Hum Genet (2001) 109 : 253–261
DOI 10.1007/s004390100576

ORIGINAL INVESTIGATION

**Rick A. Kittles · Dale Young · Sally Weinrich
Julie Hudson · George Argyropoulos · Flora Ukoli
Lucile Adams-Campbell · Georgia M. Dunston**

Extent of linkage disequilibrium between the androgen receptor gene CAG and GGC repeats in human populations: implications for prostate cancer risk

Received: 26 December 2000 / Accepted: 2 July 2001 / Published online: 4 August 2001
© Springer-Verlag 2001

Abstract While studies have implicated alleles at the CAG and GGC trinucleotide repeats of the androgen receptor gene with high-grade, aggressive prostate cancer disease, little is known about the normal range of variation for these two loci, which are separated by about 1.1 kb. More importantly, few data exist on the extent of linkage disequilibrium (LD) between the two loci in different human populations. Here we present data on CAG and GGC allelic variation and LD in six diverse populations. Alleles at the CAG and GGC repeat loci of the androgen receptor were typed in over 1000 chromosomes from Africa, Asia, and North America. Levels of linkage disequilibrium between the two loci were compared between populations. Haplotype variation and diversity were estimated for each population. Our results reveal that populations of African descent possess significantly shorter alleles for the two loci than non-African populations (P<0.0001). Allelic diversity for both markers was higher among African Americans than any other population, including indigenous Africans from Sierra Leone and Nigeria. Analysis of molecular variance revealed that approx. 20% of CAG and GGC repeat variance could be attributed to differences between the populations. All non-African populations possessed the same common haplotype while the three populations of African descent possessed three divergent common haplotypes. Significant LD was observed in our sample of healthy African Americans. The LD observed in the African American population may be due to several reasons; recent migration of African Americans from diverse rural communities following urbanization, recurrent gene flow from diverse West African populations, and admixture with European Americans. This study represents the largest genotyping effort to be performed on the two androgen receptor trinucleotide repeat loci in diverse human populations.

R.A. Kittles (✉) · D. Young · G.M. Dunston
National Human Genome Center, Howard University,
2041 Georgia Ave., Washington, DC 20060
e-mail: rkittles@howard.edu,
Tel.: +1-202-8067028, Fax: +1-202-9863972

R.A. Kittles · F. Ukoli · L. Adams-Campbell · G.M. Dunston
Cancer Center, Howard University, Washington, DC 20059, USA

R.A. Kittles · G.M. Dunston
Department of Microbiology, College of Medicine,
Howard University, Washington, DC 20059, USA

S. Weinrich · J. Hudson
Population Studies, South Carolina Cancer Center,
University of South Carolina, Columbia, SC 29203, USA

G. Argyropoulos
Department of Medicine/Endocrinology,
Medical University of South Carolina,
Charleston, SC 29403, USA

Introduction

The human androgen receptor (AR) is a ligand-dependent nuclear transcriptional factor that regulates the expression of genes necessary for the growth and development of both normal and malignant prostate tissue. The AR gene is about 90 kb and is located on chromosome Xq11–12. Exon 1 of the gene encodes the N-terminal domain, which controls transcriptional activation of the receptor. Exon 1 also encodes two polymorphic trinucleotide repeats (CAG and GGC), which code for polyglutamine and polyglycine tracts, respectively in the N-terminal domain. In vitro studies have demonstrated an inverse relationship between CAG repeat length and AR transcriptional activation ability (Chamberlain et al. 1999).

Variations in AR CAG repeat length have been associated with a number of genetic diseases. Spinal ataxia 1 (SCA1), Kennedy's disease, and Huntington's disease are examples of AR loss of function disorders that result from expansion in AR CAG repeat length (LaSpada et al. 1991; Orr et al. 1993). In addition, short CAG and GGC repeat lengths have been widely attributed to increased risk of developing prostate cancer (Giovannucci et al. 1997; Hardy et al. 1996; Platz et al. 1998). More specifically, individuals with CAG repeat lengths less than 20 and GGC repeats less than 16 have been associated with increased

risk of developing prostate cancer (Giovannucci et al. 1997; Platz et al. 1998; Stanford et al. 1997). Striking differences in CAG repeat lengths have been observed between populations. Black men tend to have significantly shorter repeats than their white counterparts (Edwards et al. 1992; Irvine et al. 1995; Sartor et al. 1999). These genetic differences may be potentially important in understanding why populations of African descent are more susceptible to developing prostate cancer. African American men have the highest incident rate of prostate cancer of any ethnic group in the United States (Brawley and Kramer 1996). Increasing evidence suggests that prostate cancer is more prevalent in populations of African descent (Glover et al. 1998; Ogunbiyi and Shittu 1999; Osegbe 1997). However, attempts to explain the disparity in risk between populations are limited. Although diet (i.e., fat intake) may help to explain the high prevalence of prostate cancer among African Americans, such an influence may be limited when considering other populations of African descent (i.e., Caribbeans and West Africans) whose diet differ considerably.

Genetic studies on the AR CAG and GGC loci have focused mainly on European American prostate cancer patients and controls. Little is known about AR haplotypic variation, especially among different human populations. Evaluating variation in AR trinucleotide repeat lengths across human populations may provide a better understanding of the ethnic disparity associated with prostate cancer. Studies have not formally evaluated variation and the extent of linkage disequilibrium between the two trinucleotide repeat loci across human populations and in particular among those that may have contributed to the African American gene pool. This would be an important prerequisite to determining if there are subpopulations of disease chromosomes segregating in high-risk groups such as African Americans.

When the occurrence of pairs of specific alleles at different loci on the same haplotype is not independent, the deviation from the independence is termed linkage disequilibrium (LD). LD is a population genetic phenomenon that has been useful for gene mapping efforts. It is usually found in populations for genetic markers that are tightly (close genetic distance) linked and can be generated by mutation, selection, or admixture of populations with different allele frequencies. Generally disequilibrium is dependent on population size, time (generations), and distance between genetic markers. Normally, the greater the distance between markers, the faster the decay of disequilibrium. However, for highly polymorphic markers such as microsatellites, the high mutation rate contributes significantly to randomizing associations of alleles.

The aim of this study was to formally evaluate variation and the extent of linkage disequilibrium between the AR gene CAG and GGC repeat loci in human populations, particularly those of African descent such as African Americans.

The African American population is genetically and culturally heterogeneous due to their unique history in the United States (Jackson 1993). While a significant portion of the African American gene pool originates from Western and Central Africa, other populations have also contributed to the present genetic makeup of the population. To better understand variation within the African American population we included comparative populations representing West Africans (Nigeria and Sierra Leone), European Americans, Chinese, and Amerindians.

Subjects and methods

Unrelated African American men (n=520) were recruited from Columbia, South Carolina, for prostate-specific antigen (PSA) screening over the past 5 years. Nigerians (n=85) representing the Edo (Bini) ethnic group were recruited in the Udo community near Benin City, Nigeria. European American men (n=90) were recruited from the Washington, DC area. The African American, Nigerian, and European American men were recruited as healthy community-based controls for prostate cancer studies. Inclusion criteria were men between 50 and 80 years of age with PSA levels less than 4.0 ng/ml and normal digital rectal examinations. In addition, unrelated men representing the Mende ethnic group from Sierra Leone (n=240), Han Chinese from Taiwan (n=60), and an Amerindian population from a community in the southwestern United States (n=103) were also included. No clinical data or medical history was collected for the Sierra Leone, Chinese, and Amerindian participants. Informed consent for genetic analysis was obtained for all subjects. Individuals of mixed ancestry were not excluded. Genomic DNA was isolated from whole-blood samples using the Puregene (Gentra Biosystems) DNA isolation kit. The trinucleotide repeat CAG and GGC loci were amplified by PCR using 50 ng genomic DNA. Primers used to amplify the CAG locus were 5′-TCC AGA ATC TGT TCC AGA GCG TG-3′ (forward) and 5′-GCT GTG AAG GTT GCT GTT CCT CAT-3′ (reverse). Primers specific for the GGC locus were 5′-CCA GAG TCG CTC GCG ACT ACT ACA ACT TTC C-3′ (forward) and 5′-GGA CTG GGA TAG GGC ACT CTG CTC ACC-3′ (reverse). Florescent dyes 6-FAM and HEX were used to label the forward primers for GGC and CAG respectively. PCR cycling conditions for the CAG locus were 35 cycles of 95°C for 30 s, 60°C for 30 s, and 72°C for 30 s. Conditions for the GGC locus was 25 cycles of 97°C for 30 s, 55°C for 30 s and 72°C for 1 min.

PCR products for both loci were then pooled and electrophoresed on an ABI 377 DNA sequencer, (ABI, Foster City, Calif., USA). Genescan and Genotyper 5.0 programs (ABI) were used to generate fragment sizes and genotypes. Due to limited genomic DNA some samples could not be typed for both loci. Statistical analyses for comparison of repeat length mean, mode, and variance among populations were performed using Origin 5.0 (Microcal Software, Northampton, Mass., USA). Heterozygosities for the two trinucleotide repeats and for the haplotypes were computed as $n(1-\Sigma p_i^2)/(n-1)$, where p_i represents the frequency of the ith allele or haplotype, and where n is the number of chromosomes drawn from the population. Standard errors were obtained by using equation 8.7 in Nei (1987). Standardized pairwise linkage disequilibrium values (D′; Lewontin 1964) were calculated for all pairs of microsatellite alleles observed within each population. The null hypothesis of linkage equilibrium (D′=0) was tested and P values obtained by Fisher's exact test using the Markov chain (Guo and Thompson 1992) implemented by the computer program Arlequin 1.1 (Schneider et al. 1997).

Differences among populations were assessed by use of the hierarchical analysis of molecular haplotype variance (AMOVA; Excoffier et al. 1992; Michalakis and Excoffier 1996) implemented by the Arlequin 1.1 package. AMOVA performs a hierarchic analysis of three genetic-variance components: ΦST, subpopulations relative to the total population; ΦSC, subpopulations relative to continental groups; and ΦCT, continental groups relative to the total population. For the analysis, three groups containing the six populations were defined: (a) populations of African descent

(Nigerians, Sierra Leoneans, and African Americans); (b) European American population; and (c) Asian descent populations (Chinese and Amerindian). The AMOVA assumed a single stepwise mutation model (DiRienzo et al. 1994; Valdes et al. 1993) for the trinucleotide repeat loci. In addition, pairwise genetic distances between populations were computed from ΦST values; D = ΦST/(1–ΦST) (Slatkin 1995). Significance levels of the genetic variance components were estimated by use of 10,000 random-permutation procedures.

Results

Allelic diversity

Tables 1 and 2 show the allelic diversity observed in the six populations for the CAG and GGC markers. African Americans possessed the greatest number of alleles for both markers, which partially may be due to the larger sample size. However, for the CAG locus 18 alleles were observed among the significantly smaller sample of Nigerians. The African American population possessed the highest gene diversity for the CAG marker of any of the other populations, while greater diversity was observed for the GGC locus in the Sierra Leone population. The number of CAG alleles observed ranged from 11 for Euroamericans to 21 for African Americans (Table 1). For the GGC locus the number of alleles ranged from 4 for both Asians and Amerindians to 17 for African Americans (Table 2). The GGC allele with 15 repeats was highly frequent in non-African populations, ranging from 55% to 80%. The 15-repeat allele was less frequent among West Africans (5–10%) and intermediate in frequency among African Americans at 23%. Strikingly low diversity was observed at the GGC locus for Chinese and Amerindians. Gene diversity for the two populations of Asian ancestry was almost one-third of that observed for the African populations (Table 2).

CAG and GGC allelic distributions are shown in Fig. 1. These distributions portray a shift in the most common allele among African versus non-African populations. Allelic distributions were either unimodal or bimodal for all populations except the Nigerians and Amerindians (Fig. 1). The multimodal CAG allele distribution for the Nigerian and Amerindian populations may be due to genetic drift. For example, among Nigerians the 17 allele at the CAG locus is rare (<0.01), unlike in the other African populations. Among the Amerindians both the 20 and 22 CAG alleles are common while the 19 and 21 alleles are less frequent (Fig. 1).

As an alternative measure of intrapopulation diversity for the microsatellite markers we calculated variances in allele sizes for each locus (Tables 1, 2). Again, genetic drift operating within the Nigerian and Amerindian populations may have contributed to the higher variance in number of CAG repeats than in the other populations. Interestingly, the same trend was not observed in the Nigerian or Amerindian populations for the GGC locus (Table 2). The lowest variance for CAG allele size was observed among the Europeans. Europeans possessed about one-fourth the variance in CAG allele size than among Nigerians.

Variance in allele size for the GGC marker was one-third that of CAG allele size variance. The populations with the lowest GGC allele size variance were Asians and Amerindians. A notable trend observed among the variances calculated was that variances for the African populations were almost twice that of the non-African populations. This is consistent with the findings of other studies that have examined microsatellite diversity. These studies reveal higher gene diversity among African populations and significant genetic differences between African and non-African populations (Jorde et al. 1995, 2000; Nei and Takezaki 1996; Reich and Goldstein 1998; Shriver et al. 1997).

Table 1 CAG allelic diversity (*N* number of chromosomes, *H* gene diversity)

Population	N	H	No. of alleles	Mean	Range	Variance
African American	516	0.951	21	17.8	9–31	10.97
Sierra Leone	230	0.918	17	17.3	10–26	7.77
Nigerian	83	0.909	18	16.7	5–28	17.28
Euroamerican	87	0.866	11	19.7	13–26	5.37
Asian	60	0.846	12	20.1	14–26	4.55
Amerindian	80	0.884	14	20.1	14–30	8.62

Table 2 GGC allelic diversity (*N* number of chromosomes, *H* gene diversity)

Population	N	H	No. of alleles	Mean	Range	Variance
African American	472	0.880	17	14.3	4–20	4.94
Sierra Leone	210	0.906	14	13.7	4–24	5.50
Nigerian	78	0.771	10	13.8	8–19	3.44
Euroamerican	80	0.628	13	15.0	2–20	5.77
Asian	60	0.322	4	14.6	10–16	1.48
Amerindian	103	0.362	4	14.6	8–16	1.91

Haplotype diversity and LD

AR CAG and GGC haplotype frequencies and D′ values were determined for each population and are available from the Human Genome Diversity Laboratory. Haplotype diversity was greatest for African populations, lower for European Americans, and lowest for Amerindians (Table 3). The 15 most common AR haplotypes, their frequencies, and LD values are shown in Table 4. Highly divergent haplotypes were observed among the African populations. All non-African populations possessed the same most common haplotype designated as 20-15 (20 CAG repeats and 15 GGC repeats). Table 4 reveals that the frequency of the 20-15 haplotype in non-African populations ranged from 13% among Euroamericans to 20% among the Asians. Populations of African descent each possessed a different common haplotype. The most common haplotype among African Americans was 16-16 at 5% frequency. Among Nigerians three common haplotypes were

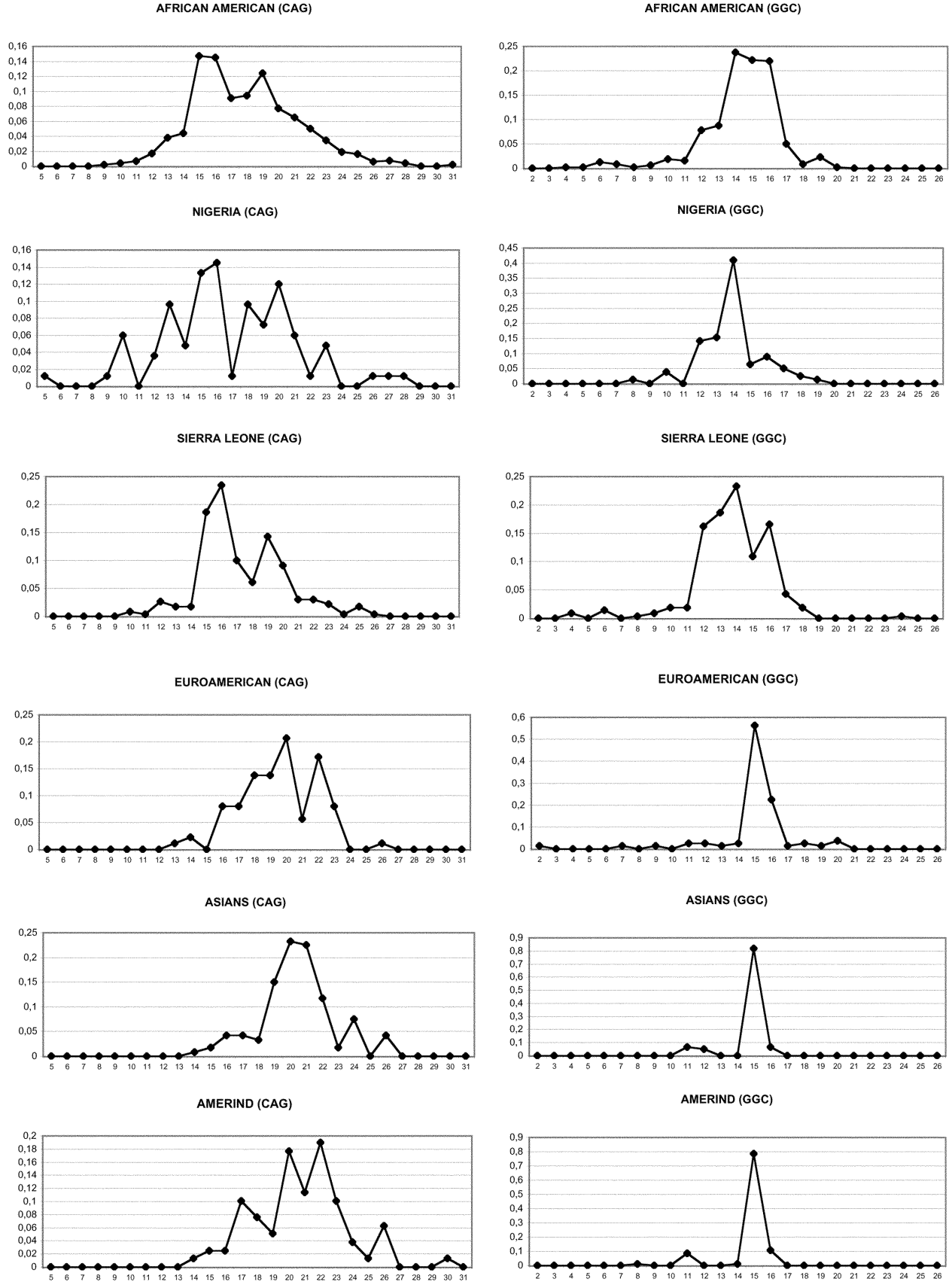

Fig. 1 Allele frequency distributions of CAG (*left*) and GGC (*right*) microsatellites from the six human populations included in this study. *X-axis* Number of repeats; *y-axis* frequency

Table 3 Androgen receptor haplotype diversity (*h*)

Population	No. of haplotypes	h
African American	132	0.991±0.015
Sierra Leone	92	0.990±0.002
Nigerian	50	0.992±0.003
Euroamerican	40	0.980±0.007
Asian	23	0.946±0.018
Amerindian	29	0.938±0.020

observed, 13-14, 19-14, and 20-14, each at a frequency of 6.5%. In Sierra Leone the most common haplotype was 16-12 at 7.4%. The non-African populations did not possess population specific haplotypes. In fact the non-African populations possessed a subset of the variation observed among the African populations. This is similar to observations using other genetic markers (Tishkoff et al. 1996, 1998).

Of all possible pairwise LD comparisons of polymorphic alleles at the two loci, 10% (35 of 357) were significant ($P<0.05$) for African Americans. The percentage of significant D′ values for comparisons of alleles among the other populations ranged from 4% for the Chinese to 8% for Amerindians and European Americans (data not shown). These differences in levels of LD could be due to recent admixture, drift, substructure, or power to detect allelic associations.

Although each non-African population shared the same common haplotype (20-15), no LD was detected between the alleles in the three populations (Table 4). This is likely due to the random effects of drift operating differently in populations after the expansion out of Africa and subsequent mutations away from the common haplotype. Drift and mutation would affect each population differently. This could explain why AR haplotype diversity is greater for the European population than populations of Asian ancestry (Table 3). Our inability to detect LD may also be due to the smaller sample sizes of the non-African populations than the African populations. Sample size likely played a role for the Chinese samples. For instance, D' for many of the common haplotypes in China was 100% but not significant (Table 4).

Table 4 also reveals significant sharing of haplotypes in the two West African populations indicative of shared ancestry. For instance, haplotype 19-14 is common in both Nigeria and Sierra Leone and in strong LD in both populations. This is also reflected in the African American population, where about 40% of the significantly associated alleles were low in frequency (<0.05 allele frequency) and appeared to be of African origin. Table 4 reveals five of these haplotypes (13-14, 15-12, 15-15, 16-12, and 16-16). This is contrary to what would be expected if the LD within the African American population were due to admixture with Euroamericans.

Table 4 Levels of linkage disequilibrium for the 15 most common androgen receptor haplotypes (most common haplotype for each population in ***boldface***, *f* haplotype frequency)

Haplotype	African Americans			Nigerians			Sierra Leoneans			European Americans			Asians			Amerindidans		
	f	D' (%)	*P*	*f*	D' (%)	*P*	*f*	D' (%)	*P*	*f*	D' (%)	*P*	*f*	D' (%)	*P*	*f*	D' (%)	*P*
13.14	0.020	37.0	<0.001	**0.065**	**36.0**	**NS**	–	–	–	–	–	–	–	–	–	–	–	–
15.12	0.002	–80.0	0.01	0.010	13.0	NS	0.042	6.0	NS	–	–	–	–	–	–	–	–	–
15.15	0.040	12.0	0.01	0.010	11.0	NS	0.030	7.0	NS	–	–	–	0.020	100.0	NS	–	–	–
15.16	0.040	8.0	NS	0.052	52.0	<0.001	0.037	2.0	NS	–	–		–	–	–	–	–	–
16.12	0.024	18.0	<0.001	0.010	23.0	NS	**0.074**	**27.0**	**<0.001**	–	–	–	–	–	–	–	–	–
16.16	**0.050**	**20.0**	**<0.001**	0.010	2.0	NS	0.047	5.0	NS	0.030	22.0	NS	–	–	–	–	–	–
17.15	0.010	–30.0	NS	–	–	–	0.005	–53.0	NS	0.050	56.0	NS	0.050	100.0	NS	0.090	48.0	NS
18.15	0.010	–43.0	NS	–	–	–	0.005	–6.0	NS	0.060	–17.0	NS	0.040	100.0	NS	0.030	–34.0	NS
19.14	0.034	3.0	NS	**0.065**	**100.0**	**<0.001**	0.069	30.0	<0.001	0.020	42.0	NS	–	–	–	–	–	–
19.15	0.020	–16.0	NS	–	–	–	0.030	12.0	NS	0.060	–17.0	NS	0.160	100.0	NS	0.040	–1.0	NS
20.14	0.030	15.0	NS	**0.065**	**15.0**	**NS**	0.027	4.0	NS	–	–	–	–	–	–	–	–	–
20.15	0.007	–47.0	NS	–	–	–	0.020	7.0	NS	**0.130**	**9.0**	**NS**	**0.200**	**–2.0**	**NS**	**0.160**	**68.0**	**NS**
21.15	0.020	7.0	NS	0.010	14.0	NS	–	–	–	0.040	45.0	NS	0.180	–33.0	NS	0.110	53.0	NS
22.15	0.010	16.0	NS	–	–	–	0.005	6.0	NS	0.100	–2.0	NS	0.110	100.0	NS	0.150	–4.0	NS
23.16	0.002	–69.0	NS	–	–	–	–	–	–	–	–	–	–	–		0.040	28.0	NS
Others	0.681			0.703			0.609			0.510			0.240			0.380		

Table 5 Genetic differentiation of populations

Type of comparison	Variance (%)	Φ Statistic	*P*
Among groups	18.5	ΦCT = 0.185	0.01
Among populations within groups	1.2	ΦSC = 0.014	<0.001
Within populations	80.3	ΦST = 0.197	<0.001

Table 6 Pairwise genetic distances based on ΦST (below the diagonal) and their significance levels (above the diagonal)

	African Americans	European Americans	Amerindians	Nigerians	Chinese	Sierra Leoneans
African Americans	–	0.000	0.000	0.019	0.000	0.029
European Americans	0.103	–	0.000	0.000	0.416	0.000
Amerindians	0.224	0.049	–	0.000	0.089	0.000
Nigeria	0.027	0.213	0.338	–	0.000	0.297
China	0.137	–0.002	0.018	0.262	–	0.000
Sierra Leone	0.011	0.212	0.351	–0.001	0.265	–

Analysis of molecular variance

Genetic variance (Φ) statistics for the AR trinucleotide repeat data are shown in Table 5. Using both molecular AR haplotypic differences based on microsatellite repeat-length and haplotype frequencies, AMOVA revealed that the AR trinucleotide repeat diversity is nonrandomly distributed across populations. The amount of genetic variance between the six populations was 19.7% ($P<0.001$). The bulk of genetic variance for the AR gene (80.3%) could be explained by individual differences within populations. The **ΦCT** estimate was 0.185, revealing that 18.5% of the genetic variance was due to differences between the African, Asian, and European descent groups (Table 5).

Pairwise genetic distances between the populations are provided in Table 6. The lowest pairwise distances (≤0.05) between populations were observed among the closely related African descendant populations (African Americans, Nigerians, and Sierra Leone) and between the European and Asian populations (Chinese and Amerindian). High genetic distance values (>0.30) were observed between divergent populations such as Amerindians and West Africans from Nigeria or Sierra Leone (Table 6). A moderate distance value of 0.10 between African Americans and European Americans suggests a shared biohistory. All but three of the population pairwise distances were significant ($P<0.05$). The nonsignificant values reflect the close genetic affinities of the three pairs of populations (Table 6).

Discussion

In order to evaluate the extent of variation and linkage disequilibrium between two trinucleotide repeat loci within the AR gene we typed alleles from both markers in six diverse human populations. Populations of African descent exhibited the highest gene diversity among the populations sampled. The African American population contained more alleles and higher gene diversity than even the indigenous West African populations from Sierra Leone and Nigeria. Asians possessed the lowest gene diversity of all populations and also contained the lowest frequency of "high-risk" short alleles for prostate cancer.

Patterns of allelic variation differed substantially between the six populations. Our data revealed that 80% of men of African decent possessed CAG alleles shorter than 20 repeats while only 50% of non-African men had these short alleles (see Fig. 1). The pattern was more pronounced for the GGC locus, where 50% of African men had GGC alleles shorter than 14 while no more than 13% of men with European and Asian ancestry possessed the short GGC alleles. Thus a greater proportion of the haplotypes defined by short alleles at both loci (<20 CAG and <14 GGC) appear to segregate in African populations than in Asian and European populations. In fact, the African American population possesses a mixture of short allele haplotypes from different African populations. This has never been explored and is quite significant since both the CAG and GGC repeat loci influence the size of the protein, which subsequently affects transactivation of the receptor. These results parallel the prevalence of prostate cancer in human populations. Populations in which shorter CAG and GGC alleles are common, such as Africans, and specifically African Americans, have the highest incidence of prostate cancer in the world. The other end of the ethnic spectrum of prostate cancer incidence reveals that prevalence among Asians, who possess larger trinucleotide alleles, may be up to 50-fold less (Ross et al. 1996). Along with other genetic and environmental factors, this could likely yield a stronger predisposition among the African American population for prostate cancer. Recently a relationship was reported between serum PSA levels and polymorphisms in the PSA and AR genes (Xue et al. 2001). Specifically, serum PSA levels increased by 7% with each decreasing AR CAG repeat allele size among individuals homozygous for a single nucleotide polymorphism in the PSA gene promoter.

Our calculation of variance in the number of trinucleotide repeats provided a reliable measure of diversity since the two markers are microsatellites that conform to a stepwise mutation model (DiRienzo et al. 1994; Valdes et al. 1993). Larger variances and higher numbers of alleles were observed for the CAG locus than the GGC locus

among all six populations. The higher diversity at the CAG locus is likely due to a higher mutation rate at the CAG locus than the GGC locus. CAG repeat variation has been shown to cause several human diseases, such as Kennedy's disease (MIM 313200), Huntington's disease (MIM 143100), and several forms of spinocerebellar ataxias: SCA1 (MIM 164400), SCA2 (MIM 183090), SCA3 (MIM 109150), and SCA7 (MIM 164500). All of these CAG repeat loci are polymorphic in normal individuals. However, there appear to be constraints on allele size in populations since disease results when the CAG repeat lengths reach a certain threshold. A recent study of the ERDA1 locus revealed that large CAG repeats are more common among Asian populations, less common in populations of European ancestry, and least common in African populations (Deka et al. 1999). This pattern is very similar to that which was observed in our study of the AR trinucleotide repeats.

To explore population genetic affinities based on the AR gene CAG and GGC repeats we performed an AMOVA. Almost 20% of AR gene variance is attributed to differences between populations. Pairwise genetic distance values were significant for all population pairs except those with shared ancestry, such as between the Asian and Amerindian populations and Nigerian and Sierra Leone populations. Much of the genetic differences between populations may be due to genetic drift. Since the AR gene is X-linked, it is more vulnerable to the effects of drift than similar markers on other autosomes. This is due to a lower recombination rate and smaller effective population size for X-linked markers (approx. three-fourths of non-X-linked autosomal markers). Although the estimate for AR genetic differentiation (**ΦST**) between populations is higher than non-X-linked autosomal markers, it is not as high as estimates for the haploid systems of mtDNA and the Y chromosome (see Jorde et al. 2000). This is because the effective population size for mtDNA and the Y chromosome is about one-third lower than for the X chromosome.

Studies that have examined the association of alleles at the AR CAG and GGC repeat loci in relation to the development of prostate cancer have been ambiguous. Irvine et al. (1995) examined AR trinucleotide repeat variation in prostate cancer cases and controls from three ethnic groups, Euroamericans, African Americans, and Asians. LD was observed within the mixed group of cases but not within any one ethnic group. Several factors may have led to this observation. The prostate cancer cases consisted of three diverse populations, and therefore it is highly likely that stratification existed when they were pooled together. Also, since the sample sizes of the three groups were low (<50) it is likely that there was not sufficient statistical power to detect LD between the two markers. Later Stanford and colleagues (1997) examined a relatively large sample of Euroamericans. Their sample size of 301 cases and 277 controls failed to detect any LD between the markers within the two groups. A larger study consisting of 582 cases and 794 controls (Platz et al. 1998) revealed significant LD in both the cases and controls. The Platz et al. (1998) finding, using mainly Euroamerican men, was significant only after they pooled alleles for the two markers into categories of fewer than 23, 23, and more than 23 repeats.

It is a general rule that strong disequilibrium indicates that two marker loci are closely spaced. However, it is not always true that two closely spaced markers show disequilibrium. The frequencies of marker alleles and sample size affect the power to detect LD. Also, recombination and/or mutation hotspots could affect LD by increasing the chance that the associated marker allele will change. Not only is haplotype variation shaped by accumulated mutation within haplotypic lineages, it is also fashioned by recombination events among the lineages. It is unlikely that the decay of LD and pattern of variability observed for the AR CAG and GGC defined haplotypes is due to a recombination hotspot between the markers since they are separated by only 1 kb, and recombination on the X chromosome occurs only in women. However, recombination cannot be ruled out, especially since recombination hotspots are more likely in areas of high GC-rich regions in the genome (Eisenbarth et al. 2000). Population history can also have an effect on the extent of LD. Our observation of no detectable LD among the non-African populations may be explained by recent population growth and the high mutation rate at the CAG repeat locus. This is in contrast to the population bottleneck explanation for higher LD levels outside of Africa (Kidd et al. 1998; Tishkoff et al. 1996, 1998). Our finding of significant LD in the African American population is due mainly to gene flow from other populations. Admixture between populations with divergent allele frequencies can generate LD extended beyond 30 cM (Lautenberger et al. 2000). Finally, genetic drift can greatly affect or reinforce existing associations. The role of genetic drift in increasing or decreasing LD may be more significant among the Amerindians since there were a smaller number of alleles observed among the Amerindians than among the other populations. This is consistent with observations of low genetic diversity among Amerindians due to their history of recent population bottlenecks (Kittles et al. 1999; Nei and Roychoudhury 1993; Urbanek et al. 1996).

As previously stated, the high level of linkage disequilibrium observed among African Americans is likely due to multiple sources of admixture. Of the significant allelic associations between the trinucleotide markers in the African American population from South Carolina, 26% appear to have originated from European Americans, while 39% were shared among West African populations from Nigeria and Sierra Leone. These results suggest that the LD generated in African Americans from Columbia, South Carolina, may be due to recent migration of African Americans from diverse rural communities following urbanization, recurrent gene flow from distinct West African populations, and admixture with European Americans. Columbia is the capital of South Carolina and is located in the center of the state. Many African Americans migrated to this region from the coastal Sea Islands and the Low Country (Berkeley, Charleston, Colleton, and Dorchester

counties) during the early 1900s. In the late 1700s the percentage of persons of African origin was quite high in the coastal areas, including the port of Charleston (ranging from 47% to 93%). It also appears that colonial South Carolinians preferred certain African ethnic groups over others as slaves (Littlefield 1981; Morgan 1998). For instance, for a period of time in South Carolina enslaved Africans from Senegambia were preferred over others (Littlefield 1981). This preference was based mainly on the Senegambian's familiarity with rice production, which was the chief crop cultivated in the Carolinas at the time. However, these preferences changed in time along with the changing slave economy in the colonies. The changing trends, along with the relative isolation of the coastal communities of South Carolina likely led to diverse South Carolina African American populations. Subsequently, divergent haplotypes were brought together as people left the rural communities for more urban areas such as Columbia.

This assessment of linkage disequilibrium in the African American population is quite significant for several reasons. First, the high level of stratification in the African American population may be a confounder in disease association studies if the substructure is not controlled for. Second, the identification of high-risk haplotypes is potentially more powerful in disease studies than single locus analyses. We intend to increase the resolution in identification of these possible high-risk haplotypes for prostate cancer by typing single nucleotide polymorphisms within the gene and performing haplotype analyses. Ultimately these studies will provide a better understanding of the role variation within the AR plays in prostate cancer etiology.

Acknowledgements We thank C. Ahaghotu, S.O.Y. Keita, J. Long and C. Rotimi for helpful discussions. We also thank two anonymous reviewers for useful critiques, and W.T. Garvey, and M. Shriver for DNA samples. This work was supported by grants RR03048-13S1 from the National Institutes of Health and DAMD17-00-1-0025 from the Department of Defense.

References

Brawley OW, Kramer BS (1996) Epidemiology of prostate cancer. In: Volgelsang NJ, Scardino PT, Shipley WU, Coffey DS (eds) Comprehensive textbook of genitourinary oncology. Williams and Wilkins, Baltimore

Chamberlain NL, Driver ED, Miesfeld RL (1994) The length and location of CAG trinucleotide repeats in the androgen receptor N-terminal domain affect transactivation function. Nucleic Acids Res 22:3181–3186

Deka R, Guangyun S, Wiest J, Smelser D, Chunhua S, Zhong Y, Chakraborty R (1999) Patterns of instability of expanded CAG repeats at the ERDA1 locus in general populations. Am J Hum Genet 65:192–198

DiRienzo A, Peterson A, Garza J, Valdes A, Slatkin M, Freimer N (1994) Mutational processes of simple-sequence repeat loci in human populations. Proc Natl Acad Sci USA 91:3166–3170

Edwards A, Hammond HA, Jin L, Caskey CT, Chakraborty R (1992) Genetic variation at five trimeric and tetrameric tandem repeat loci in four human population groups. Genomics 12: 241–253

Eisenbarth I, Vogel G, Krone W, Vogel W, Assum G (2000) An isochore transition in the NF1 gene region coincides with a switch in the extent of linkage disequilibrium. Am J Hum Genet 67:873–880

Excoffier L, Smouse P, Quattro J (1992) Analysis of molecular variance inferred from metric distances among DNA haplotypes: application of human mitochondrial DNA restriction data. Genetics 131:479–491

Giovannucci E, et al (1997) The CAG repeat within the androgen receptor gene and its relationship to prostate cancer. Proc Natl Acad Sci USA 94:3320–3323

Glover F, Coffey D, et al (1998) The epidemiology of prostate cancer in Jamaica. J Urol 159:1984–1987

Guo S, Thompson E (1992) Performing the exact test of Hardy-Weinberg proportion for multiple alleles. Biometrics 57:212–215

Hardy D, Scher H, Bogenreider T, et al (1996) Androgen receptor CAG repeat lengths in prostate cancer: correlation with age of onset. J Clin Endocrinol Metab 81:4400–4405

Irvine RA, Yu MC, Ross RK, Coetze GA (1995) The CAG and GGC microsatellites are in linkage disequilibrium in men with prostate cancer. Cancer Res 55:1937–1940

Jackson FL (1993) Evolutionary and political economic influences on biological diversity in African Americans. J Black Studies 23:539–560

Jorde L, Bamshad M, Watkins W, Zenger R, Fraley A, Krakowiak P, Carpenter K, Soodyall H, Jenkins T, Rogers A (1995) Origins and affinities of modern humans: a comparison of mitochondrial and nuclear genetic data. Am J Hum Genet 57:523–538

Jorde L, Watkins W, Bamshad M, Dixon M, Rickers C, Seielsatd M, Batzer M (2000) The distribution of human genetic diversity: a comparison of mitochondrial, autosomal, and Y-chromosome data. Am J Hum Genet 66:979–988

Kidd KK, Morar B, Castiglione CM, Zhao H, Pakstis AJ, Speed WC, Bonne-Tamir B, Lu RB, Goldman D, Lee C, Nam YS, Grandy DK, Jenkins T, Kidd JR (1998) A global survey of haplotype frequencies and linkage disequilibrium at the DRD2 locus. Hum Genet 103:211–227

Kittles R, Bergen AW, Urbanek M, Virkkunen M, Linnoila M, Goldman D, Long JC (1999) Autosomal, mitochondrial, and Y chromosome variation in Finland: evidence for a male-specific bottleneck. Am J Phys Anthropol 108:381–399

LaSpada AR, Wilson A, Lubahn D, Harding A, Fishbeck K (1991) Androgen receptor gene mutations in X-linked spinal and bulbar muscular atrophy. Nature 352:77–79

Lautenberger JA, Stephens JC, O'Brien S, Smith M (2000) Significant admixture linkage disequilibrium across 30 cM around the FY locus in African Americans. Am J Hum Genet 66:969–978

Lewontin RC (1964) The interaction of selection and linkage. I. General considerations: heterotic models. Genetics 49:49–67

Littlefield DC (1981) Rice and slaves: ethnicity and the slave trade in colonial South Carolina. University of Illinois Press, Champaign

Michalakis Y, Excoffier L (1996) A generic estimation of population subdivision using distances between alleles with special reference for microsatellite loci. Genetics 142:1061–1064

Morgan P (1998) Slave counterpoint: black culture in the eighteenth century Chesapeake and Lowcountry. University of North Carolina Press, Raleigh

Nei M, Roychoudhury K (1993) Evolutionary relationships of human populations on a global scale. Mol Biol Evol 10:927–943

Nei M (1987) Molecular evolutionary genetics. Columbia University Press, New York

Nei M, Takezaki N (1996) The root of the phylogenetic tree of human populations. Mol Biol Evol 13:170–177

Ogunbiyi J, Shittu O (1999) Increased incidence of prostate cancer in Nigerians. J Natl Med Assoc 3:159–164

Orr H, Chubg M, Banifi S, Kwiatkoski T, Servadio A, Beaudet A, McCall A, et al (1993) Expansion of an unstable trinucleotide CAG repeat in spinocerebellar ataxia type 1. Nat Genet 4: 221–226

Osegbe D (1997) Prostate cancer in Nigerians: facts and non-facts. J Urol 157:1340

Platz E, Giovannucci E, Dahl D, Krithivas K, Hennekens C, Brown M, Stampfer M, Kantoff P (1998) The androgen receptor gene GGN microsatellite and prostate cancer risk. Cancer Epidemiol Biomarkers Prev 7:379–384

Reich D, Goldstein D (1998) Genetic evidence for a Paleolithic human population expansion in Africa. Proc Natl Acad Sci USA 95:8119–8123

Ross R, Pike M, Coetzee G, Reichardt J, Yu M, Feigelson H, Stanczyk F, Kolonel L, Henderson B (1998) Androgen metabolism and prostate cancer: establishing a model of genetic susceptibility. Cancer Res 58:4497–4504

Sartor O, Zheng Q, Eastham J (1999) Androgen receptor gene CAG repeat length varies in a race-specific fashion in men without prostate cancer. Urology 53:378–380

Schneider S, Kueffer J, Roessli D, Excoffier L (1997) Arlequin ver. 1.1: a software for population genetic data analysis. Genetics and Biometry Laboratory, University of Geneva, Geneva

Shriver M, Jin L, Ferrell R, Deka R (1997) Microsatellite data support an early population expansion in Africa. Genome Res 7: 586–591

Slatkin M (1995) A measure of population subdivision based on microsatellite allele frequencies. Genetics 139:457–462

Stanford JL, Just JJ, Gibbs M, Wicklund KG, Neal CL, Blumstein BA, Ostrander EA (1997) Polymorphic repeats in androgen receptor gene: molecular markers of prostate cancer risk. Cancer Res 57:1194–1198

Tishkoff S, et al (1996) Global patterns of linkage disequilibrium at the CD4 locus and modern human origins. Science 271: 1380–1387

Tishkoff S, et al (1998) A global haplotype analysis of the mytonic dystrophy locus: implications for the evolution of modern humans and for the origin of mytonic dystrophy mutations. Am J Hum Genet 62:1389–1402

Urbanek M, Goldman D, Long JC (1996) The apportionment of dinucleotide repeat diversity in Native Americans and Europeans: a new approach to measuring gene identity reveals asymmetric patterns of divergence. Mol Biol Evol 13:943–953

Valdes A, Slatkin M, Freimer N (1993) Allele frequencies at microsatellite loci: The stepwise mutation model revisited. Genetics 133:737–749

Xue WM, Coetzee GA, Ross RK, Irvine R, Kolonel L, Henderson BE, Ingles SA (2001) Genetic determinants of serum prostate-specific antigen levels in healthy men from a multiethnic cohort. Cancer Epidemiol Biomarkers Prev 10:575–579