European Journal of Human Genetics (2010) 18, 471–478

www.nature.com/ejhg

ARTICLE

Genomic landscape of positive natural selection in Northern European populations

Tuuli Lappalainen[1,2], Elina Salmela[1,2], Peter M Andersen[3], Karin Dahlman-Wright[4], Pertti Sistonen[5], Marja-Liisa Savontaus[6], Stefan Schreiber[7], Päivi Lahermo*[,1] and Juha Kere[2,4,8]

Analyzing genetic variation of human populations for detecting loci that have been affected by positive natural selection is important for understanding adaptive history and phenotypic variation in humans. In this study, we analyzed recent positive selection in Northern Europe from genome-wide data sets of 250 000 and 500 000 single-nucleotide polymorphisms (SNPs) in a total of 999 individuals from Great Britain, Northern Germany, Eastern and Western Finland, and Sweden. Coalescent simulations were used for demonstrating that the integrated haplotype score (iHS) and long-range haplotype (LRH) statistics have sufficient power in genome-wide data sets of different sample sizes and SNP densities. Furthermore, the behavior of the F_{ST} statistic in closely related populations was characterized by allele frequency simulations. In the analysis of the North European data set, 60 regions in the genome showed strong signs of recent positive selection. Out of these, 21 regions have not been discovered in previous scans, and many contain genes with interesting functions (eg, *RAB38*, *INFG*, *NOS1AP*, and *APOE*). In the putatively selected regions, we observed a statistically significant overrepresentation of genetic association with complex disease, which emphasizes the importance of the analysis of positive selection in understanding the evolution of human disease. Altogether, this study demonstrates the potential of genome-wide data sets to discover loci that lie behind evolutionary adaptation in different human populations.

European Journal of Human Genetics (2010) **18,** 471–478; doi:10.1038/ejhg.2009.184; published online 21 October 2009

Keywords: natural selection; genetic variation; population; Europe

INTRODUCTION

A characterization of the adaptive history of human populations requires knowledge of the genes that have been affected by positive natural selection, which is also important for an analysis of the genetic causes behind human disease.[1,2] During the past decade, various statistical approaches have been developed and used for scanning the entire genome for traces of selective sweeps, and hundreds of loci with strong evidence of selection have been discovered with these methods.[3–11]

Most of these studies have used the HapMap[12] or Perlegen[13] data sets consisting of millions of single nucleotide polymorphisms (SNPs). Their good coverage of SNPs allows a relatively precise identification of genes and even the actual variants under selection,[7] but their weakness is the limited sample set that currently allows the analysis of only a few populations from each continent. Additional large data sets have recently become available through genome-wide SNP scans, providing data of a large numbers of individuals from a variety of populations. These data have also provided a powerful tool for analyzing population differentiation and structure,[14–18] but to date, only a few studies have used similar data sets for analyzing traces of positive natural selection.[5,10,19,20]

In this study, we used a genome-wide data set from Eastern and Western Finland, Sweden, Northern Germany, and Great Britain, and a combination of three statistical methods to search for loci that have been affected by recent natural selection.

MATERIALS AND METHODS

Data sets

We analyzed Affymetrix (Santa Clara, CA, USA) 250K Sty array data from Eastern and Western Finland, Sweden,[18] Northern Germany,[21] and Great Britain[22] (hereafter, 250K data), combined with 250K Nsp array data from the Germans and the British (hereafter, 500K data). We also used HapMap[12,23] 250K and 500K data provided by Affymetrix for estimating genetic differentiation between continents – selection scans for the HapMap populations have been performed previously. The quality control of the data followed common standards and is briefly described in Supplementary Methods. The data sets and samples used in this study are outlined in Table 1.

Analysis of natural selection

We employed two statistics for scanning the genome-wide data for signs of positive natural selection: integrated haplotype score test (iHS),[9] and single-SNP long-range haplotype test (LRH),[7] both based on comparing the extended haplotype homozygosity (EHH) score[24] of the ancestral and derived allele of

[1]Institute for Molecular Medicine Finland, University of Helsinki, Helsinki, Finland; [2]Department of Medical Genetics, and Folkhälsan Institute of Genetics, Biomedicum Helsinki, University of Helsinki, Helsinki, Finland; [3]Department of Neurology, Umeå University Hospital, University of Umeå, Umeå, Sweden; [4]Department of Biosciences and Nutrition, Karolinska Institutet, Huddinge, Sweden; [5]Finnish Red Cross Blood Transfusion Center, Helsinki, Finland; [6]Department of Medical Genetics, University of Turku, Turku, Finland; [7]Department of General Internal Medicine, Institute for Clinical Molecular Biology, Christian-Albrechts-University, Kiel, Germany; [8]Clinical Research Centre, Karolinska University Hospital, Huddinge, Sweden
*Correspondence: Dr P Lahermo, Institute for Molecular Medicine Finland, University of Helsinki, P.O. Box 20, Tukholmankatu 8, Helsinki 00014, Finland.
Tel: +358 9 1912 5476; Fax: +358 9 1912 5478; E-mail: paivi.lahermo@helsinki.fi
Received 15 April 2009; revised 13 August 2009; accepted 9 September 2009; published online 21 October 2009

Table 1 The data sets used in different analyses of the study, and the numbers of samples and markers in the analyses after quality control and filtering

		iHS and LRH		*F_{ST} and networks*[a]	
	Abbreviation	*250K*	*500K*	*250K*	*500K*
Eastern Finland	*FIE*				
	Samples	139	—	139	—
	SNPs	112 646	—	158 064	—
Western Finland	*FIW*				
	Samples	141	—	141	—
	SNPs	129 928	—	158 064	—
Sweden	*SWE*				
	Samples	113	—	113	—
	SNPs	130 699	—	158 064	—
Northern Germany	*GER*				
	Samples	256	252	256	252
	SNPs	135 320	302 983	158 064	350 070
Great Britain	*BRI*				
	Samples	350[b]	350[c]	350	350
	SNPs	134 846	302 091	158 064	350 070
Utah residents with ancestry from northern and central Europe	*CEU*				
	Samples	—	—	58	58
	SNPs	—	—	162 805	334 989
Han Chinese from Beijing, China, and Japanese from Tokyo	*CHB+JPT*				
	Samples	—	—	87	87
	SNPs	—	—	162 805	334 989
Yoruba from Ibadan, Nigeria	*YRI*				
	Samples	—	—	56	56
	SNPs	—	—	162 805	334 989

[a]Networks only for 250K data.
[b]For sample size testing: 700, 500, 350, 256, 140, 113, and 60 samples.
[c]For sample size testing: 700, 500, 350, 252, 140, 113, and 60 samples.

each marker. The tests are designed to detect sites in which one allele is surrounded by a much longer haplotype than expected for alleles of corresponding frequency evolving neutrally. Such a situation may arise when natural selection is driving one haplotype to high frequency, leaving recombination little time to break the haplotype. The test statistics were calculated for each marker in Sweep 1.1,[7] separately in each population.

Population differentiation across the genome was estimated by calculating the F_{ST} statistic[4,25] for each marker in several population combinations: between each European population and the other Europeans pooled together, and between different continents using the HapMap YRI, CHB+JPT, and all the European samples.

To find the genomic regions with multiple SNPs with high iHS, LRH, or F_{ST} scores, the single-SNP absolute values from each population were analyzed in 200-kb windows with a 100-kb overlap. Each window was classified as either an extreme or a suggestive outlier based on each of the three statistics: the most extreme outliers included the windows with |iHS| or |LRH| $\geqslant$3.2 for at least two or four SNPs in the 250K and 500K analysis, respectively, in any of the populations. The extreme outlier windows for F_{ST} included those with at least two or four SNPs among the highest 1/2000 of each population comparison. Similarly, a suggestive category contained windows with at least two or four SNPs with |iHS| or |LRH| $\geqslant$ 2.6, or F_{ST} among the 1/500 highest values.

To extract the windows most likely to be affected by natural selection, an overlap of at least two statistics was required, because it has been observed that overlapping false positives in the iHS and LRH tests are rare.[7] Thus, a window had to fall in the category of extreme outliers based on iHS or LRH in at least one population, and have at least a suggestive signal in any of the populations in the other test or F_{ST}. The overlapping or adjacent windows fulfilling these criteria were combined to form regions. Regions with low SNP densities or known inversions were excluded (see Supplementary Methods).

To visualize haplotype variation in the selected loci, median-joining networks were constructed with Network 4.5.0.2 (fluxus-engineering.com),[26,27] using European as well as HapMap 250K data. To analyze the extent of overlap between selection signals and association with disease, we counted the bins with and without signs of selection that contained at least one positively associating gene, as listed in the NHGRI catalog of genome-wide association studies (http://www.genome.gov/gwastudies) and Genetic Association Database (http://geneticassociationdb.nih.gov/).

Empirical analysis of different sample sizes

We estimated the effect of sample size empirically by calculating iHS and LRH for each marker in chromosomes 1–3 from the British data sampled to seven

different sizes (Table 1). We calculated the correlations of the standardized iHS and LRH values between the largest sample of 700 individuals and the smaller samples. As the analysis showed that sample size affected the reliability of the statistics (Figure 3), we wanted to assign more weight to populations with larger sample sizes. This was obtained by multiplying the iHS and LRH values of each population with the correlation coefficient of the British test sample of corresponding size before extracting outlying genomic regions described above.

Coalescent simulations

Previous studies have analyzed the performance of iHS and LRH tests for different allele frequencies, scenarios of natural selection, demographies,[7] and sample size.[20] We sought to characterize the applicability of the tests for genome-wide data sets by analyzing the joint effects of SNP density and sample size. We performed coalescent simulations using the SelSim software,[28] simulating genomic segments with a neutral model and with natural selection (see Supplementary Methods for details of the simulations and subsequent analyses). We constructed data sets of four different SNP densities corresponding to the median densities of HapMap II and Affymetrix arrays 6.0, 500K and 250K, and sample sizes of 50, 100, 150, and 200 individuals. The SNP ascertainment bias was accounted for by matching the SNP frequency spectrum of the simulations to that observed in real data.[9] We compared the power obtained with different SNP densities and sample sizes by adjusting the false discovery rate (FDR) to approximately 1% in each data set.

Allele frequency simulations

As the North European populations analyzed in this study are much more closely related than the populations in most previous scans of natural selection, we wanted to analyze the expected extent of population differences caused by positive selection in Northern Europe. For this purpose, we simulated allele frequencies with demographic models corresponding to two population pairs, one modelling an Asian and a European population, and the other modelling two North European populations. We adjusted the demographic parameters by matching the simulated allele frequency differences with empirical distributions (see Supplementary Methods). Several scenarios of selection were analyzed, including different time spans, allele frequencies at the start of selection, and selection coefficients. Selection was applied to only one of the populations of each pair. The distribution of F_{ST} was calculated from the resulting allele frequencies. The simulations were performed in R (http://www.r-project.org/), and the demographic parameters are shown in Supplementary Table 1.

RESULTS

Signs of selection across the genome

The results of the 250K analysis for all the populations are visualized in Figure 1, showing that the signal of selection is shared among several populations for many but not all the regions. Usually a high F_{ST} does not overlap with high iHS or LRH. A total of 60 regions had strong signs of selection, with at least one clearly outlying iHS or LRH score as well as at least one suggestive signal in another statistic (Table 2). However, the employed SNP density does not result in a full coverage of the euchromatin regions of the genome: in the 250K analysis, 64–74% (median 72%) of the euchromatin was covered by at least five markers per 200-kb window, and in the 500K data it was 84%. The different coverage of the populations results from differing numbers of markers becoming excluded because of minor allele frequency below 5%.

In total, the 60 regions contain 121 genes. The windows with signs of selection showed a statistically significant enrichment of genes associated with disease (χ^2 $P<10^{-4}$) – however, given the relatively large windows (200 kb), the disease-associated genes and variants are not necessarily the targets of selection. Figure 2 shows examples of median-joining networks in genes *RAB38* and *PPP2R2B*, both having a combination of characteristic signs of positive natural selection: population differentiation, enrichment of high-frequency derived alleles demonstrated by the long branches from the ancestral haplotype to the high-frequency clusters, and star-like haplotype patterns with one high-frequency haplotype surrounded by rare haplotypes.

Power and performance of iHS and LRH

The analysis of iHS and LRH values from the British, who were sampled to different sizes, showed that the reliability of both statistics improved markedly as the sample size increased, and especially the LRH statistic appeared to lack robustness for smaller samples (Figure 3). The denser marker set of the 500K data improved the correlation, but not dramatically. The differences in the robustness between the sample sizes was accounted for by scaling down the genomic values of iHS and LRH of the smaller samples, which explains the much stronger signs of selection among the British and German samples than among the Finns or Swedes (Table 2). In the

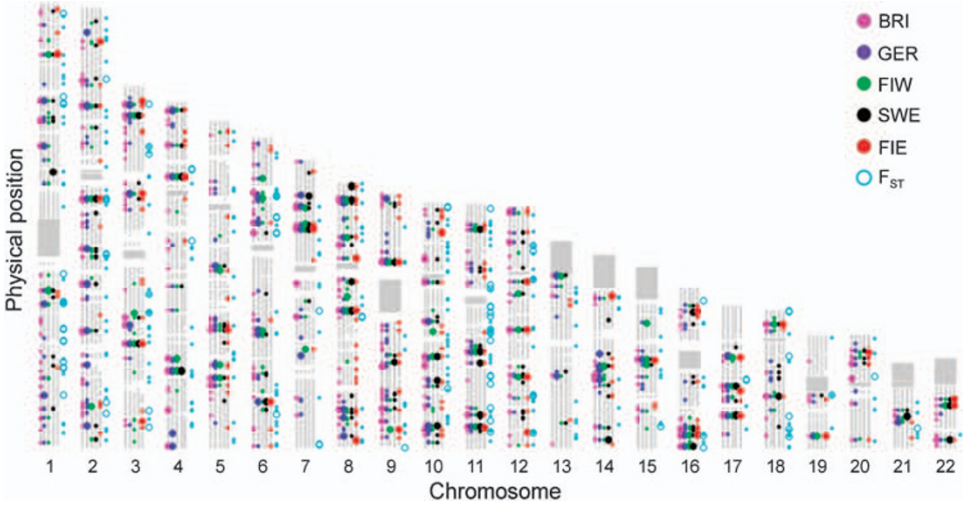

Figure 1 The iHS and LRH signals in the different populations in 200-kb windows across the genome, and F_{ST} signals over all population comparisons. The larger symbol denotes the most extreme outliers, whereas the smaller symbol denotes suggestive signals. The grey boxes and horizontal lines denote heterochromatin, centromere, and telomere regions, and the grey vertical lines correspond to windows with at least five SNPs per 200-kb window.

Table 2 The genomic regions showing the strongest signs of positive natural selection

Region (Mb)	*250K*	*500K*	*New*[a]	*iHS*	*LRH*	*FST*	*BRI*	*GER*	*FIW*	*FIE*	*SWE*	*Genes*
1:55.7–56.0	x	x	x	++	++		++	++	+		+	
1:80.6–80.9		x	x	++	+		++	++	NA	NA	NA	
1:160.1–160.4	x			++	+		+	+	++		+	ATF6,OLFML2B,NOS1AP
1:186.2–186.5		x	x	++	++		++	++	NA	NA	NA	
1:217.8–218.2	x			++	++	++	+	++				SLC30A10
2:40.1–40.4	x			+	++		++	+				SLC8A1
2:105.4–105.8	x	x		++	+	++	++	++	++	++	++	FHL2,NCK2
2:123.1–123.3		x		+	++		++	+	NA	NA	NA	
2:152.2–152.5	x	x		++	++		++	++			+	NEB,ARL5A,CACNB4
2:178.1–178.6	x			++	+	+	++	++	+		+	AGPS,TTC30B,-A,PDE11A
3:10.2–10.5	x		x	+	++	++	++	++	+	+		IRAK2,TATDN2,C3orf42,GHRL,GHRLOS, SEC13,ATP2B2
3:10.6–11.0	x	x		+	++		+	++				LOC285370,SLC6A11
3:59.3–59.5	x		x	++	+		+	++	+	++		
3:131.4–131.8		x		++	++		++	++	NA	NA	NA	COL29A1,COL6A6
3:141.5–141.8	x			++	++		++	++	++	++	++	CLSTN2
3:184.5–184.8	x		x	++	+					++		MCF2L2,KLHL6
4:24.4–24.7	x		x	+	++		++	+	+		+	SOD3,CCDC149,LGI2
4:41.6–42.0	x			++	++		++	++	++	++	++	TMEM33,WDR21B,SLC30A9,BEND4
4:141.5–141.8	x		x	+	++		++	+	++			SCOC,CLGN,ELMOD2,UCP1,TBC1D9
4:148.2–148.6		x		+	++		++	+	NA	NA	NA	
4:170.0–170.3	x			++	+		++					PALLD,CBR4,SH3RF1
4:172.8–173.0		x		++	+		+	++	NA	NA	NA	GALNTL6
5:80.2–80.5	x		x	++	+		+	++	+		+	MSH3,RASGRF2
5:115.6–115.8	x			++	+		++	+	+	++	+	COMMD10
5:134.7–135.0	x			++	+		+	++	++	+		H2AFY,C5orf20,TIFAB,NEUROG1,CXCL14
5:142.0–142.6	x	x		++	++		++	++	++	+	+	FGF1,ARHGAP26
5:145.9–146.2	x	x		++	++		++	+		+		PPP2R2B
6:46.8–47.1	x			++	+		++	++	+	+		PLA2G7,MEP1A,GPR116,-110
6:52.4–52.6	x		x	++	+		++				+	EFHC1,TRAM2
6:100.1–100.4		x	x	+	++		++	+	NA	NA	NA	CCNC,PRDM13
6:105.6–105.8		x		++	++		++	++	NA	NA	NA	LIN28B,BVES,POPDC3
6:145.1–145.4		x	x	++	++		+	++	NA	NA	NA	UTRN
6:159.4–159.7		x	x	++	+		+	++	NA	NA	NA	FNDC1
7:19.3–19.6	x	x		++	+		+	++			+	
7:36.7–37.3	x	x		++	++		++	+	++	++	+	AOAH,ELMO1
8:52.8–53.2		x		++	++		++	++	NA	NA	NA	PXDNL,PCMTD1,ST18
8:139.6–139.9	x	x		++	+		+	++			+	COL22A1
8:142.2–142.4	x		x	++		+				++		DENND3,SLC45A4
9:3.1–3.3	x			++	+		+	++				RFX3
9:16.6–16.8		x	x	++	+		+	++	NA	NA	NA	BNC2
9:118.1–118.4	x		x	++	+		++	+				PAPPA,ASTN2
9:125.2–125.5	x			+	++		++	+				DENND1A
10:4.1–4.4		x		++	+		+	++	NA	NA	NA	
10:43.6–44.0	x	x		++	+		++	+	+	+	+	HNRNPA3P1
10:65.1–65.4	x			++	+		++	++	+		+	
10:84.5–84.8	x		x	++	+		++	+	++	+	++	NRG3
11:87.4–87.6	x			++		+		+		+	++	RAB38
11:116.0–116.3	x			++	+		++	+	+	+	++	BUD13,ZNF259,APOA5,-A4,-C3,-A1,KIAA0999
12:2.8–3.0	x			+	++		++	+	+	+		ITFG2,NRIP2,FOXM1,C12orf32,TULP3,TEAD4
12:66.8–67.1	x		x	++	+	+	++	+	++	++	+	IFNG,IL26,IL22,MDM1
14:60.8–61.3	x	x		++	++		++	++	++	+	+	TMEM30B,PRKCH,HIF1A,SNAPC1
14:68.4–68.6	x				++	+	++	++	+			ACTN1,WDR22
16:77.1–78.0	x	x		++	++		++	+	+	+		WWOX
16:78.3–78.6	x	x		++	++		++	+				
16:81.3–81.6	x		x	++	+	+	++	++	++	++	+	
16:81.7–82.0	x		x	++	+		++	+	+	+		CDH13
17:60.5–61.0	x			++	++		++	++	++	++	++	RGS9,AXIN2
18:7.2–7.8	x	x		++	++		++	+	++	++	+	PTPRM
20:16.1–16.3	x		x	+	++		++	+		+		KIF16B
22:44.9–45.2	x			+	++		++	+	+	+	++	PPARA,C22orf40,PKDREJ,TTC38, CN5H6.4,GTSE1, TRMU,CELSR1

The symbol ++ denotes extreme outliers and the symbol + denotes a suggestive signal in the selection tests across the populations, and the maximum signal of iHS and LRH in each population (see Materials and Methods for definitions). The underlined genes associate with human disease or trait (see text for details). The genomic positions are according to human genome Build 36.
[a]Compared with all the reported loci in Huttley *et al*,[53] Akey *et al*,[4] Carlson *et al*,[3] Nielsen *et al*,[54] Hapmap 1[23] and 2,[12] Voight *et al*,[9] Wang *et al*,[8] Sabeti *et al*,[7] and Oleksyk *et al*[6] as listed by Oleksyk *et al*,[5] and Pickrell *et al*.[20]

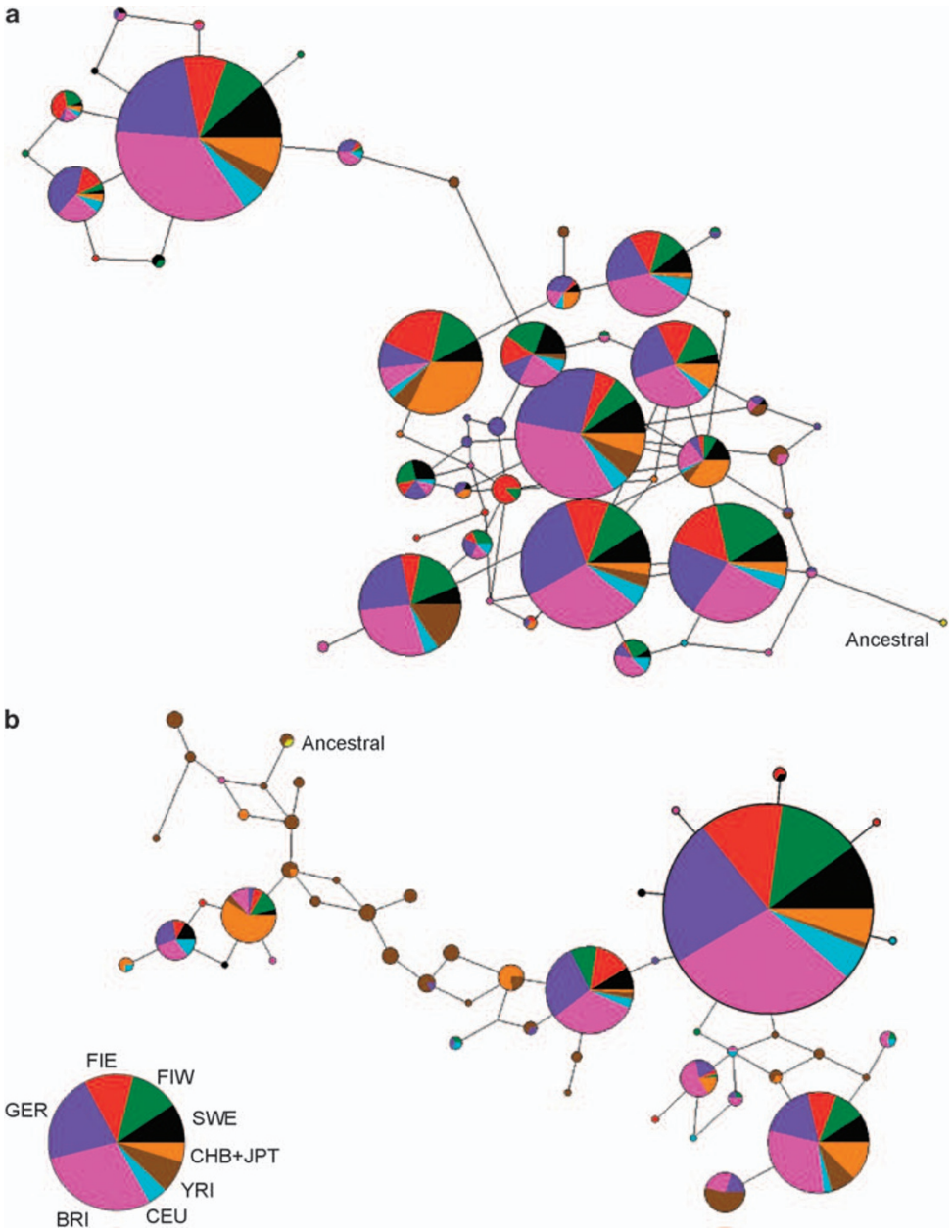

Figure 2 Median-joining networks of haplotypes in the regions Chr11:87 480 000–87 590 000 containing 15 SNPs in the *RAB38* gene (**a**), and Chr5:145 970 000–146 030 000 containing 13 SNPs in the *PPP2R2B* gene (**b**). Nodes denote the haplotypes, with their size corresponding to the overall frequency. The legend showing the color codes of the population frequencies also shows the relative sizes of the study samples in the entire data to assist the interpretation of haplotype frequency differences. The branches connecting the haplotypes denote the SNPs differing between haplotypes. The ancestral (chimpanzee) haplotype is marked in yellow.

coalescent simulations, natural selection increased both iHS and LRH values, and the values calculated from real data fell between the neutral and selected simulations, as expected (Supplementary Figure 1). The power to detect a selection signal increased with both sample size and SNP density for the iHS and LRH statistics and their combination, ranging from about 10% to over 80% (Figure 4). In the simulated selection scenario, the highest power was reached by using iHS statistic alone – however, the simulations encompass only a single population, which can underestimate the power of the combined iHS+LRH analysis performed on a data set with several populations.

Allele frequency simulations

Simulations of allele frequency differences between populations showed that for closely related populations – such as the North European population pair in the analysis – recent natural selection acting on one population has to be very strong to increase the allele frequency differences notably, because migration efficiently evens out allele frequency differences. Between continents – with a very low migration rate between the populations – even weak selection increases F_{ST}, and strong selection may lead to extreme differentiation (Figure 5, Supplementary Figure 2). Thus, as expected, the relatedness

of populations has a major effect on the possibility of using population differentiation-based tests for detecting positive selection.

DISCUSSION

In this study, we used genome-wide data from 250 000 and 500 000 SNPs to search for loci affected by recent positive natural selection in North European populations. We found convincing evidence of selection in 60 loci, 21 of which have not been discovered in previous scans for selection.

Many of the regions with strong signs of selection contain several genes with particularly interesting functions, although further studies are needed to fully determine which are the actual genes and variants behind the selective advantage. Of the two examples visualized as median-joining networks, the *RAB38* gene is expressed in melanocytes, and its disruption in mice causes oculocutaneous albinism, lung disease, and platelet deficiency.[29–31] This makes *RAB38* an interesting novel candidate locus for human pigmentation. The *PPP2R2B* gene regulates neuronal apoptosis and may affect adenovirus replication.[32,33] Yet another intriguing gene is *RGS9* that affects vision adaptation in different light conditions.[34,35] However, lack of selection signal in individual loci in this study should not be interpreted as absence of selection, as there are several interesting regions lacking sufficient SNP coverage – including the loci of, for example, the *LCT* and *OCA2* genes and the *CYP3A* region, all well-known candidates for recent natural selection.[9,36,37] In addition, the iHS and LRH tests have good power to detect selected haplotypes only in a relatively narrow frequency range, which may be the reason why some other well-established candidate genes, such as *SLC24A5*[38] and *MYOA5*,[9] show a lower signal below the chosen thresholds of this study.

Many of the selected regions contain genes associated with human disease, such as interferon gamma (*IFNG*), nitric oxide synthase 1 (neuronal) adaptor protein (*NOS1AP*), cadherin 13 (*CDH13*), and the *APOE* cluster (Table 2).[39–45] Altogether, we observed a statistically significant enrichment of genes associated with complex disease,

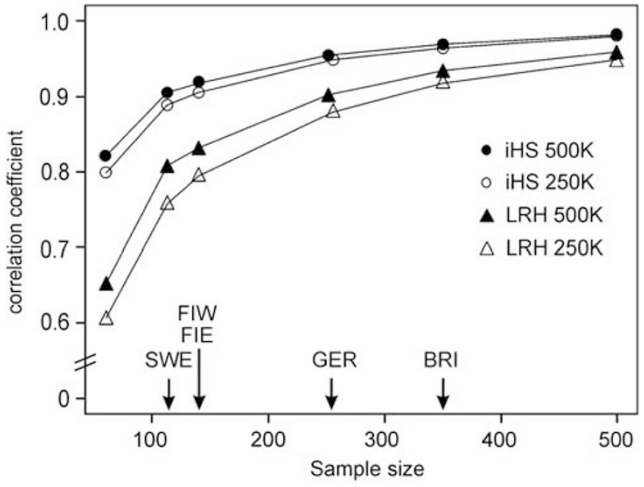

Figure 3 Correlation of the iHS and LRH values of SNPs between a sample of 700 British individuals and samples of various sizes in 250K and 500K data sets. The arrows indicate the sizes of the population samples in this study.

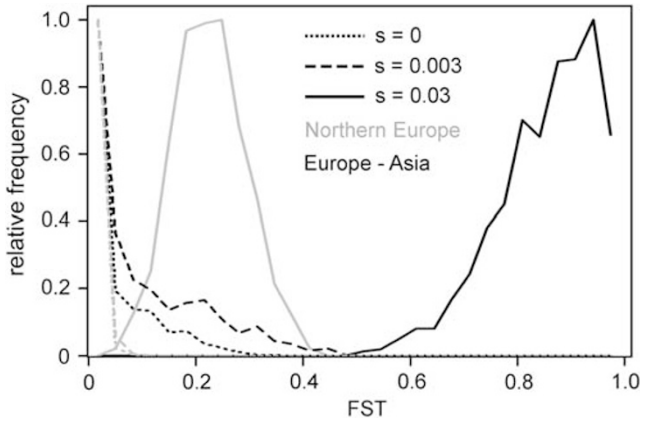

Figure 5 Distributions of F_{ST} in simulations of allele frequencies after 480 generations (about 12 000 years), with a starting allele frequency of 0.1 and three different selection coefficients (s) in two populations with demographies corresponding to a European–Asian population pair and two North European populations.

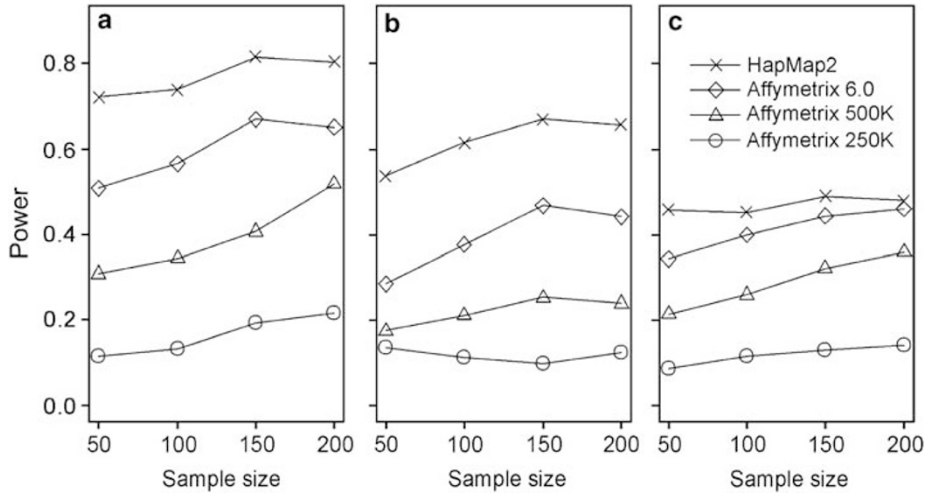

Figure 4 Power of the iHS test alone (**a**), LRH test alone (**b**), and iHS and LRH combined (**c**) in data sets of different SNP densities and sample sizes based on coalescent simulations.

which is consistent with earlier studies showing a connection between natural selection and certain types of human disease.[1,2,20,46] However, the increased population differentiation because of positive selection makes these loci particularly vulnerable to false-positive associations.[47–49] As an example, in our data set, the haplotype in the *PDE11A* gene showing very high population differentiation has been reported to associate with depression in Mexican Americans[50] – possibly due to confounding population structure. Thus, additional caution is necessary when disease association is observed in selected genes, which poses a further challenge for discovering functions of genes under natural selection by genome-wide analysis.[46]

Both empirical analyses and simulation studies showed that sample size and SNP density affect the performance of the iHS and LRH statistics. Although very dense data sets clearly yield the best power, the available genome-wide data sets with large sample sizes also seem adequate for successful selection scans, which is consistent with earlier results.[20] In reality, however, the power of the tests is likely to be lower than our simulations indicate: non-African human populations have not been of constant size, patterns of genomic variation are much more complex than in the simulations, SNP density varies between regions, and the simulated selection scenario was one in which the statistics have been shown to have good power.[7] On the other hand, the power in the simulations may be lower than in the real data because of the absence of multipopulation comparisons.

Selection signals can be compared between populations to detect adaptive differences between populations, both worldwide and locally between closely related populations.[20] The populations of Northern Europe have been distinct and subject to partly different environmental conditions for some 10 000 years, and the selective events detected by iHS scans have been estimated to be younger than that, on average $\sim$6600 years old in non-African populations.[9] Thus, some of the observed differences in the selection signals between populations may represent local adaptations. However, many are likely to arise from false negatives caused by the low power to detect selection, and differences in power between populations because of different allele frequencies, demography, and sample size.[7,9,18,51]

This study is not intended to be an exhaustive analysis of positive selection in Northern Europe. Statistical methods to scan for signs of positive natural selection are plagued by several limitations, and no method is suitable for covering the full spectrum of natural selection.[46] Specifically, the iHS and LRH statistics are efficient only when the selected haplotype has not yet reached fixation. This may be at least a partial explanation to the non-overlapping patterns of iHS or LRH with F_{ST}, which may detect more ancient selection than the other tests. However, F_{ST} alone has been suggested to be a poor indicator of selection.[52] Complementary methods that seek for fixation in one population and segregation in others,[7,10] such as XP-EHH, are unlikely to be effective for closely related populations in which the allele frequency differences, even in the presence of selection, are much smaller than between continents, as shown by our simulations. Furthermore, the SNP selection by array providers poses a limitation: although Affymetrix 500K arrays are not based on the tag-SNP approach that is problematic for LD-based selection testing due to low marker density in regions with high LD,[20] the bias toward common alleles limits the choice of statistics, makes local adaptations unlikely to be represented in the arrays, and may lead to lack of coverage in interesting regions. At present, the field is lacking comprehensive studies comparing the performance of different methods in scenarios of different kinds of positive selection, data sets, and populations, making it difficult to choose the most effective combination of statistics.

Despite these limitations, we have used data from genome-wide arrays to detect 60 regions – many previously undiscovered – with strong evidence of recent natural selection in Northern Europe, and these loci will be interesting targets for follow-up studies. Our study demonstrates the usefulness of genome-wide data sets for analysis of natural selection; particularly, the possibility to analyze large samples from a wide spectrum of human populations is a significant advantage. Furthermore, these data sets may prove useful for studying differences between populations due to local adaptations. However, the precise identification of the genes and variants behind the selective advantage will require data from genomic sequencing, as well as functional studies.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ACKNOWLEDGEMENTS

The funding for this study has been provided by the Emil Aaltonen foundation, the Research Foundation of the University of Helsinki, the Graduate School in Computational Biology, Bioinformatics, and Biometry, the Sigrid Juselius Foundation, the Academy of Finland, the Swedish Research Council, the Finnish Cultural Foundation, the National Genome Research Network (NGFN), and the PopGen Biobank, both through the German Ministry of Education and Science, and DFG excellence cluster 'inflammation at interfaces'. Funding for the WTCCC project was provided by the Wellcome Trust under Award 076113. This study makes use of the data generated by the Wellcome Trust Case Control Consortium. A full list of the investigators who contributed to the generation of the data is available from www.wtccc.org.uk. Furthermore, we thank Ingegerd Fransson and the personnel at Bioinformatics and Expression analysis core facility at Karolinska Institutet for technical support.

1 Blekhman R, Man O, Herrmann L *et al*: Natural selection on genes that underlie human disease susceptibility. *Curr Biol* 2008; **12**: 883–889.
2 Bustamante CD, Fledel-Alon A, Williamson S *et al*: Natural selection on protein-coding genes in the human genome. *Nature* 2005; **7062**: 1153–1157.
3 Carlson CS, Thomas DJ, Eberle MA *et al*: Genomic regions exhibiting positive selection identified from dense genotype data. *Genome Res* 2005; **11**: 1553–1565.
4 Akey JM, Zhang G, Zhang K, Jin L, Shriver MD: Interrogating a high-density SNP map for signatures of natural selection. *Genome Res* 2002; **12**: 1805–1814.
5 Oleksyk TK, Zhao K, De La Vega FM, Gilbert DA, O'Brien SJ, Smith MW: Identifying selected regions from heterozygosity and divergence using a light-coverage genomic dataset from two human populations. *PLoS ONE* 2008; **3**: e1712.
6 O'Reilly PF, Birney E, Balding DJ: Confounding between recombination and selection, and the Ped/Pop method for detecting selection. *Genome Res* 2008; **8**: 1304–1313.
7 Sabeti PC, Varilly P, Fry B *et al*: Genome-wide detection and characterization of positive selection in human populations. *Nature* 2007; **7164**: 913–918.
8 Wang ET, Kodama G, Baldi P, Moyzis RK: Global landscape of recent inferred Darwinian selection for Homo sapiens. *Proc Natl Acad Sci USA* 2006; **1**: 135–140.
9 Voight BF, Kudaravalli S, Wen X, Pritchard JK: A map of recent positive selection in the human genome. *PLoS Biol* 2006; **3**: e72.
10 Tang K, Thornton KR, Stoneking M: A new approach for using genome scans to detect recent positive selection in the human genome. *PLoS Biol* 2007; **7**: e171.
11 Williamson SH, Hubisz MJ, Clark AG, Payseur BA, Bustamante CD, Nielsen R: Localizing recent adaptive evolution in the human genome. *PLoS Genet* 2007; **6**: e90.
12 International HapMap ConsortiumFrazer KA, Ballinger DG *et al*: A second generation human haplotype map of over 3.1 million SNPs. *Nature* 2007; **7164**: 851–861.
13 Hinds DA, Stuve LL, Nilsen GB *et al*: Whole-genome patterns of common DNA variation in three human populations. *Science* 2005; **5712**: 1072–1079.
14 Heath SC, Gut IG, Brennan P *et al*: Investigation of the fine structure of European populations with applications to disease association studies. *Eur J Hum Genet* 2008; **12**: 1413–1429.
15 Jakobsson M, Scholz SW, Scheet P *et al*: Genotype, haplotype and copy-number variation in worldwide human populations. *Nature* 2008; **7181**: 998–1003.
16 Li JZ, Absher DM, Tang H *et al*: Worldwide human relationships inferred from genome-wide patterns of variation. *Science* 2008; **5866**: 1100–1104.
17 Novembre J, Johnson T, Bryc K *et al*: Genes mirror geography within Europe. *Nature* 2008; **7219**: 274.

18 Salmela E, Lappalainen T, Fransson I *et al*: Genome-wide analysis of single nucleotide polymorphisms uncovers population structure in Northern Europe. *PLoS ONE* 2008; **10**: e3519.
19 Kimura R, Ohashi J, Matsumura Y *et al*: Gene flow and natural selection in oceanic human populations inferred from genome-wide SNP typing. *Mol Biol Evol* 2008; **8**: 1750–1761.
20 Pickrell JK, Coop G, Novembre J *et al*: Signals of recent positive selection in a worldwide sample of human populations. *Genome Res* 2009; **5**: 826–837.
21 Krawczak M, Nikolaus S, von Eberstein H, Croucher PJ, El Mokhtari NE, Schreiber S: PopGen: population-based recruitment of patients and controls for the analysis of complex genotype-phenotype relationships. *Community Genet* 2006; **1**: 55–61.
22 Wellcome Trust Case Control Consortium: Genome-wide association study of 14 000 cases of seven common diseases and 3000 shared controls. *Nature* 2007; **7145**: 661–678.
23 International HapMap Consortium: A haplotype map of the human genome. *Nature* 2005; **7063**: 1299–1320.
24 Sabeti PC, Reich DE, Higgins JM *et al*: Detecting recent positive selection in the human genome from haplotype structure. *Nature* 2002; **6909**: 832–837.
25 Weir BS, Cockerham CC: Estimating F-statistics for the analysis of population structure. *Evolution* 1984; **38**: 1358–1370.
26 Bandelt HJ, Forster P, Rohl A: Median-joining networks for inferring intraspecific phylogenies. *Mol Biol Evol* 1999; **1**: 37–48.
27 Polzin T, Daneschmand SV: On Steiner trees and minimum spanning trees in hypergraphs. *Operations Res Lett* 2003; **31**: 12–20.
28 Spencer CC, Coop G: SelSim: a program to simulate population genetic data with natural selection and recombination. *Bioinformatics* 2004; **18**: 3673–3675.
29 Loftus SK, Larson DM, Baxter LL *et al*: Mutation of melanosome protein RAB38 in chocolate mice. *Proc Natl Acad Sci USA* 2002; **7**: 4471–4476.
30 Oiso N, Riddle SR, Serikawa T, Kuramoto T, Spritz RA: The rat Ruby (R) locus is Rab38: identical mutations in Fawn-hooded and Tester-Moriyama rats derived from an ancestral Long Evans rat sub-strain. *Mamm Genome* 2004; **4**: 307–314.
31 Osanai K, Oikawa R, Higuchi J *et al*: A mutation in Rab38 small GTPase causes abnormal lung surfactant homeostasis and aberrant alveolar structure in mice. *Am J Pathol* 2008; **5**: 1265–1274.
32 Ben-Israel H, Sharf R, Rechavi G, Kleinberger T: Adenovirus E4orf4 protein down-regulates MYC expression through interaction with the PP2A-B55 subunit. *J Virol* 2008; **19**: 9381–9388.
33 Dagda RK, Merrill RA, Cribbs JT *et al*: The spinocerebellar ataxia 12 gene product and protein phosphatase 2A regulatory subunit Bbeta2 antagonizes neuronal survival by promoting mitochondrial fission. *J Biol Chem* 2008; **52**: 36241–36248.
34 Stockman A, Smithson HE, Webster AR *et al*: The loss of the PDE6 deactivating enzyme, RGS9, results in precocious light adaptation at low light levels. *J Vis* 2008; **8**: 10.1–10.10.
35 Hartong DT, Pott JW, Kooijman AC: Six patients with bradyopsia (slow vision): clinical features and course of the disease. *Ophthalmology* 2007; **12**: 2323–2331.
36 Bersaglieri T, Sabeti PC, Patterson N *et al*: Genetic signatures of strong recent positive selection at the lactase gene. *Am J Hum Genet* 2004; **6**: 1111–1120.
37 Thompson EE, Kuttab-Boulos H, Witonsky D, Yang L, Roe BA, Di Rienzo A: CYP3A variation and the evolution of salt-sensitivity variants. *Am J Hum Genet* 2004; **6**: 1059–1069.
38 Lamason RL, Mohideen MA, Mest JR *et al*: SLC24A5, a putative cation exchanger, affects pigmentation in zebrafish and humans. *Science* 2005; **5755**: 1782–1786.
39 Arking DE, Pfeufer A, Post W *et al*: A common genetic variant in the NOS1 regulator NOS1AP modulates cardiac repolarization. *Nat Genet* 2006; **6**: 644–651.
40 Levy D, Larson MG, Benjamin EJ *et al*: Framingham Heart Study 100K Project: genome-wide associations for blood pressure and arterial stiffness. *BMC Med Genet* 2007; **8** (Suppl 1): S3.
41 Pacheco AG, Cardoso CC, Moraes MO: IFNG +874T/A, IL10 −1082G/A and TNF −308G/A polymorphisms in association with tuberculosis susceptibility: a meta-analysis study. *Hum Genet* 2008; **5**: 477–484.
42 Silverberg MS, Cho JH, Rioux JD *et al*: Ulcerative colitis-risk loci on chromosomes 1p36 and 12q15 found by genome-wide association study. *Nat Genet* 2009; **2**: 216–220.
43 Wratten NS, Memoli H, Huang Y *et al*: Identification of a schizophrenia-associated functional noncoding variant in NOS1AP. *Am J Psychiatry* 2009; **166**: 434–441.
44 Coon KD, Myers AJ, Craig DW *et al*: A high-density whole-genome association study reveals that APOE is the major susceptibility gene for sporadic late-onset Alzheimer's disease. *J Clin Psychiatry* 2007; **4**: 613–618.
45 Willer CJ, Sanna S, Jackson AU *et al*: Newly identified loci that influence lipid concentrations and risk of coronary artery disease. *Nat Genet* 2008; **2**: 161–169.
46 Nielsen R, Hellmann I, Hubisz M, Bustamante C, Clark AG: Recent and ongoing selection in the human genome. *Nat Rev Genet* 2007; **11**: 857–868.
47 Marchini J, Cardon LR, Phillips MS, Donnelly P: The effects of human population structure on large genetic association studies. *Nat Genet* 2004; **5**: 512–517.
48 Lange EM, Sun J, Lange LA *et al*: Family-based samples can play an important role in genetic association studies. *Cancer Epidemiol Biomarkers Prev* 2008; **9**: 2208–2214.
49 Freedman ML, Reich D, Penney KL *et al*: Assessing the impact of population stratification on genetic association studies. *Nat Genet* 2004; **4**: 388–393.
50 Wong ML, Whelan F, Deloukas P *et al*: Phosphodiesterase genes are associated with susceptibility to major depression and antidepressant treatment response. *Proc Natl Acad Sci USA* 2006; **41**: 15124–15129.
51 Teshima KM, Coop G, Przeworski M: How reliable are empirical genomic scans for selective sweeps? *Genome Res* 2006; **6**: 702–712.
52 Hofer T, Ray N, Wegmann D, Excoffier L: Large allele frequency differences between human continental groups are more likely to have occurred by drift during range expansions than by selection. *Ann Hum Genet* 2009; **1**: 95–108.
53 Huttley GA, Smith MW, Carrington M, O'Brien SJ: A scan for linkage disequilibrium across the human genome. *Genetics* 1999; **4**: 1711–1722.
54 Nielsen R, Williamson S, Kim Y, Hubisz MJ, Clark AG, Bustamante C: Genomic scans for selective sweeps using SNP data. *Genome Res* 2005; **11**: 1566–1575.

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