# User Manual for *csig* for Windows

Version 0.2.9

program implemented by

**Robert Yu** 

theoretical methodology developed by

Jian Wang and Sanjay Shete

August 6, 2008

easy computing with YU

# **Table of Contents**

INSTALLATION AND UNINSTALLATION	4
INSTALLATION	4 5
GET STARTED	6
Exploration Mode Batch Mode	6 7
DESCRIPTION	11
METHODOLOGY	11
DETAILED CSIG USAGE AND DESCRIPTION	11
USAGE SYNOPSIS EXPLANATION <i>I. Making Example Data</i> <i>How are the random samples created</i> <i>II. Data input</i> . <i>III. Output</i> DESCRIPTION	11 12 12 12 13 13 14 14
APPENDIX	17
How to configure console window How to quickly change directory to a desired destination in console window	17

# Installation and uninstallation

### Installation

This version is limited to Windows (2000, XP and Vista). Installation can be done in the following steps:

Upon saving a copy of the program on your local machine, click on Sciences, and follow the on-screen instructions:

😣 Composite Significance	0.2.9 Setup	Composite Significance 0.2.9 Setu	ıp	
	Welcome to the Composite Significance 0.2.9 Setup Wizard	License Agreement Please review the license terms before inst	alling Composite Significance 0.2.9.	8
	This wizard will guide you through the installation of Composite Significance 0.2.9.	Press Page Down to see the rest of the agr	eement.	
	It is recommended that you close all other applications	Licensing Terms and Conditions:		
	Televant system files without having to reboot your computer.	IMPORTANT: CAREFULLY READ THIS LICE THIS PRODUCT INDICATES YOUR ACKNOW LICENSE AND AGREE TO ITS TERMS. IF YOU DO NOT AGREE, DO NOT DOWNLC Cancer Center RESERVES THE RIGHT TO R DISCONTINE IF LEF CE AND DESTROY ALL	VSE BEFORE USING THIS PRODUCT. US WLEDGMENT THAT YOU HAVE READ THI: VAD THIS PRODUCT. U.T. M.D. Anderson VEQUEST THAT UNLICENSED USERS	ING 5 n
		If you accept the terms of the agreement, agreement to install Composite Significance	select the first option below. You must a 0.2.9. Click Next to continue.	ccept the
		<ul> <li>I accept the terms in the License Agreer</li> <li>I do not accept the terms in the License</li> </ul>	nent Agreement	
		csig Windows Installer by rky101 2008		
	Next > Cancel		< Back Next >	Cancel
8 Composite Significance	0.2.9 Setup	🛛 🙆 Composite Significance 0.2.9 Setu	up	
Choose Install Location				
Choose the rolder in which to i	nstall Composite Significance 0.2.9.		2.9 is being installed.	2
Setup will install Composite Sig folder, click Browse and select	nificance 0.2.9 in the following folder. To install in a different another folder. Click Install to start the installation.	Extract: csig.exe		
		Output folder: C:\WINDOWS		
		Extract: csig.exe		
- Dectination Folder				
C:\Program Files\Composi	te Significance Browse			
Space required: 1.9MB Space available: 89.8GB				
csig Windows Installer by rky101		csig Windows Installer by rky101 2008		
	< Back Install Cancel		< Back Next >	Cancel
	😣 Composite Significance 0.2.	9 Setup 💶 🗔 🖂		
		ompleting the Composite		
	Si Si	ignificance 0.2.9 Setup Wizard		
	Cor	mposite Significance 0.2.9 has been installed on your nputer.		
	Clic	k Finish to close this wizard.		
		Run Composite Significance 0.2.9		
		< Back Finish Cancel		

If the checkbox was checked on the last screen (above), a console window with help message will pop up.

🔤 Composite Siginicance	<u>- 🗆 ×</u>		
USAGE:	<b>_</b>		
csig -option1 -option2 []			
-v   -version show version -h   -? -help show this message -m   -man show this message			
-e¦-example making embedded example data -y¦-yes combined with "-e" to give a silent support of overwriting any existing example data			
-f¦-file input file -p1 1st p-value (a real number) -p2 2nd p-value (a real number)			
-c1 column position of p-value1 -c2 column position of p-value2			
-o ¦ -outp output filename Default is "output_CS.csv". In the exploration mode, this op will only affect log file name. User can use redirect to capture the output in a file	tion		
-d ¦ -decimal an integer representing the decimal place in output. Effo numbers are 0 to 9 (but using 0 be meaningless although csig will take.) Default is 6. Use any number greater than 9 csig will generate the maximum the current computer can express.	ective would y		
SYNOPSIS:			
I. CREATE example data files			
csig -e y (default is "n")			
This will generate 3 example data, storing in "examples" sub-directory.			
Any value other than "y" will stop making			
If any example data exist from previous making, More	<b>_</b>		

You may press "space" bar to go through the pages, or simply press "Q" to exit the message, and it should stay at "C:\". You may start to explore, or choose to quit the console to view the start menu of the installation.

1 our Windows XP	m Microsoft Office Tools	🖌 🗑 Composite Significance 🕨	M.D. Anderson Cancer Center
	microsoft Outlook with Business Contact Manager	•	📾 manual 🔹 🔀 README.pdf
	m Microsoft Works	•	Composite Significance
	m Nikon View 6	•	w Uninstall
N	m Real	•	
All Programs 🌔	Rovio	•	

At the start menu, within "Composite Significance", you have a "manual" folder containing "README.pdf" serving as user guide (this document), "Composite Significance" shortcut link to launch console window. "Uninstall" will allow you easily remove all install files without your looking for separately. A link to "M.D. Anderson Cancer Center" provides a convenience to visit its website.

### Uninstallation

Robert K. Yu (rkyu@mdanderson.org)

To remove the installation, go to "Start"  $\rightarrow$  "Programs"  $\rightarrow$  "Composite Significance"  $\rightarrow$  "Uninstall". It'll delete all the files and programs.

Solution October 2017 Series Annual of Csig.exe version 0.2.9

Robert K. Yu (rkyu@mo	danderson.	n.org)	ŀ
	Compos	site Significance 0.2.9 Uninstall 🛛 🛛 🔀	
	2	Are you sure you want to completely remove Composite Significance 0.2.9 and all of its components?	
		Yes No	
		🚟 Composite Significance 0.2.9 Uninstall 🛛 🔀	
		Composite Significance 0.2.9 was successfully removed from your computer.	
		ОК	

Note: after installations, "csig.exe" is physically sitting in "C:\WINDOWS",

Address 🗁 C:\WINDOW	S		
Name	Size	Туре	Date Modified 🔻
Scsig.exe	1,181 KB	Application	8/7/2008 11:04 PM

from where user can directly invoke it without setting it in the path. If you have your own working directory in path, such as "c:\bin", you may create a copy of csig.exe there. And you can use it without any links in the start menu. But if the installation was done through the installer, it'd be better not to change the file location otherwise it'll make later uninstallation problematic.

# Get started

### **Exploration Mode**

Before start, it might benefit to know how to configure console window in appendix.

- "Start" → "Programs" → "Composite Significance" to invoke the console. Then make and change to a working folder, e.g. "c:\myWork",
- Composite Siginicance
  C:\>mkdir myWork
  C:\>cd myWork
  C:\myWork>\_

2. To explore the program, type this

C:\myWork\csig -p1 your\_p\_value1 -p2 your\_p\_value2 [-decimal 3]

where [-decimal 3] is optional if user desires to have output values with 3 decimal places. And

 $your_p_value1$ " and " $your_p_value2$ " are the pair of p-values from your logistic regression analysis and Hardy-Weinberg Equilibrium test (order doesn't matter). For example, to calculate composite significance results using p-value 1 = 0.03 and another 0.005, use the following command

C:\myWork>csig -p1 0.03 -p2 .005

Upon pressing "Enter", the result will print to screen and be saved in a log file with default filename "log\_output\_CS.txt".

csig (version - 0.2.9 with options:	Win32) is running
chrMum column	=
decimal	=
input filename	=
location column	= log_output_CS.txt
making example	= no
output filename	=
p-value1	= 0.03
p-value2	= .005
pi column	=
snplD column	=
RESULTS	
p1	= 0.030000
p2	= 0.005000
p-value_TS	= 0.000075
[-log10(p-value_TS)]	= 4.124939
p-value_TSM	= 0.000085
[-log10(p-value_TSM)]	= 4.068769
TS-value	= 0.988750
TS-value	= 0.987929

To view the log file, you can either type "more" or "write" command with the filename. The former will print the content on screen and the latter will launch Wordpad with the file content.

C:\myWork>write log_output_CS.txt	
🗒 log_output_CS.txt - WordPad	
Eile Edit View Insert Format Help	
Running csig (version - 0.2.9 Win32)	
Thu Aug 7 23:50:34 2008	
INPUT FILE INFO	
User provided options:	
chrNum column =	
decimal = 6	
input filename =	
location column =	
making example = no	
output filename = output_CS.csv	
p-value1 = 0.03	
p-value2 = .005	
p2 column =	
snpID column =	
	Ξ
RESULTS	
	-
p1 = 0.030000	
p2 = 0.005000	
p-value_TS = 0.000075	
$[-\log 10 (p-value_TS)] = 4.124939$	
$p-value_1SM = 0.0000000000000000000000000000000000$	
TS-value = 0.988750	
TSM-value = 0.987929	
csig exit successfully.	
(Thu Aug 7 23:50:34 2008)	
Total time elapsed: O sec.	
Ear Help, proce 51	×
ror neip, press ri	

Note: using this exploration mode of csig.exe can't indicate an output file to save the result. You can use commandline redirect "> outputfile" to capture or ">> outputfile" to accumulate the result. Using commandline option "-o result.txt" here can allow user to make dedicated log filename, e.g.

"log\_result.txt" in this example.

### **Batch Mode**

 First to get the embedded example data sets (total 3) by running csig with option "-e y" or "-example yes":

C:\myWork\csig -e y

Then a screen on the right will show up. The program terminates and lists the 3 example data sets sitting in a sub folder "examples\".

csig (version - 0.2.9 Win32) is running with options:
<pre>chrNum column =</pre>
Could Not Find C:\myWork\examples
MESSAGE <making data="" sample=""> : Writing to testing.dat</making>
MESSAGE <making data="" sample=""> : Writing to sample_data2.dat 1000 lines</making>
MESSAGE <making data="" sample=""> : Writing to sample_data2.csv 5000 lines Volume in drive C has no label. Volume Serial Number is 3C9D-6646</making>
Directory of C:\myWork\examples
08/07/2008 11:59 PM 317 testing.dat 08/07/2008 11:59 PM 49,148 sample_data2.dat 08/07/2008 11:59 PM 238,655 sample_data2.csv 3 File(s) 288,120 bytes 0 Dir(s) 96,461,004,800 bytes free

#### Robert K. Yu (rkyu@mdanderson.org)

2. Now if you invoke this "-e y" option again or accidentally use it in future while you've already created the example data sets, the program will pause to ask if you want to replace the existing ones.

🖼 Composite Siginicance – csig -e y	_
csig (version - 0.2.9 Win32) is running with options:	
<pre>chrNum column =</pre>	
WARNING <making data="" sample=""> : Example data may already exist. Continue (y/ n</making>	'n)? !

If "y" is chosen, new example data sets will be created. Except for "testing.dat", in the other two data sets the p-values are generated by random although sample sizes are fixed (see detail in the help manual by typing "-h" or "-?" or see below). Here is part of "sample\_data2.dat":

p1 p2 snp		
0.0745406337082386	0.0464276634156704	snpID2000
0.15244622528553 0.4	0436440333724 snpID2001	
0.00894095934927464	0.390869622118771 snp:	ID2004
0.0162267098203301	0.511203425936401 snp:	ID2009
0.279701948165894 0.1	125540100038052 snpID2016	
0.268928272649646 0.1	130892470479012 snpID2025	
0.0720584439113736	0.0093390978872776	snnTD2O36

 To avoid such a pause, this version of csig.exe supports a silent permission of overwrite (or abort) of re-creating example data sets. To overwrite, use this

```
C:\myWork\csig -e y -y y
```

Or abort when previous make is existing

C:\myWork\csig -e y -y n

 If using "sample\_data2.dat" for testing, now at prompt,

> C:\myWork\csig -f examples\sample\_data2.dat -c1 1 -c2 2 -d 7



This requests csig to take a sample data with one set of p-values in column 1 (option -c1 1) and another in column 2 (option -c2 2), and the output result with 7 decimal places.

The screen will print out the one on the right.

And two files are generated, a log file log\_output\_CS.txt and output\_CS.csv.

5. To view the result file, type this

C:\myWork\write output\_CS.csv

p1,p2,snp,p-value\_TS,[-log10(p-value\_TS)],p-value\_TSN,[-log10(p-value\_TSN]],TS-value,TSN-value 0.0745406337082386,0.0464276634156704,smp1D2000,0.0093405,2.0296513,0.0102014,1.9913421,0.8744530,0.8680348 0.1524462258853,0.4045644033724,smp1D2001,0.1676818,0.7755142,0.1747416,0.7576037,0.4680574,0.4538292 0.00894095934927464,0.390869622118771,smp1D2004,0.0556926,1.2542024,0.0498267,1.3025375,0.66934363,0.7083503 0.0162267098203301,0.511203425936401,smp1D2009,0.0985209,1.0024715,0.0887221,1.0519683,0.55922574,0.6108239 0.279701948165894,0.125540100038052,smp1D2016,0.0999099,1.0227866,0.0994769,1.0022776,0.601914,0.5879105 0.268928272649646,0.130892470479012,smp1D2025,0.0938855,1.0274015,0.1002116,0.9990819,0.6019651,0.5863916 0.0720584439113736,0.009339078872776,smp1D2024,0.0027444,2.5615559,0.0026214,2.5814623,0.9319475,0.931041 0.001067494958752.0.00849222112447023,smn1D2049.0.000737.4,4242084.0.0006355.4,4449578.0.9921220.992120

Another way to view the result is to open it up in MS Excel, where user may take advantage of tools Excel provides, e.g. sorting and charting, etc. To open it in Excel, just go to the working folder and directly click on output\_CS.csv file. A quick way to go to the current working folder, user may type at the prompt the following

C:\myWork\start .

C:\myWork>start .\_

Address 🛅 C:\myWo	rk		
Name 🔺	Size	Туре	Date Modified
examples		File Folder	8/8/2008 12:10 AM
📃 log_output_CS.txt	2 KB	Text Document	8/8/2008 12:15 AM
🔊 output_CS.csv	107 KB	Microsoft Excel	8/8/2008 12:15 AM

Notice a little dot after "start".

	A1	▼ fx	p1						
	A	В	С	D	E	F	G	Н	
1	p1	p2	snp	p-value_TS	[-log10(p-value_TS)]	p-value_TSM	[-log10(p-value_TSM)]	TS-value	TSM-value
2	0.074540634	0.046427663	snplD2000	0.0093405	2.0296313	0.0102014	1.9913421	0.874453	0.8680348
3	0.152446225	0.404364403	snplD2001	0.1676818	0.7755142	0.1747416	0.7576037	0.4680574	0.4538292
4	0.008940959	0.390869622	snplD2004	0.0556926	1.2542024	0.0498267	1.3025375	0.6934363	0.7083503
-5	0.01622671	0.511203426	snplD2009	0.0985209	1.0064715	0.0887221	1.0519683	0.5922574	0.6108239
6	0.279701948	0.1255401	snplD2016	0.0939099	1.0272886	0.0994769	1.0022776	0.6019134	0.5879105
7	0 100010170	0.4000047	0001D2025	0 0000022	4 0074045	0.4000446	0 0000910	0 6040654	0 2000040

- 6. Now you may test how csig detects and treats errors in the input data set. Example data file "testing.dat" was provided just for this purpose. Here is what the data looks like:
  - SNPID, Chr, Loc, p1, p2 rs34221 1 23432 0.4332 0.001 rs34222 1 23442 3.3e-5 0.001 rs34223 1 23452 1.56332 p2 rs34224 1 23462 0.5674 rs34221, 4 43432 0.4332 0.005 rs674222 5, 33442 1.2e-15 -3.051 rs994223, 61, 73452, 1.4332, p2 rs764224, z 83462, 0.1332 rs764224, x 83462, 0.1332

The data has p-values pairs in columns 4 and 5. Some of them either containing one p-value or containing non-number. Also notice the mixture of the delimiters; not only mixing "," with spaces/tabs, but also some comma comes with space before and/or after it. But csig can detect and process them properly.

Run this following command,

csig -f examples\testing.dat -c2 5 -c1 4 -o result\_of\_problem\_data.txt

where –o option indicates user defined output filename. Although it contains ".txt" extension, the program changes it to ".csv".

After executing the command, the screen and log file both report the finding of errors and how many error lines were dropped, and how many valid lines yield the results.

Robert K. Yu (rkyu@mdanderson.org)

August 6, 2008

<pre>sig (version - 0.2.9 Win32) is running vith options: chrNum column = input filemae = examples\testing.dat location column = log file = log_output_CS.txt making example = no output filemae = result_of_problem_data.txt p-value2 = pi column = 4 p2 column = 5 snplD column = csig (version - 0.2.9 Win32) is running with opts: f = examples\testing.dat -o result_of_problem_data. log file = log_result_of_problem_data.txt Checking the basic file properties counting lines 11 BASIC INFORMATION Filename extension Path extension Path extension Path examples\ File size (byte) 317 &lt;- S17 Bytes Line size (byte) 28.82 C- Line complex.testing.dat: SNFID, Chr.Loc.pt.p2 ri Aug 8 00:10:37 200 Had RY OF EEROR REPORTING Total error found: 7 Total lines dropped: 8 Total lines kept: 3 LISI OF DROPPED LINES</pre>	Composite Siginicance	
chrNun column = docinal = 6 input ficine = examples\testing.dat location terms = making example = na output filemame = result_of_problem_data.txt p-value2 = prolumn = 4 p2 column = 4 p2 column = 5 snplD column = file = log_result_of_problem_data.txt Checking the basic file properties counting lines 11 BASIC INFORMATION Filemame extension Path examples\ Supression > 61 lines 11 Filemame extension Path examples\ Solution = 11 bise (byte) 317 <- File size (byte) 28.82 <- Line size (byte) 28.82 <- Line size (byte) 7235915 2000 Mode (permission) 72329 0.001 rs34221 1 23442 3.3e=5 0.001 Line size double 110 SUMMARY OF ERBOR REPORTING Total lines kept: 3 Line size doubled: 2 Line size doubled: 3 List of DROPPED LINES List of DROPPED LINES 	sig (version - 0.2.9 Win3) ith options:	2) is running
<pre>input filemane = examples\testing.dat location column = making example = no output filemane = result_of_problem_data.txt p-valued = p-valued = p-valu</pre>	chrNum column =	
location column = log file = log_output_CS.txt making example = no output filename = result_of_problem_data.txt p-value2 = pl column = 4 pcolum = 5 snplD column = csig (version - 0.2.9 Win32) is running with opts: -f examples/testing.dat -o result_of_problem_data. log file = log_result_of_problem_data.txt Checking the basic file properties counting lines 11 BRSIC INFORMATION Filename stem Filename extension Path column = 11 State size (byte) 317 <- Total number of lines 11 LIST OF DROPPED LINES - LIST OF DROPPED LINES - LIST OF DROPPED LINES - - - - - - - - - - - - -	decimal = b input filename = ex	amples\testing.dat
making example = nog_output_Ls.txt making example = no output filename = result_of_problem_data.txt p-value2 = pl column = 4 p2 column = 5 snplb column = csig (version - 0.2.9 Win32) is running with opts: -f = examples\testing.dat -o result_of_problem_data. log file = log_result_of_problem_data.txt Checking the basic file properties counting lines 11 BASIC INFORMATION 	location column =	
output filename = result_of_problem_data.txt p-value2 = pl column = 4 p2 column = 5 snplD column = csig (version - 0.2.9 Win32) is running with opts: -f = examples\testing.dat -o result_of_problem_data. log file = log_result_of_problem_data.txt Checking the basic file properties counting lines 11 BASIC INFORMATION 	making example = no	g_output_cs.txt
p value - p value - csig (version - 0.2.9 Win32) is running with opts: -f = examples\testing.dat -o result of problem_data. log file = log_result_of_problem_data.txt Checking the basic file properties counting lines 11 BASIC INFORMATION 	output filename = re:	sult_of_problem_data.txt
pl column = 4 p2 column = 5 snplD column = csig (version - 0.2.9 Win32) is running with opts: -f = examples\testing.dat -o result_of_problem_data. log file = log_result_of_problem_data.txt Checking the basic file properties counting lines 11 BASIC INFORMATION 	p-value1 = p-value2 =	
snplD column = ' snplD column = ' snplD column = ' csig (version - 0.2.9 Win32) is running with opts:	p1 column = 4 p2 column = 5	
csig (version - 0.2.9 Win32) is running with opts:	snpID column =	
Checking the basic file properties counting lines 11 BASIC INFORMATION Variable Ualue Filename extension Path testing Filesize (hyte) 317 Eytes > File size (hyte) 28.82 <- Line size (hyte) 28.82 <- Last access time Fri Aug 8 00:10:37 200 Last inode chage time Thu Aug 7 23:59:15 200 Mode (permission) 666 PRINT FIRST 3 LINES OF examples testing.dat: SMP1D.Chr.Loc.p1.p2 rs34221 1 23432 0.4332 0.001 rs34222 1 23442 3.3e-5 0.001 rs34222 1 23442 3.3e-5 0.001 rs3422 1 23442 3.3e-5 0.001 SUMMARY OF ERROR REPORTING Total lines dropped: 8 Total lines kept: 3 LIST OF DROPPED LINES 1 (header) 4 5 7 8 9 10 11	sig (version - 0.2.9 Win3: -f = examples\test log file = log_res	2) is running with opts: ing.dat -o result_of_problem_data.cs ult_of_problem_data.txt
counting lines         11         BASIC INFORMATION         Wariable         Value         Filename stem         Filename extension       dat         Path       examples         > File size (byte)       317 Gytes         File size (byte)       28.82         Line size (byte)       28.82 Generation         Last access time       Fri Aug 8 00:10:37 200         Last access time       Fri Aug 8 00:10:37 200         Last inode chage time       Thu Aug 7 23:59:15 200         Last inode chage time       Thu Aug 7 23:59:15 200         Mode (permission)       666         SUMMARY OF examples/testing.dat:         SNPID.Chr.Loc.pl.p2       rs34221 1 23432 0.4332 0.001         rs34222 1 23442 3.3e-5 0.001         SUMMARY OF ERROR REPORTING         Intal lines dropped:         %       %         1 (header)         4       5         8       9         9       10         1 (header)	hecking the basic file pr	operties
11         Wariable         Value         Filename stem Path       Calue         File size       317 Bytes         Total number of lines > Line size       11 28.82 (- 28.82 Bytes         Line size       28.82 C- Line size         Last access time Last modify time Last mode chage time Mode (permission)       Fri Aug 8 00:10:37 200 Last inode chage time Thu Aug 7 23:59:15 200 Mode (permission)         PRINI FIRST 3 LINES OF examples testing.dat: SNPID.Chr.Loc.pl.p2 rs34221 1 23442 0.4332 0.001 rs34222 1 23442 3.3e-5 0.001       Colspan="2">Colspan="2"Colspan="2"Colspan="2"Colspan="2"Colsp	counting lines	
BASIC INFORMATION Variable Ualue Filename stem testing Filename extension dat Path texamples > File size (byte) 317 <- File size 317 Bytes Total number of lines 211 Line size (byte) 28.82 <- Line size (byte) 28.82 <- Line size 28.82 Bytes 28.82 <- Last access time Fri Aug 8 00:10:37 200 Last modify time Fri Aug 8 00:10:37 200 Last modie chage time Thu Aug 7 23:59:15 200 Mode (permission) 666 PRINT FIRST 3 LINES OF examples testing.dat: SNPID, Chr. Loc. p1, p2 rs34221 1 23432 0.4332 0.001 rs34222 1 23442 3.3e-5 0.001 SUMMARY OF ERROR REPORTING Total lines dropped: 8 Total lines kept: 3 LIST OF DROPPED LINES 1 (header) 4 5 7 8 9 10 11	11	
Variable     Value       Filename stem     testing       Filename extension     .dat       Path     examples       > File size (hyte)     317 (-       File size (hyte)     317 Bytes       Total number of lines     11       > Line size (hyte)     28.82 (-       Line size (hyte)     28.82 (-       Line size (hyte)     28.82 (-       Last access time     Fri Aug 8 00:10:37 200       Last modify time     Fri Aug 8 00:10:37 200       Last inode chage time     Thu Aug 7 23:59:15 200       Mode (permission)     666       PRINT FIRST 3 LINES OF examples/testing.dat:       SNPID.Chr.Loc.pl.p2       rs34221 1 23432 0.4332 0.0001       rs34222 1 23442 3.3e-5 0.001	BASIC INFORMATION	
Filename stem       .dat         Path       examples\         > File size (byte)       317 <-	Variable	Value
Filename extension Path       .dat examples         > File size (byte)       317 Gytes         File size (byte)       317 Bytes         In size (byte)       28.82 (-         Line size (byte)       28.82 Bytes         Last access time       Fri Aug 8 00:10:37 200         Last modify time       Fri Aug 8 00:10:37 200         Last modify time       Fri Aug 8 00:10:37 200         Last inode chage time       Thu Aug 7 23:59:15 200         Mode (permission)       666         PRINT FIRST 3 LINES OF examples testing.dat:         SNPID, Chr, Loc, p1, p2         rs34222 1 23442 3.3e-5 0.001         rs34222 1 23442 3.3e-5 0.001         SUMMARY OF ERROR REPORTING	Filename stem	testing
<pre>&gt; File size (byte) 317 &lt;- File size 317 Bytes Total number of lines 11 &gt; Line size (byte) 28.82 &lt;- Line size 28.82 (- Line size 28.82 (- Last access time Fri Aug 8 00:10:37 200 Last inode chage time Fri Aug 8 00:10:37 200 Last inode chage time Thu Aug 7 23:59:15 200 Mode (permission) 666 PRINT FIRST 3 LINES OF examples\testing.dat: SNPID.Chr.Loc.p1.p2 rs34221 1 23432 0.4332 0.001 rs34222 1 23442 3.3e-5 0.001 SUMMARY OF ERROR REPORTING Total lines dropped: 8 Total lines kept: 3 LIST OF DROPPED LINES 1 (header) 4 5 8 9 10 11</pre>	Filename extension Path	.dat examples∖
Total number of lines       11 28.82         Line size (byte)       28.82 Bytes         Last access time       Fri Aug 8 00:10:37 200         Last modify time       Fri Aug 8 00:10:37 200         Last inode chage time       Fri Aug 8 00:10:37 200         Last inode chage time       Fri Aug 8 00:10:37 200         Last inode chage time       Fri Aug 8 00:10:37 200         Mode (permission)       666         PRINT FIRST 3 LINES OF examples\testing.dat:         SNPID.Chr.Loc.pl.p2         rs34221 1 23432 0.4332 0.001         rs34222 1 23442 3.3e-5 0.001         rs34221 1 2:3432 0.4332 0.001         rs34222 1 2:3442 3.3e-5 0.001         rs34221 1 2:3432 0.4332 0.001         rs34222 1 2:3442 3.3e-5 0.001         SUMMARY OF ERROR REPORTING         1         Total lines dropped:       8         7       8         9       10         10       11	File size (byte) File size	317 <- 317 Bytes
<pre>&gt; Line size (byte) 28.82 &lt;- Line size 28.82 Bytes Last access time Fri Aug 8 00:10:37 200 Last inode chage time Fri Aug 8 00:10:37 200 Last inode chage time Thu Aug 7 23:59:15 200 Mode (permission) 666 </pre>	Total number of lines	. 11
Last access time         Fri Aug 8 00:10:37 200           Last modify time         Fri Aug 8 00:10:37 200           Last inode chage time         Thu Aug 7 23:59:15 200           Mode (permission)         666           PRINT FIRST 3 LINES OF examples\testing.dat:           SNPID.Chr.Loc.pl.p2           rs34221 1 23432 0.4332 0.001           rs34222 1 23442 3.3e-5 0.001           SUMMARY OF ERROR REPORTING           Total error found:         7           Total lines dropped:         8           Total lines kept:         3           LIST OF DROPPED LINES           1 (header)         4           5           7           8           9           10           11	Line size (byte) Line size	28.82 <- 28.82 Bytes
Last modify time         Fri Mug 8 00:10:37 200           Last inode chage time         Thu Aug 7 23:59:15 200           Mode (permission)         666           PRINT FIRST 3 LINES OF examples\testing.dat:           SNPID.Chr.Loc.p1.p2           rs34221 1 23432 0.4332 0.001           rs34222 1 23442 3.3e-5 0.001           SUMMARY OF ERROR REPORTING	Last access time	Ewi Aug 8 00-10-37 200
Last inode chage time Thu Aug ? 23:59:15 200 Mode (permission) 666 PRINT FIRST 3 LINES OF examples\testing.dat: SNPID.Chr.Loc.p1.p2 rs34221 1 23432 0.4332 0.001 rs34222 1 23442 3.3e-5 0.001 SUMMARY OF ERROR REPORTING Total error found: ? Total lines dropped: 8 Total lines kept: 3 LIST OF DROPPED LINES 1 (header) 4 5 7 8 9 10 11	Last modify time	Fri Aug 8 00:10:37 200
PRINT FIRST 3 LINES OF examples\testing.dat: SNPID.Chr.Loc.p1.p2 rs34221 1 23432 0.4332 0.001 rs34222 1 23442 3.3e-5 0.001 SUMMARY OF ERROR REPORTING Total error found: 7 Total lines dropped: 8 Total lines kept: 3 LIST OF DROPPED LINES 1 (header> 4 5 7 8 9 10 11	Last inode chage time Mode (nermission)	Thu Aug 7 23:59:15 200 666
SUMMARY OF ERROR REPORTING         Total error found:       ?         Total lines dropped:       8         Total lines kept:       3         LIST OF DROPPED LINES         1 (header)         4         5         ?         8         9         10         11	RINT FIRST 3 LINES OF exa SNPID,Chr,Loc,p1,p2 rs34221 1 23432 0.4 rs34222 1 23442 3.3	mples\testing.dat: 332 0.001 e-5 0.001
SUMMARY OF ERROR REPORTING Total error found: 7 Total lines dropped: 8 Total lines kept: 3 LIST OF DROPPED LINES 1 (header) 4 5 7 8 9 10 11		
Iotal lines dropped:       8         Total lines kept:       3	SUMMARY OF	ERROR REPORTING
LIST OF DROPPED LINES 	Iotal error found Total lines dropped Total lines kent	: 7 : 8 : 3
LIST OF DROPPED LINES 		
4 5 7 8 9 10 11		1 (beaden)
5 7 8 9 10 11		4
8 9 10 11		5
9 10 11		8
<u> </u>		10
		11

Here is the list of resulting files in C:  $\mbox{myWork}$  and the output from running "more" result file:

08/08/2008 08/08/2008 08/08/2008 08/08/2008 08/08/2008	12:15 AM 12:21 AM 12:15 AM 12:21 AM 4 File(s) 0 Dir(s)	1,809 log_output_CS.txt 2,456 log_result_of_problem_data.txt 109,456 output_CS.csv 356 result_of_problem_data.csv 114,077 bytes 96,182,865,920 bytes free
C:\myWork>m	ore result_of	problem_data.csv
SNPID,Chr,L	,oc,p1,p2,p-va	ue_IS, [=log10(p=value_IS>], p=value_TSM, [=log10(p=value_TSM>], TS=value, TSM=value
rs34221,1,2	,3432,0.4332,0	001.0.063133.1.199743.0.055579.1.255087.0.673600.0.691974
rs34222,1,2	,3442,3.3e-5,0	001.0.0000006.6.421607.0.000000.6.466709.0.999200.0.999237
rs34222,4,4	,3432,0.4332,0	005.0.065475.1.183922.0.058071.1.236042.0.667600.0.685146

### Description

This program is an implementation of the following methodology, which has been published:

Wang, J. and Shete, S. A Test for Genetic Association that Incorporates Information about Deviation from Hardy-Weinberg Proportions in Cases. Volume 83, Issue 1, 53-63, 26 June 2008.

# Methodology

For assessment of genetic association between single-nucleotide polymorphisms (SNPs) and disease status, the logistic regression model or generalized linear model is typically employed, and has been applied to detect a variety of disease-causing SNPs. However, the regression approaches do not integrate information that is available from other sources, such as departure from Hardy-Weinberg proportion in cases. The Hardy-Weinberg proportion is one of the most important principles in population genetics. Deviation from Hardy-Weinberg proportion among cases (patients) may provide additional evidence for the association between SNPs and diseases. Thus, testing for deviation from Hardy-Weinberg proportion could be another approach for genetic association studies.

We developed two approaches: (1) a mean-based tail strength (*TS*) measure and (2) a median-based tail strength (*TSM*) measure, to combine evidence from Hardy-Weinberg proportion and from regression approaches to perform the case-control association study. Both measures combined two very different hypothesis tests, the logistic-regression model and the test for deviation from Hardy-Weinberg proportion in cases, and allowed dependence between these two tests. The newly developed approaches are more powerful than the traditional association study approaches, achieving higher power than each individual test and maintaining good control over type I error probabilities. We derived the exact formulas for calculation of corresponding p-values. We also proposed an approach for estimating empirical p-values with the use of a resampling procedure. For the detail of the approaches, please refer to the study of Wang and Shete.

Note: According to the simulation studies in the paper, we can see that the empirical p-values from the permutation approach and the exact p-values from the exact formulas are very similar, and the exact formulas give slightly conservative p-values. Therefore, the exact p-values are considered satisfactory. The current version of this computer program only applies to calculation of the p-values using the exact formulas. The program for the permutation method as described in the paper is under construction, and will be available soon.

### **Detailed CSig Usage and Description**

#### USAGE:

csig	-option1	-option2 []
-v -h -m	-version -? -help -man	show version show this message show this message
-e	-example	making embedded example data
-у	-yes	combined with "-e" to give a silent support of overwriting any existing example data
-f   -p1 -p2	-file	input file 1st p-value (a real number) 2nd p-value (a real number)
-c1		column position of p-value1

	ਠ User Manual of csig.exe version 0.2.9	
Robert K. Yu (rkyu@mdanderson.org)		August 6, 2008
-c2	column position of p-value2	
-o   -outp	output filename Default is "output_CS.csv". In the exploration mode, this option will only affect log file name. User can use redirect to capture the output in a file	
-d   -decimal	an integer representing the decimal place in output. Effective numbers are 0 to 9 (but using 0 would be meaningless although csig will take.) Default is 6. Use any number greater than 9 csig will generate the maximum the current computer can express.	

#### SYNOPSIS:

I. CREATE example data files

csig -e y (default is "n")

This will generate 3 example data, storing in "examples" sub-directory.

Any value other than "y" will stop making examples.

If any example data exist from previous making, csig will pause a question for permission of replacement. A negative answer will abort the step and continue to use the current ones.

A silent permission can provide at initial starting of csig instead of waiting for response:

#### csig -e y -y y

II: EXPLORATION with a pair of p-values

#### csig -p1 0.002 -p2 1.2e-5

Note: the p-value in "dot-leading" format is also acceptable, e.g. ".002" for 0.002.

#### csig -p1 .002 -p2 0.03

or

csig -p1 0.002 -p2 1.2e-5 -d 3 > output.txt

III. BATCH MODE of list of p-value pairs in file,

csig -f sample data.dat -o result.txt -c1 4 -c2 5 [-d 99]

Note: any value greater than 9 for option -d allows csig to generate result values in max decimal places that the working computer can.

#### **EXPLANATION:**

I. Making Example Data :

Command "**csig** -**e y**" will lead to create 3 example data. And if csig detects prexisting examples, it'll prompt for user's decision to generate new set replacing the old ones.

csig chooses to terminate upon this execution to allow user to choose examples for the next run.

Except for "testing.dat", the other two are generated using random numbers. Thus, each run may have different value sets.

#### How are the random samples created?

"sample\_data2.dat", containing 1000 lines, uses double random for p-value 1 and p-value 2, i.e.

```
p-value1 = rand() * rand();
p-value2 = rand() * rand();
```

"sample\_data2.csv", containing 5000 lines, uses the following for p-value 1 and p-value 2, i.e.

> p-value1 = rand() \* 0.02; p-value2 = rand() \* rand() \* 0.1.

#### II. Data input :

1. For exploration, user just needs to give a pair of p-values, and/or output decimal place requirement, e.g.

#### csig -p1 0.023 -p2 1.05e-3 [-d 3]

If option "-d" is not provided, a default 6 will be used. Any number greater than 9 will direct csig to generate whatever the computer may be able to express.

2. For batch mode, user is required to provide the list of p-value pairs in a data file using either "," or space (including tab) or mixture of both as delimiter, and then tell csig where are the column positions for p-value1 and p-value2. One example is:

```
SNPID,Chr,Loc,p1,p2
rs34221 1 23432 0.4332 0.001
rs34222 1 23442 3.3e-5 0.001
rs34223 1 23452 1.56332 p2
rs34224 1 23462 0.5674
rs34221, 4 43432 0.4332 0.005
rs674222 5, 33442 1.2e-15 -3.051
rs994223, 61, 73452, 1.4332, p2
rs764224, z 83462,,0.1332
rs764224, x 83462, ,0.1332
rs764224, XY 83462 ,2,0.1332
```

where p1 and p2 are in columns 4 and 5, and the first 3 columns are optional SNP info. But notice that a header line is always expected. This dataset was created on purpose with mixed types of delimiter, i.e. ", " and ", " and " ". Also some "invalid" p-values were added to test whether csig will detect. Upon detected, the entire SNP line will be dropped but continues to process the next line.

To run csig with this dataset (if saved as "data.dat"), the commandline typing would look like this:

csig -f data.dat -c1 4 -c2 5 -o result.csv -d 5

#### III. Output

Robert K. Yu (rkyu@mdanderson.org)

And two output files will be generated, one result file "result.csv" and a log file "log\_result.txt". The log file will list total drops of 8 lines (including the header lines) for analysis and keep 3 valid lines for output results. In "result.csv", the result looks like this:

SNPID, Chr, Loc, p1, p2, p-value-TS, [-log10(p\_TS)], p-value-TSM, [-log10(p\_TSM)], TS-value, TSM-value rs34221, 1, 23432, 0.4332, 0.001, 0.06313, 1.19974, 0.05558, 1.25509, 0.67360, 0.69197 rs34222, 1, 23442, 3.3e-5, 0.001, 0.00000, 6.42161, 0.00000, 6.46671, 0.99920, 0.99924 rs34221, 4, 43432, 0.4332, 0.005, 0.06548, 1.18392, 0.05807, 1.23604, 0.66760, 0.68515

Interpretation of output

Columns	Column head	Explanation
1	SNPID	SNP rs number (user provided)
2	Chr	Chromosome number (user provided)
3	Loc	SNP location (user provided)
4	pl	p-value 1 (user provided)
5	p2	p-value 2 (user provided)
6	p-value-TS	p-value-TS
7	[-log10(p-value)	-TS)]
		negative log10 of p-value-TS
8	p-value-TSM	p-value-TSM
9	[-log10(p-value)	-TSM)]
		negative log10 of p-value-TSM
10	TS-value	Tail Strength Mean value
11	TSM-value	Tail Strength Median value

Columns 6 and 8 are the composite p-values, and columns 7 and 9 are the log transformed version of the two composite p-values.

Notice that this is a comma delimited file with "csv" file extension, and thus user may click on it and let it automatically opened up in Excel and have better view or even take advantage of tools in Excel, e.g. basic sorting and charting results.

In exploration mode, the results are output to screen only but a log file will be generated, which also contains result. Another way to collect several runs of the exploration is to redirect the screen output, e.g.

csig -p1 0.233 -p2 0.03 -d 3 >> result explore.txt

#### **DESCRIPTION:**

This program csig is to calculate the composite significance combining both logistic regression p-value and HWE testing p-value. Please refer to README.PDF for rationale and detailed information of this methodology, or email to authors below:

August 6, 2008

Dr. Sanjay Shete: sshete@mdanderson.org Dr. Jian Wang: jianwang@mdanderson.org

Reference:

Wang, J. and Shete, S. A Test for Genetic Association that Incorporates Information about Deviation from Hardy-Weinberg Proportions in Cases. Volume 83, Issue 1, 53-63, 26 June 2008.

csig was designed to facilitate the calculation of this method for both exploration purpose and batch mode processing need.

Note: According to the simulation studies in the paper, we can see that the empirical p-values from the permutation approach and the exact p-values from the exact formulas are very similar, and the exact formulas give slightly conservative p-values. Therefore, the exact p-values are considered satisfactory. The current version of this computer program only applies to calculation of the p-values using the exact formulas. The program for the permutation method as described in the paper is under construction, and will be available soon.

csig (version 0.2.8 win32) was designed as a analysis module in our ongoing programming effort to create a SNP Analytics, which is a GUI-based software. That software will integrate other tools including LD measurements, exact HWE test, haplotype estimation, SNP simulation, association tests, etc.

Please send us your email address and we will update our new tools, or tell us your suggestions.

Version: Release:	0.2.8 win32 08/04/2008
History:	
08/07/2008	<ul> <li>Modification</li> <li>Standardized prototyping of common subroutines</li> <li>Added another option, "-yes" to combine with "-example" for silent support of overwriting any existing example data</li> </ul>
08/04/2008	<pre>Development - Release for other platforms:  * Linux (x86-32), tested on     RedHat / Fedora 8     SUSE 10 (AMD x86-64) using 32-bit format  * Solaris - SPARC 32-bit, tested on     SunOS 5.9 Generic_122300-24         sun4u sparc SUNW, Sun-Fire-880     SunOS 5.10 Generic_127111-09         sun4u sparc SUNW, Sun-Fire-V490     SunOS 5.10 Generic_120011-14         sun4u sparc SUNW, Sun-Fire-480R - Modified output messages - Changed heading by using "[-log10 ()]"     instead of "neg_log10" Revised the help/manual message</pre>

( rla u a made na la rea e na area)		August C. 2008
(rkyu@mdanderson.org)	R	August 6, 2008
07/31/2008	<ul> <li>Fix</li> <li>allow leading "." for a real number in the input, e.g. ".001" for "0.001"</li> <li>change the name "Pmean" and "Pmedian" to "p-value-TS" and "p-value-TSM"</li> <li>in batch mode, the default log file won't take the same file extension as the one of input data file (as it did in the previous versions)</li> <li>output file will always take ".csv" to reflect its comma-delimited format and force</li> </ul>	
	to open in MS Excel (since this is a Windows	
	version)	
07/30/2008	Modifications	
	- embed example data into csig	
	Two of the three examples are now	
	generated with randomness (see	
	details in manual). - Windowg installer thug changes	
	accordingly; no link to example	
	files at start menu. Meanwhile,	
	improvement on invoking console	
	- Fixed some typos in PDF user	
	- Changed heading by using "neg log10"	
	instead of "-log10". The latter	
	would have to cause problem if user	
	has trouble to decipher the leading	
	- Added an alarm bell at the end of	
	csig. It'll ring only when it runs	
	over 2 seconds. This may be helpful	
07/29/2008	Modifications	
	- allow flexible column positions	
	of p-values	
	- allow user indication of output decimal place	
07/13/2008	Initial release.	
	single module calculating composite significance.	

Robert Yu at <u>rkyu@mdanderson.org</u>

# Appendix

### How to configure console window

The default setting of the Windows console window may not be well fit your commandline program, e.g. lower screen buffer size, etc. The set it wider, for example, you can follow the steps below (and you may choose to adjust other properties, too).

Upon the console window is up running, e.g. Start $\rightarrow$ run $\rightarrow$ cmd, or through launching "Composite Significance" program, do the following:

1. Right click the top title bar, and choose "Properties"



2. Select tab "Layout", and change the value at "Width" in "Screen Buffer Size". The default is 80.

🙉 "C:\WINDOW5\system32\cmd.exe" Properties 🛛 🛛 🔋 🗙					
Options Font Layout Cold	ns				
Options Font Layout Cold Window Preview	rs   Screen Buffer Size Width: Height: Window Size Window Position Left: Iop: I♥ Let system position	110 1 300 1 51 1 242 1 242 1 242 1 tion window			
	ОК	Cancel			

3. Upon click "OK", in the next dialog box, check "Save properties for future windows with the same title". Then "OK". Ready to go.



# How to quickly change directory to a desired destination in console window

At console window, using tab-key for self completion has accelerated the typing for changing directory to a deep sit folder. However, it still takes quite a bit typing and tab-key pressings. In particular, when the desired folder sits in another drive, it may ask for one or two more steps.

Here lists a simple way to accomplish the task.

If the initial directory is at C: drive when a console is just launched, and a desired folder deeply sits at G: drive, the following steps may make it "jump":

Step 1. With launched console and opened Windows Explore in your desired folder:

🔤 Command Prompt	🌀 Back 🔹 🐑 🐇 🥬 S	iearch 😥 Folders 📔	•	. 🗆 🗙
Microsoft Windows XP [Version 5.1.2600] (C) Copyright 1985-2001 Microsoft Corp.	Address 🛅 G:\Volume2\work\Composite	eSignificance080625\CSignifi	icance\release029 🛛 💌	🔁 Go 📃 📥
C:\Documents and Settings\Robert>_	File and Folder Tasks	Name 🔺	Size Type File Folder	Date Mod 8/7/2008
	Make a new folder Publish this folder to the Web Share this folder	· \$README.doc · CSignificance.pl □ CSignificance.pl.bak ■ README.doc	1 KB Microsoft 31 KB PL File 31 KB BAK File 563 KB Microsoft	8/7/2008 8/7/2008 8/7/2008 8/7/2008

Step 2. Highlight and copy the path at the "Address" bar in Windows Explorer. Note, the "Copy" step can be done either using right click on the highlighted path or pressing Ctl+C.

ack 🝷 🕥 🕤 🏂 🔎 S	earch 😥 Folders 📑	•							
😳 G:\Volume2\work\Composite	🗁 G:\Volume2\work\CompositeSignificance080625\CSignificance\release02 <sup>ol</sup>								
	Name 🔺	Size Ty	Undo						
and Folder Tasks 🔅	🚞 examples	File	Cut						
Make a new folder	👜 ~\$README.doc	1 KB Mic	Сору						
	😹 CSignificance.pl	31 KB PL	Paste						
Web	🖻 CSignificance.pl.bak	31 KB BA	Delete						
Share this folder	README.doc	563 KB Mid	Select All						

Step 3. At console prompt, type "cd /d ", and at the console title bar, right click and choose "Edit"  $\rightarrow$  "Paste".



Notice that switch "/d" for "cd" is critical when a directory change is attempted from one drive to another.

#### Step 4. Press Enter. Done

