Analysis of Allele and Haplotype Diversity Across 25 Genomic Regions in Three Eastern European Populations

Andrey Khrunin^{a, b} Evelin Mihailov^b Tiit Nikopendius^{b, c} Kaarel Krjutškov^{b, c}
Svetlana Limborska^a Andres Metspalu^{b–d}

^aInstitute of Molecular Genetics, Russian Academy of Sciences, Moscow, Russia; ^bInstitute of Molecular and Cell Biology, University of Tartu, ^cEstonian Biocentre, and ^dEstonian Genome Project of University of Tartu, Tartu, Estonia

Key Words

SNPs • Linkage disequilibrium • Haplotypes • Haplotype diversity • Population history

Abstract

Objective: Individual population history is the main reason for the variability of linkage disequilibrium (LD) patterns and haplotype frequencies among populations. Such diversity may influence the transferability of tag SNPs from one population to another. Our goal was to compare patterns of pairwise LD and allele and haplotype frequencies in Estonian and Russian populations, to estimate the genetic variation between populations and assess the potential transferability of tag SNPs. **Methods:** 452 SNPs from 25 unlinked genomic regions on 12 chromosomes were genotyped in 140 Estonians and 207 Russians from Northern and Western regions of the European area of Russia. **Results:** The allele frequency distributions were highly consistent between populations ($R > 0.90$ for all pairwise comparisons). The overall frequency variation among populations was low ($F_{ST} = 0.0054$). The number of SNPs with high-range F_{ST} values (0.02–0.09) was most prominent for the *MC5R* genomic region. Haplotype heterogeneity among populations was low (F_{ST} values ranged within 0.00–0.01, with the exception of haploblocks in the *ADIPOR2* and *MC5R* regions). The interpopulation

proximity was also evaluated using haplotype diversity. **Conclusion:** Our data showed a high concordance between the populations studied, which may be considered as the result of their historical formation on a cognate genetic basis.

Copyright © 2009 S. Karger AG, Basel

Introduction

Linkage disequilibrium (LD) plays a central role in human genetics. One of the main reasons for conducting detailed LD studies is the uniformization of the design of association studies [1]. Dense genome-wide haplotype maps for populations of European, West African, and East Asian descent have been reported by the International HapMap Project [2, 3] and by Perlegen Sciences [4]. These haplotype maps are considered as important resources for the design of large-scale SNP-based studies, which provide detailed information about haplotype structure. This information can be used to select tag SNPs corresponding to neighboring markers or to a set of com-

A. Khrunin and E. Mihailov contributed equally to this work.

Fig. 1. Geographic locations of the populations tested.

mon haplotypes in a genomic region with high LD. A practical issue for association studies is whether tag SNPs selected using the HapMap reference population data adequately capture patterns of variation in other populations, especially in populations within the same broad geographic area. Several recent studies suggest that the HapMap data are a valuable resource for the selection of tag SNPs for additional populations [5–7], and that the same is true globally [8, 9]. In contrast, other studies have shown and emphasized that LD patterns may vary substantially among populations and ethnic groups, even in populations with a similar geographic origin [10, 11]. Furthermore, authors have concluded that case-control studies require empirical characterization of the extent of LD and haplotype structure for each region or gene of interest [12–15]. These inferences reflect the fact that variation of LD patterns and underlying haplotype frequencies is a function of the history of any given population, and that genetically distinct populations may differ in both the extent of LD and the frequencies of haplotypes [16–18].

Several studies have addressed the relationship between the Estonian population and several other European populations, including the HapMap populations [6, 15]. Here, we compared the Estonian and Russian populations, based on the analysis of SNP allelic variation, LD structure, and haplotype diversity within 25 genomic regions. We genotyped 452 SNPs in 140 individuals from Estonia and in two groups of Russian individuals (207 individuals). Taking into account that Russia is a large and diverse country, we used two populations from

northwestern Russia, geographically close to Estonia: one from Mezen (northern Russia) and one from Andreapol (western Russia) (fig. 1). The populations chosen also seem to be relevant to the whole genome association (WGA) studies conducted using samples from the Estonian Biobank, in which about 15% of the samples were collected from individuals of Russian descent (www.geenivaramu.ee). This comparative investigation is of interest from genetic and historical perspectives, as it aimed not only to assess the potential transferability of tag SNPs (LD patterns) across these populations within regions of interest, but also to analyze evolutionary relationships between Estonians and Russians. Linguistically, Russians and Estonians belong to distinct communities (Balto-Slavic and Finno-Ugric, respectively). On the other hand, the impact of Finno-Ugric communities on the ethnogenesis of Russians is well documented by anthropological, archaeological, and linguistic data [19, 20]. Here, both Russian populations selected resided in areas that were previously inhabited by ancient Finno-Ugric tribes. The settlement of these territories by Slavs was accompanied by active contacts with local populations and their subsequent assimilation by growing Slavic populations; however, the extent of the assimilation might differ from one region to another. Our Y chromosome microsatellite polymorphism data revealed significant differences between Russians from the Archangelsk region (Northern Russians) and other Russian populations [21]. Taking into account the historical and geographical specificities of the colonization of Northern European territories, it has been proposed that Northern Russians may have preserved more Finno-Ugric components when compared with Southern Russians. Our data demonstrate that Estonians and populations from northwestern areas of Russia, especially from the Andreapol region, display relatively low levels of genetic divergence.

Materials and Methods

Population Samples

Selected common SNPs from dbSNP were investigated in three populations. DNA samples from 140 unrelated Estonians were selected from a collection of 1,090 samples from the Biobank of the Estonian Genome Project (www.geenivaramu.ee). Because no population substructures were detected within the Estonian people [6], five males and five females were randomly selected from each of the 15 Estonian counties (Harju, Ida-Viru, Jõgeva, Järva, Lääne-Viru, Põlva, Pärnu, Rapla, Saaremaa, Tartu, Valga, Viljandi, and Võru; Hiiumaa and Läänemaa counties were combined). For the Russian populations, blood samples were obtained with informed consent from healthy donors of two geographi-

cally diverse locations in the European regions of Russia: the northern Mezen district of the Archangelsk region (116 samples) and the western Andreapol district of the Tver region (91 samples) (fig. 1). All individuals were unrelated and represented the native ethnic group of the regions studied (i.e., they belonged to at least the third generation living in a particular geographic region). DNA was isolated from peripheral leukocytes following standard techniques that use proteinase K treatment and phenol–chloroform extraction [22].

Genomic Regions and SNP Marker Selection

We analyzed 25 unlinked genomic regions on 12 chromosomes. The pairwise LD method was applied in this study, according to an algorithm for tag SNP selection [23]. For this, we used the Tagger program included in Haploview, according to the CEU population data from the HapMap project: the minor allele frequency threshold was set at 5%, the r^2 threshold was set at 0.8, and the LOD threshold for multimarker tests was set at 3.0.

From the 520 SNPs initially chosen and genotyped, 68 were excluded from further analysis because of insufficient quality parameters: 58 SNPs had low call rates or failed genotype calling (false signals at unexpected positions), 6 SNPs were monomorphic in at least one of the populations tested, and 4 SNPs were not in Hardy-Weinberg equilibrium (HWE; $p < 0.0001$). In total, 452 SNPs were successfully genotyped in 140 Estonian and in 207 Russian subjects. The detailed information on selected genomic regions and SNPs is listed in online suppl. table 1 (for online supplemental material, see www.karger.com/doi/10.1159/000210447).

Genotyping

SNP genotyping was performed according to the principles of the modified APEX method described in Krjutskov et al. [24], in which an initial APEX assay [25] was further developed to achieve SNP-specific primer extension on genomic DNA using APEX-2 primers carrying universal PCR primer binding sites at their 5' ends, ending with the amino link (the same primers were used for spotting the oligonucleotide arrays for APEX-2). In the first round of amplification, we carried out six individual multiplex PCR reactions using 50 ng of genomic DNA in a 20 μ l volume to generate a complete set of SNP-specific amplicons. In the second round of amplification, the PCR products from the first round were further amplified using universal primers in a 100 μ l volume including 20 μ l of first-phase PCR product. 250 μ l of pooled PCR product from the universal primer amplification was treated with shrimp alkaline phosphatase and exonuclease I, followed by sample concentration, hybridization to the microarray, and single-base extension reactions. After the APEX-2 reaction, SNPs were detected using the Genorama Quattroimager™ four-channel microarray imaging system (Asper Biotech Ltd., Estonia). All technical details for amplification conditions and APEX-2 procedures are described in online supplement X.

Statistical Analyses

Pearson's correlation coefficient (referred to herein as R , to distinguish it from the linkage disequilibrium measure r^2) was used to measure correlations in allele and haplotype frequencies, linkage disequilibrium parameters, and haplotype diversity between the populations studied [26]. All correlation analyses were performed using the STATISTICA software, version 6.0 (<http://www.statsoft.com/>).

Minor allele frequencies were estimated in Estonian subjects and compared with Russian subjects using Pearson's χ^2 test. For markers with a high extent of heterozygosity, Estonian minor allele configurations were used as reference data. For each marker used in the diversity analysis, the genotyping was complete for greater than 98% of the individuals in all populations, with an overall call rate of 99.5%. The Haploview version 3.32 [27] software was used to calculate χ^2 statistics for differences in allele frequency and HWE. A Bonferroni-adjusted nominal p value threshold of 1.1×10^{-4} was applied to correct for multiple testing, assuming 452 effective independent marker loci.

For each SNP and for SNP groups across genomic regions, we also calculated Wright's fixation indices, F_{ST} , to estimate the extent of allelic variation due to genetic differences within populations, relative to interpopulation genetic differences. F_{ST} values were calculated from our data using the algorithm of Weir and Cockerham [28, 29] by comparing all three populations as well as three pairwise population combinations. All of the above-mentioned F_{ST} calculations were performed using the PowerMarker software package (v.3.0) with 95% confidence intervals based on 1,000 bootstraps across loci [30]. Haplotype analysis was performed for all 25 genomic regions. Haplotypes were inferred and pairwise linkage disequilibrium statistics (D' and r^2) were calculated for adjacent markers using an accelerated expectation-maximization (EM) algorithm embedded in the Haploview software [27]. The haplotype block patterns in all three populations were defined using block definition based on the LD measure D' and its confidence interval [31].

Analysis of molecular variance (AMOVA, an option of the Arlequin package (v.3.1)) was used to test for heterogeneity within and among populations on the basis of estimated haplotype frequencies for individual LD blocks [32]. In addition, pairwise F_{ST} statistics based on haplotype frequencies were also used to compare the three populations studied. The statistical significance of F_{ST} values was evaluated using the permutation procedure (10,000 permutations), where significance level was set at $p < 0.05$. For each block in each sample, haplotype diversity levels were also calculated and tested in pairwise comparisons for correlations.

Results

Comparison of Allelic Variation

452 SNPs, corresponding to 25 regions on 12 chromosomes, were genotyped in Estonian samples, as well as in Russian subjects from the Mezen (Archangelsk) and Andreapol (Tver) regions. Estimated allele frequencies for all 452 SNPs are provided in online supplemental table 2. We compared allele frequencies between populations using the minor allele configurations determined in Estonians as the reference alleles (fig. 2A–C). The allele frequency distributions were highly consistent between populations ($R > 0.90$ and $p < 10^{-6}$ for all pairwise comparisons). Fourteen SNPs showed large differences in allele frequencies between the Estonian and Mezen populations, and remained significantly different after correc-

Fig. 2. Comparison of allele frequency variations. **A** Estonia versus Mezen data ($R = 0.904$, $F_{ST} = 0.0078$); **B** Estonia versus Andreapol data ($R = 0.940$, $F_{ST} = 0.0022$); **C** Andreapol versus Mezen data ($R = 0.903$, $F_{ST} = 0.0058$). SNPs that showed large differences in allele frequencies between populations and remained significant after correction for multiple testing are depicted by triangles.

tion for multiple testing ($p < 0.0001$). In contrast, allelic frequency comparison between the Estonian and Andreapol populations revealed only one significantly different SNP (fig. 2A and B, respectively). Comparison of the Mezen and Andreapol populations showed five SNPs with significant differences in allele frequency (fig. 2C).

To assess the degree of variation between populations shown by differences in allele frequency, we calculated F_{ST} values for each SNP, each genomic region, the entire SNP set across all populations, and each pairwise population comparison. The population-averaged single-locus F_{ST} values for the majority (79%) of SNPs were ≤ 0.01 . Only a minor fraction (approximately 3%) of SNPs showed F_{ST} values ≥ 0.04 ; the highest value, 0.087, was obtained for SNP 536 (rs7236531) at the *MC5R* genomic region (online supplemental table 2). In general, the entire set of 452 SNPs had a mean F_{ST} value of 0.0054, indicating a relatively low level [33] of population differentiation. As for variations between populations, a minimal mean F_{ST} value of 0.0022 was observed when comparing Estonians with Andreapol Russians. In two other pairwise comparisons, the mean F_{ST} values were

higher: 0.0058 for Andreapol Russians versus Mezen Russians, and 0.0078 for Estonians versus Mezen Russians. These two population pairs were characterized by having a smaller fraction of SNPs with pairwise F_{ST} values ≤ 0.01 (77 and 75%, respectively). When comparing Estonians with Andreapol Russians, the fraction of such SNPs was 87%. To facilitate the analysis of variation between populations, we also compared F_{ST} values across the genomic regions analyzed (fig. 3 and online supplemental table 3). The majority of the region-specific F_{ST} values in the comparison of Andreapol Russians with Estonian samples were low, and the F_{ST} values were frequently lower than the population-averaged values. The most significant pairwise differences, including a peak at the *MC5R* genomic region, were found between Mezen Russian and Estonian samples.

Comparison of Haplotypes

We measured the extent of pairwise LD between SNPs within each of the 25 genomic regions with at least five genotyped SNPs, and compared the haplotype structures observed within populations. The general patterns

Fig. 3. Region-specific F_{ST} values (population-averaged and pairwise comparisons).

Fig. 4. Example of two haploblocks inferred for the *MC5R* genomic region (A = Andreapol, E = Estonians, and M = Mezen).

of LD were very similar across samples. Pairwise comparisons based on r^2 disequilibrium coefficients [11] demonstrated a high correlation in LD strength ($R = 0.91-0.98$, $p < 10^{-6}$) between populations across the majority

of the genomic regions. Weaker correlations in interpopulation LD magnitude ($R = 0.79-0.90$, $p < 10^{-6}$) were observed for the *GHRL* and *PPARGC1A* genomic regions and in some pairwise comparisons.

A total of 73, 83, and 66 haploblocks were inferred in the Andreapol, Mezen, and Estonian samples, respectively, and most of them were either shared or identical among the three populations; however, the majority of these blocks (55–64% per population) were two-locus blocks. In order to analyze the allele frequency variation between populations in a more complex manner, we extended our blocks by including neighboring SNPs (threshold pairwise $D' \geq 0.7$). By doing this, the differences in the strength of LD could be generated by different frequencies of the same few haplotypes, and similar amount of LD could be generated by different sets of haplotype frequencies [10]. Therefore, 40 haploblocks composed of 3–13 SNPs were inferred. All haploblocks were the same in all populations (fig. 4), and the haplotype frequencies were estimated for each block in each population. The distribution of haplotypes with at least four SNPs per block and with inferred frequency values of more than 1% in at least one of the three populations displayed very similar patterns (fig. 5). For the most part, 3–5 common haplotypes per block represented approximately 90–95% of the total diversity observed among the populations; however, in some cases – most prominently for a block in the *IL6* genomic region – the number of low-frequency haplotypes (less than 5%) was notably higher, reaching 6–8 haplotypes or even more. The existence of many low-frequency haplotypes may be due to increased local recombination rates, which is true for *IL6*, *CARTPT*, and some other genomic regions (www.hapmap.org).

Haplotype frequencies across all 40 blocks were strongly positively correlated between populations with a higher concordance between Estonian and Andreapol Russian samples (mean $R = 0.98$, $p < 10^{-6}$). Slightly lower correlation values of 0.95 and 0.96 ($p < 10^{-6}$) were found between Andreapol and Mezen, and Mezen and Estonian haplotypes, respectively. For 83% of the haploblocks ($n = 33$), the pairwise correlation values for population haplotype frequencies were greater than 0.90. Lower R values (0.72–0.90) were mostly specific to the pairwise comparisons, and were found when Andreapol and Mezen populations or Mezen and Estonian populations were compared.

Haplotype variability among populations was assessed using F_{ST} values. The AMOVA percent variance in haplotype frequencies among populations varied from close to zero for blocks in the *POMC*, *ADIPOQ*, *GAD2*, *AGRP*, *MC4R*, and *LEP* genomic regions, to 1.5% for *ADIPOR2*, and was highest for the *MC5R* region (3.6%). The *ADIPOR2* and *MC5R* regions may serve as an example of factors influencing the outcomes of WGA studies performed in diverse populations in which disease risk might

be associated with population origin and/or stratification. Pairwise comparisons among our three populations demonstrated noticeable variations in F_{ST} values, generating two distinct groups: the first group consisted of the Andreapol/Mezen and Mezen/Estonian comparisons, and the second group consisted of the Andreapol/Estonian comparison (fig. 6 and online supplemental table 4). The first group of comparisons (Andreapol/Mezen and Mezen/Estonian) showed significant differences in many (>40%) cases, whereas the differences between Andreapol and Estonian samples were significant in only four cases (10%). This suggests that the Andreapol and Estonian populations are genetically more similar than the other population pairs; however, it should be noted that, in general, the majority of mean and pairwise F_{ST} values for all comparisons were ≤ 0.01 , indicating a quite low level of population differentiation.

Finally, the interpopulation proximity was also evaluated through haplotype diversity across blocks. A higher concordance was observed between Estonians and Andreapol Russians ($R = 0.96$; $p < 10^{-6}$) relative to the other pairwise comparisons ($R = 0.88$ for Andreapol versus Mezen Russians and $R = 0.91$ for Estonians versus Mezen Russians).

Discussion

The variation in both LD patterns and haplotype frequencies within and across populations has a significant impact on tag SNP selection for candidate gene or whole-genome association studies. When different populations exhibit the same or similar patterns of pairwise LD, it is relatively easy to identify tag SNPs capturing the most common haplotypes, and to use them to perform case-control or family-based association studies; however, whenever haplotype frequencies vary considerably across populations, it becomes more difficult to ensure adequate genomic coverage by predicting which SNPs will capture the existing haplotypes in all subpopulations, thereby increasing the chance of spurious findings due to confounding effects [11, 14, 34]. Several recent studies based mainly on HapMap data agree that information obtained for one population (e.g., HapMap CEU) is a valuable resource for identifying tag SNPs for additional populations of similar (e.g., European) descent [6, 7, 15]. In contrast, other studies have shown that LD patterns vary substantially among populations, and that these populations exhibit substantial genetic diversity (including haplotype diversity) as measured by Wright's fixation index,

Fig. 5. Haplotype frequencies in selected haploblocks in the populations tested (A = Andreapol, E = Estonians, and M = Mezen). Patterns represent different haplotypes from each block composed of at least four SNPs and with inferred frequency values >1% in at least one of the three populations.

F_{ST} [10, 12, 34]. Furthermore, for genetically similar (related) populations, a higher portability of tag SNPs has been demonstrated [5, 9, 18]. Taken together, individual population history and geographic variation may challenge the usefulness of a single reference population for the selection of tag SNPs for association studies, if geo-

graphic or ethnic variations in allele and haplotype frequencies are prevalent.

The goal of our study was to compare LD patterns, SNP variation, and haplotype variation in 25 genomic regions between population samples belonging to two different ethnic groups, whose population history may be consid-

Fig. 6. Haplotype-based F_{ST} values (population-averaged and pairwise comparisons).

ered quite complex, to assess the potential transferability of tag SNPs (LD patterns) across these populations. Among the 452 SNPs tested, allele distributions were highly correlated between populations. Only a small fraction of SNPs demonstrated large differences in their allele frequencies between populations, and remained significantly different after correction for multiple testing. The majority of these SNPs contributed mainly to the remarkably higher region-specific and haplotype-based F_{ST} values. SNPs in the *MC5R* genomic region were the most prominent among them. Taking into account the results of the pairwise comparisons, we can conclude that their ability to differentiate the populations was correlated with SNP allele frequencies in the Mezen sample; however, none of the F_{ST} values observed was extreme, when compared with the distribution of F_{ST} statistics provided by genome-wide studies [35, 36]. Although the possibility of the presence of selective pressure should not be fully excluded, this may be considered to be a result of random genetic drift associated with the historical and geographical specificities of population settlement in the Northern European territory of Russia [21, 37] (also see below).

Pairwise comparisons of LD magnitude (in terms of r^2) across genomic regions showed that the overall patterns of LD were highly concordant among the three populations studied, and that only two regions and a few pairwise comparisons showed a weaker concordance in LD strength. Taking into account the average SNP density across other regions, one can propose that the differences are region-specific and should be considered in

future studies. Likewise, region-specific F-statistics, haplotype-based mean F_{ST} values did not demonstrate substantial (>1%) differences in haplotype frequencies among populations for the majority of blocks analyzed; however, the haplotype-based calculations allowed us to identify one additional region (*ADIPOR2*) that might be important for the estimation of population differentiation. At the same time, it should be noted that analysis of the phylogenetic relationships among the most frequent (>5%) haplotypes for the five-loci *MC5R* block (termed *MC5R 1* in fig. 6), which had the highest locus- and haplotype-based F_{ST} values, did not reveal any haplotypes that might be described as highly specific for one of the populations or for any part of the median-joining network (data not shown). This leads us to propose that these haplotypes are quite ancient. The interpopulation proximity was also evaluated through haplotype diversity across blocks, and showed a high concordance between populations.

Our results suggest a high degree of tag SNP portability between Estonians and Russians, and lead us to predict that possible confounding effects during association studies using the genomic regions analyzed are likely to be minimal. Furthermore, the observed genetic similarity of the three populations used in this study may be due to their historical formation on a cognate genetic basis (Finno-Ugrian and/or Baltic), or may even have originated in more ancient times, when the populations that inhabited the territories to the south/southeast of the Baltic Sea coast existed as a global regional community [38, 39]. The higher concordance between Andreapol Russians

and Estonians, supported in all of our analyses, is in good agreement with the evidence of historical migrations of Estonian people to the neighboring territories of Russia during the 18th and 19th centuries [40]; some western/northwestern districts of the Tver region, including the current Andreapol region, were included in these territories. We propose that a similar explanation may underlie the specificity of Pskov Russians found by Malyarchuk et al. [41] (geographically, the Pskov region lies to the west of the Andreapol region and shares borders with Estonia and Latvia). With regard to Northern Russians, their mainly Y chromosome-based proximity to the Finno-Ugric- and Baltic-speaking populations has been described [21, 37, 41].

Acknowledgements

This study was financially supported by the Estonian Ministry of Education and Science grant 0182582s03 (A. Metspalu, P.I.), Enterprise Estonia grant EU19955, and EU FP6 project MolTools grant LSHG-CT-2004-503155. This research was also supported by the Russian Ministry of Science and Education (FTSNTP contract No. 02.444.11.7109), the Russian Academy of Sciences, Programs 'Molecular and Cell Biology', 'Basic Science for the Advantage of Medicine,' and the Russian Basic Research Foundation.

References

- 1 The International HapMap Consortium: The International HapMap Project. *Nature* 2003;426:789–796.
- 2 International HapMap Consortium: A haplotype map of the human genome. *Nature* 2005;437:1299–1320.
- 3 International HapMap Consortium: A second generation human haplotype map of over 3.1 million SNPs. *Nature* 2007;449:851–861.
- 4 Hinds DA, Stuve LL, Nilsen GB, Halperin E, Eskin E, Ballinger DG, Frazer KA, Cox DR: Whole-genome patterns of common DNA Variation in three human populations. *Nature* 2005;307:1072–1179.
- 5 Nejentsev S, Cooper JD, Godfrey L, Howson JM, Rance H, Nutland S, Walker NM, Guja C, Ionescu-Tirgoviste C, Savage DA, Undlien DE, Rønningen KS, Tuomilehto-Wolf E, Tuomilehto J, Gillespie KM, Ring SM, Strachan DP, Widmer B, Dunger D, Todd JA: Comparative high-resolution analysis of linkage disequilibrium and tag single nucleotide polymorphisms between populations in the vitamin D receptor gene. *Hum Mol Genet* 2004;13:1633–1639.
- 6 Montpetit A, Nelis M, Laflamme P, Magi R, Ke X, Remm M, Cardon L, Hudson TJ, Metspalu A: An evaluation of the performance of tag SNPs derived from HapMap in a Caucasian population. *PLoS Genet* 2006;2:e27.
- 7 Willer CJ, Scott LJ, Bonnycastle LL, Jackson AU, Chines P, Pruim R, Bark CW, Tsai YY, Pugh EW, Doheny KF, Kinnunen L, Mohlke KL, Valle TT, Bergman RN, Tuomilehto J, Collins FS, Boehnke M: Tag SNP selection for Finnish individuals based on the CEPH Utah HapMap database. *Genet Epidemiol* 2006;30:180–190.
- 8 de Bakker PI, Burt NP, Graham RR, Guiducci C, Yelensky R, Drake JA, Bersaglieri T, Penney KL, Butler J, Young S, Onofrio RC, Lyon HN, Stram DO, Haiman CA, Freedman ML, Zhu X, Cooper R, Groop L, Kolonel LN, Henderson BE, Daly MJ, Hirschhorn JN, Altshuler D: Transferability of tag SNPs in genetic association studies in multiple populations. *Nat Genet* 2006;38:1298–1303.
- 9 Conrad DF, Jakobsson M, Coop G, Wen X, Wall JD, Rosenberg NA, Pritchard JK: A worldwide survey of haplotype variation and linkage disequilibrium in the human genome. *Nat Genet* 2006;38:1251–1260.
- 10 Sawyer SL, Mukherjee N, Pakstis AJ, Feuk L, Kidd JR, Brookes AJ, Kidd KK: Linkage disequilibrium patterns vary substantially among populations. *Eur J Hum Genet* 2005;13:677–686.
- 11 Evans DM, Cardon LR: A comparison of linkage disequilibrium patterns and estimated population recombination rates across multiple populations. *Am J Hum Genet* 2005;76:681–687.
- 12 Bonnen PE, Wang PJ, Kimmel M, Chakraborty R, Nelson DL: Haplotype and linkage disequilibrium architecture for human cancer-associated genes. *Genome Res* 2002;12:1846–1853.
- 13 Crawford DC, Carlson CS, Rieder MJ, Carington DP, Yi Q, Smith JD, Eberle MA, Kruglyak L, Nickerson DA: Haplotype diversity across 100 candidate genes for inflammation, lipid metabolism, and blood pressure regulation in two populations. *Am J Hum Genet* 2004;74:610–622.
- 14 Ramírez-Soriano A, Lao O, Soldevila M, Calafell F, Bertranpetit J, Comas D: Haplotype tagging efficiency in worldwide populations in CTLA4 gene. *Genes Immun* 2005;6:646–657.
- 15 Mueller JC, Lohmussaar E, Mägi R, Remm M, Bettecken T, Lichtner P, Biskup S, Illig T, Pfeufer A, Luedemann J, Schreiber S, Pramstaller P, Pichler I, Romeo G, Gaddi A, Testa A, Wichmann HE, Metspalu A, Meitinger T: Linkage disequilibrium patterns and tagSNP transferability among European populations. *Am J Hum Genet* 2005;76:387–398.
- 16 Zhang K, Akey JM, Wang N, Xiong M, Chakraborty R, Jin L: Randomly distributed crossovers may generate block-like patterns of linkage disequilibrium: an act of genetic drift. *Hum Genet* 2003;113:51–59.
- 17 Bertranpetit J, Calafell F, Comas D, Gonzalez-Neira A, Navarro A: Structure of linkage disequilibrium in humans: genome factors and population stratification. *Cold Spring Harb Symp Quant Biol* 2003;68:79–88.
- 18 Gu S, Pakstis AJ, Li H, Speed WC, Kidd JR, Kidd KK: Significant variation in haplotype block structure but conservation in tagSNP patterns among global populations. *Eur J Hum Genet* 2007;15:302–312.
- 19 Alexeeva TI: Ethnogenesis of Eastern Slavs. Moscow, Moscow State University, 1973 (book in Russian).
- 20 Sedov VV: Origin and Early History of Slavs. Moscow, Nauka, 1979 (book in Russian).
- 21 Khrunin AV, Bebiakova NA, Ivanov VP, Solodilova MA, Limborskaia SA: Polymorphism of Y-chromosomal microsatellites in Russian populations from the northern and southern Russia as exemplified by the populations of Kursk and Arkhangel'sk Oblast. *Genetika* 2005;41:1125–1131.
- 22 Milligan BG: Total DNA isolation; in Hoelzel AR (ed): *Molecular Genetic Analysis of Populations*. London, Oxford University Press, 1998, pp 29–60.

- 23 Carlson CS, Eberle MA, Rieder MJ, Yi Q, Kruglyak L, Nickerson DA: Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. *Am J Hum Genet* 2004;74:106–120.
- 24 Krjutskov K, Andreson R, Mägi R, Nikopensius T, Khrunin A, Mihailov E, Tammekivi V, Sork H, Remm M, Metspalu A: Development of a single tube 640-plex genotyping method for detection of nucleic acid variations on microarrays. *Nucleic Acids Res* 2008;36:e75.
- 25 Kurg A, Tõnisson N, Georgiou I, Shumaker J, Tollett J, Metspalu A: Arrayed primer extension: solid-phase four-color DNA resequencing and mutation detection technology. *Genet Test* 2000;4:1–7.
- 26 Ribas G, González-Neira A, Salas A, Milne RL, Vega A, Carracedo B, González E, Barroso E, Fernández LP, Yankilevich P, Robledo M, Carracedo A, Benítez J: Evaluating HapMap SNP data transferability in a large-scale genotyping project involving 175 cancer-associated genes. *Hum Genet* 2006;118:669–679.
- 27 Barrett JC, Fry B, Maller J, Daly MJ: Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* 2005;21:263–265.
- 28 Weir BS, Cockerham CC: Estimating F-statistics for the analysis of population structure. *Evolution* 1984;38:1358–1370.
- 29 Weir BS: Genetic data analysis. Sunderland, Sinauer, 1996.
- 30 Liu K, Muse SV: PowerMarker: an integrated analysis environment for genetic marker analysis. *Bioinformatics* 2005;21:2128–2129.
- 31 Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, Blumenstiel B, Higgins J, DeFelice M, Lochner A, Faggart M, Liu-Cordero SN, Rotimi C, Adeyemo A, Cooper R, Ward R, Lander ES, Daly MJ, Altshuler D: The structure of haplotype blocks in the human genome. *Science* 2002;296:2225–2229.
- 32 Excoffier L, Laval G, Schneider S: Arlequin ver. 3.0: An integrated software package for population genetics data analysis. *Evolutionary Bioinformatics Online* 2005;1:47–50.
- 33 Hartl DL, Clark AG: Principles of Population Genetics. Sunderland, Sinauer, 2007, pp 281–285.
- 34 Beaty TH, Fallin MD, Hetmanski JB, McIntosh I, Chong SS, Ingersoll R, Sheng X, Chakraborty R, Scott AF: Haplotype diversity in 11 candidate genes across four populations. *Genetics* 2005;171:259–267.
- 35 Akey JM, Zhang G, Zhang K, Jin L, Shriver MD: Interrogating a high-density SNP map for signatures of natural selection. *Genome Res* 2002;12:1805–1814.
- 36 Kidd KK, Pakstis AJ, Speed WC, Kidd JR: Understanding human DNA sequence variation. *J Hered* 2004;95:406–420.
- 37 Balanovsky O, Rootsi S, Pshenichnov A, Kivisild T, Churnosov M, Evseeva I, Pocheshkhova E, Boldyreva M, Yankovsky N, Balanovska E, Villems R: Two sources of the Russian patrilineal heritage in their Eurasian context. *Am J Hum Genet* 2008;82:236–250.
- 38 Alexeeva TI: Ethnogenesis and ethnic history of Eastern Slavs; in Alexeeva TI (ed): Eastern Slavs. Anthropology and Ethnic History. Moscow, Scientific World, 1999, pp 307–315 (book in Russian).
- 39 Wiik K: Who are the Finns? in Suominen S, et al (eds): A Man of Measure: Festschrift in Honour of Fred Karlsson on his 60th birthday. Turku, The Linguistic Association of Finland, 2006, pp 97–108 (Special Supplement to SKY Journal of Linguistics, Vol. 19).
- 40 Manakov AG: Geocultural space for North-West of Russian plane: dynamics, structure, hierarchy. Pskov, Vozrozhdenie, 2002 (book in Russian).
- 41 Malyarchuk B, Derenko M, Grzybowski T, Lunkina A, Czarny J, Rychkov S, Morozova I, Denisova G, Miścicka-Sliwka D: Differentiation of mitochondrial DNA and Y chromosomes in Russian populations. *Hum Biol* 2004;76:877–900.