Distribution of genome-wide linkage disequilibrium based on microsatellite loci in the Samoan population

Hui-Ju Tsai,[1] Guangyun Sun,[2] Diane Smelser,[2] Satupaitea Viali,[3] Joseph Tufa,[4] Li Jin,[2] Daniel E. Weeks,[1,5]† Stephen T. McGarvey[6†] and Ranjan Deka[2†]*

[1]Department of Human Genetics, University of Pittsburgh, Pittsburgh, PA 15261, USA
[2]Department of Environmental Health, University of Cincinnati, Cincinnati, OH 45267, USA
[3]Department of Health, Apia, Samoa
[4]Department of Health, Pago Pago, American Samoa
[5]Department of Biostatistics, University of Pittsburgh, Pittsburgh, PA 15261, USA
[6]International Health Institute and Department of Community Health, Brown University, Providence, RI 02912, USA
**Correspondence to*: Tel: +1 412 624 5388; Fax: +1 412 624 3020; E-mail: dweeks@watson.hgen.pitt.edu
† These three authors have contributed equally to and co-directed this study.

Date received (in revised form): 28th May 2004

Abstract

Whole genome-wide scanning for susceptibility loci based on linkage disequilibrium (LD) has been proposed as a powerful strategy for mapping common complex diseases, especially in isolated populations. We recruited 389 individuals from 175 families in the US territory of American Samoa, and 96 unrelated individuals from American Samoa and the independent country of Samoa in order to examine background LD by using a 10 centimorgan (cM) map containing 381 autosomal and 18 X-linked microsatellite markers. We tested the relationship between LD and recombination fraction by fitting a regression model. We estimated a slope of −0.021 (SE 0.00354; $p < 0.0001$). Based on our results, LD in the Samoan population decays steadily as the recombination fraction between autosomal markers increases. The patterns of LD observed in the Samoan population are quite similar to those previously observed in Palau but markedly contrast with those observed in a non-isolated Caucasian sample, where there is essentially no marker-to-marker LD. Our analyses support the hypothesis of a recent bottleneck, which is consistent with the known demographic history of the Samoan population. Furthermore, population substructure tests support the hypothesis that self-identified Samoans represent one homogenous genetic population.

Keywords: *association mapping, linkage disequilibrium, isolated population, Samoa*

Introduction

Linkage disequilibrium (LD), the non-random association between loci, can aid in genetic mapping of complex diseases. It has been proposed that under high-density genetic maps, genome-wide LD scanning can be a powerful approach for searching for susceptibility genes determining complex diseases.[1] There have been several debates on the usefulness of isolated populations for mapping susceptibility genes determining complex diseases[2,3] and whether the extent of genome-wide LD is indeed larger in such populations compared with general populations.[4–6]. Boehnke[7] notes that negative findings from a few populations, or simulation results,[8] should be elaborated with caution, and even if the extent of LD is similar across populations, the benefits of an isolated population living in a relatively homogeneous environment and the ease of study should not be ignored.[9–13]

The Samoans of the Western Pacific represent one of the best examples of an isolated population. Archaeological and linguistic evidence[14] indicate a rapid eastward migration of populations into the western Pacific from Southern China, which took place about 4,000 to 5,000 years before the present (BP). By about 3,000 years BP, archaeological evidence indicates that Polynesian culture was established and flourished in Tonga and Samoa, well before further eastward expansion.[15,16] This Express Train model of Polynesian settlement[17] is supported by mtDNA data.[18–20] An alternative model, the Entangled Bank, proposed by Terrell,[21] asserts a neighbouring homeland for the Polynesians in Melanesia, in which the Polynesians evolved in a complex nexus of interactions among the already settled Pacific islanders. This is supported by nuclear DNA data.[22] Our previous work based on Y-chromosome SNP haplotypes, however, found no Melanesian-specific haplotypes among the Polynesians, particularly

the Samoans.[23] Samoans had only four haplotypes out of the 15 observed in the entire region of study comprising South-East Asia, Melanesia, Micronesia and Polynesia. In other work with microsatellite data, we also reported a significant reduction in genetic diversity among the Samoans compared with large cosmopolitan populations.[24–26] Further analysis of microsatellite data indicated a small effective population size and associated bottleneck events during Samoan history.[27]

Reconstruction of the population history of the Samoan islands is difficult, and estimates of population size through time, including at the time of first European contact and for the subsequent 100 years, remain debatable.[28,29] Nonetheless, it is important to describe the archaeological evidence and population genetic interpretations of Samoan demographic history. The original settlers of Polynesia were thought to be small in number, however, the ideal nutritional and disease ecology allowed rapid and sustained population growth up until the time of first European contact.[16] Detailed and well-dated archaeological and ethnohistorical evidence indicates that prior to European contact, Samoan villages were located at all levels on the mountain slopes (including at the top) in both what is now Samoa and American Samoa.[15,16] The archaeological findings of widespread population, beyond the littoral area, are consistent with estimates of the maximum population density, or human carrying capacity, derived from ethnographic work on agricultural intensity and population on other Polynesian islands.[16] Based on the wealth of archaeological and ethnographic data, contemporary scholars assert that the pre-European contact population of the Samoan islands could not have comprised fewer than 100,000 people and could have been as large as 300,000 people.[15,16]

The Samoan population suffered a significant depopulation after European contact, which has been attributed largely to the impact of introduced diseases.[16] The earliest reports of Samoan population size were made in the middle of the 19th century, many decades after European contact, which occurred first in 1722, and again in 1768 and 1787, with steady contact established only in 1836. From 1849 until 1900, the population size estimates for the Samoan islands vary between 30,000 to 40,000 individuals.[28] Throughout the 19th century, there were documented epidemics of infectious disease, chiefly influenza, and estimates of high mortality attributable to these waves of disease.[29,30] The impact of infectious diseases is assumed to have started soon after European contact in the later 18th century. By the early 20th century, the population had again increased. Approximate population sizes for the Samoan islands in the 20th century are provided by McArthur[29] for several time periods; these include: about 39,800 for the period 1900–1910; about 50,000 in 1930; and about 69,500 in 1940. Thus, it is quite clear that a massive depopulation occurred in the Samoas after European contact and that population growth was stagnant until the early 20th century, when it grew rapidly. Based on the 2000 census of American Samoa (performed with the aid of the US Census Bureau), the population of ethnic Polynesians in American Samoa is 55,704. Based on projections from the 2001 census of Samoa, the population of ethnic Samoans in 2003 was estimated at approximately 165,000 in independent Samoa.

The combination of genetic, archaeological and demographic evidence strongly suggests that the Samoan population was established by a relatively small number of founders, has been an isolated population and experienced a reduction in population size about 200–300 years ago. These population history events are very likely to have influenced the patterns of LD in the contemporary Samoan population.

In this study, we examined the distribution of genome-wide LD in Samoa as a function of recombination based on marker data from a 10 centimorgan (cM) genome-wide scan. We also tested for bottleneck effects and population substructure.

Materials and methods

Subjects

A sample of 390 individuals (201 males, 189 females) from 177 families was recruited in American Samoa using the Department of Health diabetes registry as part of a genome-wide study of type 2 diabetes susceptibility genes.[31] Subjects were asked about their Samoan ancestry in order to limit study participation to those who reported that all four grandparents were of Samoan ethnicity, without European or Asian ancestry. Study protocols were approved by the Institutional Review Board of the Miriam Hospital, Providence, RI, USA. Written informed consent was obtained from all participants.

We also sampled 96 unrelated individuals (50 males, 46 females), 40 recruited from American Samoa and 56 from Samoa. None of these individuals had diabetes and they all self-reported that all four grandparents were Samoan. These samples were derived from our previous longitudinal study of adiposity and cardiovascular disease risk factors in American Samoa and Samoa from 1990 to 1995.[32,33]

To provide a measure of marker-to-marker LD in a 'typical' outbred population, 333 unrelated North American Caucasians of European ancestry (172 males, 161 females) were selected, regardless of their disease status, from 229 pedigrees, each of which contained at least one affected individual with confirmed ankylosing spondylitis. Specifically, the 'founders' or the parents of the pedigrees were selected. The samples were collected through the North American Spondylitis Consortium (NASC). The genotyping data analysed in the present study contain information neither on individual identification nor on disease status of each individual. Genotyping protocols used in genotyping NASC individuals are identical to those used in this study.

Genotyping

Buffy coats were prepared from 10 ml of blood in the field following standard protocols and shipped on dry ice to the laboratory in Cincinnati. DNA was isolated from buffy coats using the Puregene DNA isolation kit (Gentra Systems Inc) quantitated and arrayed in 96-well microtitre plates.

The genome scan was conducted with the Applied Biosystems Inc (ABI) PRISM linkage mapping set version 2, consisting of 400 microsatellite markers, with fluorescently labelled primers, spaced at an average 10 cM distance between markers. We conducted multiplex polymerase chain reaction (PCR) amplification (three to five markers in each PCR reaction) in a 7.5 μl final reaction volume containing ~20 ng of genomic DNA and ~0.8–1.0 μl of AmpliTaq Gold™ DNA polymerase (5 U/μl). Initial incubation occurred for 12 minutes at 95°C; the first amplification was carried out for approximately 10–15 cycles in the PE GeneAmp™ 9600 thermal cycler using the following parameters: denaturation at 94°C for 15 seconds, annealing at 55°C for 15 seconds and extension at 72°C for 30 seconds. The second amplification was carried out for approximately 20–23 cycles using the following parameters: first denaturation at 89°C for 15 seconds, annealing at 55°C for 15 seconds and extension at 72°C for 30 seconds; then denaturation at 94°C for 15 seconds, annealing at 55°C for 15 seconds and extension at 72°C for 30 seconds; and then a final extension at 72°C for ten minutes and an overnight incubation at 4°C.

Amplified DNA products underwent gel electrophoresis on an ABI 377 DNA sequencer using internal standard Gene Scan-400 ROX (PE Applied Biosystems) for 2.5 hours at constant power (3000 V, 60 mA, 200 W) and at 51°C. For quality control, a negative control and two positive control samples of known genotype [Centre du Etude Polymorphisme Humain (CEPH) sample 1347-02] were run on each gel. GeneScan 3.1 and Genotyper 2.5 (PE Applied Biosystems) software were used for sizing and genotyping, respectively.

Statistical methods

Measure of LD. We computed a multi-allelic version of the D' statistic,[34] which we define here using the notation found in the GOLD program documentation.[35] Consider two markers A and B. Let $n_{i\bullet}$ be the number of haplotypes carrying allele i at locus A, $n_{\bullet j}$ be the number of haplotypes carrying allele j at locus B and n_{ij} be the number of haplotypes carrying allele i at locus A and allele j at locus B. If N is the total number of haplotypes, then the allele frequencies $p_{i\bullet}$ and $p_{\bullet j}$ and haplotypes frequencies p_{ij} can be estimated as:

$$p_{i\bullet} = \frac{n_{i\bullet}}{N}, p_{\bullet j} = \frac{n_{\bullet j}}{N}, p_{ij} = \frac{n_{ij}}{N}$$

The multi-allelic definition of D' is then:

$$D_{ij} = p_{ij} - p_{i\bullet}p_{\bullet j}$$

$$D_{ij,\max} = \begin{cases} \min\left[p_{i\bullet}p_{\bullet j}, (1-p_{i\bullet})(1-p_{\bullet j})\right] & D_{ij} < 0 \\ \min\left[(1-p_{i\bullet})p_{\bullet j}, p_{i\bullet}(1-p_{\bullet j})\right] & D_{ij} \geq 0 \end{cases}$$

$$D' = \sum_i \sum_j p_{i\bullet}p_{\bullet j}\frac{|D_{ij}|}{D_{ij,\max}}$$

Haplotype frequency estimation and LD testing

For both the Samoan and the NASC sample, we used the same sets of 381 autosomal microsatellite markers and 18 X-linked microsatellite markers to evaluate LD between all pairs of markers on the same chromosome; there were 3,531 autosomal marker pairs and 153 X-linked marker pairs. The 'ldmax' program, which is part of the GOLD package,[35] was used to estimate haplotype frequencies for each marker pair, using the expectation-maximisation (EM) algorithm,[36,37] and to calculate the multi-allelic version of the D′ statistic[34] for the autosomal marker data and females' X-linked marker data (GOLD website). Since haplotypes of X-linked data for males can be established unequivocally, we computed the D′ statistic for males' X-linked data using a function we wrote in R.[38] Haplotype frequencies were estimated ignoring familial relationships — Broman[39] has shown that such estimates, while they may be slightly less precise, are unbiased.[40,41]

When the sample size is small, Lewontin's D′ measure can be biased upwards.[42] We corrected this bias by performing a permutation analysis. We permuted the alleles at the first marker of each pair, calculated and recorded the new D′, repeated these two steps 1,000 times and took the average of the D′ over 1,000 permutations as the permuted D′. We also recorded the corresponding empirical p-value for each marker pair. We then computed an adjusted D′ by taking the difference between the observed D′ and the permuted D′; the adjusted D′, which we denote as D'_c, should be an unbiased measure of LD. Teare *et al.*[42] evaluated this permutation correction and found that it works well under the null hypothesis and is generally an over-correction under alternative hypotheses, which suggests that the levels of LD presented here may be underestimated.

To examine the distribution of LD across the entire autosomal genome, we regressed the D'_c from the autosomal markers against the inter-marker recombination fractions. We also investigated the distribution of LD on the X chromosome by using the D'_c measures obtained from male and female data.

To test for heterogeneity in LD across the autosomal genome, we performed analysis of covariance, in which the predictors in the analysis were the chromosome number and the inter-marker recombination fractions.

We used 5 Y-chromosomal short tandem repeat (STR) markers (DYS388, DYS390, DYS391, DYS394, DYS395) in 20 unrelated Samoan males to estimate the diversity of Y haplotypes, which we computed as $1 - \Sigma P_i$, where P_i is the frequency of the ith observed haplotype.

Population bottleneck

A population bottleneck followed by population expansion creates an imbalance between heterozygosity and allele size variance. Under such conditions, the variance will, for a time, be higher than expected. We measured this by computing the imbalance index (β) of Kimmel *et al.*[43] using, after data cleaning, 371 autosomal markers for the Samoan population and 380 autosomal markers for the NASC population.

Population substructure

Using 25 unlinked autosomal loci, we used a correlated allele frequency model to make inferences about the underlying population substructure. We chose the 25 unlinked loci from our microsatellite marker data; all of these loci were in Hardy–Weinberg equilibrium (it was necessary to use unlinked loci because the statistical tests we employed assume that all loci are unlinked). We picked two loci from chromosomes 1–3 that are at least 200 cM apart, so that the two loci are essentially unlinked. For the other chromosomes, we picked one locus from each chromosome. For these analyses, we used the 'structure' program (*structure* website) to fit clustering models with K = 1, 2, and 3 clusters; the 'structure' program has been shown to produce accurate population assignments with modest numbers of loci.[44]

Results

We used 381 autosomal microsatellite markers and 18 X-linked microsatellite markers to evaluate LD between all pairs of markers on the same chromosome. For each pair of markers, we computed a permutation-adjusted measure of LD, D'_c, which adjusts for biases in Lewontin's D′ measure due to small sample sizes. For the autosomal marker pairs, we found that D'_c declines with increasing recombination fractions between marker pairs. When we averaged the D'_c measures within recombination fraction intervals, the average D'_c in the Samoans decreased as the recombination fraction increased (Table 1a). In general, 15.87 per cent of the D'_c measures are significantly different from zero. By contrast, in the NASC sample, the average D'_c is essentially zero and the percentage of significant values is close to the percentage expected by chance (Table 1a). Note that the per cent significant is a function of sample size and so interpretation of apparent differences can be confounded by sample size differences; however, the Samoan sample size and the NASC sample size are of the same order of magnitude.

We fitted a regression model to examine the relationship between D'_c and recombination fraction, and found that D'_c significantly decreases as the recombination fraction increases (Figure 1); we estimated a slope of -0.021 (SE 0.00354; $p < 0.0001$). If we remove the potentially influential outlier with a D'_c of 0.38, the results remain essentially the same, with a slope of -0.019 (SE 0.00346; $p < 0.0001$).

We evaluated whether there was heterogeneity in the D'_c values across the autosomal chromosomes by using analysis of covariance. The predictors in the analysis were the chromosome number (treated as a set of indicator functions) and inter-marker recombination fractions. We did not find evidence of heterogeneity across chromosomes (F 0.96; df 21; $p = 0.509$), whereas the recombination fraction was significant (F 38.10; df 1; $p < 0.0001$).

For the 18 X-linked microsatellite marker data, we did not observe any steady pattern in the Samoans between the averaged LD measure and recombination fractions, either in the male or female data (Table 1b). The average D'_c values and the

Table 1a. Mean D'_c and per cent significant (at the 0.05 level) as a function of the recombination fraction between all possible pairs of autosomal markers drawn from the same chromosome.

Population	Statistic	Recombination fraction interval				
		<0.1	0.1–<0.2	0.2–<0.3	0.3–<0.4	>0.4
Samoa (N = 482)	Mean D'_c	0.026	0.018	0.016	0.014	0.014
	% significant	24.79	16.43	17.79	12.90	15.06
Palau* (N = 84)	Mean D'_c	0.031	0.019	0.017	0.012	0.009
	% significant	16.2	11.6	11.6	7.1	4.4
NASC (N = 333)	Mean D'_c	−0.001	0.001	0.000	0.000	0.000
	% significant	4.48	3.97	3.87	4.44	4.63

*Palaun data from Devlin *et al.*[9]

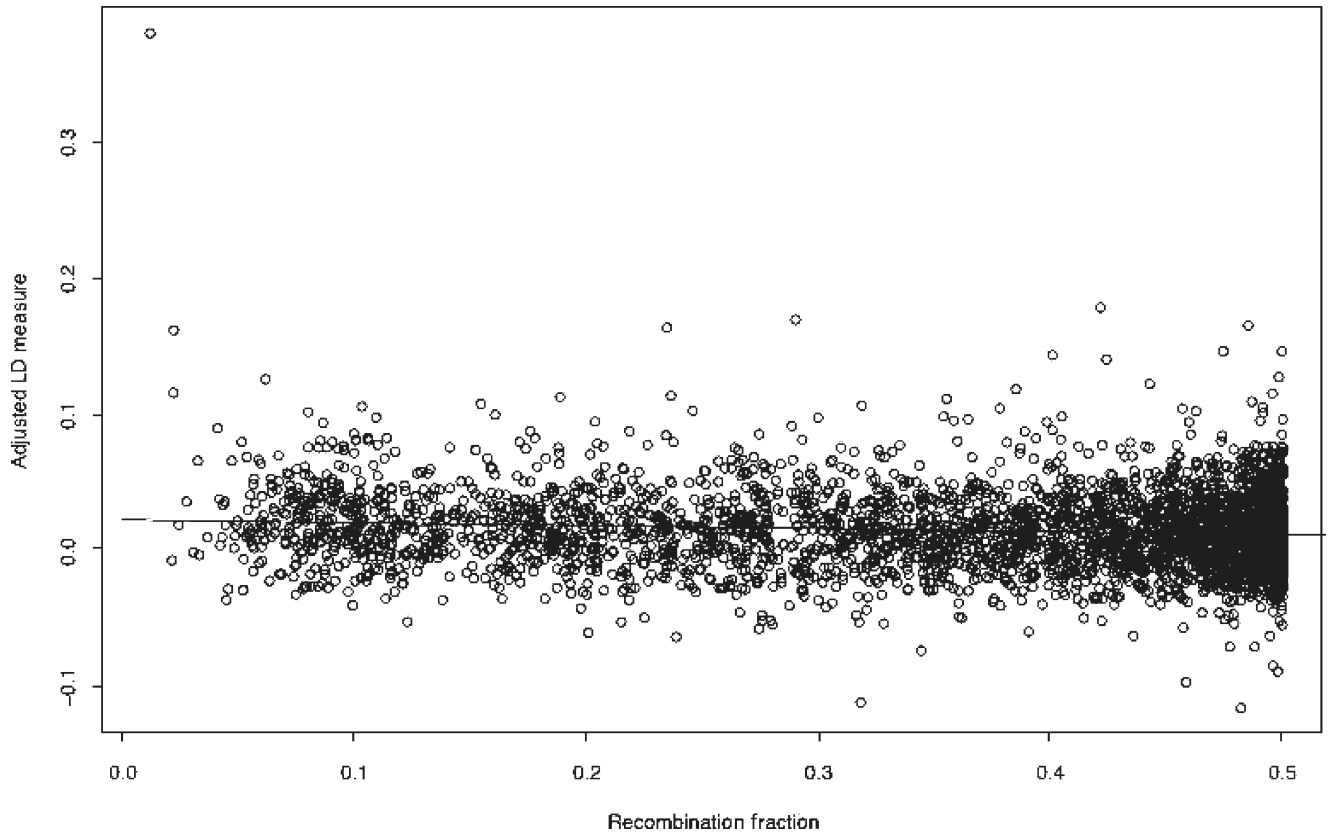

Figure 1. The adjusted linkage disequilibrium measure, D'_c, of all autosomal marker pairs versus recombination fraction.

per cent significant are much higher in the Samoans than in the NASC sample (Table 1b). We also fitted a regression model to examine the relationship between D'_c and recombination fraction in X-linked data. By regressing D'_c on recombination fractions, we obtained negative slopes in males and females, but the *p*-values of the slopes were not significant (data not shown).

We estimated the heterozygosity of the autosomal (X-linked) markers using data from the (female) unrelated individuals and the first (female) sib from each family. We then compared the Samoan heterozygote frequencies with the observed heterozygote frequencies in the CEPH and NASC families. We found the average heterozygosity (0.67) in Samoan data was 0.12 less than in the CEPH families (0.79) and 0.10 less than in the NASC families (0.77). Eighty-seven per cent of markers in the CEPH data and 83 per cent of markers in the NASC data had greater heterozygosity than the corresponding marker in the Samoan data; a sign test showed that both of these differences were highly significant ($p < 0.0001$ for both tests). A previous study in the Palauans found similar results.[9]

The imbalance index (β), an index sensitive to bottleneck events,[43] was estimated for both the Samoan and NASC populations using our autosomal markers. In the presence of a bottleneck event, β is expected to exceed 1. The imbalance index in the Samoan population is 3.86, which is much larger than that in the NASC population ($\beta = 1.31$), indicating that the bottleneck event that occurred in the Samoans is a much more recent event. The β estimated in the NASC population is very similar to an earlier estimation of 1.33 in Europeans.[43] It should be noted that the presence of population substructure would also lead to an elevated β value, which is not the case for the Samoan population, given that it is genetically more homogeneous than the other populations.

Using 25 unlinked autosomal loci, we also tested for population substructure.[44] The number of alleles at these markers ranged from four to 17, with a median of ten, in the Samoan population; and from five to 18, with a median of 11, in the NASC population. Our population substructure analyses strongly support the hypothesis that the Samoans represent one genetic population, even though members of our sample are drawn from two different countries. Based on the 'structure' results, assuming a uniform prior on the number of clusters K, we obtain a posterior probability of 1.0 that $K = 1$. Note also that we have enhanced homogeneity by selecting individuals with four Samoan grandparents. The very low level

Table 1b. Mean D'_c and per cent significant (at the 0.05 level) as a function of the recombination fraction between all possible pairs of X-linked markers.

Population	Statistic	Recombination fraction interval				
		<0.1	0.1–<0.2	0.2–<0.3	0.3–<0.4	>0.4
Samoan females (N = 201)	Mean D'_c	0.129	0.062	0.080	0.059	0.082
	% significant	55.6	33.3	56.3	48.2	48.2
Samoan males (N = 249)	Mean D'_c	0.116	0.050	0.071	0.036	0.067
	% significant	44.4	38.9	43.8	37.0	41.0
Palaun males* (N = 54)	Mean D'_c	0.123	0.041	0.041	0.020	0.026
	% significant	44.0	0	13.6	13.6	21.4
NASC females (N = 161)	Mean D'_c	−0.002	0.000	−0.002	0.006	0.002
	% significant	12.5	7.69	9.09	4.55	9.52
NASC males (N = 172)	Mean D'_c	−0.021	−0.003	−0.014	−0.018	0.000
	% significant	11.11	11.11	0	0	4.82

*Palaun data from Devlin *et al.*[9]

of non-Polynesian alleles seen in our sample[24] indicates that our selection scheme reduced admixture to very low levels.

Discussion

The results we present here are relevant to both the potential use of LD to map disease genes in the Samoan population and to inferences about the evolutionary history of the Samoan population. From our results, we find that LD, as measured using multi-allelic markers, extends over substantial distances across the whole genome in Samoans. When we divide the LD values into groups according to recombination fractions, we observe that the average LD in most intervals is quite similar to that observed in Palau[9] (Table 1a). Compared with the autosomal data, we find a higher level of LD in the X-linked data (as would be expected, based on its smaller effective population size), but when we regress LD on the recombination fractions, the slopes for the X-linked data are not significantly negative. Since we estimated LD for males and females separately, it is possible that the sample sizes were too small. The patterns of LD observed in the Samoans are quite different from those observed in the non-isolated Caucasian NASC sample (Tables 1a and 1b), with the NASC sample displaying essentially no marker-to-marker LD above that which would be expected by chance (except, perhaps, for markers close to each other on the X-chromosome).

We have also typed 43 SNPs from a fragment spanning 104 kb of DNA on chromosome 21 to study the relationship between LD and physical distance between SNP markers in Samoans and four outbred populations, namely, Benin from Nigeria, German, Japanese and Chinese, each representing the four major continental populations.[45] The results show that the Samoans have significantly elevated D′ values compared with all four continental populations throughout the 100 kb region, without any sign of attenuation. It is very likely that this pattern of LD will extend much further than 100 kb in the Samoans. By contrast, in the other four populations, D′ shows a declining slope when plotted against increasing physical distance.

Based on our Y-linked STR data, the Y haplotype diversity in Samoans is 0.855, which is lower than the observed diversity of 0.96 in the Palauan study; however, it is comparable to other isolated populations. From the five Y-linked STR data in two isolated Iberian samples (Basque and Catalan), the diversities of Y haplotypes are 0.85 and 0.89 in Basque and Catalan, respectively.[46]

Our statistical tests on our genetic data support the scenario of a recent bottleneck, which is consistent with the known demographic history of the Samoan population. Furthermore, our population substructure results support the hypothesis that self-identified Samoans represent one homogenous genetic population. These results are consistent with our prior reports on the genetic structure of the Samoans.[24]

Our finding of high levels of genome-wide LD among Samoans are consistent with the multiple lines of evidence from genetic, archaeological and demographic history research that indicate that the Samoan population was founded by a relatively small number of founders, remained isolated for about 3,000 years and experienced a reduction in population size about 200–300 years ago.[15,16,23–27] The findings of

genome-wide LD in the Samoan population are similar to levels of LD observed in another isolated Pacific population, namely, Palau of Micronesia,[9] as well as an isolated population from the Central Valley of Costa Rica.[10] The present Samoan findings and these others[9] support the hypothesis that LD extends further in isolated populations than in continental populations.[4–6]

The results presented here on patterns of LD using a relatively sparse set of markers indicate that a more thorough characterisation of LD patterns in the Samoans is likely to be of interest. In particular, it would be of interest to compare the Samoan population with other isolated populations that have grown rapidly. While we have begun to generate higher density data in limited regions, as described above for chromosome 21,[45] the results presented here provide important baseline information about the levels of background LD, but provide little information about the extent of 'useful' LD, as required for genome-wide association scans.[47] While the precise definition of what levels of LD are useful for association mapping remains debatable, in outbred northern European populations LD is thought to be useful only over a relatively short range of 10–30 kb.[47] The dramatic difference between our results for the Samoan population and those from the NASC sample favour the hypothesis that the level of useful LD will be higher in the Samoan population.

Acknowledgments

This research was supported by NIH grants AG09375, HL52611, DK55406 and DK59642. We thank the members of the Department of Health, Government of Samoa and the American Samoan Department of Health for their assistance in data collection; local political officials for their cooperation; and the participants for their patience. We would like to thank Dr John Reveille, the Principal Investigator of the North American Spondylitis Consortium (NASC) project for his generous support in sharing the genotype data. The genotyping of the NASC samples was supported by NIH grant RO1-AR46208. We thank Dr Ning Wang of the Center for Genome Information, University of Cincinnati, for her help in computing the imbalance indices. We thank the two anonymous reviewers whose comments helped us to markedly improve this manuscript.

Electronic database information

URLs for data presented herein are as follows:

GOLD, Abecasis and Cookson, http://www.sph.umich.edu/csg/abecasis/GOLD/index.html, University of Michigan (accessed 20th April, 2004).

structure, Pritchard, J.K., Stephens, M. and Donnelly, P., http://pritch.bsd.uchicago.edu/software.html, University of Chicago, [cited 2004 April 20].

References

1. Risch, N. and Merikangas, K. (1996), 'The future of genetic studies of complex human diseases', *Science* Vol. 273, pp. 1516–1517.
2. Kruglyak, L. (1999), 'Genetic isolates: Separate but equal?', *Proc. Natl. Acad. Sci. USA* Vol. 96, pp. 1170–1172.
3. Peltonen, L., Palotie, A. and Lange, K. (2000), 'Use of population isolates for mapping complex traits', *Nat. Rev. Genet.* Vol. 1, pp. 182–190.
4. Taillon-Miller, P., Bauer-Sardina, I., Saccone, N.L. *et al.* (2000), 'Juxtaposed regions of extensive and minimal linkage disequilibrium in human Xq25 and Xq28', *Nat. Genet.* Vol. 25, pp. 324–328.
5. Eaves, I.A., Merriman, T.R., Barber, R.A. *et al.* (2000), 'The genetically isolated populations of Finland and Sardinia may not be a panacea for linkage disequilibrium mapping of common disease genes', *Nat. Genet.* Vol. 25, pp. 320–323.
6. Lonjou, C., Collins, A. and Morton, N.E. (1999), 'Allelic association between marker loci', *Proc. Natl. Acad. Sci. USA* Vol. 96, pp. 1621–1626.
7. Boehnke, M. (2000), 'A look at linkage disequilibrium', *Nat. Genet.* Vol. 25, pp. 246–247.
8. Kruglyak, L. (1999), 'Prospects for whole-genome linkage disequilibrium mapping of common disease genes', *Nat. Genet.* Vol. 22, pp. 139–144.
9. Devlin, B., Roeder, K., Otto, C. *et al.* (2001), 'Genome-wide distribution of linkage disequilibrium in the population of Palau and its implications for gene flow in Remote Oceania', *Hum. Genet.* Vol. 108, pp. 521–528.
10. Service, S.K., Ophoff, R.A. and Freimer, N.B. (2001), 'The genome-wide distribution of background linkage disequilibrium in a population isolate', *Hum. Mol. Genet.* Vol. 10, pp. 545–551.
11. Shifman, S. and Darvasi, A. (2001), 'The value of isolated populations', *Nat. Genet.* Vol. 28, pp. 309–310.
12. Mohlke, K.L., Lange, E.M., Valle, T.T. *et al.* (2001), 'Linkage disequilibrium between microsatellite markers extends beyond 1 cM on chromosome 20 in Finns', *Genome Res.* Vol. 11, pp. 1221–1226.
13. Hall, D., Wijsman, E.M., Roos, J.L. *et al.* (2002), 'Extended inter-marker linkage disequilibrium in the Afrikaners', *Genome Res.* Vol. 12, pp. 956–961.
14. Bellwood, P.S. (1979), *Man's Conquest of the Pacific: The Prehistory of Southeast Asia and Oceania*, Oxford University Press, New York, NY.
15. Green, R.C., Davidson, J.M. and Bernice Pauahi Bishop Museum (1969), *Archaeology in Western Samoa*, Auckland Institute and Museum, Auckland, NZ.
16. Kirch, P.V. (2000), *On the Road of the Winds: An Archaeological History of the Pacific Islands before European Contact*, University of California Press, Berkeley, CA.
17. Diamond, J.M. (1988), 'Express train to Polynesia', *Nature* Vol. 336, pp. 307–308.
18. Sykes, B., Leiboff, A., Low-Beer, J. *et al.* (1995), 'The origins of the Polynesians: An interpretation from mitochondrial lineage analysis', *Am. J. Hum. Genet.* Vol. 57, pp. 1463–1475.
19. Redd, A.J., Takezaki, N., Sherry, S.T. *et al.* (1995), 'Evolutionary history of the COII/tRNA^{Lys} intergenic 9 base pair deletion in human mitochondrial DNAs from the Pacific', *Mol. Biol. Evol.* Vol. 12, pp. 604–615.
20. Melton, T., Clifford, S., Martinson, J. *et al.* (1998), 'Genetic evidence for the proto-Austronesian homeland in Asia: mtDNA and nuclear DNA variation in Taiwanese aboriginal tribes', *Am. J. Hum. Genet.* Vol. 63, pp. 1807–1823.
21. Terrell, J. (1988), 'History as a family tree, history as an entangled bank: constructing images and interpretations of prehistory in the South Pacific', *Antiquity* Vol. 62, pp. 642–657.
22. Martinson, J.J. (1996), 'Molecular perspectives on the colonization of the Pacific', In: Boyce, A.J. and Mascie-Taylor, C.G.N. (eds.), *Molecular Biology and Human Diversity*, Cambridge University Press, Cambridge, UK, pp. 171–195.
23. Su, B., Jin, L., Underhill, P. *et al.* (2000), 'Polynesian origins: Insights from the Y chromosome', *Proc. Natl. Acad. Sci. USA* Vol. 97, pp. 8225–8228.
24. Deka, R., McGarvey, S.T., Ferrell, R.E. *et al.* (1994), 'Genetic characterization of American and Western Samoans', *Hum. Biol.* Vol. 66, pp. 805–822.
25. Deka, R., Jin, L., Shriver, M.D. *et al.* (1995), 'Population genetics of dinucleotide (dC-dA)n.(dG-dT)n polymorphisms in world populations', *Am. J. Hum. Genet.* Vol. 56, pp. 461–474.

26. Deka, R., Shriver, M.D., Yu, L.M. *et al.* (1999), 'Genetic variation at twentythree microsatellite loci in sixteen human populations', *J. Genet.* Vol. 78, pp. 99–121.
27. Shriver, M.D., Jin, L., Ferrell, R.E. *et al.* (1997), 'Microsatellite data support an early population expansion in Africa', *Genome Res.* Vol. 7, pp. 586–591.
28. McArthur, N. (1967), *Island Populations of the Pacific*, Australian National University Press, Canberra, Australia.
29. McArthur, N.A. (1956), *The Populations of the Pacific.* Part III American Samoa and Part IV Western Samoa and the Tokelau Islands, Australian National University, Department of Demography, Canberra, Australia.
30. Gilson, R.P. (1970), *Samoa 1830 to 1900: The Politics of a Multi-cultural Community*, Oxford University Press, New York, NY.
31. Tsai, H.-J., Sun, G., Weeks, D.E. *et al.* (2001), 'Type 2 diabetes and three calpain-10 gene polymorphisms in Samoans: No evidence of association', *Am. J. Hum. Genet.* Vol. 69, pp. 1236–1244.
32. McGarvey, S.T., Levinson, P.D., Bausserman, L. *et al.* (1993), 'Population-change in adult obesity and blood-lipids in American-Samoa from 1976–1978 to 1990', *Am. J. Hum. Biol.* Vol. 5, pp. 17–30.
33. Galanis, D.J., McGarvey, S.T., Quested, C. *et al.* (1999), 'Dietary intake of modernizing Samoans: Implications for risk of cardiovascular disease', *J. Am. Diet Assoc.* Vol. 99, pp. 184–190.
34. Lewontin, R.C. (1964), 'The interaction of selection and linkage. I. General considerations; heterotic models', *Genetics* Vol. 49, pp. 49–67.
35. Abecasis, G.R. and Cookson, W.O. (2000), 'GOLD — Graphical overview of linkage disequilibrium', *Bioinformatics* Vol. 16, pp. 182–183.
36. Excoffier, L. and Slatkin, M. (1995), 'Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population', *Mol. Biol. Evol.* Vol. 12, pp. 921–927.
37. Long, J.C., Williams, R.C. and Urbanek, M. (1995), 'An E-M algorithm and testing strategy for multiple-locus haplotypes', *Am. J. Hum. Genet.* Vol. 56, pp. 799–810.
38. Ihaka, R. and Gentleman, R. (1996), 'R: A language for data analysis and graphics', *J. Comput. Graph. Stat.* Vol. 5, pp. 299–314.
39. Broman, K.W. (2001), 'Estimation of allele frequencies with data on sib-ships', *Genet. Epidemiol.* Vol. 20, pp. 307–315.
40. Chakraborty, R. (1978), 'Number of independent genes examined in family surveys and its effect on gene frequency estimation', *Am. J. Hum. Genet.* Vol. 30, pp. 550–552.
41. Chakraborty, R. (1991), 'Inclusion of data on relatives for estimation of allele frequencies', *Am. J. Hum. Genet.* Vol. 49, pp. 242–244.
42. Teare, M.D., Dunning, A.M., Durocher, F. *et al.* (2002), 'Sampling distribution of summary linkage disequilibrium measures', *Ann. Hum. Genet.* Vol. 66, pp. 223–233.
43. Kimmel, M., Chakraborty, R., King, J.P. *et al.* (1998), 'Signatures of population expansion in microsatellite repeat data', *Genetics* Vol. 148, pp. 1921–1930.
44. Pritchard, J.K., Stephens, M. and Donnelly, P. (2000), 'Inference of population structure using multilocus genotype data', *Genetics* Vol. 155, pp. 945–959.
45. Deka, R., McGarvey, S.T., Weeks, D.E. *et al.* (2003), 'Genetic variation in an isolated population, the Samoans of Polynesia: Implications for mapping complex traits', *Am. J. Hum. Genet.* Vol. 73(Suppl.), pp. 187.
46. Perez-Lezaun, A., Calafell, F., Seielstad, M. *et al.* (1997), 'Population genetics of Y-chromosome short tandem repeats in humans', *J. Mol. Evol.* Vol. 45, pp. 265–270.
47. Ardlie, K.G., Kruglyak, L. and Seielstad, M. (2002), 'Patterns of linkage disequilibrium in the human genome', *Nat. Rev. Genet.* Vol. 3, pp. 299–309.