Genetic characterization of northeastern Italian population isolates in the context of broader European genetic diversity

Tõnu Esko1,2,3,28, Massimo Mezzavilla4,28, Mari Nelis1,3, Christelle Borel5, Tadeusz Debniak6, Eveliina Jakkula7, Antonio Julia8, Sena Karachanak9, Andrey Khrunin10, Peter Kisfali11, Veronika Krulisova12, Zita Aušrelė Kučinskaienė13, Karola Rehnström14, Michela Tragliola15, Liene Nikitina-Zake16, Fritz Zimprich17, Stylianos E Antonarakis5, Xavier Estivill18, Damjan Glavač19, Ivo Gut20, Janis Klovins16, Michael Krawczak21, Vaidutis Kučinskas13, Mark Lathrop22,23, Milan Macke12, Sara Marsal18, Thomas Meitinger24,25, Béla Melegh11, Svetlana Limborska10, Jan Lubinski9, Aarno Paolotie7,14, Stefan Schreiber21, Draga Toncheva9, Daniela Toniolo15, H-Erich Wichmann26,27, Alexander Zimprich17, Mait Metspalu2,3, Paolo Gasparini4,28, Andres Metspalu*1,2,3,28 and Pio D’Adamo4,28

Population genetic studies on European populations have highlighted Italy as one of genetically most diverse regions. This is possibly due to the country’s complex demographic history and large variability in terrain throughout the territory. This is the reason why Italy is enriched for population isolates, Sardinia being the best-known example. As the population isolates have a great potential in disease-causing genetic variants identification, we aimed to genetically characterize a region from northeastern Italy, which is known for isolated communities. Total of 1310 samples, collected from six geographically isolated villages, were genotyped at >145 000 single-nucleotide polymorphism positions. Newly genotyped data were analyzed jointly with the available genome-wide data sets of individuals of European descent, including several population isolates. Despite the linguistic differences and geographical isolation the village populations still show the greatest genetic similarity to other Italian samples. The genetic isolation and small effective population size of the village populations is manifested by higher levels of genomic homozygosity and elevated linkage disequilibrium. These estimates become even more striking when the detected substructure is taken into account. The observed level of genetic isolation in Friuli-Venezia Giulia region is more extreme according to several measures of isolation compared with Sardinians, French Basques and northern Finns, thus proving the status of an isolate.

European Journal of Human Genetics (2013) 21, 659–665; doi:10.1038/ejhg.2012.229; published online 19 December 2012

Keywords: population genetics; isolated population; genetic distance

INTRODUCTION

Human complex traits arise from the new mutations as well as from the interplay between existing genetic variants and exposure to environmental conditions. This means that it is desirable to study populations with decreased genetic variability, such as isolated populations1–2 and large homogeneous populations2–3 as more power is gained for genetic association mapping studies.4 Population isolates are by definition characterized by small effective population size (Ne), which results in stronger effects of random genetic drift leading to decreased genetic variability.5 These processes can be temporarily

1Estonian Genome Center, University of Tartu, Tartu, Estonia; 2Department of Biotechnology, Institute of Molecular and Cell Biology, University of Tartu, Tartu, Estonia; 3Estonian Biocentre, Tartu, Estonia; 4Medical Genetics, Department of Reproductive Sciences and Development, IRCCS-Burlo Garofolo, University of Trieste, Trieste, Italy; 5Department of Genetic Medicine and Development, University of Geneva Medical School, Geneva, Switzerland; 6International Hereditary Cancer Center and Department of Genetics, Pomeranian Medical University, Szczecin, Poland; 7Institute for Molecular Medicine Finland (FIMM) and National Institute for Health and Welfare, Helsinki, Finland; 8Unit of Investigation of Rheumatology, Vall d’Hebron Hospital, Barcelona, Spain; 9Department of Medical Genetics, Medical University of Sofia, Sofia, Bulgaria; 10Department of Molecular Bases of Human Genetics, Institute of Molecular Genetics, Russian Academy of Science, Moscow, Russia; 11Department of Medical Genetics, University of Pécs, Pécs, Hungary; 12Department of Biology and Medical Genetics, University Hospital Motol and Faculty of Medicine, Charles University Prague, Prague, Czech Republic; 13Department of Human and Medical Genetics, Vilnius University, Vilnius, Lithuania; 14Wellcome Trust Sanger Institute, Hinxton, UK; 15Division of Genetics and Cell Biology, San Raffaele Research Institute, Milano, Italy; 16Lithuanian Biomedical Research and Study Center, Riga, Latvia; 17Department of Clinical Neurology, Medical University of Vienna, Vienna, Austria; 18Center for Genomic Regulation (CRG-UPF) and CIBERESP, Barcelona, Spain; 19Department of Molecular Genetics, University of Ljubljana, Ljubljana, Slovenia; 20Centre Nacional d’Anàlisis Genòmica, Barcelona, Spain; 21PopGen Biobank, University Hospital Schleswig-Holstein, Campus Kiel, Germany; 22Commissariat à l’Energie Atomique, Institut Genomique, Centre National de Génotypage, Evry, France; 23McGill University and Genome Quebec Innovation Center, Montreal, Canada; 24Institute of Human Genetics, Helmholtz Zentrum München, German Research Center for Environmental Health, Neuherberg, Germany; 25Institute of Human Genetics, Technische Universität München, Klinikum rechts der Isar, Munich, Germany; 26Energie

2These authors contributed equally to this work.

Received 16 March 2012; revised 21 August 2012; accepted 4 September 2012; published online 19 December 2012
continuous and/or woven into a chain of discrete bottleneck events in the demographic history of an isolate. Several population isolates, like Sardinians,6 northern Finns,4 Amish7 and Icelanders8 have successfully been used in both linkage and genome-wide association studies for pinpointing DNA sequence variants for disease predispositions (OMIM database: www.omim.org and NHGRI GWAS Catalog9). Another favorable aspect of population isolates is the largely shared environment. The isolation usually arises from geographical barriers and therefore everyone is exposed to the same factors, which enables to effectively design powerful gene–environment interaction studies.3

In recent years, several studies have started to shed light on the structure of the genetic variation on global,10,11 continental,12,13 regional,1,14 ethnic group (such as Jews,15 Indians,16 Brazilians17) and country2,18,19 level, but very rarely are the studied sample representative for a whole country.2,3 The analyses on the sub-population level20,21 focus mostly on population isolates and demonstrate a decreased genetic variability within, but also elevated diversity between neighboring regions and source populations.2,15,21,22 It is important to characterize any putative population isolate for events of recent admixture and presence of sub-structure as these could disrupt the genetic homogeneity and lead to possible spurious associations in gene mapping studies.2,3 Therefore, a correct sampling strategy is fundamental.

Italians, on the example of Sardinians, are one of the most studied European populations next to Saami people and Basques where the population structure has been analyzed in depth by using both the haplotype loci (mtDNA and Y chromosome)24,25 and the autosomal markers.2,20 It is hypothesized that Italy may be enriched for population isolates because of complex demographic history and topographic variability (Italian Network of Genetic Isolates: http://www.netgene.it/ita/ingi.asp).

This study aims to test this hypothesis by genetically characterizing a hilly part of Friuli-Venezia Giulia (FVG) county located in north-eastern Italy. The region is particularly interesting, – while covering a total area of only 7858 km² several distinct dialects are spoken and several villages sport traditions and/or surnames linking them to ethnic groups further away rather than in FVG or in Italy for that matter. All in all, it seems the area is characterized by complex demographic history. For example, people in the village of Resia speak an archaic proto-Slavic language, known as Resian, but their surnames are Italian or Italianized. In the village of Illegio, not far from Resia, people speak another local language – called friuliano – of the Rhaeto-Romance language sub-family, which, during the Middle Ages, was widespread – from modern Switzerland to Slovenia. Illegio is further characterized by a limited number of surnames, what could be interpreted as evidence for marginal immigration. Another layer of specific cultural heritage is added by characteristic local symbols found engraved on local houses. The inhabitants of village Sauris speak an archaic dialect of German origin and according to legends the locals have their ancestral roots near Tyrol. Until 50 years ago, before a flood devastated the valley of origin, the main spoken language in the village of Ertro was a Latin dialect called ertzano, while current population is an admixture of the former inhabitants of Ertro and migrants from the nearby regions. The village of San Martino del Carso (SMC) is the only Italian-speaking village in the Slovenian-speaking Carso region. Finally, the village of Clauzetto is located in a remote valley where people speak friuliano.

The rich history and linguistic diversity of these six populations allows to predict elevated levels of intragroup genetic homogeneity, higher intergroup differentiation between the villages, and suggests at least some degree of ancestry with the neighboring regions of Slovenia, Germany and Italy. In this study, we analyze the genetic variation within the FVG region in order to understand (a) the relationship between the villages inhabitants and the other Italians, and Europeans as a whole, which may offer unique insight into the complex demographic history of the region, and (b) to evaluate if any of the village populations represent a genuine population isolate, as these would have a great potential in genetic epidemiological studies. In order to answer these questions, high-density genotype data of >1400 newly analyzed samples from FVG region and Slovenia were combined with several publicly available data collections.5,11,15,16,26

This enables us to directly contrast the hypothetical cultural origins of FVG region villages against the data-driven genetic similarity patterns. Furthermore, we compare the genetic diversity and genomic homogeneity found in the six FVG villages with other well-known geographical and cultural population isolates, such as Sardinians, French Basques and northern Finns.

**MATERIALS AND METHODS**

**Samples**

We genotyped 1310 samples from six geographically isolated villages in the Italian FVG region. We combined this data with the 96 newly genotyped Slovenian samples and with five published population-based collections14,15,16,26 (see Supplementary Table 1), with an emphasis on populations with European ancestry. After data quality control and exclusion of close relatives, 3091 samples from 72 populations were used for the analyses (Supplementary Table 1). This set included 739 samples of Italian ancestry from 11 populations: (1) Borbera Valley (North-West Italy) and Apulia region (South-East Italy; Carfantino)3 and (2) Sardinians, Tuscans and northern Italians from Human Genetic Diversity Panel11 and (3) six villages from the FVG region. Figures 1a and b shows the valleys and isolated villages from where the Italian samples were collected. Hereafter, the nation name is used for general populations and a village or region name for more isolated populations.

A written informed consent for participation was obtained from all newly genotyped subjects.

**Single-nucleotide polymorphism (SNP) genotypes and quality control**

The six Italian cohorts were genotyped using the Human370CNV and the Slovenian sample with the HumanOmnipExpress beadchips according to the manufacturer's protocol (Illumina Inc., San Diego, CA, USA). The published data sets used in the analysis had been genotyped with different versions of Illumina beadchips.

The genotype data from different platforms were first merged and then filtered according to the standard genotype quality control metrics using PLINK v1.07 software.27 Only the SNPs of the 22 autosomal chromosomes with minor allele frequency of >1%, Hardy–Weinberg equilibrium P-value > 10⁻⁶, and genotyping success rate >95% were included, leaving approximately 145 000 SNP markers. Only the individuals with genotyping success rate >95% were used. Cryptic relatedness was tested with the same software and from the detected relative pairs (up to the second cousins) only one was randomly chosen for the subsequent analyses. Depending on the analyses (e.g., computational optimization and bias in sample sizes) very distant ethnic groups were excluded and the population sample sizes were trimmed to 50 or 24 samples. Background linkage disequilibrium (LD) can induce bias in the principal component28 and structure-like analyses; therefore the set of 145 000 SNPs was thinned by excluding markers in strong LD (pairwise genotype correlation r² > 0.4) in a window of 200 SNPs (sliding window overlap 25 SNPs at a time) leaving approximately 101 000 SNPs for the subsequent analyses.

**Statistical analyses**

An unsupervised, maximum likelihood-based clustering algorithm assembled in ADMIXTURE29 software was applied to the European and Near-Eastern ancestry (Jews, Palestinians, Adyghes and Druzes) population samples (n = 1975) to identify the putative ancestral clusters within the samples as
using EIGENSOFT 31 package. The same software was used to calculate the entire sample. The pairwise Fst values that were further analyzed with hierarchical agglomerative clustering. The analyses were performed on two sets of samples: (1) all available samples (n = 3091), and (2) only the trimmed European and Near-Eastern ancestry populations (n = 1008). Spatial ancestry analysis (SPA) was applied on the latter set, to explicitly model the spatial distribution of each marker in order to describe the degree of population stratification in each of the putative population isolates into the European genetic background.32

Principal component ancestry informative markers (PCAIMs)33 panel was constructed of 250 SNPs (out of 101 000 SNPs) retained from the top five PCs (50 SNPs from each)17 ranked by the absolute loading scores. PCAIMs panel was the input to a non-model-based multivariate approach, a discriminant analysis of principal components (DAPCs), which is implemented into the R package adegenet ver1.3-0.36 All principal components were included to the analysis of principal components (DAPCs), which is implemented into the R package adegenet ver1.3-0.36 All principal components were included to the k-means clustering algorithm from K_{DAPC} = 1 to K_{DAPC} = 14, and the best-fitting K_{DAPC} was selected using the Bayesian information criterion. DAPC was used to assign individuals into the predicted clusters and real populations were used as identifiers. Over fitting was avoided by estimating the difference between the proportion of successful reassignments and values obtained using random grouping.

Pairwise LD extent was calculated as the genotype correlation ($r^2$) between marker pairs < 100 kb apart using the PLINK v1.07 software. A custom Perl script was applied to categorize the $r^2$ values according to inter-marker distances (0–5 kb, 5–10 kb, and so on) and mean $r^2$ was calculated for each category. Calculations were restricted only to the established population isolates, the FVG region samples, and a set of geographical reference populations (Estonians, Slovenians and Swiss).

The genomic runs of homozygosity (gROH) and the inbreeding coefficient ($F_{INH}$) were estimated using PLINK v1.07 with established parameters.34 ROH was defined as a sequence of at least 25 consecutive homozygous SNPs.
compute quantitative estimates for individual ancestry in extended LD range and elevated levels of gROH (Supplementary ancestry component was found to be highly correlated with both rather than specific ancestry. This interpretation gains further support is representing the intra-population differences in level of inbreeding particular village. The results revealed a substantial level of intra-
region, although on higher values almost all of the FVG populations

spanning at least 1500 kb, with a maximum gap of 100 kb between the adjacent SNPs and a density of SNPs within the run of no >20 kb/SNP34 (for each individual the gROH was defined as the sum of above defined genomic regions). The \( P_{in} \) was estimated for each sample based on the ratio between observed and expected number of the homozygous genotypes.

RESULTS

FVG region in the context of broader European genetic diversity

As the first step in understanding the place of the FVG region populations on the genetic canvas of Europe, we applied a model-based structure-like algorithm assembled in ADMIXTURE29 to compute quantitative estimates for individual ancestry in \( K \) hypothetical ancestral populations. The best predictive accuracy was observed for a model with \( K = 9 \) (Supplementary Figure 1). In general, the ancestry proportions distributions among FVG village samples were very similar to other populations in the geographical region, although on higher \( K \) values almost all of the FVG populations became dominated by a single component largely specific to any one particular village. The results revealed a substantial level of intra-population structure in most of the FVG region populations manifested in elevated variability in membership to the village-specific ancestry component (Figure 1c). It is likely that this structure is representing the intra-population differences in level of inbreeding rather than specific ancestry. This interpretation gains further support in subsequent analyses where the presence of the village-specific ancestry component was found to be highly correlated with both extended LD range and elevated levels of gROH (Supplementary Figure 7; see below). For further analyses, the village populations were split into sub-populations (referred as GS and IS) based on the prevalence of the village-specific ancestry component (described in Materials and Methods section) and considered as independent populations to more accurately represent the genetic diversity within FVG villages.

Next, we conducted a PCA on the European and Near-Eastern ancestry populations (on the global PCA plot11 all FVG samples clustered onto Europeans; Supplementary Figure 2A). The two first principal components accounted for 1.5% and 0.7% of the genetic variation, whereas clearly separating the population isolates, like Sardinians,22 French Basques22 and northern Finns,2 but also the more homogenous FVG sub-populations (Figure 1; Supplementary Figure 2B). A more detailed picture emerged from the SPA32 analyzes as follows: (a) SMC has the widest spread, clustering onto the German-speaking populations and partially overlapping with the Slovenians, (b) the GS sub-populations of Clauzetto, Sauris and Illegio cluster between SMC and northern Italians, from which Borbera and Tuscanos positioned farther ‘south’ and (c) GS sub-populations of Erto and Resia are shifted toward population isolates but cluster less distinctly (Figure 2). The clustering of the FVG region samples in both PCA and SPA projections were roughly representative to their geographical location.

To further minimize the bias from within- and maximize the between-group variance, we applied the DAPC35 to more precisely identify the genetically closest populations to the FVG groups. For the combined European and Near-Eastern populations, the SNP PCAIM panel had the best fit for \( K_{DAPC} = 7 \) (Supplementary Figure 3) and the DAPC clearly out-clusters the IS sub-populations from Illegio (C1), Sauris (C3) and Resia (C6), including some of the respective GS sub-population samples as well (Figure 3). The rest of the FVG populations clustered mostly with C2 (dominated by central Europeans) and C5 (dominated mainly by Italians), although a minor fraction was also assigned to C4 (dominated by northern Europeans), like Sauris and SMG, and C7 (dominated by Near-Eastern ancestry

Figure 2 Model-based mapping convergence with SPA. Label position indicates the (a) specific PC1 and PC2 coordinate values for each individual and (b) the mean PC1 and PC2 coordinate values for each population. For (a, b), the colors have a following meaning: (1) dark blue color: a homogeneous fraction of the FVG population; a blue color: more general fraction of the FVG population; a red color: other Italian samples; a violet color: Basques; an orange color: Slovenians; and green color: all other populations. For (a, b), the following population abbreviation labels are used: AT, Austrians; BA, French Basques; BG, Bulgarians; BO, Borbera; CA, Carlantino; CL, Clauzetto; CH, Swiss; CZ, Czechs; GR, Germans; ER, Erto; ES, Spaniards; FR, French; HU, Hungarians; IL, Illegio; IT, Italians; JW_A, Ashkenazy Jews; JW_S, Sephardic Jews; OR, Orcadians; RE, Resia; RO, Romanians; SA, Sardinians; SA_, Sauris; SMC, San Martino del Carso; SI, Slovenians; TU, Tuscans. The extra abbreviations: N, northern; S, southern; I, a more homogeneous sub-population; G, a more general sub-population.
samples. In population names: N, northern; S, southern; I, a more fitting cluster. The size of the rectangle scales with the number of assigned comparisons were very distant from their geographical neighbors. Genetic isolation of FVG populations was further highlighted by DAPC, which assigned three of the IS sub-populations (Illegio, Sauris and Resia) exclusively to respective single village-specific clusters, whereas all other FVG populations were distributed between several clusters (Figure 3).

As may be expected from the observed genetic structure, the between populations pair-wise Fst values of the IS sub-populations were extremely high compared with all the other populations (overall mean Fst: Resia 0.023 and Sauris 0.021 vs SMC 0.006) and elevated compared with the previously characterized population isolates (overall mean Fst: Sardinia 0.014; French Basques 0.013 and northern Finns 0.18) (Supplementary Figure 6).

The results from the LD analysis revealed a striking difference between the sub-populations drawn from a single village compared with the established population isolates, whereas even the (GS) sub-population of Resia and Illegio showed a strong deviation in LD from the expected southern decline (Figure 4). The LD block length was the shortest for the GS sub-populations of Clauzetto and Sauris, where the extent was comparable with values observed for Slovenians and Swiss.

Finally, similar large deviation from the expected values was observed for population mean gROH in IS sub-populations (Figure 5). For example, the mean gROH for the latter set was 47 Mb, compared with 11 Mb in the population isolates cluster, 5.5 Mb in the GS sub-populations and roughly 2 Mb in the European reference populations (Figure 5). The latter is similar to previously published estimates. The same tendency was observed for population mean gROH in IS sub-populations (overall mean Fst: Sardinia 0.014; French Basques 0.013 and northern Finns 0.18) (Supplementary Figure 6).

DISCUSSION

In this study, we were interested to see how the FVG village populations fit into the general paradigm of spatial structure of the genetic variation of European populations using 145,000 autosomal SNPs. To this end, we compared the genetic variation of the FVG village populations against a reference panel composed of other Italian, a wider range of European, and a set of Near-Eastern populations. We observed a striking level of genetic structure among the FVG populations (Figure 1c), which likely arises from increased genetic similarity within specific subsets of samples from respective villages and not difference in genetic origin. The FVG-specific components were present in the background profile of all Europeans, which is consistent with the FVG-specific components representing a fraction of the overall genetic variability in Europe.

Based on the composition of ancestral components the FVG populations were split into two sets for the subsequent analyses: (1) more general sub-population (GS), which resembled other Italians and populations in close proximity, and (2) more diverse and very homogeneous set (IS). The GS sets were used to further understand the genetic position of the FVG populations on the European genetic canvas and to look for the genetic roots of cultural differences. The

FVG village populations as genetic isolates

Some of the FVG villages reside in very remote valleys and this can lead to genetic isolation, so we next investigated measures of genetic diversity that might reflect a history of relative isolation, elevated levels of inbreeding and/or recent historical bottlenecks. To compare the depth of the isolation in the FVG village sub-populations, we compared the observed values with the established population isolates (like French Basques, Sardinians and northern Finns) but also with the other reference populations.

A high level of inner-structure in the FVG populations was revealed by ADMIXTURE (Figure 1c), which was further confirmed by the PCA, SPA and DAPC (Supplementary Figure 2B; Figures 2 and 3). For example, in both the PCA and SPA analyses, the IS FVG sub-populations (except Clauzetto) positioned very distant from their geographical neighbors (Supplementary Figure 2B). Genetic isolation of FVG populations was further highlighted by DAPC, which assigned three of the IS sub-populations (Illegio, Sauris and Resia) exclusively to respective single village-specific clusters, whereas all other FVG populations were distributed between several clusters (Figure 3).

Hierarchical agglomerative clustering analysis based on the between populations pair-wise Fst distances revealed that the IS sub-populations were very distant from all other populations (Supplementary Figure 4). The GS sub-populations showed very small pair-wise Fst distances from the genetically close populations (Supplementary Figure 5). Two main observations were made: (1) the SMC was equally distant from Sauris and Clauzetto as well as from its geographical neighbors, such as Slovenians, and (2) Sauris resembled Clauzetto the most but also the Swiss. All the other FVG sets resembled a larger list of populations (Supplementary Figure 5).

Like many populations, like Clauzetto, Illegio and Sauris but also other southern Europeans (Figure 3). The size of the rectangle scales with the number of assigned samples. In population names: N, northern; S, southern; I, a more homogeneous sub-population; G, a more general sub-population.
two first components in both the PCA and the SPA analyses positioned the FVG samples close within the variation of Italian populations as reported by Li et al.\(^{11}\) In addition, FVG region showed the smallest Fst distances with the populations in close geographical proximity. For example the SMC, while being an only Italian-speaking village in a Slovenian-speaking region, had equally distant Fst values from Slovenians and other FVG populations, but at the same time had a very broad clustering in the SPA and was consistent with varied level of admixture. This suggests that the original Italian founders of the village have genetically mixed with the neighboring Slovenians and thus on PC plot position in between Italians and Slovenians but have maintained their Italian cultural heritage.

From the discriminant analyses of principal components, which effectively minimized the large within-group variability,\(^{33}\) we found some hints for the origin of Sauris, the historical founders of which are believed to be German speakers from the region adjacent to Tyrol (Supplementary Note). In the DAPC, the Sauris samples positioned to several clusters, mainly Italian-specific C5 and southern European-specific C2 but also in the northern European cluster C4 (also slightly present only for SMC), together with the Germans and Swiss. The smallest Fst distance with the Swiss next to Clauzetto further strengthened the Sauris link with the German-speaking populations.

In addition to being intriguing subjects in studies on population history, linguistics and ethnogenesis, population isolates are useful tools for genetic epidemiology studies in quest for finding disease susceptibility alleles. Using isolates is advantageous because of very similar exposure to environmental factors and effects of small Ne – enrichment of some harmful sequence variants because of more pronounced random genetic drift, a higher level of consanguinity and overall genetic homogeneity.\(^{5,6}\) In this study, we detected a strong
signal for genetic isolation for subsets of the village populations that could only be discovered by analyzing a representative fraction of the village populations. The more homogeneous FVG sub-populations showed more extreme values for all considered measures of isolation compared with Sardinia,\(^{11,36}\) French Basques\(^{13,36}\) and northern Finns.\(^{2}\)

Most of the extensive genetic homogeneity is explained by the very small \(N_e\) within the FVG villages (eg, 105 inhabitants in the village of Ertos), which has led to high but varying levels of consanguinity suggested also by the trice as large inbreeding coefficients in GS compared with IS sub-population clusters. Not surprisingly also the more general sub-populations of the FVG, except Clauzetto, demonstrated values in the direction that is indicative of isolation relative to the European reference set for \(Fst\) distances,\(^{3}\) in the extent of LD\(^{4,36}\) and levels of genomic background homozgyosity.\(^{7}\)

This study has yielded two main results. We have deeply characterized and confirmed the status of genetic isolate for four further populations, which provide a useful tool for gene mapping studies. Second, the genetic structure analyses clearly highlights the need to analyze a large and representative sample to precisely estimate the intragroup variability in a population. Our results show the extent of the genetic diversity and variation within and between populations sampled from northeastern part of Italy while highlighting an extreme level of isolation compared with other genetic isolates.

**CONFLICT OF INTEREST**

The authors declare no conflict of interest.

**ACKNOWLEDGEMENTS**

EGCUT received financing from FP7 grants (ENGAGE 201413, OPENGENE 245536), Estonian Basic Research grant SF0180142s08, Estonian Research Roadmap through Estonian Ministry of Education and Research (3.2.0301.11-0312), Center of Excellence in Genomics (EXCEGEN) and Development Fund 245536), Estonian Basic Research grant SF0180142s08, Estonian Research.

EGCUT received financing from FP7 grants (ENGAGE 201413, OPENGENE 245536), Estonian Basic Research grant SF0180142s08, Estonian Science Foundation grant (8973). We acknowledge EGCUT personnel, especially Mr V Soo. Data analyzes were carried out in part in the High Performance Computing Center of University of Tartu. M Mack was supported by CZ.2.16/3.1.00/24022.

---

1 Price A, Helgason A, Palsson S et al: The effect of divergence time on the nature of population structure: an example from Iceland. *PLoS Genet* 2009; 5; e1000505.
2 Jakula E, Reinfurlo K, Varilo T et al: The genome-wide patterns of variation expose significant substructure in a founder population. *Am J Hum Genet* 2008; 83; 787–794.
3 Nelis M, Esko T, Ma¨gi R; et al. Genetic structure of Europeans: a view from the North-East. *PLoS One* 2009; 4; e4572.
4 Peltonen L, Jalkano A, Varilo T: Molecular genetics of the Finnish disease heritage. *Hum Mol Genet* 1999; 8; 1913–1923.
5 Kristiansson K, Naukkarinen J, Peltonen L: Isolated populations and complex disease gene identification. *Genome Biol* 2008; 9; 109.
6 Orri S, Goura G, Loizzi giorno A, Cox DW, Contu L: 24 bp deletion and Ala1278 to Val mutation of the ATP7B gene in a Sardinian family with Wilson disease. *Hum Mutat* 1997; 10; 84–85.
7 Pufferberger EG: Genetic heritage of the Old Order Mennonites of southeastern Pennsylvania. *Am J Med Genet C Semin Med Genet* 2003; 121C; 18–31.
8 Holm H, Gudbjartsson DF, Sulem P et al: A rare variant in MYH6 is associated with high risk of sick sinus syndrome. *Nat Genet* 2011; 43; 316–320.
9 Hindorff LA, Sethupathy P, Junkins HA et al: Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci USA* 2009; 106; 9362–9367.