BMC Genetics

Research article

Study of human SP-A, SP-B and SP-D loci: allele frequencies, linkage disequilibrium and heterozygosity in different races and ethnic groups

Wenlei Liu[1], Christy M Bentley[1] and Joanna Floros*[2,3,4]

Address: [1]Department of Health Evaluation Sciences, Penn State College of Medicine, Hershey, PA 17033, USA, [2]Department of Cellular and Molecular Physiology, Penn State College of Medicine, Hershey, PA 17033, USA, [3]Department of Pediatrics, Penn State College of Medicine, Hershey, PA 17033, USA and [4]Department of Obstetrics and Gynecology, Penn State College of Medicine, Hershey, PA 17033, USA

Email: Wenlei Liu - wliu@hes.hmc.psu.edu; Christy M Bentley - cbentley@hes.hmc.psu.edu; Joanna Floros* - jfloros@psu.edu

* Corresponding author

Published: 11 August 2003

BMC Genetics 2003, **4**:13

Received: 12 June 2003
Accepted: 11 August 2003

This article is available from: http://www.biomedcentral.com/1471-2156/4/13

Abstract

Background: SP-A, SP-B, and SP-D are pulmonary surfactant proteins. Several linkage and association studies have been done using these genes as markers to locate pulmonary disease susceptibility genes, but few have studied the markers systematically in different ethnic groups. Here we studied eight markers in *SP-A*, *SP-B*, and *SP-D* genes in seven ethnic groups from three races (Caucasian, Black and Hispanic). We measured the similarity of the marker distribution among the ethnic groups in order to see whether people in different ethnic groups or races could be mixed together for linkage and association studies. To evaluate the usefulness of these markers, we estimated the informativeness of each marker loci in the seven ethnic groups by assessing their heterozygosity and PIC values. We also conducted linkage disequilibrium (LD) analysis to identify associated marker loci and to estimate the haplotype frequencies in each of the seven ethnic groups in an attempt to find valuable haplotypes so that the level of polymorphism of the "markers" could be increased.

Results: Our findings indicate that allele and genotype frequencies may be different between different ethnic groups, especially between ethnic groups from different races. The markers are in general polymorphic in a variety of study groups, especially for the two SP-A1 and SP-A2 markers. Two-locus LD analysis reveals that three pairs of loci are strongly associated together: B-18(A/C) with B1013(A/C), DA11(C/T) with DA160(A/G), SP-A1 with SP-A2. Three-locus LD analysis suggests that B-18(A/C), B1013(A/C) and B1580(C/T) are strongly associated with each other.

Conclusions: Allele and genotype frequency differences imply that different ethnic groups should be mixed with extreme caution before performing linkage and association studies. The associated markers could be used together to increase the level of polymorphism and the informativeness of the "markers".

Background

Pulmonary surfactant is a lipoprotein complex essential for normal lung function. Deficiency of surfactant or derangement of surfactant activity, may lead to respiratory distress syndrome (RDS) or congenital alveolar proteinosis (CAP) [1–3]. The pulmonary surfactant proteins, SP-A, SP-B, and SP-D play important roles in surfactant function, structure and metabolism. Genetic variants of SP-A and SP-B have been found to associate with RDS and the SP-B locus has been linked to the pathogenesis of CAP [2,4–8]. Therefore, study of differences among surfactant protein variants may help explain individual variability in susceptibility to pulmonary disease, and the genetic variants of surfactant proteins may serve as valuable markers for disease gene mapping.

Linkage and association mapping are important tools for gene discovery. Both utilize available marker information to infer the location of disease susceptibility genes. They have been successfully used to map disease susceptibility genes such as Duchenne muscular dystrophy [9,10] and cystic fibrosis [11,12]. However, population stratification may confound linkage or association based mapping. The definition of population is arbitrary and is usually based on linguistic, cultural or geographic classification of the individuals. Population subdivision is influenced by complex interactions between social organization, dispersal tendencies and environmental factors. Pooling samples without regard to ethnicity where the allele frequencies are different in different ethnic groups may lead to problems relevant to population admixture. When the population is mixed, spurious association may be found between a disease phenotype and arbitrary markers that have no true linkage [13,14]. Population subdivisions that are not appropriately accounted for can lead to high false positive rates and may invalidate standard tests in association mapping [14]. The presence of admixture may also complicate traditional linkage analysis. The key assumptions of Hardy-Weinberg equilibrium (HWE) and linkage equilibrium (LE) invoked in linkage analysis may be violated by population admixture, and incorrect inference for linkage may result [15].

Several studies have focused on linkage and association mapping of pulmonary disease susceptibility variants using *SP-A*, *SP-B*, and *SP-D* as markers, but few have investigated the degree of genetic differentiation among the different ethnic groups and races before mixing them together. To avoid potential problems caused by population stratification that may confound linkage or association studies, we systematically characterized the *SP-A*, *SP-B*, and *SP-D* allele and genotype distributions in seven ethnic groups from three races: White American, Greek, German, Netherlands, Black American, Nigerian, and Mexican. The first four ethnic groups are Caucasian. The next two groups are Black. The last ethnic group, Mexican, is from a Hispanic population. We examined the level of genetic differentiation by comparing allele frequencies and genotype frequencies among different ethnic groups and races.

To show the level of informativeness of the SP-A, SP-B, and SP-D markers, we computed the heterozygosity and polymorphism information content (PIC) values for each of the eight loci we studied. In general, with more alleles at a marker locus, the marker is more informative. Most of the markers we studied have only two alleles. If alleles at nearby loci cosegregate as a unit, then the haplotypes could serve as a single "allele" and the total number of "alleles" will increase. Therefore, the informativeness of the "marker" will be improved. To increase the informativeness of the markers, we investigated linkage disequilibrium between loci on the same chromosome in an attempt to find valuable haplotypes for disease gene mapping.

Results

To conduct linkage and association analysis correctly, the study group should be homogeneous. The degree of genetic differentiation among the different ethnic groups and races were evaluated using Weir and Cockerham's [16] estimator (θ) of Wright's F_{ST}. Table 1 lists the estimated F_{ST} values among the ethnic groups within a single race as well as the estimated F_{ST} values among races. F_{ST} values were not estimated for the Hispanic population because there is only one ethnic group within the Hispanic population in our sample set. F_{ST} measures the amount of genetic variation in the whole population that is attributable to genetic differentiation among subpopulations. The larger the F_{ST} value, the more different the populations are. When the true F_{ST} value is close to zero, bias or sampling variation could lead to negative estimates. If the alleles from different populations are more related to each other than the alleles within the same population, the true F_{ST} value will be negative. In both cases, the populations are highly similar. Thus negative F_{ST}estimates suggest that the study groups are not different from one another (SP-A1, B1580(C/T) and DA11(C/T) in Black study group and DA11(C/T) in the whole study group). Combining the information from all available marker loci, relatively low levels of differentiation were detected among the four White ethnic groups (F_{ST} = 0.0036) and between the two black ethnic groups (F_{ST} = 0.0136), and a relatively high level of differentiation (F_{ST} = 0.0503) was found among the three races. F_{ST} analysis indicates that the differences among ethnic groups within a single race are smaller than among racial groups within the entire sample.

To find out whether the different ethnic groups or races have significantly different allele and/or genotype frequencies, we carried out a set of chi-square goodness of fit tests. Tables 2 lists the *p*-values of the χ^2 tests. Significant *p*-values (after sequential Bonferroni correction, $p' < 0.05$) are indicated with asterisks. In general, the allele tests and genotype tests gave us similar results, except for DA11(C/

Table 1: Genetic differentiation within and among different races

Locus	Among White F_{ST}	Between Black F_{ST}	Among Three Races F_{ST}
SP-A1	.0034	-.0029	.0323
SP-A2	.0004	.0251	.0611
B-18(A/C)	.0042	.0039	.0076
B1013(A/C)	.0063	.0086	.0617
B1580(C/T)	.0023	-.0105	.0394
B9306(A/G)	.0027	.1096	.0472
DA11(C/T)	.0056	-.0070	-.0024
DA160(A/G)	.0042	.0737	.1509
overall	.0036	.0136	.0503

Weir and Cockerham's [16] estimator of Wright's F_{ST} was calculated in computer program GDA, both within White and Black race and among White, Black, and Hispanic. For the among races calculation, different ethnic groups within the same race are combined together into one study group.

Table 2: p-values for tests of identical allele and genotype frequencies

Tests		*SP-A1*	*SP-A2*		*SP-B*			*SP-D*	
		SP-A1[a]	*SP-A2*[a]	B-18(A/C)	B1013(A/C)	B1580(C/T)	B9306(A/G)	DA11(C/T)	DA160(A/G)
Among	Genotype	-	-	.3451	.0303	.0997	.3542	**<.0001***	.1321
White	Allele	**.0082***	.1444	.1057	.0567	.2137	.1840	.1153	.1466
Among	Genotype	-	-	.2517	.1287	.7789	**<.0001***	.8690	**.0027***
Black	Allele	.0394	**<.0001***	.2412	.1722	.8111	**.0004***	.5983	**.0038***
All	Genotype	-	-	.0462	**<.0001***	**<.0001***	**<.0001***	.7751	**<.0001***
Races	Allele	**<.0001***	**<.0001***	.0218	**<.0001***	**<.0001***	**<.0001***	.8160	**<.0001***

(a) The rare alleles (alleles not listed in Table 5) are grouped together and treated as a single allele. Allele frequencies are compared within and among races. * After sequential Bonferroni correction, $p' < 0.05$.

T) in white study group. At this locus, we detected significant allele frequency difference among the four white ethnic groups ($p < 0.0001$) but no genotype frequency difference. The discrepancy of allele and genotype tests could be explained if we look at the heterozygosity values of this locus. The observed heterozygosity is much higher than the expected heterozygosity in Greeks, and the Greek study group deviates from HWE. Although the allele frequencies show no difference between the four ethnic groups, excess heterozygosity in the Greek group leads to the difference in genotype frequencies. After removing the Greek ethnic group, both p values from the genotype test and allele test become non-significant ($p = 0.3382$ for genotype test and $p = 0.1029$ for allele test). Table 2 shows that the allele and genotype frequencies differ significantly among different races and some even differ among different ethnic study groups within the same race. Although DA11(C/T) has been found to have different genotype frequencies among the four white ethnic groups, there was no evidence showing that the allele or genotype frequencies are different among races. This could be due to the effect of population admixture. The Greek sample has a small size (N = 71) and contributes relatively small effect to the combined group. We also performed pairwise comparison of allele and genotype frequencies among the seven ethnic groups (not shown). In general, ethnic groups in one race differ significantly from all the ethnic groups in another race.

To determine the degree of polymorphism of the eight marker loci and then infer their informativeness, we computed their heterozygosity and PIC values in each of the seven ethnic groups. We found that the observed (H_{Obs}) and expected (H_E) heterozygosity and PIC values varied in different ethnic groups though many confidence intervals of the observed heterozygosity overlapped. The two multi-allelic SP-A markers are highly polymorphic in all seven ethnic groups (H_{Obs} and PIC range from 0.48 to 0.81). The degree of polymorphism in general is also relatively high for the six biallelic SP-B and SP-D markers (H_{Obs} and PIC range from .32 to .80), except for a few cases, such as DA160(A/G) in Black American and Nigerian, where H_{Obs} are .21 and 0 respectively. The generally high degree of polymorphism implies that each marker locus is informative and could be used in disease gene mapping. The difference in heterozygosities in different ethnic groups suggests that the allele and genotype distributions are different among groups. We also calculated confidence intervals for the observed heterozygosities in the ethnic groups and compared the expected heterozygosities with the observed ones. In most cases, the expected heterozygosity fell within the confidence interval of the corresponding observed value. Only the expected heterozygosities of SP-A1 in the German and DA11(C/T) in the Greek study groups were found to fall outside the confidence intervals of the observed values, suggesting that these marker loci are not in Hardy-Weinberg equilibrium in these ethnic groups. Fisher's exact tests of departure from HWE verified that SP-A1 in the German ($p = 0.0031$) and DA11(C/T) in the Greek study groups ($p < 0.0001$) are not in HWE. In addition, exact tests indicated that SP-A2 in the white American ($p < 0.0001$) and in the German study groups ($p = 0.0009$) are indeed in Hardy-Weinberg disequilibrium.

All seven ethnic groups exhibit significant LD between SP-A1 and SP-A2 marker loci. When haplotype frequencies and LD were estimated using Arlequin software, we assumed that all the ethnic groups were in HWE. Since SP-A1 and SP-A2 in the German study group and SP-A2 in the White American study group deviate from HWE, the estimates of haplotype frequencies may be biased and significant associations between SP-A1 and SP-A2 could be also due to departure from HWE in these two study groups. Thus, the German and White American study groups were tested again for significant associations between SP-A1 and SP-A2 using GDA software. The within-locus Hardy-Weinberg disequilibrium can be prevented from affecting the significance of two loci disequilibrium by telling GDA to preserve the genotypes [17]. Significant associations were still found in these two study groups using GDA. Six (Black American, Mexican, German, Greek, Netherlands, and White American) out of the seven groups show significant LD between DA11(C/T) and DA160(A/G) loci. Again, since DA11(C/T) in the Greek study group deviates from HWE, association in this group was tested using GDA. Two groups (Mexican and German) have significant LD between DA160(A/G) and SP-A2 and four groups (Mexican, Netherlands, German, and White American) have significant LD between DA11(C/T) and SP-A2 (study groups not in HWE were tested using GDA). Thus the orientation of SP-D markers cannot be inferred by the LD analysis. LD analysis indicates that SP-A1 strongly associates with SP-A2 and DA11(C/T) strongly associates with DA160(A/G). Therefore, each of these two pairs of loci can be treated together as a single locus. Of the seven ethnic groups, we found six of them (Black American, Mexican, German, Greek, Netherlands, and White American) to demonstrate significant LD between marker loci B-18(A/C) and B1013(A/C), three (Mexican, German, and White American) display significant LD between B1013(A/C) and B1580(C/T), and three (Nigerian, Greek, and White American) exhibit significant LD between B1580(C/T) and B9306(A/G). These results imply that marker B-18(A/C) and B1013(A/C) are strongly associated with each other and the two markers together could form valuable haplotypes to be used as new marker alleles.

Since higher order linkage disequilibria is potentially more informative than pairwise linkage disequilibrium, we also conducted three-marker LD analysis. Three ethnic groups (Mexican, White American, German) display significant three-marker LD among DA11(C/T), DA160(A/G) and SP-A2. Four ethnic groups (Mexican, Greek, White American and German) demonstrate significant three-marker LD among DA160(A/G), SP-A2 and SP-A1. Four ethnic groups (Mexican, White American, German and Netherlands) demonstrate significant three-marker LD among DA11(C/T), SP-A2 and SP-A1. Six out of seven ethnic groups (White American, German, Greek, Mexican, Netherlands and Black American) exhibit significant three-marker LD among B-18(A/C), B1013(A/C) and B1580(C/T). Two ethnic groups (Greek and Mexican) show significant three-marker LD among B1013(A/C), B1580(C/T) and B9306(A/G). Therefore, three-marker LD analysis suggests that B-18(A/C), B1013(A/C) and B1580(C/T) are strongly associated together. Tables 3 and 4 enumerate the estimated two and three-locus haplotype frequencies and linkage disequilibrium coefficients for the associated marker loci in each of the seven ethnic groups. The estimated two-locus and three-locus haplotype frequencies and LD coefficients are different among the seven ethnic groups, which reflect differences among their evolutionary histories. LD coefficients measure the non-random associations between alleles at different loci, where associations are generated by some stochastic processes such as natural selection [18], mutation, sampling in a finite population, and certain forms of geographical structure. Recombination events during meiosis could break down the associations between alleles over time. Since the seven ethnic groups diverged a long time ago, they could have undergone different evolutionary events and these events could have led to differences in their LD. For example, the seven ethnic groups could have experienced different mutation events after their divergence which in turn may explain differences in their LD coefficients. Moreover, bottleneck generations during their evolution could lead to large sampling variation and this may as well result in differences in their haplotype frequencies and LD coefficients.

Discussion

Although family-based control methods such as the transmission disequilibrium test (TDT) [19] use the non-trans-

Table 3: Estimated two-locus haplotype frequencies and linkage disequilibrium coefficients in the seven ethnic groups

Loci	Haplotype	Black Am		Nigerian		Mexican		German		Greek		Netherlands		White Am	
		freq(a)	*LD*(b)	freq	LD	freq	LD	freq	LD	freq	LD	freq	LD	freq	LD
B-18(A/C)	AC	0.63	0.08	0.57	0	0.52	0.20	0.55	0.17	0.52	0.22	0.71	0.16	0.52	0.16
--	CA	0.13	0.08	0.06	0	0.40	0.20	0.32	0.17	0.42	0.22	0.23	0.16	0.32	0.16
B1013(A/C)	CC	0.24	-0.08	0.24	0	0.07	-0.20	0.08	-0.17	0.03	-0.22	0.04	-0.16	0.10	-0.16
	AA	0	-0.08	0.13	0	0.02	-0.20	0.05	-0.17	0.03	-0.22	0.02	-0.16	0.06	-0.16
DA11(C/T)	CG	0.40	0.04	0.38	0	0.39	0.18	0.38	0.12	0.34	0.05	0.27	0.14	0.40	0.16
--	TA	0.10	0.04	0	0	0.48	0.18	0.36	0.12	0.26	0.05	0.54	0.14	0.39	0.16
DA160(A/G)	TG	0.49	-0.04	0.62	0	0.12	-0.18	0.20	-0.12	0.27	-0.05	0.17	-0.14	0.2	-0.16
	CA	0	-0.04	0	0	0.01	-0.18	0.07	-0.12	0.13	-0.05	0.03	-0.14	4E-3	-0.16
SP-A2	$1A^{0}6A^{2}$	0.19	0.07	0.20	0.08	0.47	0.22	0.55	0.23	0.50	0.23	0.60	0.18	0.50	0.20
--	$1A^{1}6A^{3}$	0.11	0.02	0.11	-0.07	0.09	0.07	0.15	0.11	0.17	0.11	0.07	0.06	0.13	0.10
SP-*A1*(c)	$1A^{2}6A^{3}$	0.14	0.07	0.18	0.09	0.05	0.04	0.05	0.03	0.06	0.03	0.03	0.02	0.04	0.02
	$1A^{0}6A^{3}$	0.10	-0.01	0.08	-0.03	0.04	-0.06	0.03	-0.14	0.03	-0.13	0.04	-0.08	0.03	-0.10

(a) Estimated haplotype frequency (b) Estimated linkage disequilibrium coefficient (c) only common haplotypes are shown here.

Table 4: Estimated three-locus haplotype frequencies and three-locus LD coefficients at B-18(A/C)-B1013(A/C)-B1580(C/T) in the seven ethnic groups

haplotype	Black Am		Nigerian		Mexican		German		Greek		Netherlands		White Am	
	freq(a)	*LD*(b)	freq	LD	freq	LD	freq	LD	freq	LD	freq	LD	freq	LD
A A C	0	-0.011	0	-9.5E-4	0.007	0.007	0.012	0.008	0.011	0.011	0.019	0.038	0.017	-0.003
A A T	0	0.011	0.105	9.5E-4	0.014	-0.006	0.036	-0.008	0.018	-0.011	0	-0.038	0.038	0.003
A C C	0.198	0.011	0.231	9.5E-4	0.394	-0.006	0.273	-0.008	0.274	-0.011	0.465	-0.038	0.274	0.003
A C T	0.444	-0.011	0.325	-9.5E-4	0.124	0.006	0.274	0.008	0.248	0.011	0.249	0.038	0.249	-0.003
C A C	0.064	0.011	0	9.5E-4	0.159	-0.006	0.105	-0.008	0.145	-0.011	0.070	-0.038	0.132	0.003
C A T	0.063	-0.011	0.088	-9.5E-4	0.238	0.006	0.219	0.008	0.274	0.011	0.162	0.038	0.190	-0.003
C C C	0.074	-0.011	0.060	-9.5E-4	0.038	0.006	0.046	0.008	0.030	0.011	0.036	0.038	0.040	-0.003
C C T	0.157	0.011	0.191	9.4E-4	0.027	-0.006	0.035	-0.008	0	-0.011	0	-0.038	0.061	0.003

(a) Estimated three-locus haplotype frequency (b) Estimated three-locus linkage disequilibrium coefficient

mitted allele from the heterozygous parents as controls to avoid problems of population stratification, these studies are more difficult to conduct because they require cooperation with parents and availability of samples from the parents. Therefore, case control association tests are still being widely used to map disease susceptibility genes [20–22]. However, spurious association may result, if the study groups are not well matched in ethnicity. For example, a study of prostate cancer in African Americans [23] showed significant association between *CYP3A4-V* locus and prostate cancer in African Americans when the population stratification was uncorrected (p = 0.007). However, after correcting for population stratification, there was no longer significant association (p = 0.25).

SP-A, SP-B, and SP-D markers may play crucial roles in locating genes causing lung diseases. In this project, we systematically studied these markers in seven ethnic groups from three races. Both the F_{ST} analysis and χ^2 tests indicate a high level of population differentiation among races. Some markers demonstrate significant allele and genotype frequency differences even between ethnic groups within the same race. For many marker loci, there is no evidence to show significant difference between ethnic groups within a single race, although a definitive conclusion can not be reached at present. For example, a chi-square test did not detect significant allele frequency difference between the two black ethnic groups at B1013(A/C) locus (p = .1722). However, the power of the test is only 27.6% at 0.05 significance level (computed using SAS version 8). Therefore, it is possible that there is significant allele frequency difference between the two groups, but the test did not have enough power to detect it given the small sample sizes. Our findings indicate that individuals from different races cannot be joined together into one sample. However, individuals in different ethnic

Table 5: Markers and study groups analyzed

			White			Black		Hispanic
		German	Greek	Netherlands	White Am	Black Am	Nigerian	Mexican
SP-A	*SP-A1**	176	103	28	301	65	49	103
	*6A***	.034	.083	.071	.093	.031	.082	.267
	6A^{2}**	.568	.515	.661	.562	.423	.449	.490
	6A^{3}**	.293	.316	.179	.243	.392	.408	.199
	6A^{4}**	.065	.083	.089	.076	.085	.061	.034
	*SP-A2**	174	100	28	298	63	52	98
	*1A***	.0489	.110	.071	.102	.127	.077	.286
	1A^{0}**	.566	.525	.643	.530	.278	.269	.515
	1A^{1}**	.132	.165	.071	.143	.222	.433	.102
	1A^{2}**	.078	.085	.054	.076	.167	.221	.077
SP-B	*B-18(A/C)**	157	87	28	308	70	58	103
	*A***	.5955	.552	.732	.576	.629	.698	.539
	*B1013(A/C)**	157	87	28	307	68	56	103
	*A***	.3726	.448	.250	.378	.132	.196	.418
	*B1580(C/T)**	157	87	28	304	69	37	103
	*C***	.4363	.460	.589	.462	.341	.324	.597
	*B9306(A/G)**	156	87	28	308	45	58	103
	*A***	.9103	.954	.964	.916	.8	.957	.791
SP-D	*DA11(C/T)**	177	71	24	146	68	42	103
	*T***	.554	.528	.708	.596	.596	.560	.597
	*DA160(A/G)**	177	65	24	145	68	38	103
	*G***	.573	.615	.438	.603	.897	1.000	.515

* the numbers in the row are numbers of individuals in the samples analyzed ** the numbers in the row are estimated allele frequencies for the study groups

groups within a single race should be combined with extreme caution. Allele and genotype frequency tests should be performed to verify that there is no difference before combining them to avoid potential problems of population admixture.

We estimated the informativeness of each of the eight marker loci in the seven ethnic groups by assessing their heterozygosity and PIC values. In general, the markers are polymorphic. To increase the informativeness of the markers, especially for the biallelic marker, we carried out linkage disequilibrium analysis in an attempt to find valuable haplotypes. Our results revealed that three pair of loci are strongly associated together: B-18(A/C) with B1013(A/C), DA11(C/T) with DA160(A/G), and SP-A1 with SP-A2. In addition to two-marker LD analysis, we also performed three-marker LD analysis in an attempt to extend haplotype blocks. Extending haplotype blocks from two marker loci to three marker loci greatly increases the possible allele number (from four to eight for biallelic loci). Using haplotypes as a single allele will improve the informativeness of the marker. We reported the estimated haplotype frequencies and LD coefficients for the associated markers in each of the seven ethnic groups. Differences in estimated haplotype frequencies and LD imply that the seven ethnic groups have different evolutionary history. Although there has been no study on haplotype frequencies at these loci in these ethnic groups, some studies [24,25] have shown haplotype frequency differences in some of these ethnic groups at other loci. The published data are in general consistent with the present findings with regards to differences in evolutionary history among ethnic groups.

Conclusions

Differences in allele and genotype frequency suggest that caution should be taken when trying to combine individuals from different ethnic groups or races to avoid results with high false positive rates in linkage or association analysis. SP-A, SP-B, and SP-D markers are polymorphic in a variety of study groups and could serve as good markers for genetic studies, especially for the associated markers.

Methods

Samples

All samples and genotyping are from previously published studies [6–8,26,27] or ongoing studies in our laboratory, and all samples in the present study are from unrelated healthy controls. Individuals came from seven different ethnic groups (White American, Greek, German, Netherlands, Black American, Nigerian, and Mexican) within three racial categories (Caucasian, Black, and Hispanic). The subjects in the Mexican study group were nonsmokers. The smoking status in others was unknown. Genotypic data were available for eight marker loci within *SP-A*, *SP-B* and *SP-D* loci. Allele frequencies were derived from maximum likelihood estimation based on the observed genotypic data. *SP-A1* and *SP-A2* are two functional *SP-A* genes and more than 30 alleles have been characterized partially or entirely [28]. The *SP-A1* alleles are denoted as $6A^n$ and are classified based on five single nucleotide polymorphisms (SNPs) at codons for amino acids, AA19, AA50, AA62, AA133, and AA219. The first A stands for *SP-A* and the second A stands for amino acid. Number 19, 50, etc denotes the number of the SP-A amino acid sequence prior to the cleavage of the signal peptide. The *SP-A2* alleles are denoted as $1A^n$ and are classified based on four SNPs at codons for amino acids, AA9, AA91, AA140, and AA223. Therefore, the *SP-A1* and *SP-A2* alleles are haplotypes whereas the *SP-B* and *SP-D* alleles described below are individual SNPs. B-18(A/C), B1013(A/C), B1580(C/T) and B9306(A/G) are four marker loci located within the *SP-B* gene. The number following B refers to the nucleotide position of the polymorphism. The letters inside the parenthesis denote the alternative nucleotides at these positions. For *SP-D*, marker loci DA11 and DA160 have been characterized. These correspond to codons A(C/T)G and (A/G)CA, respectively. D stands for SP-D and A stands for amino acid. The numbers 11 and 160 denote the amino acid number of the SP-D amino acid sequence after the cleavage of the signal peptide (or of the first 20 amino acids). Therefore, amino acid 11 is part of the amino terminal sequence of SP-D and not part of the signal peptide. Thus, the six *SP-B* and *SP-D* SNP markers have two different alleles at each locus. Previous studies have demonstrated that the *SP-A* and *SP-D* markers are located on chromosome 10 in the order of DA11(C/T)/DA160(A/G), *SP-A2*, *SP-A1* (the orientation of the SP-D markers is not certain) and the SP-B markers are located on chromosome 2 in the order of B-18(A/C), B1013(A/C), B1580(C/T) and B9306(A/G) [29–32]. Table 5 shows the number of individuals studied and the estimated allele frequencies at each of the eight marker loci in each ethnic group. Allele frequencies for four common *SP-A1* alleles and four common *SP-A2* alleles are listed. For the six *SP-B* and *SP-D* SNP markers, allele frequencies for one allele are listed.

Statistical analysis

Each of the eight marker loci was tested for deviations from HWE within each ethnic group using Fisher's exact tests in the computer program genetic data analysis (GDA) [33]. The seven tests at each marker locus for the seven different ethnic groups were considered a family of tests. Significance was evaluated after applying the sequential Bonferroni correction for multiple testing [34].

To find out whether different ethnic groups could be mixed together, the degree of genetic differentiation among the different ethnic groups within a single race and among the three different races was quantified using Weir and Cockerham's [16] estimator (θ) of Wright's F_{ST}, as calculated by the GDA software. Ethnic groups within the same race were combined together into one study group when the within race F_{ST} was estimated. F_{ST} was calculated for each single locus as well as across all loci. Population genetic structure was also examined by testing the null hypothesis that the allele frequencies and genotype frequencies are identical across all ethnic groups and races using chi-square goodness of fit tests which is equivalent to Wright's F_{ST} tests in fixed populations [17]. The χ^2 test statistics were computed using SAS version 8. When the expected allele or genotype counts were less than 5, Fisher's exact test was performed through a Monte Carlo procedure, using computer software StatXact-4 version 4.0.1. At each of the eight marker loci, the allele frequencies and genotype frequencies were first compared among different ethnic groups within a single race, and then the different ethnic groups within the same race were combined together and the allele and genotype frequencies were compared among different races. For marker loci with multiple alleles (SP-A1 and SP-A2 loci), the rare alleles (alleles not listed in Table 5) were grouped together and treated as a single allele. Allele frequencies of the four common alleles and the rare allele group were compared within and among races. The null hypothesis of identical genotype frequencies was not tested because the total number of possible genotypes was too large. For each marker locus, significance levels were adjusted using a sequential Bonferroni correction for multiple comparisons [34]. All the tests for a single locus were considered as a family of tests. To study genetic differentiation of ethnic groups across the races, allele and genotype frequencies were also compared between pairs of ethnic groups in the same race as well as in different races. Similarly, significance was evaluated after applying sequential Bonferroni correction at each marker locus.

To explore the level of polymorphism and then infer the informativeness of each of the eight markers, the observed heterozygosity value $(H_0 = \sum_{i=1}^{m-1}\sum_{j=i+1}^{m}\frac{n_{ij}}{N})$, expected heterozygosity $(H_E = 2\sum_{i=1}^{m-1}\sum_{j=i+1}^{m}p_ip_j)$ and 95% asymptotic confidence intervals circumscribing the observed heterozygosity were computed for each ethnic group using SAS version 8. Another often used measure of polymorphism, the Polymorphism Information Content (PIC), was also calculated based on the estimated allele frequencies. PIC value is defined as the probability of inferring with certainty which parental allele is passed to the offspring $(PIC = 2\sum_{i=1}^{m-1}\sum_{j=i+1}^{m} p_i p_j (1 - p_i p_j))$. The PIC value of each of the marker loci was also computed for all the seven ethnic groups using SAS version 8.

Finally, linkage disequilibrium (LD) and haplotype frequencies were explored using the computer program Arlequin version 2.000 [35]. Maximum-likelihood haplotype frequencies were estimated using an Expectation-Maximization (EM) algorithm [17,36]. LD coefficients were computed based on the estimated haplotype frequencies and allele frequencies. Pairwise linkage disequilibrium was tested using likelihood-ratio tests [37]. When pairwise associations were tested using Arlequin, the loci were assumed to be in HWE. If either one of the loci is not in HWE, significant LD could be also due to departure from HWE. To avoid this problem, associations between loci were tested using computer program GDA at loci not in HWE. Genotypes were preserved in GDA to prevent the within-locus Hardy-Weinberg disequilibrium from affecting the significance of disequilibrium in two-locus LD [17]. For loci not in HWE, LD coefficients were estimated without assuming HWE using GDA and haplotype frequencies were calculated based on the estimated allele frequencies and LD coefficients. Three-marker LD significance tests were performed in GDA. Similarly, to prevent the within-locus Hardy-Weinberg disequilibrium from affecting the significance of disequilibrium in three-locus LD, genotypes were preserved when the loci involved in the haplotypes were not in HWE. When estimating the two or three-marker haplotype frequencies and LD coefficients, several individuals were not genotyped at all the loci involved in the haplotypes and were eliminated from the analysis. Three-locus linkage disequilibria coefficients were derived based on the estimated allele frequencies and two-locus LD coefficients using the formula presented by Weir [17].

Authors' contributions

Author WL performed the statistical analysis and drafted the manuscript. Author CMB participated in performing the χ^2 tests and computing the heterozygosities. JF conceived of the study, and participated in its design and coordination. All authors read and approved the final manuscript.

Acknowledgements

The authors thank Drs Bruce S. Weir, Hung-Mo Lin, and Gary Chase, and Susan DiAngelo for their helpful suggestions and Drs Marietta Xanthou and Costas Papagaroufalis for samples and Jelena Pavlovic for genotyping. This work was supported by NIH Grant R37 HL34788 and the General Clinical Research Center, PSU, College of Medicine.

References

1. Avery ME and Mead J: **Surface properties in relation to atelectasis and hyaline membrane disease.** *Am J Dis Child* 1959, **97:**517-523.
2. Floros J and Kala P: **Surfactant proteins: Molecular genetics of neonatal pulmonary diseases.** *Annu Rev Physiol* 1998, **60:**365-384.
3. deMello DE and Lin Z: **Pulmonary alveolar proteinosis: a review.** *Pediatr Pathol Mol Med* 2001, **20:**413-432.
4. deMello DE, Nogee LM, Heyman S, Krous HF, Hussain M, Merritt TA, Hsueh W, Haas JE, Heidelberger K, Schumacher R and Coten HR: **Molecular and phenotypic variability in the congenital alveolar proteinosis syndrome associated with inherited surfactant protein B deficiency.** *J Pediatr* 1994, **125:**43-50.
5. Floros J, Veletza SV, Kotikalapudi P, Krizkova L, Karinch AM, Friedman C, Buchter S and Marks K: **Dinucleotide repeats in the human surfactant protein B gene and respiratory distress syndrome.** *Biochem J* 1995, **305:**583-590.
6. Kala P, Ten Have T, Nielsen H, Dunn M and Floros J: **Association of pulmonary surfactant protein A (SP-A) gene and respiratory distress syndrome: interaction with SP-B.** *Pediatr Res* 1998, **43:**169-177.
7. Floros J, Fan R, Matthews A, DiAngelo S, Luo J, Nielsen H, Dunn M, Gewolb IH, Koppe J, van Sonderen L, Farri-Kostopoulos L, Tzaki M, Ramet M and Merrill J: **Family-based transmission disequilibrium test (TDT) and case-control association studies reveal surfactant protein A (SP-A) susceptibility alleles for respiratory distress syndrome (RDS) and possible race differences.** *Clin Genet* 2001, **60:**178-187.
8. Floros J, Fan R, Diangelo S, Guo X, Wert J and Luo J: **Surfactant protein (SP) B associations and interactions with SP-A in white and black subjects with respiratory distress syndrome.** *Pediatr Int* 2001, **43:**567-576.
9. Murray JM, Davies KE, Harper PS, Meredith L, Mueller CR and Williamson R: **Linkage relationship of a cloned DNA sequence on the short arm of the X chromosome to Duchenne muscular dystrophy.** *Nature* 1982, **300:**69-71.
10. Monaco AP and Kunkel LM: **Cloning of the Duchenne/Becker muscular dystrophy locus.** *Adv Hum Genet* 1988, **17:**61-98.
11. Kerem B, Rommens JM, Buchana JA, Markiewicz D, Cox TK, Chakravarti A, Buchwald M and Tsui L-C: **Identification of the cystic fibrosis gene: genetic analysis.** *Science* 1989, **245:**1073-1080.
12. Dean M, Drumm ML, Stewart C, Gerrard B, Perry A, Hidaka N, Cole JL, Collins FS and Iannuzzi MC: **Approaches to localizing disease genes as applied to cystic fibrosis.** *Nucleic Acids Res* 1990, **18:**345-350.
13. Lander ES and Schork NJ: **Genetic dissection of complex traits.** *Science* 1994, **265:**2037-2048.
14. Ewens WJ and Spielman RS: **The transmission/disequilibrium test: history, subdivision, and admixture.** *Am J Hum Genet* 1995, **57:**455-464.
15. Barnholtz-Sloan JS, de Andrade M, Dyer TD and Chakraborty R: **Admixture effects in the traditional linkage analysis of admixed families.** *Ethn Dis* 2002, **12:**411-419.
16. Weir BS and Cockerham CC: **Estimating F-statistics for the analysis of population structure.** *Evolution* 1984, **38:**1358-1370.
17. Weir BS: *Genetic Data Analysis II Sinauer Associates, Inc. Sunderland, MA, USA*; 1996:119-201.
18. Strobeck C: **Expected linkage disequilibrium for a neutral locus linked to a chromosomal arrangement.** *Genetics* 1983, **103:**545-555.
19. Spielman RS, McGinnis RE and Ewens WJ: **Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM).** *Am J Hum Genet* 1993, **52:**506-516.
20. Liphaus Bde L, Goldberg AC, Kiss MH and Silva CA: **Analysis of human leukocyte antigens class II-DR in Brazilian children and adolescents with systemic lupus erythematosus.** *Rev Hosp Clin Fac Med Sao Paulo* 2002, **57:**277-282.
21. Tomer Y, Concepcion E and Greenberg DA: **A C/T Single-Nucleotide Polymorphism in the Region of the CD40 Gene is Associated with Graves' Disease.** *Thyroid* 2002, **12:**1129-1135.
22. Talmud PJ, Palmen J, Nicaud V, Tiret L and European Atherosclerosis Research II Study: **Association of the hormone sensitive lipase -60C > G variant with fasting insulin levels in healthy young men.** *Nutr Metab Cardiovasc Dis* 2002, **12:**173-177.
23. Kittles RA, Chen W, Panguluri RK, Ahaghotu C, Jackson A, Adebamowo CA, Griffin R, Williams T, Ukoli F, Adams-Campbell L, Kwagyan J, Isaacs W, Freeman V and Dunston GM: **CYP3A4-V and prostate cancer in African Americans: causal or confounding**

association because of population stratification? *Hum Genet* 2002, **110:**553-560.

24. Jurevic RJ, Chrisman P, Mancl L, Livingston R and Dale BA: **Single-nucleotide polymorphisms and haplotype analysis in beta-defensin genes in different ethnic populations.** *Genet Test* 2002, **6:**261-269.
25. Moraes MO, Santos AR, Schonkeren JJ, Vanderborght PR, Ottenhoff TH, Moraes ME, Moraes JR, Sampaio EP, Sarno EN and Huizinga TW: **Interleukin-10 promoter haplotypes are differently distributed in the Brazilian versus the Dutch population.** *Immunogenetics* 2003, **54:**896-899.
26. Krizkova L, Sakthivel R, Olowe SA, Rogan PK and Floros J: **Human SP-A: genotype and single-strand conformation polymorphism analysis.** *Am J Physiol* 1994, **266:**L519-527.
27. Guo XX, Lin HM, Lin Z, Montano M, Sansores R, Wang G, DiAngelo S, Pardo A, Selman M and Floros J: **Surfactant protein genes A, B, and D marker alleles in COPD of a Mexican population.** *Eur Respir J* 2001, **18:**482-490.
28. DiAngelo S, Lin Z, Wang G, Phillips S, Ramet M, Luo J and Floros J: **Novel, non-radioactive, simple and multiplex PCR-cRFLP methods for genotyping human SP-A and SP-D marker alleles.** *Dis Markers* 1999, **15:**269-281.
29. Bruns G, Stroh H, Veldman GM, Latt SA and Floros J: **The 35kd pulmonary surfactant associated protein is encoded on chromosome 10.** *Human Genetics* 1987, **76:**58-62.
30. Glasser SW, Korfhagen TR, Weaver T, Pilot-Matias T, Fox JL and Whitsett JA: **cDNA and deduced amino acid sequence of human pulmonary surfactant-associated proteolipid SPL(Phe).** *Proc Natl Acad Sci U S A* 1987, **84:**4007-4011.
31. Vamvakopoulos NC, Modi WS and Floros J: **Mapping the human pulmonary surfactant-associated protein B gene (SFTP3) to chromosome 2p12->p11.2.** *Cytogenet Cell Genet* 1995, **68:**8-10.
32. Hoover RR and Floros J: **Organization of the human SP-A and SP-D loci at 10q22-q23. Physical and radiation hybrid mapping reveal gene order and orientation.** *Am J Respir Cell Mol Biol* 1998, **18:**353-362.
33. Lewis PO and Zaykin D: **Genetic Data Analysis: Computer program for the analysis of allelic data. Version 1.0 (d16c).** *Free program distributed by the authors over the internet from* 2001 [http://lewis.eeb.uconn.edu/lewishome/software.html].
34. Rice WR: **Analyzing tables of statistical tests.** *Evolution* 1989, **43:**223-225.
35. Schneider S, Roessli D and Excoffier L: **Arlequin ver 2.000: a software for population genetic data analysis.** *Geneva, Switzerland: Genetics and Biometry Laboratory, University of Geneva*; 2000.
36. Excoffier L and Slatkin M: **Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population.** *Mol Biol Evol* 1995, **12:**921-927.
37. Slatkin M and Excoffier L: **Testing for linkage disequilibrium in genotypic data using the EM algorithm.** *Heredity* 1996, **76:**377-383.