Genetic Differences between Five European Populations

Original Paper

Human Heredity

Hum Hered 2010;70:141–149
DOI: 10.1159/000313854

Received: January 15, 2010
Accepted after revision: April 17, 2010
Published online: July 8, 2010

Genetic Differences between Five European Populations

Valentina Moskvina[a] Michael Smith[a] Dobril Ivanov[a] Douglas Blackwood[b] David StClair[c] Christina Hultman[d] Draga Toncheva[e] Michael Gill[f] Aiden Corvin[f] Colm O'Dushlaine[f] Derek W. Morris[f] Naomi R. Wray[g] Patrick Sullivan[h] Carlos Pato[i] Michele T. Pato[i] Pamela Sklar[j] Shaun Purcell[j] Peter Holmans[a] Michael C. O'Donovan[a] Michael J. Owen[a] George Kirov[a] International Schizophrenia Consortium[1]

[a]MRC Centre for Neuropsychiatric Genetics and Genomics, Department of Psychological Medicine and Neurology, School of Medicine, Cardiff University, Cardiff, [b]Division of Psychiatry, School of Molecular and Clinical Medicine, University of Edinburgh, Edinburgh, and [c]Institute of Medical Sciences, University of Aberdeen, Aberdeen, UK; [d]Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden; [e]Department of Medical Genetics, University Hospital Maichin Dom, Sofia, Bulgaria; [f]Neuropsychiatric Genetics Research Group, Department of Psychiatry and Institute of Molecular Medicine, Trinity College Dublin, Dublin, Ireland; [g]Queensland Institute of Medical Research, Brisbane, Qld., Australia; [h]Department of Genetics, University of North Carolina, Chapel Hill, N.C., [i]Department of Psychiatry and the Behavioral Sciences, Franz Alexander Chair in Psychiatry, Keck School of Medicine at USC, Los Angeles, Calif., [j]Broad Institute of Harvard and MIT, Cambridge, Mass., and Center for Human Genetic Research, Massachusetts General Hospital, Boston, Mass., USA

Key Words
Population · Gene · Stratification · Pigmentation · Immunity

Abstract
Aims: We sought to examine the magnitude of the differences in SNP allele frequencies between five European populations (Scotland, Ireland, Sweden, Bulgaria and Portugal) and to identify the loci with the greatest differences. ***Methods:*** We performed a population-based genome-wide association analysis with Affymetrix 6.0 and 5.0 arrays. We used a 4 degrees of freedom χ^2 test to determine the magnitude of stratification for each SNP. We then examined the genes within the most stratified regions, using a highly conservative cutoff of $p < 10^{-45}$. ***Results:*** We found 40,593 SNPs which are genome-wide significantly ($p \leq 10^{-8}$) stratified between these populations. The largest differences clustered in gene ontology categories for immunity and pigmentation. Some of the top loci span genes that have already been reported as highly stratified: genes for hair color and pigmentation *(HERC2, EXOC2, IRF4)*, the *LCT* gene, genes involved in NAD metabolism, and in immunity (HLA and the Toll-like receptor genes TLR10, TLR1, TLR6). However, several genes have not previously been reported as stratified within European populations, indicating that they might also have provided selective advantages: several zinc finger genes, two genes involved in glutathione synthesis or function, and most intriguingly, *FOXP2*, implicated in speech development. ***Conclusion:*** Our analysis demonstrates that many SNPs show genome-wide significant differences within European pop-

[1] The full details of this consortium are given in the online supplementary materials. For all online supplementary materials, see www.karger.com/doi/10.1159/000313854.

KARGER

Fax +41 61 306 12 34
E-Mail karger@karger.ch
www.karger.com

0001–5652/10/0702–0141$26.00/0

Accessible online at:
www.karger.com/hhe

Valentina Moskvina and George Kirov
MRC Centre for Neuropsychiatric Genetics and Genomics
Department of Psychological Medicine and Neurology, School of Medicine
Cardiff University, Heath Park, Cardiff CF14 4XN (UK), Tel. +44 292 074 6611
Fax +44 292 068 7068, E-Mail MoskvinaV1@cardiff.ac.uk or Kirov@cardiff.ac.uk

ulations and the magnitude of the differences correlate with the geographical distance. At least some of these differences are due to the selective advantage of polymorphisms within these loci.

Introduction

Traditionally, genetic differences between populations have been identified with genetic markers on the Y-chromosome, mitochondrial DNA, alleles that have reached fixation in certain populations, and classical genetic markers such as blood groups. Recent studies have provided genome-wide information on differences between populations and have shown that despite close genetic similarities among white Europeans, some subtle but informative differences exist. When using genome-wide arrays, the cumulative effects of these differences in allele frequencies allow the place of birth of individual subjects to be predicted quite well [1–4], to the point that 90% of individuals can be placed to within 700 km of their reported origin by using SNP genotypes [4, 5], or within single countries, e.g. Finland [3], Iceland [4] and the UK [6]. Therefore, the use of these arrays has opened up possibilities to explore the population history of very closely related ethnic groups. The loci showing the highest stratification contain some very likely candidate genes that can account for these differences via effects on selection [6]. The Wellcome Trust Case Control Consortium (WTCCC) was the first study to report a set of highly differentiated SNPs clustered in several genomic regions, which had different allele frequencies even within the boundaries of a single European country, the UK. The strongest signals came from loci harboring genes involved in immunity, lactose metabolism, and the gene encoding for NAD synthetase 1, which might have a role in the prevention of pellagra. More recently, some of these loci were confirmed [3], while another study [7] revealed a different set of informative markers for European ancestry.

Here we describe the genetic differences between five European populations (Scotland, Ireland, Sweden, Bulgaria and Portugal) using the data from a genome-wide association (GWA) study [8]. We show that many SNPs reach genome-wide significant differences between these populations, even after Bonferroni correction for multiple testing, and as expected, the differences correlate with the geographical distance between these populations. Some of these SNPs are within genes which were already reported in previous publications, such as genes involved in immunity and pigmentation. We have also identified a number of genes for which a selective advantage has not been clearly described before. Our study differs from previous ones on this topic in that it contains larger sample sizes from each population studied, allowing the generation of unequivocal statistical significance, and the identification of a larger number of stratified loci. Identifying these loci can help elucidate some of the selection factors that have shaped the recent population history of Europe. We focus on the specific genes that show the largest genetic stratification present across Europe.

Materials and Methods

Subjects

We used five samples studied in a recent GWA study on schizophrenia [8, 9]: Irish, Scottish, Swedish, Bulgarian and Portuguese. The genetic background of participants of these five samples is more stable due to relatively low historic rates of immigration into these countries (we did not include the population recruited in London, which is likely to include many migrants). These populations are from four corners of the European continent (North, North-West, South-West and South-East), maximizing our ability to detect differences. The North-West/South-East gradient has been shown to be the strongest gradient for genetic differences in Europe [1, 7, 10].

Sample sizes were as follows: 1,129 individuals from Bulgaria (482 of these had been cases diagnosed with schizophrenia, and 647 controls); 1,142 from Ireland (275 cases and 866 controls); 656 from Scotland (369 cases and 287 controls); 620 from Sweden (390 cases and 230 controls); and 563 from Portugal (347 cases and 216 controls). For our primary analysis, we combined cases and controls in this study because we observed that the magnitude of the genetic differences that exist between cases and controls [9] are >20 orders of magnitude lower than the cutoff criterion that we used in our analysis ($p < 10^{-30}$). We also repeated the analysis restricted only to controls to confirm that the results are not biased due to the inclusion of patients. The top regions (described in the Discussion) remained the same, and the correlation between the $-\log_{10}$(p values) in the combined and the control-only analyses for those 11 loci was r = 0.95. Therefore, we provide the results on the full samples, as larger sample sizes produce less chance fluctuation (we provide the controls-only p values for the top-stratified SNP within every gene in online suppl. table 1).

Genotyping

Genotyping was performed at the Broad Institute, USA, on Affymetrix 6.0 and 5.0 arrays. Details are given in our primary GWA study paper [9]. The Affymetrix 6.0 and 5.0 arrays provide genotypes for 906,600 and 500,568 SNPs, respectively. For the analysis, we used 363,411 SNPs passing the standard quality control criteria [8], excluding missing genotypes >10% per individual and minor allele frequencies <1% and being present on both

arrays. We also excluded individuals that were shown to be related, and those who were population outliers (more information on the filtering criteria are presented in [8, 9]). The exclusion of population outliers effectively excluded subjects that might have been migrants to that country, or offspring of parents from different countries.

Statistical Analysis

We compared each population against each other pair-wise, using the Armitage trend test with 1 degree of freedom (d.f.). The different sample sizes result in more significant p values between the larger samples, even when the distribution of genotypes is the same. In order to account for this, we scaled down the genotype counts, making each population the same size, so that all differences were directly comparable. In order to achieve this, we multiplied the observed genotype counts by the ratio between the smallest sample size and the current sample size $\tilde{n}_k = N_s/N_c \cdot n_k$, where k indicates 11, 12 or 22 genotypes, n_k and $\tilde{n}_k$ are the observed and adjusted counts of genotype k; N_c is the current sample size and $N_s = 563$ is the sample size of the smallest population (Portugal). This scaling-down produces more conservative estimates of the differences.

In order to find the SNPs that are most highly differentiated between all five populations, we applied a 4 d.f. χ^2 test for a 5×2 contingency table (5 – number of populations, 2 – number of alleles). The populations were not scaled by size for this analysis.

Another way to examine the patterns of genetic variation between populations is through the Wright's fixation index F_{ST} [11]. F_{ST} was estimated according to Wright's approximate formula $F_{ST} = (H_T - H_S)/H_T$, where H_T represents expected heterozygosity per locus of the total population and H_S is calculated as weighted average over populations of expected heterozygosity of each subpopulation (weighted by sample size). In the current study, they ranged from 0 to 0.061. However, F_{ST} values correlated almost perfectly ($r = 0.999$) with the negative $\log_{10}$ of the p values from the above χ^2 test, indicating that the two tests provide the same measures for the genetic variation. This effect has been observed before [12]. For the rest of the paper we use the p value results, as these are more intuitive for readers who are not population geneticists. They are also easier to use in comparisons of the different analyses that we performed and give a more familiar measure for population stratification magnitude in GWA studies (F_{ST} results for the best SNP within every gene are given in online suppl. table 1).

Having processed the data and identified the SNPs that displayed the largest differences between populations, we determined the genes to which these SNPs mapped. As these genes were found to cluster in discrete loci, we selected only the most significant loci in the genome in order to limit our discussion to the top hits, and used an arbitrary cutoff of $p < 10^{-45}$ for a SNP association with ethnic origin derived with the χ^2 test. We defined the region involved, again arbitrarily, as flanked by SNPs that were stratified at $p < 10^{-30}$, with a gap of >500 kb distance that contains no such SNPs, as defining the end of the region on either side.

We also assessed population structure within the data using principal components analysis as implemented in EIGENSTRAT [13]. Eigenvectors were calculated based on a linkage disequilibrium (LD)-pruned subset of 101,532 SNPs with $r^2 \leq 0.5$. LD pruning was performed using PLINK version 1.06 [14]. We show the plot of the first two principal components extracted from EIGENSTRAT.

Gene Ontology Analysis

Standard methods for testing of enrichment of gene ontology (GO) categories on a gene list could not be used, since these rely on there being a single measurement per gene, whereas GWA study data consists of different numbers of SNPs per gene, each with a measure of significance of differentiation. These are not independent, due to LD. We therefore used the ALIGATOR program [15] to test enrichment of GO categories on lists of significantly-differentiated SNPs. SNPs were assigned to genes if their physical position (NCBI SNP build 129) lay between the start and end points of the gene (as defined by NCBI sequence build 36.3). A list of significant genes was defined as those genes that contain a SNP that is stratified at a conservative $p < 10^{-30}$, in order to minimize the noise. Each gene was counted only once, regardless of the number of significant SNPs it contained, thereby correcting for bias caused by multiple significant SNPs in a gene arising from LD. As described by Holmans et al. [15], 50,000 random gene lists of the same length were simulated, and the number of genes in each category present on each simulated gene list compared to that observed on the actual list of significant genes. Thus, an empirical p value for enrichment was obtained for each category. The gene lists were simulated by sampling SNPs at random, thus correcting for variable numbers of SNPs per gene. An empirical distribution for the number of significantly enriched categories was also obtained, enabling a test for an excess of such categories in the real data to be performed. All GO categories containing 3 or more genes were tested, but a minimum of 2 significant genes was required for a category to count as over-represented (to prevent small categories being over-represented on the basis of one chance hit).

Results

There are ten pair-wise comparisons of the five populations. Given the stratification between populations, not unexpectedly, there was a substantial excess of SNPs at a 5% significance level for the number of SNPs tested and the 10 comparisons performed (table 1). We opted to use a conservative Bonferroni correction (although many of the tests are not independent) and multiplied each of the p values by $363{,}411 \cdot 10$ (number of SNPs by number of comparisons). Despite this, there were still large numbers of genome-wide significant results in all comparisons. We present both the original results, to demonstrate the true number of significant results in the study, and the rescaled results (to the sample size of the smallest population, see Material and Methods), in order to provide a measure for the relative differences between the populations. The rescaled results are of course more conservative.

As expected, the number of significant differences correlated with the distance between the countries, but

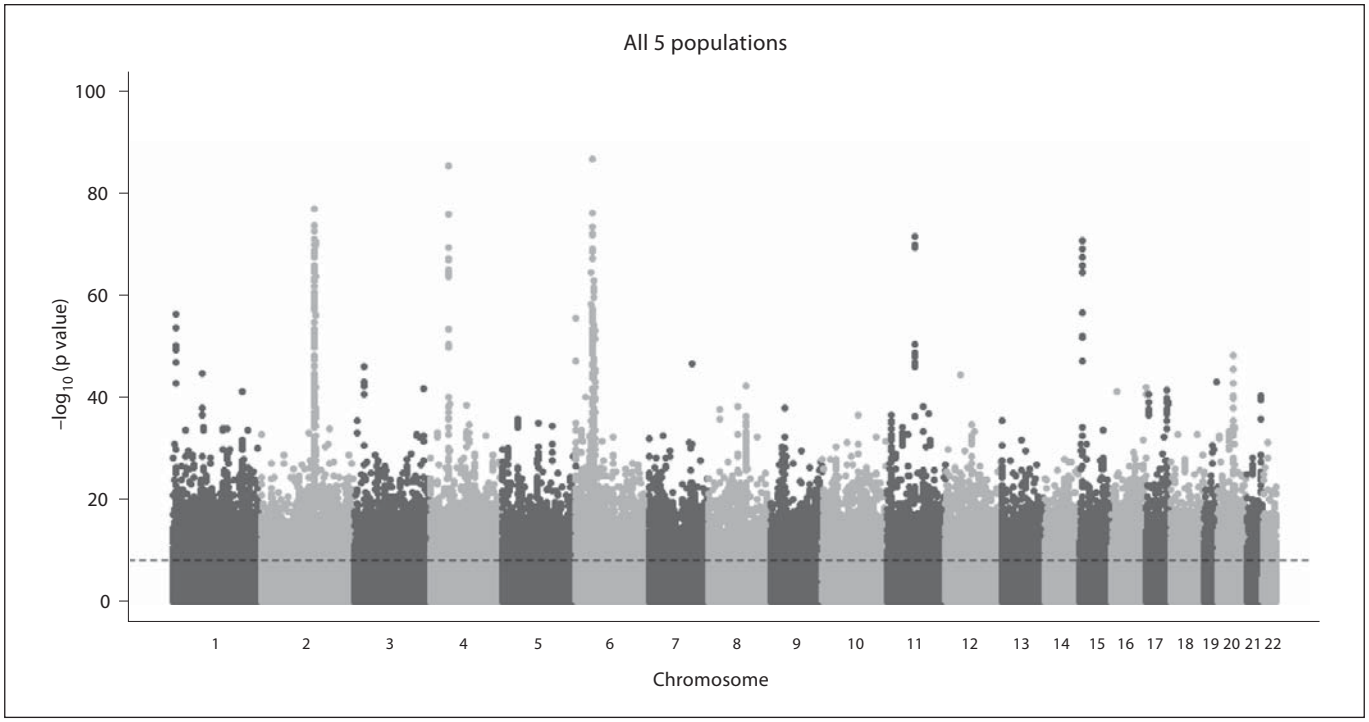


Fig. 1. Significance of the population stratification between the five populations for each SNP versus its genomic position.

Table 1. Pair-wise comparisons between five European populations

Comparison	Original results				Rescaled results			
			Bonferroni corrected				Bonferroni corrected	
	SNPs with $p \leq 0.05$, n	min. p value	SNPs with $p \leq 0.05$, n	min. p value	SNPs with $p \leq 0.05$, n	min. p value	SNPs with $p \leq 0.05$, n	min. p value
Bulgaria–Ireland	184,367	3.3×10^{-65}	21,603	1.2×10^{-58}	126,714	4.2×10^{-33}	2,946	1.5×10^{-26}
Bulgaria–Sweden	165,431	3.1×10^{-40}	11,091	1.1×10^{-33}	135,978	3.0×10^{-26}	3,542	1.1×10^{-19}
Bulgaria–Scotland	149,935	1.2×10^{-50}	7,114	4.4×10^{-44}	116,599	9.2×10^{-35}	1,895	3.3×10^{-28}
Scotland–Ireland	37,024	3.0×10^{-13}	14	1.1×10^{-6}	17,063	3.9×10^{-9}	2	0.014
Scotland–Sweden	89,071	7.9×10^{-22}	306	2.9×10^{-15}	78,367	2.6×10^{-19}	159	9.3×10^{-13}
Ireland–Sweden	125,217	5.2×10^{-34}	2,309	1.9×10^{-27}	93,828	1.0×10^{-22}	455	3.7×10^{-16}
Bulgaria–Portugal	126,493	2.1×10^{-28}	2,796	7.5×10^{-22}	101,079	3.1×10^{-17}	659	1.1×10^{-10}
Scotland–Portugal	128,232	3.2×10^{-33}	2,637	1.2×10^{-26}	121,903	2.1×10^{-30}	1,912	7.7×10^{-24}
Ireland–Portugal	154,435	5.7×10^{-50}	8,446	2.1×10^{-43}	128,950	1.6×10^{-39}	2,823	5.8×10^{-33}
Sweden–Portugal	159,353	9.9×10^{-33}	8,188	3.6×10^{-26}	155,455	2.0×10^{-30}	7,048	7.4×10^{-24}

even the neighboring Scottish and Irish samples had 14 SNPs with alleles that differed in frequency at a genome-wide significant level after Bonferroni correction.

Figure 1 shows the negative $\log_{10}$ of the p values produced by the χ^2 test for the five populations (y-axis) according to their genomic positions (x-axis). Chromosomes are indicated in shades of grey. The black horizontal line indicates the genome-wide significance level ($p = 10^{-8}$). The corresponding figures for each pair-wise comparison between all populations are presented in online

Table 2. Regions with the highest differences between European populations

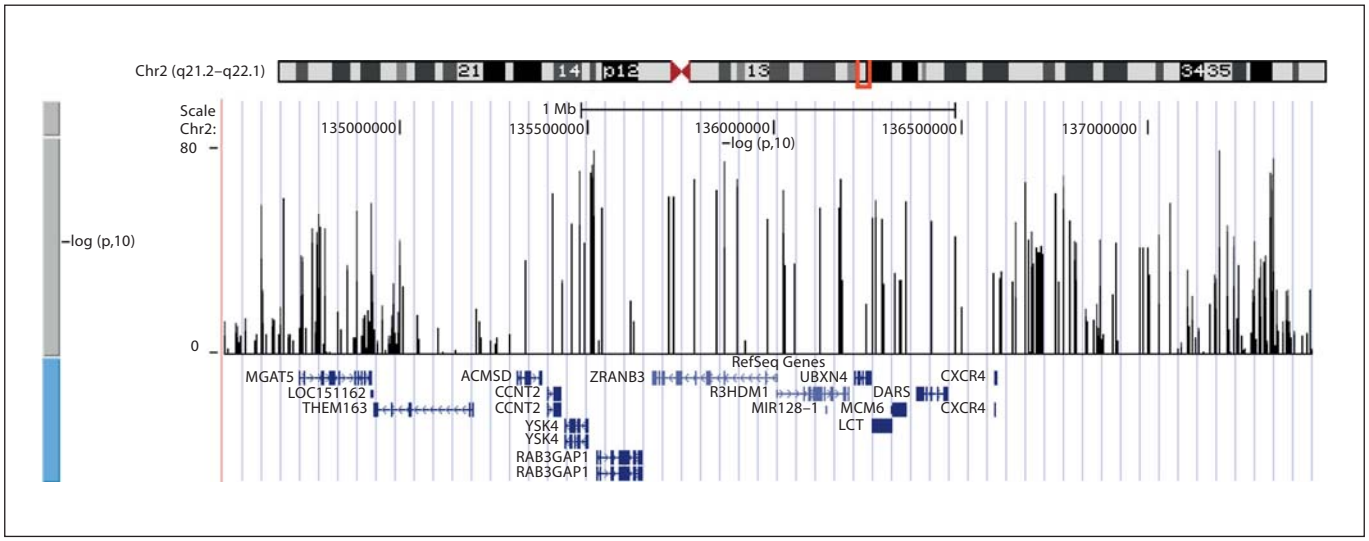
Region flanked by SNPs with $p < 10^{-30}$	Relevant gene function	Genes within the regions	Corresponding region in studies [3] or [6]	Region size kb	SNPs with $p < 10^{-30}$ n	Most significant SNP and allele frequencies (Bg/Ir/Sc/Sw/Port)	Significance $-\log_{10}(p)$
Chr1: 8.27–8.66	Arginine-glutamic acid dipeptide repeats	RERE		390	8	rs12136766 0.499/0.321/0.316/ 0.297/0.485	56.31
Chr2: 134.63–137.34	Immunity (CXCR4), NAD (ACMSD), lactase (LCT)	MGAT5, TMTM163, **ACMSD**, CCNT2, YSK4, RAB3GAP, UBXN4, **LCT**, MCM6, DARS, CXCR4	134.75–137.46	2,710	109	rs7582192 0.082/0.286/0.268/ 0.252/0.141	76.79
Chr4: 38.38–38.58	Immunity	**TLR10, TLR 1, TLR 6** (Toll-like receptors)	38.53–38.74*	200	20	rs6835514 0.420/0.164/0.134/ 0.286/0.363	85.14
Chr6: 0.33–0.49	Skin, hair, eye color	**IRF4, EXOC2**	0.33–0.49	160	9	rs6920655 0.404/0.207/0.318/ 0.403/0.398	55.43
Chr6: 28.5–28.64	Zink fingers; glutathione peroxidase family	ZSCAN3, ZSCAN12, ZSCAN23, **GPX5, GPX6**		210	3	rs13215804 0.179/0.401/0.392/ 0.251/0.316	64.30
Chr6: 29.41–35.28	Immunity	**HLA** region	31.1–31.6*	5,870	114	rs486416 0.169/0.450/0.392/ 0.367/0.181	86.61
Chr6: 37.86–38.24	Zink finger, protein-protein interactions	**ZFAND3, BTBD9**		380	10	rs2281266 0.255/0.433/0.438/ 0.433/0.289	51.40
Chr7: 113.70–114.14	Speech	FOXP2		440	2	rs1378769 0.104/0.021/0.028/ 0.018/0.069	46.50
Chr11: 70.81–70.90	NAD	DHCR7, **NADSYN1**	70.78–70.93	90	10	rs2276360 0.353/0.160/0.190/ 0.338/0.372	71.22
Chr15: 25.69–26.20	Skin, hair, eye color	**OCA2**, **HERC2**	26.20	510	12	rs8041209 0.139/0.035/0.042/ 0.015/0.160	70.51
Chr20: 33.00–33.31	Glutathione synthetase; protection of cells from oxidative damage	**GSS**, MYH7B, TRPC4AP, EDEM2, PROCR, MMP24		310	9	rs619865 0.035/0.149/0.138/ 0.086/0.051	48.08

The genes discussed in the present study are printed in bold. * = regions overlapping with those in [3]. Note the absence of LCT in [3].

supplementary figure 1. They give a direct comparison of the population differences, which are greater between more distant populations. This conclusion is confirmed by the Principal Component analysis (online suppl. fig. 2a) which placed the two-dimensional positions of the five populations quite accurately over the map of Europe (online suppl. fig. 2b).

We then identified the list of genome-wide significant SNPs between all five populations using the 4 d.f. χ^2 test, i.e. we tested for differences in the allele counts between all five populations. There were 40,593 SNPs that were genome-wide significant with p values $\leq 10^{-8}$.

GO analysis on the top-ranked SNPs (with p values $<10^{-30}$, which lay in a total of 100 genes) using ALIGATOR revealed a number of significantly over-represented categories; 43 categories were over-represented at $p < 0.01$ (not shown), and 18 at $p < 0.001$ (online suppl. table 2). Both numbers are significantly greater than the number of categories expected (13.23 and 1.98 at $p < 0.01$ and $p < 0.001$, respectively) generated from random sets of SNPs



Color version available online

Fig. 2. Genetic region around the *LCT* gene. Vertical bars indicate the strength of the stratification for each SNP on a scale of 0–80 for the –log(p,10).

of the same length (p = 0.024 and p = 0.005, respectively). Several of the categories are related to the MHC region on Chr6, which is known to exhibit considerable long-range LD. It is therefore possible that the significance of immunity-related categories is inflated. We therefore removed all SNPs and genes in the MHC region and repeated the analysis (results not presented). The excess number of over-represented categories was not solely due to the MHC region, and the top-ranking GO categories remained largely unchanged.

Regions containing the most significantly differentiated SNPs (at least one SNP at $p < 10^{-45}$ and the boundary defined with lack of p value $<10^{-30}$ for an interval of >500 kb, see Materials and Methods) and the genes within these intervals are shown in table 2. The relevance of these genes and comparisons with previous findings follow in the Discussion. Figure 2 shows the significance of SNPs and the positions of genes in one of these intervals, which is one of the best-known stratified regions: around the *LCT* gene on Chr 2. Similar figures for the 11 top-ranked loci are presented in online supplementary figure 3.

Discussion

A trend for a NW/SE gradient for genetic differences between European populations was observed over 30 years ago using a limited number of genetic markers [10]. The recent GWA studies provided a much more detailed picture of these differences and have shown a remarkably close relationship between genetic similarity and place of birth in the European continent (see Introduction). In this paper we confirmed the genetic relationship between populations and their geographic distribution in Europe [1, 3–5]. This is demonstrated by the magnitude of differences between the five populations (online suppl. fig. 1) and the results of the Principal Component analysis (online suppl. fig. 2).

The main aim of this paper was, however, to identify the most highly stratified genes and the mechanisms that might have contributed to these differences. The large sample size and availability of populations from four corners of Europe allowed us to obtain extremely high significance levels to confidently identify the top hits. Some of the genes within these regions have already been identified in previous research (e.g. [3, 6], indicated in a separate column in table 2) and have plausible biological effects on selection, while others have not been within the top hits in these studies, and for some genes we are not aware of any obvious effects on selection within Europe. Most differences are likely to have been caused by ancient differences between formerly isolated groups during the population of Europe [10]. The largest differences are more likely to have been caused by selective forces operating differently in different parts of the continent, e.g. by epidemics and nutritional factors. We now discuss the main groups of genes within our top 11 regions from table 2 and online supplementary figure 3.

Genes Involved in Hair, Skin, and Eye Color

Our GO categories analysis placed the genes for pigmentation at the top of our results. The region on Chr6: 0.33–0.49 Mb, including *IRF4* and *EXOC2*, was previously associated with hair color, freckles and skin sensitivity in a GWA study in 2,986 Icelanders and replicated in samples of 2,718 Icelanders and 1,214 Dutch people [16]. Another group [17] showed that rs12203592 within the *IRF4* gene is associated with hair, skin and eye color and tanning ability. The association between *IRF4* genotypes at rs12203592 with eye and particularly with hair color was confirmed in a US cohort [18]. Our best SNP in this region is rs6920655 ($p = 3.7 \times 10^{-56}$), ~30 kb away from *IRF4* and ~45 kb from *EXOC2*. *IRF4* is expressed in melanocytes and is suggested as a sensitive marker for metastatic melanomas and benign melanocytic nevi [19], making it the more plausible candidate.

The gene for oculocutaneous albinism II (OCA2), and *HERC2* on Chr15: 25.69–26.20 Mb are also implicated in skin, hair, and eye color [17] and are within our most significant regions. Association of eye and hair color with SNPs in *OCA2* (rs7495174, rs6497268, rs11855019) and *HERC2* (rs1667394) was found by Sulem et al. [16]. These authors argue that since the link between *OCA2* and pigmentation is quite well established, the association with the *HERC2* gene is due to LD. However, rs12913832 in *HERC2* remained significant after adjusting for *OCA2* SNPs [17]. Our best SNP in this region also lies within *HERC2*: rs8041209 ($p = 3.1 \times 10^{-71}$). It is of course possible that the responsible gene is still *OCA2*, but there are regulatory elements for it nearby.

Some genes implicated before in the genetics of hair and skin color did not reach our 11 most significant hits, but were also strongly stratified: *TYR* (tyrosinase precursor) [20] at Chr11: 88.55–88.67 Mb reached a best p value of 10^{-39} and *SLC45A2 (MATP)* on Chr11: 33.98–34.02 Mb, implicated in hair color formation [16, 19, 21], reached $p = 3.1 \times 10^{-36}$.

Immunity Genes

These are well-known factors for selection in Europe, and GO categories related to immunity dominated our top GO findings together with those related to pigmentation (online suppl. table 2). Among our most significant regions (table 2) is the HLA region on Chr6, and the cluster of Toll-like receptors *TLR10, TLR1, TLR6* on Chr4: 38.38–38.59 Mb. Toll-like receptors play a role in pathogen recognition and activation of innate immunity in, for example, defense against tuberculosis. Both regions were stratified in the WTCCC study [6]. HLA was the top region identified in our GWA study paper as conferring risk for schizophrenia [9]. That result is, however, not caused by population stratification as it was derived by the Cochran-Mantel-Haenszel test, which accounts for population differences, and we did not reach similar results for the other top hits in the current study, such as *TLR10* or *LCT*.

Genes Involved in NAD Metabolism

This mechanism was proposed in the WTCCC study [6] which found the NAD synthetase 1 gene *(NADSYN1)* to be stratified. A role in prevention of pellagra was postulated. Pellagra is caused by a lack of niacin (Vitamin B3) and can result from nutritional deficit of niacin or tryptophan. It is possible that genetic variation in the genes involved in its metabolism can contribute to the development of the illness among populations with limited amounts of niacin or tryptophan in their diet. We also find this gene among our top hits, with the associated region overlapping that in the WTCCC study, thus strengthening the initial observation. Interestingly, another gene *(ACMSD)* that plays a role in NAD metabolism is also within our top regions: close to the lactase gene *(LCT)*. *ACMSD* is an intermediate in the de novo synthesis pathway of NAD from tryptophan. It is possible that the presence of *ACMSD* within this locus has increased the strength of the signal around *LCT*. The effect from NAD metabolism is further supported by the GO analysis (but only if we repeated the analysis with a relaxed cutoff of $p < 10^{-20}$, data not presented), as the category 'oxidoreductase activity, acting on the CH-OH group of donors, NAD or NADP as acceptor' reached 12th place on the list, with 8 genes in this category reaching that cutoff.

Lactase Gene

Lactase gene *(LCT)* is one of the best-known genes that have provided selective advantage around the world, because of the ability of farming communities to consume the milk of domesticated animals [22, 23]. It is also among our top hits.

Overlap with Previous Studies

As shown in table 2, five of the 11 top loci coincide with those identified in the WTCCC study [6] as stratified within the UK. The boundaries of the loci also overlap closely, indicating that the same factors have operated within the UK and within Europe. Three of the loci also coincide with those identified by McEvoy et al. [3]. That work was conducted on populations of Northern Euro-

pean descent and did not find the *LCT* gene to be stratified. The authors reasoned that there had been a similar strength of selection in these regions of Europe regarding milk tolerance. The strongest signal in that study was for immunity genes (with the HLA locus the top hit), and although pigmentation genes were represented with one locus *(OCA2* and *HERC2)*, these genes did not come up so prominently as in our study (presumably again because no populations from the South were included). In contrast, Bauchet et al. [7] suggested a list of 20 best European ancestry informative markers, which are not among our top hits. One possible explanation is the small sample size in that study (a total of 297 individuals).

New Findings

There were several stratified loci that included genes for which there is no obvious mechanism for a role in selection in these populations. Of those, the zinc finger genes *ZSCAN3, ZSCAN12* and *ZSCAN23* on Chr6 were most stratified. Most intriguingly, we also find SNPs lying within 20 kb from *FOXP2* among our top hits. This gene has been implicated in the development of language in humans [24]. There is evidence that it has been subject to positive selection when human and primate genomes are compared [21]; however, the two human-specific amino-acid changes are likely to have occurred more than 300,000 years ago [25]. It should be pointed out that the signal in *FOXP2* just reached our inclusion criteria and did not involve many SNPs. The true relevance of this finding will therefore have to be tested in other studies, preferably including more populations. Another finding involves three genes of the glutathione peroxidase system: glutathione peroxidase 5 and 6 *(GPX5, GPX6)* in the Chr6: 28.43–28.64 Mb locus, and glutathione synthetase *(GSS)* within the Chr20: 33.00–33.31 Mb locus. Glutathione is part of the hydrogen peroxide scavenging system and is important for the protection of cells from oxidative damage by free radicals. The effect may be coming instead from other genes in these loci, e.g. the zinc finger genes *ZSCAN3* and *ZSCAN23* on Chr6.

Our paper focuses on the top 11 loci and suggests plausible mechanisms for most of them. However, the total number of genome-wide significant SNPs is >50,000 and the top hits clustered in several GO categories. We cannot judge which ones are due to the effects of selection or to other mechanisms. We present a full list of genes with the best and median p values for SNPs within them (separately for the full sample and for controls only), so that others can make use of this information in future studies (online suppl. table 1).

References

1 Heath SC, Gut IG, Brennan P: Investigation of the fine structure of European populations with applications to disease association studies. Eur J Hum Genet 2008;16:1413–1429.

2 Lao O, Lu TT, Nothnagel M, Junge O, Freitag-Wolf S, Caliebe A, Balascakova M, Bertranpetit J, Bindoff L, Comas D: Correlation between genetic and geographic structure in Europe. Curr Biol 2008;18:1241–1248.

3 McEvoy BP, Montgomery GW, McRae AF, Ripatti S, Perola M, Spector TD, Cherkas L, Ahmadi KR, Boomsma D, et al: Geographical structure and differential natural selection among North European populations. Genome Res 2009;19:804–814.

4 Price AL, Helgason A, Palsson S, Stefansson H, St Clair D, Andreassen OA, Reich D, Kong A, Stefansson K: The impact of divergence time on the nature of population structure: an example from Iceland. PLoS Genet 2009; 5:e1000505.

5 Novembre J, Johnson T, Bryc K, Kutalik Z, Boyko AR, Auton A, Indap A, King KS, Bergmann S, et al: Genes mirror geography within Europe. Nature 2008;456:98–101.

6 Wellcome Trust Case Control Consortium: Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 2007;447:661–678.

7 Bauchet M, McEvoy B, Pearson LN, Quillen EE, Sarkisian T, Hovhannesyan K, Deka R, Bradley DG, Shriver MD: Measuring European population stratification with microarray genotype data. Am J Hum Genet 2007;80: 948–956.

8 International Schizophrenia Consortium: Rare chromosomal deletions and duplications increase risk of schizophrenia. Nature 2008;455:237–241.

9 International Schizophrenia Consortium: Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. Nature 2009;460:748–752.

10 Menozzi P, Piazza A, Cavalli-Sforza L: Synthetic maps of human gene frequencies in Europeans. Science 1978;201:786–792.

11 Wright S: The genetical structure of populations. Nature 1950;166:247–249.

12 Weir BS: Genetic Data Analysis II: Methods for Discrete Population Genetic Data. Sinauer Associate, 1996, p 167.

13 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.

14 Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ, Sham PC: PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 2007;81:559–575.

15 Holmans P, Green EK, Pahwa JS, Ferreira MA, Purcell SM, Sklar P, Wellcome Trust Case-Control Consortium, Owen MJ, O'Donovan MC, Craddock N: Gene ontology analysis of GWA study data sets provides insights into the biology of bipolar disorder. Am J Hum Genet 2009;85:13–24.

16 Sulem P, Gudbjartsson DF, Stacey SN, Helgason A, Rafnar T, Magnusson KP, Manolescu A, Karason A, Palsson A, Thorleifsson G, et al: Genetic determinants of hair, eye and skin pigmentation in Europeans. Nat Genet 2007;39:1443–1452.

17 Han J, Kraft P, Nan H, Guo Q, Chen C, Qureshi A, Hankinson SE, Hu FB, Duffy DL, Zhao ZZ, et al: A genome-wide association study identifies novel alleles associated with hair color and skin pigmentation. PLoS Genet 2008;16:4.
18 Gathany AH, Hartge P, Davis S, Cerhan JR, Severson RK, Cozen W, Rothman N, Chanock SJ, Wang SS: Relationship between interferon regulatory factor 4 genetic polymorphisms, measures of sun sensitivity and risk for non-Hodgkin lymphoma. Cancer Causes Control 2009;20:1291–1302.
19 Sundram U, Harvell JD, Rouse RV, Natkunam Y: Expression of the B-cell proliferation marker MUM1 by melanocytic lesions and comparison with S100, gp100 (HMB45), and MelanA. Mod Pathol 2003;16:802–810.
20 Stokowski RP, Pant PV, Dadd T, Fereday A, Hinds DA, Jarman C, Filsell W, Ginger RS, Green MR, van der Ouderaa FJ, Cox DR: A genomewide association study of skin pigmentation in a South Asian population. Am J Hum Genet 2007;81:1119–1132.
21 Branicki W, Brudnik U, Draus-Barini J, Kupiec T, Wojas-Pelc A: Association of the SLC45A2 gene with physiological human hair colour variation. J Hum Genet 2008;53:966–971.
22 Kelley JL, Swanson WL: Positive selection in the human genome: from genome scans to biological significance. Ann Rev Genomics Hum Genet 2008;9:143–160.
23 Enattah NS, Jensen TG, Nielsen M, Lewinski R, Kuokkanen M, Rasinpera H, El-Shanti H, Seo JK, Alifrangis M, Khalil IF, et al: Independent introduction of two lactase-persistence alleles into human populations reflects different history of adaptation to milk culture. Am J Hum Genet 2008;82:57–72.
24 Enard W, Przeworski M, Fisher SE, Lai CS, Wiebe V, Kitano T, Monaco AP, Pääbo S: Molecular evolution of FOXP2, a gene involved in speech and language. Nature 2002;418:869–872.
25 Krause J, Lalueza-Fox C, Orlando L, Enard W, Green R, Burbano H, Hublin J, Hänni C, Fortea J, de la Rasilla M, et al: The derived FOXP2 variant of modern humans was shared with Neandertals. Curr Biol 2007;17:1908–1912.