



Genes predict village of origin in rural Europe

James F Wilson, Colm O'Dushlaine, Ruth Mcquillan, Michael E Weale, Daniel J M Crouch, Åsa Johansson, Yurii Sergeevich Aulchenko, Christopher S Franklin, Ozren Polašek, Christian Fuchsberger, et al.

► To cite this version:

James F Wilson, Colm O'Dushlaine, Ruth Mcquillan, Michael E Weale, Daniel J M Crouch, et al.. Genes predict village of origin in rural Europe. European Journal of Human Genetics, 2010, 10.1038/ejhg.2010.92 . hal-00549892

HAL Id: hal-00549892

<https://hal.science/hal-00549892>

Submitted on 23 Dec 2010

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24

Genes predict village of origin in rural Europe

Colm O’Dushlaine¹, Ruth McQuillan², Michael E Weale³, Daniel JM Crouch³, Åsa
Johansson⁴, Yurii Aulchenko⁵, Christopher S Franklin², Ozren Polašek⁶, Christian
Fuchsberger⁷, Aiden Corvin¹, Andrew A Hicks⁷, Veronique Vitart⁸, Caroline
Hayward⁸, Sarah H Wild², Thomas Meitinger^{9,10}, Cornelia M van Duijn⁵, Ulf
Gyllenstein⁴, Alan F Wright⁸, Harry Campbell², Peter P Pramstaller⁷, 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.

1 8 MRC Human Genetics Unit, Institute of Genetics and Molecular Medicine,
2 Edinburgh, EH4 2XU, Scotland.

3 9 Institute of Human Genetics, Helmholtz Zentrum München, German Research
4 Centre for Environmental Health, Neuherberg, D-85764, Germany

5 10 Institute of Human Genetics, Klinikum rechts der Isar, Technische Universität
6 München, Germany

7 11 Croatian Centre for Global Health, University of Split, 21000 Croatia.

8

9 **Correspondence:**

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

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

13 Tel: +44 131 651 1643

14 Fax +44 131 650 6909

15

16 **Keywords:**

17 Population structure, principal components, genome-wide genotyping

18

1

2 **Abstract**

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

9

10

11 **Introduction.** High-density genome-wide scans have revealed a considerable degree
12 of geographic structure among populations across the globe¹. Even within Europe, the
13 continent with the lowest among-population genetic diversity, populations separated
14 by less than 500 km, such as the English and Irish², Italians from Lombardy and
15 Tuscany¹, Finns from neighbouring regions³ and Estonians from different counties⁴
16 can be differentiated. However, it remains to be seen whether the populations of
17 villages a few km apart can be distinguished.

18

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

1 subgroup. Principal components analysis (PCA) was performed using Eigensoft⁵ and
2 model-based clustering using Frappe⁶.

3

4 We employed PCA plus linear discriminant analysis (LDA) to predict subpopulation
5 membership using the genetic data⁷. We used all SNPs except for those on the X
6 chromosome and regions of high linkage disequilibrium identified in Table 1 of
7 Price⁸. PC scores were obtained according to described methods⁵, substituting each
8 SNP with the residuals produced by linear regression on the 3 previous SNPs. A
9 double cross-validation procedure was used to correct for overfitting of the validation
10 set, using the ratio of PC scores between the validation and training samples to
11 calculate a scaling factor. A second validation cycle trained an LDA step which was
12 used to classify a separate outgroup of individuals, corrected with the scaling factor.
13 The predicted classes (using the first three principal components) were compared with
14 the known geographic groups to calculate an error rate. Written informed consent was
15 obtained from all subjects.

16

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

1 89% of individuals correctly classified on the basis of their genetic data for Italy, the
2 Scottish Isles and Croatia, respectively. In each area, when individuals with
3 grandparents from more than one village were included in PC analysis, they were
4 scattered among the clusters, consistent with mixed origins (not shown).

5

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

20

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

1 preserved considerable separation between the villages (not shown). Using only the
2 most discriminating 300 SNPs, little structure could be observed between the two
3 Croatian villages, however in Scotland and Italy one of the three settlements included
4 in each location remained completely differentiated from the other two (not shown).
5 We note that these results are only indicative of the minimum number of SNPs
6 required to separate these populations, since by necessity SNPs have been selected
7 intrinsically on the basis of F_{ST} within the same dataset, rather than extrinsically from
8 other data.

9
10 The slightly lower differentiation of the Croatian villages is not surprising given the
11 fact that they are physically the closest of those considered here, being 8 km apart,
12 with only low hills separating them. In contrast the settlements in the Scottish Isles
13 and Italy are separated by 15 to 30 km, of sea in the former case, and of 3000 m
14 mountains in the latter, although there are deep connecting valleys.

15
16 Such fine-scale differentiation is consistent with the highly non-random nature of
17 human mate choice over the millennia. The average distance between the birthplaces
18 of spouses in rural parts of Finland, the Po valley in northern Italy and the isles of
19 Scotland in the 19th century was $\sim 1.5\text{--}3\text{ km}^{10}$. Such close endogamy was probably the
20 norm in rural Europe due to lack of transport or economic opportunities. The
21 breakdown of these isolates has since dramatically altered population structure¹¹.

22
23 The exquisite structure preserved in the genomes of people with all grandparents from
24 the same settlement demonstrates that very detailed genetic and geographic ancestry
25 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 discovery¹². 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.

1 9 McQuillan R, Leutenegger AL, Abdel-Rahman R, *et al*: Runs of homozygosity in
2 European populations. *Am. J. Hum. Genet.* 2008 **83**, 359-372.
3 10 Cavalli-Sforza LL, Menozzi P, Piazza A: The History and Geography of Human
4 Genes, Princeton, Princeton University Press, 1994.
5 11 Rudan I, Carothers AD, Polasek O, *et al*: Quantifying the increase in average
6 human heterozygosity due to urbanisation. *Eur. J. Hum. Genet.* 2008 **16**, 1097-1102.
7 12 Vitart V, Rudan I, Hayward C, *et al*: SLC2A9 is a newly identified urate
8 transporter influencing serum urate concentration, urate excretion and gout. *Nat.*
9 *Genet.* 2008 **40**, 437-432.

10

11 **Titles and legends to figures.**

12

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

19





