pseudoreplication, however, because each simulation was used only once per combination.

EXCLUSION RISK

In each simulation setting—that is, with a given population history, haplotype number, initial LD, sample size, and genotyping density, the exclusion threshold was calculated based on the distribution of allele frequency differences in genotyped loci between subpopulation samples. Intuitively, if a locus has a large frequency difference between the total subpopulations, it is likely to exhibit a large difference in the genotyping of the subpopulation samples, and thus has a lower risk of being excluded than a locus whose frequency difference is small. To allow comparison between simulation settings, we calculated for each setting the true frequency difference that a locus should have between the total subpopulations in order to have a given risk of exclusion.

Firstly, we recorded for each simulated locus the frequency difference that would be observed when genotyping every mth locus from subpopulation samples. For a non-genotyped locus, this observed difference was taken to be the larger of the absolute sample frequency differences in its two closest genotyped loci (thus all loci between two genotyped ones have the same observed difference). For a genotyped locus, the observed difference was taken to be the largest of the absolute sample frequency differences in the locus itself and the two closest genotyped loci. In this manner, an observed difference was calculated for all possible placements of genotyped loci at m locus intervals (that is, for locus subsets beginning from simulated locus 1, 2, ..., m); thus, all simulated loci had altogether m observed differences. The loci were then divided according to their true absolute frequency difference into overlapping categories of width 0.05 at 0.01 intervals (that is, 0.00…0.05, 0.01…0.06, and so forth). In each category, the proportion of observed frequency differences that remain below the exclusion threshold of the simulation setting determines the exclusion risk. To inspect the true frequency difference corresponding to an exclusion risk of 20% (that is, gene detection power of 0.80), the 20th percentile of the observed frequency differences was calculated in each category. The highest category where the 20th percentile did not exceed the exclusion threshold was found, and from the category immediately above that, the median of the true frequency differences was recorded as the frequency difference with a 20% exclusion risk for the simulation setting in question.

PENETRANCE MODEL

Denote by q_x the frequency of a disease predisposing, dominant allele in subpopulation x (where x is h or l for high or low incidence, respectively), and by c_x the carrier frequency in subpopulation x, that is, the proportion of individuals in subpopulation x who have at least one copy of the predisposing allele. In addition, denote by p_a and p_n the risk of a carrier and a non-carrier for developing the disease (that is, penetrance and phenocopy rate), respectively. Now (see fig 2B) the disease incidence in subpopulation x is $i_x = c_x \times p_a + (1-c_x) \times p_n$, and, in Hardy-Weinberg equilibrium, $c_x = 1-(1-q_x)^2$. As frequencies and probabilities, p_a, p_n, i_x, c_x and q_x are all between 0 and 1. Because the allele is assumed to be predisposing, $p_n < p_a$; furthermore, i_x is between p_n and p_a.

From above, the carrier frequency difference between the two subpopulations is $\Delta c = c_h - c_l = (i_h - i_l)/(p_a - p_n) = \Delta i/\Delta p$. Thus, a large Δc can result from a large incidence difference Δi or a small disease risk difference Δp. For given Δp and disease allele frequency difference $\Delta q = q_h - q_l$, the corresponding Δi will in turn be largest when $q_l = 0$.

ESTIMATING AMI INCIDENCES

As the penetrance model above has no time component, the closest real life counterpart of its incidences would be lifetime incidences. In the case of AMI, these could be estimated by transforming the age standardised yearly incidences i_{x1} of literature to n year incidences $i_{xn} = 1-(1-i_{x1})^n$. The incidences from fig 1(A) thus become $i_{l20} = 0.0844$ and $i_{h20} = 0.1346$ and those from Tuomilehto *et al*[24] $i_{l30} = 0.1419$ and $i_{h30} = 0.2556$; these can be considered minimum estimates of regional lifetime incidences. A part of their differences (0.0502...0.1136) is, however, undoubtedly due to differences in nongenetic risk factor levels between east and west, and we therefore use 0.04 (from $i_l = 0.06$ and $i_h = 0.10$) as an estimate of the genetic lifetime incidence difference of AMI.

APPENDIX B

INCIDENCE AND ALLELE FREQUENCY MAPS

The maps in fig 1 were drawn using Gibbs sampling on Bayesian regression models that describe the geographical variation in incidence and in allele frequency. Autoregressive regression models allow separating the random variation in the data from the spatial trends without need to specify a parametric model for the spatial trend. Gibbs sampling,

which is a Markov chain Monte Carlo (MCMC) method, yields numerical samples from the posterior distribution of the model.

In the incidence model, the cases are placed on a 10×10 km grid. In each grid cell, the number of cases depends on several factors: the baseline risk, the size and age structure of the population in the cell, the geographical effect that we want to extract, and random variation. The geographical effect, denoted by λ_i where i is the number of the grid cell, was assigned an autoregressive model. Denoting by λ_{-i} the mean value of λ in the neighbouring cells of grid cell i and by m_i the number of neighbouring cells, the model has the form

$$\lambda_i \sim N\left(\bar{\lambda}_{-i}, \frac{1}{\tau m_i}\right),$$

where τ is a precision (inverse variance) parameter.

The number of cases in a grid cell is assumed to have a Poisson distribution whose mean depends on the known variables and the geographical effect. Denoting the mean of grid cell i by μ_i, the model has the form

$\log(\mu_i) = \alpha + \beta_k + \lambda_i + \log(N_i)$

where α is the baseline, β_k is the age group effect, λ_i is the local deviation of the risk from the baseline, and N_i is the population at risk in the location i.

The incidence data was collected as described in Rytkönen *et al*[25] and Karvonen *et al*.[26] Pooled data from years 1983, 1988, and 1993 were used for AMI and from 1987 to 1996 for diabetes. Age standardised incidences were calculated using the mean population structure of Finland from respective years,

$$\mathrm{SR}_i = e^{\alpha + \sum w_k \beta_k + \lambda_i} 100000.$$

The allele frequency maps were drawn using a similar model, with the genotype data described in Material and Methods. The 10 modelled alleles were chosen by visual inspection of province-level frequency differences in χ^2 significant alleles; four alleles were selected for presentation based on the pattern and reliability of the modelled frequencies. In this model, the allele frequency p_i of municipality i, has a logistic regression structure of the form

$$\log\left(\frac{p_i}{1-p_i}\right) = \alpha + \lambda_i,$$

where α is again the baseline and λ_i is the local deviation from the baseline. The local deviations were again assumed to have a geographical dependency of a similar form as in the incidence model. There are two differences: firstly, instead of a grid structure, neighbours were defined as municipalities having a common border, or in isolated cases, as the closest municipality; secondly, the model needs another random effect because of uncertainty about the allele origins. The reason for the uncertainty is that for each subject, only the grandparents' birthplaces are known, not which of the grandparents carried the subject's alleles. For the two alleles in a genotype, there are eight possible combinations of municipalities where one allele originates from the birthplace of a maternal grandparent and the other allele from the birthplace of a paternal grandparent. The probabilities of these combinations were considered equal and fixed, and the stochastic origin of the alleles was modelled by sampling a random combination of origins of individuals during each iteration.

In both models, the parameters α and τ were given uninformative priors, $p(\alpha)1$ and $\tau \sim \gamma(0.05,0.0005)$. Posterior means of incidence and allele frequency were based on 10000 iterations after an initial burn in of 5000 iterations. No parallel chains were run as hierarchical models with a large number of random effects converge fast. We used WinBUGS 1.4[27] for the MCMC estimation and ArcView (version 8.3 Esri, Redlands, CA, USA) for the mapping.