Leverage SAS/Genetics Package to perform Biomarker Drug Response Analysis on Pharmacogenomic Data.

Deepali Gupta, GCE Solutions, Mansfield, MA
Shirish Nalavade, Independent Consultant, Mansfield, MA

ABSTRACT

The primary objective of genetic analysis is to infer how an individual's genetic profile affects subject's response to drug treatment and moreover how safe and effective treatment dosage varies depending on their respective genetic makeup. ALLELE procedure analysis serve to characterize the markers themselves or the population from which they were sampled, and can also further serve as the basis for joint analysis on markers and traits. This procedure uses the notation and concepts described by Weir (1996) in reference for all equations and methods. This paper will provide an introduction to PROC ALLELE and will demonstrate on how to adopt this procedure to analyze pharmacogenomics (marker) data. We will illustrate on how to construct tables of allele, genotype frequencies, Hardy-Weinberg Equilibrium (HWE) analyses and linkage disequilibrium between each pair of markers and look at key statistical estimates used for inferences analysis.

INTRODUCTION

PROC ALLELE calculates the PIC, heterozygosity, and allelic diversity measures that serve to give an indication of marker informativeness. Such measures can be useful in determining which markers to use for further linkage or association testing with a trait. High values of these measures are a sign of marker informativeness, which is a desirable property in linkage and association tests.
Associations between markers might also be of interest. PROC ALLELE provides tests and various statistics for the association, also called the linkage disequilibrium, between each pair of markers. These statistics can be formed either by using haplotypes that are given in the data, by estimating the haplotype frequencies, or by using only genotypic information.
For more information on PROC ALLELE readers are encouraged to see reference.

GETTING STARTED: ALLELE PROCEDURE

PROC ALLELE provides the flexibility of analyzing data in either horizontal or vertical format by using different options. Pharmacogenomics data should have information about alleles for each gene/marker or data can be in form of genotypes instead of alleles. The ALLELE procedure shown in the following code calculates some basic summary statistics for each marker included in the analysis.

Source Dataset:

As example of variables in dataset:

ALLELE 1 NUCLEOTIDE	ALLELE 2 NUCLEOTIDE	GENE NAME	MARKER NAME	GENOTYPES

In Table 1. Source Data ,ALLELE 1 NUCLEOTIDE for VEGFA/ RS1570360 ALLELE1 is variable VEGFA__RS1570360_ALLELE1 and so on.

Example 1: As seen in figure below, we have gene and marker information for subjects. Second and Third columns are illustrating the set of alleles for the first marker, the next two columns are for the second marker, and so on. There is one row per each individual subject.

PT	VEGFA /RS1570360 ALLELE1	VEGFA /RS1570360 ALLELE2	VEGFA /RS2010963 ALLELE1	VEGFA /RS2010963 ALLELE2	VEGFA /RS3025039 ALLELE1	VEGFA /RS3025039 ALLELE2
2001	G	G	G	G	C	C
1001	G	G	C	G	C	C
1002	A	G	G	G	C	C

Table 1. Source data

```
PROC ALLELE DATA=H_ALLELE OUTSTAT=LD PREFIX=MARKER PERMS=10000 BOOT=1000 SEED=123;
VAR VEGFA__RS1570360_ALLELE1 VEGFA__RS1570360_ALLELE2
VEGFA__RS2010963_ALLELE1 VEGFA__RS2010963_ALLELE2 VEGFA__RS3025039_ALLELE1
VEGFA__RS3025039_ALLELE2;
RUN;
```

Example 2: Horizontal data is in form of genotypes instead of allele.

PT	VEGFA /RS1570360 GENOTYPE	VEGFA /RS2010963 GENOTYPE	VEGFA /RS3025039 GENOTYPE
2001	G/G	G/G	C/C
1001	G/G	C/G	C/C
1002	A/G	G/G	C/C

Table 2. Horizontal data

```
PROC ALLELE DATA=H_GENO OUTSTAT=LD PREFIX=MARKER PERMS=10000 BOOT=1000
SEED=123 GENOCOL DELIMITER='/';
VAR VEGFA__RS1570360_GENOTYPE VEGFA__RS2010963_GENOTYPE
VEGFA__RS3025039_GENOTYPE;
RUN;
```

Example 3: Vertical or Tall data.

PT	GENNAME	MARKER	ALLELE1	ALLELE2	GENOTYPE
2001	VEGFA	RS1570360	G	G	G/G
2001	VEGFA	RS2010963	G	G	G/G
2001	VEGFA	RS3025039	C	C	C/C
1001	VEGFA	RS1570360	G	G	G/G
1001	VEGFA	RS2010963	C	G	C/G
1001	VEGFA	RS3025039	C	C	C/C
1002	VEGFA	RS1570360	A	G	A/G
1002	VEGFA	RS2010963	G	G	G/G
1002	VEGFA	RS3025039	C	C	C/C

Table 3. Vertical data

```
PROC SORT DATA=VERT ;
  BY GENNAME MARKER;
RUN;

PROC ALLELE DATA=VERT PERMS=10000 SEED=123 GENOCOL;
  BY GENNAME MARKER;
  VAR GENOTYPE;
RUN;
```

By default, three tables are created. The first is a marker summary table, containing measures of marker informativeness:the polymorphism information content(PIC), heterozygosity, and allelic diversity; the number of alleles and number of individuals typed at each marker; and statistics for the HWE test. Test for HWE: ChiSq (the chi-square statistic), DF(the degrees of freedom for the chi-square test), ProbChiSq (the P-value for the chi-square test) and ProbExact an estimate of the exact P-value for the HWE test (only if the PERMS= option is specified in the PROC ALLELE statement).
Here is the "Marker Summary" table shown for above examples (1, 2 & 3) in output 1

The ALLELE Procedure

Marker Summary

Locus	Number of Indiv	Number of Alleles	PIC	Hetero-zygosity	Allelic Diversity	Test for HWE: Chi-Square	DF	Pr > ChiSq	Prob Exact
MARKER1	3	2	0.2392	0.3333	0.2778	0.1200	1	0.7290	1.0000
MARKER2	3	2	0.2392	0.3333	0.2778	0.1200	1	0.7290	1.0000
MARKER3	3	1	0.0000	0.0000	0.0000	0.0000	0	.	.

Output 1. Marker Summary

The next table is the Allele Frequency table. In addition to the frequency estimates themselves, these tables contain the standard error of the frequencies. When BOOTSTRAP= option is included in the PROC ALLELE statement, the bootstrap lower and upper limits of the confidence interval for the frequency based on the Confidence level determined by the ALPHA= option of the PROC ALLELE statement (0.95 by default). These tables have been suppressed in this analysis with the NOFREQ option. Here is the "Allele Frequencies" table shown for above examples (1, 2 & 3) in output 2.

Locus	Allele	Count	Frequency	Standard Error	CONTROLVAR	95% Lower Confidence Limit	95% Upper Confidence Limit
MARKER1	A	1	0.1667	0.1361	1	0.0000	0.5000
MARKER1	G	5	0.8333	0.1361		0.5000	1.0000
MARKER2	C	1	0.1667	0.1361	1	0.0000	0.5000
MARKER2	G	5	0.8333	0.1361		0.5000	1.0000
MARKER3	C	6	1.0000	0.0000	1	1.0000	1.0000

Output 2. Allele frequencies

Third table is Genotype frequency table. The "Genotype Frequencies" table lists all the observed genotypes (denoted by the two alleles separated by a "/") for each marker, with the observed genotype count and frequency, an estimate of the disequilibrium coefficient d, the standard error of the estimate, and when the BOOTSTRAP= option is specified, the lower and upper limits of the bootstrap confidence interval for d based on the confidence level determined by the ALPHA= option of the PROC ALLELE statement (0.95 by default). Here is the "Genotype Frequencies" table shown for above examples (1, 2 & 3) in output 3

Locus	Genotype	Count	Frequency	HWD Coefficient	Standard Error	CONTROLVAR	95% Lower Confidence Limit	95% Upper Confidence Limit
MARKER1	A/G	1	0.3333	-0.0278	0.0454	1	-0.2500	0.0000
MARKER1	G/G	2	0.6667	-0.0278	0.0454		-0.2500	0.0000
MARKER2	C/G	1	0.3333	-0.0278	0.0454	1	-0.2500	0.0000
MARKER2	G/G	2	0.6667	-0.0278	0.0454		-0.2500	0.0000
MARKER3	C/C	3	1.0000	0.0000	0.0000	1	0.0000	0.0000

Output 3. Genotype frequencies

Testing for linkage disequilibrium using Proc Allele

To get the statistics for Linkage Disequilibrium between pairs of markers use OUTSTAT=option in the PROC ALLELE statement as used in example1 & 2. The OUTSTAT= data set contains the following variables: the BY variables, if any,Locus1 and Locus2,(which contain the pair of markers for which the disequilibrium statistics are calculated), NIndiv contains the number of individuals that have been genotyped at both the markers listed in Locus1 and Locus2 (that is, the number of individuals that have no missing alleles for the two loci), Test indicates which disequilibrium test is performed, HWE for individual markers (when Locus1 and Locus2 contain the same value) or LD for marker pairs, ChiSq, which contains the chi-square statistic for testing for disequilibrium, DF contains the degrees of freedom for the chi-square test, ProbChi contains the P-value for the chi-square test, ProbEx contains an estimate of the exact P-value for testing the pair of markers in Locus1 and Locus2 for disequilibrium. This variable is included in the OUTSTAT= data set only when the PERMS= parameter in the PROC ALLELE statement is a positive integer and HAPLO=EST is not specified. Here is the “LD” dataset shown for above examples (1 & 2) in output 4.

LOCUS1	LOCUS2	NINDIV	DISTANCE	TEST	CHISQ	DF	PROBCHI	PROBEX
MARKER1	MARKER1	3	0	HWE	0.12	1	0.72903	1
MARKER1	MARKER2	3	1	LD	0.48	1	0.48842	1
MARKER1	MARKER3	3	2	LD	0.00	0	.	.
MARKER2	MARKER2	3	0	HWE	0.12	1	0.72903	1
MARKER2	MARKER3	3	1	LD	0.00	0	.	.
MARKER3	MARKER3	3	0	HWE	0.00	0	.	.

Output 4. Testing for linkage disequilibrium

Example4: LD Stat can be obtained from vertical data structure using TALL option.

```
PROC SORT DATA=VERT;
BY MARKER PT ;
RUN;

ODS OUTPUT LDMEASURES=LDMEASURES;
PROC ALLELE DATA=VERT TALL INDIV=PT MARKER=MARKER HAPLO=NONE CORRCOEFF DPRIME
OUTSTAT=LD MAXDIST=3 PERMS=10000 ;
VAR ALLELE1 ALLELE2;
RUN;
ODS OUTPUT CLOSE;
```

Same “LD” dataset will be created as shown in Output 4.

HAPLO=OPTION for linkage disequilibrium calculations and tests.

PROC ALLELE in SAS/GENETICS provides an effective tool for calculating LD coefficients and testing allelic association respectively. 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, Lewontin’s D’, the proportional difference d, and Yule’s Q.

This paper illustrates the use of HAPLO option of the PROC ALLELE.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. 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.

By default or when HAPLO=NONE or NONEHWD is specified, the composite linkage disequilibrium (CLD) coefficient is used in place of the usual linkage disequilibrium (LD) coefficient. In addition, the composite haplotype frequencies are used to form the linkage disequilibrium measures indicated by the options CORRCOEFF and DPRIME. When HAPLO=EST, the maximum likelihood estimates of the haplotype frequencies are used to calculate the LD test statistic as well as the LD measures. The HAPLO=GIVEN option indicates that the haplotypes have been observed, and thus the observed haplotype frequencies are used in the LD test statistic and measures.

Table 4 displays how the HAPLO= option of the PROC ALLELE statement interacts with the linkage disequilibrium calculations.

HAPLO= Option	LD Test Statistic	LD Exact Test	Estimate of Haplotype Freq
GIVEN	$\tilde{D}_{uv}$	Permutes alleles to form new 2-locus haplotypes	Observed freq, $\tilde{p}_{uv}$
EST	$\hat{D}_{uv}$	Not performed	Estimated freq, $\hat{p}_{uv}$
NONE	$\tilde{\Delta}_{uv}$	Permutes alleles to form new 2-locus genotypes	Composite freq, $\tilde{p}^*_{uv}$
NONEHWD	$\tilde{\Delta}_{uv}$	Permutes genotypes to form new 2-locus genotypes	Composite freq, $\tilde{p}^*_{uv}$

Table 4. Interaction of HAPLO= Option with LD Calculations

SAS codes Proc allele with HAPLO option

HAPLO=GIVEN

```
ODS OUTPUT LDMEASURES=LDMEASURES_GIVEN;
PROC ALLELE DATA=H_ALLELE OUTSTAT=LD_GIVEN PREFIX=MARKER CORRCOEFF       DPRIME
HAPLO=GIVEN NOFREQ;
      VAR VEGFA__RS1570360_ALLELE1 VEGFA__RS1570360_ALLELE2 VEGFA__RS1570360_ALLELE1
VEGFA__RS1570360_ALLELE2
          VEGFA__RS3025039_ALLELE1 VEGFA__RS3025039_ALLELE2
;
RUN;
ODS OUTPUT CLOSE;

PROC PRINT DATA = LDMEASURES_GIVEN WIDTH =MIN;
RUN;

PROC PRINT DATA = LD_GIVEN WIDTH =MIN;
RUN;
```

Outputs from ldmeasures_given and ld_given datasets are shown in output 5 and 6 below respectively.

Linkage Disequilibrium Measures

Locus1	Locus2	Number of Indiv	Haplotype	Count	Frequency	LD Coeff	Corr Coeff	Lewontin's D'
MARKER1	MARKER2	3	A-A	1	0.1667	0.1389	1.0000	1.0000
MARKER1	MARKER2	3	G-G	5	0.8333	0.1389	1.0000	1.0000
MARKER1	MARKER3	3	A-C	1	0.1667	0.0000	.	.
MARKER1	MARKER3	3	G-C	5	0.8333	0.0000	.	.
MARKER2	MARKER3	3	A-C	1	0.1667	0.0000	.	.
MARKER2	MARKER3	3	G-C	5	0.8333	0.0000	.	.

Output 5. Output from LDMEASURES_GIVEN dataset

LOCUS1	LOCUS2	NINDIV	DISTANCE	TEST	CHISQ	DF	PROBCHI
MARKER1	MARKER1	3	0	HWE	0.12	1	0.72903
MARKER1	MARKER2	3	1	LD	6.00	1	0.01431
MARKER1	MARKER3	3	2	LD	0.00	0	.
MARKER2	MARKER2	3	0	HWE	0.12	1	0.72903
MARKER2	MARKER3	3	1	LD	0.00	0	.
MARKER3	MARKER3	3	0	HWE	0.00	0	.

Output 6. Output from LD_GIVEN dataset

HAPLO=EST

```
ODS OUTPUT LDMEASURES=LDMEASURES_EST;
PROC ALLELE DATA=H_ALLELE OUTSTAT=LD_EST PREFIX=MARKER CORRCOEFF DPRIME HAPLO=EST
NOFREQ;
      VAR VEGFA__RS1570360_ALLELE1 VEGFA__RS1570360_ALLELE2 VEGFA__RS1570360_ALLELE1
VEGFA__RS1570360_ALLELE2
          VEGFA__RS3025039_ALLELE1 VEGFA__RS3025039_ALLELE2
;
RUN;
ODS OUTPUT CLOSE;

PROC PRINT DATA =LDMEASURES_EST WIDTH =MIN;
RUN;

PROC PRINT DATA =LD_EST WIDTH =MIN;
RUN;
```

Here is the “ldmeasures_est” and ld _est datasets shown for above examples in output 7 and 8 respectively.

Linkage Disequilibrium Measures

Locus1	Locus2	Number of Indiv	Haplotype	Frequency	LD Coeff	Corr Coeff	Lewontin's D'
MARKER1	MARKER2	3	A-A	0.1667	0.1389	1.0000	1.0000
MARKER1	MARKER2	3	G-G	0.8333	0.1389	1.0000	1.0000
MARKER1	MARKER3	3	A-C	0.1667	0.0000	.	.
MARKER1	MARKER3	3	G-C	0.8333	0.0000	.	.
MARKER2	MARKER3	3	A-C	0.1667	0.0000	.	.
MARKER2	MARKER3	3	G-C	0.8333	0.0000	.	.

Output 7. Output from LDMEASURES_EST dataset

LOCUS1	LOCUS2	NINDIV	DISTANCE	TEST	CHISQ	DF	PROBCHI
MARKER1	MARKER1	3	0	HWE	0.12	1	0.72903
MARKER1	MARKER2	3	1	LD	3.00	1	0.08326
MARKER1	MARKER3	3	2	LD	0.00	0	.
MARKER2	MARKER2	3	0	HWE	0.12	1	0.72903
MARKER2	MARKER3	3	1	LD	0.00	0	.
MARKER3	MARKER3	3	0	HWE	0.00	0	.

Output 8. Output from LD_EST dataset

HAPLO=NONE

```
ODS OUTPUT LDMEASURES=LDMEASURES_NONE;
PROC ALLELE DATA=H_ALLELE OUTSTAT=LD_NONE PREFIX=MARKER CORRCOEFF DPRIME HAPLO=NONE
NOFREQ;
      VAR VEGFA__RS1570360_ALLELE1 VEGFA__RS1570360_ALLELE2 VEGFA__RS1570360_ALLELE1
VEGFA__RS1570360_ALLELE2
          VEGFA__RS3025039_ALLELE1 VEGFA__RS3025039_ALLELE2
;
RUN;
ODS OUTPUT CLOSE;

PROC PRINT DATA =LDMEASURES_NONE WIDTH =MIN;
RUN;

PROC PRINT DATA =LD_NONE WIDTH =MIN;
RUN;
```

Here is the “ldmeasures_none” and ld_none datasets shown for above examples in output 9 and 10 respectively..

Linkage Disequilibrium Measures

Locus1	Locus2	Number of Indiv	Haplotype	Frequency	LD Coeff	Corr Coeff	Lewontin's D'
MARKER1	MARKER2	3	G-G	0.7500	0.1111	0.8000	1.0000
MARKER1	MARKER3	3	A-C	0.1667	0.0000	.	.
MARKER1	MARKER3	3	G-C	0.8333	0.0000	.	.
MARKER2	MARKER3	3	A-C	0.1667	0.0000	.	.
MARKER2	MARKER3	3	G-C	0.8333	0.0000	.	.

Output 9. Output from LDMEASURES_NONE dataset

LOCUS1	LOCUS2	NINDIV	DISTANCE	TEST	CHISQ	DF	PROBCHI
MARKER1	MARKER1	3	0	HWE	0.12	1	0.72903
MARKER1	MARKER2	3	1	LD	1.92	1	0.16586
MARKER1	MARKER3	3	2	LD	0.00	0	.
MARKER2	MARKER2	3	0	HWE	0.12	1	0.72903
MARKER2	MARKER3	3	1	LD	0.00	0	.
MARKER3	MARKER3	3	0	HWE	0.00	0	.

Output 10. Output from LD_EST dataset

ODS TABLE NAMES

PROC ALLELE assigns a name to each table it creates, and you must use this name to reference the table when using the Output Delivery System (ODS). These names are listed in Table 5.

Table 5 ODS Tables Created by the ALLELE Procedure

ODS Table Name	Description	Statement	Option
MarkerSumm	Marker summary		
AlleleFreq	Allele frequencies		
GenotypeFreq	Genotype frequencies		
LDMeasures	Linkage disequilibrium measures	PROC	CORRCOEFF, DELTA, DPRIME, PROPDIFF, RHO, or YULESQ
PopulationSummary	Population summary	POP	
CombinedFStats	Combined F statistics	POP	
MarkerFStats	Marker F statistics	POP	INDIVLOCI

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 procedures to test for LD, PROC ALLELE provides different LD measures and LD test statistics between two loci.

REFERENCES

SAS/Genetics(TM) 9.2 User's Guide. Available at
http://support.sas.com/documentation/cdl/en/geneug/59659/HTML/default/viewer.htm#geneug_allele_sect001.htm

ACKNOWLEDGMENTS

Many thanks to Vikash Jain at eClinical Solutions, A Division of Eliassen Group for influencing us and help support compile this paper and review it.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Please feel free to contact the author at:

Name: Deepali Gupta
Phone: 508-406-8082
E-mail: Deepali78@gmail.com

Name: Shirish Nalavade
Phone: 508 594 6337
E-mail: snalavade@eclinicalsol.com