European Journal of Human Genetics (2005) 13, 798–806

www.nature.com/ejhg

ARTICLE

Genetic stratification of pathogen-response-related and other variants within a homogeneous Caucasian Irish population

Ciara Dolan[1], Aisling O'Halloran[1], Daniel G Bradley[2], David T Croke[1], Alun Evans[3], John K O'Brien[1], Patrick Dicker[1] and Denis C Shields*[,1]

[1]*Institute for Biopharmaceutical Sciences, Royal College of Surgeons in Ireland, 123 St Stephen's Green, Dublin 2, Ireland;* [2]*Department of Genetics, Smurfit Institute, Trinity College, Dublin 2, Ireland;* [3]*Department of Epidemiology and Public Health, Queen's University Belfast, UK*

Selection pressures from pathogens impact on the worldwide geographic distribution of polymorphisms in certain pathogen-response-associated genes. Such gene-specific effects could lead to confounding by geographic disease associations. We wished to determine if such constraints impinge on the genetic structure of a population of Irish patients and whether variants associated with responses to pathogens showed greater stratification. The counties of origin of each subject's grandparents were used as the geographic variable. F_{st}, proportional to the extent of population structure, was low (mean $F_{st} = 0.004$ across 25 SNPs, range 0.001–0.008) and it was not significantly higher for pathogen response SNPs ($F_{st} = 0.004$) than for other SNPs ($F_{st} = 0.003$, $P = 0.21$). Correspondence analysis revealed weak trends primarily in approximately northeast to southwest and secondarily in northwest to southeast directions. One-dimensional spatial autocorrelation analysis revealed a weak (Moran's *I* autocorrelation of −0.10) tendency for SNP frequencies to diverge with greater distance. Two-dimensional autocorrelation indicated a northeast to southwest gradient that was similar for both the pathogen response and other SNPs. The southeastern county, Wexford, showed a distinctive pattern, perhaps consistent with Anglo-Norman settlements. In conclusion, these results indicate that pathogen response SNPs do not exhibit significantly more population structure than other SNPs within this Caucasian population. This suggests that the specific population structure of particular genes may not typically be a cause of strong confounding in genetic studies where population structure is controlled.
European Journal of Human Genetics (2005) **13,** 798–806. doi:10.1038/sj.ejhg.5201382
Published online 23 March 2005

Keywords: population structure; F_{st}; correspondence analysis; selection pressures; cardiovascular disease

Introduction

Selection pressures from pathogens are known to impact on the worldwide geographic distribution of genetic variants in certain pathogen-response-associated genes. Such gene-specific effects could potentially lead to confounding by geographic disease associations. We wished to determine if such constraints impact on the genetic structure of more homogeneous populations.

The Irish population lies at the extreme of Europe, with some evidence that it lies at the end of cline of gene frequencies across Europe.[1] There is also evidence that clines of gene frequencies exist across Ireland. Anthropological studies in the 1950s showed that there was clear

*Correspondence: Dr DC Shields, Department of Clinical Pharmacology, Royal College of Surgeons in Ireland, 123, St Stephens Green, Dublin 2, Ireland. Tel: +353 1 4022381; Fax: +353 1 4022388;
E-mail: dshields@rcsi.ie
Received 20 July 2004; revised 23 December 2004; accepted 12 January 2005

evidence of a gradient of genetic variation across Ireland, for example the west of Ireland has a higher freckle density than in the east.[2] Early serological studies of the ABO and Rh blood groups showed that the west of Ireland has the highest frequency of the blood group O in Europe (>75%), with the frequency of group A being highest in the east of the country (>30%). This eastern region corresponds well with the area of greatest settlement by the Anglo-Normans, in which population there is a high frequency of the A group.[3] The frequency of Rh negative was found to be high in the east and lower in the west, ranging from 12 to 20%. However, a more recent study of genetic structure of the British Isles concluded that there was very little spatial genetic structure when studying several genetic systems including the ABO blood groups.[4] A study of Y chromosomal variation also showed an east-west pattern, with haplogroup 1 (R1b3)[5] at a frequency of 98.3% in the western-most Irish province Connaught, which declines easterly to 73.3% in Leinster.[6] Therefore, it has been shown that for some systems there does appear to be population structure in Ireland, suggesting the need to test for such structure in the investigation of any candidate gene investigations of disease in Irish populations. We set out to assess the geographic variation in Irish autosomal polymorphisms and whether such variation was greater for pathogen response genes.

One feature of Irish population structure has been the collapse of the population during the Great Irish Famine (1847–1851). Gene frequencies are more likely to be dominated by small bottlenecks during population foundation, rather than by the halving over 60 years caused by death and emigration. Nevertheless, it is of interest to investigate whether gene frequencies might be associated with the distribution of patterns of mortality during this extreme event. Some genetic variants have been shown to influence susceptibility to certain diseases that were prevalent during the Great Irish famine, such as tuberculosis (TB).[7] We therefore contrasted the allele frequencies of SNPs associated with TB response[8–10] with documented mortality attributed to TB both during the Great Famine[7] and more reliable mortality figures at a later time point in the early 20th century,[11–13] to investigate whether genetic background had a detectable influence on historic TB mortality.

Materials and methods

Subjects

There were 1600 patients in all, 1163 of whom participated in a study of the genetics of acute coronary syndromes (ACS)[14] and have either suffered from myocardial infarction, unstable angina or stable angina, and 437 of whom are part of a study on the genetics of early-onset ACS who have suffered from either myocardial infarction or unstable angina before the age of 55 (males) or 60 (females). Subjects have provided information about the Irish county of origin of each of their grandparents. These data were used as the geographic variable, so any grandparents who came from outside Ireland or whose Irish county of origin was unknown were excluded. For example, if the origin of only two of the four possible grandparents was known, these data were used in the analysis. This reduced the data set to 5645 grandparents (Table 1).

Geographic information

Each of the 32 counties of Ireland, from both Northern Ireland and the Republic of Ireland, were represented in the data set. The latitude and longitude were obtained for the most central location of each of the counties (Table 1). Counties with low representation (less than 50 grandparents) were merged, creating a final data set of 23 sample locations. While this ascertainment procedure does not provide an unbiased estimate of allele frequencies in Ireland (since candidate genes may differ in frequency in this disease population compared to the general population), cardiovascular disease is at a high incidence in all counties, and therefore the differences between grandparental counties of origin are likely to be highly representative of the differences seen in the population as a whole.

SNPs and genotyping

Each patient has been genotyped for 25 different SNPs in genes that can be categorised into two groups: those selected as candidate genes that may modulate platelet-mediated thrombosis,[15–21] and those where there is documented evidence in the literature that the SNP variant appears to modulate significantly responses to pathogens in population studies[8–10,22–29] (Table 2).

Genotyping was carried out using the Amplifluor™ method by K Biosciences (www.kbioscience.co.uk). As part of quality control, 87 duplicate samples were genotyped. The majority of the SNP assays had a 0% error rate. The HLA-A SNP had an error rate of 1.1%, the SULT1A1 SNP had an error rate of 2.3% and the TLR4 exon 3 T399I SNP has an error rate of 1.1%. This gives an overall error rate of 0.2%. Tests for Hardy–Weinberg equilibrium (HWE) were also carried out using the G statistic.

Statistical analysis

F_{st} F_{st} is a standardised measure of the genetic variance among populations.[30] It is a measure of heterozygote deficiency, reflecting the probability that two alleles drawn at random are identical relative to the combined population. F_{st} is proportional to the variance in allele frequencies between populations. It was calculated for each SNP over all sample locations, and a multilocus F_{st} was calculated for each sample location. The pathogen response SNPs as a group were compared to the thrombosis SNPs as a group, and the differences between the groups were assessed using

Table 1 Description of each of the 23 Irish sample locations, indicating latitude, longitude, number of individuals and F_{st} values

Location	*Longitude*	*Latitude*	*No. of grand parents*	*% of total*	*Ave. overall F_{st}*	*Ave. thrombosis F_{st}*	*Ave. pathogen F_{st}*
Antrim	6.17	54.83	573	10.2	0.0037	0.0035	0.0038
CL	7.08	52.83	106	1.9	0.0049	0.0030	0.0062
Clare	9.0	52.83	100	1.8	0.0059	0.0048	0.0067
Cork	8.68	51.92	712	12.6	0.0042	0.0028	0.0052
Derry	6.75	54.83	95	1.7	0.0064	0.0030	0.0089
Donegal	7.92	54.85	94	1.7	0.0052	0.0058	0.0047
Down	5.92	54.33	212	3.8	0.0069	0.0092	0.0051
Dublin	6.18	53.33	1285	22.8	0.0036	0.0033	0.0038
Galway	8.83	53.26	282	5.0	0.0036	0.0026	0.0043
Kerry	9.67	52.08	187	3.3	0.0052	0.0033	0.0067
Kildare	6.76	53.17	101	1.8	0.0056	0.0032	0.0072
LAM	6.67	54.17	223	4.0	0.0048	0.0024	0.0066
LFCL	7.75	54.08	135	2.4	0.0056	0.0074	0.0042
Limerick	8.75	52.5	293	5.2	0.0049	0.0047	0.0050
Meath	6.67	53.68	97	1.7	0.0067	0.0102	0.0040
MRS	8.5	53.75	154	2.7	0.0048	0.0032	0.0060
Offaly	7.68	53.25	112	2.0	0.0080	0.0022	0.0123
Tipperary	7.93	52.59	207	3.7	0.0062	0.0071	0.0056
Tyrone	7.17	54.58	137	2.4	0.0057	0.0049	0.0063
Westmeath	7.43	53.5	136	2.4	0.0059	0.0077	0.0046
Wexford	6.51	52.42	122	2.2	0.0090	0.0085	0.0093
Wicklow	6.25	53.0	123	2.2	0.0061	0.0031	0.0083
WK	7.33	52.25	159	2.8	0.0053	0.0038	0.0065
Total			5645	100.0			

CL: Carlow, Laois; LAM: Louth, Armagh, Monaghan; LFCL: Leitrim, Fermanagh, Cavan, Longford; MRS: Mayo, Roscommon, Sligo; WK: Waterford, Kilkenny.

the Kolmogorov–Smirnoff test. Weir and Cockerham F statistics were estimated using FSTAT software[31,32] according to the following equation:

$$\theta = \frac{\delta_a^2}{\delta_a^2 + \delta_b^2 + \delta_w^2}$$

where δ_a is the among-sample variance component, δ_b is the between-individual within-sample variance and δ_w is the within-individual component. We used Mantel test statistics[33] to evaluate the correlations between the multilocus F_{st} values of all SNPs with the corresponding distance in kilometres between each of the sample locations.

Correspondence analysis Correspondence analysis[34] is a statistical ordination method that collapses the major source of correlation variation within a group of variables into an axis (CA axis 1), and similarly defines subsequent axes explaining any residual variation. Correspondence analysis was performed using ADE4 software for the R statistical package.[35] Minor allele counts for each SNP in each sample location were used as the input.

Spatial autocorrelation The spatial autocorrelation coefficient[36] Moran's *I* was used to test whether the allele frequencies are independent of the allele frequencies at a neighboring location within specified distance classes. The significance of SA correlograms corrected for multiple tests.[37] Spatial autocorrelation analysis was performed using PASSAGE software.[38] For 1-D correlograms, both statistics are plotted against distance classes. 2-D (Windrose) correlograms take into account compass bearings as well as distance and are used to see if the data are anisotropic. The five distance classes for the 1-D analysis had upper limits of 84, 127, 167, 219 and 385 km. The five distance classes for the 2-D analysis had upper limits of 40, 85, 160, 265 and 400 km.

TB data

Three of the pathogen response SNPs have been shown in other populations to modulate susceptibility to TB. These are the SLC11A1[10,39] SNP, the VDR FokI[8,40] SNP and the IL12RB1[9] SNP. Death rates from TB during the famine year of 1847 on a per county basis obtained from the census of 1871[7] were used to look for any correlations between such death rates and frequency of the susceptibility alleles of the SNPs. As such data have been shown to be unreliable,[12,13] they are only analysed as an exploratory exercise. We also investigated the more reliable death rates on a county basis obtained in the Irish Registrar Generals' decennial summaries for the years 1901–1910.[11] This report detailed death rates from several different forms of TB, as well as death rates from all forms of TB per county. Correlations were sought between the allele frequencies of each of the three variants and the death rates from TB per county per 100 000 of population in the years 1901–1910. Stepwise multiple regression analysis[41] considered whether the TB

Table 2 Details of the 25 SNPs

Gene symbol	*Gene name*	*Function*	*SNP*[a] *(minor allele underlined)*	*Ave. minor allele freq.*	*Min. minor allele freq.*	*Max. minor allele freq.*	F_{st}	*HWE test (P-value)*	*HWE over all locations 23 d.f. (P-value)*
ABO	ABO blood group	Path response	G235S, G/A 021640 (Seattle SNP)	0.07	0.02	0.14	0.005	0.001[b]	0.001[b]
CCR2	Chemokine receptor 2	Path response	V64I, G/A rs1799864	0.06	0.02	0.08	0.001	0.571	0.999
CCR5	Chemokine receptor 5	Path response	−503, G/A 601373.0006 (OMIM)	0.45	0.36	0.57	0.007	0.078	0.756
CX3CR1	Chemokine (C-X3-C motif) receptor 1	Path response	T280M, C/T rs3732378	0.18	0.12	0.25	0.003	0.208	0.078
CXCL12	Chemokine (C-X-C motif) ligand 12	Path response	3′ 801, G/A 600835.0001 (OMIM)	0.21	0.13	0.30	0.008	0.746	0.497
FCGR2a	Fc fragment of IgG, low affinity IIa, receptor for (CD32)	Path response	R165H, G/A rs1801274	0.41	0.34	0.53	0.002	0.178	0.426
FUT2	Fucosyltransferase 2	Path response	W157X, G/A rs601338	0.40	0.25	0.50	0.007	0.081	0.623
HLA-A	Major histocompatibility complex, class I, A	Path response	A269V, C/T rs1272176	0.06	0.02	0.10	0.007	<0.001[b]	0.067
IL12 RB1	Interleukin 12 receptor, beta 1	Path response	Q214R, A/G rs11575934	0.36	0.22	0.48	0.005	0.183	0.934
IL13	Interleukin 13	Path response	R144Q, G/A rs20541	0.18	0.12	0.26	0.005	0.836	0.549
SLC11A1	Solute carrier family 11, member 1	Path response	Intron 4, G/C rs3731865	0.29	0.21	0.37	0.006	0.765	0.252
TLR4	Toll-like receptor 4	Path response	D299G, A/G rs4986790	0.08	0.05	0.11	0.001	0.321	0.985
TLR4	Toll-like receptor 4	Path response	T399I, C/T rs4986791	0.08	0.05	0.12	0.003	0.182	0.978
VDR	Vitamin D (1,25-dihydroxyvitamin D3) receptor	Path response	Intron 8, T/G VDDRAPA1 (GeneCanvas)	0.47	0.41	0.52	0.003	0.993	0.427
VDR	Vitamin D (1,25-dihydroxyvitamin D3) receptor	Path response	T1M fok1, C/T rs2228570	0.38	0.33	0.53	0.000	0.891	0.983
ALAD	Aminolevulinate, delta-, dehydratase	Thrombosis	K68N, G/C rs1800435	0.08	0.03	0.12	0.003	0.307	0.953
CD109	CD109 antigen	Thrombosis	Y703S, A/C rs10455097	0.46	0.39	0.56	0.003	0.553	0.578
Col1A1	Collagen, type I, alpha 1	Thrombosis	Intron 1, G/T 120150.0051 (OMIM)	0.20	0.14	0.31	0.003	0.521	0.779
Col3A1	Collagen, type III, alpha 1	Thrombosis	A531T, G/A 120180.0007 (OMIM)	0.23	0.17	0.33	0.003	0.504	0.694
GalNT4	UDP-*N*-acetyl-alpha-D-galactosamine: polypeptide *N*-acetylgalactosaminyltransferase 4	Thrombosis	V506I, G/A rs2230283	0.36	0.27	0.45	0.003	0.306	0.565
GP6	Glycoprotein VI	Thrombosis	S219P, T/C rs1613662	0.15	0.10	0.22	0.003	0.370	0.952

Table 2 (*Continued*)

Gene symbol	Gene name	Function	SNP[a] (minor allele underlined)	Ave. minor allele freq.	Min. minor allele freq.	Max. minor allele freq.	F_{st}	HWE test (*P*-value)	HWE over all locations 23 d.f. (*P*-value)
GSTO1	Glutathione *S*-transferase omega 1	Thrombosis	A140D, C/A rs4925	0.30	0.23	0.40	0.004	0.492	0.908
RABggta	Rab geranylgeranyltransferase, alpha subunit	Thrombosis	T420A, A/G rs729421	0.37	0.27	0.43	0.003	0.668	0.669
SULT1a1	Sulphotransferase family, cytosolic, 1A, phenol-preferring, member 1	Thrombosis	R213H, G/A rs9282861	0.34	0.25	0.42	0.006	0.452	0.774
THBS1	Thrombospondin 1	Thrombosis	N700S, A/G rs2228262	0.12	0.04	0.17	0.002	0.511	0.837

[a]rs number = http://www.ncbi.nlm.nih.gov/SNP/; Seattle SNP = http://pga.gs.washington.edu/data/abo/abobg.csnps.txt; OMIM = http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db = OMIM; Gene Canvas = http://genecanvas.idf.inserm.fr/.
[b]SNPs showing significant Hardy–Weinberg disequilibrium ($P < 0.05$).

mortality rates per county correlated with allele frequency of the three SNPs, considering latitude, longitude, NW-SE and SW-NE directional trends and correspondence analysis axis 1 as covariates (successively dropping terms from the model, with $P = 0.10$ as the cutoff for retaining terms).

Results

Hardy–Weinberg equilibrium

The two SNPs that appear in significant Hardy–Weinberg disequilibrium are the ABO SNP and the HLA-A SNP. To determine if this could be accounted for by regional variation among sample locations, a local HWE test was summed over the 23 locations (23 d.f.). The HLA test was no longer significant, implying that the variant is at equilibrium once regional variation was taken into account. The ABO variant was still highly significant: this was not simply a consequence of small expectations for some calculations, since it remained significant ($P = 0.001$) when calculated using Yates' correction. The genotype frequencies are 85.2% GG and 14.8% GA in the overall population. There were no subjects who were homozygote for the minor allele A. It is possible that there is further unmeasured substructure for this variant in the Irish population, or that some other factor is impacting on HWE. Independent genotyping by a different method would be required to confirm this.

F_{st} values

The average F_{st} value for the 25 SNPs was 0.004, which is very low, given that the average F_{st} value calculated using more widely sampled populations is greater than 0.120.[42,43] This indicates little variation among the SNPs among the 23 sample locations. However, pathogen response SNPs tend to have higher F_{st} values (mean 0.004, SE ± 0.0007) than the thrombosis SNPs (mean 0.003, SE ± 0.0003), but this difference is not significant ($P = 0.210$).

Multilocus F_{st} values across pairs of sample locations for the 25 SNPs over the 23 sample locations were calculated, and the means for each county are shown in Table 2. It is noticeable that the counties containing larger urban settlements (Antrim, Dublin, Galway, Limerick, Cork) have a low mean F_{st} (0.004–0.005), consistent with a homogenisation of gene frequencies in urban areas. Certain counties (Offaly, Wexford) have a high mean F_{st} (0.008–0.009). This may relate to substantial and persistent patterns of settlement in the Wexford area after the Anglo-Norman invasions of the 12th century,[44] and to the settlement of Offaly from England and Scotland by 17th century plantations[44] and/or a low rate of migration between these counties and other areas. There were very small and insignificant correlations between the multilocus F_{st} values between each sample location and the distance in kilometres between each sample location for all

SNPs ($r=0.05$, $P=0.279$), for the pathogen response SNPs ($r=0.068$, $P=0.241$) and for the thrombosis SNPs ($r=-0.014$, $P=0.504$).

Correspondence analysis

Correspondence analysis displays the relationship between the 23 sample locations and between the 25 SNPs simultaneously. For ease of visualisation, the SNPs have been displayed separately (Figure 1). The first axis accounts for 16% of the variance and the second axis accounts for 12% of the variance. Axes 3, 4 and 5 accounted for 11, 10 and 9% of the variance, respectively.

We plotted the values of the first, second and third axes for the 23 sample locations on a map of Ireland. These maps show some general trends (Figure 2a–c), with axis 1 showing an NE-SW trend and axis 2 showing an NW-SE trend, although with notable exceptions. The first axis indicates that the South-Western County, Kerry, may represent the most distinctive extreme with the eastern seaboard (Wexford, Dublin and Antrim) lying at the other extreme. Wexford, consistent with this county having a slightly more distinct genetic make-up, showed sharp contouring with neighbours in each plot and dominated the third axis, consistent with its relatively high F_{st} value (Table 2).

On both the first and second axes of Figure 1, there are three pathogen response SNPs that stand apart from the rest. They are the CXCL12 SNP, the HLA-A SNP and the ABO SNP, which are all at quite low minor allele frequencies in the population.

Spatial autocorrelation analysis

Figure 3 displays the 1-D correlogram for the average of the 25 SNPs, for the 15 pathogen response SNPs and for the 10 thrombosis SNPs. The low Moran's *I* values (largest $I=0.05$, smallest $I=-0.1$ on a scale of –1 to 1) indicate low levels of spatial autocorrelation, either positively or negatively. This indicates lower genetic population structure than seen in a study of the British Isles, which itself only showed minor spatial genetic structure.[4]

Pathogen response SNPs showed a slightly stronger correlation between nearby regions, although the patterns observed for each group of SNPs were largely the same (Figure 3).

The main pattern to emerge across all SNPs from the 2-D autocorrelation was largely the same in both the pathogen

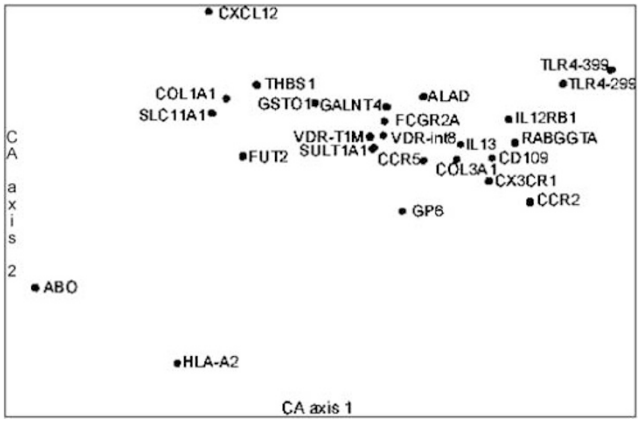

Figure 1 Correspondence analysis of the 25 SNPs. Axis 1 is proportional to darkness of shading in Figure 2a and axis 2 is proportional to the darkness of shading in Figure 2b.

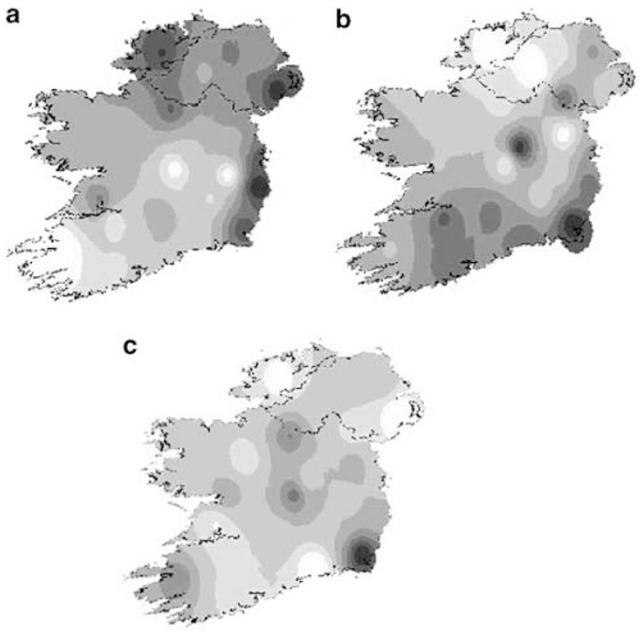

Figure 2 Map of correspondence analysis. (**a**) Axis 1 values, showing an approximate NE-SW gradient. (**b**) Axis 2 values, showing an approximate NW-SE gradient. (**c**) Axis 3 values.

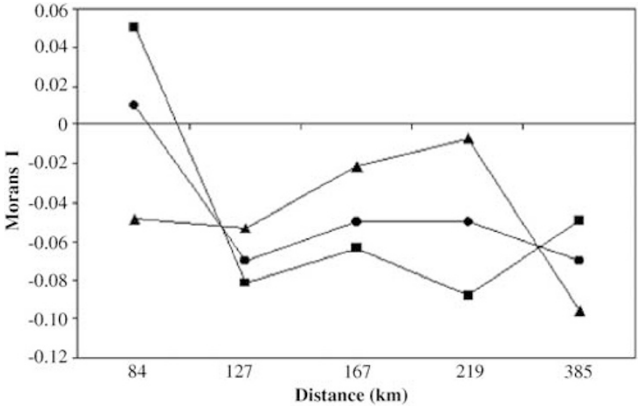

Figure 3 1-D spatial autocorrelograms using the Moran's *I* statistic. Distances indicated are upper limits of the distance classes. —●— Average of 26 SNPs; —■— average of 15 pathogen response SNPs; —▲— average of 10 thrombosis SNPs.

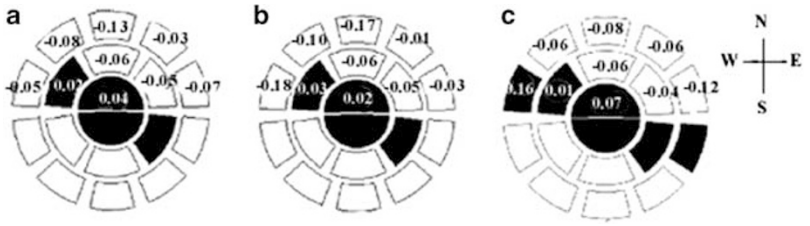

Figure 4 (**a**) Average 2-D spatial autocorrelogram of the 25 SNPs. (**b**) Average 2-D spatial autocorrelogram of the 15 pathogen response SNPs. (**c**) Average 2-D spatial autocorrelogram of the 10 thrombosis SNPs. In each segment, the average value for *I* is shown. Inner annulus represents 0–43 km, second annulus represents 43–172 km and third annulus represents 172–387 km. Black segments are positively spatially autocorrelated ($0<I<1$) and white segments are negatively spatially autocorrelated ($-1<I<0$).

response and other SNPs (Figure 4a–c). The general pattern is positive spatial autocorrelation in a northwest to southeast direction and negative spatial autocorrelation elsewhere, indicating a general northeast to southwest gradient of allele frequencies across Ireland. This is consistent with the northeast to southwest trend seen in the first axis of the correspondence analysis. While there are some differences, overall there is no strong indication from this analysis of a radically different pattern in pathogen response variants compared to the others.

Test for correlation between allele frequencies and TB deaths

Using linear regression analysis, there is a significant association between the east-west gradient (longitude) and TB deaths from 1901–1910 ($P=0.017$) and also for TB deaths during the famine year of 1847 ($P<0.001$). No statistically significant correlations were found between the TB deaths from 1901–1910 and any of the allele frequencies per county level ($\alpha=0.05$). There is a suggestive correlation between the SLC11A1 allele ($r=-0.29$, $P=0.18$), the VDR FokI allele ($r=0.07$, $P=0.75$) and the IL12Rb1 allele ($r=-0.23$, $P=0.29$) with TB deaths. There is a significant correlation between TB deaths during the famine year of 1847 and the SLC11A1 C allele ($r=0.42$, $P=0.05$), although this is not significant for either of the other two SNPs ($P=0.11$ for VDR FokI and $P=0.59$ for IL2Rb1). Stepwise regression analysis considered whether TB deaths from 1901–1910 were influenced by the frequencies of the minor allele for any of the three SNPs, also considering correspondence analysis axis 1 and directional trends as dependent variables in the model. This found no significant predictors for the TB death rates from 1901–1910, although correspondence analysis 1 ($P=0.018$) and correspondence analysis 3 ($P=0.001$) are strong predictors of TB deaths during the famine year of 1847 (Figure 5b), which may largely reflect the association of certain genetic backgrounds with densely populated eastern areas.

The analysis considered up to four counties for each person, which are clearly not independent. For the data exploration aspects of this study, this does not introduce any particular bias. For the hypothesis testing aspects, an appropriate statistical correction would be required. However, since all the *P*-values were not significant, no such adjustments are necessary.

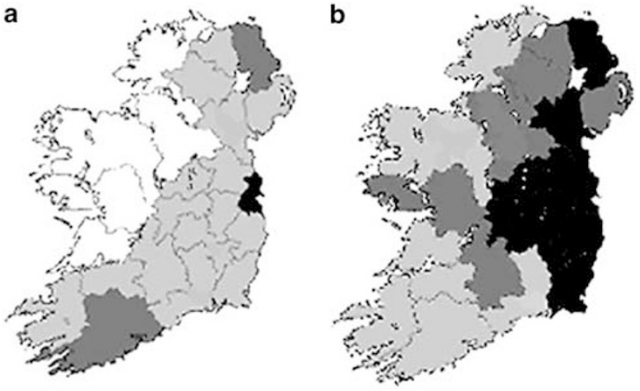

Figure 5 (**a**) Death rates from TB from 1901–1910 from the registrar generals' report. (**b**) Death rates from TB during the famine year of 1847. Black: 400–500 deaths per 100 000; dark grey: 300–400 deaths per 100 000; light grey: 200–300 deaths per 100 000; white: <100 deaths per 100 000.

Discussion

This analysis shows that pathogen response variants, which we anticipated might show a stronger population structure than other variants, show a trend towards slightly stronger structure, but that this trend is very weak and not significant. Geographic trends appear very similar for the two groups of SNPs. The overall pattern of genetic variation is suggestive of a trend distinguishing the SW and midlands from the north and east, but this is not a dominating pattern, since a secondary weaker independent trend from NW to SE is indicated from the correspondence analysis. Much larger sample sizes of subjects will be required to define such trends more accurately. One county, Wexford at the southeastern tip, has a distinct pattern, perhaps

consistent with its history as the earliest entry point (1169 AD) for the substantial and persistent migrations from Britain, which are a major characteristic of Irish historical demography.[45] Some indication of the durability of the early Wexford settlement can be found in the persistence of a distinctive early English dialect from the time of the Anglo-Norman invasion until the middle of the last century.[46]

How informative are the grandparental origins in relation to genetic structure of Ireland? The industrial revolution had a lesser impact in Ireland than in the UK, with a predominantly rural agricultural economy well into the mid-20th Century. Census data indicate that of the 8.1 million strong population in 1841, only some 400 000 were living outside the counties in which they were born, with most migration to the urban centres of Dublin and Belfast.[47] Therefore, the choice of grandparental counties of origin as an estimate of allelic origin should capture a significant proportion of the genetic structure in the Irish population, outside of the main urban regions.

A previous study has shown that the intron 4 G/C variant in the SLC11A1 gene is associated with TB susceptibility. While we found that this SNP was associated with TB mortality in the year 1847 during the Great Irish Famine, this association is not significant after correction for multiple testing. It may be that TB deaths in the Famine and in later historic periods investigated were dominated by nongenetic factors.

Our study indicates that selection pressures on pathogen response modulating variants have not created a markedly different population structure within a homogeneous Caucasian population. A separate issue is whether the allele frequencies themselves have been influenced by selection. Our analysis cannot answer this question, except to note that the impact of such selection pressures, if they exist, has not been confined to a particular geographic area.

What are the implications for problems of confounding genetic associations with disease through population structure? In general, it appears that selection may play quite a weak role in altering allele frequencies of common polymorphisms within a homogeneous population. The Irish may represent a good population to study weak genetic risks conferred by genes involved in disease responses with low danger of confounding through population structure. Although the population structure is small, it is detectable, and this supports proposals that such structure should be corrected for in analysis of weak genetic associations with disease.[48] Even if some random genes showed marginally less structure than those genes such as HLA-A2, which show slightly greater structure, the trend will be in the same direction, and a reasonable evaluation of the data will allow for confounding caused by genetic substructure.

Acknowledgements

We thank Professor Cormac O'Grada, University College Dublin, for helpful advice on the history of the Irish population. This work was supported by the Health Research Board, the Northern Ireland R & D office and the Programme for Research in Higher Level Institutions administered by the Higher Education Authority in Ireland, and the European Union QLG2 CT-2002-01254 GenomeUTwin project.

References

1 Rosser ZH, Zerjal T, Hurles ME *et al*: Y-chromosomal diversity in Europe is clinal and influenced primarily by geography, rather than by language. *Am J Hum Genet* 2000; **67**: 1526–1543.
2 Hooton E, Dupertuis C: *The Physical Anthropology of Ireland: Papers of the Peasbody Museum of Archeology and Ethnology*, Cambridge, MA 1955.
3 Dawson G: The frequencies of the ABO and Rh(D) blood groups in Ireland from a sample of 1 in 18 of the population. *Ann Hum Genet* 1964; **28**: 49.
4 Falsetti AB, Sokal RR: Genetic structure of human populations in the British Isles. *Ann Hum Biol* 1993; **20**: 215–229.
5 Ngo KY, Vergnaud G, Johnsson C, Lucotte G, Weissenbach J: A DNA probe detecting multiple haplotypes of the human Y chromosome. *Am J Hum Genet* 1986; **38**: 407–418.
6 Hill EW, Jobling MA, Bradley DG: Y-chromosome variation and Irish origins. *Nature* 2000; **404**: 351–352.
7 Kennedy L, Ell P, Crawford E, Clarkson L: *Mapping the Great Irish Famine*. Dublin: Four Courts Press Ltd, 1999.
8 Wilkinson RJ, Llewelyn M, Toossi Z *et al*: Influence of vitamin D deficiency and vitamin D receptor polymorphisms on tuberculosis among Gujarati Asians in west London: a case–control study. *Lancet* 2000; **355**: 618–621.
9 Akahoshi M, Nakashima H, Miyake K *et al*: Influence of interleukin-12 receptor beta1 polymorphisms on tuberculosis. *Hum Genet* 2003; **112**: 237–243.
10 Bellamy R, Ruwende C, Corrah T, McAdam KP, Whittle HC, Hill AV: Variations in the NRAMP1 gene and susceptibility to tuberculosis in West Africans. *N Engl J Med* 1998; **338**: 640–644.
11 Registrar G: *Supplement to the 47th report of the registrar general in Ireland of births, marriages and deaths in Ireland containing decennial summaries for the years 1901–1910*, House of Commons parliamentary papers, 1914.
12 Mokyr J, O'Grada C: *Famine disease and famine mortality: lessons from the Irish experience, 1845–1850: Famine Demography*, Oxford, 2002.
13 Mokyr J, O'Grada C: What do people die of during famines? The Great Irish Famine in comparative perspective. *Eur Rev Econ History* 2002; **6**: 339–364.
14 Muckian C, Fitzgerald A, O'Neill A, O'Byrne A, Fitzgerald DJ, Shields DC: Genetic variability in the extracellular matrix as a determinant of cardiovascular risk: association of type III collagen COL3A1 polymorphisms with coronary artery disease. *Blood* 2002; **100**: 1220–1223.
15 Zafarullah K, Kleinert C, Tromp G *et al*: G to A polymorphism in exon 31 of the COL3A1 gene. *Nucleic Acids Res* 1990; **18**: 6180.
16 Topol EJ, McCarthy J, Gabriel S *et al*: Single nucleotide polymorphisms in multiple novel thrombospondin genes may be associated with familial premature myocardial infarction. *Circulation* 2001; **104**: 2641–2644.
17 Schuh AC, Watkins NA, Nguyen Q *et al*: A tyrosine703serine polymorphism of CD109 defines the Gov platelet alloantigens. *Blood* 2002; **99**: 1692–1698.
18 Engelke CE, Meinl W, Boeing H, Glatt H: Association between functional genetic polymorphisms of human sulfotransferases 1A1 and 1A2. *Pharmacogenetics* 2000; **10**: 163–169.
19 Wetmur JG, Kaya AH, Plewinska M, Desnick RJ: Molecular characterization of the human delta-aminolevulinate

dehydratase 2 (ALAD2) allele: implications for molecular screening of individuals for genetic susceptibility to lead poisoning. *Am J Hum Genet* 1991; **49**: 757–763.
20 Grant SF, Reid DM, Blake G, Herd R, Fogelman I, Ralston SH: Reduced bone density and osteoporosis associated with a polymorphic Sp1 binding site in the collagen type I alpha 1 gene. *Nat Genet* 1996; **14**: 203–205.
21 Whitbread AK, Tetlow N, Eyre HJ, Sutherland GR, Board PG: Characterization of the human Omega class glutathione transferase genes and associated polymorphisms. *Pharmacogenetics* 2003; **13**: 131–144.
22 Ali S, Niang MA, N'Doye I *et al*: Secretor polymorphism and human immunodeficiency virus infection in Senegalese women. *J Infect Dis* 2000; **181**: 737–739.
23 McDermott DH, Fong AM, Yang Q *et al*: Chemokine receptor mutant CX3CR1-M280 has impaired adhesive function and correlates with protection from cardiovascular disease in humans. *J Clin Invest* 2003; **111**: 1241–1250.
24 Smith MW, Dean M, Carrington M *et al*: Contrasting genetic influence of CCR2 and CCR5 variants on HIV-1 infection and disease progression. Hemophilia Growth and Development Study (HGDS), Multicenter AIDS Cohort Study (MACS), Multicenter Hemophilia Cohort Study (MHCS), San Francisco City Cohort (SFCC), ALIVE Study. *Science* 1997; **277**: 959–965.
25 Yee AM, Phan HM, Zuniga R, Salmon JE, Musher DM: Association between FcgammaRIIa-R131 allotype and bacteremic pneumococcal pneumonia. *Clin Infect Dis* 2000; **30**: 25–28.
26 Yamamoto F, Clausen H, White T, Marken J, Hakomori S: Molecular genetic basis of the histo-blood group ABO system. *Nature* 1990; **345**: 229–233.
27 Winkler C, Modi W, Smith MW *et al*: Genetic restriction of AIDS pathogenesis by an SDF-1 chemokine gene variant. ALIVE Study, Hemophilia Growth and Development Study (HGDS), Multicenter AIDS Cohort Study (MACS), Multicenter Hemophilia Cohort Study (MHCS), San Francisco City Cohort (SFCC). *Science* 1998; **279**: 389–393.
28 Graves PE, Kabesch M, Halonen M *et al*: A cluster of seven tightly linked polymorphisms in the IL-13 gene is associated with total serum IgE levels in three populations of white children. *J Allergy Clin Immunol* 2000; **105**: 506–513.
29 Arbour NC, Lorenz E, Schutte BC *et al*: TLR4 mutations are associated with endotoxin hyporesponsiveness in humans. *Nat Genet* 2000; **25**: 187–191.
30 Wright S: The genetical structure of populations. *Ann Eugen* 1951; **15**: 323–354.
31 Weir B, Cockerham C: Estimating F-statistics for the analysis of population structure. *Evolution* 1984; **38**: 1358–1370.
32 Goudet J. FSTAT, a program to estimate and test gene diversities and fixation indices, 2001.
33 Smouse P, Long J, Sokal RR: Multiple regression and correlation extensions of the Mantel test of matrix correspondence. *Syst Zool* 1986; **35**: 627–632.
34 Greenacre M: *Theory and Applications of Correspondence Analysis*. London: Academic Press, 1984.
35 Thioulouse JCD, Dolédec S, Olivier JM: ADE-4: a multivariate analysis and graphical display software. *Stat Comput* 1997; **7**: 75–83.
36 Cliff A, Ord J: *Spatial Autocorrelation*. London: Pion, 1973.
37 Oden NL: Assessing the significance of a spatial correlogram. *Geograph anal* 1984; **16**: 1–16.
38 Rosenberg MS: *PASSAGE. Pattern Analysis, Spatial Statistics and Geographic Exegesis. Version 1.0*. Department of Biology, Arizona State University: Tempe, AZ, 2001.
39 Cervino AC, Lakiss S, Sow O, Hill AV: Allelic association between the NRAMP1 gene and susceptibility to tuberculosis in Guinea-Conakry. *Ann Hum Genet* 2000; **64**: 507–512.
40 Bellamy R, Ruwende C, Corrah T *et al*: Tuberculosis and chronic hepatitis B virus infection in Africans and variation in the vitamin D receptor gene. *J Infect Dis* 1999; **179**: 721–724.
41 Stata C: *Stata Statistical Software: Release 8.0*. College Stationm, TX: Stata Corp. LP, 2003.
42 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.
43 Bowcock AM, Kidd JR, Mountain JL *et al*: Drift, admixture, and selection in human evolution: a study with DNA polymorphisms. *Proc Natl Acad Sci USA* 1991; **88**: 839–843.
44 Barry T: *A History of Settlement in Ireland*. London: Routledge, 2000.
45 Smith B: The conquest of Ireland; in Duffy S (ed): *Atlas of Irish History*. Dublin: Gill and MacMillan, 2000, pp 32–49.
46 Dolan TP, O'Muirithe D: *The Dialect of Forth and Bargy, Co. Wexford, Ireland*. Dublin: Four Courts Press, 1996.
47 Census of Ireland for the year of 1841, 1841; 446–449.
48 Cardon LR, Palmer LJ: Population stratification and spurious allelic association. *Lancet* 2003; **361**: 598–604.