Population stratification across Europe

Matthias Wjst*, Rafael de Cid, Francesc Castro, Michael Abramson, Deborah Jarvis, Josep Anto, Xavier Estivill, Ursula Ackermann-Liebrich, Peter Burney, Isabel Cerveri, Sue Chinn, Roberto de Marco, Thoarinn Gislason, Joachim Heinrich, Christer Janson, Nino Künzli, Benedicte Leyneart, Francoise Neukirch, Jan Schouten, Jordi Sunyer, Amund Gulsvik, Ernst Omenaas, Cecilie Svanes, Paul Vermeire, Manolis Kogevinas for the ECRHS Genetics Group

to whom all correspondence should be addressed: Matthias Wjst, Comprehensive Pneumology Center (CPC), Institute of Lung Biology and Disease, Helmholtz Zentrum München, Ingolstädter Landstrasse 1, Neuherberg / Munich, Germany

Abstract

Background: Genetic disease association studies across Europe require a detailed knowledge of population structure and genetic ancestry. Extent of population stratification, number and type of SNPs necessary for correction is only partially known so far.

Methodology/Principal Findings: The European Community Respiratory Health Survey was performed in representative individuals of 20- to 44-year-old men and women where 5,350 DNA samples from 19 centers ranging from Umea in the North and Huelva in the South of Europe were analyzed. 22 random telomeric marker from nearly all chromosomes were tested as well as well as four SNPs in LCT and OCA2, two genes known to be under selective pressure. Several SNPs showed a strong correlation with geographical latitude and a high inflation factor λ indicating population stratification. One third of the telomeric SNPs even showed a strong association with body height. Mixed effects models including city of origin, multiple regression analysis including principal components from the covariance matrix as well as genomic control could all correct for stratification although to varying extent.

Conclusions/Significance: There is clear indication of genetic stratification across Europe that can be mostly corrected by current methods while several open questions remain.

Background

European population history is far from being understood on a genetic level [1] although a detailed knowledge might be necessary to adjust for population stratification. Population stratification is usually referred to as spurious false positive (or negative) association due to genetically different subpopulations where also disease prevalence differs [2]. Fear of population stratification has dominated genetic studies in the last decade when family based approaches were *en vogue* where population stratification is not a problem [3]. Current expectations, however, are that effects of population stratification are less important and will not lead to a distortion [2], [4] if dealt with properly. Support for this opinion is still scarce with smaller studies pointing towards non negligible effects of population stratification [5].

Attempts to correct for stratification have been made by direct or indirect phenotypical characteristics like birth place, self-reported ancestry, anthropometric measures, or even linguistic features. Unfortunately many of these labels are ambiguous and a more elegant way would be to using genetic marker itself. Genetic marker used so far were mainly derived by mtDNA [6] or Y chromosomal variation [7], however, due to the low discriminatory power and the sex restriction, both systems have only limited value for population based studies. Microsatellite marker would be more informative, but they are also less favoured than SNP marker as they do not allow to be automatically scored in a high throughput genotyping setting. SNPs may offer this advantage while extent of stratification, number of necessary SNPs, information content of SNPs and optimal

correction strategy is still under discussion [8], [9], [10], [11], [12], [13]. To get an unbiased estimate of European population stratification, we genotyped an European cohort by a set of random SNP including some ancestry informative SNPs that are likely to indicate stratification.

Methods

Sample. The methods for ECRHS I were published earlier with protocols and questionnaires available from the study website www.ecrhs.org. Briefly, participating centres were each selected from an area defined by pre-existing administrative boundaries, with a population of at least 150000 people. An up-to-date sampling frame was used to randomly select at least 1500 men and 1500 women aged 20 to 44 years carried out in 1991-1993. All subjects were sent a questionnaire enquiring about respiratory symptoms and attacks of asthma in the last 12 months, current use of asthma medication and nasal allergies including hayfever (ECRHS I screening). This sample consists of 54 centres with 200,682 participants [14]. During ECRHS II follow-up in 1999-2002 subjects answered again a standardized questionnaire administered by trained interviewers and underwent lung function as well as blood tests [15]. Ethical approval was obtained for each centre from the appropriate institutional ethics committee and written consent was obtained from each participant.

Genotyping and data handling. Batches of frozen blood samples were sent by courier to Munich, where they were registered in a LIMS system and cross checked with the clinical ECRHS database in London. DNA was manually extracted with Puregene kits without proteinase K (Gentra, Minneapolis), tested for purity and quantified

with picogreen (Invitrogen, Carlsbad) before being aliquoted. Genotyping of the population stratification marker was performed as part of a larger program on genetics of respiratory health using the SNPlex™ platform in Barcelona (Applied Biosystems). Genotyping was based on multiplex OLA/PCR and capillary electrophoresis where we followed the manufacturer's protocol for the SNPlex™ Multiplex Genotyping Systems (Applied Biosystems, Foster City, CA). 4 ng of each DNA sample was fragmented by boiling at 99ºC 6 for 10 minutes and allowed to dehydrate. Using custom multiplex pools designed, synthesized, and formulated by the manufacturer, the manufacturer's specifications are followed through phosphorylation, oligonucleotide ligation, exonuclease clean-up, PCR, and hybridization steps (Applied Biosystems, Foster City, CA). All reactions are set up in a 384 well format using various robotic systems. Samples with lower DNA concentration (<3 ng/ul or < 5ng/ul after whole genome amplification) were discarded from analysis. Average genotyping rate for automatically called genotypes by Genemapper™ for the overall genotyping project was of 93 %. Genotyping quality was controlled, in addition to internal positive and negative controls provided by AB's manufacturer, for both genotype concordance and Mendelian inheritance where 6 duplicated samples (NA10860, NA10861, NA11992, NA11993, NA11994 and NA11995) of two HapMap reference trios (CEPH131 and CEPH132) were incorporated in the genotyping process. Each trio was genotyped twelve times on average. A high concordance rate was observed with reference genotypes for these samples (http://www.hapmap.org, rel21_NCBI_Build35) when available with only one discordant genotype. No major discrepancies between duplicated samples were observed. No Mendelian inheritance error were observed in informative trios. Moreover specific male marker from Y chromosome (rs3894) was analysed and conflictive samples for Y marker genotype were discarded. Genotype concordance was tested using SNPator, a web-based tool for genotyping management and SNP analysis developed by CEGEN (http://www.CEGEN.org).

Analysis. R software 2.0.1 was used for further data handling and analysis (www.r-project-org) including the R libraries allelic, car, gmodels, genetics, lattice, nlme and RODBC as well as some user defined libraries. Additional scripts included population genetic F statistics (David Duffy, www.qimr.edu.au/davidD/R/fstat.R), and genomic control (Bernie Devlin, wpicr.wpic.pitt.edu/WPICCompGen/genomic_control/genomic_control.htm). Stratification was further tested with structure software 2.1 (David Pritchard, pritch.bsd.uchicago.edu/software.html). The range between k=2 and k=15 was tested, with 10,000 burn-in followed by 10,000 iterations where summary statistics of α, Fst and the likelihoods converged. As parameter for the ancestry model, admixture was chosen without where individual i inherited part of his genome from ancestors in population k without prior knowing population k. Conditional on the ancestry vector $q^{(i)}$ the origin of each allele is independent. Admixmap 3.4 (David O'Donnell www.ucd.ie/genepi/admixmap/) was also used for analysis, however, estimates were unreliable for unknown reasons and are therefore not reported here.

Results

5,868 ECRHS DNA samples were intended for genotyping. These included even 969 samples of low DNA amount that were prior subject to WGA. Following genotyping 16 females with Y chromosome derived amplification products were excluded, as well as all genotypes of the two SNPs with less than 80% call rate (rs964681, rs1528460) as well as all individuals with less than 80% of all possible genotypes (N=510). 5,350 samples from 19 center (of these 18 European) typed for 26 SNPs finally remained under analysis of population stratification. Table 1 shows details of the included center and table 2 all tested marker. Except 2 SNPs close to LCT (lactase) and 2 located in OCA2

(oculocutaneous albinism II or pink-eye dilution homolog in the mouse) SNPs were located on all human chromosomes except X and Y chromosome.

All SNPs were polymorphic (table 2) and in Hardy-Weinberg equilibrium except of 4 SNPs: rs4988235, rs309125 (both LCT), rs1800404 (OCA2), rs719366 (random SNP at telomere). When Hardy-Weinberg equilibrium was analyzed by city, centers had between zero and five marker that significantly violated HWE. The strongest deviation ($p<0.001$) was seen in Ipswich with an OCA marker, in Melbourne with both LCT marker and in Paris with 2 random marker.

When genotype frequencies of single center were compared with European averages differences, mainly the two LCT marker showed significant differences (table 4). Galdakao showed differences for both OCA marker, one of these differences shared with Huelva and Uppsala. Tartu had 2 and Umea 1 significant different genotype frequency at the telomeric SNP sites. Only LCT ($R^2=0.84$) and OCA marker were in LD ($R^2=0.72$) while all other marker were independent from each other.

We then constructed a matrix of genetic distance measure Fst values for all marker between single centers. In this analysis, most marker showed Fst values <0.05, while values for both LCT marker scattered with Fst values up to 0.31. Bergen showed high distances at rs1900758 to all other center while SNP rs873196 and rs1024116 (table 5) showed a Fst = 0.78 n Galdako and Paris.

SNP allele frequency was highly correlated with

geographical latitude (both LCT SNPs rs4988235 R^2=-0,72 and rs309125 R^2=0,72; also both OCA SNP rs1900758 R^2=0,76, as well as several random marker, see Figure 1). When trying to find subpopulations in the whole data set by the structure algorithm, however, no clear pattern emerged (Figure 2). Neither by the recommended parameter α nor by visual inspection, clear subgroups could be identified that would allow a better classification than by city of origin. As a further descriptive analysis we also applied hierarchical cluster analysis (also known als single linkage method being closely related to a minimal spanning tree algorithm, Figure 3). Grouping by cluster analysis clearly was more informative, resembling closely related populations like both German or both Belgian centers.

In a last step we compared different methods that have been recommended for control of population stratification (table 7). Stratification is clearly present in this sample with an estimate of λ ranging between 2.3 and 5.2. A linear regression using body height as a dummy trait showed 9 associated marker (table 7). Bonferroni corrected p-value cut-off value would be 0,0019 and FDR 0,0027 in this setting, leading to 5 significantly associated marker even after FDR correction. A further adjustment of the association by factors derived from structure and admixmap algorithms largely failed with the first principal component derived from the SNP covariance matrix being also ineffective for this purpose. This was in contrast to a mixed effect model that including city as fixed term and performed rather well as did the genomic control approach.

Discussion

We have shown that allele frequencies of several SNPs differed significantly from the European average with many SNPs showing a strong correlation with geographical latitude. As a result 9 of the 26 SNPs showed spurious associations with body height. MCMC clustering could not resolve major European groups, however, a mixed effects model including city of origin, a multiple regression including principal components from the SNP covariance matrix as well as the genomic control approach could correct for population stratification.

Reasons for departure from HWE

In addition to largely differing allele frequencies we found a deviation from Hardy-Weinberg equilibrium for both LCT SNPs. This effect was largely influenced by the center Melbourne [16]. Being the only non European city, Melbourne represents a diverse and multicultural city with almost a quarter of the population being born overseas. Although the ECRHS population sample was predominantly Anglo-Celtic and of Caucasians origin, genetic results may reflect the fact that Melbourne is home to residents from >200 countries. Most likely, the observed HWE deviation therefore does not indicate any genotyping error but reflect the violation of implicit HWE assumptions like infinite population size, random mating, no migration and genotypes of equal fitness [17]. These assumptions might also hold for the LCT variants as indicated by recent historical [18], geographical [19] or disease-oriented research [20]. In addition departure of Hardy-Weinberg may also occur by chance where it may be kept in mind that

evolution would not occur if the Hard-Weinberg law would always hold. The only European city showing also two marker with HWE violation is Paris – the second most densely populated area in the Western World.

European heritage

The Paris center also showed another unusual observation as the Fst pattern at two SNP marker resembled Galdakao, a Basque community in Northern Spain. A retrospective analysis of study participants in Paris indeed showed Basque surnames, which makes it likely that a Basque subgroup has been identified by these two markers. Bergen, the second largest city of Norway, also had a SNP marker separating this city from the rest of the study population; the reason is unclear but may be found with the particular population history that goes back to the Hanseatic League.

The European population tree constructed from the hierarchical clustering of our SNP marker closely resembles previous data obtained by mtDNA and Y chromosome variation [1], [7], [21]: Hamburg, Erfurt and Basel were correctly classified into the West/High Germanic cluster, Spanish and French cities into the Latin cluster, British center into the West/Low Germanic cluster and Bergen and Uppsala belonging to the North Germanic cluster. The Basque community separates from the rest of the European population [6]. A more detailed work up of SNP and haplotype flow may help to answer open questions of prehistoric admixture on the genome of Europeans [22]. Our data at least confirm the previous observation that European populations are hybrids, containing variable proportions derived from Palaeolithic settlers and Neolithic

immigrants. Matching these data with hominin fossils as currently been undertaken [23], [24] will be challenging and further delineate genes with selective sweeps [25].

Extent of stratification

The extent of population stratification has been mainly examined between different continental populations [26] while subcontinental differences in diallelic marker have only recently be tested on a larger scale [27], [28]. Seldin genotyped 928 European Americans of different regional origin. Fst values [29] have been small but there was evidence for major difference in population structure of "Southern" and "Northern" Europe . Compared with the estimates of other smaller studies [30] our assessment of λ of 2.3 to 5.2 seems to be rather high. Differences may be accounted to a larger number of samples or a larger number of informative marker included here. Further studies will need to confirm if there is really such a high level of stratification across European samples.

Control of stratification

While extent of stratification can be measured by an inflation of association results [31] the assignment of individuals to subgroups is widely determined with the structure algorithm using Monte Carlo markov chains. Unexpectedly the structure algorithm could not discriminate our population even by using different starting parameters. Many other statistical procedures, however, have been described [8], [9], [10] to detect and control population stratification. The frequently used "genomic control" approach uses multiple polymorphisms where the

degree of overdispersion generated by population substructure can be estimated [32]. As the association of all of our random SNP marker with body height is definitely spurious we may conclude that genomic control effectively eliminates such associations. The mixed effects model with city as fixed as fixed and genotype as random effect also provided reasonable results. In contrast to previous observations [33] the inclusion of principal component analysis factors left the original regression coefficients unchanged. This is also in contrast to another recent study [12] that compared traditional case-control tests, structured association, genomic control and principal components analysis under various simulated stratification levels. The authors found that the performance of PCA was very stable under various scenarios. The difference to the empirical results in this study not fully clear but may relate to extent of stratification and numerical properties of our marker panel.

It is also an open question if the original statistic may be deflated by a common factor regardless of the marker being tested [11] as there could be variable stratification in genomic regions that are under high selective pressure. Further work may therefore include also known associated marker and examine stratification on a genome-wide scale. As we did not include any true positive height marker, it may also be tested if any of the methods provides over-correction.

SNP panels

It is an open question how many SNPs are being necessary for adjustment of population stratification as ccepted

standards for SNP panels do not exist at the moment. The Frudakis set 2003 [34] was probably the first SNP based panel to infer ethnic origin. It included 56 SNPs derived from an initial set of 211 SNPs of pigmentation and xenobiotic response genes and could differentiate individuals in Florida by self-reported African, Asian and European descent. The Freedman set [30] included between 24 and 48 SNPs and a total of 11 studies, from various sources: noncoding SNPs, missense SNPs and ancestry informative SNPs. The Smith set [35] was based on 3,011 SNPs extracted of ~450,000 database entries and further validated in 78 European Americans, 120 sub-Saharan Africans,109 AfricanAmericans, 40 Cantonese Chinese, and 29 Mexican Amerindians. The Lao set [36] used 10 out of 8,491 SNPs of the Affymetrix 10K array to differentiate 76 human individuals from 21 sampling locations representing Africa, South Africa, America, Asia, North Asia and Europe (YCC panel). The Seldin set [28] used 2,657 SNPs in 500 kB distance (basically an Illumina linkage panel) in 681 Americans of reported European ancestry. Finally, the Sanchez set 2006 [37], that we used here in part is a forensic panel of 52 SNPs reported to be polymorphic in European, Asian and African populations, located on distal autosomes and at least 100 kB distant from known genes. Our current SNP panel was able to control for population structure, however, as the results still show some residual stratification it may be necessary to increase size and quality of the SNP marker panel.

Tables and Figures

Table 1: Study centers

city		continent	inhabitants	latitude (degree)	participants (corrected)
Albacete	Spain	Europe	148900	38	313
Antwerp South	Belgium	Europe	NA	51	226
Antwerp City	Belgium	Europe	452500	51	292
Barcelona	Spain	Europe	1503900	41	218
Basel	Switzerland	Europe	166000	47	410
Bergen	Norway	Europe	209400	60	492
Erfurt	Germany	Europe	217100	50	258
Galdakao	Spain	Europe	29500	43	363
Grenoble	France	Europe	150800	45	240
Hamburg	Germany	Europe	1793800	53	247
Huelva	Spain	Europe	142300	37	250
Ipswich	Great Britain	Europe	129700	52	147
Melbourne	Australia	Austr/NewZ	3660000	-37	375
Norwich	Great Britain	Europe	169800	52	241
Oviedo	Spain	Europe	201200	43	195
Paris	France	Europe	2152400	48	324
Tartu	Estonia	Europe	113400	58	267
Umea	Sweden	Europe	72700	63	198
Uppsala	Sweden	Europe	126600	59	458

Table 2: SNPs analyzed

Gene	Code	Build	Chromosome	Position	Alleles	Genotyped %	analyzed
random	rs1490413	dbSNP125	1	4277696	AG	99,4	yes
random	rs10495407	dbSNP125	1	234765349	AG	99,2	yes
random	rs876724	dbSNP125	2	104974	CT	98,9	yes
LCT-13910CT	rs4988235	dbSNP125	2	136442378	CT	99,0	yes
LCT-	rs309125	dbSNP125	2	136477287	CT	99,1	yes
random	rs907100	dbSNP125	2	239345579	CG	98,6	yes
random	rs1357617	dbSNP125	3	936782	AT	99,5	yes
random	rs1979255	dbSNP125	4	190693229	CG	99,4	yes
random	rs717302	dbSNP125	5	2932395	AG	99,0	yes
random	rs1029047	dbSNP125	6	1080939	AT	99,5	yes
random	rs917118	dbSNP125	7	4230244	CT	98,3	yes
random	rs2056277	dbSNP125	8	139468298	CT	99,7	yes
random	rs1015250	dbSNP125	9	1813774	CG	99,8	yes
random	rs964681	dbSNP125	10	132588409	CT	75,3	no
random	rs2076848	dbSNP125	11	134172756	AT	98,8	yes
random	rs2111980	dbSNP125	12	104830721	AG	99,6	yes
random	rs1335873	dbSNP125	13	19799724	AT	99,3	yes
random	rs873196	dbSNP125	14	97915284	CT	86,5	yes
OCA2	rs1900758	dbSNP125	15	25903692	AG	91,0	yes
OCA2	rs1800404	dbSNP125	15	25909368	AG	99,4	yes
random	rs1528460	dbSNP125	15	52997997	CT	58,4	no
random	rs729172	dbSNP125	16	5546198	AC	99,6	yes
random	rs740910	dbSNP125	17	5647347	AG	98,1	yes
random	rs1024116	dbSNP125	18	73561374	AG	86,4	yes
random	rs719366	dbSNP125	19	33155177	CT	98,9	yes
random	rs1005533	dbSNP125	20	38920524	AG	99,7	yes
random	rs722098	dbSNP125	21	15607469	AG	99,4	yes
random	rs2040411	dbSNP125	22	46156931	AG	98,7	yes

Table 3: SNP genotype frequencies and Hardy-Weinberg equilibrium

marker	SNP	genotype	N	% genotype		N	% genotype		N	%	HWE
random	rs1490413	AA	1030	18,7	AG	2699	48,9	GG	1754	31,8	0,891
random	rs10495407	AA	642	11,6	AG	2444	44,3	GG	2387	43,3	0,674
random	rs876724	CC	2618	47,5	CT	2315	42,0	TT	530	9,6	0,590
LCT-13910CT	rs4988235	CC	1157	21,0	CT	2412	43,7	TT	1908	34,6	0,000
LCT	rs309125	CC	2395	43,4	CT	2307	41,8	TT	765	13,9	0,000
random	rs907100	CC	1085	19,7	CG	2668	48,4	GG	1678	30,4	0,680
random	rs1357617	AA	2832	51,4	AT	2203	40,0	TT	449	8,1	0,484
random	rs1979255	CC	2443	44,3	CG	2428	44,0	GG	614	11,1	0,785
random	rs717302	AA	1309	23,7	AG	2776	50,3	GG	1371	24,9	0,194
random	rs1029047	AA	768	13,9	AT	2481	45,0	TT	2239	40,6	0,055
random	rs917118	CC	2813	51,0	CT	2199	39,9	TT	412	7,5	0,543
random	rs2056277	CC	2955	53,6	CT	2155	39,1	TT	396	7,2	0,918
random	rs1015250	CC	222	4,0	CG	1761	31,9	GG	3523	63,9	0,933
random	rs2076848	AA	992	18,0	AT	2659	48,2	TT	1801	32,7	0,846
random	rs2111980	AA	1074	19,5	AG	2646	48,0	GG	1773	32,2	0,125
random	rs1335873	AA	381	6,9	AT	2046	37,1	TT	3053	55,4	0,137
random	rs873196	CC	780	14,1	CT	2224	40,3	TT	1788	32,4	0,046
OCA2	rs1900758	AA	2236	40,6	AG	2205	40,0	GG	591	10,7	0,185
OCA2	rs1800404	AA	3355	60,8	AG	1830	33,2	GG	301	5,5	0,015
random	rs729172	AA	929	16,8	AC	2663	48,3	CC	1895	34,4	0,911
random	rs740910	AA	2705	49,1	AG	2244	40,7	GG	465	8,4	1,000
random	rs1024116	AA	1611	29,2	AG	2287	41,5	GG	886	16,1	0,139
random	rs719366	CC	794	14,4	CT	2459	44,6	TT	2209	40,1	0,011
random	rs1005533	AA	1592	28,9	AG	2736	49,6	GG	1165	21,1	0,892
random	rs722098	AA	3554	64,5	AG	1738	31,5	GG	188	3,4	0,179
random	rs2040411	AA	2392	43,4	AG	2413	43,8	GG	646	11,7	0,319

Table 4: Signficance value for differences in local genotype distribution compared to European average

	Albacete	Antwerp City	Antwerp South	Barcelona	Basel	Bergen	Erfurt	Galdakao	Grenoble	Hamburg	Huelva	Ipswich	Melbourne	Norwich	Oviedo	Paris	Tartu	Umea	Uppsala
rs1490413	1,2E-03	8,5E-01	2,7E-01	7,0E-01	6,4E-01	9,9E-02	9,7E-01	1,2E-01	2,2E-01	1,0E+00	5,4E-01	1,3E-01	2,9E-02	6,4E-01	6,1E-02	7,4E-01	3,5E-03	3,3E-02	1,1E-01
rs10495407	8,6E-02	7,3E-01	2,4E-01	8,2E-01	1,4E-01	4,1E-01	1,2E-01	4,8E-02	9,5E-01	9,6E-01	7,1E-01	5,2E-01	6,4E-01	3,5E-01	9,1E-01	8,0E-01	9,4E-01	4,0E-01	3,2E-01
rs876724	5,9E-01	2,8E-02	9,8E-01	8,2E-01	8,4E-02	9,7E-02	8,3E-01	5,9E-01	7,5E-01	8,7E-01	1,9E-01	2,6E-01	8,2E-01	1,3E-02	4,5E-02	2,8E-01	1,4E-01	7,0E-04	9,4E-03
rs4988235	1,8E-13	2,2E-06	2,4E-05	6,3E-16	5,5E-05	4,2E-47	3,6E-01	2,7E-01	7,8E-16	5,0E-01	1,4E-17	4,6E-04	5,8E-06	3,0E-06	2,3E-03	5,7E-09	1,3E-08	8,1E-08	1,6E-14
rs309125	7,7E-11	3,1E-06	2,9E-04	2,7E-16	1,2E-04	1,0E-34	1,3E-01	1,3E-01	5,2E-11	5,8E-01	6,5E-14	1,1E-02	3,6E-05	2,7E-05	1,1E-02	1,6E-07	2,1E-07	7,0E-07	3,5E-13
rs907100	2,1E-01	8,1E-01	9,6E-01	4,3E-01	2,0E-01	3,2E-01	7,0E-01	9,3E-01	1,2E-01	6,5E-01	7,1E-01	2,0E-01	8,5E-02	4,6E-02	5,0E-01	1,9E-02	4,2E-01	1,4E-03	6,3E-03
rs1357617	2,1E-02	4,8E-01	1,5E-01	7,9E-01	2,4E-02	9,6E-01	3,1E-01	7,2E-03	3,9E-01	7,0E-01	6,9E-02	6,7E-02	5,4E-01	1,8E-01	6,9E-01	7,3E-01	4,2E-01	3,7E-01	9,3E-01
rs1979255	2,3E-02	9,1E-01	7,3E-01	3,7E-02	6,0E-01	4,0E-02	3,0E-01	1,7E-01	4,7E-03	3,3E-01	4,7E-02	7,4E-01	1,4E-02	8,4E-01	2,2E-01	1,3E-01	7,9E-06	1,6E-03	3,3E-03
rs717302	3,7E-01	4,0E-02	1,3E-01	5,3E-01	7,7E-01	6,6E-01	6,9E-01	2,2E-04	5,1E-01	5,4E-01	5,2E-01	1,6E-01	8,9E-01	1,1E-01	6,8E-01	5,9E-01	4,0E-02	2,4E-01	7,8E-02
rs1029047	3,1E-02	2,1E-01	3,8E-01	8,6E-01	3,2E-01	4,8E-02	2,7E-01	4,2E-03	3,8E-01	8,7E-01	3,0E-02	9,9E-01	2,5E-01	6,4E-01	4,3E-01	6,5E-04	1,2E-03	9,1E-01	6,5E-01
rs917118	8,7E-01	6,7E-01	5,0E-01	2,0E-01	8,1E-02	7,5E-02	1,8E-01	5,6E-01	3,4E-01	6,5E-02	5,2E-01	1,2E-01	2,8E-01	5,6E-01	9,6E-01	3,8E-01	2,0E-01	7,2E-01	1,8E-01
rs2056277	3,5E-01	4,4E-01	8,9E-01	1,2E-01	2,2E-01	2,0E-01	5,4E-01	1,1E-04	3,7E-01	7,2E-01	9,8E-03	4,0E-01	3,0E-01	4,1E-01	7,1E-01	4,0E-01	8,3E-01	5,0E-03	6,4E-01
rs1015250	5,0E-05	2,2E-01	7,4E-01	5,2E-01	7,1E-01	1,0E-01	7,4E-01	8,1E-01	4,5E-01	1,7E-02	2,2E-02	9,2E-01	6,8E-01	4,2E-01	1,1E-01	8,7E-01	9,3E-01	1,7E-01	8,4E-02
rs2076848	4,9E-01	3,8E-01	1,5E-02	9,4E-02	2,7E-02	1,0E-01	4,2E-01	1,7E-04	8,1E-01	1,2E-01	2,1E-02	6,5E-01	2,8E-01	6,1E-02	6,6E-01	8,5E-01	2,0E-02	1,3E-09	2,1E-04
rs2111980	7,1E-01	3,2E-01	1,8E-01	5,4E-01	6,7E-01	4,1E-01	4,8E-01	2,3E-01	6,7E-01	7,7E-01	8,5E-02	9,4E-01	6,4E-01	3,2E-01	1,7E-01	7,0E-01	1,8E-01	7,5E-03	8,2E-01
rs1335873	3,0E-03	6,7E-01	2,7E-01	2,7E-01	9,1E-01	2,0E-03	8,7E-01	6,6E-01	1,1E-01	5,6E-01	5,3E-02	5,1E-01	3,3E-01	8,5E-02	2,0E-02	8,0E-01	2,3E-01	2,1E-01	5,8E-01
rs873196	2,9E-01	5,8E-01	4,7E-01	6,5E-01	6,7E-02	4,8E-01	8,2E-02	7,5E-01	5,9E-01	7,3E-01	1,5E-03	8,0E-01	2,8E-01	8,6E-01	4,6E-01	7,2E-01	8,4E-01	2,5E-03	6,4E-01
rs1900758	5,8E-04	1,2E-01	4,3E-02	2,6E-05	1,4E-02	4,7E-01	9,8E-04	5,1E-07	3,5E-01	6,9E-04	6,0E-09	8,4E-01	2,0E-01	5,8E-02	2,7E-03	3,5E-05	1,8E-02	1,4E-02	5,8E-08
rs1800404	3,2E-02	2,7E-01	1,4E-01	4,2E-03	4,7E-01	1,3E-04	1,4E-04	1,6E-08	6,5E-01	9,2E-03	3,1E-04	5,4E-04	2,7E-01	5,3E-01	1,6E-02	4,0E-04	1,0E-01	6,1E-01	8,4E-04
rs729172	6,9E-01	6,3E-01	1,0E+00	9,5E-02	7,5E-01	9,2E-01	9,8E-01	5,2E-01	4,3E-02	8,2E-02	1,0E-01	4,7E-02	4,2E-01	4,0E-01	3,3E-01	2,6E-01	5,1E-01	1,2E-01	7,9E-01
rs740910	1,8E-01	8,0E-01	7,4E-01	3,5E-01	1,1E-01	2,7E-02	6,7E-01	1,5E-01	9,0E-01	6,3E-01	9,1E-02	9,8E-01	3,3E-02	9,0E-02	4,4E-01	5,0E-01	3,2E-02	1,4E-02	3,0E-01
rs1024116	1,5E-01	3,4E-01	5,4E-01	5,2E-01	3,4E-02	9,9E-01	9,1E-01	4,6E-01	8,1E-01	4,1E-01	1,1E-01	7,1E-02	7,5E-01	3,9E-01	1,9E-01	2,8E-01	2,2E-02	5,2E-01	8,2E-01
rs719366	3,6E-03	3,7E-01	1,6E-02	3,7E-01	7,3E-01	7,2E-04	7,4E-01	1,3E-01	1,5E-01	3,2E-01	2,2E-04	8,1E-02	1,5E-01	1,1E-01	5,3E-04	9,7E-02	5,4E-01	2,6E-02	1,1E-04
rs1005533	9,8E-01	2,4E-01	2,0E-01	1,0E+00	1,8E-01	6,8E-01	6,5E-01	5,0E-03	3,9E-01	5,7E-02	4,1E-02	9,4E-01	8,7E-01	3,0E-01	5,5E-01	2,2E-01	1,5E-01	1,9E-01	1,4E-02
rs722098	6,6E-01	3,5E-01	3,2E-01	5,1E-02	6,9E-01	8,3E-01	5,0E-02	1,3E-01	4,1E-01	8,2E-01	6,3E-01	5,9E-01	5,0E-01	6,9E-01	9,4E-02	2,4E-01	4,2E-07	8,0E-02	7,2E-01
rs2040411	9,4E-02	2,9E-02	6,2E-01	7,9E-01	1,4E-01	9,4E-01	6,0E-01	1,4E-02	8,3E-01	9,7E-01	9,5E-02	7,1E-01	9,5E-03	6,4E-01	1,9E-01	9,3E-01	1,3E-01	2,6E-01	5,8E-01

Table 5: Fst matrix of SNP rs1024116 between center comparison

	Antwerp City	Antwerp South	Barcelona	Basel	Bergen	Erfurt	Galdakao	Grenoble	Hamburg	Huelva	Ipswich	Melbourne	Norwich	Oviedo	Paris	Tartu	Umea	Uppsala
Albacete	0,07	0,07	0,01	0,00	0,00	0,00	0,25	0,00	0,00	0,00	0,02	0,00	0,00	0,01	0,71	0,00	0,00	0,05
Antwerp S		0,00	0,07	0,07	0,06	0,07	0,07	0,05	0,06	0,06	0,06	0,06	0,06	0,06	0,51	0,07	0,07	0,00
Antwerp C	0,00		0,07	0,07	0,06	0,07	0,07	0,05	0,06	0,06	0,06	0,07	0,06	0,07	0,51	0,07	0,07	0,00
Barcelona	0,07	0,07		0,00	0,00	0,00	0,25	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,70	0,01	0,00	0,05
Basel	0,07	0,07	0,00		0,00	0,00	0,25	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,70	0,01	0,00	0,05
Bergen	0,06	0,06	0,00	0,00		0,00	0,24	0,00	0,00	0,00	0,01	0,00	0,00	0,00	0,69	0,01	0,00	0,04
Erfurt	0,07	0,07	0,00	0,00	0,00		0,25	0,00	0,00	0,00	0,01	0,00	0,00	0,00	0,70	0,00	0,00	0,05
Galdakao	0,07	0,07	0,25	0,25	0,24	0,25		0,22	0,24	0,24	0,23	0,24	0,24	0,24	0,30	0,26	0,25	0,10
Grenoble	0,05	0,05	0,00	0,00	0,00	0,00	0,22		0,00	0,00	0,01	0,00	0,00	0,00	0,68	0,01	0,00	0,03
Hamburg	0,06	0,06	0,00	0,00	0,00	0,00	0,24	0,00		0,00	0,01	0,00	0,00	0,00	0,70	0,00	0,00	0,05
Huelva	0,06	0,06	0,00	0,00	0,00	0,00	0,24	0,00	0,00		0,00	0,00	0,00	0,00	0,69	0,01	0,00	0,04
Ipswich	0,06	0,06	0,00	0,00	0,01	0,01	0,23	0,01	0,01	0,00		0,00	0,01	0,00	0,68	0,03	0,01	0,04
Melbourne	0,06	0,07	0,00	0,00	0,00	0,00	0,24	0,00	0,00	0,00	0,00		0,00	0,00	0,69	0,01	0,00	0,04
Norwich	0,06	0,06	0,00	0,00	0,00	0,00	0,24	0,00	0,00	0,00	0,01	0,00		0,00	0,69	0,00	0,00	0,04
Oviedo	0,06	0,07	0,00	0,00	0,00	0,00	0,24	0,00	0,00	0,00	0,00	0,00	0,00		0,69	0,02	0,00	0,04
Paris	0,51	0,51	0,70	0,70	0,69	0,70	0,30	0,68	0,70	0,69	0,68	0,69	0,69	0,69		0,71	0,70	0,54
Tartu	0,07	0,07	0,01	0,01	0,01	0,00	0,26	0,01	0,00	0,01	0,03	0,01	0,00	0,02	0,71		0,00	0,05
Umea	0,07	0,07	0,00	0,00	0,00	0,00	0,25	0,00	0,00	0,00	0,01	0,00	0,00	0,00	0,70	0,00		0,05

Table 6: Summary of the structure analysis for k=2 up to k=15 strata

K	Ln P(D)	Var[LnP(D)]	α1	Fst_1	Fst_2	Fst_3	Fst_4	Fst_5	Fst_6	Fst_7	Fst_8	Fst_9	Fst_10	Fst_11	Fst_12	Fst_13	Fst_14	Fst_15
2	-174399	2742	0,26	0,27	0,01													
3	-172259	3223	0,05	0,07	0,22	0,09												
4	-171606	3977	0,06	0,09	0,10	0,12	0,14											
5	-174426	8850	0,24	0,28	0,23	0,02	0,02	0,13										
6	-174215	9356	0,14	0,04	0,10	0,23	0,03	0,25	0,12									
7	-173871	9664	0,06	0,12	0,13	0,13	0,11	0,04	0,28	0,04								
8	-177212	14953	0,20	0,21	0,02	0,38	0,03	0,29	0,01	0,02	0,29							
9	-178047	17651	0,12	0,04	0,29	0,13	0,33	0,05	0,04	0,17	0,14	0,14						
10	-194843	51185	0,12	0,17	0,03	0,27	0,14	0,03	0,13	0,04	0,31	0,28	0,05					
11	-189094	39861	0,11	0,02	0,03	0,20	0,14	0,03	0,04	0,29	0,33	0,15	0,11	0,27				
12	-178094	18092	0,10	0,27	0,19	0,32	0,29	0,16	0,02	0,01	0,19	0,04	0,03	0,01	0,15			
13	-176972	15539	0,11	0,33	0,02	0,34	0,02	0,00	0,26	0,02	0,16	0,25	0,03	0,02	0,28	0,19		
14	-179171	19761	0,12	0,29	0,31	0,21	0,19	0,00	0,01	0,37	0,02	0,01	0,36	0,01	0,03	0,03	0,30	
15	-198670	59531	0,10	0,34	0,24	0,25	0,03	0,25	0,02	0,06	0,03	0,02	0,03	0,19	0,27	0,02	0,37	0,18

Table 7: Effect of different methods to control for population stratification in a model for body weight

type	SNP	LR	PCA	ME	GC
random	rs1490413	6,6E-01	6,1E-01	4,0E-02	4,0E-01
random	rs10495407	9,2E-01	9,0E-01	5,0E-01	6,2E-01
random	rs876724	1,5E-01	5,8E-01	3,8E-01	6,5E-01
LCT	rs4988235	4,3E-16	2,7E-14	1,3E-03	2,2E-02
LCT	rs309125	6,1E-15	2,6E-15	4,2E-03	4,0E-02
random	rs907100	3,2E-02	4,5E-02	1,3E-01	1,3E-01
random	rs1357617	8,1E-02	1,6E-01	4,8E-01	9,3E-01
random	rs1979255	8,4E-02	1,8E-01	5,0E-01	8,9E-01
random	rs717302	7,9E-01	7,1E-01	2,1E-01	7,3E-01
random	rs1029047	1,9E-01	2,6E-01	9,5E-01	5,6E-01
random	rs917118	5,6E-01	9,7E-01	6,7E-01	8,1E-01
random	rs2056277	1,8E-02	1,1E-01	2,5E-01	6,6E-01
random	rs1015250	4,6E-02	2,7E-01	6,4E-01	6,6E-01
random	rs2076848	5,5E-03	5,8E-03	2,7E-01	4,2E-01
random	rs2111980	9,3E-01	5,8E-01	6,0E-01	6,7E-01
random	rs1335873	2,3E-01	9,2E-01	6,9E-01	4,4E-01
random	rs873196	2,9E-01	3,8E-01	3,3E-01	7,1E-01
OCA2	rs1900758	3,1E-08	4,8E-05	4,8E-01	4,7E-02
OCA2	rs1800404	1,5E-07	7,9E-05	2,1E-01	3,5E-02
random	rs729172	8,0E-01	4,4E-01	2,2E-01	5,3E-01
random	rs740910	5,0E-01	5,2E-01	3,7E-01	7,0E-01
random	rs1024116	7,2E-01	8,2E-01	8,7E-01	9,2E-01
random	rs719366	3,8E-04	3,3E-03	3,1E-01	2,2E-01
random	rs1005533	4,9E-01	8,4E-01	4,5E-01	7,2E-01
random	rs722098	7,6E-01	9,2E-01	2,1E-01	9,9E-01
random	rs2040411	1,3E-01	2,1E-01	9,8E-01	7,9E-01

LR: baseline linear regression for height including sex as covariateof homocygous status
PCA, like LR but including also first factor of principal component of SNP matrix as covariables (Price 2006)
ME: mixed effects model with city as fixed and genotype as well as sex as random effect (Bates & Pinheiro 1998)
GC, genomic control (Devlin & Roeder 1999)
STRAT, structured association testing conditional of ancestry of individuals (Pritchard 2001) has not been performed as it allows only to take into accout case-control status
ADMIXMAP, (hybrid of Bayesian and classical approach, McKeigue 2000) has been performed but estimates were unreliable for unknown reason

Figure 1: City averaged minor allele frequency of SNP rs719366 by geographical latitude. Melbourne is classified with absolute latitude.

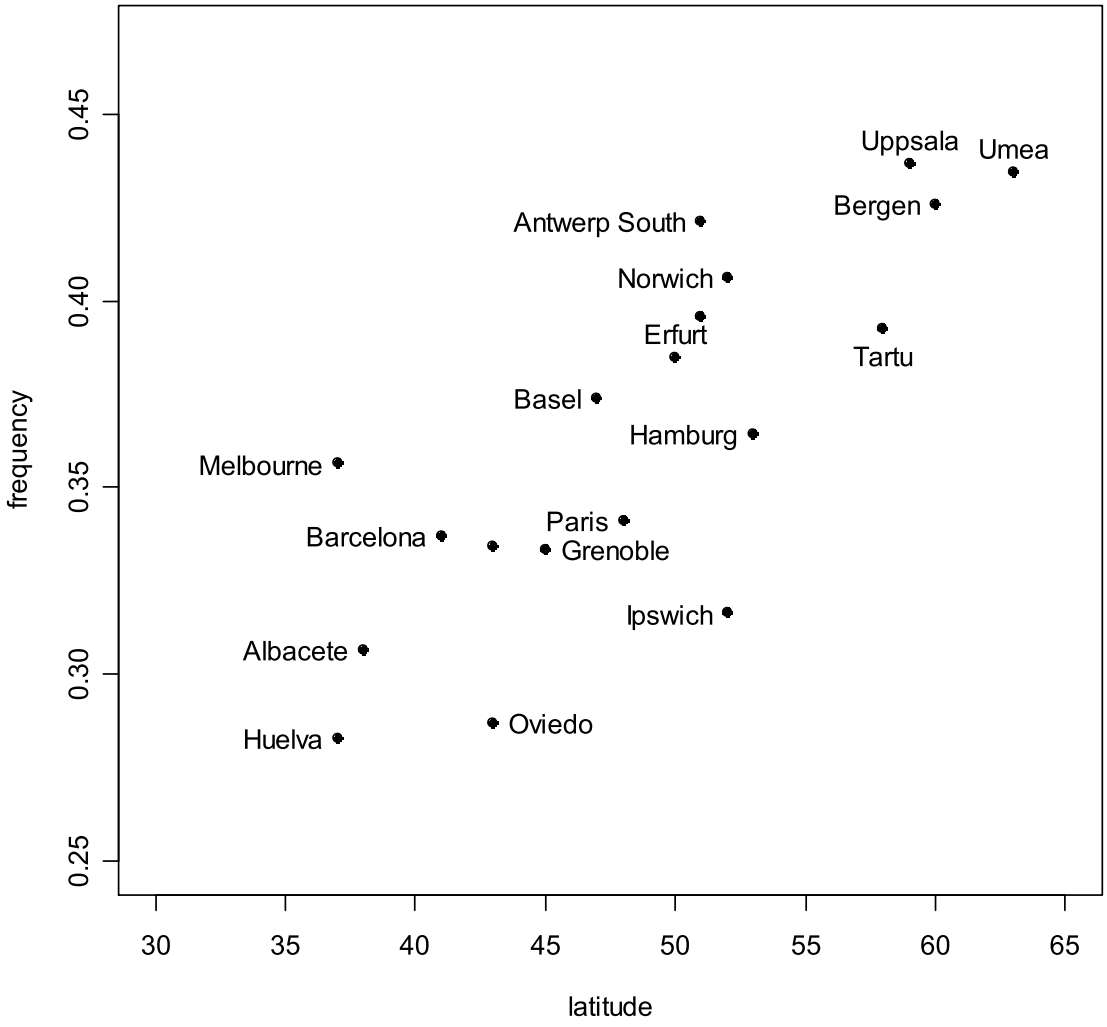


Figure 2: Structure analysis: Compressed bar plot of proportional membership to five groups in all individuals

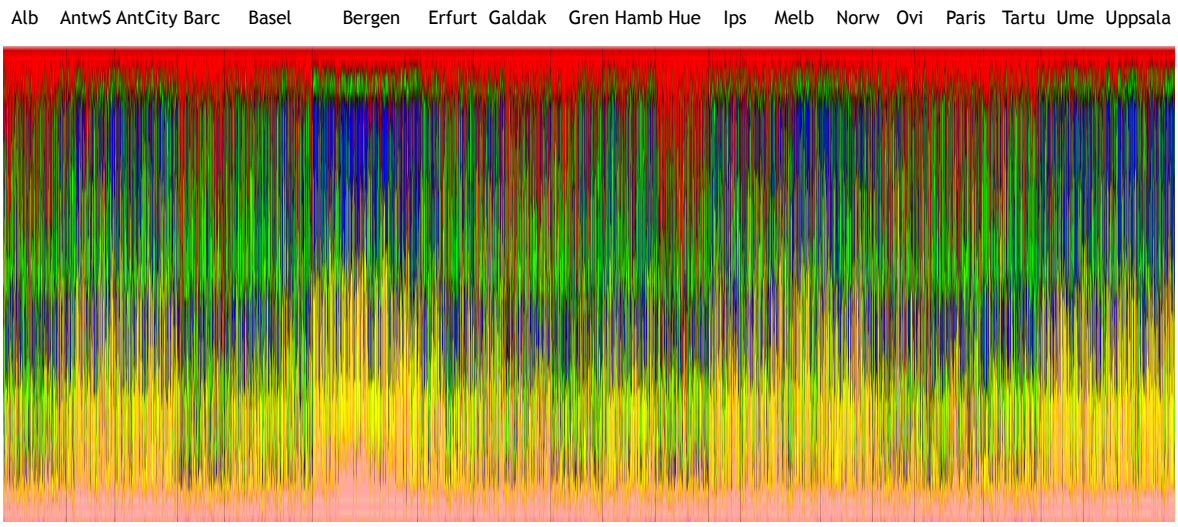
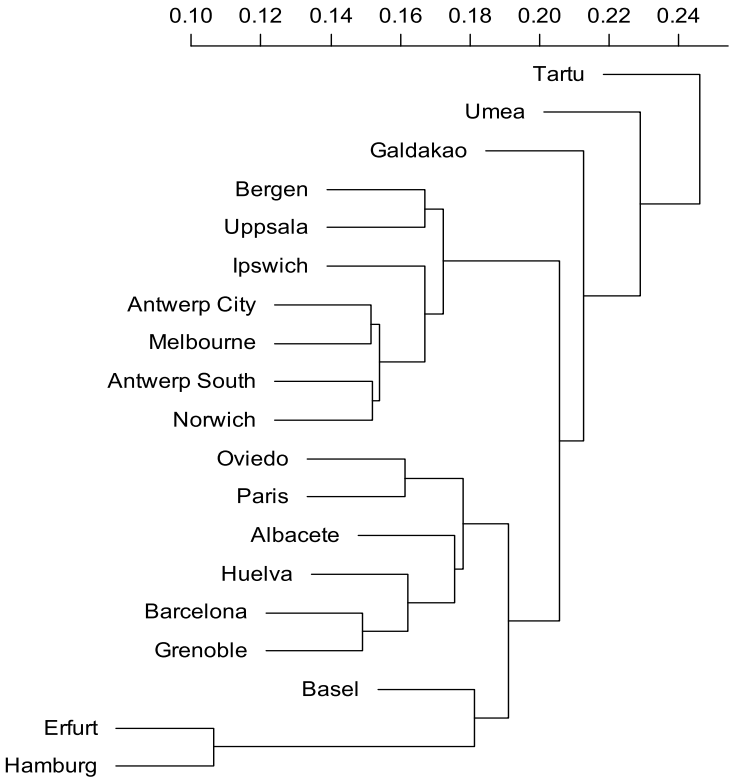


Figure 3: Hierarchical clustering of cities by SNP marker



Authors Contributions

All authors participated in the local study management and reviewed the paper. PB, SC and DJ planned and organised the ECRHS study center. MW is running the ECRHS biobank, conducted statistical analysis and wrote the paper. RdC and XE planned the genetic study, did genotyping and quality control, FC and DB did data analysis, MK planned the genetic study, wrote the grant application and supervised the analysis. All authors have read and approved the final version of the manuscript.

Competing Interests

All authors declare that they do not have any competing interests with the contents of this manuscript.

Acknowledgments

We wish to thank Michelle Emfinger for proof-reading of the manuscript, Christine Braig, Anika Luze and David Kutschke for excellent laboratory work. The ECRHS study is joint project by many participants and funded by many sources.

Project Leader: Peter Burney; Statistician: Sue Chinn; Principal Investigator: Deborah Jarvis; Project Co-ordinator: Jill Knox; Principal Investigator: Christina Luczynska; Assistant Statistician: J Potts; Data Manager: S Arinze.

Steering Committee: Josep M Antó, Institut Municipal d'Investigació Mèdica (IMIM-IMAS), Universitat Pompeu Fabra (UPF); Peter Burney, King's College London; Isa Cerveri, University of Pavia; Susan Chinn, King's College London; Roberto de Marco, University of Verona; Thorarinn Gislason, Iceland University Hospital; Joachim Heinrich, GSF – Institute of Epidemiology; Christer Janson, Uppsala University; Deborah Jarvis, King's College London; Jill Knox, King's College London; Nino Künzli, formerly University of Basel, now University of

Southern California Los Angeles; Bénédicte Leynaert, Institut National de la Santé et de la Recherche Médicale (INSERM); Christina Luczynska, King's College London; Françoise Neukirch, Institut National de la Santé et de la Recherche Médicale (INSERM); J Schouten, University of Groningen; Jordi Sunyer, Institut Municipal d'Investigació Mèdica (IMIM-IMAS), Universitat Pompeu Fabra (UPF); Cecilie Svanes, University of Bergen; Vermeire, University of Antwerp; Matthias Wjst, GSF Institute of Epidemiology.

Principal Investigators and Senior Scientific Team

Australia: Melbourne (M Abramson, EH Walters, J Raven, S Dharmage). Belgium: South Antwerp & Antwerp City (P Vermeire, J Weyler, M Van Sprundel, V Nelen). Canada: Halifax (D Bowie), Hamilton (MR Sears, HC Siersted), Montreal (MR Becklake, P Ernst), Prince Edward Island (L Sweet, L Van Til), Vancouver (M Chan-Yeung, H Dimich-Ward), Winnipeg (J Manfreda, NR Anthonisen). Estonia: Tartu (R Jogi, A Soon). France: Paris (F Neukirch, B Leynaert, R Liard, M Zureik), Grenoble (I Pin, J Ferran-Quentin). Germany: Erfurt (J Heinrich, M Wjst, C Frye, I Meyer). Iceland: Reykjavik (T Gislason, E Bjornsson, D Gislason, T Blondal, KB Jorundsdottir). Italy: Turin (M Bugiani, P Piccioni, E Caria, A Carosso, E Migliore, G Castiglioni), Verona (R de Marco, G Verlato, E Zanolin, S Accordini, A Poli, V Lo Cascio, M Ferrari), Pavia (A Marinoni, S Villani, M Ponzio, F Frigerio, M Comelli, M Grassi, I Cerveri, A Corsico). Netherlands: Groningen and Geleen (J Schouten, M Kerkhof). Norway: Bergen (A Gulsvik, E Omenaas, C Svanes, B Laerum). Spain: Barcelona (JM Antó, J Sunyer, M Kogevinas, JP Zock, X Basagana, A Jaen, F Burgos), Huelva (J Maldonado, A Pereira, JL Sanchez), Albacete (J Martinez-Moratalla Rovira, E Almar), Galdakao (N Muniozguren, I Urritia), Oviedo (F Payo). Sweden: Uppsala (C Janson, G Boman, D Norback, M Gunnbjornsdottir), Goteborg (K Toren, L Lillienberg, AC Olin, B Balder, A Pfeifer-Nilsson, R Sundberg), Umea (E Norrman, M Soderberg, K Franklin, B Lundback, B Forsberg, L Nystrom). Switzerland: Basel (N Künzli, B Dibbert, M Hazenkamp, M Brutsche, U Ackermann-Liebrich). United Kingdom: Norwich (D Jarvis, B Harrison), Ipswich (D Jarvis, R Hall, D Seaton).

Funders

Financial support for ECRHS I centres: Allen and Hanbury, Belgian Science Policy Office, National Fund for Scientific Research; Ministère de la Santé, Glaxo France, Institut Pneumologique d'Aquitaine, Contrat de Plan Etat-Région Languedoc-Rousillon, CNMATS, CNMRT (90MR/10, 91AF/6), Ministre delegué de la santé, RNSP, France; Health Canada, Province of Prince Edward Island, Glaxo Canada; GSF, and the Bundesministerium für Forschung und Technologie, Bonn, Germany; Ministero dell'Università e della Ricerca Scientifica e Tecnologica, CNR, Regione Veneto grant RSF n. 381/05.93, Italy; Norwegian Research Council project no. 101422/310; Dutch Ministry of Wellbeing, Public Health and Culture, Netherlands; Ministero Sanidad y Consumo FIS (grants #91/0016060/00E-05E and #93/0393), and grants from Hospital General de Albacete, Hospital General Juan Ramón Jiménenz, Consejeria de Sanidad Principado de Asturias, Spain; The Swedish Medical Research Council, the Swedish Heart Lung Foundation, the Swedish Association against Asthma and Allergy; Swiss National Science Foundation grant 4026-28099; National Asthma Campaign, British Lung Foundation, Department of Health, South Thames Regional Health Authority, UK; United States Department of Health, Education and Welfare Public Health Service (grant #2 S07 RR05521-28) and Victorian Asthma Foundation.

References

1. Cavalli-Sforza LL, Menozzi, P., Piazza, A. (1996) The history and geography of human genes. Princeton University Press.
2. Marchini J, Cardon LR, Phillips MS, Donnelly P (2004) The effects of human population structure on large genetic association studies. Nat Genet 36: 512-517.
3. Altmuller J, Palmer LJ, Fischer G, Scherb H, Wjst M (2001) Genomewide scans of complex human diseases: true linkage is hard to find. Am J Hum Genet 69: 936-950.
4. Steffens M, Lamina C, Illig T, Bettecken T, Vogler R, et al. (2006) SNP-Based Analysis of Genetic Substructure in the German Population. Hum Hered 62: 20-29.
5. Campbell CD, Ogburn EL, Lunetta KL, Lyon HN, Freedman ML, et al. (2005) Demonstrating stratification in a European American population. Nat Genet 37: 868-872.
6. Simoni L, Calafell F, Pettener D, Bertranpetit J, Barbujani G (2000) Geographic patterns of mtDNA diversity in Europe. Am J Hum Genet 66: 262-278.
7. Rootsi S, Magri C, Kivisild T, Benuzzi G, Help H, et al. (2004) Phylogeography of Y-chromosome haplogroup I reveals distinct domains of prehistoric gene flow in europe. Am J Hum Genet 75: 128-137.
8. Li T, Li Z, Ying Z, Zhang H Influence of population stratification on population-based marker-disease association analysis. Ann Hum Genet 74: 351-360.
9. Price AL, Zaitlen NA, Reich D, Patterson N New approaches to population stratification in genome-wide association studies. Nat Rev Genet 11: 459-463.
10. Sillanpaa MJ Overview of techniques to account for confounding due to population stratification and cryptic relatedness in genomic data association analyses. Heredity.
11. Wang K (2009) Testing for genetic association in the presence of population stratification in genome-wide association studies. Genet Epidemiol 33: 637-645.
12. Zhang F, Wang Y, Deng HW (2008) Comparison of population-based association study methods correcting for population stratification. PLoS One 3: e3392.
13. Lee WC, Wang LY (2009) Reducing population stratification bias: stratum matching is better than exposure. J Clin Epidemiol 62: 62-66.
14. Wjst M, Dharmage S, Andre E, Norback D, Raherison C, et al. (2005) Latitude, birth date, and allergy. PLoS Med 2: e294.
15. Chinn S, Jarvis D, Melotti R, Luczynska C, Ackermann-Liebrich U, et al. (2005) Smoking cessation, lung function, and weight gain: a follow-up study. Lancet 365: 1629-1635; discussion 1600-1621.
16. Abramson M, Kutin J, Czarny D, Walters EH (1996) The prevalence of asthma and respiratory symptoms among young adults: is it increasing in Australia? J Asthma 33: 189-196.
17. Trikalinos TA, Salanti G, Khoury MJ, Ioannidis JP (2006) Impact of violations and deviations in Hardy-Weinberg equilibrium on postulated gene-disease associations. Am J Epidemiol 163: 300-309.
18. Burger J, Kirchner M, Bramanti B, Haak W, Thomas MG (2007) Absence of the lactase-persistence-associated allele in early Neolithic Europeans. Proc Natl Acad Sci U S A 104: 3736-3741.
19. Tishkoff SA, Reed FA, Ranciaro A, Voight BF, Babbitt CC, et al. (2007) Convergent adaptation of human lactase persistence in Africa and Europe. Nat Genet 39: 31-40.
20. Sladek R, Rocheleau G, Rung J, Dina C, Shen L, et al. (2007) A genome-wide association study identifies novel risk loci for type 2 diabetes. Nature 445: 881-885.

21. Garrigan D, Hammer MF (2006) Reconstructing human origins in the genomic era. Nat Rev Genet 7: 669-680.
22. Dupanloup I, Bertorelle G, Chikhi L, Barbujani G (2004) Estimating the impact of prehistoric admixture on the genome of Europeans. Mol Biol Evol 21: 1361-1372.
23. Green RE, Krause J, Ptak SE, Briggs AW, Ronan MT, et al. (2006) Analysis of one million base pairs of Neanderthal DNA. Nature 444: 330-336.
24. Green RE, Krause J, Briggs AW, Maricic T, Stenzel U, et al. A draft sequence of the Neandertal genome. Science 328: 710-722.
25. Sabeti PC, Schaffner SF, Fry B, Lohmueller J, Varilly P, et al. (2006) Positive natural selection in the human lineage. Science 312: 1614-1620.
26. Rosenberg NA, Pritchard JK, Weber JL, Cann HM, Kidd KK, et al. (2002) Genetic structure of human populations. Science 298: 2381-2385.
27. Yang N, Li H, Criswell LA, Gregersen PK, Alarcon-Riquelme ME, et al. (2005) Examination of ancestry and ethnic affiliation using highly informative diallelic DNA markers: application to diverse and admixed populations and implications for clinical epidemiology and forensic medicine. Hum Genet 118: 382-392.
28. Seldin MF, Shigeta, R., Villoslada, P., Selmi, C., Tuomilehto, J., Silva, G., Belmont, J.W., Klareskog, L., Gregersen, P.,K. (2006) European population substructure: clustering of Northern and Southern populations. PLOS Genetics 2: e143.
29. Holsinger KE, Weir BS (2009) Fundamental concepts in genetics: Genetics in geographically structured populations: defining, estimating and interpreting FST. Nature Publishing Group 10: 639-650.
30. Freedman ML, Reich D, Penney KL, McDonald GJ, Mignault AA, et al. (2004) Assessing the impact of population stratification on genetic association studies. Nat Genet 36: 388-393.
31. Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. Genetics 155: 945-959.
32. Devlin B, Roeder K, Wasserman L (2001) Genomic control, a new approach to genetic-based association studies. Theor Popul Biol 60: 155-166.
33. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, et al. (2006) Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet 38: 904-909.
34. Frudakis T, Venkateswarlu K, Thomas MJ, Gaskin Z, Ginjupalli S, et al. (2003) A classifier for the SNP-based inference of ancestry. J Forensic Sci 48: 771-782.
35. Smith MW, Patterson N, Lautenberger JA, Truelove AL, McDonald GJ, et al. (2004) A high-density admixture map for disease gene discovery in african americans. Am J Hum Genet 74: 1001-1013.
36. Lao O, van Duijn K, Kersbergen P, de Knijff P, Kayser M (2006) Proportioning whole-genome single-nucleotide-polymorphism diversity for the identification of geographic population structure and genetic ancestry. Am J Hum Genet 78: 680-690.
37. Sanchez JJ, Phillips C, Borsting C, Balogh K, Bogus M, et al. (2006) A multiplex assay with 52 single nucleotide polymorphisms for human identification. Electrophoresis 27: 1713-1724.