PharmaSUG 2013 - Paper SP05

'Case-Control' the analysis of Biomarker data using SAS ® Genetic Procedure

JAYA BAVISKAR, INVENTIV HEALTH CLINICAL, MUMBAI, INDIA

ABSTRACT

Genetics aids to identify any susceptible diseases at the root cause level due its inherent ability to provide an array of information very specific to the building blocks of life. Hence, analysis of genetic data like Biomarkers or the genetic make-up is a crucial task and to derive accurate inferences adds complexity to it further. Taking this into consideration; SAS® has introduced new procedures that helps analyze genetic data so as to arrive at accurate conclusions.

The CASE CONTROL procedure is designed to handle such biomarker data there-by aiding to analyze and assess data more effectively and efficiently. Backed with statistical concepts and inbuilt options; the procedure allows to focus more on the interpretation of data in question through 3 readily available 'Chi-square' tests. The paper will discuss on bringing data in the desired format; applying SAS® options available; statistical computations to be considered and deriving correct inferences.

INTRODUCTION

A randomly selected Biomarker data can be assessed or interpreted for identifying correlation between a disease and the biomarkers representing it. The statistical analysis of the data helps in targeting medicines to correct the disease/ailment at the root cause level.

The procedure uses the same analogy as seen in the analysis of 'Case' and 'Control' studies where population is classified in sub-groups of 'Affected by disease/condition' and those 'Not affected by disease/condition'. It provides allelic/genotypic frequencies of similarity or variation that signify association of disease with the biomarker(s) of interest. Single nucleotide polymorphism (SNPs) may also be an observed phenomenon for a selected sub-population distinguished by environmental, demographical or geographical factors. Hence correct inferences can be drawn taking this into account.

The CASE CONTROL procedure provides an array of in-built options that makes it relatively easier to focus directly on inferences of the biomarker data. The results are displayed using three types of Chi-Square tests that are designed considering that evaluation is of very specific genotypic data.

	Number of M_1 Alleles			
	0	1	2	Total
Case	r_0	r_1	r_2	R
Control	s_0	s_1	s_2	S
Total	n_0	n_1	n_2	N

Source: See References 1 and 7

The following statistical computation of M_1 alleles is derived by considering 'N' number of total population classified by 'R' number of 'Cases' as opposed to 'S' number of 'Controls'. Thus the resultant 'Chi-square' values are computed considering the same.

1) **The Genotype Case Control Test** – helps evaluate the amount of correlation between the biomarker/allele and the disease it represents. Nielsen and Weir (See reference 4) obtained the following statistic for genotype case-control test that takes both additive and non-additive effects of allele in consideration.

$$X_G^2 = \sum_{i=0}^{2} \left[\frac{(Nr_i - Rn_i)^2}{NRn_i} + \frac{(Ns_i - Sn_i)^2}{NSn_i} \right]$$

Source: See References 4 and 7

Sasieni (See reference 5) computed the following statistics for the trend and allele case-control test. Here the assumption that 'Variance Inflation Factor' (VIF) is constant across a genome. Bacanu, Devlin and Roeder (See reference 2) further adjusted the trend statistics so that the variance correction can be applied to bi-allelic markers. This can be done by including the 'VIF' option while using this procedure.

2) **The Allele Case Control Test –** helps evaluate any additional correlations that exist between more than one biomarker/allele with respect to the disease.

$$X_A^2 = \frac{2N[2N(r_1 + 2r_2) - 2R(n_1 + 2n_2)]^2}{(2R)2(N - R)[2N(n_1 + 2n_2) - (n_1 + 2n_2)^2]}$$

Source: See References 2 and 7

3) **The Linear Trend Test -** helps evaluate any additional correlations that exist between more than one biomarker/allele with respect to the disease.

$$X_T^2 = \frac{N[N(r_1 + 2r_2) - R(n_1 + 2n_2)]^2}{R(N - R)[N(n_1 + 4n_2) - (n_1 + 2n_2)^2]}$$

Source: See References 2 and 7

METHOD:

It is important to bring data in the desirable format hence it is important to check if it follows the 'Hardy-Weinberg Equilibrium' (HWE). The HWE equates that alleles received from each parent have to be independent of each other in an ideal population. Disequilibrium is observed due to genetic mutations or any other factors such as migration, selection etc.

The PROC ALLELE procedure (based on concepts presented by Weir, 1996) can be used to verify if the data qualifies the HWE principle. The procedure helps in highlighting any associative attributes that result to identify potential markers. Hence, a high 'Polymorph Info Content '(PIC), allelic diversity, linkage disequilibrium are useful to shortlist markers or traits for the PROC CASECONTROL. Thus the overall data can be preened further by considering only such alleles or markers in the PROC CASECONTROL which appear to be promising in the said aspects. This helps to avoid suppression of any potential attributes (markers/traits) that get negated under 'noise' created by disassociated attributes.

Statistical computations exclude any 'missing values' from the analysis of 'bi-allelic' or multi-allelic' markers. While handling such data; one may consider applying formats so as to bring data in the desired format.

It is to be noted that the course of this paper is focused on explaining the usage of in-built SAS Genetic procedures hence the data used to derive output and analysis is based on 'dummy data' (Created by assigning value 1, 2, 3,..n to represent an allele or 'A', 'T', 'G', 'C' to represent genotypes). Usually the procedures are applicable for a wide range of biomarkers such as oncological, immunological or any other area of interest. Genetic sequences of disease/condition or biomarkers of interests are available at genomic databases NCBI, EBI (See references 8 and 9) and other search engines that are updated and maintained in a timely manner.

```
/*************************************************************************
      Following dummy data is categorized by status of allele that are
'Affected' (A) / 'Not Affected' (N). The 'Bi-allelic' markers are represented
by values '1' and '2' respectively.
*************************************************************************/

DATA ONCO_MARKER;
      INPUT AFFECTED $ MARKER1-MARKER16;
DATALINES;

N 2 1 1 1 2 1 1 1 2 2 1 1 1 1 1 1
N 2 1 2 2 . . 1 1 2 1 1 1 1 1 1 1
N 2 2 . . 2 1 1 1 2 1 2 1 1 1 2 1
N 2 1 . . 2 2 1 1 2 2 1 1 1 1 2 1
N 2 1 . . 2 2 1 1 2 1 . . 2 1 1 1
N 2 2 . . 2 2 1 1 . . 2 1 1 1 2 1
N 1 1 . . 2 2 1 1 1 1 2 1 1 1 2 1
N 1 1 . . 2 2 1 1 1 1 . . 1 1 2 1
N 2 1 . . 2 2 1 1 1 1 . . 2 1 2 1

A 2 1 2 1 2 1 1 1 1 1 2 1 . . 2 1
A 2 1 2 1 2 2 1 1 2 1 1 1 . . 1 1
A 2 2 2 1 2 2 1 1 2 2 . . . . 2 1
A 2 1 2 2 2 1 1 1 2 1 2 1 . . 2 2
A . . 2 2 2 1 . . 1 1 2 2 . . 2 1
A 1 1 1 1 2 1 1 1 2 1 1 1 . . 2 2
A 2 1 1 1 2 2 1 1 1 1 2 1 . . 2 1
A 2 1 2 2 2 2 1 1 2 2 . . . . 2 2
A 2 1 1 1 2 2 1 1 2 1 2 1 . . 1 1
A 2 1 2 2 2 1 1 1 2 1 2 1 . . 2 2

      ---// more data lines ---//

A 1 1 1 1 2 2 1 1 2 1 2 1 . . 2 2
A 2 1 2 1 2 1 1 1 2 1 2 2 . . 2 1
A 2 2 2 2 1 1 1 1 2 1 2 1 . . 2 2
A 1 1 1 1 2 1 . . 2 1 2 2 . . 2 2
A 1 1 2 1 2 1 1 1 2 1 2 1 . . 2 2
A 2 2 1 1 2 2 1 1 2 1 1 1 . . 2 1  ;
RUN;
```

Note: The data needs to be sorted as per status of disease (Affected/Not Affected), any individual identifiers or any other variables that are to be preferred.

```
PROC ALLELE DATA=ONCO_MARKER OUTSTAT=ONCO_ALLELE PREFIX=MARKER HAPLO=EST
      CORRCOEFF DPRIME;
      VAR MARKER1-MARKER16;
RUN;
```

Here the PROC ALLELE procedure accepts the input dataset 'ONCO_MARKER' and outputs to a desired dataset 'ONCO_ALLELE'. 'HAPLO' can be 'EST' (when haplotype frequencies are expected) or can have values 'NONE', 'GIVEN', 'NONEHWD'. 'CORRCOEFF' provides correlation coefficient and 'DPRIME' displays 'Lewontin's 'D'. 'VAR' contains analysis variables (Refer 7 for more details).

The SAS System

The ALLELE Procedure

Marker Summary

Locus	Number of Indiv	Number of Alleles	Polymorph Info Content	Heterozygosity	Allelic Diversity	Test for HWE		
						Chi-Square	DF	Pr > ChiSq
marker1	57	2	0.3571	0.4912	0.4654	0.1759	1	0.6749
marker2	51	2	0.3379	0.3529	0.4306	1.6590	1	0.1977
marker3	57	2	0.3267	0.5439	0.4114	5.9140	1	0.0150
marker4	50	2	0.0739	0.0800	0.0768	0.0868	1	0.7683
marker5	57	2	0.3749	0.7368	0.4998	12.8140	1	0.0003
marker6	52	2	0.3674	0.6346	0.4850	4.9466	1	0.0261
marker7	18	2	0.1780	0.2222	0.1975	0.2812	1	0.5959
marker8	57	2	0.3545	0.4737	0.4606	0.0460	1	0.8302

OUTPUT 1 A: The higher the PIC value, the greater the potential to have an association to a trait or disease.

Linkage Disequilibrium Measures

Locus1	Locus2	Number of Indiv	Haplotype	Frequency	LD Coeff	Corr Coeff	Lewontin's D'
marker1	marker2	50	1-1	0.4907	0.0357	0.1633	0.1830
marker1	marker2	50	1-2	0.1593	-0.0357	-0.1633	-0.1830
marker1	marker2	50	2-1	0.2093	-0.0357	-0.1633	-0.1830

//////////~~~~~~~~More Output Here ~~~~~~~~~~~~~~~~~~~~//////////

marker6	marker8	51	1-1	0.2865	0.0880	0.3756	0.6087
marker6	marker8	51	1-2	0.2919	-0.0880	-0.3756	-0.6087
marker6	marker8	51	2-1	0.0566	-0.0880	-0.3756	-0.6087
marker6	marker8	51	2-2	0.3650	0.0880	0.3756	0.6087
marker7	marker8	17	1-1	0.5177	0.0081	0.0578	0.1651
marker7	marker8	17	1-2	0.3941	-0.0081	-0.0578	-0.1651
marker7	marker8	17	2-1	0.0412	-0.0081	-0.0578	-0.1651
marker7	marker8	17	2-2	0.0471	0.0081	0.0578	0.1651

OUTPUT 1 B: The information related to Heterozygosity, Linkage Disequllibrium, Allelic diversity etc. help understand the genetic makeup and interaction between alleles.

```
PROC CASECONTROL DATA=ONCO_MARKER PREFIX=MARKER PERMS=10000;
        VAR MARKER1-MARKER16;
        TRAIT AFFECTED;
RUN;
```

The 'VAR' will usually contain the 'Marker alleles' to be considered or the 'Genotype Marker' if the GENOCOL= option is to be implemented. Assuming that interest is on identifying the 'Affected' as opposed to 'Not Affected' individuals; the 'TRAIT' will display statistics of key focus.

The output is segregated amongst three individual 'Chi-square' values and their respective 'Probabilities'. That is to say the 'Genotype', 'Allelic' and 'Trend' of data is captured by the procedure independently. The 'Degrees of Freedom' for 'Genotype' will be '2' in a 'Bi-allelic' marker; '1' for 'Allele' and '1' for 'Trend'.

APPLICATIONS OF PROC CASECONTROL PROCEDURE:

APPLICATION ON A BI-ALLELIC MARKER DATA:

The SAS System

Locus	NumTraitA	NumTraitN	ChiSqGenotype	ChiSqAllele	ChiSqTrend	dfGenotype	dfAllele	dfTrend	ProbGenotype	ProbAllele	ProbTrend
Marker1	39	18	2.423	1.307	1.384	2	1	1	0.298	0.253	0.239
Marker2	40	11	4.232	0.974	0.825	2	1	1	0.120	0.324	0.364
Marker3	40	17	7.149	4.778	7.048	2	1	1	0.028	0.029	0.008
Marker4	32	18	7.729	7.407	7.729	1	1	1	0.005	0.006	0.005
Marker5	40	17	13.77	0.486	0.924	2	1	1	0.001	0.486	0.337
Marker6	38	14	1.532	0.067	0.097	2	1	1	0.465	0.796	0.755
Marker7	0	18	0.000	0.000	0.000	0	0	0	.	.	.
Marker8	40	17	9.043	8.346	8.590	2	1	1	0.011	0.004	0.003

OUTPUT 2: Output of a bi-allelic marker that provides the Genotype, Allele and Trend probabilities listed.

PREFIX= option assigns the prefix value that has been provided. Default prefix 'M' gets assigned to markers used.

OUTSTAT= option helps redirecting output of Chi-square and 3 test-probabilities in SAS Dataset.

PERMS= 'number' per-mutates the trait as many times as the 'number' specified in the option. A 'Monte Carlo' estimate of exact *p*- values is provided when the option is used. Accuracy is directly proportional to the number of permutations used but it comes at the cost of longer processing time when very large data is in consideration. Hence it is suggested to adjust the number of permutations sufficient to be considerably accurate.

SEED= option starts generating random number from the 'seed' number provided (Initial number that user provides).

OUTPUT DISPLAYING INCREASED PRECISION BY USAGE OF PERMS= AND SEED= OPTIONS

Locus	NumTraitA	NumTraitN	ChiSqGenotype	ChiSqAllele	ChiSqTrend	dfGenotype	dfAllele	dfTrend	ProbGenotype	ProbAllele	ProbTrend
Marker1	39	18	2.4232	1.30688	1.38375	2	1	1	0.29771	0.25296	0.23946
Marker2	40	11	4.2324	0.97374	0.82495	2	1	1	0.12049	0.32375	0.36374
Marker3	40	17	7.1491	4.77766	7.04784	2	1	1	0.02803	0.02883	0.00794
Marker4	32	18	7.7295	7.40741	7.72947	1	1	1	0.00543	0.00650	0.00543
Marker5	40	17	13.7710	0.48565	0.92353	2	1	1	0.00102	0.48587	0.33655
Marker6	38	14	1.5316	0.06708	0.09699	2	1	1	0.46497	0.79564	0.75547
Marker7	0	18	0.0000	0.00000	0.00000	0	0	0	.	.	.
Marker8	40	17	9.0425	8.34572	8.58966	2	1	1	0.01088	0.00387	0.00338

The SAS System

Locus	NumTraitA	NumTraitN	ChiSqGenotype	ChiSqAllele	ChiSqTrend	dfGenotype	dfAllele	dfTrend	ProbGenotype	ProbAllele	ProbTrend
Marker1	39	18	2.4232	1.30688	1.38375	2	1	1	0.33803	0.29810	0.28948
Marker2	40	11	4.2324	0.97374	0.82495	2	1	1	0.11368	0.43871	0.48268
Marker3	40	17	7.1491	4.77766	7.04784	2	1	1	0.02365	0.04092	0.01256
Marker4	32	18	7.7295	7.40741	7.72947	1	1	1	0.01322	0.01493	0.01322
Marker5	40	17	13.7710	0.48565	0.92353	2	1	1	0.00108	0.54266	0.40687
Marker6	38	14	1.5316	0.06708	0.09699	2	1	1	0.55922	0.82622	0.79481
Marker7	0	18	0.0000	0.00000	0.00000	0	0	0	.	.	.
Marker8	40	17	9.0425	8.34572	8.58966	2	1	1	0.00977	0.00543	0.00453

OUTPUT 3: Marker8 in above differs considerably when the PERMS= option along with SEED= is used. For PERMS=10000; precision of all 3 tests increased.

APPLICATION ON A MULTI-ALLELIC MARKER DATA:

Multi-allelic markers are processed in a similar way as bi-allelic markers and differ only at the number of alleles into consideration (may have more than two situated either on the same or a different loci).

Let us assume that 'M1'and 'M2' comprises of multiple alleles ('A_1', 'A_2',…,'A_N') grouped at two different loci such that their similarity and variation is segregated across the data. So after applying the procedure; the output results in probabilities that are not significant at α = 0.05. (Refer **OUTPUT 4A**). Hence it may be erroneously concluded that there is no association between marker and disease by observing the probabilities of 'Markers' (M1, M2).

Given a case where previous research identifies certain allele(s) of the Biomarkers to show a possible correlation between allele and disease exists. So it becomes necessary to cross-verify the results from a new perspective and identify if such a correlation does exist.

So in such a scenario; it becomes necessary to identify such alleles of interest and map them to form a 'Bi-allelic' pattern. This can now be checked for probable association with the disease. The same data now shows significance at α = 0.05. (Refer **OUTPUT 4B**) implying that association may exist.

Thus, the probability of the marker became significant when 'alleles' of no relative significance (that contributed to the noise in overall statistics) were excluded from the analysis. Hence, careful observation and any background information related to the alleles, markers or disease of interest is highly recommended.

The SAS System

Locus	NumTrait1	NumTrait2	ChiSqGenotype	ChiSqAllele	ChiSqTrend	dfGenotype	dfAllele	dfTrend	ProbGenotype	ProbAllele	ProbTrend
M1	30	30	27.333	4.441	5.039	24	7	7	0.2892	0.7278	0.6552
M2	30	30	18.077	8.772	13.244	15	7	7	0.2586	0.2694	0.0664

OUTPUT 4A: Markers indicate no association with disease due to no relative significance at α = 0.05

The SAS System

Locus	NumTrait1	NumTrait2	ChiSqGenotype	ChiSqAllele	ChiSqTrend	dfGenotype	dfAllele	dfTrend	ProbGenotype	ProbAllele	ProbTrend
M1	30	30	12.193	6.599	10.103	2	1	1	0.0023	0.0102	0.0015

OUTPUT 4B: The marker probability changed considerably when the over-shadowing effect of disassociated alleles was removed.

OUTPUT USING TALL= OPTION

The TALL= option allows data to be processed column-wise as opposed to row-wise there-by providing leverage of reading very large and continuous data. The data needs to be sorted by including the 'Marker ID' and then by 'Individual ID' variable in the 'BY' statement. This option requires to specify the 'MARKER=' and 'INDIV=' to display the probabilities and provides the same output as seen in '**OUTPUT 2**'.

STRATIFIED ANALYSIS:

Similar analogy follows for data classified into 'categories' or 'STRATA'. This can be done by using the 'STRATA' statement that helps to specify the categorization of data by treatment, gender or any other classification.

Stratification also helps to assess 'X-linked' biomarkers and their association with the disease.

The 'MISSING=' option allows to consider observations where the classification is absent or missing.

ODS Table Name	Description	Statement / Option
StrataLevels	Strata levels	STRATA
StrataInfo	Strata information	STRATA / INFO

Source: Reference 7

The PROC CASECONTROL uses the mentioned table names when using the 'Output Delivery System' (ODS).

CONCLUSION

Thus by utilizing in-built SAS genetic procedures; the focus can be shifted towards analyzing any association between a disease and marker; towards prognosis; towards identifying prospective markers amongst a group of markers or to associate any other significant finding of interest.

RECOMMENDED READING AND REFERENCES

1. **Armitage, P. (1955), "Tests for Linear Trends in Proportions and Frequencies," Biometrics, 11, 375–386.**
2. **Bacanu, S-A., Devlin, B., and Roeder, K. (2000), "The Power of Genomic Control," American Journal of Human Genetics, 66, 1933–1944.**
3. **Devlin, B. and Roeder, K. (1999), "Genomic Control for Association Studies," Biometrics, 55, 997–1004.**
4. **Nielsen, D.M. and Weir, B.S. (1999), "A Classical Setting for Associations between Markers and Loci Affecting Quantitative Traits," Genetical Research, 74, 271–277.**
5. **Sasieni, P.D. (1997), "From Genotypes to Genes: Doubling the Sample Size," Biometrics, 53, 1253–1261.**
6. **Base SAS® Procedures Guide**
7. **SAS Genetics 9.2 User's Guide**
   http://support.sas.com/documentation/onlinedoc/genetics/genetics.pdf
8. **National Center For Biotechnology Information (NCBI) www.ncbi.nlm.nih.gov/**
9. **EMBL EUROPEAN BIOINFORMATICS INSTITUTE www.ebi.ac.uk/**

ACKNOWLEDGMENT

I am grateful to **Dr. Prashant Kirkire,** Country Manager, India, Inventiv Health Clinical; for his encouragement provided throughout and for the careful review of the paper. **Sandeep Sawant** is greatly appreciated for his invaluable feedback and suggestions at a very short notice.

CONTACT INFORMATION

Your comments and questions are greatly appreciated and encouraged. Contact the author at:

NAME: **JAYA BAVISKAR**

ENTERPRISE: **INVENTIV HEALTH CLINICAL,**

ADDRESS: **INVENTIV HEALTH CLINICAL,**
GROUND FLOOR,
A WING, MARWAH CENTRE,
KRISHNALAL MARWAH MARG,
ANDHERI(EAST),

MUMBAI - 400072,

COUNTRY : **INDIA**

E- MAIL: **Jaya.Baviskar@inventivhealth.com; JBaviskar@yahoo.com**