Statistical Analysis of Molecular Data Using Software Packages



Statistical analysis of genetic association studies among unrelated individuals *First edition*

Dr. Kuldeep Kumar Tyagi



Department of Animal Genetics & Breeding College of Veterinary & Animal Sciences Sardar Vallabhbhai Patel University of Agriculture & Technology Modipuram, Meerut- 250 110

Published Online:

First edition: 2021 Total pages: 18

Published by:

Department of Animal Genetics & Breeding College of Veterinary & Animal Sciences Sardar Vallabhbhai Patel University of Agriculture and Technology Meerut- 250 110, Uttar Pradesh, India

Publication No.

SVP/2021/06/02/231 Dated: July 14th, 2021 (for official use)

How to cite this Lecture notes

Tyagi, K 2021, *Statistical analysis of genetic association studies among unrelated individuals*, lecture notes series *Statistical Analysis of Molecular Data Using Software Packages*, Training on "Application of Molecular and Bioinformatics Tools in Agriculture and Allied Sciences" Sardar Vallabhbhai Patel University of Agriculture & Technology, Meerut, Uttar Pradesh- 250110, India. Delivered 14th July 2021. Retrieved online from <u>https://vepub.com</u>

Address Correspondence to:

Dr. Kuldeep Kumar Tyagi Associate Professor & OIC Department of Animal Genetics & Breeding COVAS, SVPUAT, Meerut- 250110 (U.P.) India drtyagivet@gmail.com +91 9601283365 (M)



This work is licensed under a Creative Commons Attribution 4.0 International License. Copyright © 2021 K K Tyagi

<u>ABOUT</u>

These lecture notes on "Statistical analysis of genetic association studies among unrelated individuals" were prepared under my lecture series on "Statistical Analysis of Molecular" Data Using Software Packages". This lecture was delivered to trainees attending 14 day training on "Application of Molecular and Bioinformatics Tools in Agriculture and Allied Sciences" organized by CST, UP Centre of Excellence in Agriculture Biotechnology in Collaboration with DBT funded Bioinformatics Infrastructure Facility, College of Biotechnology, SVPUAT during 7-20th July 2021. I have tried to explain how molecular data of a population can be analyzed using SPSS and HW_Test software packages. Overall, population association studies in which unrelated individuals for different economic traits are typed at a number of Single Nucleotide Polymorphism (SNP) markers excluding family-based association studies, admixture mapping or linkage studies have been introduced and explained. Use of diagrams, explanatory boxes, examples and tables have deliberately been used throughout the notes to create an interest among the trainees. Once through with these lecture notes readers will be able to understand the basics and application of statistical analysis of genetic association studies among unrelated individuals using statistical packages. I had tried my level best to simplify the concept in easy to understand language. Further constructive suggestions to improve this lecture notes are always welcome from readers on my email and whatsapp.

[KULDEEP KUMAR TYAGI]

DISCLAIMER

These lecture notes on "Statistical analysis of genetic association studies among unrelated individuals" under the lecture series on "Statistical Analysis of Molecular Data Using Software Packages" have been compiled from various resources available in the public domain. Only excerpts from the original works have been used. This is being done for educational purposes in the interest of developing a concise and updated reading material for students and trainees with no intent of commercial benefits. References to the source of material used have been included in the footnotes. The author does not claim any ownership of any copyrighted material included by chance in the lecture. If due to inability to trace the original source any copyrighted material got included, it may please be brought to the notice of the author for rectification.

[KULDEEP KUMAR TYAGI]

OTHER LECTURE NOTES BY THE AUTHOR

As on: 13-07-2021

	Series: Animal Genetics & Breeding										
S.No.	Title Date of Publication Editions			Pages	Views	Down Ioads	Access				
1	Introduction and importance of statistics and biostatistics	12-10-2020	First	20	40.1k	254	<u>Download</u>				
2	Probability	20-08-2020	First	33	12.6k	193	<u>Download</u>				
3	Probability	24-11-2020	Second	24	29.5k	280	<u>Download</u>				
4	Probability Distribution	07-12-2020	First	33	12.6k	193	Download				
5	Introduction of population genetics	23-02-2021	First	16	22.8k	128	<u>Download</u>				
6	Genetic structure of population	23-02-2021	First	40	22.2k	211	Download				
7	Hardy Weinberg Law	10-07-2021	First	42	49	0	Download				

ACKNOWLEDGEMENT

The author wishes to thank Dr. Pankaj Kumar and organizing committee of training "Application of Molecular and Bioinformatics Tools in Agriculture and Allied Sciences" for giving me an opportunity to prepare and deliver this lecture to the participants. The help rendered by Dr. Atul Gupta while preparing this manuscript was commendable. This manuscript may not have taken this present shape without his critical review and inputs. I also want to thank Dr. Rajbir Singh, Dean, College of Veterinary and Animal Sciences for his continuous persuasion and support. I am blessed that he considers me befitted for lecture deliveries at various platforms and always routes to me such opportunities related to my expertise. Thanks to my wife, Dr. Surbhi Tyagi; son, Om Tyagi and daughter, Somya Tyagi for their selfless love, affection, support and sparing me from my various household duties. They adjusted and spared me the time from our daily routine for preparation of this lecture note.

[KULDEEP KUMAR TYAGI]

TABLE OF CONTENTS

Statistical analysis of genetic association studies among unrelated individuals	1
1. Coding the genotypes	1
1.1. Feeding data in excel workbook	1
2.SPSS data sheet	3
2.2. Importing data to SPSS	3
3. Analysis	6
3.1. Frequency table using SPSS	6
3.2. Testing for Hardy Weinberg Equilibrium (HWE)	7
3.2.1 Analysis using HW_TEST software	8
3.3. Descriptive Statistics using SPSS	10
3.3. Analysis of Variance (ANOVA) and Duncan test using SPSS	12
4. Summary	18

Statistical analysis of genetic association studies among unrelated individuals

People all around the world have been conducting genetic association studies for a long time. Still the literature to practically understand the statistical analysis of such data using software packages is scarce. Statistical analysis of molecular data is a vast topic. So we will be undertaking one of the important basic and preliminary topics *"Statistical analysis of genetic association studies among unrelated individuals"* in this lecture. You will be able to analyse molecular data of a population using SPSS and HW_Test software packages after going through this lecture. Overall, we will be discussing population association studies in which unrelated individuals for different economic traits are typed at a number of Single Nucleotide Polymorphism (SNP) markers excluding family-based association studies, admixture mapping or linkage studies.

1. Coding the genotypes

To demonstrate practical hands on we will be making use of dummy data. This dummy data is available in Excel workbook and can be <u>downloaded online</u>. This data sheet represents SNP data of six genes say (Gene 1, Gene 2, Gene 3, Gene 4, Gene 5 and Gene 6) on 550 lactating buffaloes and their associated data on traits of economic importance like Age at First Calving (AFC) in days, Service Period (SP) in days and Standard 300 day Lactation yield (SLY) in Kg.

1.1. Feeding data in excel workbook

First data can be feeded in Excel workbook for the ease of data feeding. The header row will define the variables. The first column can be used to either assign the serial number or identification number of the buffalo. Columns thereafter constitute the variables like various genes and traits of economic importance. Data pertaining to each individual is assigned in a single row. Say for example serial number 101, the genotypes (coding) were AB(2), AA(1), AB(2), CD(9), AB(2), AB(2) for Gene 1, Gene 2, Gene 3, Gene 4, Gene 5 and Gene 6 respectively. Whereas the values for corresponding traits of economic importance Age at First Calving (AFC) is 1077 days, Service Period (SP) is 121 days and

Standard 300 day Lactation yield (SLY) is 2225 Kg respectively. Similarly, data pertaining to each buffalo is feeded into the excel workbook. Save your excel workbook on your computer in a folder designated for this analysis.

 Table 1: Number of buffaloes pertaining to each genotype and their corresponding coding.

Ganas	Number o	fbı	uffalo	oes f	for e	ach		C	odin	g for	' eacl	า
Genes		ger	noty	be					ge	noty	ре	
Gene 1	Two alleles		А	В					А	В		
	Ν	А	110	266				А	1	2		
	550	В		174				В		3		
				-								
Gene 2	One Allele		А						А			
	Ν	А	550					А	1			
	550											
Gene 3	Three Alleles		А	В	С				А	В	С	
	Ν	А	47	129	115			А	1	2	3	
	550	В		77	119			В		4	5	
		С			63			С			6	
Gene 4	Four Alleles		А	В	С	D			А	В	С	D
	Ν	А	19	84	59	67		А	1	2	3	4
	550	В		48	94	75		В		5	6	7
		С			29	51		С			8	9
		D				24		D				10
						-						
Gene 5	Two Alleles		А	В					А	В		
	Ν	А	107	264				А	1	2		
	550	В		179				В		3		

Gene 6	Three Alleles		А	В	С			А	В	С
	Ν	А	36	174	177		А	1	2	3
	550	В		28	101		В		4	5
		С			34		С			6

2.SPSS data sheet2.2. Importing data to SPSS

We will be importing data to SPSS once our excel workbook is ready. Open SPSS program installed on your computer. A dialog box as shown in the figure below will open.



Click on type in data to open the SPSS data view. Go to $File \Rightarrow Read Text Data..$ It will prompt you to locate your data file as shown below.

ta Open Data							×
Look in: 🕂 Dow	nloads	•	🔮 🛅				
5X	CoA_UG COB_UG COVAS Data Set DST Sun Emailing	_lind Year_l S_VPB324_S UG_AGB_S XISX /ey Data Dr/ Final Electo	Elective_U SetI.xlsx etI.xlsx ajit.xlsx ajit.xlsx	GE-223_S	et-I.xlsx .00 P.M.) (1)).xlsx	
4							•
File name: Dat	a Set.xlsx						Open
Files of type: Exc	el (*.xls, *.xlsx	*.xlsm)				-	Paste
Minimize Lot Syli	el (*.xls, *.xlsx us (*.w*) k (*.slk) ase (*.dbf)	,*.xlsm)					Cancel <u>H</u> elp
SAS	6 (*.sas7bdat,	*.sd7, *.sd2	, *.ssd01,	*.ssd04, *	xpt)		
Sta	ta (*.dta) t (* txt * dat * /	csv)					
All I	Files (*.*)					~	

Select the File type from the drop down menu and then locate your file and click open. This will open another dialog box as shown below.

Opening Exc	el Data Source	×							
C:\Users\919	C:\Users\91960\Downloads\Data Set.xlsx								
<table-cell> Read varia</table-cell>	able names from the first row of data								
Worksheet	Worksheet: Data Set [A1:J1001]								
Range:	B1:J551								
Maximum wid	th for string columns: 32767								
OK Cancel Help									

Define your data range (B1:J551) and then click OK. This will import all your excel data

into the SPSS data view. Now we will be defining variables by clicking the variable view tab given at the bottom left of SPSS.

	-		-	-	-	-			
28	2	1	3	5	2	5	1146	122	1771
29	2	1	6	4	2	6	1018	123	2033
30	2	1	2	6	3	3	1016	124	2078
31	3	1	1	8	3	4	1121	119	2044
32	3	1	4	7	1	2	1222	119	2185
33	1	1	6	9	1	5	1260	109	2082
34	1	1	3	10	2	1	1110	128	1898
36	2	1	2	6	2	2	1125	101	2011
36	2	1	5	9	2	3	1176	129	1938
37	2	1	3	5	3	5	1062	113	2130
	4								
Data View	Variable View								
4 P	Type here to sear	ch	0	Hi 💽	🗧 💼 😺	s 💼			

This will open the variable view of the SPSS as shown below.

🚰 Untitl	ed2.si	av [DataSe	et1] - 16	IM SPSS Stat	istics [ata Editor								
<u>Elle</u>	dit	View D	ata	Transform	Anal	yze Dire	ct <u>Marketing</u>	Graphs Utilities	Add- <u>o</u> ns <u>W</u>	indow Help				
2				1	2		≛ ≢	H 86	¥ =	₫ 🛄		•	j	
		Nam	10	Туре		Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
1		Gene1		Numeric	1	2	0	Gene 1	None	None	12	Right Right	🚓 Nominal	🔪 Input
2		Gene2		Numeric	1	2	0	Gene 2	None	None	12	T Right	🚴 Nominal	S Input
3		Gene3		Numeric	1	12	0	Gene 3	None	None	12	Right	🚓 Nominal	S Input
4		Gene4		Numeric	1	12	0	Gene 4	None	None	12	Right	🚓 Nominal	S Input
5		Gene5		Numeric	1	2	0	Gene 5	None	None	12	Right	🚓 Nominal	S Input
6		Gene6		Numeric	1	2	0	Gene 6	None	None	12	Right Right	🚓 Nominal	S Input
7		AFCday:	s	Numeric	1	2	0	AFC (days)	None	None	12	Right	ne Scale	S Input
8		SPd		Numeric	1	2	0	SP(d)	None	None	12	Right	🥔 Scale	S Input
9		SLYKg		Numeric	1	2	0	SLY Kg	None	None	12	Right	🥔 Scale	S Input
10														
11							1	🧃 Value Labels				×		
12														
13								Value Labels						
14								Value: 1				Spelling		
15								Label: AA						
16											_			
17								Add						
18								200						
19								Change						
20								Remove						
21														
22														
23									OK C	Cancel Help				

Click in the cell of the column labeled "Values". This will open a dialog box in which value and their associated genotype (Label) can be added. Click on Ok once you complete feeding all the genotype labels corresponding to a gene. The final dataset file can be <u>downloaded here</u>.

Now each genotype will be visible in the data view using the toggle button provided at the top menu ribbon as shown in the figure below. So basically if you consider gene 1, which has two alleles (A and B) will be having three genotypes AA, AB and BB and their corresponding coding will be 1, 2 and 3 respectively. Similarly, coding patterns for other genes are also followed in a similar fashion. Coding our data in this way will help us to analyse our data with ease using above stated softwares.

3. Analysis

3.1. Frequency table using SPSS

Go to Analyze \Rightarrow Descriptive Statistics \Rightarrow Frequencies. This will open a dialog box, select and put all the variable named genes into the variables box and click OK. This will give you an output file with frequency of genotypes for all the genes. The file can be <u>downloaded online</u>



Below is the screenshot of the output file. These frequencies are being feeded in table 1 and will be used subsequently in Hardy Weinberg equilibrium testing. The benefit of making this table is to have handy absolute frequencies for each genotype ready to be feeded in HW_TEST software. The format prevents mistakes while feeding the value of each genotype because it simulates the data feeding design available in the HW_TEST software.

http://www.commenterscore.com/second statistics for the second statist

File	Edit	View	Data	Transform	m <u>I</u> nsert	Format	Analyze	Direct Mark	eting <u>G</u> raph	s <u>U</u> tilities	Add-ons	Window	Help
6					Ш,		~]		* =	0	b 🗄		
	Outp	ut Log Frequer E Titl	ncies e	1	Freq	uencie Setl] C	S:\Users\	HP\Downl	oads\Untit	led2.sav	,		
		- Act	ive Data tistics	set					Statistics				
	÷	🕒 Fre	quency	Table			Gene 1	Gene 2	Gene 3	Gene 4	Gene 5	Gene 6	٦
		- 6] Title		N	Valid	55	0 550	0 550	550	550	550	1
		- C	Gene	2		Missing	1	0 0	0 0	0	0	0	
			Gene Gene Gene Gene	3 4 5 5	Freq	uency	Table						
								Gene 1	1				
							Frequency	Percent	Valid Perce	Cum ent Pe	ulative rcent		
					Valid	AA	110	20.0	20	0.0	20.0		
						AB	266	48.4	48	1.4	68.4		
						BB	174	31.6	31	.6	100.0		
						Total	550	100.0	100				
								Gene 2	2				
				-						Cum	ulative		
					Mallet		Frequency	Percent	Valid Perce	ent Pe	rcent		
					valid	~	550	100.0	100		100.0		
								Gene 3	3				
							Frequency	Percent	Valid Perce	Cum ent Pe	ulative rcent		
					Valid	AA	47	8.5	6	3.5	8.5		
						AB	129	23.5	23	0.5	32.0		
						AC	115	20.9	20	.9	52.9		
						BB	77	14.0	14	.0	66.9		
						BC	119	21.6	21	.6	88.5		
						CC	63	11.5	11	.5	100.0		
						Total	550	100.0	100).0			

3.2. Testing for Hardy Weinberg Equilibrium (HWE)

Molecular data on a population are generally subjected to be tested for Hardy Weinberg equilibrium (HWE) as a measure to find preliminary anomalies. HWE test is used extensively on a routine basis to exclude samples with gross molecular typing defects from the usually very large sets of genetic markers presently used in various types of population genetic analyses. Natural or domesticated populations are generally

subjected to inbreeding, population stratification, selection or sometimes disease association may lead to deviations from HWE. Sometimes deviation from HWE may be an indication of presence of a common deletion polymorphism, a mutant PCR-primer site or because of a tendency to miscall heterozygotes as homozygotes. So far researchers¹ have tested for HWE primarily as a data quality check and have discarded loci that, for example, deviate from HWE among controls at significance level $\alpha = 10^{-3}$ or 10^{-4} . However, the possibility that a deviation from HWE is due to a deletion polymorphism or a segmental duplication that could be important in disease causation should now be considered before discarding loci.

We will be using user friendly, window based executable open source software HW_TEST for testing HWE². The program can be obtained free of charge directly from the github internet repository as an standalone executable zip file <u>https://github.com/Lemes-RenanB/HardyWeinbergTesting</u>. The downloadable zip file contains a well defined user manual. The Pearson test is easy to compute, but the χ^2 approximation can be poor when there are low genotype counts, and it is better to use a Fisher exact test, which does not rely on the χ^2 approximation.

3.2.1 Analysis using HW_TEST software

Hardy Weinberg Equilibrium Test File	_		×
HARDY WEINBERG EQUILIBRI	јм те	ST	
Copyright 2006-2019 by Fernando A. Bautzer Santos, Renan B. Ler Departamento de Genetica e Biologia Evolutiva, Universida Rua do Matao, 277 05508-090 - Sao Paulo, SP, Brazil otto@usp.br	mes and P de de Sao	aulo A. Ott Paulo	0
Number of alleles: ³			
Next			

¹ Balding DJ. A tutorial on statistical methods for population association studies. Nat Rev Genet. 2006 Oct;7(10):781-91. doi: 10.1038/nrg1916. PMID: 16983374.

² Santos, F., Lemes, R. B., & Otto, P. A. (2020). HW_TEST, a program for comprehensive HARDY-WEINBERG equilibrium testing. Genetics and molecular biology, 43(2), e20190380. https://doi.org/10.1590/1678-4685-GMB-2019-0380

The opening window of software HW_TEST asks for the number of alleles in your data set. Let us consider a case of gene 3 which has three alleles. Put this value into the designated field and press next.

Now refer to the structural representation of gene 3 and their corresponding genotypes given in Table 1. The values corresponding to each genotype need to be filled in designated space provided in HW_TEST software.

🕴 Hardy Weinberg Equilibrium File	Test		_	×
_ Statistical Tests	HARDY WEINBERG	EQUILIBRIUM TEST		
Statistical F58ts	Chi-square without correction	Exact probability by simulation		
Number of Genoty	A B C A 47 B 77 C 63			
		Run		

Once all the values are entered correctly, click on "Run". This will open the output file with results. The first result output provides us an idea about the observed and expected genotypes frequencies. One can see that the sum of these genotypic frequencies equals unity. Now, the second output provides us with the estimates of allelic frequencies. Finally the P value for χ^2 estimate is provided. If the value of "P" is greater than 0.05 indicates that our null hypothesis of no difference between observed and expected genotypic frequencies is accepted. Therefore, for the given gene the population is in

Hardy Weinberg equilibrium. We can assume for the given gene evolutionary forces like mutation, migration and selection are not operating in the given population at the time of screening.



3.3. Descriptive Statistics using SPSS

To obtain the descriptive statistics we click on Analyze followed by comparing means and then means ie. Analyze \Rightarrow Comparing Means \Rightarrow Means. This will open a dialog box. In the dialog box put all the traits of economic importance into the dependent variable box and all the genes into the independent list as shown in the figure given below.

tans Means		\times
	Dependent List:	Options Bootstrap
	Layer 1 of 1 Previous <u>N</u> ext	
	Independent List	
ОК	Paste Reset Cancel Help	

Now this will open another dialog box in which various parameters for descriptive statistics can be defined.

t	Means: Options			\times
	Statistics: Median Grouped Median Sum Range First Last Kurtosis Std. Error of Kurtosis Skewness Std. Error of Skewness Harmonic Mean Geometric Mean Percent of Total Sum Percent of Total N	*	Cell Statistics: Mean Number of Cases Standard Deviation Std. Error of Mean Minimum Maximum Variance	
	Statistics for First Layer Statistics for First Layer Anova table and eta Test for linearity Continue	Cancel	Help	

Click on continue to save the changes and then Ok to obtain the results of descriptive statistics. The output file can be <u>downloaded here</u> for your ready reference to match your results when you practice this exercise.

i Descriptive Statistics.spv [Document1] - IBM SPSS Statistics Viewer										
<u>File Edit View Data Transform</u>	n <u>I</u> nsert	F <u>o</u> rmat <u>A</u> nalyze	Direct <u>M</u> arketin	ng <u>G</u> raphs	s <u>U</u> tilities A					
😂 🖩 🖨 🗟 🤌		r 7								
Jt DD	AFC (days) SP(d) SLY Kg * Gene 1									
/leans	Gene	1	AFC (days)	SLY Kg						
📺 Title	AA	Mean	1154.28	126.60	2065.25					
📑 Notes		N	110	110	110					
Active Dataset		Std. Deviation	98.570	10.494	140.987					
AFC (days) SP(d) SLY Kg * Gen		Std. Error of Mean	9.398	1.001	13.443					
AFC (days) SP(d) SLY Kg * Gen		Minimum	900	86	1688					
🛱 AFC (days) SP(d) SLY Kg * Gen		Maximum	1411	149	2518					
AFC (days) SP(d) SLY Kg * Gen		Variance	9716.039	110.114	19877.219					
AFC (days) SP(d) SLY Kg * Gen	AB	Mean	1173.55	126.73	2072.97					
AFC (days) SF(d) SET Kg Gen		Ν	266	266	266					
		Std. Deviation	97.540	11.303	153.680					
		Std. Error of Mean	5.981	.693	9.423					
		Minimum	949	86	1698					
	•	Maximum	1423	158	2627					
	·	Variance	9514.067	127.751	23617.535					
	BB	Mean	1194.81	126.26	2087.71					
		N	174	174	174					
		Std. Deviation	104.815	10.973	148.846					
		Std. Error of Mean	7.946	.832	11.284					
		Minimum	867	95	1680					
		Maximum	1535	154	2537					
		Variance	10986.224	120.412	22154.995					
	Total	Mean	1176.42	126.56	2076.09					
		N	550	550	550					
		Std. Deviation	100.959	11.023	149.663					
		Std. Error of Mean	4.305	.470	6.382					
		Minimum	867	86	1680					
		Maximum	1535	158	2627					

The output provides, mean, number of observations, standard deviation, standard error of mean, minimum, maximum and variance pertaining to each genotype.

3.3. Analysis of Variance (ANOVA) and Duncan test using SPSS

Now we will be subjecting our data for finding any association between various genes studied and traits of economic importance available with us for each individual.

Go to $Analyze \Rightarrow General Linear Model \Rightarrow Univariate$. This will open another dialog box. Put one quantitative trait (say AFC) into the dependent variable box and all the genes in the Fixed factor box as shown below. Then click on the model which will open another dialog box to define our model for ANOVA.



Click on the custom radio button, select all the fixed factors and choose main effects type from the drop down menu of build terms. Click on the arrow to put all fixed factors into the model box and to activate the continue button. Click on continue to save your changes.

🭓 Univariate: Model

Specify Model O Full factorial	ustom						
Federe & Covariates:	Build Term(s) Type: Main effects						
Sum of sguares: Type III 👻 Include intercept in model							
Continue Cancel Help							

Now click on the Post Hoc button to carry out post hoc tests in your analysis.

🔄 🤤 Univariate			×
SP(d) [SPd] SLY Kg [SLYKg]	*	Dependent Variable: AFC (days) [AFCdays] Fixed Factor(s): Gene 1 [Gene1] Gene 2 [Gene2] Random Factor(s):	Model Contrasts Plots Post Hoc Save Options Bootstrap
	*	Covariate(s):	
ОК Ра	ste	Reset Cancel Help	

This will open another dialog box as shown below.

Provide the second seco	\times								
Factor(s): Post Hoc Tests for: Gene2 Gene1 Gene3 Gene2 Gene4 Gene3 Gene6 Gene5	4								
Equal Variances Assumed LSD S-N-K Waller-Duncan Bonferroni Tukey Type I/Type II Error Ratio: 100 Sidak Tukey's-b Dunnett Scheffe Duncari Control Category: Last R-E-G-W-F Hochberg's GT2 Test R-E-G-W-Q Gabriel @ 2-sided < Control © > Control									
Equal Variances Not Assumed Tamhane's T2 Dunnett's T3 Games-Howell Dunnett's C Continue Cancel Help									

Put all the fixed factors into the post hoc box, select Duncan checkbox and click on continue to save your changes. Now click on Ok to obtain the results.

 Univariate SP(d) [SPd] SLY Kg [SLYKg] 	•	Dependent Variable: AFC (days) [AFC days] Fixed Factor(s):	X Model ontrasts Plots
	*	Random Factor(s):	ost <u>H</u> oc Save ptions potstrap
	*	<u>C</u> ovariate(s): <u>W</u> LS Weight:	
ОК	Paste	Reset Cancel Help	

This will open the output window with desired results. Similarly other two traits of economic importance will be put in the dependent variable box one by one and click ok to obtain the results. The output file can be <u>downloaded here</u> for your ready reference to match your results when you practice this exercise.

The first table we need to lookup in the output file is "Test of between subject effect". This is the ANOVA table.

i ANOVA.spv [Document2] - IBM SPSS Statistics Viewer																
<u>F</u> ile <u>E</u> dit	View	Data	Trans	sform	Insert	F <u>o</u> rmat	Analyze	Direc	t <u>M</u> ark	eting	<u>G</u> raphs	Utilities	Add-ons	Wir	ndow	<u>H</u> elp
) 🖻					2			Ł		0	•	7		
Cutput Cog Cutput Cup				Tests of Between-Subjects Effects Dependent Variable: AFC (days)											_	
	- Ad	tive Data amings	set		Source		Type of So	Type III Sum of Squares		df	Mean Square		F		Sig.	
	Between-Subje		Correc	Corrected Model		235629.529 ^a 23		10244.762		1.005	5	.456	1			
-			Interce	Intercept		06685.0		1	304306685.0		29861.737	·	.000	1		
ė	🚺 Po	st Hoc T	Hoc Tests		Gene1	Gene1		97386.253		2	48693.126		4.778		.009	1
	E	Title			Gene2			.000		0						1
	-	Gene '	1		Gene3		46	937.857		5	93	87.571	.921		.467	1
		. С н	omo		Gene4		27	702.804		9	30	78.089	.302	2	.974	1
	□ ···· (∰ T ····· (∰ A □ ···· (∰ A □ ···· (∰ A		Gene5		12	412.451		2	62	06.225	.609		.544	1		
			Gene6 Error		30	30790.643 5 5360214.609 526		6158.129		.604		.697	1			
					5360			526	10190.522					1		
		<u>e</u> T	tle	-	Total		7667	78406.0		550						1
			omo	1	Correc	ted Total	5595	844.138		549						1
	e -6	Gene	a A 4		a. R	Squared	= .042 A	djusted I	R Squ	ared = .	000)			_		-

The table shown above is for the dependent variable AFC. The first column shows the sources of variation which in our case are different genes. Now in this table we have to look for the source of variation which is significantly affecting our dependent variable. The gene 1 in this case is contributing a significant effect with p = 0.009 which is less than 0.05. No other gene is causing a significant effect on our dependent variable AFC. The R squared value of 0.042 represents that around 4.2 % variation in our dependent variable is explained by our model. Adjusted R-squared is a modified version of R-squared that has been adjusted for the number of predictors in the model. The adjusted R-squared increases when the new term improves the model more than would be expected by chance. It decreases when a predictor improves the model by less than expected.

Since only gene 1 is contributing significant effect on dependent variable the post-hoc tests pertaining to this gene will only be considered. The output results of post-hoc Duncan test are summarized under the heading of "Homogeneous Subsets".



The output table for Duncan test is quite helpful in determining the superscripts of means that differed significantly (α = 0.05). Thus the mean AFC (days) for genotype AA, AB and BB will bear superscripts a, ab and b respectively. Similarly results for all the genes can be reported in similar fashion. This enables us to analyze molecular data for testing Hardy Weinberg equilibrium along the identification of genes affecting dependent variables significantly. The method also gives us an idea about the percent variation explained by our predictor variables in the model.

4. Summary

This lecture on "Statistical analysis of genetic association studies among unrelated individuals" teached us one of the methods of coding our genotype data. Thereafter data is feeded as each column representing a variable and each row representing observations made on an individual. It has been observed that HW_TEST software is user friendly and proved to be an indispensable tool for testing a population for Hardy Weinberg equilibrium. This software also provides us the estimates of genotypic and allelic frequency in a matter of a few seconds. Further usage of SPSS software has been discussed for finding absolute frequencies of various genotypes pertaining to each gene. SPSS also proved to be user friendly menu driven software to obtain descriptive statistics and ANOVA with post - hoc tests for dependent variables pertaining to various genotypes. Final interpretation of ANOVA helped us to determine genes causing significant variation in dependent variables. Thus, the conjunction of HW_TEST and SPSS softwares can successfully be used in statistical analysis of genetic association studies among unrelated individuals.

ABOUT THE AUTHOR



Dr. Kuldeep Kumar Tyagi had completed his B.V.Sc & A.H. in the year 2006 from Guru Angad Dev Veterinary and Animal Sciences University, Ludhiana, Punjab India. He got admission in a master program in the subject of Animal Genetics and Breeding at Indian Veterinary Research Institute, Bareilly, Uttar Pradesh, India after securing 6th rank in All India ICAR-JRF examination. He had completed his Masters in the year 2008 and carried out research on competent fibroblast cells used in somatic cell nuclear transfer. He qualified CSIR Net in his first attempt during the final semester of the masters program itself. He got selected as Assistant Professor in the year 2009 at

College of Veterinary Science & A.H. at Navsari Agricultural University, Navsari, Gujarat, India. He enriched his practical knowledge and expertise in the subject of Animal Breeding while disbursing his duties as Scheme Incharge at Livestock Research Station of the same university for 9 years. During the same tenure he also accumulated practical expertise on various aspects of field level breeding programs while heading "All India Coordinated Research Project on Goat Improvement -Surti Field Unit" as Principal Investigator. He completed his Ph.D. in the year 2016 from the same university as an inservice candidate. He had worked on gene expression studies on mammary epithelial cells of buffaloes during his Ph.D. degree program. He had been selected as Associate Professor in the department of Animal Genetics & Breeding, College of Veterinary and Animal Science, Sardar Vallbhbhai, Patel University of Agriculture & Technology, Meerut, Uttar Pradesh, India in the year 2018. Since then he has been heading the same department as Officer-Incharge. He had handled 5 externally funded and 27 institutionally funded research projects. He had coguided two masters students. He has in his credit 62 research papers, 14 research recommendations, 7 lecture notes and 4 success stories. He is a member of 4 professional societies and attended 21 conferences/ symposiums/ workshops. He has remained on a panel of experts for framing question papers for National level, State level examination bodies and various Universities. He is hosting a google site for online teaching https://sites.google.com/view/learnagb and can be reached at drtyagivet@gmail.com for initiating a conversation.

Published by:

Department of Animal Genetics & Breeding College of Veterinary & Animal Sciences Sardar Vallabhbhai Patel University of Agriculture and Technology Meerut- 250 110, Uttar Pradesh, India

<u>To cite this lecture notes:</u>

Tyagi, K 2021, *Statistical analysis of genetic association studies among unrelated individuals*, lecture notes series *Statistical Analysis of Molecular Data Using Software Packages*, Training on "Application of Molecular and Bioinformatics Tools in Agriculture and Allied Sciences" Sardar Vallabhbhai Patel University of Agriculture & Technology, Meerut, Uttar Pradesh- 250110, India. Delivered 14th July 2021. Retrieved online from <u>https://vepub.com</u>



This work is licensed under a Creative Commons Attribution 4.0 International License. Copyright © 2021 K K Tyagi