

Available online at www.sciencedirect.com

SCIENCE @ DIRECT®

Gene xx (2006) xxx–xxx

GENE

SECTION
EVOLUTIONARY GENOMICSwww.elsevier.com/locate/gene

Regional differences among the Finns: A Y-chromosomal perspective

Tuuli Lappalainen^{a,c,*}, Satu Koivumäki^{b,1}, Elina Salmela^a, Kirsi Huoponen^c,
Pertti Sistonen^d, Marja-Liisa Savontaus^{b,c}, Päivi Lahermo^a

^a Finnish Genome Center, University of Helsinki, Finland

^b Department of Biology, Laboratory of Genetics, University of Turku, Finland

^c Department of Medical Genetics, University of Turku, Finland

^d Finnish Red Cross Blood Transfusion Center, Helsinki, Finland

Received 15 August 2005; received in revised form 10 March 2006; accepted 12 March 2006

Received by J.G. Zhang

Abstract

Twenty-two Y-chromosomal markers, consisting of fourteen biallelic markers (YAP/DYS287, M170, M253, P37, M223, 12f2, M9, P43, Tat, 92R7, P36, SRY-1532, M17, P25) and eight STRs (DYS19, DYS385a/b, DYS388, DYS389I/II, DYS390, DYS391, DYS392, DYS393), were analyzed in 536 unrelated Finnish males from eastern and western subpopulations of Finland. The aim of the study was to analyze regional differences in genetic variation within the country, and to analyze the population history of the Finns. Our results gave further support to the existence of a sharp genetic border between eastern and western Finns so far observed exclusively in Y-chromosomal variation. Both biallelic haplogroup and STR haplotype networks showed bifurcated structures, and similar clustering was evident in haplogroup and haplotype frequencies and genetic distances. These results suggest that the western and eastern parts of the country have been subject to partly different population histories, which is also supported by earlier archaeological, historical and genetic data. It seems probable that early migrations from Finno-Ugric sources affected the whole country, whereas subsequent migrations from Scandinavia had an impact mainly on the western parts of the country. The contacts between Finland and neighboring Finno-Ugric, Scandinavian and Baltic regions are evident. However, there is no support for recent migrations from Siberia and Central Europe. Our results emphasize the importance of incorporating Y-chromosomal data to reveal the population substructure which is often left undetected in mitochondrial DNA variation. Early assumptions of the homogeneity of the isolated Finnish population have now proven to be false, which may also have implications for future association studies.

© 2006 Elsevier B.V. All rights reserved.

Keywords: Y chromosome; Biallelic; STR; Variation; Substructure

1. Introduction

The genetic background of the Finnish population has been subject to wide interest from various perspectives. Its role as a genetic isolate has stimulated a variety of research on both rare

Mendelian disorders and common diseases with complex etiology. In addition to clinically oriented studies, a considerable amount of attention has been focused on unravelling the origins and history of the Finnish population. However, a refined picture of the regional variation and population structure within Finland has previously not been obtained, and the full potential of the modern Y-chromosomal analysis has not been used for gaining information about the origins of the Finnish population.

The oldest archaeological evidence of settlement in Finland dates back to approximately 10 500 years ago (Takala, 2004) and coincides with the time when the ice had finally retreated from this area after the last glacial maximum. The earliest inhabitants arrived in the southern part of Finland from the south and south-east, and in the northern parts of the country along the North Sea coast. Archeologists assume that Finland was settled by repeated

Abbreviations: AGU, aspartylglukosaminuria; AMOVA, analysis of molecular variance; DHPLC, denaturing high-performance liquid chromatography; mtDNA, mitochondrial DNA; PCA, principal component analysis; RFLP, restriction fragment length polymorphism; STR, short tandem repeat; YBP, years before present.

* Corresponding author. P.O. Box 63 / Haartmaninkatu 8, 00014, University of Helsinki, Finland. Tel.: +358 9 19125480; fax: +358 9 19125478.

E-mail address: tuuli.lappalainen@helsinki.fi (T. Lappalainen).

¹ Tuuli Lappalainen and Satu Koivumäki have contributed equally to this article.

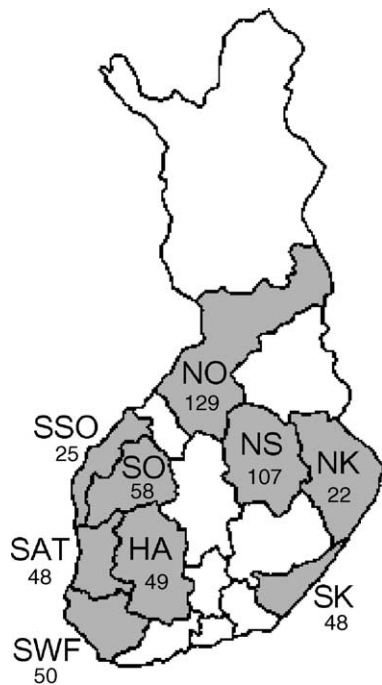


Fig. 1. Geographical locations of the provinces. Western Finland: Southern Ostrobothnia (SO), Häme (HA), South-Western Finland (SWF), Swedish-Speaking Ostrobothnia (SSO), Satakunta (SAT) and Eastern Finland: Northern Karelia (NK), Southern Karelia (SK), Northern Savo (NS), and Northern Ostrobothnia (NO). Numbers denote sample sizes.

migration waves with associated genetic and/or cultural effects, including the arrival of the comb ceramic culture from the east (6900–4900 YBP) and the corded ware culture at the western parts of the country (approximately 5200–4800 YBP). Later, Finland received important influences especially from the south and west, e.g. from the Baltic region and Scandinavia (Huurre, 1995). After the initial settlement, the total population size of Finland remained low, which, combined with local isolation of small population units and demographic crises, has caused major bottleneck events during the Finnish population history. Although the Finnish population has often been considered to be rather homogenous, many cultural (Talve, 1972; Huurre, 1995) and linguistic (Rapola, 1961) phenomena have a uniform

southeast–northwest borderline, in addition to epidemiological borderlines of e.g. cardiovascular disease and stroke (Forsen et al., 1982; Jousilahti et al., 1998; Tuomilehto et al., 1992).

Autosomal markers (Nevanlinna, 1972) and control region sequences of mitochondrial DNA (Lahermo et al., 1996) indicate that Finns are genetically homogenous and that their genetic background does not differ significantly from other European populations. However, earlier studies on Y-chromosomal variation have supported the scenario of duality of the population structure (e.g. Kittles et al., 1998; Lahermo et al., 1999) in addition to a significant eastern contribution to the paternal gene pool of the Finns, prominent especially in the eastern parts of Finland. Most notably, the presence of the DYF155S2 deletion polymorphism and Tat C allele indicates eastern affinities (Kittles et al., 1998; Lahermo et al., 1999).

In order to obtain a more refined picture of the population substructure and to study the paternal origins of the Finnish population, we have analyzed Y-chromosomal variation in an extended and representative population sample, comprised of a total of 536 unrelated males from the eastern and the western subpopulations of Finland, using a combination of rapidly evolving STR loci and slowly evolving biallelic loci.

2. Materials and methods

2.1. Population and DNA samples

DNA was extracted with standard methods from the blood samples of 536 healthy unrelated men from the eastern and the western subpopulations of Finland, collected by the Finnish Red Cross Blood Transfusion Service (Fig. 1, Table 1). The regional origin of the donors was defined on basis of their grandparental birthplace. Geographically, Northern Ostrobothnia is in the west but it was populated from Eastern Finland during the 1500s, and is thus genetically regarded as an eastern province.

2.2. STR markers

Six tetranucleotide Y-chromosomal polymorphisms (DYS19, DYS385a/b, DYS389I/II, DYS390, DYS391, DYS393) and two trinucleotide Y-chromosomal polymorphisms (DYS388,

Table 1
The haplogroup frequencies within the eastern and western provinces

Population (abbreviation)	n	A	Y*(xA,D,E,I,J,K)	DE	I*(xIIa, IIb, IIc)	IIa	IIb	IIc
Southern Karelia (SK)	48	0 (0%)	0 (0%)	0 (0%)	0 (0%)	8 (16.67%)	0 (0%)	1 (2%)
Northern Karelia (NK)	22	0 (0%)	0 (0%)	0 (0%)	0 (0%)	4 (18.18%)	0 (0%)	0 (0%)
Northern Savo (NS)	107	0 (0%)	0 (0%)	0 (0%)	0 (0%)	16 (14.95%)	0 (0%)	1 (2%)
Northern Ostrobothnia (NO)	129	0 (0%)	1 (0.78%)	0 (0%)	0 (0%)	30 (23.26%)	0 (0%)	0 (0%)
East	306	0 (0%)	1 (0.33%)	0 (0%)	0 (0%)	58 (18.95%)	0 (0%)	2 (0.65%)
Southern Ostrobothnia (SO)	58	0 (0%)	1 (1.72%)	0 (0%)	0 (0%)	27 (46.55%)	0 (0%)	0 (0%)
Swedish-Speaking Ostrobothnia (SSO)	25	0 (0%)	1 (4%)	0 (0%)	0 (0%)	9 (36%)	0 (0%)	0 (0%)
Satakunta (SAT)	48	0 (0%)	1 (2.08%)	0 (0%)	0 (0%)	25 (52.08%)	0 (0%)	1 (2%)
Häme (HA)	49	0 (0%)	0 (0%)	0 (0%)	0 (0%)	17 (34.69%)	1 (2.04%)	0 (0%)
South-Western Finland (SWF)	50	0 (0%)	0 (0%)	2 (4%)	0 (0%)	14 (28%)	0 (0%)	1 (2%)
West	230	0 (0%)	3 (1.3%)	2 (0.87%)	0 (0%)	92 (40%)	1 (0.43%)	2 (0.87%)
Total	536	0 (0%)	4 (0.75%)	2 (0.37%)	0 (0%)	150 (27.99%)	1 (0.19%)	4 (0.75%)

Boldfacing denotes summary data from different subpopulations.

DYS392) were analyzed. PCR amplifications of the STR loci were carried out in 25 or 10 μ l reaction volumes, using 5 ng of template DNA, 1 U AmpliTaqGold (Applied Biosystems, Forster City, CA), 200 μ M of each dNTP, 1 μ M primers, 2 mM MgCl₂, and 1/10 reaction volume AmpliTaq Gold or GeneAmp PCR buffer I (Applied Biosystems) (containing 100 mM Tris–HCl pH 8.3, 500 mM KCl). Each forward primer was 6-FAM, TET, HEX, VIC or NED labeled. The primer sequences and reaction conditions for all STR loci used in the study can be found at the Genome Database URL (<http://www.gdb.org>). The amplified fragments were pooled and electrophoresis runs were performed using an ABI Prism 377 or 3730 sequencer (Applied Biosystems) under standard conditions. The fragment lengths were analyzed using Genemapper or GeneScan software (Applied Biosystems). Standardization of STR allele sizing for most markers was assured by use of allelic ladders kindly provided by Peter de Knijff (University of Leiden).

2.3. Biallelic markers

The YAP_{Alu} insertion (DYS287) locus was typed as described in Hammer and Horai (1995) and the 12f2 was typed according to Rosser et al. (2000). The Tat (Zerjal et al., 1997), SRY-1532 (originally called SRY10831, also known as SRY-1533) (Santos et al., 1999; Whitfield et al., 1995), M253 (Rootsi et al., 2004), P43 (Karafet et al., 2000) and M17 (Underhill et al., 1997) polymorphisms were typed using restriction enzyme digestion and RFLP analysis as described in the original articles. The P37 polymorphism was amplified with primers from YCC (2002) and digested with HpyCH4III, and primers from Hammer et al. (2000) and digestion with HpyCH4V were used to genotype the P25 polymorphism. The primers and conditions used for producing an amplicon flanking M9 were obtained from the SNP database (<http://www.ncbi.nlm.nih.gov/SNP/>), and the site was typed using digestion with BsmI (C was cut; G was uncut). The 92R7 was typed using primer sequences and conditions published in Mathias (1994). All the digested fragments were run on 2% agarose gel and visualized with ethidium bromide, with the exception of M17, for which an ABI Prism 377 DNA sequencer was used for electrophoresis. M233 and P36 were genotyped by sequencing the amplicons produced with primers from YCC (2002). The M170 was typed with DHPLC Wave™ equipment using primers published in Underhill et al. (1997).

2.4. Statistical and phylogenetic analysis

According to the common practice, we use the terms haplogroup and haplotype for Y-chromosomal lineages defined by biallelic and STR markers, respectively. Additionally, we use the term compound haplotype for lineages identified by the combination of biallelic and STR marker information.

Haplogroups were constructed from the biallelic locus data according to The Y Chromosome Consortium (2002) and haplogroup frequencies were calculated for both the eastern and western subpopulations. Both STR and biallelic data were used to calculate diversity values of different subpopulations, and to measure the Y-chromosomal diversity distributions within and between the eastern and the western subpopulations by doing an analysis of molecular variance (AMOVA) (Excoffier et al., 1992). The extent of differentiation between eastern and western population groups within biallelic haplogroups was estimated also by calculation of R_{ST} (or Φ_{ST}) values (Slatkin, 1995) from STR haplotype data. Significance levels of the genetic variance components as well as R_{ST} values were estimated by the use of 10 000 permutations. All of these calculations were performed with Arlequin 2.001 software (Schneider et al., 2000).

Phylogenetic relationships between the compound haplotypes were reconstructed as a median-joining network using the program Network 4.112 (Bandelt et al., 1999); software available from the Life Sciences and Engineering Technology Solutions Web-site (<http://www.fluxus-engineering.com/>). Each STR marker was assigned a weight that was inversely proportional to the variation of the marker in our sample, and the biallelic markers were given a weight of ten times the average of the STR markers.

Principal component analysis calculations were performed using MatLab (Math-Works, Inc. Natick, MA). Allele frequency matrices of combined biallelic and STR data and both marker types separately were transformed to correlation matrices and covariance matrices. A standardization of the allele frequencies was performed by dividing them by $\sqrt{\bar{p}(1-\bar{p})}$, where \bar{p} is the mean allele frequency of the whole sample. To avoid the bias effect of rare alleles, >1% and >5% frequency fractiles in addition to the whole dataset were also used. PC analyses were performed for all matrices by permutation both with all the above-mentioned data modifications and without them. The effects of different sample sizes were tested by permutating the resampling method where all subpopulation samples were

J	K*(xN2,N3,P)	N2	N3	P* (xR1a, R1b, Q)	R1a* (xR1a1)	R1a1	R1b	Q
0 (0%)	0 (0%)	0 (0%)	34 (70.83%)	1 (2.1%)	0 (0%)	4 (8.3%)	0 (0%)	0 (0%)
0 (0%)	0 (0%)	0 (0%)	15 (68.18%)	0 (0%)	0 (0%)	1 (4.55%)	1 (4.55%)	1 (4.55%)
0 (0%)	0 (0%)	0 (0%)	84 (78.50%)	0 (0%)	0 (0%)	5 (4.68%)	1 (0.93%)	0 (0%)
0 (0%)	0 (0%)	0 (0%)	84 (65.12%)	0 (0%)	0 (0%)	8 (6.2%)	6 (4.65%)	0 (0%)
0 (0%)	0 (0%)	0 (0%)	217 (70.92%)	1 (0.33%)	0 (0%)	18 (5.88%)	8 (2.61%)	1 (0.33%)
0 (0%)	0 (0%)	2 (3.45%)	15 (25.86%)	0 (0%)	0 (0%)	11 (18.96%)	2 (3.45%)	0 (0%)
0 (0%)	0 (0%)	0 (0%)	10 (40%)	0 (0%)	0 (0%)	3 (12%)	2 (8%)	0 (0%)
0 (0%)	1 (2.08%)	0 (0%)	13 (27.08%)	0 (0%)	0 (0%)	4 (8.33%)	3 (6.25%)	0 (0%)
0 (0%)	0 (0%)	0 (0%)	27 (55.10%)	0 (0%)	0 (0%)	1 (2.04%)	3 (6.12%)	0 (0%)
0 (0%)	0 (0%)	0 (0%)	30 (60.00%)	0 (0%)	0 (0%)	1 (2%)	2 (4%)	0 (0%)
0 (0%)	1 (0.43%)	2 (0.87%)	95 (41.30)	0 (0%)	0 (0%)	20 (8.70%)	12 (5.22%)	0 (0%)
0 (0%)	1 (0.19%)	2 (0.37%)	312 (58.21%)	1 (0.19%)	0 (0%)	38 (7.09%)	20 (3.73%)	1 (0.19%)

randomly reduced to a sample size of 22 individuals, which represents the smallest population sample in this study.

3. Results

3.1. Haplogroup frequency distributions

Biallelic haplogroup frequencies for the Finnish subpopulations are presented in Table 1. Statistically significant differences (χ^2 -test with P -values < 0.001) in allele frequencies between the eastern and the western provinces were seen in two haplogroups, N3 and I1a. The most common haplogroup of the Finns, N3, was found in 58% of the samples (312/536), and it is particularly common in the eastern part of the country. The second largest haplogroup among the Finns was haplogroup I1a, represented by 150 individuals (28% of the population sample), most of them from Western Finland. The frequencies of the other haplogroups were under 10%. A maximum parsimony tree illustrates the haplogroup frequencies among the eastern and the western provinces (Fig. 2).

3.2. STR haplotypes

To get a more detailed picture of the Finnish male population we genotyped moderately- and fast-evolving (Heyer et al., 1997; Kayser et al., 2000) STR markers in order to construct haplotypes within each biallelic haplogroup. Full data, with haplo-

types, frequencies of compound haplotypes and distribution in different subpopulations, are given in Supplementary Table 1, and a maximum parsimony tree of a median-joining network of compound haplotypes is presented in Fig. 3. The total number of compound haplotypes was 263, of which 205 were found only in a single male. Of the remaining 58 compound haplotypes that were observed in multiple individuals, two major ones showed evident geographical association. Compound haplotype number 72 harbors 65 individuals of which 82% are from Eastern Finland, whereas haplotype 10, with 37 individuals, is more common among western males, who make up 81% of the total frequency of this compound haplotype. Most of the other haplotypes represent this bisection topology of the network. Structuring of the STR variation into eastern and western pools even within biallelic haplogroups is supported by R_{ST} values calculated from the STR haplotype frequencies within haplogroups. R_{ST} values show significant genetic distances between eastern and western samples within haplogroups N3 and I1a ($P < 0.001$), whereas the difference is not statistically significant for smaller haplogroups (data not shown). Statistically significant differences (χ^2 -test with P -values < 0.001) in allele frequencies between the eastern and the western provinces were seen in six STRs (DYS 385a/b, DYS 389I, DYS390, DYS391, DYS392, and DYS393, data not shown).

Significant differences in both the STR and biallelic marker allele frequencies were seen between Northern Ostrobothnia and the western subpopulations, but not between this province and the eastern subpopulations (data not shown), which thus justifies the inclusion of Northern Ostrobothnia in the eastern gene pool despite its more western location.

3.3. Y-chromosomal diversity

Y-chromosomal haplotype diversity values (Table 2) were calculated using three different classifications: 1) based solely on the frequency of fourteen biallelic markers, 2) based on the frequency of eight STRs, and 3) based on the frequencies of all biallelic markers and STRs combined. The mean diversity values of biallelic and STR markers were 0.636 and 0.974 among the western provinces, and 0.472 and 0.964 among the eastern provinces, respectively. The diversity values of compound haplotypes were slightly higher or equal to the values based solely on STR data.

3.4. AMOVA

By using AMOVA, we estimated the different components of the observed genetic variance (Table 3). Most (~88%) of the genetic variance present among our samples could be explained by intrapopulation differences. 9.2% of the variation was observed between the groups and 2.3% between populations within the groups.

3.5. Principal component analysis

A principal component analysis was performed based on 9 different combinations of the statistical analysis elements (see

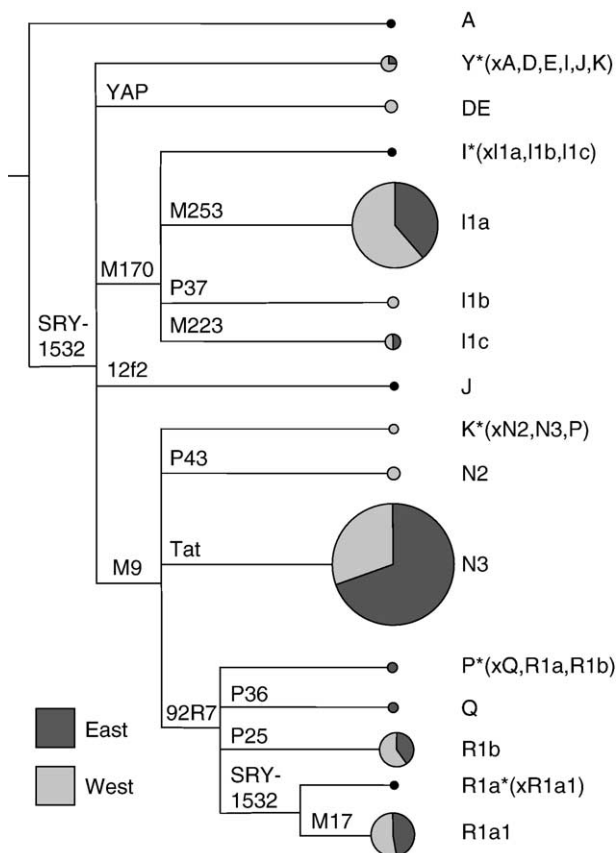


Fig. 2. The maximum parsimony tree of haplogroups that are represented in Finns. The size of each circle corresponds to a categorical frequency. Colours denote geographical distribution within each haplogroup.

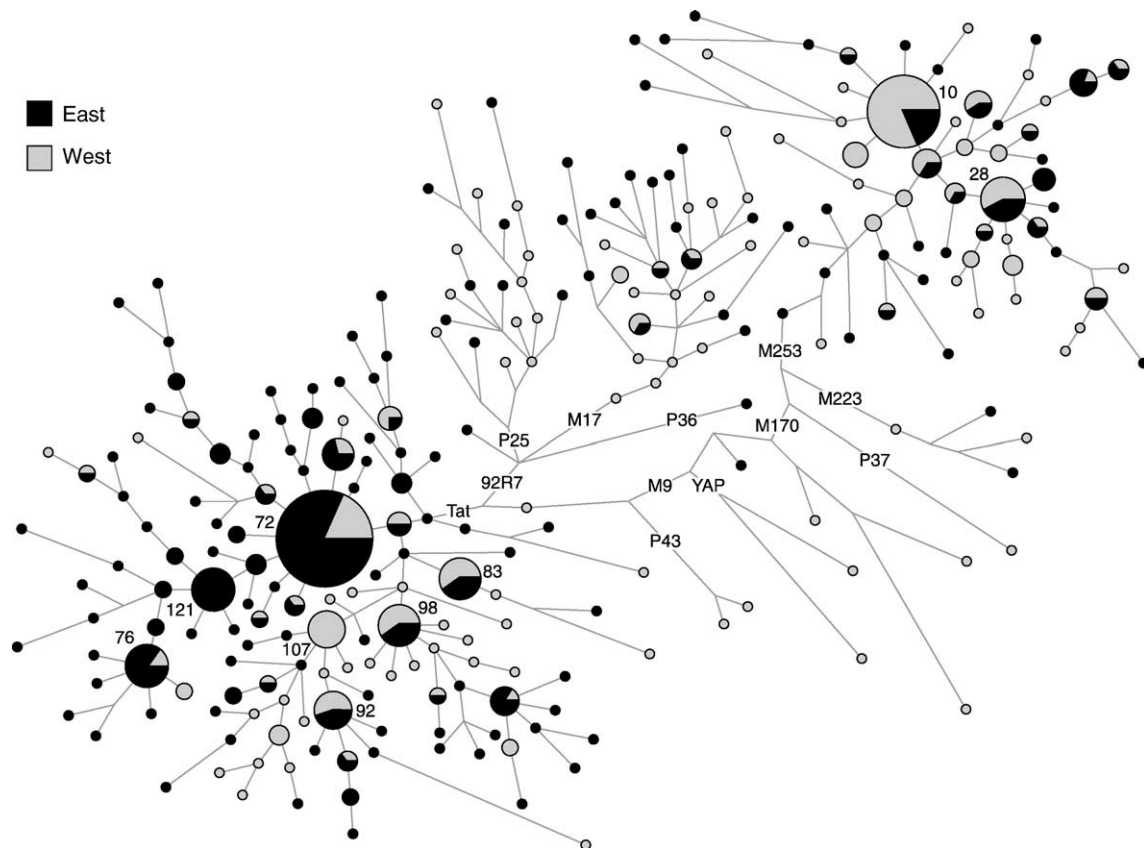


Fig. 3. A maximum parsimony tree of a median-joining network of compound haplotypes. Circles denote the compound haplotypes with the area of the circle being proportional to the number of individuals, and colours denote geographical distribution within each haplotype. Length of the links between compound haplotypes corresponds to the number of mutations, of which only biallelic polymorphisms are given. Numbering of the most common haplotypes with a frequency ≥ 10 as in Supplementary Table 1.

Section 2.4.) including correlation matrix-based PCA for allele frequency data of STR and biallelic markers with standardization (Fig. 4). The first principal component clusters populations into eastern and western affiliations. The Häme, South-Western Finland and Swedish-Speaking Ostrobothnia groups are intermediating populations between the eastern and western clusters. The two principal components explain 84% of the allele frequency variation.

4. Discussion

The aim of this study was to gather genetic evidence of the substructure of the Finnish population, to evaluate with a more detailed sample material the theory of a dual origin for the Finns earlier supported by the genetic findings of Kittles et al. (1998), and to further elucidate the contribution of European, Finno-Ugric or other founders by using a set of phylogenetically informative markers. To achieve this aim, we analyzed Y-chromosomal variation of 536 unrelated men from 9 provinces of Finland with 14 biallelic markers and 8 STRs.

Classical studies with blood group antigen markers such as ABO, Rh+/- and MN in Finns have shown large differences in frequencies at the county level, but not at the level of whole regions of the country (Mustakallio, 1989; Nevanlinna, 1973; Workman et al., 1976). Studies with maternally inherited

mtDNA have drawn a picture of high homogeneity in the regional variation of the Finnish population, together with a clear European pattern (Finnila et al., 2001; Sajantila et al., 1995; Vilkki et al., 1988). In contrast, patrilineal markers of the Y chromosome have propounded both reduced variation and an east–west frequency gradient (Hedman et al., 2004; Kittles et al., 1998, 1999a,b; Raitio et al., 2001). In the present study, we found further support for the proposed sharp distinction between the eastern and western male subpopulation groups in Finland (Kittles et al., 1998), together with some local differences within the two geographical groups. Furthermore, our analysis of the phylogenetically informative biallelic markers revealed interesting features about the Finnish population history.

4.1. The genetic substructure of the Finnish population

A parsimony tree of biallelic haplogroups and a median-joining network of compound haplotypes defined by biallelic and STR data both had similar bipolar structures, thus reflecting a division into eastern and western subpopulations. A similar scenario of two major star-shaped haplotype clusters has been found with different markers in previous studies of smaller Finnish population samples (Kittles et al., 1998, 1999a,b). This is supported also by the R_{ST} values that suggest clear differences

Table 2
The Y-chromosomal diversity values and their standard deviations

Population (n)	No. (diversity ± SD) of haplotype		
	SNP	STR	SNP + STR
Southern Karelia (48)	5 (0.473±0.078)	33 (0.942±0.024)	33 (0.942±0.024)
Northern Karelia (22)	5 (0.520±0.114)	17 (0.970±0.024)	17 (0.970±0.024)
Northern Savo (107)	5 (0.374±0.053)	70 (0.976±0.007)	70 (0.976±0.007)
Northern Ostrobothnia (129)	5 (0.520±0.041)	78 (0.946±0.016)	78 (0.946±0.016)
East (306)	8 (0.472±0.716)	167 (0.964±0.007)	168 (0.964±0.007)
Southern Ostrobothnia (58)	6 (0.690±0.040)	36 (0.923±0.030)	36 (0.923±0.030)
Swedish-speaking Ostrobothnia (25)	5 (0.717±0.056)	19 (0.977±0.018)	19 (0.977±0.018)
Satakunta (48)	7 (0.657±0.055)	35 (0.981±0.010)	35 (0.981±0.010)
Häme (49)	5 (0.583±0.048)	33 (0.975±0.011)	33 (0.975±0.011)
South-Western Finland (50)	6 (0.534±0.057)	39 (0.985±0.008)	40 (0.988±0.007)
West (230)	10 (0.636±0.051)	127 (0.974±0.006)	128 (0.974±0.006)
Total (536)	12 (0.492±0.060)	260 (0.976±0.003)	263 (0.976±0.003)

Boldfacing denotes summary data from different subpopulations.

in STR variation between the geographical groups even within the most common haplogroups. This is interesting particularly in comparison to studies of deeply rooted structuring of STR variation within haplogroups (Bosch et al., 1999), where the majority of genetic variability was found among haplogroups, and only a small fraction could be attributed to differences among populations. Furthermore, the proportion of variance between eastern and western groups of 9.2% in the AMOVA analysis can be considered to represent substantial genetic differentiation in a culturally and historically fairly homogeneous country.

χ^2 -, AMOVA and PCA test results yield interesting information on the internal history of the Finnish population and its subpopulations. Firstly, the genetic traces of the forced internal migration movement from Savo towards the northern and eastern parts of the country in the 1500s are clear. Secondly, Häme, South-Western Finland and Swedish-Speaking Ostrobothnia form intermediate regions between the eastern and

Table 3
The AMOVA results of the 9 provincial populations, based on the 14 biallelic and the 8 STR loci

Source of molecular variation	Variance (FST statistic)	
	(%)	P
Among groups	9.24	<.001
Among populations within groups	2.34	<.001
Within populations	88.42	<.001

Groups consist of subpopulations classified into eastern and western Finns as in Table 1.

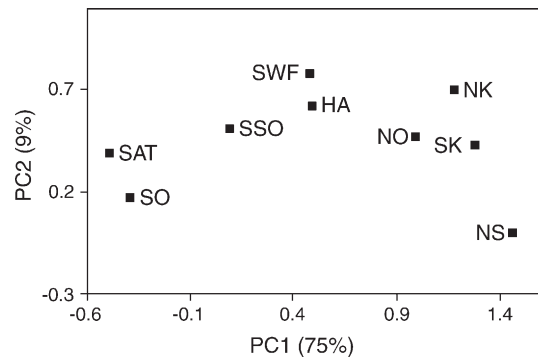


Fig. 4. Two-dimensional plot of covariance matrix-based principal component analysis with standardization of all the STR and biallelic marker allele frequencies. The numbers in parentheses represent the percent of variance explained by the respective principal components.

western subpopulation pools. For Häme and South-Western Finland, these results could easily be explained as an admixture since both these provinces have been administratively important districts during historical times. For Swedish-Speaking Ostrobothnians the explanation is less clear, and the finding could be a result of several factors ranging from low sample saturation to genetic drift and population admixture. Interestingly, however, these three western regions are the areas of the oldest permanent settlement in Finland. The special nature of the genetic variation of this area is supported by the distribution of some diseases belonging to the Finnish disease heritage, such as AGU, which is considerably rare in South-Western Finland and Ostrobothnia (Norio, 2003).

Our study confirms the existence of population substructure within a population that has earlier been considered homogeneous based mainly on mtDNA data and assumptions drawn from known population history. Finland is not the first population where such conclusions have proven to be false, as the Icelandic population has also been shown to have significant substructuring (Helgason et al., 2005). These could be seen as warning examples regarding a priori assumptions of internal homogeneity for small, isolated and culturally uniform populations.

Our evidence of the population substructure in the paternal lineages of Finns may have implications for association studies, especially for the Y-chromosomal loci. Undetected population structure is a well-known potential source of error in case-control association studies, especially if the sampling strategy is different for cases and controls in terms of geographical origin. In such cases, differences in allele frequencies between cases and controls may be due to geographical stratification of allele frequencies instead of an association to a disease, which can thus lead to false positive and negative results (Freedman et al., 2004; Marchini et al., 2004). It is vital that this is taken into account in association studies carried out with Finnish population samples. However, Y-chromosomal variation is geographically more stratified than autosomal variation due to the effects of patrilocality and enhanced genetic drift caused by the small effective population size of the Y chromosome (e.g. Jobling and Tyler-Smith, 2003). The sensitivity of the Y chromosome to the population substructure makes it a good

marker for detecting regional variation, but similar patterns are not necessarily present in autosomal variation. In this case, a set of autosomal loci should be used to assess the substructure in the Finnish population in more detail to evaluate its potential practical consequences for the methodology of association studies.

4.2. The genetic relationship of Finns to other European or Asian populations

Earlier studies have shown that the majority of the Finnish Y-chromosomal gene pool is shared with other European populations, with some notable similarities to more eastern populations which the other European populations lack (Kittles et al., 1998; Lahermo et al., 1999; Zerjal et al., 1997). The presence of eastern features in the Y-chromosomal data of the Finns is in conflict with the mtDNA data, where the Finns are indistinguishable from other Europeans (Lahermo et al., 1996; Sajantila et al., 1995). On the other hand, some eastern influences on a predominantly European background have also been observed in the blood group variation of the Finns (e.g. Cavalli-Sforza and Piazza, 1993; Guglielmino et al., 1990; Virtaranta-Knowles et al., 1991).

The eastern influence on the Finnish gene pool is manifested in the presence of the Tat and M17 polymorphisms. Haplogroup N3, defined by the Tat C allele, is the most characteristic haplogroup in Eastern Finland. It is present at a high frequency in northern Eurasia, including most of the populations speaking Finno-Ugric languages and the Balts, but it is nearly absent in Central Europe and Scandinavia, except for northern Norway (Lahermo et al., 1999; Laitinen et al., 2002; Raitio et al., 2001; Dupuy et al., 2001; Zerjal et al., 1997). The present knowledge of this polymorphism supports a common component of Finno-Ugric genetic background among these populations rather than a major Asian component among the Finns or the Balts. This is supported by the very low frequency of West Siberian haplogroups N2 and Q among the Finns. Another marker with an eastern association is M17, defining haplogroup R1a1. This polymorphism has been suggested to have originated in the steppe north of the Black Sea about 5000 YBP, and it has probably increased in frequency in association with the taming of the horse and expansion of the so-called Kurgan culture, with which the Indo-European languages have possibly spread. M17 is particularly common among the Slavic populations and Central Asians (Wells et al., 2001), and its presence in Finland probably represents relatively recent gene flow from the Russians.

In addition to eastern affiliations, close ties can be seen between Finland and the neighboring populations in the Baltic region and Scandinavia as well as other European populations. The most evident link to the Scandinavian region is the high frequency of haplogroup I1a only in Scandinavia (Rootsi et al., 2004) and Western Finland, where this haplogroup reaches its highest reported frequency of 40%. This suggests a major Swedish influence in the western parts of Finland. The low frequency of R1b, the most common Y-chromosomal haplogroup in Western Europe, is intriguing, as it is considerably

more common in Sweden (Tambets et al., 2004). This distribution pattern may indicate that the migrations bringing R1b to Scandinavia were relatively late and had only minor effect in Finland. The absence of haplogroups J and especially I1b commonly found in the Balkans and Eastern Europe demonstrate the lack of gene flow from southeastern Europe to Finland.

STR haplotype frequencies of the Finns were also compared to other European populations in the Y-STR Haplotype Reference Database (<http://www.ystr.org/index.html>). Most of the common haplotypes among the Finns can be found exclusively in Northern Europe, especially in Sweden and Estonia. This could partly be a reflection of the known strong Scandinavian influences in the Baltic region, possibly during the Viking era. The distribution of haplotype 10 suggests contacts reaching further to Central and even Southern Europe, as the haplotype is widespread in Europe with the highest frequencies in Western Finland and Sweden. However, STR haplotype frequency comparisons cannot be used to detect contacts with Finno-Ugric populations east from Finland due to lack of information from these populations.

The overall haplotype diversity values of the Finns were lower than in other European populations. The average biallelic haplogroup diversity value for Scandinavian and Baltic populations is 0.68 (Zerjal et al., 2001), while the mean value in Finland is just 0.492, even though a larger and more polymorphic marker set was used. The difference is not so drastic when the STR loci are included: a 0.99 average European value (Roewer et al., 2000) compared to the 0.976 average value of Finns. Previous studies of the Finnish Y-chromosomal variation have yielded lower diversity values than ours, but the differences are probably caused by different sample sizes and marker sets (Hedman et al., 2004; Kittles et al., 1999a,b). Low diversity values have been usually explained by means of a scenario postulating a recent bottleneck event or events during the settlement of Finland.

Altogether, the STR and biallelic Y-chromosomal variation among the populations of eastern and western provinces of Finland support a scenario where eastern and western parts of the country have been populated by people with partly different affinities, with Scandinavian influences more pronounced in the west and Finno-Ugric affinities stronger in the east. Furthermore, STR variation indicates more recent close contacts with neighboring areas in Baltica and Scandinavia, with possible Finno-Ugric associations remaining unknown. Based on the biallelic Y-chromosomal variation, it would seem that this scenario is probably best explained by repeated small migration waves, or partly more or less continuous influences, some of them from Finno-Ugric sources, others from similar sources to those behind the settlement of other Northern European and Scandinavian regions. The high Tat C frequency in the areas of the oldest settlement could suggest that early Finno-Ugric migrations — possibly associated to the comb ceramic culture — affected the whole country, and in the western parts of Finland this population was partly replaced by subsequent migrations from Scandinavia that had less effect in the most densely-populated regions within the west. The Scandinavian influence may be linked to the corded

ware culture that only affected the western parts of the country. The lower genetic diversity of the eastern population versus the western parts of the country may be explained by the internal population history, namely small population size and recent settlement of the sparsely-populated parts of the country. On the other hand, it could also reflect a situation where the Finno-Ugric settlement had been an earlier event, or possibly more of a one time event, of a small migrant population, with a possibly more diluted effect in the more densely-populated western parts of the country where the European contacts have been more continuous.

The differences in the mtDNA and Y-chromosomal data could be explained by different population histories for the males and females (Oota et al., 2001; Seielstad et al., 1998), although a major impact of patrilocality in Finland cannot easily find solid support, at least during historical times, and there is little indication of past polygamic traditions. Rather, it is plausible that a part of the sources of male influences between the regions has been different, as reviewed in the previous paragraph. Other possible explanations for the discrepancy between the Y chromosome and mtDNA data could be the more recent time-scale of detectable events in the Y-chromosomal variation due to more pronounced genetic drift. In this scenario, the more European characteristics of mtDNA variation of Finns could indicate European sources for the earliest, post-glacial migrations into Finland, whereas the eastern affiliations evident in the Y-chromosomal variation would be due to later influences. In addition, it should be noted that differences in mtDNA variation between the regions of Finland have not yet been extensively studied, and the homogeneity of maternal genetic variation may prove to be false. Finally, the lack of a refined analysis of mtDNA sequence variation among most of the Finno-Ugric populations and therefore the lack of mtDNA markers typical to these populations may also be a reason why no unique connection visible in mtDNA has been found between the Finns and the other Finno-Ugric populations.

4.3. Conclusions

In conclusion, the Finnish population is an excellent example of how incorporation of Y-chromosomal data into existing knowledge of mitochondrial DNA variation will reveal entirely novel features of the population history. Our results suggest that there are regional differences in the genetic structure of the Finnish male population. Most notably, genetic variation follows the partitioning of the country into western and eastern subpopulations, thus supporting the theory that these regions have of two major Y-chromosomal founder lineages in Finland that have resulted in partly different genetic histories. Although a part of the genetic background of the subpopulations is clearly different, it would probably be wrong to assume that either region has been populated by a single separate migration wave only, but rather that they have received different influences from partly similar and partly different sources over several millennia. In addition, connections between different parts of Finland have existed throughout the nation's history, although these connections have been hindered by different cultural, linguistic and religious backgrounds. The precise history or origin of the

difference between the eastern and the western gene pool in Finland and the exact origin of the differences or time scales and extents of migration events into Finland await further elucidation. Much of the Finnish Y-chromosomal gene pool is shared with the other European and especially Scandinavian populations, with some eastern, possibly Finno-Ugric, components remaining as a separating factor, especially in the eastern parts of the country.

Acknowledgements

This study was financially supported by the Academy of Finland (Grant No. 38826) and the Emil Aaltonen Foundation.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at [doi:10.1016/j.gene.2006.03.004](https://doi.org/10.1016/j.gene.2006.03.004).

References

- Bandelt, H.J., Forster, P., Rohl, A., 1999. Median-joining networks for inferring intraspecific phylogenies. *Mol. Biol. Evol.* 16, 37–48.
- Bosch, E., et al., 1999. Variation in short tandem repeats is deeply structured by genetic background on the human Y chromosome. *Am. J. Hum. Genet.* 65, 1623–1638.
- Cavalli-Sforza, L.L., Piazza, A., 1993. Human genomic diversity in Europe. A summary of recent research and prospects for the future. *Eur. J. Hum. Genet.* 1, 3–18.
- Dupuy, B.M., et al., 2001. Y-chromosome variation in a Norwegian population sample. *Forensic Sci. Int.* 117, 163–173.
- Excoffier, L., Smouse, P.E., Quattro, J.M., 1992. Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. *Genetics* 131, 479–491.
- Finnila, S., Hassinen, I.E., Majamaa, K., 2001. Phylogenetic network for European mtDNA. *Am. J. Hum. Genet.* 68 (6), 1475–1484.
- Forsen, T., Eriksson, J., Qiao, Q., Tervahauta, M., Nissinen, A., Tuomilehto, 1982. Incidence and prognosis of ischemic heart disease with respect to geographical area. An epidemiological study of middle-aged Finns. *Acta Med. Scand.* 212, 355–360.
- Freedman, M.L., et al., 2004. Assessing the impact of population stratification on genetic association studies. *Nat. Genet.* 36 (4), 388–393.
- Guglielmino, C.R., Piazza, A., Menozzi, P., Cavalli-Sforza, L.L., 1990. Uralic genes in Europe. *Am. J. Phys. Anthropol.* 83, 57–68.
- Hammer, M.F., Horai, S., 1995. Y chromosomal DNA variation and the peopling of Japan. *Am. J. Hum. Genet.* 56, 951–962.
- Hammer, M.F., et al., 2000. Out of Africa and back again: nested clastic analysis of human Y chromosome variation. *Mol. Biol. Evol.* 15 (4), 427–441.
- Hedman, M., Pimenoff, V., Lukka, M., Sistonen, P., Sajantila, A., 2004. Analysis of 16 Y STR loci in the Finnish population reveals a local reduction in the diversity of male lineage. *Forensic Sci. Int.* 142, 37–43.
- Helgason, A., Yngvadottir, B., Hrafnkelsson, B., Gulcher, J., Stefansson, K., 2005. An Icelandic example of the impact of population structure on association studies. *Nat. Genet.* 37 (1), 90–95.
- Heyer, E., Puymirat, J., Dieltjes, P., Bakker, E., de Knijff, P., 1997. Estimating Y chromosome specific microsatellite mutation frequencies using deep rooting pedigrees. *Hum. Mol. Genet.* 6, 799–803.
- Huurre, M., 1995. 9000 vuotta Suomen esihistoriaa (9000 years of Finnish prehistory). Otava, Helsinki.
- Jobling, M.A., Tyler-Smith, C., 2003. The human Y chromosome: an evolutionary marker comes of age. *Nat. Genet.* 4, 598–612.

- Jousilahti, P., Vartiainen, E., Tuomilehto, J., Pekkanen, J., Puska, P., 1998. Role of known risk factors in explaining the difference in the risk of coronary heart disease between eastern and southwestern Finland. *Ann. Med.* 30, 481–487.
- Karafet, T.M., Osipova, L.P., Gubina, M.A., Posukh, O.L., Zegura, S.L., Hammer, M.F., 2000. High levels of Y-chromosome differentiation among native Siberian populations and the genetic signature of a boreal hunter-gatherer way of life. *Hum. Biol.* 75, 761–789.
- Kayser, M., et al., 2000. Characteristics and frequency of germline mutations at microsatellite loci from the human Y chromosome, as revealed by direct observation in father/son pairs. *Am. J. Hum. Genet.* 66, 1580–1588.
- Kittles, R.A., et al., 1998. Dual origins of Finns revealed by Y chromosome haplotype variation. *Am. J. Hum. Genet.* 62, 1171–1179.
- Kittles, R.A., et al., 1999a. Autosomal, mitochondrial, and Y chromosome DNA variation in Finland: evidence for a male-specific bottleneck. *Am. J. Phys. Anthropol.* 108, 381–399.
- Kittles, R.A., et al., 1999b. Cladistic association analysis of Y chromosome effects on alcohol dependence and related personality traits. *Proc. Natl. Acad. Sci. U. S. A.* 96, 4204–4209.
- Lahermo, P., et al., 1996. The genetic relationship between the Finns and the Finnish Saami (Lapps): analysis of nuclear DNA and mtDNA. *Am. J. Hum. Genet.* 58, 1309–1322.
- Lahermo, P., et al., 1999. Y chromosomal polymorphisms reveal founding lineages in the Finns and the Saami. *Eur. J. Hum. Genet.* 7, 447–458.
- Laitinen, V., Lahermo, P., Sistonen, P., Savontaus, M.L., 2002. Y-chromosomal diversity suggests that Baltic males share common Finno-Ugric-speaking forefathers. *Hum. Hered.* 53, 68–78.
- Marchini, J., Cardon, L.R., Phillips, M.S., Donnelly, P., 2004. The effects of human population structure on large genetic association studies. *Nat. Genet.* 36 (5), 512–517.
- Mathias, N., 1994. Highly informative compound haplotypes for the human Y chromosome. Bayés, M and Tyler-Smith, C *Human Molecular Genetics* 3 (1), 115.
- Mustakallio, E., 1989. Veriryhmät O, A, B, AB ja MN Suomessa ja asutuksen leviäminen. *Ann Universitatis Turkuensis* C77.
- Nevanlinna, H.R., 1972. The Finnish population structure. A genetic and genealogical study. *Hereditas* 71, 195–236.
- Nevanlinna, H.R., 1973. Further evidence of the inhabitation in a marginal population. *Hereditas* 74, 127–131.
- Norio, R., 2003. The Finnish disease heritage. III: the individual diseases. *Hum. Genet.* 112, 470–526.
- Oota, H., Settheetham-Ishida, W., Tiwawech, D., Ishida, T., Stoneking, M., 2001. Human mtDNA and Y-chromosome variation is correlated with matrilineal versus patrilineal residence. *Nat. Genet.* 29 (1), 20–21.
- Raitio, M., et al., 2001. Y-chromosomal SNPs in Finno-Ugric-speaking populations analyzed by minisequencing on microarrays. *Genome Res.* 11, 471–482.
- Rapola, M., 1961. Johdatus suomen murteisiin (An Introduction to the Finnish Dialects). Suomalaisen kirjallisuuden seura, Helsinki.
- Roewer, L., et al., 2000. A new method for the evaluation of matches in non-recombining genomes: application to Y-chromosomal short tandem repeat (STR) haplotypes in European males. *Forensic Sci. Int.* 114, 31–43.
- Roots, S., et al., 2004. Phylogeography of Y-chromosome haplogroup I reveals distinct domains of prehistoric gene flow in Europe. *Am. J. Hum. Genet.* 75 (1), 128–137.
- Rosser, Z.H., et al., 2000. Y-chromosomal diversity in Europe is clinal and influenced primarily by geography, rather than by language. *Am. J. Hum. Genet.* 67, 1526–1543.
- Sajantila, A., et al., 1995. Genes and languages in Europe: an analysis of mitochondrial lineages. *Genome Res.* 5, 42–52.
- Santos, F.R., et al., 1999. The central Siberian origin for native American Y chromosomes. *Am. J. Hum. Genet.* 64, 619–628.
- Schneider, S., Roessli, D., Excoffier, L., 2000. Arlequin: a Software for Population Genetics Data Analysis. Ver 2.000. Genetics and Biometry Lab, Dept. of Anthropology, University of Geneva.
- Seielstad, M.T., Minch, E., Cavalli-Sforza, L.L., 1998. Genetic evidence for a higher female migration rate in humans. *Nat. Genet.* 20, 278–280.
- Slatkin, M., 1995. A measure of population subdivision based on microsatellite allele frequencies. *Genetics* 139 (1), 457–462.
- Takala, H., 2004. The Ristola Site in Lahti and the Earliest Postglacial Settlement of South Finland. Lahti City Museum, Lahti. 205 pp.
- Talve, I., 1972. Suomen kulttuurirajoista ja alueista (Aspects on cultural boundaries and regions in Finland). Finnish Academy of Sciences, Helsinki.
- Tambets, K., et al., 2004. The Western and eastern roots of the Saami—the story of genetic “outliers” told by mitochondrial DNA and Y chromosomes. *Am. J. Hum. Genet.* 74, 661–682.
- Tuomilehto, J., et al., 1992. Acute myocardial infarction (AMI) in Finland — baseline data from the FINMONICA AMI register in 1983–1985. *Eur. Heart J.* 13, 577–587.
- Underhill, P.A., et al., 1997. Detection of numerous Y chromosome biallelic polymorphisms by denaturing high-performance liquid chromatography. *Genome Res.* 7, 996–1005.
- Vilkki, J., Savontaus, M.L., Nikoskelainen, E.K., 1988. Human mitochondrial DNA types in Finland. *Hum. Genet.* 80, 317–321.
- Virtaranta-Knowles, K., Sistonen, P., Nevanlinna, H.R., 1991. A population genetic study in Finland: comparison of the Finnish- and Swedish-speaking populations. *Hum. Hered.* 41, 248–264.
- Wells, R.S., et al., 2001. The Eurasian heartland: a continental perspective on Y-chromosome diversity. *Proc. Natl. Acad. Sci. U. S. A.* 98 (18), 10244–10249.
- Whitfield, L.S., Sulston, J.E., Goodfellow, P.N., 1995. Sequence variation of the human Y chromosome. *Nature* 378, 379–380.
- Workman, P.L., Mielke, J.H., Nevanlinna, H.R., 1976. The genetic structure of Finland. *Am. J. Phys. Anthropol.* 44, 341–367.
- The Y Chromosomal Consortium, 2002. A nomenclature system for the tree of human Y-chromosomal binary haplogroups. *Gen. Res.* 12, 339–348.
- Zerjal, T., et al., 1997. Genetic relationships of Asians and Northern Europeans, revealed by Y-chromosomal DNA analysis. *Am. J. Hum. Genet.* 60, 1174–1183.
- Zerjal, T., et al., 2001. Geographical, linguistic, and cultural influences on genetic diversity: Y-chromosomal distribution in Northern European populations. *Mol. Biol. Evol.* 18, 1077–1087.