J. Protozool., 35(1), 1988, pp. 81–85

Linkage Disequilibrium in Natural Populations of *Trypanosoma cruzi* (Flagellate), the Agent of Chagas' Disease[1]

QIFA ZHANG,[2] M. TIBAYRENC,[3] and F. J. AYALA
Department of Genetics, University of California, Davis, California 95616

ABSTRACT. We have studied linkage disequilibrium in natural populations of *Trypanosoma cruzi*, the agent of Chagas' disease, by analyzing (i) a set of 524 stocks from the whole geographical range of the parasite, characterized at four gene loci coding for enzymes; (ii) a subsample of 121 stocks characterized at 12 enzyme loci; and (iii) a subset of 386 stocks from six locations in Bolivia, characterized by four enzyme loci. Our results show that the linkage disequilibrium reaches the maximum possible value, given the observed allelic frequencies, for almost all the locus pairs. This result is most consistent with the hypothesis that genetic recombination is absent or very rare in *T. cruzi* natural populations. Partition of the linkage disequilibrium variance for the six Bolivian populations shows that both inter- and intrapopulation components are substantial and that the relationships among the components are $D^2_{IS} < D^2_{ST}$, and $D'^2_{IS} < D'^2_{ST}$. These inequalities are interpreted as the result of an interplay between genetic drift, rare or absent mating, and clonal selection in generating linkage disequilibrium in *T. cruzi* populations.

We have earlier proposed on the basis of isozyme data that Mendelian sexuality is absent or very rare in natural populations of *Trypanosoma* (*Schizotrypanum*) *cruzi* (8), and that these populations exhibit a complex multiclonal structure (12). These hypotheses are supported by the presence of a strong linkage disequilibrium in that some isozyme genotypes are ubiquitous and are sampled numerous times in places very distant from each other and in very different ecosystems, whereas the vast majority of the theoretically possible recombinants are lacking. We present here the results of a statistical study of this linkage disequilibrium.

MATERIALS AND METHODS

Two data sets were obtained from studies of isozyme variation in *Trypanosoma cruzi* stocks collected from many localities of North, Central, and South America (Table I). The first set was a study of four enzyme loci in each of 524 stocks. The four loci studied are as follows: glucose-6-phosphate isomerase (*Gpi*), isocitrate dehydrogenase (*Idh*), malic enzyme 2 (*Me-2*), and phosphoglucomutase (*Pgm*) (see ref. 12). Twenty different genotypes were found (Table II); the genotypes are inferred on the assumption that the genome of *T. cruzi* is diploid (7, 9, 12). The second set of data consists of a subsample of 121 of the 524 stocks, assayed for 11 additional isozyme loci; the 15 loci combined yield 43 different genotypes (12). The 11 additional loci are aconitase (*Acon*), adenylate kinase (*Adk*), glucose-6-phosphate dehydrogenase (*G6pd*), glutamate dehydrogenase NAD+ (*Gdh Nad+*), glutamate dehydrogenase NADP+ (*Gdh Nadp+*), leucine aminopeptidase (*Lap*), malate dehydrogenase (*Mdh*), malic enzyme 1 (*Me-1*), peptidase 1 (*Pep-1*), peptidase 2 (*Pep-2*), and 6-phosphogluconate dehydrogenase (*6Pgd*).

A linkage disequilibrium variance was calculated for the four loci assayed in the total sample of 524 individuals and for 12 loci in the subsample of 121 stocks. Among these 121 stocks, four exhibited mixtures of two different genotypes (12). In the calculations, different genotypes identified within the same host were counted as separate stocks. Three loci were not used for the linkage disequilibrium study, because they are monomorphic or nearly so: *Acon*, *Adk*, and *Mdh*.

Our samples include more than 25 stocks from each of six localities in Bolivia, which made it possible to evaluate the relative magnitude of linkage disequilibrium due to geographical differentiation. Linkage disequilibrium parameters proposed by Ohta (5) were calculated for these six localities using the four-locus data. A randomization test (4) was performed for each parameter in order to determine the statistical significance of the estimates.

The many alleles observed at most of the loci studied yielded an enormous number of possible linkage disequilibrium associations. To bring the data to a manageable size, we have recorded the most frequent allele of each locus as "allele 1" and lumped together all the remaining alleles as "allele 2." This collapsing of alleles usually leads to underestimating linkage disequilibria (13, 14). The data show that at nearly every locus the most frequent allele has a frequency above 0.50 whereas other alleles have frequencies below 0.20. Thus, the loss of information due to lumping alleles is not particularly large in the present analysis.

RESULTS

Linkage disequilibrium among four allozyme loci in the total sample. The first set of data included 524 stocks assayed for variation at four enzyme loci (*Gpi*, *Idh*, *Me-2*, and *Pgm*) (see Tables I and II). The frequencies of allele 1 at these four loci were 0.601, 0.607, 0.590, and 0.601, respectively. All the pairwise linkage disequilibrium values were at, or near, the maximum possible for the observed allelic frequencies (Table III).

The statistical significance of the observed linkage disequilibrium values was tested as follows. First, we obtained a single statistic linkage disequilibrium variance for the population as:

$$D^2_P = \frac{2}{n(n-1)} \sum_{i=j+1} D^2_{ij} \qquad (A)$$

for all i, j = 1, 2, . . . , n, where D_{IJ} is the usual linkage disequilibrium value between the i^{th} and the j^{th} loci (1), and n is the number of loci studied. Second, because the statistical properties of either D_{IP} or D^2_P are unknown, we performed a randomization test (4) to determine the statistical significance of the observed D^2_P. The allelic combinations of the four loci were randomized by permutating the allelic composition of each locus

[1] We are indebted to C. Camacho and L. Echalar (IBBA, La Paz, Bolivia) for valuable technical collaboration in growing *T. cruzi* stocks. We thank the following scientists for some of the *T. cruzi* stocks: J. P. Dedet (Institut Pasteur in Cayenne, French Guiana), P. Desjeux (IBBA in La Paz, Bolivia), C. La Fuente (CENETROP in Santa Cruz, Bolivia), J. L. Lemesre and F. Le Pont (IBBA in La Paz, Bolivia), D. Le Ray (IMT "Prince Leopold" in Antwerp, Belgium), M. A. Miles (London School of Tropical Medicine, U. K.), and J. Theis (U. C. Davis, California). The experimental part of this study was performed at the Instituto Boliviano de Biologia de Altura (IBBA) in La Paz, Bolivia, with Drs. G. Antezana, Y. Carlier, and P. Desjeux as directors and with financial support from the French Technical Cooperation and from the Ministère de l'Industrie et de la Recherche (PVD/81/L-1423).

[2] Present address: Department of Agronomy, Huazhong Agricultural University, Wuhan, China.

[3] Present address: ORSTOM, 2051 Avenue du Val de Montferrand, BP 5045, 34032 Montpellier Cedex, France.

TABLE I. *Geographic origin of the 524 stocks of* Trypanosoma cruzi *studied at four gene loci and their composite isozyme genotype or zymodene.*[a]

Locality	Zymodeme	Number of stocks
Colombia		
Puerto Ele	2	5
Ecuador		
Guayaquil	10	1
Peru		
Ucayali	10	1
Bolivia		
Yungas	2	10
	39	11
Chiwisivi	2	141
	39	2
Cochabamba	2	15
	39	10
	40	1
Comarapa	2	9
	39	8
	32	4
Santa Cruz	2	7
	16	1
	28	2
	39	24
	40	3
Sucre	2	23
	10	1
	32	8
	39	16
Camiri	2	4
	37	2
	38	1
	39	6
	40	7
Tupiza	2	45
	32	9
	39	31
	40	14
Tarija	2	8
	38	1
	39	20
	40	4
Alto Beni	2	1
	25	1
Potosi	39	1
Vallegrande	2	8
Venezuela		
Miranda	2	1
Portuguese	2	1
Cojedes	2	4
Aragua	2	1
Barinas	2	1
Carabobo	2	1
French Guiana		
Montjoly	1	1
Montsinery	2	5
Cacao	2	2
	3	1
	8	1
Panamana	2	1
Chile		
Cachicuyu	2	2
	39	3
Cucumen	32	1
Monte Patria	32	1
	39	1
Arrayan	32	1
Chanaral	32	1
Locality X[b]	40	1
Brazil		
Goias Goiana	2	1
	31	2
Locality X[b]	3	1
Bélem	2	2
	27	1
	35	1
	36	1
Espirito Santo	2	1
Sao Paulo	2	4
	31	3
Minas Gerais	2	1
Locality X[b]	10	1
Valle Grande	39	1
Rio Grande	40	2
Bahia	30	1
Honduras		
Tegucigalpa	10	1
Mexico		
Locality X[b]	2	1
USA		
Locality X[b]	14	2

[a] See Table II.
[b] Exact location is not known.

independently of the other three loci and D^2_P was calculated for each set of the randomized data. This procedure was repeated 150×, and the resulting D^2_P values were plotted in a distribution histogram. Then the observed D^2_P was compared to the distribution obtained by the random permutation. If fewer than 5% of the values of D^2_P in the distribution were larger than the observed D^2_P, the observed D^2_P would be considered statistically significantly greater than zero at the probability level $P = 0.05$. This procedure is believed to provide a nearly exact test for the null hypothesis. The largest value of D^2_P derived from the randomization test was 0.000375, whereas the D^2_P estimated from the experimental data was 0.0547, about 145× as large as the largest value obtained based on the null hypothesis that alleles of different loci are associated at random. Thus, linkage disequilibrium is indeed highly significant in the total sample of *T. cruzi*.

Linkage disequilibrium for 12 enzyme loci. In order to study the extent of linkage disequilibrium among loci representing a larger portion of the genome, we analyzed data for 12 polymorphic loci in 121 *T. cruzi* stocks. The linkage disequilibrium values were calculated for all the 66 pairwise comparisons possible for a data set of 12 loci. Most of these linkage disequilibria were nearly at the maximum values for the observed sets of allelic frequencies. The disequilibrium variance (D^2_P) for this data set was 0.0153, whereas the D^2_P value drived from random permutation fell between 0.000189 and 0.000568. Therefore, the linkage disequilibrium among these 12 enzyme loci is again highly significant.

TABLE II. *The 20 genotypes (zymodemes) inferred from the study of four isozyme loci. Alleles at each locus are numbered consecutively starting with 1.*[a]

Genotype	*Gpi*	*Idh*	*Me-2*	*Pgm*
1	*5/5*	*1/1*	*4/4*	*2/7*
2	*5/5*	*1/1*	*4/4*	*3/3*
3	*5/6*	*1/1*	*4/4*	*3/3*
8	*5/5*	*1/1*	*4/4*	*8/8*
10	*5/5*	*1/1*	*4/7*	*3/3*
14	*5/5*	*1/1*	*7/7*	*3/3*
16	*6/6*	*1/1*	*2/4*	*1/3*
25	*5/5*	*1/1*	*2/4*	*1/3*
26	*4/4*	*2/2*	*3/3*	*8/8*
27	*4/4*	*2/2*	*3/3*	*9/9*
28	*4/4*	*2/2*	*3/3*	*4/4*
30	*1/3*	*2/2*	*6/6*	*11/11*
31	*3/3*	*2/2*	*6/6*	*11/11*
32	*3/3*	*2/2*	*6/6*	*10/12*
35	*4/4*	*2/2*	*5/5*	*5/5*
36	*4/4*	*2/2*	*5/5*	*9/9*
37	*4/4*	*2/2*	*5/5*	*6/10*
38	*2/4*	*2/2*	*5/5*	*10/10*
39	*2/4*	*2/2*	*5/5*	*6/10*
40	*3/4*	*2/2*	*5/5*	*4/11*

[a] Numbering of the zymodemes is according to ref. 12.

Linkage disequilibrium among geographically separated populations. We evaluated the geographical component of linkage disequilibrium using the data for six Bolivian localities for which suitable sample sizes were available. The localities are Chiwisivi, Cochabamba, Santa Cruz, Sucre, Tupiza, and Tarija, with 143, 26, 37, 48, 99, and 33 stocks, respectively (Table I). We followed the calculation procedures and designations of linkage disequilibrium components given by Ohta (5). For a subdivided large population, the total linkage disequilibrium variance can be partitioned into a component due to genetic differentiation among subpopulations and a component due to linkage disequilibria within subpopulations. Ohta (5) derived two ways of partitioning the total linkage disequilibrium variance of a large population into corresponding components based on two alternative views of the genetic architecture of subdivided populations.

According to the first view, the total variance of linkage disequilibrium includes a component due to allelic differentiation at paired loci (D^2_{ST}) and a component due to linkage disequilibria within subpopulations (D^2_{IS}), which Ohta defined as follows:

$$D^2_{IS} = E\left[\sum_{i,j} (g_{ij,k} - x_{i,k}y_{j,k})^2\right]$$

and

$$D^2_{ST} = E\left[\sum_{i,j} (x_{i,k}y_{j,k} - \bar{x}_i\bar{y}_j)^2\right]$$

where $g_{ij,k}$ is the frequency of the gametic type consisting of the i^{th} allele of the first locus and the j^{th} allele at the second locus in the k^{th} population, and $x_{i,k}$ and $y_{j,k}$ are the frequencies of the i^{th} allele at the first locus and the j^{th} allele of the second locus, respectively. The expectation, E, is taken over the subpopulations.

Alternatively, the total variance of linkage disequilibrium can be viewed as composed of a portion accounted for by gametic differentiation at paired loci among subpopulations (D'^2_{IS}), and a portion representing linkage disequilibrium in the total population (D'^2_{ST}), which are defined:

$$D'^2_{IS} = E\left[\sum_{i,j} (g_{ij,k} - \bar{g}_{ij})^2\right]$$

and

$$D'^2_{ST} = E\left[\sum_{i,j} (\bar{g}_{ij} - \bar{x}_i\bar{y}_j)^2\right].$$

The total variance (D^2_{IT}) was defined by:

$$D^2_{IT} = E\left[\sum_{i,j} (g_{ij,k} - \bar{x}_i\bar{y}_j)^2\right].$$

It can be shown that these parameters are related as follows:

$$D^2_{IT} > D^2_{IS} + D^2_{ST} \quad \text{and} \quad D^2_{IT} = D'^2_{IS} + D'^2_{ST}. \qquad (B)$$

Because D^2_{IT} is the total variance, significant values greater than zero for any one of the four components will ensure significant differences from zero for D^2_{IT}. Thus, it is not necessary to test D^2_{IT} for statistical significance if the other components are significant. A randomization test was applied to each of the four components. Each randomization test corresponded to a particular null hypothesis and involved a specific permutation procedure. For example, in testing the component D^2_{IS}, the null hypothesis specified was that there was no linkage disequilibrium within any local population. The permutation procedure randomized the genotype at each locus independently of all the other loci within each local population; D^2_{IS} was calculated for the randomized data. This procedure was repeated 150× as before.

The observed disequilibrium for the intrapopulation component (D^2_{IS}) was about 20× as large as the upper limit observed for this component in the randomization test (0.1585 vs. 0.0083, see Table IV). The variance of the disequilibrium for the total

TABLE III. *Observed pairwise linkage disequilibrium (D_{OBS}) and their theoretically maximum values (D_{MAX}) for the observed allelic frequencies. Six pairwise comparisons are made between the four loci assayed in 524 stocks of* T. cruzi.

	D_{OBS}	D_{MAX}	D_{OBS}/D_{MAX}
D_{12}	0.2362	0.2362	1.00
D_{13}	0.2314	0.2354	0.98
D_{14}	0.2360	0.2398	0.98
D_{23}	0.2319	0.2319	1.00
D_{24}	0.2363	0.2363	1.00
D_{34}	0.2314	0.2354	0.98

TABLE IV. *Observed variance components of linkage disequilibrium for four allozyme loci in six Bolivian populations of* T. cruzi[a] *and range obtained in the randomization tests for the same components.*

	Observed	Range obtained in the randomization tests
D^2_{IS}	0.1585	0.0000 to 0.0083
D^2_{ST}	0.2068	0.0000 to 0.0417
D'^2_{IS}	0.1681	0.0000 to 0.0300
D'^2_{ST}	0.2188	0.0000 to 0.0015
D^2_{IT}	0.3856	

[a] See Table I.

population (D'^2_{ST}) was 0.2188 (Table IV) which was more than 100× larger than the largest D'^2_{ST} obtained in the 150 random permutations. Thus, both of these variance components are very highly significant, and the disequilibrium component for the total population is larger than the intrapopulation component. The observed D^2_{ST} was 0.2068 (about 5× larger than the largest value in the random permutation test); the observed D'^2_{IS} was 0.1681 (5.6× as large as the largest value in the randomization test), indicating that the geographical differentiation due to gametic differences as well as that due to allelic differences for paired loci are also highly significant. Detailed information regarding the relative importance of each component can be obtained from the partitioning proposed by the inequality and equation (B). Using the inequality, 43% of the joint variance (i.e. $D^2_{IS} + D^2_{ST}$) is due to linkage disequilibrium within local populations, and about 57% is due to geographical differentiation. The partitioning by the second method (the equation) indicates that about 43% of the total variance is due to geographical differentiation (D'^2_{IS}) whereas 57% can be accounted for by the disequilibrium in the total population (D'^2_{ST}).

DISCUSSION

We have analyzed linkage disequilibrium between various pairs of gene loci coding for enzymes in a large number of stocks derived from various hosts and representing a significant eco-geographical range of the distribution of the species *T. cruzi* (see ref. 12 and Table I). The analyses demonstrate that there is large linkage disequilibrium between alleles at paired loci. The disequilibrium obtains for all 12 polymorphic loci and hence, presumably, over a substantial portion of the genome of the parasite. The linkage disequilibrium values for almost all the locus pairs studied are at about the maximum possible for the observed allelic frequencies, which indicates that associations among alleles at different loci are almost complete and that recombination is rare or absent even within local populations. In this connection, we note that in the wild barley, *Hordeum spontaneum,* a predominantly self-fertilizing species (selfing rate above 99%), Zhang, Saghai-Maroff, & Allard (unpubl. data) have found that although there were large amounts of linkage disequilibria, the observed number of different genotypes was much larger for any given number of loci than the number observed in the present study. This indicates that the small amount of outcrossing observed in the wild barley (under 1%) is sufficient to generate and maintain new gametic types in the population. Thus, the results obtained in the present study support the proposition that genetic exchanges are rare or absent in *T. cruzi* natural populations and that these populations have an essentially clonal structure (8, 12).

The hypothesis that the apparent absence of recombination can be accounted for by founder effects that dispel the opportunity for genetic exchange between different *T. cruzi* genotypes (3) is not corroborated by careful analysis of the microdistribution of *T. cruzi* genotypes in Bolivia; radically dissimilar genotypes very often occur in close sympatry (same human host, same insect vector), which provides maximum opportunity for mating (2, 6, 9–11).

Ohta (5) has suggested that the relationships among the four components of total disequilibrium are useful for inferring the role of the evolutionary processes yielding the observed non-random associations between alleles at different loci. She has shown that the relationships $D^2_{IS} < D^2_{ST}$ and $D'^2_{IS} > D'^2_{ST}$ should hold whenever migration among subdivisions is limited. On the contrary, when natural selection is primarily responsible for linkage disequilibrium but not for local differentiation, one would predict $D^2_{ST} < D^2_{IS}$ and $D'^2_{IS} < D'^2_{ST}$, because the number of gametes with favorable combinations of alleles would increase in every population. The relationships in our analysis of four isozyme loci in six localities of Bolivia are $D^2_{ST} > D^2_{IS}$ and $D'^2_{IS} < D'^2_{ST}$. The relationship $D^2_{ST} > D^2_{IS}$ suggests that allele frequencies among local populations are very different as a result of genetic drift due to limited migration. But the relationship $D'^2_{IS} < D'^2_{ST}$ indicates that the differentiation of gametic frequencies is small compared to the linkage disequilibrium in the total population. In fact, all the allelic associations are in the same direction in all the six localities for all the six possible locus pairs. Although absence of recombination is probably largely responsible for the observed complete associations, mutations at one locus or another could break such perfect associations. Thus, the population structure reflected by $D'^2_{IS} < D'^2_{ST}$ suggests that clonal selection favoring particular gene combinations (particular natural clones) is partly responsible for the observed linkage disequilibria.

In conclusion, the two relationships observed, $D^2_{IS} < D^2_{ST}$ and $D'^2_{IS} < D'^2_{ST}$, seem to indicate an interplay between genetic drift, absence of recombination, and clonal selection. Genetic drift would lead to differentiation of allele frequencies among local populations, whereas rare or absent recombination would maintain favored gene combinations in all the subpopulations. Clonal selection could interfere in both directions; some gene combinations would be favored in many or all populations ("generalist genotypes"), but other gene combinations could be selected for by local factors (leading to differentiated populations). We have shown (12) that some *T. cruzi* genotypes are widespread over very large geographical areas and various ecological conditions and so behave like "generalist genotypes." This situation could be due to clonal selection favoring some gene combinations over many populations. On the other hand, we have shown in Bolivia (11) a highly significant correlation between the frequencies of different *T. cruzi* genotypes on the one side and altitude and longitude on the other side. This is compatible with the hypothesis of differential local adaptation of *T. cruzi* genotypes to climatic factors; in this case, clonal selection would lead to differentiated populations.

LITERATURE CITED

1. Ayala, F. J. & Kiger, J. A. 1984. *Modern Genetics,* 2nd ed. Benjamin/Cummings, Menlo Park, California, pp. 847–848.

2. Brénière, S. F., Tibayrenc, M., Antezana, G., Pavon, J., Carrasco, R., Selaès, H. & Desjeux, P. 1985. Résultats préliminaires en faveur d'une relation faible ou inexistante entre les formes cliniques de la maladie de Chagas et les souches isoenzymatiques de *Trypanosoma cruzi. C. R. Acad. Sci. Paris,* **300**: 555–558.

3. Cibulskis, R. E. 1985. The microdistribution of *Trypanosoma cruzi. Trans. R. Soc. Trop. Med. Hyg.,* **79**: 138–139.

4. Kempthorne, O. 1955. The randomization theory of experimental inference. *J. Amer. Stat. Assoc.,* **50**: 946–967.

5. Ohta, T. 1982. Linkage disequilibrium due to random genetic drift in finite subdivided populations. *Proc. Natl. Acad. Sci. USA,* 79: 1940–1944.

6. Tibayrenc, M. 1985. On the microdistribution and sexuality of *Trypanosoma cruzi. Trans. R. Soc. Trop. Med. Hyg.,* **79**: 882–883.

7. Tibayrenc, M., Cariou, M. L. & Solignac, M. 1981. Interprétation génétique des zymogrammes de flagellés des genres *Trypanosoma* et *Leishmania. C. R. Acad. Sci. Paris,* **292**: 623–625.

8. Tibayrenc, M., Cariou, M. L., Solignac, M. & Carlier, Y. 1981. Arguments génétiques contre l'existence d'une sexualité actuelle chez *Trypanosoma cruzi*; implications taxinomiques. *C.R. Acad. Sci. Paris,* **293**: 207–209.

9. Tibayrenc, M., Cariou, M. L., Solignac, M., Dedet, J. P., Poch, O. & Desjeux, P. 1985. New electrophoretic evidence of genetic variation and diploidy in *Trypanosoma cruzi,* the causative agent of Chagas' disease. *Genetica,* **67**: 223–230.

10. Tibayrenc, M., Echalar, L., Dujardin, J. P., Poch, O. & Desjeux,

P. 1984. The microdistribution of isoenzymic strains of *Trypanosoma cruzi* in Southern Bolivia: new isoenzyme profiles and further arguments against Mendelian sexuality. *Trans. R. Soc. Trop. Med. Hyg.*, **78**: 519–525.

11. Tibayrenc, M., Hoffmann, A., Poch, O., Echalar, L., Le Pont, F., Lemesre, J. L., Desjeux, P. & Ayala, F. J. 1986. Additional data on *Trypanosoma cruzi* isozymic strains encountered in Bolivian domestic transmission cycles. *Trans. R. Soc. Trop. Med. Hyg.*, **80**: 442–447.

12. Tibayrenc, M., Ward, P., Moya, A. & Ayala, F. J. 1986. Natural populations of *Trypanosoma cruzi,* the agent of Chagas' disease, have a complex multiclonal structure. *Proc. Natl. Acad. Sci. USA,* **83**: 115–119.

13. Weir, B. S. & Cockerham, C. 1978. Testing hypothesis about linkage disequilibrium with multiple alleles. *Genetics,* **88**: 633–642.

14. Zouros, E., Golding, G. B. & Mackey, T. F. C. 1977. The effect of combining alleles into electrophoretic classes on detecting linkage disequilibrium. *Genetics,* **85**: 543–556.

Received 24 II 87; accepted 14 X 87