Testing for linkage disequilibrium using SAS/GENETICS and SAS/STAT

Amina Barhdadi and Marie-Pierre Dubé
Statistical Genetics Research Group at the Montreal Heart Institute, Montreal, QC, Canada
October 2011

ABSTRACT

Testing for the presence of linkage disequilibrium (LD) and measuring its value is important in statistical genetics. LD deals with the correlation of genetic variation at two or more loci in the genome within a given population. PROC ALLELE in SAS/GENETICS provides a variety of pair-wise LD measures that are related to the well-known Pearson correlation r. Different statistical tests of linkage disequilibrium are performed using PROC ALLELE. PROC HAPLOTYPE offers LD test statistics for multiple loci.

In this paper, we clarify differences between LD measures obtained using PROC ALLELE and show how the HAPLO=OPTION of this procedure interacts with the linkage disequilibrium calculations and tests. Moreover, we compare PROC CORR and PROC ALLELE in terms of correlation coefficients of genotypic data.

INTRODUCTION

Linkage disequilibrium, the non random association of alleles from different loci, is often the basis for evaluating the association of genomic variation with human traits among unrelated subjects. LD plays a central role in mapping genes relevant for specific traits of interest mainly in humans. In this approach, genetic variation at a set of marker loci in a sample of individuals is tested for a given phenotype. If such an association is found between a particular marker locus and the phenotype, it suggests that either the variation at that marker locus affects the phenotype of interest, or that the variation of that marker locus is in LD with the true phenotype-related locus, which was not genotyped.

Linkage disequilibrium calculations and tests are based on haplotype frequencies estimation. A haplotype is a combination of alleles at multiple loci on a single chromosome. If one locus has alleles A_1 and A_2, and a second locus has alleles B_1 and B_2, the observed genotype $A_1A_1B_1B_1$ must contain two haplotypes of type A_1B_1; genotype $A_1A_2B_1B_1$ must contain haplotypes A_1B_1 and A_2B_1 and so on. In some situation, there is ambiguity and it is not possible to know which two haplotypes constitute any individual genotype. For example the genotype $A_1A_2B_1B_2$ may have haplotypes A_1B_1 and A_2B_2 or haplotypes A_1B_2 and A_2B_1.

There are a variety of LD measures available. The most commonly used one is the linkage disequilibrium coefficient D, for which the notations are shown in Table 1 and values in equation (1). In order to compare LD quantities among different pairs of loci with differing allele frequencies, several standardization methods have been proposed (see [1] and [2]). One way of standardization is provided by dividing the coefficient D by its maximum value given the allele frequencies to obtain D' coefficient described in equation (2).

Table1: Association between two alleles at each of two loci. The marginal frequencies represent the allele's frequencies.

		Alleles at locus A		
		A_1	A_2	
Alleles at locus B	B_1	$p_{A_1B_1}$	$p_{A_2B_1}$	p_{B_1}
	B_2	$p_{A_1B_2}$	$p_{A_2B_2}$	p_{B_2}
		P_{A_1}	P_{A_2}	

$$D = p_{A_1B_1} - p_{A1}p_{B_1} = p_{A_2B_2} - p_{A_2}p_{B_2} = -(p_{A_1B_2} - p_{A_1}p_{B_2}) = -(p_{A_2B_1} - p_{A_2}p_{B_1}) \qquad (1)$$

$$D' = \frac{D}{D_{max}},\ D_{max} = \begin{cases} \min(p_{A_1}p_{B_2}, p_{A_2}p_{B_1}) & \text{if } D > 0 \\ \max(-p_{A_1}p_{B_1}, -p_{A_2}p_{B_2}) & \text{if } D < 0 \end{cases} \qquad (2)$$

Linkage disequilibrium coefficient D is related to the well-known Pearson correlation coefficient r as follows

$$r = \frac{D}{\sqrt{p_{A_1}p_{A_2}p_{B_1}p_{B_2}}} \qquad (3)$$

The squared coefficient of determination $r^2$is often used to remove the arbitrary sign introduced, when the marker alleles are arbitrary labeled.

PROC ALLELE and PROC HAPLOTYPE in SAS/GENETICS provide an effective tool for calculating LD coefficients and testing allelic association respectively. This paper illustrate the use of HAPLO option of the PROC ALLELE and LD option of the PROC HAPLOTYPE to to test for linkage disequilibrium and to calculate its value. A comparison between PROC ALLELE , PROC HAPLOTYPE and PROC CORR will be made.

METHODS

Significance testing for LD coefficient D follows testing for independence in a 2x2 table as shown in Table1. The usual methods for this type of test are chi-square test, likelihood ratio test and Fisher exact test. The significance of observed values of any statistics can alternatively be obtained by permuting the alleles of one of the loci with respect to the other locus alleles, keeping the allele frequencies constant. In this case, the p-value for the statistic is the proportion of permutations, which result in equal to or more extreme values of the statistic.

1. Testing for linkage disequilibrium

1.1. LD between two loci

1.1.1. PROC ALLELE

PROC ALLELE in SAS/GENETICS offers five different measures of linkage disequilibrium, namely the linkage coefficient D, the correlation coefficient r, the population attributable risk δ, Lewontin's D', the proportional difference d, and Yule's Q. These measures are calculated only for pairs of markers at most *d* markers apart, where *d* is the integer specified in MAXDIST=option of the PROC ALLELE statement.

Linkage disequilibrium coefficients computation interacts with the Haplo option which affects all linkage disequilibrium tests and measures. This option indicates whether haplotype frequencies should not be used, haplotype frequencies should be estimated, or observed haplotype frequencies in the data should be used.

The **HAPLO=GIVEN** option indicates that the haplotypes have been observed, and the haplotype frequencies are used in the LD test statistics and measures. When **HAPLO=EST** is specified, the maximum likelihood estimates of the haplotype frequencies are used to compute LD statistics and measures. The LD coefficient D is estimated as $\tilde{D}_{uv} = \hat{p}_{uv} - \tilde{p}_u\tilde{p}_v$ where the estimate $\hat{p}_{uv}$ is calculated according to the method described by Weir and Cockerham in [3]. By default or when **HAPLO=NONE** is specified, the composite linkage disequilibrium (CLD) coefficient Δ is used instead of the LD coefficient D. Moreover, the composite haplotype frequencies are used to form the linkage disequilibrium measures and tests. The maximum likelihood estimate $\tilde{\Delta}$ of Δ can be calculated as described by Weir in [4].

For each option of HAPLO statement, PROC ALLELE calculates an overall chi-square test that linkage disequilibrium D between two markers is zero. Table2 shows how the LD coefficients are estimated and the related statistics are computed.

Table2: Interaction between Haplo=option of proc allele and LD computations and tests

HAPLO=OPTION	Estimate of Haplotype frequencies	LD coefficient	LD test statistic
GIVEN	Observed freq, $\tilde{p}_{uv}, \tilde{p}_u$ and $\tilde{p}_v$	$\tilde{D}_{uv} = \tilde{p}_{uv} - \tilde{p}_u\tilde{p}_v$; $u = A_1, A_2$; $v = B_1, B_2$	$\chi^2_T = \sum_{u=1}^{2}\sum_{v=1}^{2} \frac{(2n)\tilde{D}^2_{uv}}{\tilde{p}_u\tilde{p}_v}$
EST	Estimated freq, $\hat{p}_{uv}, \hat{p}_u$ and $\hat{p}_v$	$\tilde{D}_{uv} = \hat{p}_{uv} - \tilde{p}_u\tilde{p}_v$; $u = A_1, A_2$; $v = B_1, B_2$	$\chi^2_T = \sum_{u=1}^{2}\sum_{v=1}^{2} \frac{n\hat{D}^2_{uv}}{\tilde{p}_u\tilde{p}_v}$
NONE	Composite freq, $\tilde{p}^{\circ}_{uv}, \tilde{p}^{\circ}_u$ and $\tilde{p}^{\circ}_v$	$\tilde{\Delta}_{uv} = \left(\frac{n_{uv}}{n}\right) - 2\tilde{p}^{\circ}_u\tilde{p}^{\circ}_v$; $u = A_1$; $v = B_1$	$\chi^2_T = \frac{\tilde{\Delta}^2_{A_1B_1}}{\tilde{p}_{A_1}\tilde{p}_{B_1}}$

$n_{A_1B_1} = 2X_{A_1A_1B_1B_1} + X_{A_1A_1B_1B_2} + X_{A_1A_2B_1B_1} + \frac{1}{2}X_{A_1A_2B_1B_2}$. X is a count of the number of subjects with the phenotype indicated by its subscript.

1.1.2. PROC CORR

The correlation r between alleles at a pair of genetic loci is a measure of linkage disequilibrium and is described in equation (3). Strictly speaking, the correlation is for the indicator variables *X* and *Y* defined as

$$X = \begin{cases} 1, \text{Allele1 is } A_1 \\ 0, \text{Allele 1 is } A_2 \end{cases}; \quad Y = \begin{cases} 1, \text{Allele 2 is } B_1 \\ 0, \text{Allele 2 is } B_2 \end{cases}.$$

This correlation coefficient has drawn much attention during recent years especially in testing association between markers and genetic diseases.

PROC CORR in SAS/STAT computes Pearson correlation coefficient by default, but also offers Spearman's correlation and other measures of association, such as Cronbach's alpha and Kendall's tau. Probabilities associated with these statistics are also provided by PROC CORR.

Test for the null hypothesis $H_0 : \rho = 0$ based on Pearson's correlation *r* has test statistic

$t = \sqrt{(n-2)\frac{r^2}{1-r^2}}$, which has a t-distribution with *n*-2 degrees of freedom. Same test is used if *r* is a Spearman correlation.

1.2. LD between multiple loci

1.2.1. PROC HAPLOTYPE

One application of PROC HAPLOTYPE is determining whether there is linkage disequilibrium (LD), or association between loci. PROC HAPLOTYPE performs a likelihood ratio test to test the hypothesis of no LD between marker loci.

When the **LD** option is specified in the PROC HAPLOTYPE statement, haplotype frequencies are calculated using the expectation maximization (EM) algorithm. Under the null hypothesis of no LD, haplotype frequencies are simply the product the product of the individual allele frequencies. The log likelihood under the null hypothesis, $logL_0$, is calculated based on these haplotype frequencies with degrees of freedom $df_0 = \sum_{i=1}^{m}(k_i - 1)$, where *m* is the number of loci and $k_i$is the number of alleles for the i^{th} locus (see [5]). Under the alternative hypothesis, the log likelihood $logL_1$ is calculated from the EM estimates of the haplotypes frequencies with degree of freedom $df_1 = \text{number of haplotypes} - 1$. A likelihood ratio test is used to test this hypothesis as follows

$$2(logL_1 - logL_0) \sim \chi^2_v$$

where $\lambda = df_1 - df_0$ is the difference between the number of degrees of freedom under the null hypothesis and the alternative.

2. APPLICATION

We have simulated genotypes of 500 individuals at 100 single nucleotide polymorphisms (SNPs) with linkage disequilibrium. These data was used to compare LD coefficients and test statistics obtained with each of PROC ALLELE, PROC HAPLOTYPE and PROC CORR.

2.1. SAS CODES

2.1.1. PROC ALLELE with HAPLO option

/*Data set one, in tall format, contains genotype data of 500 individuals at 100 SNPs. Variable indiv is the individual id, variable locusname indicates the name of each marker and variables a1 and a2 are the two alleles of each genotype. Maxdist=100 specifies the maximum number of marker s for which the linkage disquilibrium calculations should be performed. Outstat option is used to name the output data set containing the disequilibrium statistics.*/

```
ods output ldmeasures=ldmeas_none;
ods output allelefreq=allelefreq_none;
proc allele data=one tall indiv=indiv marker=locusname haplo=none
corrcoeff dprime yulesq outstat=ld_none maxdist=100 ;
var a1 a2;
run;
ods output close;
ods output close;

ods output ldmeasures=ldmeas_given;
ods output allelefreq=allelefreq_given;
proc allele data=one tall indiv=indiv marker=locusname haplo=given
outstat=ld_given corrcoeff dprime yulesq maxdist=100;
var a1 a2;
run;
ods output close;
ods output close;

ods output ldmeasures=ldmeas_est;
ods output allelefreq=allelefreq_est;
proc allele data=one tall indiv=indiv marker=locusname haplo=est
corrcoeff dprime yulesq outstat=ld_est maxdist=100;
var a1 a2;
run;
ods output close;
ods output close;
```

2.1.2. PROC HAPLOTYPE with LD option

By default, all possible haplotypes from the sample are considered in PROC HAPLOTYPE. We provide here a SAS macro to compute haplotypes from a pair of markers.

```
%macro haplo;
%let N=100;/* N total number of markers*/
%do i=1 %to %eval(&N);
    dm output 'clear' continue; /*to clean the output window*/
%do j=%eval(&i)+1 %to %eval(&N);
data two(where=(markernum in (%eval(&i), %eval(&j))));/*markernum
indicates the number of each marker*/
set one; /*In this data set markers are numbered*/
run;
data _null_;
call symput ('first','SNP%eval(&i)');
call symput ('second', 'SNP%eval(&j)');
run;

ods output ldtest=ld;
ods output haplotypefreq=freq_haplo;
proc haplotype data=two(where=(locusname in ("&first", "&second")))
tall indiv=indiv marker=locusname ld;
var a1 a2;
run;
ods output close;
ods output close;

data freq_haplo;set freq_haplo;
combination=compress("&first"||"&second");
run;

data ld;set ld;
combination=compress("&first"||"&second");
run;

proc append base=ld_haplotype data=ld;
run;
proc append base=freq_haplotype data=freq_haplo;
run;

%end;
%end;
%mend haplo;
%haplo;
```

2.1.3. PROC CORR

As was mentioned earlier, data one is in tall format, we transposed it to use PROC CORR as follows

```
proc sort data=one out=two;by indiv; run;

proc transpose data=two out=two; var genotype; by indiv; id locusname;
run;

ods output pearsoncorr=corr_pearson;
proc corr data=two pearson outp=statistic_pearson;
var SNP1-SNP100; /*SNP1-SNP100 are values of the variable locusname in
the dataset one*/
run;
ods output close;
```

To make the output data set from PROC CORR dataset (here corr_pearson) in tall format we provide the following SAS macro

```
%macro corr_tall;
%do i=1 %to 100;
data _null_;
set corr_pearson;
call symput ('second','SNP%eval(&i)');
run;
data snp(keep=variable snp%eval(&i) psnp%eval(&i) combination
rename=(snp%eval(&i)=corr_pearson psnp%eval(&i)=pval_pearson));
set corr_pearson;
combination=compress(variable||"&second");
run;
proc append base=coeff_pearson data=snp;
run;
%end;
%mend corr_tall;
%corr_tall;
```

Two data sets were created. The first one, called LDmeasures, is a merge of resulting datasets of LD measures from PROC ALLELE and PROC CORR. The second one, called LDstatistics is a merge of resulting datasets of LD statistics from PROC CORR, PROC ALLELE and PROC HAPLOTYPE. We applied PROC CORR to these two created datasets as following

```
ods html;
ods graphics on;
proc corr data=ldmeasures plots=matrix;
Var  corrcoeff_none corrcoeff_given corrcoeff_est corr_pearson;
run;
ods graphics off;
ods html close;

ods html;
ods graphics on;
ods output pearsoncorr=pearsoncorr;
proc corr data=ldstatistics plots=matrix Pearson;
var  chisq_none chisq_given chisq_est t_square chisq_haplo run;
```

```
ods output close;
ods graphics off;
ods html close;
```

2.2. RESULTS

Table 3 displays correlation between LD coefficient R obtained by PROC CORR and PROC ALLELE using HAPLO option. This table shows a high correlation between correlation coefficients obtained using PROC ALLELE with HAPLO option none and est and Pearson correlation coefficient obtained using PROC CORR. Table 4 summarizes correlation between test statistics of LD obtained using each of the ALLELE ,CORR and HAPLOTYPE procedure. Table 4 shows a high correlation between the three following statistics, Chisq-est which is the chi square test used by PROC ALLELE with Haplo option est, T-square the square of Student statistic used by PROC CORR and chisq-haplo which is the likelihood ratio test of LD used by PROC HAPLO with LD option.

Table 3: Correlation between LD measures obtained using PROC ALLELE and PROC CORR.

Pearson Correlation Coefficients, N = 4950, Prob > |r| under H0: Rho=0				
	Coeff-none	**Coeff-given**	**Coeff-est**	**Coeff-Pearson**
Coeff-none	1.00000	0.72747, <0.0001	0.99630, <.0001	0.99956, <0.0001
Coeff-given	0.72747 <.0001	1.00000	0.72615 <.0001	0.72769 <.0001
Coeff-est	0.99630, <.0001	0.72615, <0.0001	1.00000	0.99751, <0.0001
Coeff-Pearson	0.99956, <0.0001	0.72769, <0.0001	0.99751, <0.0001	1.00000

Table 4: Correlation between LD statistic tests using PROC ALLELE, PROC HAPLOTYPE and PROC CORR.

Pearson Correlation Coefficients, N = 4950, Prob > |r| under H0: Rho=0					
	Chisq-none	**Chisq-given**	**Chisq-est**	**T-square**	**Chisq-haplo**
Chisq-none	1.00000	0.18368, <0001	0.98007, <0.0001	0.99622, <0.0001	0.99374 <.0001
Chisq-given	0.18368, <0.0001	1.00000	0.21187, <0.0001	0.19236 <.0001	0.18509 <.0001
Chisq-est	0.98007, <0.0001	0.21187, <0.0001	1.00000	0.98653, <0.0001	0.99462, <0.0001
T-square	0.99622, <0.0001	0.19236, <0.0001	0.98653, <0.0001	1.00000	0.99613, <0.0001

Chisq-haplo	0.99374, <0.0001	0.18509, <0.0001	0.99462, <0.0001	0.99613, <0.0001	1.00000

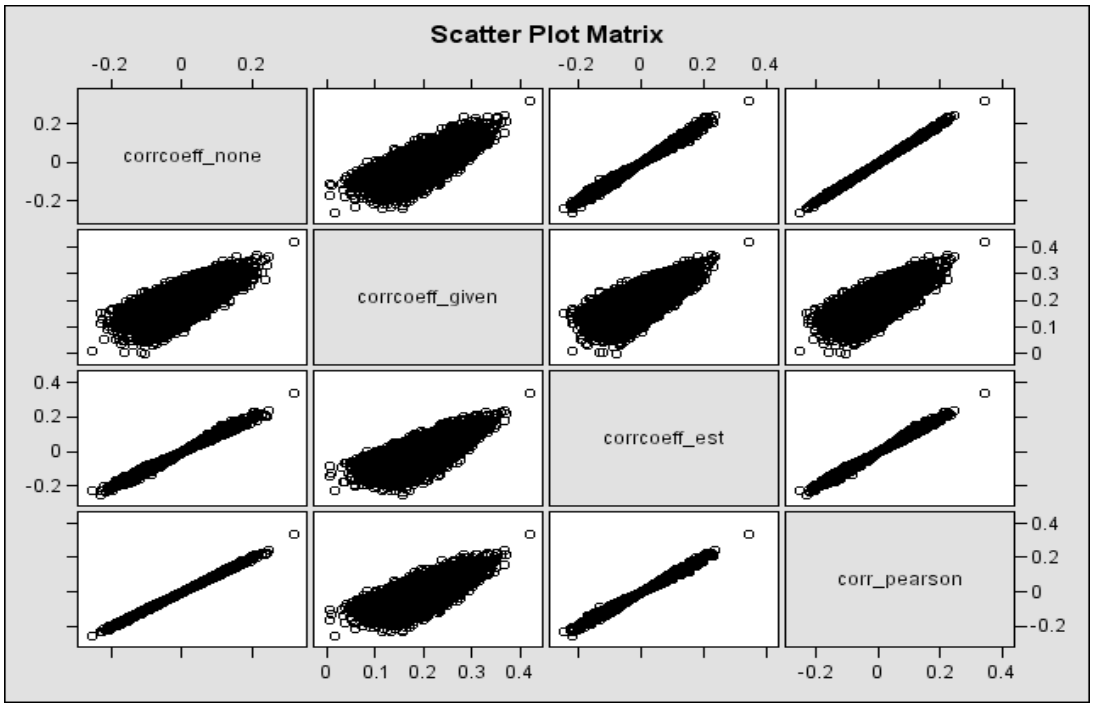

Figure 1: Scatter plot matrix of measures obtained using PROC CORR and PROC ALLELE with the three different Haplo options. Corr-Pearson refers to the Pearson correlation coefficient obtained using PROC CORR. Corrcoeff-none, Corrcoeff-est and Corrcoeff-given refer to the correlation coefficient obtained using PROC ALLELE with Haplo=none, Haplo=est and Haplo=given respectively.

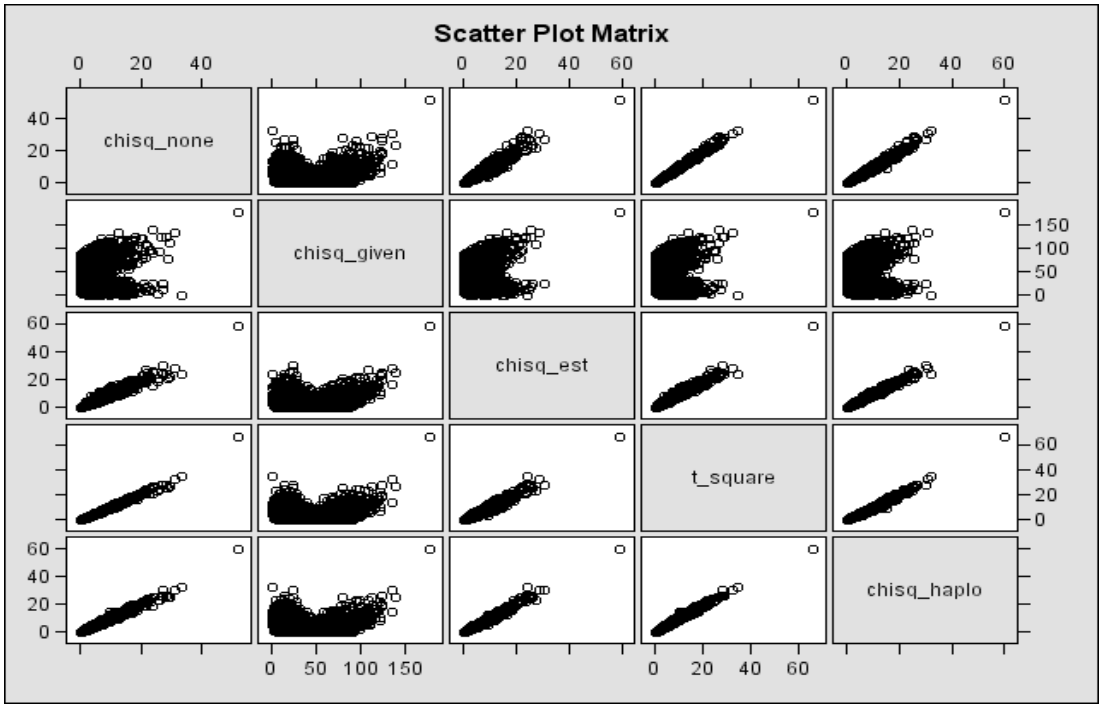

Figure 2: Scatter plot matrix of LD test statistics obtained using PROC CORR, PROC ALLELE and PROC HAPLOTYPE. T-square refers to the square of Student statistic used by PROC CORR to test for LD, chisq-none, chisq-est and chisq-given refer to the chi-square statistic using PROC ALLELE with Haplo=none, Haplo=est and Haplo=given respectively. Chisq-haplo stands for the chi-square statistic using PROC HAPLO with LD option.

3. CONCLUSION

Testing for the presence of linkage disequilibrium and measuring its value are two important tools of statistical genetics that have recently received much more attention. SAS/GENETICS provides two procedures to test for LD, PROC ALLELE provides different LD measures and LD test statistics between two loci. PROC HAPLOTYPE provide allele association test between multiple loci when the LD option is specified. In this paper we compared results of these two procedures with results obtained with SAS/STAT CORR procedure. Results of this comparison will be useful to researchers engaged in genetic studies.

4. REFERENCES

1. Lewontin, R.C., *On measures of gametic disequilibrium.* Genetics, 1988. **120**(3): p. 849-52.
2. Devlin, B. and N. Risch, *A comparison of linkage disequilibrium measures for fine-scale mapping.* Genomics, 1995. **29**(2): p. 311-22.

3. Weir, B.S. and Cockerham.C., *Estimation of Linkage Disequilibrium in randomly Mating Populations.* Heredity, 1979. **42**: p. 105-111.
4. Weir, B.S., *Inferences about linkage disequilibrium.* Biometrics, 1979. **35**(1): p. 235-54.
5. Zhao, J.H., CurtisD. and Sham, P.C., *Model free analysis and permutation tests for allelic associations.* Human Heredity, 2000. **50**: p. 133-139.

CONTACT INFORMATION

Contact the authors at:

Marie-Pierre Dubé
Montreal Heart Institute
5000 Belanger
Montreal, QC H1T 1C8:
Work phone: (514) 376-3330 #2298
E-mail: marie-pierre.dube@umontreal.ca
Web: www.statgen.org

Amina Barhdadi
Work phone: (514) 376-3330 #3303
E-mail: amina.barhdadi@statgen.org