Assessing DNA Sequence Variations in Human ESTs in a Phylogenetic Context Using High-Density Oligonucleotide Arrays

Jian-Bing Fan,1,* Dominik Gehl,2 Linda Hsie,1 Naiping Shen,1 Kerstin Lindblad-Toh,4 Jean-Philippe Laviolette,4 Elisabeth Robinson,4 Rob Lipshutz,1 David Wang,4,† Thomas J. Hudson,4,5 and Damian Labuda2,3,‡

1Affymetrix, Inc., 3380 Central Expressway, Santa Clara, California 95051, USA
2Centre de Recherche, Hôpital Sainte Justine, and 1Département de Pédiatrie, Université de Montréal, 3175 Côte Ste-Catherine, Montréal, Qc, H3T 1C5, Canada
3Département de Pédiatrie, Université de Montréal, 3175 Côte Ste-Catherine, Montréal, Qc, H3T 1C5, Canada
4Whitehead Institute, MIT Center for Genome Research, Cambridge, Massachusetts 02142, USA
5Montreal Genome Centre, McGill University Health Centre, Montreal, Qc, H3G 1A4, Canada

*Present address: Illumina, Inc., 9885 Towne Centre Drive, San Diego, California 92121, USA
†Present address: First Genetic Trust, Inc., Three Parkway North Center, Suite 150 North, Deerfield, Illinois 60015, USA
‡To whom correspondence and reprint requests should be addressed. Fax: (514) 345-4731. E-mail: damian.labuda@umontreal.ca.

We have analyzed human genomic diversity in 32 individuals representing four continental populations of Homo sapiens in the context of four ape species. We used DNA resequencing chips covering 898 expressed sequence tags (ESTs), corresponding to 109 kb of sequence. Based on the intra-species data, the neutral hypothesis could not be rejected. However, the mutation rate was two times lower than typically observed in functionally unconstrained genomic segments, suggesting a certain level of selection. The worldwide diversity (297 segregating sites and nucleotide diversity of 0.054%) was partitioned among continents, with the greatest amount of variation observed in the African sample. The long-term effective population size of the human population was estimated at 13,000; a similar figure was obtained for the African sample and a 20% lower estimate was obtained for the other continents. Africans also differed in having a higher number of continental-specific polymorphisms contributing to the higher average nucleotide diversity. These results are consistent with the existence of two distinct lineages of modern humans: amalgamation of these lineages in Africa led to the higher present-day diversity on that continent, whereas colonization of other continents by one of them gave the effect of a population bottleneck.

Key Words: sequence variation, DNA chip, phylogenetic analysis, nucleotide diversity, human genome diversity, SNP, single-nucleotide polymorphism, base substitution

INTRODUCTION

Understanding human diversity, both normal diversity and that directly responsible for inherited disease, will advance our knowledge of the genetic basis of human health, which in turn will contribute to medical genetics programs and to the development of new diagnostic and research tools. In the last decade, most studies were either oriented towards elucidation of the hereditary basis of single mendelian diseases or carried out with a practical goal of producing DNA markers for linkage mapping and forensic studies. These polymorphic markers were usually ascertained in a small number of samples, primarily of European descent, causing a bias in the assessment of variability in worldwide populations. Today, with the draft of the nuclear DNA sequence already available, attention is focusing on the challenge of a large-scale characterization of human genetic variation.

The present-day diversity of the human genome was shaped by genetic and demographic dynamics throughout the evolution of the ancestral population. The genomic information from worldwide populations documents, through mutations and recombinations that the human species has
experienced, the effects of natural selection, as well as those of prehistoric and historic migrations, past social structure, and demography. The extent of the present-day diversity cannot be understood without this historical perspective, which also requires considering the context of other primate genomes or even more extant phylogenetic branches. For the dynamic interpretation of population parameters, it is necessary to know the mutation rate characterizing the investigated genetic system, as it may differ from one region or functional unit to another. The polymorphic positions at sites with low mutational turnover (10^-8 per generation) are typically biallelic. Alleles that are identical by state with the corresponding site in the DNA sequences of our closest primate relatives, chimpanzee, gorilla, gibbon, and orangutan, can be considered identical by descent, thus representing ancestral rather than the new (or derived) alleles [1,2]. This rule applies because, given the overall substitution rate, the probability of parallel mutation along different branches in the primate phylogenetic tree is very low. However, in some sites with a higher mutation rate, such as methylable CpG dinucleotides [3], the assignment of the new or ancestral allelic status based on phylogenetic comparison might be less certain. Nevertheless, the information about the ancestral state of polymorphic sites and the extent of their conservation across species complements that on these sites' variability within human populations, assisting us in genetic epidemiological studies. A low frequency of a new allele suggests a recent origin of a polymorphism. Typically, young polymorphisms have restricted continental range and linkage disequilibrium spanning a large chromosome interval. In contrast, a polymorphism with a high frequency of its new allele is likely to be ancient, especially if geographically widespread, and its linkage disequilibria will span shorter distances [4].

DNA-chip technology offers an opportunity for a rapid and unbiased survey of a large sequence portion representing a wide range of genomic regions. The DNA chips were first used to determine DNA sequence variations in human ESTs [5]. They were also used to characterize the single-nucleotide polymorphism (SNP) patterns in the coding regions of human genes [6,7] with diverse human populations. Furthermore, the DNA chips were used to determine ancestral alleles for 214 human SNPs [8], as well as to identify evolutionarily conserved sequences between human and closely related species [9,10]. We conducted a large-scale survey of SNPs within 898 EST sequences covering over 109 kb of DNA. We studied 32 individuals from four continental human populations as well as single genomic samples of four species of apes.

RESULTS

Critical Evaluation of the DNA Chip Data

Using the variant detector arrays (VDA), we resequenced 898 primarily 3’ end ESTs covering 109,307 bp of human DNA sequence [5]. We examined the sequence in four groups of eight individuals that are of African, European, Asiatic, and Amerindian descents. Subsequently, we compared these four European samples with the corresponding DNA sequences from common chimpanzee, gorilla, orangutan, and gibbon (one individual per species). In the first experiment using the panel of 32 human samples, 388 putative polymorphic sites have been called, from which 297 positions were retained for further analysis. We eliminated positions when the following criteria were met: (i) more than 3, and up to 16, genotypes were called as unknown (25 sites); (ii) the second allele was only represented by homozygous genotypes (13 sites); (iii) more than 20, and up to 32, genotypes were heterozygous (16 sites); and (iv) 12 to 20 heterozygotes and only homozygotes for one of the alleles were called (16 sites). Seven more sites were eliminated by combining criteria (ii) and (iv). Finally, 29 positions were eliminated when the same polymorphism was present in more than one species (that is, when the same heterozygotes were found in human and ape, or in two apes). In the second experiment that added four ape species, 3839 sites were called and 3361 retained. As above, we conservatively did not accept the same polymorphism across species, considering it as an indication of an ambiguous hybridization (315 sites); we also excluded human “polymorphisms” that were eliminated (81) or not reported in the first analysis (216 sites). Many of the sites were excluded simultaneously by different criteria, which explains why their total number is smaller than the sum of sites excluded under each criterion. This also indicates that false calls, rather than being random, are related to the sequence context, rendering the analysis of certain sites unreliable. Although our exclusion criteria were at some points very conservative, relaxing some of them would only slightly increase the set of polymorphisms included in further analysis and would change the estimated reported parameters by less than 10%.

To independently validate the chip-derived data, we resequenced DNA fragments on ABI377 sequencers. These DNA fragments comprised a random sample of variant sites detected in the hybridization-based screening. Because the reliability of the polymorphic calls within a panel of human sequences has already been demonstrated [5–7], we focused on verifying the non-human sequences. We tested several orthologous ape sequences that were polymorphic in humans (67 positions), segments including the ape sites with fixed differences among species (83 positions), and, finally, a sample of segments carrying the sites found polymorphic in apes (82 positions). By ABI377 sequencing, we confirmed calls of fixed interspecific differences in 81 of 83 positions tested, which indicates an overall reliability in predicting the sequence divergence in apes at the level of 97.4% (± 1.7%). The two discordant positions were both found in orangutan: in one, ABI sequencing revealed a G rather than an A and in the second, a C call turned out to represent a C/T heterozygote. The latter suggests that the number of polymorphic calls in apes may be underestimated. While
testing ape calls at positions found to be polymorphic in humans, we obtained 2 discordant sequences out of 67, indicating an uncertainty level of about 3.0% (3.0 ± 2.1). Taken together, we estimate the resequencing of ape DNA using EST chips based on the human sequence to be ~ 97% reliable. This compares well with an experimental failure in PCR amplification of up to 6% in apes and human samples, using the set of the corresponding “human” primers. The DNA chips designed with human sequences performed very well in resequencing ape DNAs, in finding fixed substitutions, and in defining the ancestral alleles of human polymorphisms. In contrast to this, we found that the analysis performed relatively poorly in predicting the polymorphic sites in ape genomes. Among 82 ape polymorphisms tested, only 26% (21/82) were confirmed (ranging from 19%, 4/21, in gibbon to 35%, 6/17, in gorilla) and 11% (9/82) were shown to have different polymorphisms, such as misread insertions or deletions. The remaining sites were shown to be monomorphic: 35% were false-positives, identical with the human sequence, and 28% were different from the human equivalent. The latter represented false-negatives of fixed substitutions, taking into account that we did not observe other inconsistencies between the sequenced ape and the corresponding human segments. They accounted for only a small fraction, < 5%, of a total count of fixed differences observed, thus affecting very little the count of fixed differences and mutation rate estimates. The variance in the number of false calls was large, such that no significant differences in performance between ape genomes could be stated; however, the impression is that gibbon, the most extant species, performed less well than the great apes. We concluded that in the experimental set-up used here, we could not reliably estimate the level of nucleotide polymorphisms in the ape genomes.

Summary Statistics and Phylogenetic Analysis of Human Variability Data

Our survey of a worldwide sample of 32 individuals or 64 human haploid genome equivalents (n = 64) revealed 297 polymorphisms in 109 kb of ESTs analyzed, or one segregating site (S) every 368 bp. Nucleotide diversity (π), corresponding to the probability of observing two different allelic nucleotides in two chromosomes randomly chosen from the population, was 0.054% for the world sample. About 20% more variation was observed in Africa (S = 203 and π = 0.052%) than in Europe (S = 161, π = 0.042%), Asia (S = 164, π = 0.044%), or Amerindians (S = 162, π = 0.043%). Quantities S and π can be used to estimate the parameter θ = 4Nμ, where N corresponds to the long-term effective population size and μ denotes the average mutation rate. The two estimates of θ are θS = S / Σi–1 (1 / i), where Σ is the number of samples, and θπ, where π represents the average pairwise sequence difference, equivalent to π when divided by the total number of sites experimentally examined (~105,000 rather than the theoretical 109,309 sites, taking into account efficiency of amplification by PCR). By comparing θS and θπ estimates, we can test, according to Tajima [11], whether the analyzed segments evolve neutrally. The test statistics D based on the analyzed data set had a negative value (Table 1), but the difference from zero was not significant and thus the null hypothesis assuming neutrality could not be rejected. However, Tajima’s test was conservative here because the examined polymorphic sites were not linked, representing loci dispersed all over the genome. The negative D values, rather than the departure from neutrality, may also indicate population expansion, consistent with the recent demographic history of human populations. On the other hand, the overall mutation rate characterizing the EST segments of 1.04 × 10–8 per site per generation is about half that reported

### Table 1: Summary statistics of human variability data

<table>
<thead>
<tr>
<th></th>
<th>World</th>
<th>Africa</th>
<th>Europe</th>
<th>Asia</th>
<th>Americas</th>
</tr>
</thead>
<tbody>
<tr>
<td>S</td>
<td>297</td>
<td>203</td>
<td>161</td>
<td>164</td>
<td>162</td>
</tr>
<tr>
<td>Shared</td>
<td>193</td>
<td>149</td>
<td>138</td>
<td>149</td>
<td>150</td>
</tr>
<tr>
<td>Specific</td>
<td>104</td>
<td>54</td>
<td>23</td>
<td>15</td>
<td>12</td>
</tr>
<tr>
<td>Hs = θS/S (%)</td>
<td>19.0</td>
<td>26.9</td>
<td>27.4</td>
<td>27.9</td>
<td>28.1</td>
</tr>
<tr>
<td>Shared</td>
<td>26.3</td>
<td>28.8</td>
<td>29.2</td>
<td>29.3</td>
<td>29.1</td>
</tr>
<tr>
<td>Specific</td>
<td>5.6</td>
<td>21.6</td>
<td>16.7</td>
<td>13.6</td>
<td>16.2</td>
</tr>
<tr>
<td>π (%) ± SD</td>
<td>0.054 ± 0.025</td>
<td>0.052 ± 0.026</td>
<td>0.042 ± 0.022</td>
<td>0.044 ± 0.022</td>
<td>0.043 ± 0.021</td>
</tr>
<tr>
<td>θπ</td>
<td>56.4</td>
<td>54.6</td>
<td>44.2</td>
<td>45.7</td>
<td>45.6</td>
</tr>
<tr>
<td>θπ ± SD</td>
<td>62.8 ± 16.7</td>
<td>61.2 ± 22.0</td>
<td>48.5 ± 17.6</td>
<td>49.4 ± 17.9</td>
<td>48.8 ± 17.7</td>
</tr>
<tr>
<td>D (Tajima)</td>
<td>-0.36</td>
<td>-0.47</td>
<td>-0.39</td>
<td>-0.33</td>
<td>-0.28</td>
</tr>
<tr>
<td>Nπ/SD</td>
<td>12,900 ± 5,900</td>
<td>12,500 ± 6,100</td>
<td>10,100 ± 5,200</td>
<td>10,500 ± 5,200</td>
<td>10,400 ± 5,000</td>
</tr>
</tbody>
</table>

* π = 4Nμ using μ = 1.04 ± 0.04 × 10–8 per nucleotide per generation

Based on SD of nucleotide diversity, thus including sampling and stochastic processes, but not including uncertainty in the mutation rate estimate related to the uncertainty in divergence times and only approximate conversion of years to generations.
for the intronic or intergenic segments that are believed to evolve neutrally [12,13]. This would suggest a significant effect of selection acting on these sequences that are likely to be functionally constrained. However, this selection is appreciable here at phylogenetic time scale rather than at the intraspecific level [14].

Knowing $\pi$ and the mutation rate, we estimated the effective population size $N_e$ of the human population at almost 13,000 (Table 1). The same estimate based on the African sample is only slightly lower, whereas those based on the samples of the three remaining continents, of $\sim$ 10,000, are consistently lower (Table 1). These estimations assume random mating and constant population size for the substantial time period of human’s lineage history, thus neglecting the possible effect of population bottlenecks and recent demographic expansion. Nevertheless, this result further emphasizes that Africa weighs more in the overall world diversity than other continents [15].

By comparing human polymorphisms with their orthologous sites in the ape DNAs, we gained information about the status, new or ancestral, of the corresponding alleles. The allele that is found in apes (that is, which is identical by descent) is considered ancestral because there is little chance that a site polymorphic in the human lineage underwent exactly the same mutation independently in distinct ape lineages. The other allele is therefore derived, the new allele resulting from a mutation of the ancestral one. Knowing the ancestral state, we can root phylogenetic trees of human populations, establish the direction of mutations, and estimate the age of the polymorphisms [16,17]. Figure 1 compares the frequency distribution of the new alleles among continental samples compared with their frequency in the total sample (note that polymorphisms were ordered according to their new allele frequency, from the most to the least frequent, and at the same time from the oldest to the youngest). The frequency profiles of non-African populations are very similar, which is reflected in the similar concentration of relatively high new allele frequencies on the left of the plots. In the African sample, the distribution of new allele frequencies, although not strikingly different, appears flatter and extended to the right of the plot. In the phylogenetic analysis of the four populations using the frequencies of their polymorphisms, the Africans are closest to the root in the maximum likelihood tree and are followed by Europeans, Asians, and Amerindians (Fig. 2A). A DNA-distance matrix between individual genotypes was used for similarity clustering of samples by the neighbor-joining method. This analysis almost ideally groups the individuals of the

FIG. 1. Frequency spectrum of new alleles in four continental population samples compared with the total sample. Note that ordering of polymorphisms along the x-axis is arbitrary from the most to the least frequent in the world sample and the same order is kept for the continents below.

![Diagram showing frequency spectrum of new alleles in four continents](http://www.idealibrary.com)
same continental origin, with the exception of one Amerindian who falls within the Asiatic group (Fig. 2B). This demonstrates the relatively high-resolution power of such a large-scale polymorphic survey to reveal continental affinities of the individual genomes. The “bushy” tree in Fig. 2B illustrates well that the differences among collective populations’ characteristics are smaller than between individuals within these populations. The \( F_{ST} \) for the world (composed of four analyzed continents) is 0.153, indicating that only ~ 15.3% of the variance is due to differences among these populations, while the remaining portion of the variance is due to differences among individuals within these populations. The pairwise \( F_{ST} \) values of the continental samples analyzed are reported in Table 2. The greatest intercontinental variance is seen in all combinations with Africa (~ 19%) and the lowest is seen between Asia and the Americas (~ 8%), consistent with the maximum likelihood tree (Fig. 2A).

### Table 2: Continentally restricted (specific) polymorphisms in pairwise comparisons between continents

<table>
<thead>
<tr>
<th></th>
<th>Africa</th>
<th>Europe</th>
<th>Asia</th>
<th>Americas</th>
</tr>
</thead>
<tbody>
<tr>
<td>Africa</td>
<td>-</td>
<td>95</td>
<td>95</td>
<td>90</td>
</tr>
<tr>
<td>Europe</td>
<td>53</td>
<td>-</td>
<td>51</td>
<td>55</td>
</tr>
<tr>
<td>Asia</td>
<td>56</td>
<td>54</td>
<td>-</td>
<td>39</td>
</tr>
<tr>
<td>Americas</td>
<td>49</td>
<td>56</td>
<td>37</td>
<td>-</td>
</tr>
<tr>
<td>Total specific</td>
<td>54</td>
<td>23</td>
<td>15</td>
<td>12</td>
</tr>
</tbody>
</table>

*Populations in the left column are compared with those on the top and vice versa (number of specific polymorphisms given above and below diagonal, respectively). Note, that in contrast to shared polymorphisms, the relation is not symmetric: for example, there are 95 specific polymorphisms in Africa when compared with Europe, but only 53 such sites are found in Europe compared with Africa.*

### African “Identity”

Table 1, as well as Figs. 1 and 2, demonstrates the distinctiveness of Africans as compared with other continental populations, either by their greater genetic variability or by their position relative to the root and the frequency profile of new alleles. Samples from other continents display relatively similar characteristics, as far as diversity indices and the number of shared and specific polymorphisms are concerned. Where does this marked difference between Africa and the remaining continents come from?

The number of new alleles shared between continental groups is 193 per 297 polymorphisms analyzed, and remains in the range of 138–150 in the four continental samples (Table 1). In contrast, the number of continent-specific polymorphisms is about twice more in the African sample than on other continents: for example, Africans exhibit 54 specific sites while Amerindians exhibit only 12 (Table 1). This is essentially due to the presence of new alleles (note, however, that some continent-specific polymorphisms are due to new alleles fixed everywhere but in one continent). In a pairwise continent-by-continent analysis, the number of African-specific polymorphisms, compared with Europe, Asia, and Americas, stays within the narrow range of 90–95, whereas the corresponding levels of European, Asian, and Amerindian-specific polymorphisms in pairwise comparisons excluding Africa are only between 37 and 56 (Table 3).

To further analyze the continental distribution of polymorphisms, we compared the frequency classes in the total sample and in the contributing populations (Fig. 3). Frequency class 1 groups all the sites where the new allele occurs only

### Table 3: Population pairwise \( F_{ST} \) values

<table>
<thead>
<tr>
<th></th>
<th>Africa</th>
<th>Americas</th>
<th>Asia</th>
</tr>
</thead>
<tbody>
<tr>
<td>America</td>
<td>0.198</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Asia</td>
<td>0.191</td>
<td>0.078</td>
<td></td>
</tr>
<tr>
<td>Europe</td>
<td>0.193</td>
<td>0.113</td>
<td>0.108</td>
</tr>
</tbody>
</table>

*\( F_{ST} \) value for the world sample composed of four continents is 0.153.*
once; frequency class 2 groups all different sites with the new allele occurring twice, etc. To facilitate comparison between continents, the world sample is also subdivided into 16 frequency classes, 1–4, 5–8, etc. In Fig. 3A, bars represent the total contribution of the observed polymorphisms within given frequency classes, while the theoretical curve shows the expected equilibrium distribution of new alleles [11]. We did not detect any significant deviation of the observed frequencies from the theoretical expectation. However, in Africa we observed an imbalance, with a small relative excess of new alleles at low frequency classes and their relative shortage at the higher frequency classes. Furthermore, in Africa the contribution of shared and specific alleles at low frequency classes was almost equal, and there was a relative shortage of shared polymorphisms at high frequency classes (Fig. 3B). Most of the specific polymorphisms outside Africa occurred as singletons, that is, in the frequency class 1.

Similar to the number of polymorphic sites, we calculated the average per polymorphic site heterozygosity $H_s$ for all the polymorphisms and separately for the shared and the specific sites (Table 1). Understandably, $H_s$ for all polymorphic sites was lowest in the world sample, where continent-specific polymorphisms counted little because of their low global frequency. Among continental samples, $H_s$ was slightly lower in Africa, corroborating the “flat” frequency profile of the new alleles on this continent (Fig. 1) and their relatively higher concentration in low frequency classes (Fig. 3). The highest nucleotide diversity ($\pi$) observed in Africa results thus from a collective contribution of the greater number of segregating sites rather than from their greater average per polymorphic site heterozygosity $H_s$. All continental groups, including Africa, are surprisingly homogenous as far as the number of shared segregating sites and their $H_s$ values are concerned. It is thus both the number and $H_s$ in African specific sites that distinguish this continent from other continental samples [7,18] (Table 1).

**Phylogenetic Analysis of Primate Lineages**

As reported above, our results concerning fixed differences between the human and the great and the lesser ape lineages were sufficiently robust and reliable to allow the interspecies phylogenetic analysis. All the methods applied, that is, maximum likelihood, maximum parsimony, and distance-based methods (followed by neighbor-joining analysis), provide a consistent pattern of the tree topology, with gibbon, orangutan, and gorilla lineages branching off before the divergence of the chimpanzee and the human lineage (Fig. 4). Assuming the divergence times between the human lineage and chimpanzee, gorilla, orangutan (great apes), and gibbon (lesser ape) lineages of 5, 7, 12, and 16 millions years ago, respectively, and assuming the average generation time of 20 years, we obtained a mutation rate of $1.04 \times 10^{-8} \pm 0.04$ per nucleotide per generation. The uncertainty of this estimate is primarily related to that of the assumed divergence times between the lineages and is not reflected in the standard error, which only takes into account the number of divergent sites considered. The uncertainty due to the uncertainty of divergence times is on the order of 20% or more [19–21]. The generation time within our species corresponds to 30 years per generation within historical times [22,23]. It was certainly
shorter in earlier hominids [24], as it is in the presently living apes and presumably in their ancestors as well [25]. Assuming a generation time of 20 years seems a good compromise as well as a good point for comparison with other studies using the same denominator [13,16].

**DISCUSSION**

Our resequencing approach, using Affymetrix resequencing DNA chips (VDA) designed with human EST sequences, proved to be reliable in characterizing human diversity. By introducing additional acceptance criteria, we could lower the level of spurious calls to less than 3%. The analysis using DNA chips was also found to be at least 97% reliable in defining fixed differences among apes and human lineages. However, its performance in identifying polymorphic sites within the ape DNAs using a single sample per species was relatively poor and, therefore, we did not analyze these data further.

The phylogenetic analysis of lesser ape (gibbon), greater apes (orangutan, gorilla and chimpanzee), and human ESTs provided yet another demonstration of their relative evolutionary position, placing chimpanzee rather than gorilla as a sister species of *Homo sapiens* [26]. The inter-species genetic distances estimated from different pairwise human-ape or ape-ape comparisons were consistent (Table 4). Although the divergence times between ape species and humans are still only approximate [21,27], the estimated mutation rates for the human lineage were consistent, yielding an average of about $0.5 \times 10^{-9}$ per site per year. Interestingly, the distances in this study were two times lower than the corresponding values for intergenic sequences obtained in a large sequencing survey [13], therefore resulting in a 50% reduction in the mutation rate. Note that this conclusion is not affected by uncertainty in the actual divergence times between apes and humans because it is the sequence divergence between the corresponding species that is directly compared. Because the intergenic sequences are believed to evolve neutrally, this rate reduction may suggest selection pressure having acted on EST segments [14].

We characterized human genomic diversity in a sample representing four continents looking in a set of 898 expressed genomic segments. The segments originated from different chromosomes, which substantially reduced the bias of looking only at a single locus. By using samples from different continents, we also eliminated the ascertainment bias typical for analysis of polymorphisms within a single human population. The overall level of genomic diversity was about half that estimated for fourfold degenerated coding sites and in the range of that seen at twofold degenerated sites and in 5'- and 3'-untranslated regions [28]. These values were also in the nucleotide diversity range reported for a variety of nuclear non-Y chromosome DNA segments [6,8,16,29–33]. More importantly, the discrepancies in the diversity levels observed in different genetic systems narrow down when population effective size $N_e$ is estimated, that is, when the diversity is being scaled with the mutation rate. Our $N_e$ estimate of about 13,000 is close to Kimura predictions [34] and a previous estimate of 10,000 [28] from sequence diversity data, subsequently confirmed in several studies (see references above). Thus, in spite of the sequence-dependent variation in the nucleotide diversity as well as in the phylogenetic mutation rate, the ratio of these two statistics seems to remain relatively constant all over the genome. Thus, it is the mutation rate that controls the diversity level, although in the case of the ESTs studied here it is most likely under control of purifying selection.

![FIG. 4. Maximum likelihood tree based on fixed inter-species differences in the 898 human ESTs. Boxes represent confidence intervals at the external and the internal nodes.](image-url)
At the population level, we observed an important excess of the diversity among Sub-Saharan Africans. Greater diversity in Africa (both \( S \) and \( \pi \)) is as striking as the similarity among other continental samples (Table 1). This is reflected in the estimates of \( N_e \) using different continental samples (note, however, that \( N_e \) estimate refers to the whole species population even if a local sample is being used [34]). Greater diversity among Sub-Saharan Africans was used as an argument for the recent-out-of-Africa model of the origins of modern humans [35]. Indeed, the source population can be expected to preserve greater genetic diversity than its daughter populations issued from migration events. Typically, two mutually non-exclusive scenarios are considered. The first one assumes greater long-term effective population size in Africa that promoted a build-up of extra diversity compared with other continents with populations of smaller size. The second scenario invokes the out-of-Africa bottleneck that caused reduction of the previously accumulated diversity. Under the first scenario, the extra African diversity is expected to be younger, accumulated after fission of populations outside Africa. On the other hand, if new mutations are relatively infrequent (due to the recent bottleneck or a small population size), the demography will only affect the old diversity. The relative shortage of a new diversity suggests such a bottleneck to be very young (low frequency continent specific polymorphisms had no time to arise \textit{de novo}). Therefore, both scenarios, of greater population size in Africa and of the out-of-Africa bottleneck, find support in our data. In our previous paper on chromosome X diversity in worldwide populations we proposed another scenario that may explain observed distribution of genetic diversity in continental populations [18]. While analyzing the structure of SNP-based haplotypes in the dystrophin gene, we found evidence that the African-specific polymorphisms, which qualified as “young” because of their low new allele frequency, actually were old polymorphisms representing admixture [18]. In other words, our data suggested that the major historical partition occurred not between Africa and other continents, as proposed by the “bottleneck” scenario, but within Africa itself. According to this interpretation of \textit{Homo sapiens} prehistory, at least two, presumably African lineages had experienced independent evolution for a significant time period. One of them subsequently expanded, colonizing the world as well as Africa, where it mixed with the other, local lineage. Amalgamation of these lineages in Africa resulted in the excess of diversity, leaving an impression of a greater long-term effective population size on this continent, while purported demographic and range expansion of one of these lineages left non-African populations with characteristics of the recent population bottleneck.

It is primarily the star-like phylogeny of mitochondrial trees that suggests that Upper Paleolithic colonization of the continents was accompanied by demographic expansion [37,38]. Range and demographic expansion was also documented in paleontological record from both Upper Paleolithic and Neolithic sites [39]. Recent archeological data suggest the emergence of symbolic art in South Africa as early as \(77,000\) years ago, preceding its outburst in European Upper Paleolithic by almost \(40,000\) years [40]. Importantly, by diminishing effects of the genetic drift, demographic growth slows down the fixation of old, high frequency alleles as well as the increase of new allele frequency of \textit{de novo} appearing polymorphisms. This is in contrast to what would be expected if the population remained at small and constant size. According to this interpretation, the distribution of the new allele frequency classes in Fig. 3 is consistent with greater expansion in Africa than outside. On the other hand, if new mutations are relatively infrequent (due to the recent bottleneck or a small population size), the demography will only affect the old diversity.

Indeed, there is little or no signal of population expansion in our data, considering the fact that the slightly negative value of Tajima \( D \) can be ascribed to selection rather than to demographic growth. In our data, selection was suggested by a \(50\%\) reduction in the average mutation rate in the analyzed ESTs compared with intergenic regions believed to evolve neutrally. The reduction of diversity in \(3^\prime\) ESTs was observed previously [5]. Chen and Li [13] have also found the divergence in intronic regions reduced by \(25\%\) and that in the synonymous codon positions by \(30-40\%\). In turn, an
earlier analysis reported the divergence in the 3′-UTR to evolve with only two-thirds of the rate of fourfold degenerated sites [12] suggesting an even lower mutation rate in 3′-UTRs than in the coding region. In this respect, our 3′ ESTs are most similar to the latter regions and our data are therefore consistent with these earlier estimates [6,7,12,19]. The generally observed difference between autosomal and mitochondrial data in their ability to record the star-like phylogeny can be due to a combination of factors, including demographic history and density of genetic events [41–43]. The greater historical depth of African samples, due to the contribution of local lineages, may result in the slightly higher expansion signal at the level of autosomal diversity (Fig. 3). With a mutation rate of 1.04 × 10⁻⁸ per nucleotide per generation, one mutation is expected to occur in a segment of 105 kb every 1000 generations (that is, every 30,000 years in the history of modern humans). In a continental sample of eight individuals tested for the segment of 105 kb, we would expect to observe 16 new unique polymorphisms (that is, at frequency class 1; Fig. 3) having occurred during such a period of time, assuming star-like expansion. If the analyzed segments mutated at a rate of the mitochondrial DNA [38], 1600 or more polymorphic sites would be expected. Assuming a significant reduction of the “background” diversity by a population bottleneck preceding the expansion, the Tajima D is expected to be different in mitochondrial and nuclear systems; therefore, while revealing in the mitochondrial system, it could be non-informative in the nuclear one. This may provide a tentative explanation for the apparent insensitivity of the Tajima’s test to the recent human demography when applied to nuclear-autosomal data, in contrast to what can be seen in paleontological record and in the mitochondrial diversity. The results in Fig. 2B and Table 3 indicate much greater inter-individual than inter-group variation. Given the same number of shared polymorphisms across all continents and the very similar frequency profile in all but the African sample, the results are indeed consistent with the recent expansion. In Europe, Asia, and the Americas, the number of unique new polymorphisms is less than 16, and continentally restricted polymorphisms, outside frequency class 1 or 15, are barely existent (Fig. 3B).

Our data clearly indicate that the independent history of different continental populations is relatively recent. Although the nuclear genome lacks the power of resolution typical of mitochondrial system, our results are consistent with the recent out-of-Africa and within-Africa expansion of one of the African lineages. This explains the shortage of specific non-African polymorphisms and the relative excess of new low-frequency alleles in Africa. Our data provide a framework to test different scenarios of human population (pre-)history to explain the frequency profile of new alleles in the contemporary populations. Even if mtDNA and the Y chromosome seem to tell different stories, they actually represent different sides of the same story whose interpretations have to be consistent and converge with what can be read from X-chromosome and autosomal data.

**Materials and Methods**

Human DNA samples purchased from Coriell Cell Repository were NA00707, NA01057, NA07340, NA10116, NA10307, NA13111, NA13112, NA13114, NA13121A, NA13122, NA13123, NA13124, NA13125, NA1587A, NA11590, NA11589, NA12060, NA12061, NA12062, NA12063, NA12065, NA12066, NA12068, NA12069, and NA10469A, NA10470, NA10471, NA10472A, NA10492, NA10493, NA10494, NA10495A. They represent eight European descent, eight South-East Asians, eight Amerindians, and eight Sub-Saharan Africans (African Pygmies). The genomic sample from gibbon (Hylobates lar) was from the ATCC HB8370 cell line, whereas those of common chimpanzee (Pan troglodytes), gorilla (Gorilla gorilla), and orangutan (Pongo pygmaeus) were isolated from the peripheral blood obtained form Granby Zoo (Granby, Quebec). We amplified STSs from specific human genomic regions using corresponding PCR primers as described [44]. The same primer sets, originally designed for the human genomic DNA sequences, were used to amplify the homologous regions from the great apes. We screened a total of 32 human genomic DNAs (64 haploid equivalents).

**Chip design and assay.** Eight variant detector arrays (VDA), corresponding to 898 human 3′ EST sequences (109,307 bp in length), were designed as described [5]. Experimental methods for DNA sample preparation, VDA hybridization, and scanning were as described [5].

We estimated PCR amplification failure in the human and ape samples from the array hybridization results. For any given EST of any individual sample, if the average hybridization intensity was lower than the experimental threshold (10 counts above background) or if the correct re-sequencing call rate was lower than 40%, we considered the amplification as a failure. Using this cutoff, an average failure was up to 6.4% in the analysis of 32 human samples, and at similar level in chimpanzee, gorilla, orangutan, and gibbon DNAs.

**Chip data analysis.** Candidate SNPs were identified by comparative analysis of the VDA hybridization patterns in 32 human samples, using a program called “Ulysses” developed at Affymetrix. Scans from four great apes were analyzed together with four Caucasian scans. Once the polymorphic sites in each sample were identified, the Ulysses computed a P value for each locus. If the computed hybridization value was defined as the percentage of the mutant allele (B allele), as opposed to the allele used in the chip design (A allele). Since both sense-strand and anti-sense-strand were tilled on the chip, two P values, Psense and Panti-sense were obtained at each locus. Genotypes were deduced based on the P values as follows: if the Absolute [Psense - Panti-sense] ≥ 0.5, no genotype was called; if the absolute [Psense - Panti-sense] < 0.5, an average value was computed as [Psense + Panti-sense] / 2. If the average P value was > 0.75, a B/B genotype was called; if the average P value was < 0.25, an A/A genotype was called; if the average P value was between 0.25 and 0.75, an A/B genotype was called.

**Statistical analysis.** Computer-assisted analyses of the chip genotyping data were done in Microsoft ACCESS and using Arlequin program package v. 2.0 [45] (http://anthropologie.unige.ch/arlequin). Intra-specific and inter-specific phylogenetic analyses based on substitution polymorphisms and fixed differences, respectively, were carried out using Phylip package v. 3.57 [46].

We estimated the mutation rate from interspecific genetic distances determined from the comparison of human and ape ESTs, according to the Kimura two-parameter model, using the DNADIST and assuming the human-chimpanzee, human-gorilla, human-orangutan, and human-gibbon divergence times of 5, 7, 12, and 18 million years, respectively. To correct for the missing data due to PCR failure (see above), we used 105 kb rather than 109 kb to calculate the genetic distances and to estimate the mutation rate. To convert the latter estimate from units per year to units per generation, we used the generation time of 20 years. This generation time does not apply to modern humans [22,23] but approximates better the average generation time along ape and human lineages since their divergence [13].

**Acknowledgments**

We thank Eva Zietkiewics (Polish Academy of Sciences) for discussion and comments on the manuscript; Clément Lanthier from Zoological Garden in Granby, QC, for the great apes blood samples; Jacques Bernier (Les Productions Jacques Bernier) for figures; and Dominika Kozubskas (Centre de Recherche Hopital Sainte-Justine) for secretarial assistance. T.J.H. is a recipient of a Clinician-Scientist Award from the Canadian
RECEIVED FOR PUBLICATION FEBRUARY 20;
ACCEPTED JUNE 24, 2002.

REFERENCES