J Hum Genet (2006) 51:129–136
DOI 10.1007/s10038-005-0336-7

ORIGINAL ARTICLE

Esther Esteban · Natalia Rodon · Marc Via
Emili Gonzalez-Perez · Josep Santamaria
Jean-Michel Dugoujon · Farha El Chennawi
Mohamed Melhaoui · Mohamed Cherkaoui
Giuseppe Vona · Nourdin Harich · Pedro Moral

Androgen receptor CAG and GGC polymorphisms in Mediterraneans: repeat dynamics and population relationships

Received: 30 August 2005 / Accepted: 19 October 2005 / Published online: 20 December 2005

Abstract Microsatellite variation (CAG and GGC repeats) of the androgen receptor (AR) gene shows remarkable differences among African and non-African populations. In vitro studies have demonstrated an inverse relationship between the length of both microsatellites and AR activity. This fact may explain the observed association of the AR gene with prostate cancer and the strong ethnic differences in the incidence of this cancer. CAG and GGC genetic variation has been tested in a large set of populations from the Ivory Coast as well as 12 Mediterranean samples whose variation is described for the first time. The pattern of frequencies observed in the Ivory Coast agrees with data previously reported for other Sub-Saharan populations. Concerning the Mediterranean variation, Sardinian samples are characterised by low genetic diversities, and Egyptian Siwa Berbers by a particular pattern of GGC frequencies. High and Middle Atlas Moroccan Berbers are the most closely related to the Sub-Saharan variation. For both the CAG and GGC repeats, the Ivory Coast and some Moroccan samples exhibit high frequencies of low size alleles (CAG under 18 repeats, and GGC under 15 repeats) that have been associated with prostate cancer.

Keywords Polyglutamine · Polyglycine · Androgen receptor · Prostate cancer · Population genetics

Electronic Supplementary Material Supplementary material is available for this article at http://dx.doi.org/10.1007/s10038-005-0336-7 and is accessible for authorized users.

E. Esteban · N. Rodon · M. Via · E. Gonzalez-Perez
J. Santamaria · P. Moral (✉)
Department of Animal Biology-Anthropology,
Faculty of Biology, University of Barcelona,
Avda. Diagonal, 645, 08028 Barcelona, Spain
E-mail: pmoral@ub.edu
Tel.: +34-934-021461
Fax: +34-934-035740

J.-M. Dugoujon
Centre of Anthropology, UMR 8555, CNRS, University Paul Sabatier, Toulouse, France

F. E. Chennawi
Immunology Unit, Mansoura Faculty of Medicine, Mansoura, Egypt

M. Melhaoui
General Ecology, Faculty of Sciences, University Mohamed Ist, Oujda, Morocco

M. Cherkaoui
Human Ecology, Faculty of Sciences, University Cadi Ayyad, Marrakech, Morocco

G. Vona
Department of Experimental Biology, University of Cagliari, Cagliari, Italy

N. Harich
Department of Biology, Faculty of Sciences,
University Chouaïb-Doukkali, El Jadida, Morocco

Introduction

The androgen receptor gene (AR; OMIM 313700) has been proposed as a strong candidate for risk and progression of prostate cancer (Porkka and Visakorpi 2004). This gene, located on the X chromosome (Xq11-12), belongs to the superfamily of nuclear hormone receptor (NR) genes that mediate the action of lipophilic ligands. The AR gene product regulates expression of the genes necessary for growth and development of many target tissues, including male and female reproductive organs. Protein functional domains include a variable N-terminal domain (NTD), a highly conserved DNA-binding domain, a hinge region, and a C-terminal ligand-binding domain (LBD). Transcriptional activation involves a ligand-induced intramolecular interaction between NTD and LBD domains. Upon binding the hormone ligand (dihydrotestosterone), the receptor dissociates from accessory proteins, translocates into the

nucleus, dimerises, and then stimulates transcription of androgen-responsive genes.

The N-terminal domain of the protein is encoded by exon 1 (Brinkmann et al. 1989) and contains two polymorphic short tandem repeats (STRs) (CAG and GGC) separated by approximately 1.1 kb. The CAG microsatellite encodes a polyglutamine tract that ranges in size from 5 to 33 glutamine (Gln) residues, with an average length of 20 trinucleotide repeats. The GGN tract ($GGT_3GGG_1GGT_2GGC_n$) encodes a polyglycine tract with a length of 22–24 glycines (average length of 16 GGC repeats). In vitro studies (Chamberlain et al. 1994; Ding et al. 2004) have demonstrated an inverse relationship between the length of both repeats and AR activity. Low size CAG (<19 repeats) and GGC (<15 repeats) alleles result in higher receptor activity, and have been associated with prostate cancer, earlier age of onset, and a higher grade and more advanced stage of prostate cancer at the time of diagnosis (Visakorpi 2003). Alleles of small GGC size have also been associated with oesophageal cancer (Dietzsch et al. 2003). In contrast, long CAG and GGC alleles are associated with decreased transactivation function in the AR receptor and have been associated with breast (Suter et al. 2003) and endometrial (Sasaki et al. 2003a) cancers.

CAG and GGC allele length variation has been described in populations (Kittles et al. 2001; Sasaki et al. 2003b) of the main ethnic groups (African, American of African descent, Asian and European) but, until now, information about CAG and GGC allele distribution in Mediterranean countries has been practically non-existent (Hadjkacem et al. 2004). For both STRs, remarkable allele size differences have been observed between African and non-African populations. These differences may be important to clarify differences in ethnic incidences of some cancers (Ferlay et al. 2004). Prostate cancer incidence in American men of African descent (age standardized rate (ASR) $=271/100{,}000$) are higher than those observed in American men of European descent (ASR $=167/100{,}000$), whereas the incidences of prostate cancer and other steroid-related cancers are remarkably low in Asia.

This work intends to report data describing CAG and GGC allele variation in a sub-Saharan African sample of the Ivory Coast as well as in 12 Mediterranean groups. The peopling of the Mediterranean is a profusely studied anthropological topic (Plaza et al. 2003; Quintana-Murci et al. 2003; Esteban et al. 2004). Rather than going deeper into this subject, although the populations and markers will allow us to do so, the main goal here is to obtain new data about the distribution of the above polymorphisms in South Europeans (seven samples from Spain, Italy, Greece and Turkey) and North Africans (five samples from Morocco, Algeria and Egypt). The proposed associations among the AR CAG and GGC STRs, and risk for several cancers make them particularly interesting. Knowledge of the current population variation of these two markers may be extremely informative as a baseline for the design of future epidemiological studies. It is well known that population stratification due to population differences in allele frequencies may lead to confounding results in association studies (Ardlie et al. 2002). This fact must be seriously considered in a region such as the Mediterranean, with important South–North migratory movements, both in the past and recently. Moreover, different genetic analyses based on uniparental and autosomal markers (Plaza et al. 2003; Quintana-Murci et al. 2003; Esteban et al. 2004) indicate a certain degree of Sub-Saharan African admixture in some North African populations. To test the magnitude of this contribution in terms of the CAG and GGC STR variation may contribute relevant information to population-based research in genetic epidemiology.

Materials and methods

Populations sampled

Blood samples were collected from healthy and unrelated males and females from 13 different groups within eight countries. Samples were obtained with the informed consent of the participants. All participants had their four grandparents born in the same region. With the exception of the Greek sample, participants were representatives of rural areas of geographically well-defined regions. Some of these geographical regions are the homeland of several anthropologically well-defined Berber groups in Morocco, Algeria, and Egypt. Three Spanish samples were analysed from northwest Spain (number of chromosomes $n=106$), southeast Spain ($n=112$), and the Basque Country ($n=62$). The island of Sardinia was sampled in two different localities: inner Sardinia ($n=68$) and coastal Sardinia ($n=86$). Greece (Athica region, $n=66$) and Turkey (Anatolia Peninsula, $n=150$) completed the North Mediterranean sampling. The Moroccan samples consisted of three Berber groups from High Atlas ($N=88$), Middle Atlas ($N=118$), and Northeast Atlas ($n=90$). Berbers from Mzab in Algeria ($n=74$) and from the Siwa Oasis in Egypt ($n=82$) were also analysed. Finally, a sample from the Ahizi ethnic group ($n=88$) from the Ivory Coast was genotyped to include a representation of the Sub-Saharan African variation.

Genetic analyses

CAG and GGC STRs were amplified using the methodology defined in previous studies (Kittles et al. 2001; Sasaki et al. 2003b). PCR products were pooled and electrophoresed on an ABI PRISM 3700 DNA sequencer (Applied Biosystems, Foster City, CA). Genescan and Genemapper 3.0 programs (ABI PRISM, Applied Biosystems) were used to generate fragment sizes and genotypes. Different male individuals selected by their CAG and GGC alleles were sequenced to confirm size lengths.

Statistical analyses

Trinucleotide allele frequencies were determined by direct gene counting. Allele frequencies in both sexes were compared by means of a G-test. Male and female data were grouped since no sex differences were observed.

The mean number of repeats and its variability (heterozygosity and variance of the allele distribution) were used as indicators of repeat dynamics. Two statistical parameters, β and ρ, proposed by Deka et al. (1999) to detect demographic and microsatellite mutational traits, were estimated. The β parameter is known as the imbalance of heterozygosity and allele size variance; the β value may differentiate diverse population demographic situations (constant population size, population growth after bottleneck or after equilibrium) or detect allele size constraints. The ρ parameter is the slope of the regression line between the mean and variance of allele size. This parameter is independent of demographic effects and provides information about the existence of a contraction or expansion bias in the STR. Theoretical predictions of these two parameters interpreted together concerning demographic and mutational aspects are detailed in previous works (Deka et al. 1999; Andrés et al. 2002).

Mean, variance, median, and ρ parameters were calculated using SPSS v.10.0 and STATISTICA software packages. The β parameter and its confidence intervals (CI; determined by re-sampling) were estimated using Microsoft Excel 2000 software as described in Andrés et al. (2002). For each STR, overall β and ρ values were computed by pooling all populations. Standard diversity indices, population comparisons (exact test of population differentiation), and hierarchical analyses of molecular variance (AMOVA) were implemented by the Arlequin v 2.0 package (Schneider et al. 2002).

Results

AR CAG and GGC allele size distributions in the 13 studied populations are available as Electronic Supplementary Material. In general, CAG and GGC distributions are in Hardy–Weinberg equilibrium; however, departures from equilibrium (tested in the female subsamples) were found in 6 cases out of 26 tests. The low females' number in some samples, together with the high number of alleles in both STRs are probably the cause of these Hardy–Weinberg equilibrium departures. Allele diversity values and other statistical parameters for the CAG and GGC markers are shown in Tables 1 and 2, respectively.

CAG variation

The CAG repeat shows notable levels of within-population variation (heterozygosity values from 84 to 89%) in all samples. Variance in the allele sizes is more heterogeneous; some populations, such as Siwa, Middle and High Atlas, and South Spain, show high values (10.37–9.28), whereas other groups, including the Ivory Coast, Turkey, Basque Country, and Greece, exhibit the lowest values (5.91–5.03). The high number of different alleles found in all groups (10–17) makes difficult to ascribe a global population variation pattern. In an attempt to summarise allele variation, CAG allele frequencies were grouped (see Table 1) into three different categories defined by the observed median repeat values: short alleles of less than 19 repeats, medium alleles of 19 to 21 repeats, and long alleles of more than 21 repeats. This summarised information reveals that our European samples are characterised by low (mean average frequency of 7%) and high (51%) frequencies of the short and long alleles, respectively. The Ivory Coast shows the

Table 1 Statistics and genetic variance parameters of the CAG trinucleotide in the androgen receptor (AR) gene in 13 populations. Absolute and relative (in parentheses) repeat frequencies are grouped into alleles of less than 19 repeats, alleles of 19 to 21 repeats, and alleles of more than 21 repeats. *n* Number of gene copies (chromosomes), *H* heterozygosities; *no. alleles* number of different alleles

Population	*n*	*H*	Mean	Range	Median	Variance	No. alleles	Alleles < 19	Alleles 19–21	Alleles > 21	β[a]
South Spain	112	0.8842	22.25	14–31	22	9.28	17	7 (6.25%)	44 (39.28%)	61 (54.47%)	0.26 (0.20–0.48)
Basque Country	62	0.8413	21.42	15–27	21	5.07	11	4 (6.45%)	28 (45.16%)	30 (48.39%)	0.26 (0.25–0.79)
North Spain	105	0.8736	21.89	14–31	21	7.01	14	4 (3.81%)	50 (47.61%)	52 (48.58%)	0.23 (0.17–0.39)
Inner Sardinia	68	0.8888	21.84	14–28	22	8.76	12	8 (11.76%)	23 (33.82%)	37 (54.42%)	0.22 (0.20–0.47)
Coast Sardinia	85	0.8772	21.39	16–29	21	8.31	13	11 (12.94%)	36 (42.35%)	39 (44.71%)	0.25 (0.22–0.54)
Greece	66	0.8586	21.33	16–26	21	5.03	10	5 (7.57%)	30 (45.45%)	31 (46.98%)	0.29 (0.26–0.70)
Turkey	149	0.8662	22.59	14–31	22	5.97	15	5 (3.35%)	51 (34.23%)	94 (62.42%)	0.22 (0.17–0.35)
Siwa Berbers	81	0.8733	20.58	14–30	20	10.37	14	29 (35.80%)	17 (20.98%)	36 (43.22%)	0.34 (0.19–0.63)
Mzab Berbers	74	0.8820	21.11	14–26	21	7.39	12	13 (17.56%)	26 (35.13%)	35 (47.31%)	0.21 (0.15–0.34)
North-East Atlas	92	0.8790	21.51	14–28	21	8.87	14	11 (11.96%)	35 (38.04%)	46 (50.00%)	0.26 (0.17–0.50)
Middle Atlas	117	0.8917	21.01	12–28	21	9.61	16	21 (17.95%)	45 (38.46%)	52 (43.59%)	0.23 (0.17–0.37)
High Atlas	87	0.8751	20.54	10–29	20	9.39	15	17 (19.54%)	37 (42.53%)	34 (37.93%)	0.30 (0.19–0.55)
Ivory coast	89	0.8749	19.13	14–24	19	5.91	11	32 (35.95%)	45 (50.56%)	13 (13.49%)	0.19 (0.18–0.43)

[a]Values in parentheses are 99% confidence intervals (CI) of the β values

Table 2 Statistics and genetic variance parameters of the GGC trinucleotide in the androgen receptor (AR) gene in 13 populations. Absolute and relative (in parentheses) repeat frequencies are grouped into alleles of less than 15 repeats, alleles of 15 to 17 repeats, and alleles of more than 17 repeats

Population	*n*	*H*	Mean	Range	Median	Variance	No. alleles	Alleles < 15	Alleles 15–17	Alleles > 17	β[a]
South Spain	112	0.6508	15.85	12–19	16	2.02	8	18 (16.07%)	91 (81.25%)	3 (2.68%)	0.56 (0.40–0.81)
Basque Country	62	0.5146	15.65	10–18	16	2.30	9	8 (12.90%)	53 (85.48%)	1 (1.61%)	1.42 (0.67–3.12)
North Spain	105	0.5894	15.96	07–22	16	3.02	10	9 (8.57%)	90 (85.71%)	6 (5.71%)	1.22 (0.48–2.37)
Inner Sardinia	68	0.3598	16.24	16–17	16	0.18	2	–	68 (100%)	–	0.25 (0.18–0.35)
Coast Sardinia	85	0.0913	15.95	11–20	16	0.69	5	2 (2.35%)	82 (96.47%)	1 (1.18%)	6.51 (0.47–12.20)
Greece	66	0.5331	16.23	12–17	16	1.10	3	3 (4.54%)	63 (95.46%)	–	0.61 (0.15–0.83)
Turkey	149	0.4596	15.92	11–18	16	1.36	8	11 (7.38%)	137 (91.95%)	1 (0.67%)	1.12 (0.60–1.89)
Siwa Berbers	81	0.6954	14.95	09–26	15	5.15	7	13 (16.05%)	67 (82.72%)	1 (1.23%)	1.05 (0.47–1.29)
Mzab Berbers	74	0.7100	15.35	11–17	16	3.14	6	12 (16.21%)	62 (83.79%)	–	0.58 (0.34–0.95)
North-East Atlas	92	0.6309	15.57	07–18	16	3.61	9	14 (15.22%)	74 (80.43%)	4 (4.35%)	1.14 (0.41–2.16)
Middle Atlas	117	0.5996	16.02	09–20	16	1.52	10	8 (6.84%)	105 (89.74%)	4 (3.42%)	0.58 (0.35–1.20)
High Atlas	87	0.7419	15.61	07–18	16	4.87	10	17 (19.54%)	63 (72.42%)	7 (8.04%)	0.69 (0.26–0.92)
Ivory Coast	89	0.7863	15.11	08–19	15	3.87	10	23 (25.84%)	62 (69.67%)	4 (4.49%)	0.37 (0.19–0.55)

[a]Values in parentheses are 99% CI of the β values

opposite trend (36% of short and 13% of long alleles). North African populations exhibit intermediate values of short alleles (17%), and the lowest frequencies of the medium category (35%). A recent work (Buchanan et al. 2004) describes an allele range (from 16 to 29 repeats) considered as the critical size for maintaining a protein NTB/LBD interaction that ensures adequate AR activity. Under this criterion, the Ivory Coast (8.98% of alleles < 16 repeats) and, to a lesser degree, High Atlas (5.75%) and Middle Atlas (4.27%) Berbers, show at least twice as many alleles under this critical size than those observed in the remaining groups.

The imbalance of heterozygosity and allele size variance, expressed as the β parameter, is shown in Table 1 for each population. The global β yields a pooled population value of 0.183 (99% CI: 0.175–0.198) whereas the ρ parameter ($\rho = -0.047$) is not significantly different from zero.

GGC variation

In contrast to CAG, the GGC locus (Table 2) is less variable. The majority of samples analysed have 6–10 different alleles, with the exception of inner Sardinia and Greece, which have more reduced variation (2 and 3 different alleles). The repeat size pattern shows a highly frequent allele of 16 repeats in all populations except those of the Ivory Coast and Siwa Berbers. In these two latter groups, the 15-repeat allele is found at high frequencies. As can be seen in Table 2, more than 80% of the total allele frequencies are explained by the presence of the 15–17 repeat alleles in all samples except those of the Ivory Coast and High Atlas Berbers. Alleles of less than 15 repeats are found only in notable frequencies (12–25%) in the African groups (except Middle Atlas) and in the Iberian samples of South Spain (16.07%) and the Basque Country (12.90%). High heterozygosities are observed in African samples (from 79% in the Ivory Coast to 60% in Middle Atlas Berbers) and South Spain (65%). Population heterozygosity values were compared by means of a non-parametric ANOVA. Significant differences ($\chi^2_1 = 6.33$, $P = 0.012$) were found in the comparison between north (seven samples) and south (five samples) Mediterraneans. These differences remain significant ($\chi^2_1 = 6.00$, $P = 0.014$) when the Northeast Mediterranean (inner Sardinia, coast Sardinia, Greece and Turkey) samples are compared with North Africans, but not for the North African versus Iberian Peninsula comparison ($\chi^2_1 = 2.68$, $P = 0.101$).

As for the imbalance index, β values for each population are indicated in Table 2. The overall β value (0.792, 99% CI: 0.636–1.002) is not significantly different from 1. The ρ value is -3.28 ($P = 0.00035$).

Population relationships

Pairwise population comparisons for the CAG locus report only 14 significant results out of 78 comparisons. In contrast, the GGC locus is more heterogeneous since 40 out of 78 comparisons are significant. Leaving aside some punctual population differences, coastal Sardinia, the Ivory Coast, and the Siwa Berbers show a pattern of GGC allele frequencies that differs significantly from the remaining groups.

The proportion of the genetic variance attributable to differences among groups shows remarkable differences between the CAG and GGC markers. For the CAG locus, only a global F_{ST} of 0.52% (13 samples, $P = 0.039$) yields a weak but significant result. However, when CAG frequencies are grouped into the short, medium, and long categories, the hierarchical analyses indicated in Table 3 shows a significant geographical structure related to these allele size categories. When the North Mediterranean group is compared with the African groups, both including and non-including the Ivory

Table 3 Variability analysis in the CAG and GGC loci

Non-hierarchical analysis			Hierarchical F_{ST} among populations			
	n	F_{ST}	Compared groups	Within groups	Among groups	Total F_{ST}
CAG[a]			CAG[a]			
All populations	13	0.029***	North Mediterraneans/ Africans	0.014*	0.029***	0.042***
Mediterranean populations	12	0.008 NS	North Mediterraneans/ North African	0.000 NS	0.019**	0.017 NS
North Mediterranean populations	7	0.000 NS				
North African populations	5	0.002 NS				
GGC			GGC			
All populations	13	0.090***	North Mediterraneans/ Africans	0.059***	0.058***	0.011***
Mediterranean populations	12	0.063***	North Mediterraneans/ North African	0.046***	0.033*	0.078***
North Mediterranean populations	7	0.049***	North Mediterraneans (without Sardinians)	0.011 NS	0.007 NS	0.019*
North African populations	5	0.043***	/North Africans (without Siwa Berbers)			

*$P < 0.05$; **$P < 0.01$; *** $P < 0.001$; *NS* non-significant
[a]CAG alleles have been grouped into short, medium and long categories

Coast, significant among–groups values (2.9 and 1.9%, respectively) are observed.

For the GGC locus, the amount of total genetic variance for the 13 populations examined shows a high value (F_{ST} of 9.01%; $P < 0.001$). Decreasing, but significant intra-groups F_{ST} diversity values are also observed (Table 3) in the different geographic Mediterranean groups. Hierarchical analyses indicate an apparently noticeable geographic structure related to this repeat. However, at the sight of the particular pattern of GGC frequencies showed by Siwa Berbers and Sardinians, genetic variances have been recalculated excluding these two samples. In this case, genetic variances change to non-significant values although a global F_{ST} of 1.9% ($P = 0.0237$) remains significant.

To summarise population relationships, different principal component (PC) analyses based on the allele frequencies of CAG and GGC have been checked. A plot (not shown) drawn with complete allele frequencies of the two repeats in the 13 populations analysed here, and also including two samples from Japan and Germany (Sasaki et al. 2003b), shows a population distribution around two main axes (43.6% of total variation) that clearly separates the Ivory Coast and Siwa Berbers from the remaining populations, with an intermediate position of some North African samples. The low percentage of variation expressed by the two main axes, in spite of the different populations included, may be due to both the low number of loci and the extreme allele dispersion of these STRs. A PC re-analysis based on

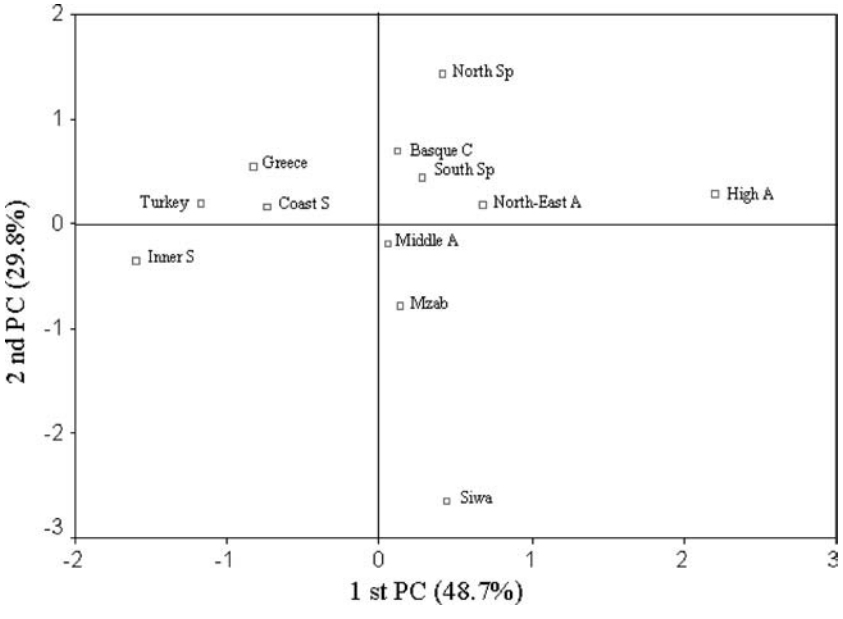

Fig. 1 Plot of the frequencies of the two principal components (*PC*) of CAG and GGC alleles of the androgen receptor (AR) gene in 12 Mediterranean populations

grouped allele frequencies (Tables 1, 2) for CAG and GGC among the 12 Mediterranean samples is represented in Fig. 1,explaining 78.5% of the genetic variance. The first axis (48.7%) distinguishes two clusters, one comprising Sardinian, Greek, and Turkish samples, and another composed of the remaining groups with the exception of High Atlas Berbers, which occupy an extreme position. The second axis (29.8%) separates the Siwa Berbers. The first axis is influenced by GGC diversity: allele frequencies of GGC < 15 repeats correlate best (87%) with the High Atlas differentiation, whereas the 15–17 repeats GGC frequencies are highly correlated (92.4%) with the East Mediterranean cluster. Also, the Siwa differentiation may be related to short CAG allele frequencies (75.2% correlation).

Discussion

This paper attempts to give a picture of the general population distribution of two STRs of the AR gene in a well-defined geographic area: the Mediterranean region. In this region, CAG data show similarly high within-population variation but the population distribution of short, medium, and long alleles is remarkably different between South Europeans and Sub-Saharan Africans; North Africans exhibit intermediate frequencies. As a result of this, the CAG allele distribution in the 12 Mediterranean samples shows a significant geographical (North-South) structure.

On the other hand, the variation at GGC locus allow us to distinguish two groups, one comprising African and Iberian Peninsula samples characterised by both high heterozygosities (H average $= 0.6577 \pm 0.0845$) and allele diversities (average number of different alleles $= 8.78 \pm 1.48$), and the other formed by Sardinia and Greece, with lesser diversities ($H = 0.3280 \pm 0.2226$; different alleles average $= 3.33 \pm 1.53$). The relatively general Mediterranean homogeneity for the GGC marker seems to be disrupted for some specific populations rather than by any North–South geographical structure. Thus, the Sardinian samples are characterised by low genetic diversities that may be interpreted as the result of genetic drift and geographic isolation, as evidenced by other autosomal (Cavalli-Sforza et al. 1994) and uniparental (Morelli et al. 2000) markers. Another well differentiated group are the Siwa; this Berber-speaking group is characterised by high endogamy levels (Fakhry 1973) and geographic isolation from the remaining Egyptian populations. In our case, the Siwa are differentiated mainly by the particular pattern of GGC allele sizes rather than by low genetic diversities, in accordance with the only previous genetic study conducted in this population that demonstrated important genetic diversity levels contrasting with the idea of an isolated population (Amory et al. 2004).

Finally, High Atlas Berbers are, among Moroccan Berbers, the most closely related to the Sub-Saharan variation. Also true to this trend, Southern Spaniards, among the Iberian samples, are the most related to the North African variation. These findings suggest genetic influences from South to North, both across the Sahara and the Mediterranean, as previously reported in other studies based on mt-DNA, Y-chromosome, and autosomes (Plaza et al. 2003; Quintana-Murci et al. 2003; Esteban et al. 2004).

The results regarding repeat dynamics and population demography reveal different patterns for CAG and GGC. The global CAG pattern ($\beta < 1$, $\rho \approx 0$) is compatible with a theoretical model of an unbiased constrained repeat; the $\beta < 1$ values observed in each sample (Table 1) may be interpreted as constraints in size (a gene particularity) rather than a population effect. The GGC pattern ($\beta \approx 1$, $\rho < 0$) agrees with a model of a constrained repeat with an expansion bias in expanding populations after a bottleneck. This demographic information is concordant with the known genetic evidence regarding populations of non-African origin. For this marker, the Ivory Coast sample, which shows (Table 2) an individual β of 0.37 (99% CI: 0.19–0.55), has a value compatible with a constant population model, as previous studies (Deka et al. 1999; Andrés et al. 2002) have suggested for other Sub-Saharan African groups.

From the point of view of the possible relationship between the dynamics of these two STRs and disease, the CAG locus analysed here has a pattern of high within-population variation and not significant ($\rho \approx 0$) correlation of mean and variance of allele sizes. A similar pattern has been described in disease-causing trinucleotides (Deka et al. 1999), even though the subjects analysed are disease-free and lie within the normal size ranges. On the contrary, the considerably low heterozygosities and within-population variances of the GGC STR, together with a significantly negative ρ value, is similar to the dynamics described for GC-rich gene-associated and anonymous loci.

With regard to the relationship between STR allele sizes, androgen receptor activity, and cancer risk, short CAG repeats (≤ 18 repeats relative to ≥ 26) have been associated with an increased risk (relative risk $= 2.14$) of advanced prostate cancer (Giovanucci et al. 1997). A more recent study (Buchanan et al. 2004) reveals that under a critical size of 16 CAG repeats, the conformational structure resulting from the short polyglutamine tract encoded by these repeats could enhance the binding of specific transcriptional coactivators, resulting in higher AR activity even at lower androgen concentrations. The Ivory Coast and all North African samples show high frequencies of CAG alleles ≤ 18 repeats. Furthermore, in the particular case of the Ivory Coast, High Atlas, and Middle Atlas Moroccan Berbers, the proportion of alleles under the size of 16 repeats (9, 6, and 4%, respectively) is at least two or three times higher than that observed in other samples.

Short GGC alleles have also been correlated with an increased risk of prostate cancer (Chang et al. 2002). A recent study (Ding et al. 2004) revealed that GGC repeat

sizes are directly correlated with cell AR protein levels: a GGC STR of 16 repeats (GGC16) yielded, on average, 1.7 times more AR protein than did GGC17, and GGC13 yielded 2.7 times more AR protein than did GGC17. The Ivory Coast, several North African samples, and South Spain exhibit high frequencies of alleles under the size of 15 repeats. The proportion of these alleles is 3.5–7 times higher than in Europeans.

Considering the chromosomal location (Xq12) of the AR gene, the effects of androgen in males is mediated by a single AR allele whereas in females, with two different AR alleles, random X-chromosome inactivation leads to effects of different alleles in different cells. The Ivory Coast, and High and Middle Atlas Moroccans show remarkably high frequencies of CAG and GGC alleles of low sizes. This fact may be translated into a considerable proportion of males, and, to a lesser degree, of females, in these populations carrying AR alleles of low sizes. To go beyond this affirmation, and to speculate about the possible relationship between the distributions of these repeats in the populations analysed here and cancer risk would be less than prudent. However, the data here reported must be discussed in relation to prostate cancer incidences.

Prostate cancer incidences (Ferlay et al. 2004), as indicated by means of ASR are: 19.7/100,000 in the Ivory Coast, 6.4/100,000 in Morocco, 5.6/100,000 in Algeria, and 4.4/100,000 in Egypt. Among South Europeans, the ASR is 35.9/100,000 in Spain, 40.5/100,000 in Italy, 26.2/100,000 in Greece, and 8.0/100,000 in Turkey. These data indicate higher incidences in the Ivory Coast than in North African countries. However, these rates also show incidence values at least five times lower in North Africans than in South Europeans and, hence, do not confirm the correlation between low size alleles and high prostate cancer incidence previously reported in other populations. In any case, these prostate cancer rates must be interpreted with caution due to the significant differences among African and European countries in some important aspects concerning cancer epidemiology: access to health services and diagnostic procedures [prostate-specific antigen (PSA); Liu et al. 2001], differences in several dietary factors [lycopene, total fat, and long-chain (*n*-3) fatty acids; Terry et al. 2004], and the remarkable divergence in population age structure among these countries.

We are aware of the extreme difficulty of genetic epidemiology studies, in particular those related to cancer, in which different genetic and environmental factors may contribute with hundreds of pieces to a very complicated puzzle. Given this fact, and in the light of the remarkable differences in CAG and GGC frequencies among the Ivory Coast, several Moroccan groups, and South Europeans, new studies with a multi-ethnic perspective may contribute to clarifying the risk factors associated with prostate cancer. Since CAG and GGC STRs are located in the same exon, further analyses of the degree of linkage disequilibrium should be strongly considered.

Acknowledgements We are grateful to all the donors for providing blood samples and to the people who contributed to their collection. In particular, we thank Prof. André Chaventré and Dr. Gil Bellis (for the samples from Ivory Coast), Dr. Francisco Luna (samples from South Spain), Dr. Angelica Saetta (samples from Greece), and Dr. Nisrine Bissar (samples from Turkey). We acknowledge with thanks the help of Prof. Josep Anton Sanchez (Departament d'Estadística, Facultat de Biologia, UB). We are grateful to the Unit of Descriptive Epidemiology at the International Agency for Research on Cancer (http://www.dep.iarc.fr) for access to the GLOBOCAN 2002 database. This research was supported by the Ministerio de Ciencia y Tecnología projects BMC2002-01224 and BSO2002-10225-E. This work, as part of the European Science Foundation EUROCORES Programme OMLL, was supported by funds from the Spanish Ministerio de Educacion y Ciencia BSO2002-10225-E, CNRS (Centre National de la Recherche Scientifique, France) and the EC Sixth Framework Programme under contract no. ERAS-CT-2003-980409. The sampling of the Berbers from Morocco and Egypt was supported by the Conseil Régional de Midi-Pyrénées, Toulouse (France).

References

Amory S, Dugoujon JM, Despiau S, Roubinet F, El Chenawi F, Blancher A (2004) Identification de trois nouveaux allèles *0* dans une population berbère de Siwa (Egypte). Antropo 7:105–112

Andres AA, Lao O, Soldevila M, Calafell F, Bertranpetit J (2002) Dynamics of CAG repeat loci revealed by the analysis of their variability. Hum Mut 21: 61–70. DOI 10.1002/humu.10151

Ardlie KG, Lunetta KL, Seielstad M (2002) Testing for population subdivision and association in four case-control studies. Am J Hum Genet 71:304–311

Brinkmann AO, Faber PW, van Rooij HC, Kuiper GG, Ris C, Klaassen P, van der Korput JA, Voorhorst MM, van Laar JM, Mulder E et al (1989) The human androgen receptor (domain structure, genomic organization and regulation of expression. J Steroid Biochem 34:307–310

Buchanan G, Yang M, Cheong A, Harris JM, Irvine RA, Lambert PF, Moore NL, Raynor M, Neufing PJ, Coetzee GA, Tilley WD (2004) Structural and functional consequences of glutamine tract variation in the androgen receptor. Hum Mol Genet 13:1677–1692. DOI 10.1093/hmg/ddh181

Cavalli-Sforza LL, Menozzi P, Piazza A (1994) The history and geography of human genes. Princeton University Press, Princeton, NJ

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

Chang BL, Zheng SL, Hawkins GA, Isaaca SD, Wiley KE, Turner A, Carpten JD, Bleecker ER, Walsh PC, Trent JM, Meyers JA, Isaacs WB, Xu J (2002) Polymorphic GGC repeats in the androgen receptor gene are associated with hereditary and sporadic prostate cancer risk. Hum Genet 110:122–129. DOI 10.1007/s00439-001-0662-6

Deka R, Guangyun S, Smelser D, Zhong Y, Kimmel M, Chakraborty R (1999) Rate and directionality of mutations and effects of allele size constraints at anonymous, gene-associated, and disease-causing trinucleotide loci. Mol Biol Evol 16:1166–1177

Dietzsch E, Laubscher R, Parker MI (2003) Esophageal cancer risk in relation to GGC and CAG trinucleotide repeat lengths in the androgen receptor gene. Int J Cancer 107:38–45. DOI 10.1002/ijc.11314

Ding D, Xu L, Menon M, Veer Reddy GP, Barrack ER (2004) Effect of GGC (Glycine) repeat length polymorphism in the human androgen receptor on androgen action. Prostate 9999:1–7. DOI 10.1002/pros.20128

Esteban E, González-Pérez E, Harich N, López-Alomar A, Via M, Luna F, Moral P (2004) Genetic relationships among Berbers and South Spaniards based on CD4 microsatellite/Alu haplotypes. Ann Hum Biol 31:202–212. DOI 10.1080/03014460310001652275

Fakhry A (1973) Siwa oasis. The American University in Cairo Press, Cairo

Ferlay J, Bray F, Pisani P (2004) GLOBOCAN 2002: Cancer incidence, mortality and prevalence worldwide, ver 2.0. IARC: Cancer Base no. 5, IARC Press, Lyon

Giovannuci E, Stampfer MJ, Krithivas K, Brown M, Dahl D, Brufsky A, Talcott J, Hennekens CH, Kantoff PW (1997) The CAG repeat within the androgen receptor gene and its relationship to prostate cancer. Proc Natl Acad Sci USA 94:3320–3323

Hadjkacem L, Hadj-Kacem H, Boulila A, Bahloul A, Ayadi H, Ammar-Keskes L (2004) Androgen receptor gene CAG repeats length in fertile and infertile Tunisian men. Ann Genet 47:217–224. DOI 10.1016/j.anngen.2004.03.010

Kittles RA, Young D, Weinrich S, Hudson J, Argyropoulos G, Ukoli F, Adams-Campbell L, Dunston GM (2001) Extent of linkage disequilibrium between the androgen receptor gene CAG and GGC repeats in human populations: implications for prostate cancer risk. Hum Genet 109:253–261. DOI 10.1007/s004390100576

Liu L, Cozen W, Bernstein L, Ross RK, Deapen D (2001) Changing relationship between socioeconomic status and prostate cancer incidence. J Natl Cancer Inst 93:705–709

Morelli L, Grosso MG, Vona G, Varesi L, Torroni A, Francalacci P (2000) Frequency distribution of mitochondrial DNA haplogroups in Corsica and Sardinia. Hum Biol 72:585–595

Plaza S, Calafell F, Helal A, Bouzerna N, Lefranc G, Bertranpetit J, Comas D (2003) Joining the pillars of Hercules: mtDNA sequences show multidirectional gene flow in the Western Mediterranean. Ann Hum Genet 67:312–328. DOI 10.1046/j.1469-1809.2003.00039.x

Porkka KP, Visakorpi T (2004) Molecular mechanisms of prostate cancer. Eur Urol 45:683–691. DOI 10.106/jeururo.2004.01.012

Quintana-Murci L, Veitia R, Fellous M, Semino O, Poloni ES (2003) Genetic structure of the Mediterranean populations revealed by Y-chromosome haplotype analysis. Am J Phys Anthropol 121:157–171. DOI 10.1002/ajpa.10187

Sasaki M, Sakuragi N, Dahiya R (2003a) The CAG repeats in exon 1 of the androgen receptor gene are significantly longer in endometrial cancer patients. Biochem Biophys Res Commun 305:1105–1108. DOI 10.1016/s0006-291x(03)00883-0

Sasaki M, Kaneuchi M, Sakuragi N, Fujimoto S, Carroll PD, Dahiya R (2003b) The polyglycine and polyglutamine repeats in the androgen receptor gene in Japanese and Caucasian populations. Biochem Biophys Res Commun 312:1244–1247. DOI 10.1016/j.bbrc.2003.11.075

Schneider S, Roessli D, Excoffier L (2002) A software for population genetics data analysis. Genetics and Biometry Laboratory, University of Geneva, Switzerland

Suter NM, Malone KE, Daling JR, Doody DR, Ostrander EA (2003) Androgen receptor (CAG)(n) and (GGC)(n) polymorphisms and breast cancer risk in a population-based case-control study of young women. Cancer Epidemiol Biomarkers Prev 12:127–135

Terry PD, Terry JB, Rohan TE (2004) Long-chain (*n*-3) fatty acid intake and risk of cancers of the breast and the prostate: recent epidemiological studies, biological mechanisms, and directions for future research. J Nutr 134:3412S–3420S

Visakorpi T (2003) The molecular genetics of prostate cancer. Urology 62:3-10. DOI 10.1016/s0090-4295(03)00776-3