Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa

Sohini Ramachandran*, Omkar Deshpande†, Charles C. Roseman§, Noah A. Rosenberg¶, Marcus W. Feldman*, and L. Luca Cavalli-Sforza†

Departments of *Biological Sciences and †Computer Science, Stanford University, Stanford, CA 94305; ‡Department of Anthropology, University of Illinois at Urbana–Champaign, 209 Davenport Hall, 607 South Matthews Avenue, Urbana, IL 61801; §Department of Human Genetics, Bioinformatics Program and Life Sciences Institute, University of Michigan, 2017 Palmer Commons, 100 Washenaw Avenue, Ann Arbor, MI 48109-2218; and ¶Department of Genetics, School of Medicine, Stanford University, Stanford, CA 94305-5120

Contributed by L. Luca Cavalli-Sforza, September 2, 2005

Equilibrium models of isolation by distance predict an increase in genetic differentiation with geographic distance. Here we find a linear relationship between genetic and geographic distance in a worldwide sample of human populations, with major deviations from the fitted line explicable by admixture or extreme isolation. A close relationship is shown to exist between the correlation of geographic distance and genetic differentiation (as measured by $F_{ST}$) and the geographic pattern of heterozygosity across populations. Considering a worldwide set of geographic locations as possible sources of the human expansion, we find that heterozygosities in the globally distributed populations of the data set are best explained by an expansion originating in Africa and that no geographic origin outside of Africa accounts as well for the observed patterns of genetic diversity. Although the relationship between $F_{ST}$ and geographic distance has been interpreted in the past as the result of an equilibrium model of drift and dispersal, simulation shows that the geographic pattern of heterozygosities in this data set is consistent with a model of a serial founder effect starting at a single origin. Given this serial-founder scenario, the relationship between genetic and geographic distance allows us to derive bounds for the effects of drift and natural selection on human genetic variation.

A regular decrease of genetic similarity with increasing geographic distance has been predicted by the theory of isolation by distance (1) and by the stepping-stone model (2), under the assumption that movement connected with mating is usually restricted to short distances (3, 4). Data on genetic polymorphisms have confirmed a strong association between genetic and geographic distance; early studies were generally limited to short geographic ranges and within-regional analyses (5, 6), but later studies have been extended to wider areas (7–9). Here, we reexamine a measure of genetic differentiation on geographic distance at the global level using 783 microsatellite loci from the Human Genome Diversity Project–Centre d’Etude du Polymorphisme Humain (HGDP-CEPH) worldwide sample of populations (10, 11). We then use simulations to examine a serial founder effect scenario as a possible explanation for the observed relationship between genetic and geographic distance.

Materials and Methods

Data. The data set that we analyzed consists of 1,027 individuals from the HGDP-CEPH Human Genome Diversity Cell Line Panel (10). Several individuals from the collection of 1,056 individuals studied by Rosenberg et al. (11) were excluded from the present analysis. These included the following: (i) no. 1026, who was studied by Rosenberg et al. (11) but who was not in the HGDP-CEPH panel; (ii) nos. 770 and 980, who were identified by Rosenberg et al. (11) as likely labeling errors; (iii) nos. 589, 652, 659, 826, 979, 981, 1022, 1025, 1087, 1092, 1154, and 1235, each of whom was identified by Mountain and Ramakrishnan (12) as a duplicate sample of another individual included in the panel; (iv) nos. 111 and 220, who were identified by Mountain and Ramakrishnan (12) as duplicates of each other but whose population labels differed; and (v) 21 individuals from the Surui population, an extreme outlier in a variety of previous analyses (11, 13, 14). Individuals not studied by Rosenberg et al. (11) but analyzed here included the following: (i) no. 1331, whose genotypes had been unavailable at the time of the Rosenberg et al. (11) study; (ii) nos. 993, 994, 1028, 1030, 1031, 1033, 1034, and 1035, who were previously excluded as members of populations with small sample sizes but who were grouped for the present analysis into Southwestern Bantu (individuals no. 1028, 1031, and 1035) and Southeastern Bantu (individuals no. 993, 994, 1030, 1033, and 1034) populations. Thus, the present data set includes two additional populations along with all populations studied by Rosenberg et al. (11) except Surui for a total of 53 populations.

Each of the 1,027 individuals was genotyped for 783 autosomal microsatellite loci, which included the 377 loci from Marshfield Screening Set no. 10 that were previously studied by Rosenberg et al. (11), as well as 406 additional loci from Marshfield Screening Sets no. 13 and 52. The complete data set used in this study is available from the authors upon request.

Geographic locations of the samples were reported by Cann et al. (10). For populations where ranges of coordinates were provided, the mean of the latitudes and the mean of the longitudes of the reported region were used to characterize the population’s location. For the Northern Han of East Asia, the coordinate pair used was (39N, 114E); 39N is the northern extreme of locations where Han individuals were sampled, whereas 114E fell in the middle of the interval of latitudes at which Han individuals were sampled.

Genetic Distance. GENETIC DATA ANALYSIS (GDA) (15) was used to compute pairwise genetic distances, as measured by $F_{ST}$ (16), for all pairs of populations. We refer to $F_{ST}$ as a “genetic distance,” although strictly speaking it does not satisfy the triangle inequality (17). A pairwise matrix of $R_{ST}$ values (18) also was computed, as were matrices for several other genetic distances. All distances were found to be highly correlated (Table 1), and, consequently, only $F_{ST}$ was used in further analysis.

Geographic Distance. For each pair of populations, we calculated geographic distance in kilometers based on great circle distances.

Abbreviation: HGDP-CEPH, Human Genome Diversity Project–Centre d’Etude du Polymorphisme Humain.

*To whom correspondence may be addressed. E-mail: sohini@stanford.edu or cavalli@stanford.edu.

© 2005 by The National Academy of Sciences of the USA
Table 1. Mantel correlations between various distances

<table>
<thead>
<tr>
<th>Correlation</th>
<th>0.9688</th>
<th>0.8914</th>
<th>0.9758</th>
<th>0.9424</th>
<th>0.9408</th>
<th>0.9484</th>
<th>0.7707</th>
<th>0.8291</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\delta_{ij}$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$F_{ST}$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$G_{ST}$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PSA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$R_{ST}$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Geographic distance</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

References for measures of genetic differentiation are as follows: $D_1$ (19), $(\delta_{ij})^2$ (20), $F_{ST}$ (16), $G_{ST}$ (21), proportion of shared alleles (PSA) (22), and $R_{ST}$ (18).

using the haversine (23), according to which the distance $D$ between two points specified by (latitude, longitude) coordinates $(\alpha_1, \delta_1)$ and $(\alpha_2, \delta_2)$, with a central angle of $\theta$ between the two points is

$$D = 2R \arctan \left( \frac{\hav(\theta)}{\sqrt{1 - \hav(\theta)}} \right),$$ \[1\]

where $\hav(\theta) = \sin^2 \left( \frac{\delta_1 - \delta_2}{2} \right) + \cos \delta_1 \cos \delta_2 \sin^2 \left( \frac{\alpha_1 - \alpha_2}{2} \right), \[2\]

and $R$ is the radius of the Earth, which we assume to be 6,371 km.

In addition to great circle geographic distances, we also calculated pairwise geographic distances using five obligatory waypoints. Waypoints were used to make our between-continent distance estimates more reflective of human migration patterns, taking into account the belief that until recently humans did not generally cross large bodies of water while migrating. These waypoints were as follows: Anadyr, Russia (64N, 177E); Cairo, Egypt (30N, 31E); Istanbul, Turkey (41N, 28E); Phnom Penh, Cambodia (11N, 104E); and Prince Rupert, Canada (54N, 130W). The distance between two waypoints is then the sum of the great circle distances between the points and the waypoint(s) in the path connecting them, plus the great circle distance(s) between waypoints if two or more waypoints are needed. Including the waypoints in between-continent distance calculations forced movement, for example, to Oceania via Southeast Asia, and to America via the Bering Strait and western coast of North America (see Fig. 6, which is published as supporting information on the PNAS web site).

Because there may have been an important expansion route along the southern coasts of Africa and Eurasia, and blue diamonds represent comparisons with America and Oceania (Ramachandran et al. (2004)).

**Figure 1.** Scatterplot of $F_{ST}$ and geographic distance. Red dots denote within-region comparisons, green triangles indicate comparisons between populations in Africa and Eurasia, and blue diamonds represent comparisons with America and Oceania. (A) The relationship between $F_{ST}$ and geographic distance computed using great circle distances. $R^2$ for the linear regression of genetic distance on geographic distance is 0.5882. (B) The correction for large bodies of water produces a different scatterplot ($R^2 = 0.7835$). The regression line fitted to the data $(F_{ST} = 4.35 \times 10^{-3} + (6.28 \times 10^{-4}) \times \text{geographic distance in kilometers})$ is drawn in black.
had mean 0 and SD 1) independently within each type of data (genetic and geographic). Because the sign of a principal coordinate is arbitrary, we adjusted the first principal coordinate of the genetic distance matrix by multiplying by $-1$ so that projection on a common set of coordinates would better visually reflect the patterns of geographic association.

**Origin of the Human Expansion.** Regressions on geographic distance from a center were performed by using each of 4,210 centers drawn from the surface of the earth as follows. By using a lattice of 200 longitudes and 79 latitudes constructed so that each lattice point represented an equal area, 4,210 lattice points on land were identified (excluding Antarctica and islands farther south than the southern tip of South America). Rivers and all lakes other than Huron, Michigan, Superior, Victoria, and the Caspian and Aral Seas were treated as land.

**The Relationship Between $F_{ST}$ and Heterozygosities.** Taking equation 5.12 from Weir (16), if $u_i$ denotes allele $u$ in population $i$, $l$ is the locus under consideration, and $\tilde{p}_{il}$ is the frequency at locus $l$ of allele $u$ in population $i$, then $\hat{\theta}$ (the estimator for $F_{ST}$) is

$$\hat{\theta} = \frac{\sum_i \frac{1}{2} \sum_u (\tilde{p}_{ilu} - \tilde{p}_{ilu})^2 - \frac{1}{2(2n-1)} \left( \sum_u \left( \tilde{p}_{ilu}^2 + \tilde{p}_{ilu}^2 \right) \right)}{\sum_i (1 - \sum_u \tilde{p}_{ilu} \tilde{p}_{ilu})}.$$  

{[3]}

Restricting our computations to one locus (removing $l$ from Eq. 3), we obtain for a sample of size $n$

$$\hat{\theta} = \frac{2n}{2(2n-1)} \sum_u \tilde{p}_{ilu}^2 + \frac{2n}{2(2n-1)} \sum_u \tilde{p}_{ilu}^2 - \sum_u \tilde{p}_{ilu} \tilde{p}_{ilu} - \frac{1}{2n-1}.$$

{[4]}

$\sum_u \tilde{p}_{ilu}^2$ is the homozygosity in population $i$ and is therefore equal to $1 - H_i$, where $H_i$ is the heterozygosity in population $i$. Assuming $2n/(2[2n-1]) \approx 1/2$, then Eq. 4 reduces to

$$\hat{\theta} = 1 - \frac{1}{1 - \sum_i \tilde{p}_{ilu} \tilde{p}_{ilu}} \left( \frac{H_i + H_2}{2} \right) \approx 1 - \frac{1}{1 - \sum_i \tilde{p}_{ilu} \tilde{p}_{ilu}},$$

{[5]}

assuming $(1/[2n-1])/(1 - \sum_i \tilde{p}_{ilu} \tilde{p}_{ilu})$ is small. If we fix population 1 as Africa and denote it by $\alpha$, and if we write

$$\gamma_i = 1 - \sum_u \tilde{p}_{ilu} \tilde{p}_{ilu},$$

{[6]}

then equating $F_{ST} = a + b \times (\text{geographic distance})$ with Eq. 5 it follows that an estimate for the heterozygosity in population $i$ is

$$H_i \approx 2 \gamma_i [1 - a - b \times (\text{geographic distance}) - H_\alpha].$$

{[7]}

It is only because geographic distance is a good predictor of $F_{ST}$ that this calculation can be made.

---

**Fig. 2.** Populations influencing the linear regression. The two plots are identical except that different features are highlighted in $A$ and $B$. The number representing each population is the rank of its influence on the regression, with 1 indicating the population whose removal from the data alters the regression by the greatest amount (see Materials and Methods and Table 2). All other points not involving comparisons with the populations of greatest influence are in gray. (A) Red 1 denotes comparisons including Karitiana; green 2, Maya; navy blue 3, Pima; and purple 4, Colombia. Black squares show comparisons between the American populations. Comparisons involving the Maya (labeled as 2) tend to produce smaller $F_{ST}$ values than are predicted by the regression line, and excluding the Maya from analysis increases $R^2$ to 0.8183. The slight increase in the error sum of squares of the regression when the Maya are included in the data set shows that they have little influence on the observed pattern. (B) Orange 5 denotes comparisons including Kalash; brown 6, San; and blue 7, Mbuti Pygmies. The black circle is the comparison between the San and Mbuti Pygmies. The black triangles are comparisons of the Kalash to the San and Mbuti. The Kalash have been identified as a genetic isolate (11) from the rest of Pakistan; here, comparisons of the Kalash with other Central/South Asian and East Asian populations produce large residuals, whereas comparisons with European and Middle Eastern groups do not, consistent with the closer relationships of the Kalash to groups in these regions than to groups in East Asia or to other groups in Pakistan (11, 27). The high $F_{ST}$ values observed in comparisons with the Mbuti Pygmies or the San, both hunter-gatherer populations, are likely to be a consequence of the deep genetic structure believed to exist in Africa and of the amount of genetic isolation these groups have experienced from other African populations (8, 28).
between migration, mutation, and drift. There clearly has not been time to reach equilibrium between the extremes of man’s inhabited range, or even within continents, in the very short evolutionary history of modern humans (29). An expansion of modern humans outward from a single center is an alternative way of producing a global correlation between geographic and genetic distances. Geological expansion events may have happened in many small steps, with each such migration involving a sampling from the previous subset of the original population. This sampling would have led to a stepwise increase in genetic drift and a concomitant decrease in genetic diversity: a serial founder effect (30, 31).

Genetic data are found to be in strong agreement with this expansion model. The rank order of continents by genetic diversity for Y-chromosomal and chromosome-21 polymorphisms correlates with the archaeologically estimated order in which modern humans entered into continents (32, 33), and expected heterozygosity (calculated by using 377 loci from the HGDP-CEPH data set) has been found to decrease linearly with distance from a possible site for the geographic origin of modern humans in East Africa (34). We have confirmed the latter observation by augmenting the 377 loci previously studied with 406 additional microsatellites from the same individuals (Fig. 4A).

Assuming that there was an initial site from which the human expansion occurred, Fig. 5 shows that the pattern of expected heterozygosities in the data set is best explained by an expansion originating in Africa. For each of 4,210 points on a lattice of latitudes and longitudes (see Materials and Methods), we regressed expected heterozygosity in the HGDP-CEPH populations on geographic distance to the lattice point (Fig. 5). The 936 locations in Africa used as origins resulted in $R^2$ values ranging from 0.757 to 0.870 (the SD of $R^2$ within Africa was 0.017), whereas $R^2$ using the 3,274 non-African locations as origins ranged from $1.67 \times 10^{-7}$ to 0.744 (the SD of $R^2$ outside of Africa was 0.245). Thus, no origin outside of Africa had the explanatory power of an origin anywhere in Africa (see also ref. 37). Because sampling was not very dense in Africa, especially in Eastern and Northern Africa, a larger sample might enable this approach to further localize the specific origin of the expansion.

Regressions based on origins in South America had the highest $R^2$ values of the non-African locations, but the correlation of expected heterozygosities with geographic distance to South America is positive, indicating that whereas heterozygosity decreases linearly with distance from Africa, it increases with distance from South America. These observations, together with the high genetic diversity in Africa and low diversity in the Americas, are consistent with an expansion from Africa, with South America being among the last places reached by migrating populations.

The linear relationships observed in Figs 4A and 1B are different depictions of the same phenomenon because pairwise $F_{ST}$ is directly related to the homozygosities of each population in the comparison (see Eq. 7) and is therefore inversely related to the populations’ heterozygosities (38). Suppose $i$ is any non-African population and $a$ is a fixed African population. We can regard $\gamma$ (see Eq. 6) as an index of the similarity of alleles between populations $i$ and $a$ where $p_u$ is the frequency of allele $u$ in the fixed African population $a$, $p_{iu}$ is the frequency of the allele $u$ in population $i$, and the sum is taken over all alleles $u$. Pooling all of the sub-Saharan African populations in the data set and averaging $\gamma$ across loci between Africa ($a$) and each non-African population ($i$) in the sample, we find that the mean of $\gamma_i$ is $\tilde{\gamma} = 0.196$ with a coefficient of variation of 1%. Substituting $\tilde{\gamma}$ and the values of $b$ and the intercept $a$ from the regression of $F_{ST}$ on geographic distance from Africa into Eq. 7, the estimate of the expected heterozygosities of all populations in the sample is within 3% of the observed values; the difference between the estimate $\hat{H}_i$ and the observed value has a SD of 0.0207. Thus, we can “transform” Figs 4A and 1B into each other almost without loss of
information, as is reflected by the similar explanatory power of the linear regressions of $F_{ST}$ ($R^2 / H = 0.78$) and of expected heterozygosity ($R^2 / H = 0.76$) on geographic distance.

Testing whether a serial founder effect could give rise to the decay of expected heterozygosity with distance observed in Fig. 4A requires appropriate demographic models for calculating the effect of drift. We performed simulations of evolutionary processes to assess whether we could recover a similar pattern to what was computed from the data as shown in Fig. 4A (37). Assume for simplicity that we begin with a parental population, and there are $n$ serial bottleneck episodes starting at the origin (the location of the parental population). In each bottleneck, a sample of individuals of size $N_b$ founds the next colony, which is established at some distance from the previous colony and which remains isolated from all other colonies. This subsampling generates a succession of colonies in time, each of which grows to a large size $K$ before generating the next colony in the chain. Each bottleneck episode decreases expected heterozygosity in the new colony by a factor of $1/(1 + 2N_b)$ (39). To be precise, this computation includes the drift effect only of the first generation after the bottleneck.

Based on this simple model of $n$ bottlenecks with $N_b$ founders at each bottleneck, an approximation for the total loss of expected heterozygosity from the beginning to the end of the expansion from the parental population due to the sequence of bottlenecks alone will be

$$\Delta H = n/(2N_b).$$

Fig. 4. The decay of heterozygosity plotted against geographic distance between populations and a possible origin of expansion. (A) Heterozygosity in the HGDP-CEPH populations against distance from Addis Ababa, Ethiopia (9N, 38E). Distances were corrected for large bodies of water. The equation of the regression line is heterozygosity $= 0.7682 + (6.52 \times 10^{-4}) \times$ (distance from Addis Ababa). $R^2 = 0.7630$. (B) Simulation results of the decay of heterozygosity with distance using a model of a serial founder effect. The simulation is based exclusively on mutation at a realistic rate and drift, as described in more detail in Supporting Text. The parameter values generating the simulation were chosen so as to fit the observed $\Delta H$ of $A$. The number of bottlenecks is $n = 100$, and the number of founders per bottleneck, $N_b$, is 250, which approximates the effective population size of a population of hunter–gatherers (35, 36). Other pairs of values of $n$ and $N_b$ in the same ratio would fit the data equally well, because their ratio is the main quantity affecting the slope. The equation of the fitted line is heterozygosity $= 0.8761 - 0.0012 \times$ (distance from the parental colony). $R^2 = 0.8587$.

Fig. 5. The origin of the human expansion. The color or shade of each of the 4,210 locations (shown as dots) indicates either a correlation coefficient $r$ or an $R^2$ value for the regression of expected heterozygosities in 53 HGDP-CEPH populations on geographic distance (corrected for large bodies of water) to the location displayed. Note that, for a simple linear regression, $r^2 = R^2$. Grayscale points indicate $R^2$ values, as shown by the gradient on the right, and correlation coefficients $r$ are displayed in Africa and South America to reflect the sign of the relationship between heterozygosity and geographic distance to locations in these continents. $R^2$ values range from 0.757 to 0.870 in Africa and from 0.519 to 0.659 in South America. The maximum value of $r$ ($=0.812$) is observed when the origin is (30S, 50.2W); the minimum value of $r$ (approximately $=0.933$) is observed when the origin is (4.3N, 12.8E).
Regressing heterozygosity on distance from the parental colony, we can estimate $\Delta H$ by calculating the difference between the intercept of the regression line and the fitted value for the last population in the expansion (the furthest population from the origin). In Fig. 4A, the observed $\Delta H$ is 0.12. Because $n$ and $N_0$ are unknown, Eq. 8 only allows the estimation of their ratio. Moreover, this simple model assumes no intermigration among colonies after their founding; it only accounts for genetic drift that occurs as a result of the bottlenecks in the serial founder effect, ignoring genetic drift during the growth period where the founding population increases in size to carrying capacity and (ii) while the population stays at carrying capacity as the subsequent colonies are formed. These components will increase the amount of drift experienced by populations over that which would ensue from a population of constant size $K$.

Simulation enables the evaluation of these components of the evolutionary process by using estimable quantities, such as the mutation rate of microsatellites and the sizes of populations (see Supporting Text, which is published as supporting information on the PNAS web site, for more discussion). Fig. 4B shows that simulation can produce heterozygosity values similar to those observed in the data set, giving a simulated value for $\Delta H$ of 0.12, very close to the observed value. $\Delta H_{\text{sim}}$ will differ from $\Delta H$ in Eq. 8 (see Supporting Text). The main assumption in the simulation (Fig. 4B) is that $N_0$, the number of founders at each bottleneck, is of the order of a hunter–gatherer tribe (35, 36).

**Discussion**

Geographic distance is a good predictor of genetic distance on a global scale (Fig. 1). The pattern’s robustness is indicated by our ability to reasonably explain anomalies (Fig. 2) based on what is generally believed to have occurred during the past 100,000 years of modern human history (29). We also find a close relationship between the correlation of $F_{ST}$ and geographic distance (Fig. 1) and the geographic pattern of heterozygosity across populations (Fig. 4A). An increase in genetic distance with geographic distance has been observed in the past and has been attributed to equilibrium models of isolation by distance, but simulation results show that the geographic pattern of heterozygositites in the HGDP-CEPH populations is consistent with a serial founder effect starting at a single origin. Further, the observed pattern of within-population diversity is best explained by an origin in Africa (Fig. 5). By studying the relationship between genetic and geographic distance, we can assess the relative importance of genetic drift and natural selection in determining the genetic variation observed among human populations. The average contribution of drift generated by the serial founder effect might be estimated from the properties of the regression in Figs. 1B and 4A. Because our regressions explain 76–78% of the observed genetic variation, this quantity is therefore an estimate of the minimum influence that drift, due to the serial founder effect, has on the total variation observed. In other words, the fraction of the variation in heterozygosity across human populations that is explained by drift is at least 76–78%.

We thank Saurabh Mahajan for bioinformatics assistance and Lynn Jorde and Montgomery Slatkin for helpful comments on the manuscript. This work was supported by National Institutes of Health Grants GM28216 and GM28428. S.R. is supported by a National Defense Science and Engineering Graduate fellowship. C.C.R. is supported by a fellowship from the Morrison Institute for Population and Resource Studies. N.A.R. is supported by a Burroughs Wellcome Fund Career Award in the Biomedical Sciences.