 Human Molecular Genetics, 2002, Vol. 11, No. 12 1381–1390

Population differences in haplotype structure within a human olfactory receptor gene cluster

Idan Menashe, Orna Man, Doron Lancet* and Yoav Gilad

Department of Molecular Genetics and the Crown Human Genome Center, The Weizmann Institute of Science, Rehovot 76100, Israel

Received January 11, 2002; Revised and Accepted April 9, 2002

We investigated the population differences in patterns of single nucleotide polymorphisms (SNPs) for a 400 kb olfactory receptor (OR) gene cluster on human chromosome 17p13.3. Samples were drawn from 35 individuals, of four different ethnogeographical origins: Pygmies, Bedouins, Yemenite Jews and Ashkenazi Jews. Of the 74 SNPs identified, two segregated between pseudogenized and intact ORs, while a third involved a change in a highly conserved motif proposed to mediate ligand-induced signal transduction. Linkage disequilibrium (LD) was computed based on phase inference across the cluster using Clark's haplotype subtraction algorithm. We also calculated LD directly from the genotypes using the expectation-maximization (EM) algorithm. Both methods yielded very similar results. Our analyses revealed substantial differences in nucleotide diversity, haplotype distribution and LD patterns among the different human populations. In particular, the two Jewish populations had low haplotype diversity and negligible decay of LD across the entire genomic region. Intriguingly, the three functional SNPs segregated at different frequencies in the different ethnogeographical groups, with the Pygmies having higher frequencies of the intact OR genes. Our data suggests that OR genes may have evolved to create different functional repertoires in distinct human populations.

INTRODUCTION

Olfactory receptor (OR) proteins are G-protein-coupled receptors (GPCRs) expressed mainly in the olfactory epithelium (1–3). There are more than 900 OR genes in the human genome (4), of which only about 40% seem to be intact. Humans are able to detect and possibly discriminate between millions of different odorants using this OR gene repertoire. Nevertheless, it was previously reported that there is great variation within humans in their olfactory sensitivity, potentially related to specific polymorphisms in OR genes (5). Cases of inability to perceive specific odorants, known as specific anosmia, were reported as a hereditary traits (5–7). Thus, variations in olfactory sensitivity among individuals may be accounted for by genetic differences. If so, it may be possible to associate olfactory sensitivity to specific OR gene allelic backgrounds. A first step in this endeavor is the examination of patterns of polymorphism at olfactory receptors.

OR genes are disposed along the entire genome in dozens of clusters (4). One of the best characterized OR gene cluster is located on human chromosome 17p13.3 (8–11) (Fig. 1). We previously reported the existence of 59 single-nucleotide polymorphisms (SNPs) (30 of them in coding regions) identified within this cluster (10,11). These SNPs were detected by resequencing PCR amplification products of 12 OR coding regions and 7 introns from 30 unrelated individuals. Levels of variation within intact genes in this cluster were significantly reduced relative to pseudogenes and introns, but there was no decrease in the rate of divergence relative to chimpanzee. These results suggest that the intact genes had evolved under positive selection (10).

The present study addresses two issues. First, we ask whether the patterns of polymorphisms are different in different ethnogeographical groups. In particular, we attempt to find out whether such differences exist in OR variations that might affect chemosensory function. Second, an attempt is made to broaden our knowledge about patterns of polymorphism and linkage disequilibrium (LD) across this OR cluster, in the context of population history. This is because such knowledge would provide a suitable foundation for future phenotype–genotype association studies that would elucidate the functional role of OR genes in this region.

RESULTS

Nucleotide diversity

SNP scoring was performed by resequencing of 12 OR coding regions and segments within three OR introns. All of these

*To whom correspondence should be addressed. Tel: + 972 8 9343683 or 9344121; Fax: + 972 8 9344487; Email: doron.lancet@weizmann.ac.il

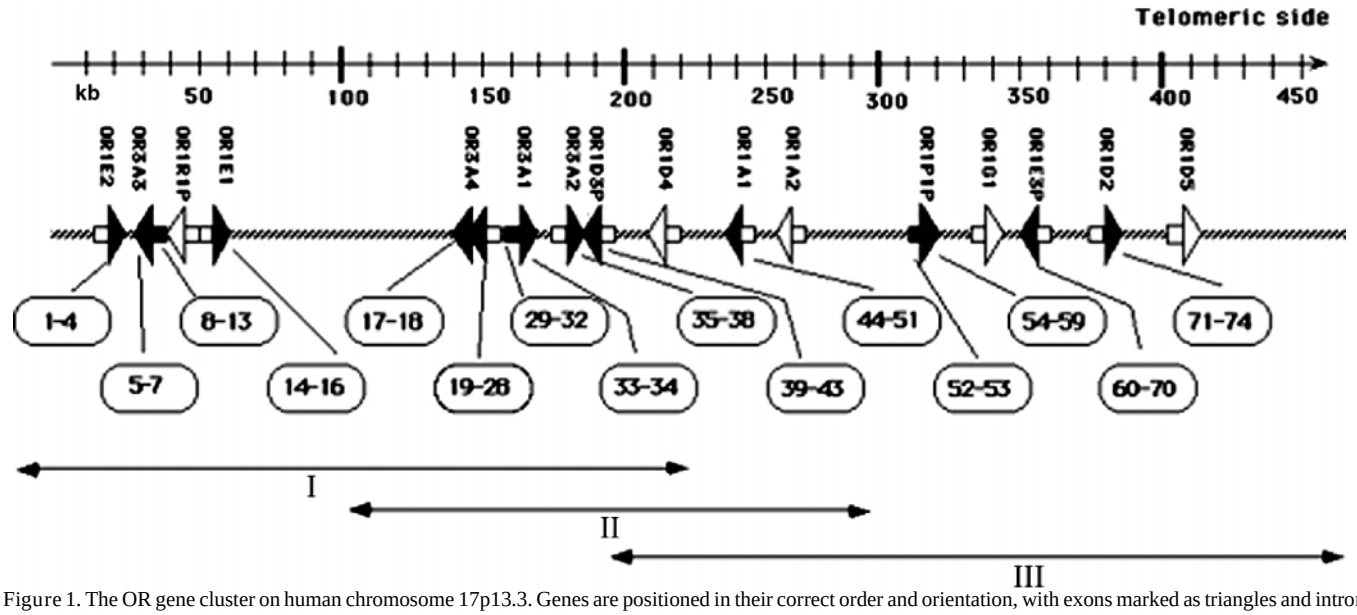

Figure 1. The OR gene cluster on human chromosome 17p13.3. Genes are positioned in their correct order and orientation, with exons marked as triangles and introns as squares. The regions that were resequenced are filled. The serial numbers of individual SNPs analyzed within each OR gene (http://bioinfo.weizmann.ac.il/~menashe/OR17_SNPs.html) are shown inside ovals. The span of the three intervals used for Clark's algorithm is indicated, marked with Roman numerals I–III.

segments were already investigated by us previously (10,11). However, in the present study, an effort was made to genotype each individual along the entire OR cluster, and subjects were selected from four disparate ethnogeographical origins.

We found a total of 74 polymorphic sites (http://bioinfo.weizmann.ac.il/~menashe/OR17_SNPs.html), 31 of which were novel. Two of the SNPs identified, in pseudogenes OR1P1P and OR1E3P, segregated between the pseudogenized and intact forms. Curiously, the variability within the same two ORs displayed significant deviations from Hardy–Weinberg equilibrium, whereby 7 out of 10 SNPs (singletons excluded) were not at equilibrium. The overall values of $\theta_w = 15.2$ (0.10% per bp) and $\pi = 0.12\%$ (Table 1) are well within the range previously reported for the presently studied cluster (10) and elsewhere (12–14).

Distinct differences were seen among the human populations. The highest nucleotide diversity was found in the Pygmy population ($\pi = 0.14\%$), while the lowest value was in the Ashkenazi Jews ($\pi = 0.08\%$). Consequently, the singletons were unequally distributed, with as many as eight in the African Pygmies and only one in the Ashkenazi Jews. These values are consistent with a historically small population size for Ashkenazi Jews and with the previously reported high variability in Africans (15–17).

We computed Tajima's D-statistic (18) to compare the observed frequency spectrum of SNPs with neutral model expectations (Table 1). The values for the entire dataset, as well as for the individual populations, did not represent a statistically significant deviation from neutrality. Yet, the two extreme values ($D = 0.88$ for the Pygmy sample and $D = -0.32$ for the Ashkenazi Jews) are noteworthy, as discussed below.

Haplotype reconstruction

Since our main goal was to calculate LD, we used only the 40 intermediate frequency sites, where the frequency of the rare

Table 1. Sequence diversity and neutrality test for the OR gene cluster on chromosome 17p13.3

	n[a]	S[b]	Singletons	π (%)	θ_w (%)	Tajima D	P (of Tajima D)
Total	35	74	19	0.12	0.10	0.42	>0.1
Pygmies	7	57	8	0.14	0.12	0.88	>0.1
Bedouin	8	55	6	0.12	0.12	0.12	>0.1
Yemenite Jews	10	46	4	0.10	0.09	0.51	>0.1
Ashkenazi Jews	10	46	1	0.08	0.08	-0.32	>0.1
Average	8.75	51	4.75	0.11	0.10	0.30	
SD	1.50	5.83	2.99	0.03	0.02	0.52	

Population variability values are given for the total data set and for each population group separately. π is the nucleotide diversity, θ_w is the population mutation rate. P-values for Tajima's D are calculated using DnaSP v. 3.12.
[a]Sample size.
[b]Number of SNPs

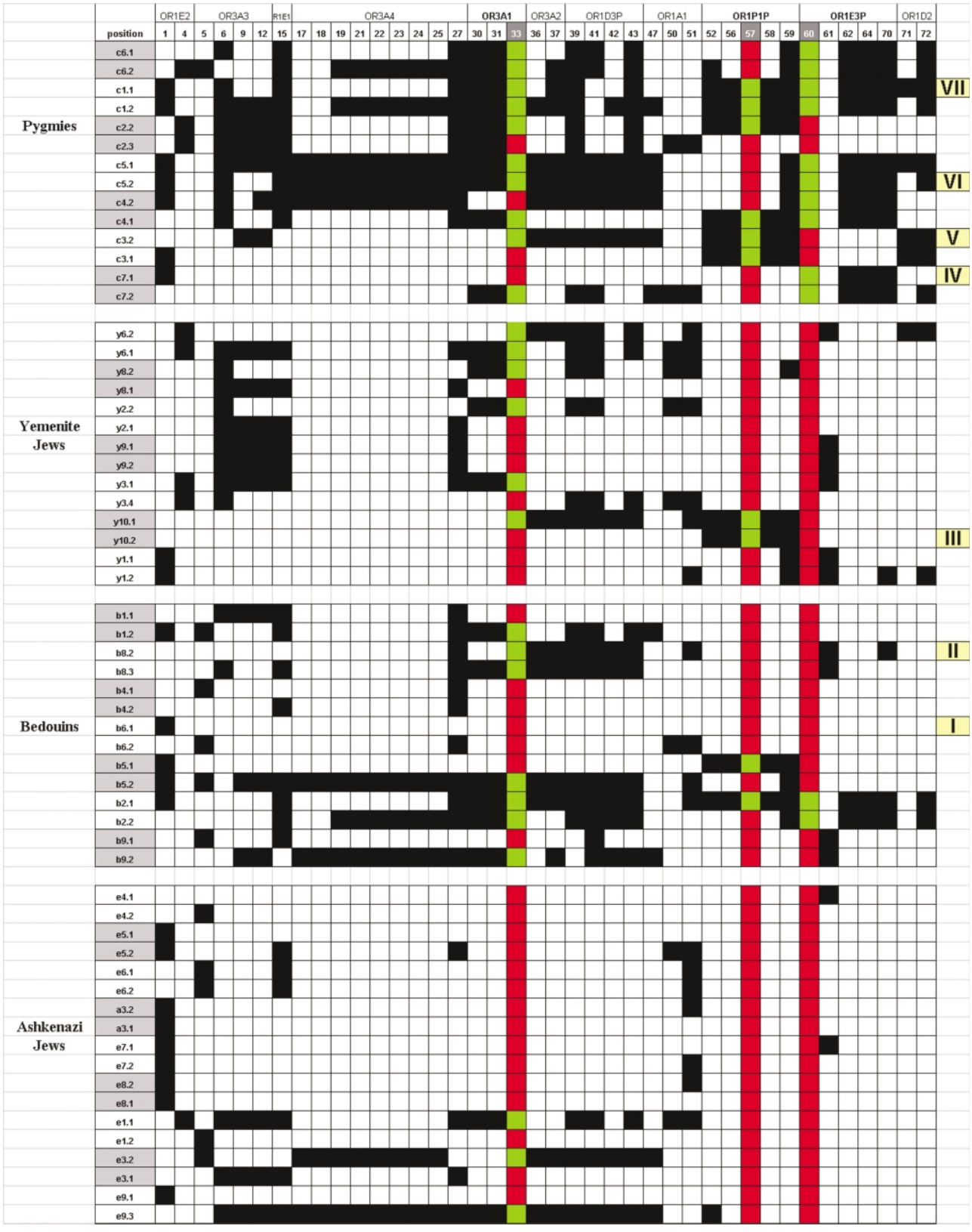

Figure 2. Haplotypes of 30 diploid individuals, as marked in the left column. The 40 polymorphic sites are indicated by their position in the different genes on the top row. The rare variant is represented by a dark square. The intact and disrupted alleles of the putatively functional SNPs are marked by green and red squares respectively. The segregating pseudogene haplotypes are shown in order of their increasing disruption level from top to bottom. The normalized fraction of each haplotype in the Pygmies was as follows: I, 0.1; II, 0.0; III, 0.6; IV, 1.0; V, 0.9; VI, 0.9; VII, 0.9.

Table 2. S_{nn} results for the total dataset and all pairwise group comparisons: the S_{nn} value is given as well as the corresponding P-value

	Yemenite Jews	Bedouin	Ashkenazi Jews	All
Pygmies	0.77	0.79	0.90	0.86
	($P < 0.01$)	($P < 0.01$)	($P < 0.01$)	($P < 0.01$)
	Yemenite Jews	0.73	0.83	0.82
		($P = 0.01$)	($P < 0.01$)	($P < 0.01$)
		Bedouin	0.69	0.74
			($P = 0.02$)	($P = 0.02$)
			Ashkenazi Jews	0.79
				($P < 0.01$)

allele was equal to or higher than 0.15. This was done because rare alleles have a higher probability to be found in LD by chance than common alleles, and hence studies generally exclude low-frequency SNPs from their LD analyses (14,19,20). Clark's algorithm (21) was used to infer haplotype phases. Forty-seven haplotypes from 30 individuals were successfully elucidated (Fig. 2). The algorithm failed for five individuals (14%) owing to ambiguities in their genotypes – a fraction not unexpected for the sample size used (21). These ambiguities are probably not the result of non-specific PCR amplifications of two highly similar ORs, since in this case excess of heterozygous genotypes should be seen, in contrast to our observations (http://bioinfo.weizmann.ac.il/~menashe/OR17_SNPs.html).

Under neutrality, where haplotype frequency is governed only by genetic drift, and assuming no recombination, the expected mean number of haplotypes for our variability values is 24(22). The observation of 47 different haplotypes – a considerably larger number (Fig. 2) – indicates that recombination events contributed to the observed haplotypes. Calculations with the DnaSP package (23), assuming no recurrent mutations, showed that the minimal number of recombination events (R_m) (24) was 18.

Population subdivision

Observation of the haplotype distribution of the whole sample revealed a notable differentiation between the ethnogeographic groups (Fig. 2). This was indicated by applying the nearest-neighbor statistic (S_{nn}) (25) to the haplotype data, which is especially suitable for relatively small populations and long genomic intervals. The results indicated a population substructure that clearly correlated with ethnogeographic origin (Table 2). All pairwise comparisons were significant, with the highest value being obtained for the comparison between Pygmies and Ashkenazi Jews.

Another pairwise comparison used was the F_{st} statistic for the diversity data (26) (Table 3). The results did not show a significant population substructure. This is perhaps expected, since the F_{st} test is known to have a low power with small sample sizes and in the presence of substantial recombination. The F_{st} results still displayed a good correlation with the S_{nn} result. Here, too, the highest value was observed for the comparison between Ashkenazi Jews and Pygmies.

Linkage disequilibrium

Based on the inferred haplotype information, the parameter D′ (27) was calculated for all pairs of sites in our sample. Fisher's exact test (FET) was used to examine the significance of LD (Fig. 3). Since Clark's method tends to reduce the number of haplotypes and thus can somewhat inflate the estimate of LD, we validated our results by calculating D′ and assessing its significance using the EM algorithm (28). This was applied directly to the genotype data, using the same frequency cutoff and the same 30 individuals. The similarity between the two algorithms was extremely high for the first two regions of the cluster (Fig. 3), with respectively 94.3% and 98.6% of the results being found to be in agreement. For the third region, the concordance level was only 64.8%. Here we are more confident with Clark's algorithm haplotypes, since the EM algorithm resolution in this region is revealed through discrepancies in intragenic haplotypes as previously determined using cloning of PCR products (10). The poor performance of the EM algorithm in this region is probably due to deviations from Hardy–Weinberg equilibrium in the two OR pseudogenes in this region.

Next, D′-values were plotted against pairwise physical distance. For the entire dataset, a decay to $D' = 0.5$ was seen at a distance of 124 kb (Fig. 4A). The observation of population substructure in the haplotypes of our sample prompted us to compare the level of LD between the different groups. A similar analysis for the four populations indicated that the

Table 3. F_{st} results for all pairwise group comparisons

	Yemenite Jews	Bedouin	Ashkenazi Jews	All
Pygmies	0.17	0.12	0.27	0.19
	Yemenite Jews	0.01	0.03	0.02
		Bedouin	0.04	0.00
			Ashkenazi Jews	0.08

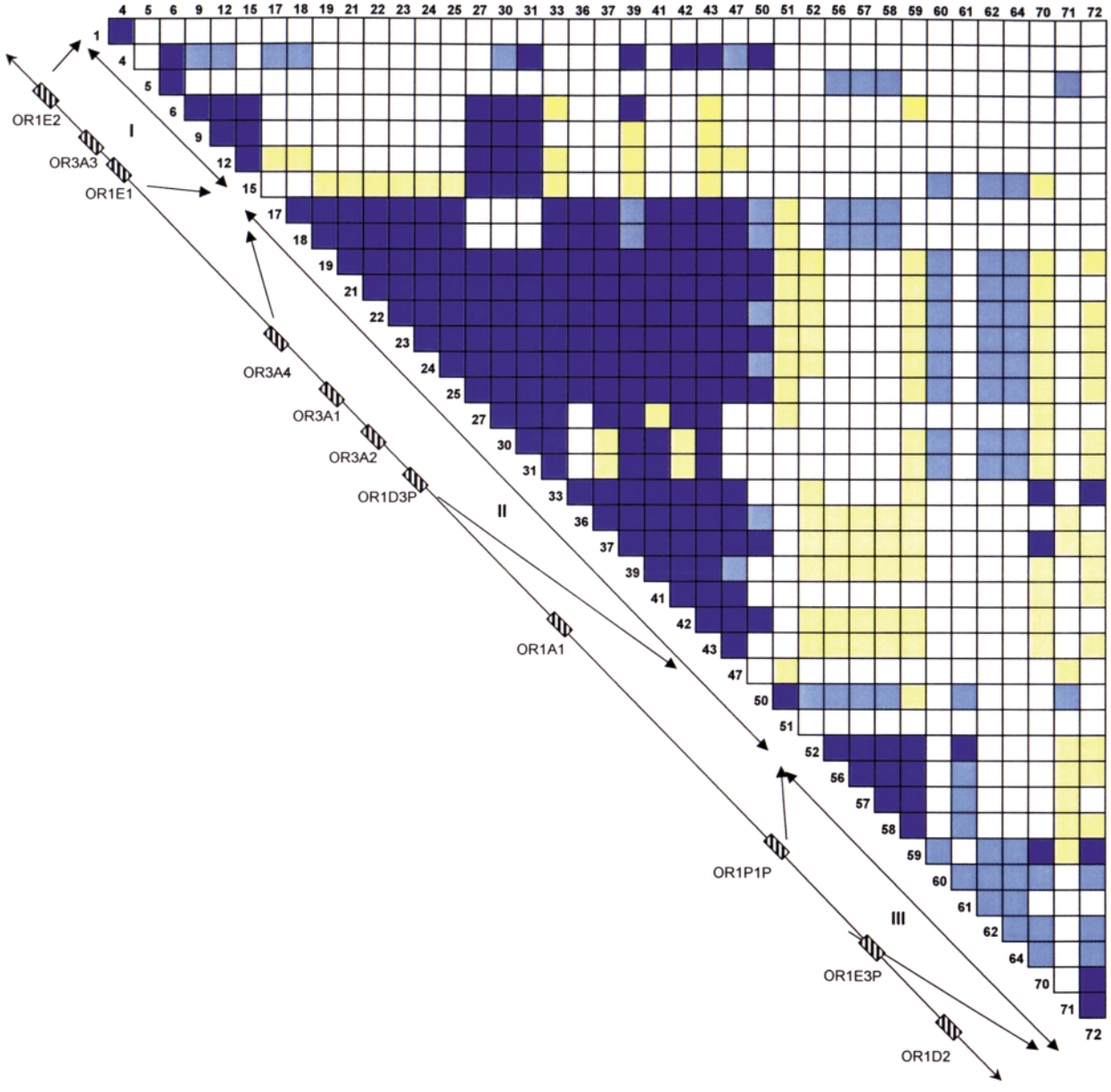

Figure 3. A linkage disequilibrium diagram across the OR cluster. SNP numbers are as in Figures 1 and 2. Dark blue squares indicate pairs in significant LD indicated by both the Clark and EM algorithms. Other colors indicate pairs in significant LD indicated only by one algorithm: Clark's (light blue) or EM (yellow). The significance level in all pairs labeled in color was $P < 0.05$ using Fisher's exact test. The positions of the SNPs relative to the OR genes and the Clark's algorithm analysis intervals are indicated near the diagonal.

Pygmies had the steepest decay of LD, with a value of $D' = 0.5$ at 210 = kb, while for the Bedouins ($D' = 0.5$ at 320 kb), Ashkenazi ($D' = 0.5$ at 400 = kb) and Yemenite Jews ($D' = 0.5$ at about 1.1 Mb), we observed considerably higher values (Fig. 4B). The minimal recombination values varied roughly in accordance, and were $R_m = 12$, 11, 5 and 6, respectively. We also found differences between populations in the distribution of LD across regions (data not shown). However, given the small sample sizes, these differences in spatial patterns of LD may only be taken as indicative and not statistically proven.

Segregating pseudogenes

Five pseudogenes were included in this analysis, two of which (OR1E3P and OR1P1P) have an open reading frame interrupted at only one position. The coding region of OR1E3P is

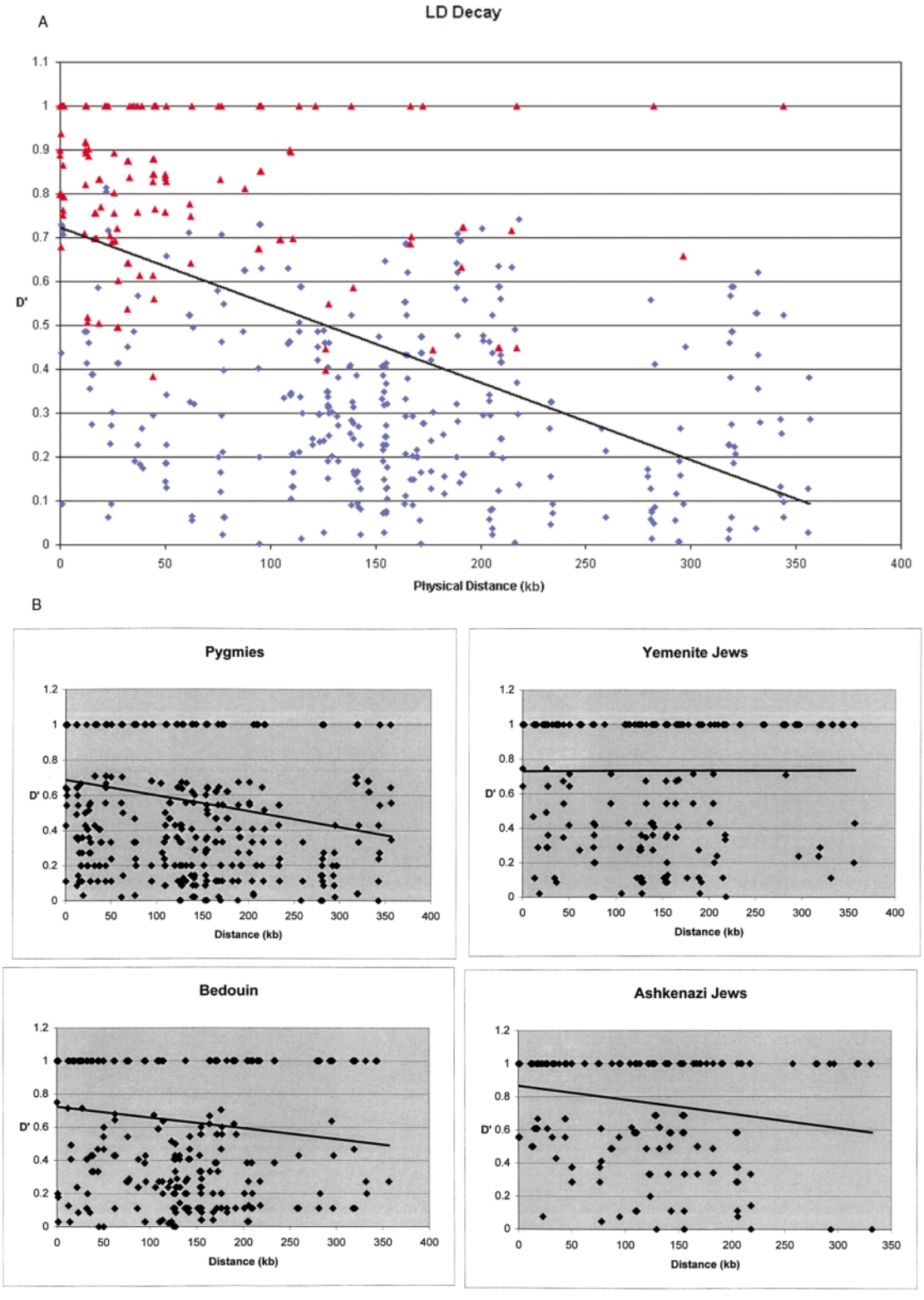

Figure 4. Scatterplot of the decay of LD with physical distance. Each point is the D′-value for a pair of sites separated by a physical distance indicated on the abscissa. A linear-fit trend-line for the data is plotted. The significance of the data was calculated using a permutation test. (A) Decay of LD to $D' = 0.5$ for the entire sample was at a distance of 124 kb ($R = -0.53$, $P < 0.01$). Pairs with significant ($P < 0.05$) D′-pairs are in red triangles. (B) Decay of LD for the individual population groups. Pygmy and Bedouin populations had significant decays to $D' = 0.5$ at a distance of 210 kb ($R = -0.24$, $P < 0.01$) and 320 kb ($R = -0.16$, $P < 0.01$), respectively, whereas the Ashkenazi Jews and Yemenite Jews displayed no significant decay [$D' = 0.5$ at about 400 kb, ($R = -0.11$, $P = 0.08$) and $D' = 0.5$ at about 1.1 Mb ($R = -0.01$, $P = 0.41$), respectively].

interrupted by a single-base deletion (in nucleotide 54) that causes a frameshift and results in a premature stop codon. The coding region of OR1P1P is interrupted by a nonsense mutation (T→A at nucleotide 553). These two mutations were found to be polymorphic in our sample: a single-base deletion of OR1E3P was absent in 12 of 70 (17%) of the chromosomes, and a missense mutation in OR1P1P was not seen in 14 of 70 (20%) of the chromosomes.

Another notable segregation is in OR3A1, where a single nucleotide substitution (G→A) yields an amino acid replacement (arginine → glutamine) in a highly conserved position in ORs and other GPCRs [the DRY motif (29)]. This residue has been suggested to play a crucial role in signaling-related conformational changes in seven-helix receptors (30). We therefore propose that the non-conservative replacement R125Q may have functional consequences, potentially leading to receptor inactivation. The intact R form was present in 48% of all chromosomes.

A most striking finding was the unequal disposition of all three segregating pseudogenes in the different ethnogeographical groups (Fig. 2). Inspection of the haplotypes for these three SNPs revealed seven out of the expected eight haplotypes. Four of the chromosomes contained the all-intact haplotype, including one homozygote pygmy individual. There was a clear trend whereby the more intact haplotypes were prevalent among the pygmies while the disrupted haplotypes were found in higher proportions in the non-African populations (Fig. 2). The highest levels of haplotype disruption were found in the Ashkenazi Jews, where 84% of the haplotypes were completely damaged.

DISCUSSION

The present resequencing study, which included intact OR coding regions, OR pseudogenes and OR intronic sequences, uncovered 31 new SNPs and confirmed 43 that were previously published. The nucleotide-diversity value for the entire data set was consistent with other human population studies (12–14,31), with clear interpopulation differences. In the two extremes, the Pygmies had a π-value roughly twice that of the Ashkenazi Jews. This trend is in general agreement with the findings of Gilad et al. (31) for the MAO-A locus, obtained for the same human DNA samples, but the discrepancy in the present case was significantly larger. The observation that the Ashkenazi Jews have the lowest π-value in our sample is consistent with other reports (16,32), which argue for the maintenance of small population size through human history.

A statistically significant deviation from neutral expectations, using Tajima's D-test of the SNP frequency spectrum, was found neither for the entire sample nor for the individual populations. That the most positive Tajima D-value occurred in the Pygmies was somewhat surprising. Published data suggest that African samples tend to have less positive Tajima D-values than non-African ones (17,33) – a trend that is attributed to possible bottleneck events in the history of the latter. This discrepancy may be explained by selection on the olfactory repertoire (10).

Other studies have indicated a substantial variation in LD distribution along the human genome (12,19,20). Numerous factors were proposed to account for this variation – most notably differences in recombination rate (34). The OR cluster studied here was found to have a particularly slow decay of LD, whereby D′ decreases to 0.5 at an average distance of 124 kb, and no decay is observed when the Pygmies are omitted from the analysis. In contrast, a whole-genome study (20) reported an average decay of LD distance of 60 kb, and less than 5 kb in the Yoruba population. This difference is too large to be accounted for by our somewhat smaller sample size. Also, a much lower recombination rate is an unlikely explanation, since a polymorphism-independent estimate suggests that the presently studied region on chromosome 17, around D17S1798 has experienced average rates of genetic exchange: 1.7 cM/Mb (35). Instead, the higher levels of LD in the OR cluster could reflect the pooling of ethnogeographic populations that show substantial differentiation, as evident by the S_{nn} comparison. It has been argued that an isolated ethnogeographical group should display more LD than the general population due to increased genetic drift (36,37). An admixture of such small isolated populations would yield regions with more LD than would be expected under a model of panmictic population (34,37,38).

Previous comparisons of particular African and non-African populations demonstrated that the non-African samples exhibited higher levels of LD, as exemplified by the study of Chinese and Italians versus the Hausa (33), and individuals from Utah versus a sample of Yorubas (20). Consistent with these studies, we find that the Pygmies showed the steepest decay of LD of the four ethnogeographic groups. Thus, the combined studies support a general African/non-African dichotomy, rather than random ethnic differences. This further strengthens the conclusion (20) that association mapping might best be carried out in non-African samples, with subsequent fine-scale mapping efforts conducted in African samples.

Our analysis of the S_{nn}-values for pairwise comparison between the two Jewish groups and the Bedouin population indicates that at this particular OR cluster, the two Jewish populations are genetically closer to the Bedouins than to each other (Table 2). This appears to correlate with the respective geographical distances, since the eastern Mediterranean Bedouins are intermediates between the European Ashkenazi and the South Arabian Yemenites. A study of Y-chromosome haplotypes (15) similarly demonstrated a correlation between the genetic and geographic distances of related populations, although Bedouins were not included. The intermediate position of the Bedouins is also reflected in the LD results (Fig. 4B). Many researchers have pointed out the need for preliminary analyses of population structure before embarking on association studies (39,40). The present results imply that association studies carried out on Israeli populations should take care not to pool different ethnicities, for fear of creating spurious associations due to stratification.

A close inspection of the data revealed considerable differences in the spatial distribution of LD across populations. In particular, the Ashkenazi Jews were the only ethnogeographic group for which pairs in significant LD were observed across 90 kb of the telomeric end of the cluster. Both Jewish samples showed no decay of LD throughout the entire cluster ($\sim$400 kb) – an observation consistent with a historically small

Jewish population and/or inbreeding. These findings highlight the need for samples of several populations in the initial panels used to choose SNPs that will be of use in association studies.

The LD results were calculated based on Clark's haplotype subtraction algorithm (21). This algorithm was shown to perform poorly when determining the phase for rare alleles (41), but it performed adequately in our system after omitting the rarest alleles. We performed a systematic comparison with the LD values computed by the more commonly used EM algorithm, based directly on genotype data. The high similarity observed in this comparison attests to the mutual complementarity of the two methods, and further underlines the validity of our results.

A significant result in the present study is that of three putative SNP-related functionally segregating pseudogenes. Seven of the potential eight haplotypes formed by this segregation were observed in the individuals studied. It is likely that a functional loss at a given OR locus results from a homozygous state of the pseudogene, i.e. that olfactory dysfunction is a recessive trait (5). Thus, individuals carrying at least one intact gene variant may have an extended olfactory ability compared with individuals who carry two pseudogene variants in the same locus. Accordingly, 10 of the 30 individuals in Fig. 2 might be hyposmic or anosmic in relation to holding both disrupted alleles in all three segregating OR loci. At the other extreme, three members of our sample are expected to be functionally intact at all three loci. Other individuals constitute intermediate cases with one or two potential functional disruptions. Nevertheless, it should be noted that further studies are needed in order to verify the functional loss of these genes due a single SNP.

Three other segregating pseudogenes were seen previously in the MHC-linked OR gene cluster on human chromosome 6 (42), suggesting that the phenomenon of SNPs that segregate between pseudogenes and intact OR loci is not restricted to the presently described cluster. These observations therefore suggest that different human individuals have different chemosensory repertoires (Y. Gilad et al., manuscript in preparation).

This predicted heterogeneity of olfactory phenotypes is in line with previously reported observations (43). An extension of the present study to a larger number of loci covered by the dozens of expected segregating OR pseudogenes could thus form a fertile ground for genotype–phenotype correlations. More specifically, the results of the present study constitute a solid basis for an association study between genetic variation within the OR gene cluster on human chromosome 17p13.3 and specific olfactory phenotypes. The conspicuous population substructure and the long stretches with significant LD found within these populations might help to use only a few markers within this cluster for the study of this kind.

An intriguing phenomenon is that the OR pseudogenes appear to segregate at different frequencies in different populations. In particular, considerably lower levels of disrupted OR genes are seen in the Pygmy population. While this phenomenon might be the result of genetic drift in isolated populations, it may also results from a specific selective pressure. If these findings are confirmed on a wider scale, they may provide key insight about the genetic basis of olfactory variability.

MATERIALS AND METHODS

Population samples and DNA sequencing

We resequenced 12 OR coding regions and 3 OR introns ($\sim$1000 bp each) in 35 unrelated individuals from 4 different ethnogeographic populations: 10 Ashkenazi Jews, 10 Yemenite Jews, 8 Bedouins (all from the Israeli National Laboratory for the genetics of human populations, Tel Aviv University) and 7 Pygmies (Coriel Cell Repertoires, Camden NJ). For each individual, we sequenced a total of 15 different regions scattered along about 400 kb of the OR gene cluster on human chromosome 17 (15 kb in total) (Fig. 1).

PCR amplification was performed in a volume of 25 µl, containing 0.2 µM of each deoxynucleotide (Promega), 50 pMol of each primer, PCR buffer containing 1.5 mM $MgCl_2$, 50 mM KCl, 10 mM Tris–HCl pH 8.3, one unit of Taq DNA polymerase (Boehringer Mannheim) and 50 ng of genomic DNA. PCR conditions were as follows: 35 cycles of denaturation at 94°C, annealing at either 55°C or 60°C and extension at 72°C, each step for 1 min. The first step of denaturation and the last step of extension were 3 and 10 min long, respectively. PCR products were separated on a 1% agarose gel to view their size, and purified using the High Pure PCR Product Purification Kit (Boehringer).

Sequencing reactions were performed on PCR products in both directions with a dye-terminators cycle sequencing kit (Perkin Elmer) on an ABI 3700 automated DNA sequencer. After base-calling with the ABI Analysis Software (version 3.0), the analyzed data were edited using the Sequencher program (GeneCodes Corp., version 4.0).

We sequenced each approximately 1 kb genomic segment from both ends for each individual and used the Sequencher software to assemble the sequences and to identify DNA polymorphisms. We repeated the sequencing reaction for any individual segment containing a singleton.

Statistical analysis

We used two summaries of the nucleotide variability calculated for each ethnogeographic group; Watterson's θ_w (44), which is based on the number of segregating sites in the sample, and the nucleotide diversity π (45), which is the average number of differences between all pairs of sequences in the sample. We used the Tajima D-test (46) to estimate whether the frequency spectrum of alleles deviates significantly from the expectations of a standard neutral model. The Tajima D-value is positive when there is an excess of intermediate frequency alleles, and negative when there is an excess of rare alleles. Positive Tajima D-values may be caused by recent bottlenecks or balancing selection, while negative D-values may be caused by recent selective sweep, purifying selection or population expansion.

Haplotype inference

The samples were sequenced on diploid DNA, and therefore provided ambiguous haplotype data for multiple heterozygotes. To resolve the haplotype structure of our sample, we used Clark's haplotype subtraction algorithm (21). This algorithm resolves the haplotypes following three steps: (i) identifying all

unambiguous haplotypes (all homozygous and sequences with one heterozygous site) and considering them as 'resolved'; (ii) determining whether each of the resolved haplotypes could be one of the alleles in the remaining ambiguous sequences; (iii) each time a possible phase of a double heterozygote is identified as one of the resolved ones, the phase is assumed to be known, and the remaining haplotype is added to the resolved haplotype set. The rationale for this algorithm is that homozygous haplotypes are probably common and that a double heterozygote is likely to contain known common haplotypes. This algorithm has been used previously and, in particular cases, has proven to be reliable by comparison with haplotypes obtained by direct molecular methods (12,47).

As the main present interest is in population substructure and patterns of LD, all rare variants (< 0.15) were excluded from the data before applying the algorithm. Then the algorithm was applied separately to the three major parts of the cluster (Fig. 1), allowing 6–12 overlapping SNPs between each pair of segments. Once all the haplotypes were resolved, we constructed the full-length haplotype using the overlapping SNPs and compared subsets of the heterozygous sites with the experimentally determined haplotype information of Gilad et al. (10).

Population stratification

The nearest-neighbor statistic (S_{nn}) (25) was used to test for population substructure. This method is a measure of how often a pair of nearest haplotypes (based on sequence similarity) belongs to the same ethnogeographical population group. The S_{nn}-value approaches unity when the populations at the two localities are highly differentiated, and is 0.5 when the populations are part of the same panmictic population (25). A permutation test is used to assess whether S_{nn} is significantly large for a particular sample, indicating that the populations at the two localities are differentiated. For genotype data in a small number of individuals with extensive recombination, this method was shown to perform better than alternative ones (25). The commonly used F_{st}-statistic (26) was also calculated for all pairwise population groups.

LD and recombination

The coefficient D′ (27) was used as a measure of LD between polymorphic sites, using the Graphical Overview of Linkage Disequilibrium (GOLD) software (48), applying Fisher's exact test (FET) for statistical significance. An alternative method used for pairwise LD computation was the expectation maximization (EM) algorithm (28) using the Arlequin software (http://lgb.unige.ch/arlequin).

ACKNOWLEDGEMENTS

We thank A. Clark for kindly providing us the code for his haplotype subtraction algorithm. We thank B. Shenhav for programming and M. Przeworski for valuable comments on the manuscript. Doron Lancet holds the Ralph and Lois Silver Chair in Human Genomics. This work was supported by the Crown Human Genome Center at the Weizmann Institute of Science, by the Alfried Krupp foundation, Germany, by the US National Institutes of Health (DC00305), by the German–Israeli Foundation for Scientific Research and Development and by an Israel Ministry of Science grant to the National Laboratory for Genome Infrastructure.

REFERENCES

1. Buck, L. and Axel, R. (1991) A novel multigene family may encode odorant receptors: a molecular basis for odor recognition. Cell, 65, 175–187.
2. Lancet, D. (1986) Vertebrate olfactory reception. Annu. Rev. Neurosci., 9, 329–355.
3. Mombarts, P. (1999) Seven-transmembrane proteins as odorant and chemosensory receptors. Science, 286, 707–711.
4. Glusman, G., Yanai, I., Rubin, I. and Lancet, D. (2001) The complete human olfactory subgenome. Genome Res., 11, 685–702.
5. Whissell-Buechy, D. and Amoore, J.E. (1973) Odour-blindness to musk: simple recessive inheritance. Nature, 245, 157–158.
6. Gross-Isseroff, R., Ophir, D., Bartana, A., Voet, H. and Lancet, D. (1992) Evidence for genetic determination in human twins of olfactory thresholds for a standard odorant. Neurosci. Lett., 141, 115–118.
7. Wysocki, C.J. and Beauchamp, G.K. (1984) Ability to smell androstenone is genetically determined. Proc. Natl Acad. Sci. USA, 81, 4899–4902.
8. Glusman, G., Sosinsky, A., Ben-Asher, E., Avidan, N., Sonkin, D., Bahar, A., Rosenthal, A., Clifton, S., Roe, B., Ferraz, C. et al. (2000) Sequence, structure and evolution of complete human olfactory receptor gene cluster. Genomics, 63, 227–245.
9. Sharon, D., Glusman, G., Pilpel, Y., Khen, M., Gruetzner, F., Haaf, T. and Lancet, D. (1999) Primate evolution of an olfactory receptor cluster: diversification by gene conversion and recent emergence of pseudogenes. Genomics, 61, 24–36.
10. Gilad, Y., Segre, D., Skorecki, K., Lancet, D. and Sharon, D. (2000) Dichotomy of single-nucleotide polymorphism haplotypes in olfactory receptor genes and pseudogenes. Nat. Genet., 26, 221–224.
11. Sharon, D., Gilad, Y.G.G., Khen, M., Lancet, D. and Kalush, F. (2000) Identification and characterization of coding single-nucleotide polymorphisms within a human olfactory receptor gene cluster. Gene, 260, 87–94.
12. Clark, A.G., Weiss, K.M., Nickerson, D.A., Taylor, S.L., Buchanan, A., Stengard, J., Salomaa, V., Vartiainen, E., Perola, M., Boerwinkle, E. et al. (1998) Haplotype structure and population genetic inference from nucleotide sequence variation in human lipoprotein lipase. Am. J. Hum. Genet., 63, 595–612.
13. Fullerton, S.M., Clark, A.G., Weiss, K.M., Nickerson, D.A., Taylor, S.L., Stengard, J.H., Salomaa, V., Vartiainen, E., Perola, M., Boerwinkle, E. et al. (2000) Apolipoprotein E variation at the sequence haplotype level: implications for the origin and maintenance of a major human polymorphism. Am. J. Hum. Genet., 67, 881–900.
14. Subrahmanyan, L., Eberle, M.A., Clark, A.G., Kruglyak, L. and Nickerson, D.A. (2001) Sequence variation and linkage disequilibrium in the human T-cell receptor beta (TCRB) locus. Am. J. Hum. Genet., 69, 381–395.
15. Hammer, M.F., Redd, A.J., Wood, E.T., Bonner, M.R., Jarjanazi, H., Karafet, T., Santachiara-Benerecetti, S., Oppenheim, A., Jobling, M.A., Jenkins, T. et al. (2000) Jewish and Middle Eastern non-Jewish populations share a common pool of Y-chromosome biallelic haplotypes. Proc. Natl Acad. Sci. USA, 97, 6769–6774.
16. Kobyliansky, E., Micle, S., Goldschmidt-Nathan, M., Arensburg, B. and Nathan, H. (1982) Jewish populations of the world: genetic likeness and differences. Ann. Hum. Biol., 9, 1–34.
17. Przeworski, M., Hudson, R.R., Di Rienzo, A., Goddard, K.A., Hopkins, P.J., Hall, J.M. and Witte, J.S. (2000) Adjusting the focus on human variation. Trends Genet., 16, 296–302.
18. Tajima, F. (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics, 123, 585–595.
19. Moffatt, M.F., Traherne, J.A., Abecasis, G.R. and Cookson, W.O. (2000) Single nucleotide polymorphism and linkage disequilibrium within the TCR alpha/delta locus. Hum. Mol. Genet., 9, 1011–1019.
20. Reich, D.E., Cargill, M., Bolk, S., Ireland, J., Sabeti, P.C., Richter, D.J., Lavery, T., Kouyoumjian, R., Farhadian, S.F., Ward, R. et al. (2001) Linkage disequilibrium in the human genome. Nature, 411, 199–204.
21. Clark, A.G. (1990) Inference of haplotypes from PCR-amplified samples of diploid populations. Mol. Biol. Evol., 7, 111–122.

22. Ewens, W.J. (1982) On the concept of the effective population size. Theor. Population Biol., 21, 373–378.
23. Rozas, J. and Rozas, R. (1995) DnaSP, DNA sequence polymorphism: an interactive program for estimating population genetics parameters from DNA sequence data. Comput. Appl. Biosci., 11, 621–625.
24. Hudson, R.R. and Kaplan, L.L. (1985) Statistical properties of the number of recombination events in the history of a sample of DNA sequence. Genetics, 111, 147–164.
25. Hudson, R.R. (2000) A new statistic for detecting genetic differentiation. Genetics, 155, 2011–2014.
26. Wright, S. (1951) The genetical structure of populations. Ann. Eugenics, 15, 323–354.
27. Lewontin, R.C. (1964) The interaction of selection and linkage. I. General considerations; heterotic models. Genetics, 49, 49–67.
28. Excoffier, L.a.M.S. (1995) Maximum-likelihood estimation of molecular haplotype frequencies in diploid population. Mol. Biol. Evol., 12, 921–927.
29. Probst, W.C., Snyder, L.A., Schuster, D.I., Brosius, J., Sealfon, S.C., Alewijnse, A.E., Timmerman, H., Jacobs, E.H., Smit, M.J., Roovers, E. et al. (1992) Sequence alignment of the G-protein coupled receptor superfamily. DNA Cell Biol., 11, 1–20.
30. Alewijnse, A.E., Timmerman, H., Jacobs, E.H., Smit, M.J., Roovers, E., Cotecchia, S., Leurs, R., Bonne-Tamir, B., Karlin, S. and Kenett, R. (2000) The effect of mutations in the DRY motif on the constitutive activity and structural instability of the histamine H(2) receptor. Mol. Pharmacol., 57, 890–898.
31. Gilad, Y., Rosenberg, S., Przeworski, M., Lancet, D. and Skorecki, K. (2002) Evidence for positive selection and population structure at the human MAO-A gene. Proc. Natl Acad. Sci. USA, 99, 862–867.
32. Bonne-Tamir, B., Karlin, S. and Kenett, R. (1979) Analysis of genetic data on Jewish populations. I. Historical background, demographic features, and genetic markers. Am. J. Hum. Genet., 31, 324–340.
33. Frisse, L., Hudson, R.R., Bartoszewicz, A., Wall, J.D., Donfack, J. and Di Rienzo, A. (2001) Gene conversion and different population histories may explain the contrast between polymorphism and linkage disequilibrium levels. Am. J. Hum. Genet., 69, 831–843.
34. Pritchard, J.K., Przeworski, M., Hammer, M.F., Redd, A.J., Wood, E.T., Bonner, M.R., Jarjanazi, H., Karafet, T., Santachiara-Benerecetti, S., Oppenheim, A. et al. (2001) Linkage disequilibrium in humans: models and data. Am. J. Hum. Genet., 69, 1–14.
35. Payseur, B.A. and Nachman, M.W. (2000) Microsatellite variation and recombination rate in the human genome. Genetics, 156, 1285–1298.
36. Laan, M. and Paabo, S. (1997) Demographic history and linkage disequilibrium in human populations. Nat. Genet., 17, 435–438.
37. Service, S.K., Ophoff, R.A. and Freimer, N.B. (2001) The genome-wide distribution of background linkage disequlibrium in population isolate. Hum. Mol. Genet., 10, 545–551.
38. Wilson, J.F. and Goldstein, D.B. (2000) Consistent long-range linkage disequilibrium generated by admixture in a Bantu–Semitic hybrid population. Am. J. Hum. Genet., 67, 926–935.
39. Cardon, L.R. and Bell, J.I. (2001) Association study designs for complex diseases. Nat. Rev. Genet., 2, 91–99.
40. Reich, D.E. and Goldstein, D.B. (2001) Detecting association in a case-control study while correcting for population stratification. Genet. Epidemiol., 20, 4–16.
41. Stephens, M., Smith, N.J. and Donnelly, P. (2001) A new statistical method for haplotype reconstruction from population data. Am. J. Hum. Genet., 68, 978–989.
42. Ehlers, A., Beck, S., Forbes, S.A., Trowsdale, J., Volz, A., Younger, R. and Ziegler, A. (2000) MHC-linked olfactory receptor loci exhibit polymorphism and contribute to extended HLA/OR-haplotypes. Genome Res., 10, 1968–1978.
43. Amoore, J.E. (1974) Evidence for the chemical olfactory code in man. Ann. NY Acad. Sci., 237, 137–143.
44. Watterson, G.A. (1975) On the number of segregating sites in genetical models without recombination. Theor. Population Biol., 7, 256–276.
45. Nei, M. and Li, W.H. (1979) Mathematical model for studying genetic variation in terms of restriction endonucleases. Proc. Natl Acad. Sci. USA, 76, 5269–5273.
46. Tajima, F. (1993) Measurement of DNA polymorphism. In Takahata, N. and Clark, A.G. (eds), Mechanisms of Molecular Evolution. Japan Scientific Societies Press, Tokyo, pp. 37–60.
47. Rieder, M.J., Taylor, S.L., Clarck, A.G. and Nickerson, D. (1999) Sequence variation in the human angiotensin converting enzyme. Nat. Genet., 22, 59–62.
48. Abecasis, G.R. and Cookson, W.O. (2000) GOLD – graphical overview of linkage disequilibrium. Bioinformatics, 16, 182–183.