Genes predict village of origin in rural Europe

Colm O'Dushlaine[1], Ruth McQuillan[2], Michael E Weale[3], Daniel JM Crouch[3], Åsa Johansson[4], Yurii Aulchenko[5], Christopher S Franklin[2], Ozren Polašek[6], Christian Fuchsberger[7], Aiden Corvin[1], Andrew A Hicks[7], Veronique Vitart[8], Caroline Hayward[8], Sarah H Wild[2], Thomas Meitinger[9,10], Cornelia M van Duijn[5], Ulf Gyllensten[4], Alan F Wright[8], Harry Campbell[2], Peter P Pramstaller[7], Igor Rudan[2,6,11], James F Wilson[2]*

Running title: Genes predict village of origin in rural Europe

1 Neuropsychiatric Genetics Research Group, Department of Psychiatry, Trinity College Dublin, Ireland.

2 Centre for Population Health Sciences, University of Edinburgh, Teviot Place, Edinburgh, EH8 9AG, Scotland.

3 Dept of Medical and Molecular Genetics, King's College London, WC2R 2LS, England.

4 Department of Genetics and Pathology, Uppsala University, SE-75185, Sweden.

5 Department of Epidemiology, Erasmus University Medical Center, Rotterdam, 3000, The Netherlands.

6 Gen-Info Ltd, Zagreb, Croatia.

7 Institute of Genetic Medicine, European Academy Bozen/Bolzano (EURAC), 39100 Italy.

8 MRC Human Genetics Unit, Institute of Genetics and Molecular Medicine, Edinburgh, EH4 2XU, Scotland.

9 Institute of Human Genetics, Helmholtz Zentrum München, German Research Centre for Environmental Health, Neuherberg, D-85764, Germany

10 Institute of Human Genetics, Klinikum rechts der Isar, Technische Universität München, Germany

11 Croatian Centre for Global Health, University of Split, 21000 Croatia.

Correspondence:

Dr James F Wilson, Centre for Population Health Sciences, University of Edinburgh, Teviot Place, Edinburgh, EH8 9AG, Scotland.

Email: jim.wilson@hgu.mrc.ac.uk

Tel: +44 131 651 1643

Fax +44 131 650 6909

Keywords:

Population structure, principal components, genome-wide genotyping

Abstract

The genetic structure of human populations is important in population genetics, forensics and medicine. Using genome-wide scans and individuals with all four grandparents born in the same settlement, we here demonstrate remarkable geographic structure across eight to thirty kilometres in three different parts of rural Europe. After excluding close kin and inbreeding, village of origin could still be predicted correctly on the basis of genetic data for 89-100% of individuals.

Introduction. High-density genome-wide scans have revealed a considerable degree of geographic structure among populations across the globe[1]. Even within Europe, the continent with the lowest among-population genetic diversity, populations separated by less than 500 km, such as the English and Irish[2], Italians from Lombardy and Tuscany[1], Finns from neighbouring regions[3] and Estonians from different counties[4] can be differentiated. However, it remains to be seen whether the populations of villages a few km apart can be distinguished.

Methods. Data are from Illumina Human Hap300 genome-wide scans. We made use of only the subset of each present day population with all four grandparents from one location and this was reduced further when exclusions were made on the basis of kinship and inbreeding. First, second and third degree relatives were removed, using genomic sharing estimates based on identity-by-state with a cut off of 0.1 (in R). We also used more stringent thresholds until no more individuals remained in each

subgroup. Principal components analysis (PCA) was performed using Eigensoft[5] and model-based clustering using Frappe[6].

We employed PCA plus linear discriminant analysis (LDA) to predict subpopulation membership using the genetic data[7]. We used all SNPs except for those on the X chromosome and regions of high linkage disequilibrium identified in Table 1 of Price[8]. PC scores were obtained according to described methods[5], substituting each SNP with the residuals produced by linear regression on the 3 previous SNPs. A double cross-validation procedure was used to correct for overfitting of the validation set, using the ratio of PC scores between the validation and training samples to calculate a scaling factor. A second validation cycle trained an LDA step which was used to classify a separate outgroup of individuals, corrected with the scaling factor. The predicted classes (using the first three principal components) were compared with the known geographic groups to calculate an error rate. Written informed consent was obtained from all subjects.

Results. Using 300,000 single nucleotide polymorphisms (SNP) and only unrelated, non-inbred individuals with all four grandparents from the same valley, village or isle, we here show genomic differentiation across eight to thirty km in three disparate areas of rural Europe, using genetic information alone (Fig 1). Principal components (PC) analysis of genomic sharing and model-based clustering (not shown) both allow separation of individuals with grandparents from each of three small Scottish isles, three alpine valleys in the north of Italy and two villages on one small island in Croatia. We employed a supervised classification approach to predict subpopulation membership. Highly reliable levels of prediction were achieved with 100%, 96% and

89% of individuals correctly classified on the basis of their genetic data for Italy, Scotland and Croatia, respectively. In each area, when individuals with grandparents from more than one village were included in PC analysis, they were scattered among the clusters, consistent with mixed origins (not shown).

Discussion. It is interesting to consider the time depth of this differentiation. By removing first, second and third degree relatives, we controlled for structure arising from mating behaviour in the last ~120 years, and so focused on the patterning arising earlier than this. The signal of structure is stronger when we include close relatives, but also persists when we use more stringent thresholds for relatedness (not shown), indicating that the patterns arise from ancient shared ancestry within the villages compared to their neighbouring subpopulations. Inbreeding and more distant shared parental ancestry will also contribute to among-population differences. We used the genomic measure F_{ROH}[9] to remove all individuals with total shared parental ancestry equivalent to one second cousin pedigree loop (F_{ROH}>0.015), whereas including inbred subjects led to more obvious structuring (not shown). Thus the observed structure in each of the populations arises partly from recent relatedness and shared parental ancestry and partly from deeper patterns of kinship within the subpopulations, overlaid with mating among them and now with immigrants.

To explore how many markers are required to recover these fine scale patterns of structure, we ranked SNPs by F_{ST} among villages and repeated the PCA analysis for the most differentiated subsets of 30,000, 10,000, 3,000 and 300 SNPs in each population. In all three populations 10,000 or more high F_{ST} SNPs recovered an essentially identical picture to that using the full data set, and even 3000 SNPs

preserved considerable separation between the villages (not shown). Using only the most discriminating 300 SNPs, little structure could be observed between the two Croatian villages, however in Scotland and Italy one of the three settlements included in each location remained completely differentiated from the other two (not shown).

The slightly lower differentiation of the Croatian villages is not surprising given the fact that they are physically the closest of those considered here, being 8 km apart, with only low hills separating them. In contrast the settlements in Scotland and Italy are separated by 15 to 30 km, of sea in the former case, and of 3000 m mountains in the latter, although there are deep connecting valleys.

Such fine-scale differentiation is consistent with the highly non-random nature of human mate choice over the millennia. The average distance between the birthplaces of spouses in rural parts of Finland, the Po valley in northern Italy and the isles of Scotland in the 19th century was ~1.5-3 km[10]. Such close endogamy was probably the norm in rural Europe due to lack of transport or economic opportunities. The breakdown of these isolates has since dramatically altered population structure[11].

The exquisite structure preserved in the genomes of people with all grandparents from the same settlement demonstrates that very detailed genetic and geographic ancestry information can be obtained by genome-wide SNP analyses. This provides novel opportunities, under certain circumstances, to predict the micro-geographic origin of an individual. Genetic association studies that include rural populations must also model this genetic structure, but it is not a barrier to gene discovery[12]. When whole genome sequences become widely available, the ability to use many more variants,

including rarer ones, to identify short shared genomic segments will perhaps allow routine identification of regional ancestries, given a suitably large and carefully collected reference sample.

Acknowledgements.

We thank the study volunteers in each population. EUROSPAN (European Special Populations Research Network) was supported by European Union FP6 grant number 018947 (LSHG-CT-2006-01947). For the MICROS study, we thank the primary care practitioners and the Department of Laboratory Medicine, Hospital of Silandro. The study was supported by the Ministry of Health and Department of Educational Assistance, University and Research of the Autonomous Province of Bolzano and the South Tyrolean Sparkasse Foundation. ORCADES was supported by the Scottish Government Chief Scientist Office and the Royal Society. DNA extractions were performed at the Wellcome Trust Clinical Research Facility (WTCRF) in Edinburgh. We acknowledge the invaluable contributions of L Anderson and the research nurses and the administrative team in Edinburgh. The Croatian study was supported through grants from the Medical Research Council UK and Ministry of Science, Education and Sport of the Republic of Croatia (number 108-1080315-0302). We thank Prof P Rudan and the staff of the Institute for Anthropological Research in Zagreb; Genotyping of the Croatian samples was carried out at the WTCRF, Edinburgh. CO'D was funded by a post-doctoral fellowship from the Irish Research Council for Science Engineering and Technology and AC from Science Foundation Ireland.

Conflict of Interest

The authors declare no conflict of interest.

References

1 Li JZ, Absher DM, Tang H, *et al*: Worldwide human relationships inferred from genome-wide patterns of variation. *Science* 2008 **319**, 1100-1104.

2 Novembre J, Johnson T, Bryc K, *et al*: Genes mirror geography within Europe. *Nature* 2008 **456**, 98-101.

3 Sabatti C, Service SK, Hartikainen AL, *et al*: Genome-wide association analysis of metabolic traits in a birth cohort from a founder population. *Nat. Genet.* 2009 **41**, 35-46.

4 Nelis M, Esko T, Mägi R *et al*. Genetic structure of Europeans: a view from the North-East. *PLOS One* 2009 **4**, e5742.

5 Patterson N, Price AL, Reich D: Population structure and eigenanalysis. *PLoS Genet.* 2006 **2**, e190.

6 Tang H, Peng J, Wang P, Risch NJ: Estimation of individual admixture: analytical and study design considerations. *Genet. Epidemiol*. 2005 **28**, 289-301.

7 Egeland T, Bøvelstad HM, Storvik GO, Salas A: Inferring the most likely geographical origin of mtDNA sequence profiles. *Ann. Hum. Genet.* 2004 **68**, 461-471.

8 Price AL, Weale ME, Patterson N, *et al*: Long-range LD can confound genome scans in admixed populations. *Am. J. Hum. Genet* 2008 **83**, 132-135.

9 McQuillan R, Leutenegger AL, Abdel-Rahman R, *et al*: Runs of homozygosity in European populations. *Am. J. Hum. Genet.* 2008 **83**, 359-372.

10 Cavalli-Sforza LL, Menozzi P, Piazza A: The History and Geography of Human Genes, Princeton, Princeton University Press, 1994.

11 Rudan I, Carothers AD, Polasek O, *et al*: Quantifying the increase in average human heterozygosity due to urbanisation. *Eur. J. Hum. Genet.* 2008 **16**, 1097-1102.

12 Vitart V, Rudan I, Hayward C, *et al*: SLC2A9 is a newly identified urate transporter influencing serum urate concentration, urate excretion and gout. *Nat. Genet.* 2008 **40**, 437-432.

Titles and legends to figures.

Fig 1. Fine-scale genetic structure in rural European populations illustrated using principal components analysis. Individuals are a subset of participants from the EUROSPAN project, with all four grandparents from the same isle (Scotland; n = 36), village (Croatia, n = 157) or valley (Italy, n = 57). Individuals with ancestry in different settlements are coloured red, blue and/or green. (**A**) Italy, (**B**) Scotland, (**C**) Croatia.