Identification of SNP Targeted Pathways From Genome-wide Association Study (GWAS) Data

Burcu Bakir-Gungor (burcub@gatech.edu)
Sezerman Lab, Sabanci University
Osman Ugur Sezerman
Sabancı University

Method Article

Keywords: Genome-wide association study, GWAS, human complex disease, single nucleotide polymorphism, SNP, pathway, protein protein interaction network, SNP functional information

Posted Date: May 22nd, 2019

DOI: https://doi.org/10.21203/rs.2.1035/v2

Version of Record: A version of this preprint was published at Protocol Exchange on May 29th, 2012. See the published version at https://doi.org/10.1038/protex.2012.019.

Abstract

The identification of the variants that explain familial risk of a specific disease is important since it enables the development of genetic risk prediction tests, diagnosis tools and therapeutical applications. One possible reason of multifactorial diseases is the alterations in the activity of biological pathways, where a series of mutations occur in distinct genes. While each of these variations extends slightly the likelihood of having the disease, they work together to give birth to the perturbations in normal biological processes. We provide a protocol \(termed PANOGA, Pathway and Network Oriented GWAS \(Genome-wide association study) Analysis) to devise functionally important pathways through the identification of genes within these pathways, where these genes are targeted by single nucleotide polymorphisms \(SNPs) obtained from the GWAS analysis. Additionally, PANOGA helps to identify other disease related genes, not targeted by the SNPs, which are also located within these affected pathways. The program accepts tab delimited or excel file containing SNP rsIDs vs. genotypic p-values and is available at: http://akademik.bahcesehir.edu.tr/~bbgungor/panoga_protocol.zip

Introduction

In this protocol, starting with a list of SNPs found to be associated with a disease in a GWAS, we propose a novel methodology to determine disease related \(SNP targeted) pathways through the identification of SNP targeted genes within these pathways. Multiple factors act on complex diseases. Since each factor would have modest effect on the disease development mechanism, it is challenging to identify significant individual factors. One possible reason of multifactorial diseases is the alterations in the activity of biological pathways. In this method, we hypothesize that these factors are crippling similar pathways in individuals. That's why pathways have higher explanatory power towards understanding disease development mechanism. To this end, our methodology starts with functionalization of several significant SNPs to identify effected genes. We then map these genes to a protein-protein interaction \(PPI) network and determine the connected subnetworks targeted by the SNPs. Next, we find the KEGG pathways in these subnetworks and determine the significance of the modifications on these pathways. The pathways are ranked according to the significance scores and are referred as the SNP targeted pathways. The protocol is outlined in "Figure 1":http://www.nature.com/protocolexchange/system/uploads/2146/original/PANOGA_NatProtExch_Figure1.doc?1338222198

Equipment

A computer with Windows or Linux OS and internet access. EQUIPMENT SETUP Hardware requirements: We recommend a 1 GHz CPU or higher, a high-end graphics card, 500MB of available hard disk space, at least 1 GB of free physical RAM and a minimum screen resolution of 1,024 x 768. Java 2 platform: Standard Edition, version 5.0 or higher \(Java SE 5 or higher). \(http://java.sun.com/javase/downloads/index.jsp). Data files: This protocol begins with a GWAS dataset containing SNP rsIDs vs. genotypic p-values in a tab-delimited text file or excel file, as detailed in "Figure 2":http://www.nature.com/protocolexchange/system/uploads/2149/original/PANOGA_NatProtExch_Figure2.doc?1338222756 . As a result of the preprocessing step of PANOGA, four additional files in SPOT, F-SNP, SNPnexus and SNPinfo input file formats are created. A sample PANOGA input file and SPOT, F-SNP, SNPnexus and SNPinfo input files are made available in PANOGA_protocol/data/ as sample_panoga_input.txt, sample_spot_input.txt, sample_fsnp_input.txt, sample_snpnexus_input.txt, sample_snpinfo_input.txt). In addition to the sample GWAS

data, several additional data files are available for readers wishing to follow this protocol as a tutorial: humanPPI.sif: contains a protein–protein interaction \(PPI) network in a sif file format, as detailed in "Figure 2":http://www.nature.com/protocolexchange/system/uploads/2149/original/PANOGA_NatProtExch_Figure2.doc?1338222756 . This file illustrates the SIF, which offers a straightforward means to import networks into Cytoscape as text. sample_spot_fsnp_snpnexus.pvals: contains SPOT and F-SNP weighted p-values \(Pw-values) for each SNP associated gene. Each of these two Pw values combine functional information of a SNP and the genotypic p-value of a SNP, that is found to be significant in GWAS. Cytoscape: Cytoscape is an open source network data integration, analysis, and visualization platform. Subnetwork identification and functional enrichment steps of PANOGA protocol are realized by Cytoscape plugins. Hence, to follow PANOGA protocol, users need to install Cytoscape version 2.6.3 by on a local computer by following the steps in Box 2 of the Cytoscape paper, published in Nature protocols. Although Cytoscape has newer versions, jActiveModules and ClueGO plugins are verified to work in Cytoscape version 2.6.3. External tools: PANOGA utilizes four external web-servers to functionalize SNPs, i.e., SPOT, F-SNP, SNPnexus, SNPinfo; jActiveModules plugin of Cytoscape to identify sub-networks; ClueGO plugin 60 of Cytoscape 68 for functional enrichment of the identified sub-networks. All of these web-servers, programs and plugins are freely available for academic use. PANOGA: PANOGA is composed of nine consecutive steps, plus a preprocessing step. The preprocessing step of PANOGA is realized by a java script \(createpanogainput.jar). SNP functionalization steps of PANOGA are realized via sending the input files into four different web-servers \(SPOT, F-SNP, SNPnexus, SNPinfo). The subnetwork identification step of PANOGA is realized by the jActiveModules plugin of Cytoscape. The remaining steps are performed via running java executable programs. The above-mentioned jar files of PANOGA can be downloaded at: http://akademik.bahcesehir.edu.tr/~bbgungor/panoga_protocol.zip

Procedure

Install PANOGA 1) Set up necessary environment to run PANOGA \(as detailed in EQUIPMENT SETUP). 2) Download the PANOGA files at http://akademik.bahcesehir.edu.tr/~bbgungor/panoga_protocol.zip. Unzip the downloaded PANOGA_protocol.zip file and extract it. The executable jar files of PANOGA are found at: PANOGA_protocol/. Preprocess GWAS data 3) Pick a disease name for your project, which can be any disease name \(e.g., diabetes), not necessarily a standard OMIM disease name. In the following steps of PANOGA procedure, we will refer to this disease name as $DISEASE_NAME. CRITICAL STEP Do not use space in the $DISEASE_NAME since it will corrupt the further steps of PANOGA procedure. 4) Create a folder with your disease name under PANOGA_protocol/data/ and under PANOGA_protocol/out/ via typing the following commands: >cd PANOGA_protocol/data >mkdir $DISEASE_NAME >cd ../out >mkdir $DISEASE_NAME >cd .. Replace $DISEASE_NAME above with the disease name that you specified in Step 3. 5) Format GWAS results input file following the instructions in "Figure 2":http://www.nature.com/protocolexchange/system/uploads/2149/original/PANOGA_NatProtExch_Figure2.doc?1338222756 , and save this file under PANOGA_protocol/data/$DISEASE_NAME/ using any input file name. e.g., PANOGA_protocol/data/diabetes/diabetes_panoga_input.txt or bipolar_gwas_result.xls. sample_panoga_input.txt file is also provided under: PANOGA_protocol/data/sample/. 6) Run the java script “createpanogainput.jar” to create four separate input files that will be used in SNP to gene assignment and SNP functionalization steps of PANOGA: Replace $INPUT_FILE_NAME with your input file name, e.g. \(diabetes_panoga_input.txt), $DISEASE_NAME with your disease name and $PVALUE_THRESHOLD with genotypic p-value threshold that you would like to use to restrict your SNPs based on their significance for disease. The default $PVALUE_THRESHOLD

is 0.05. >java -jar createpanogainput.jar $INPUT_FILE_NAME $DISEASE_NAME $PVALUE_THRESHOLD e.g. java -jar createpanogainput.jar sample_panoga_input.txt sample 0.05 This run generates $DISEASE_NAME_spot_input.txt, $DISEASE_NAME_fsnp_input.txt, $DISEASE_NAME_snpnexus_input.txt, $DISEASE_NAME_snpinfo_input.txt files under PANOGA_protocol/data/$DISEASE_NAME. CRITICAL STEP Using an input filename with an extension other than .txt or .xls interferes this step. Assign SNPs to Genes 7) PANOGA procedure uses SPOT webserver 49 to assign SNPs to genes. Go to the SPOT webserver at: https://spot.cgsmd.isi.edu/submit.php. 8) Click into “Upload SNP File” button; select SPOT input file, i.e. $DISEASE_NAME_spot_input.txt. 9) Change “Maximum SNPs to output:” parameter to 50,000 in SPOT webserver. 10) If your $PVALUE_THRESHOLD \(from Step 6) is different than 0.05, change it in the “p-value threshold: “ parameter of SPOT webserver. 11) Under “Linkage Disequilibrium \(LD) options” select the appropriate HAPMAP sample among the available options in SPOT webserver. 12) Click into “Run” button and download the result under “Primary Results” section. Save the SPOT output as Tab-delimited file under PANOGA_protocol/data/ $DISEASE_NAME/$DISEASE_NAME_spot_output.txt. 13) At this step, the users need to choose one of the following two options: option A to proceed with the full PANOGA procedure, including network oriented stages and functional information of SNPs; option B to proceed with only pathway oriented steps of PANOGA procedure. We highly recommend the users to follow the full PANOGA procedure \(option A). \(A) Proceed with the full PANOGA procedure Continue with Step 14. \(B) Proceed with only pathway oriented steps of PANOGA procedure \(i) Run the java script “parsespotoutput.jar” to get a list of gene symbols assigned into typed SNPs. >java -jar parsespotoutput.jar $DISEASE_NAME This run will create the gene symbol file \ ($DISEASE_NAME_partial_panoga_gene_ symbols.txt) under PANOGA_procedure/ClueGO/data/ and $DISEASE_NAME_ partial_panoga_gene2snp.txt file under PANOGA_procedure/data/ $DISEASE_NAME/. \(ii)Type the following command to perform functional enrichment of identified gene symbols: >cd ClueGO Replace $DISEASE_NAME below with the disease name that you specified in Step 3. >java -jar ClueGO_v1.4.command-line.jar -props clueGO.props -file1 data\$DISEASE_NAME_partial_panoga_gene_symbols.txt -analysis-name $DISEASE_NAME_partial_panoga -out out At the end of this step, enrichment results of the gene symbols are saved under PANOGA_procedure/ClueGO/out/ \(iii) Run the java script “analyzecluegooutput.jar” to create SNP targeted pathway lists and gene list for identified SNP targeted pathways. >cd .. >java –jar analyzecluegooutput.jar $DISEASE_NAME At the end of this step pathway based lists and gene list are created as explained in the “Anticipated Results” section and “PANOGA’s Application to Human Complex Diseases” subsection of Introduction section. Install Cytoscape and its plugins 14) Install Cytoscape version 2.6.3 by following its installation guide 69. Follow Cytoscape installation instructions to get the executable file. CRITICAL STEP Although Cytoscape has newer versions, jActiveModules and ClueGO plugins are verified to work in Cytoscape version 2.6.3. 15) Install jActiveModules and ClueGO version 1.4 plugins of Cytoscape. These plugins should be installed into Cytoscape_v2.6.3/plugins/ using the following options: option A to install jActiveModules plugin; option B to install ClueGO version 1.4 plugin: \(A) Installing jActiveModules plugin jActiveModules plugin is used to identify active sub-networks. Copy jActiveModules plugin from: PANOGA_protocol/EXTERNAL_TOOLS/jActiveModules.jar and save under Cytoscape_v2.6.3/plugins/. \(B) Installing ClueGO version 1.4 plugin \(i) ClueGO plugin is used in the functional enrichment step of PANOGA. Copy .cluegoplugin, provided under PANOGA_protocol/ClueGO/ into the home directory of the user. \(ii) Obtain ClueGO licence from its website \(http://www.ici.upmc.fr/cluego/cluegoLicense.shtml) and save .lf file under home/.cluegoplugin/License/.l/ and .lcf file under home/.cluegoplugin/License/.lc/. CRITICAL STEP Before running PANOGA, ensure that Cytoscape, jActiveModules and ClueGO plugins are working properly. Obtain Functional Information of SNPs 16) PANOGA procedure utilizes SPOT 49, F-SNP 51, SNPnexus 73 and SNPinfo

50 webservers to functionalize SNPs. SNP functional information through SPOT web-server 49 is already obtained in the previous step while assigning SNPs to genes. Run “runfsnp.jar” to obtain SNP functional information from F-SNP webserver 51: Replace $DISEASE_NAME with the disease name that you specified in Step 3. >java -jar runfsnp.jar $DISEASE_NAME This step will save the F-SNP output into PANOGA_procedure/data/$DISEASE_NAME/ $DISEASE_NAME_fsnp_output.txt. 17) Go to the SNPnexus webserver at: http://www.snp-nexus.org/. Under “Batch Query” option, Browse SNPnexus input file, i.e. $DISEASE_NAME_snpnexus_input.txt. 18) Under “Annotation Categories”-> “Regulatory Elements”, select ”Conserved Transcription Factor Binding Sites \(TFBS)” option and click “Run” button. 19) Download the result under “Regulatory Elements” section via clicking into TXT icon. Save the SNPnexus output as text file under PANOGA_procedure/data/$DISEASE_NAME/$DISEASE_NAME_snpnexus_output.txt. 20) Go to the SNPinfo webserver at: http://snpinfo.niehs.nih.gov/snpfunc.htm. Browse and upload SNPinfo input file, i.e. $DISEASE_NAME_snpinfo_input.txt. 21) Click “Submit” button and download the results via clicking into “Export To Excel” button under “SNP Function Prediction Results”. Save the SNPInfo output as csv file under PANOGA_procedure/data/$DISEASE_NAME/$DISEASE_NAME_snpinfo_output.csv. Prepare the Gene Attributes data 22) PANOGA needs the attributes file \(in .pvals format) to identify the sub-networks \(using jActive Modules plugin 59). This file has two weighted P-values \(SPOT Pw and F-SNP Pw values) for each gene, where the weighted P-value combines the genotypic p-value of a SNP with the functional information of a SNP that is associated with the gene. The following steps of the PANOGA procedure will create an attributes file similar to the provided sample_panoga_spot_fsnp.pvals file at PANOGA_procedure/. Run “combinespotfsnp.jar” to combine SPOT and F-SNP output files: Replace $DISEASE_NAME with the disease name that you specified in Step 3. >java -jar combinespotfsnp.jar $DISEASE_NAME 23) Run “incorporatesnpnexus.jar” to incorporate functional information from SNPnexus. Replace $DISEASE_NAME with the disease name that you specified in Step 3. >java -jar incorporatesnpnexus.jar $DISEASE_NAME This run will create the gene attributes file \ ($DISEASE_NAME_spot_fsnp_snpnexus.pvals) under PANOGA_procedure/data/$DISEASE_NAME/. Obtain network data 24) Decide which human protein-protein interaction \(PPI) dataset you would like to use as your initial network—follow option A to use the default human PPI network or option B to use a customized PPI network. \(A) Using the default human PPI network A user can work with the default human PPI network supplied in the PANOGA installation package. The default human PPI network is available in sif format in: PANOGA_protocol/data/humanPPI.sif. \(B) Using another PPI network A user can work with their own human PPI network. Since Cytoscape 68 accepts networks in many different file formats \(e.g., .sif, .gml, .xgmml, .xls, SBML, BioPAX, PSI-MI.), the user has the option to choose the network that they want to work with. Load network data 25) Start Cytoscape via following option A for Windows users, option B for Linux users. \(A) Windows Users Run Cytoscape.exe. \(B) Linux Users Run ./cytoscape.sh. 26) Decide how you would like to load network data into Cytoscape. Cytoscape allows to import networks from a local or remote computer, or from Web Services—follow option A to import a network file from a local computer, option B from a remote computer or option C to use Web Services. We recommend PANOGA users to follow option A to load network data. \(A) Loading the default human PPI network from a local computer \(i) Assemble your network data into a SIF file, as described in "Figure 2":http://www.nature.com/protocolexchange/system/uploads/2149/original/PANOGA_NatProtExch_Figure2.doc?1338222756 \(ii) Import human PPI network using File->Import->Network commands of Cytoscape. The user is free to load any human PPI network, as long as the official HUGO gene symbols are used as node identifiers. A sample human PPI network is also provided at: PANOGA_procedure/data/humanPPI.sif. \(B) Loading a PPI network from a remote computer Follow the procedure described at: http://wiki.cytoscape.org/Cytoscape_User_Manual/#Cytoscape_User_Manual .2BAC8-

Creating_Networks.Load_Networks_from_a_Remote_Computer_.28URL_import.29 \(C) Loading a PPI network using Web Services Follow the procedure described at: http://wiki.cytoscape.org/Cytoscape_User_Manual/ImportingNetworksFromW ebServices Import gene attributes 27) Assign values \(two attributes for each identified gene) to nodes \(genes) using File->Import->Attribute/Expression Matrix commands of Cytoscape and selecting the gene attributes file \($DISEASE_NAME_spot_fsnp_snpnexus.pvals) that is created in Step 23. A sample gene attributes file \(sample_spot_fsnp_snpnexus.pvals) is also provided at PANOGA_procedure/data/sample/. Identify sub-networks 28) Start jActiveModules plugin from Cytoscape->Plugins->jActiveModules. 29) Select SPOTPvaluesig and FSScorePvaluesig from Expression Attributes panel. 30) In the General Parameters panel, set “Number of Modules” parameter as 1000. “Overlap Threshold” parameter defines max percent of overlap between any two identified subnetworks. The default value used in PANOGA_protocol is 0.5. 31) Click “Find Modules” to identify active sub-networks. 32) Save the result as text file into PANOGA_procedure/data/$DISEASE_NAME/$DISEASE_NAME_jactivemodules_output.txt via clicking into “Save All Results” button on “Results Panel”. Replace $DISEASE_NAME with the disease name that you specified in Step 3. Parse jActiveModules output 33) Create a folder with your disease name under PANOGA_protocol/ClueGO/data/ and under PANOGA_protocol/ClueGO/out/ via typing the following commands: >cd ClueGO/data >mkdir $DISEASE_NAME >cd ../out >mkdir $DISEASE_NAME >cd ../.. Replace $DISEASE_NAME above with the disease name that you specified in Step 3. 34) Run “parsejactivemodulesoutput.jar” to create individual files containing gene symbols for each identified sub-network: Replace $DISEASE_NAME with the disease name that you specified in Step 3. >java -jar parsejactivemodulesoutput.jar $DISEASE_NAME At the end of this step, for the sub-networks with scores higher than 3, individual files containing gene symbols are saved under PANOGA_procedure/ClueGO/data/$DISEASE_NAME/ and the number of subnetworks created is printed on the screen. Functional enrichment of subnetworks 35) Decide which pathway resource you would like to use for the functional enrichment of the identified subnetworks. ClueGO 60 assigns a set of genes into KEGG 61 or BioCarta pathways—follow option A to assign genes into KEGG pathways, option B to assign genes into Biocarta pathways. \(A) Identifying KEGG pathways Use the clueGO.props file provided under PANOGA_procedure/ClueGO/. In order to identify KEGG pathways, make sure that under the “Select Ontologies” title “SelectedOntologySources=KEGG_14.03.2012” in the ClueGO properties file \(clueGO.props). \(B) Identifying BioCarta pathways In order to identify BioCarta pathways, under the “Select Ontologies” title change “SelectedOntologySources = REACTOME_BioCarta_07.04.2011” in the ClueGO properties file \(PANOGA_procedure/ClueGO/clueGO.props). 36) At this step, the users need to choose one of the following two options, depending on their operating systems: Windows Users, follow option A; Linux Users, follow option B. For both options, replace $DISEASE_NAME with the disease name that you specified in Step 3, $NUMBER_OF_SUBNETWORKS with the number of subnetworks, as created in Step 34. Type the following command to perform functional enrichment for each of the identified sub-networks using the clueGO.props file created in Step 35: \(A) Windows Users >java –jar functionalenrichment.jar $DISEASE_NAME $NUMBER_OF_SUBNETWORKS \(B) Linux Users >./functionalenrichment.sh $DISEASE_NAME $NUMBER_OF_SUBNETWORKS \($OPTIONAL_JAVA_PATH) If java is installed as root user, skip $OPTIONAL_JAVA_PATH and run as: e.g. ./functionalenrichment.sh diabetes 508 If java is already installed on a different path, specify $OPTIONAL_JAVA_PATH and run as: e.g. ./functionalenrichment.sh diabetes 508 ../../jre1.7.0_04/bin At the end of this step, enrichment results of each of the identified sub-networks are saved under PANOGA_procedure/ClueGO/out/$DISEASE_NAME/ for both options. Combine functional enrichment results 37) Run the java script “combinesubnetworkpathways.jar” to create SNP targeted pathway lists and gene list for identified SNP targeted pathways. Replace $DISEASE_NAME with the disease name that you specified in

Step 3, $NUMBER_OF_SUBNETWORKS with the number of subnetworks, as created in Step 34. >java –jar combinesubnetworkpathways.jar $DISEASE_NAME $NUMBER_OF_SUBNETWORKS At the end of this step pathway based lists and gene list are created as explained in the “Anticipated Results” section and “PANOGA’s Application to Human Complex Diseases” subsection of Introduction section. Visualize SNP targeted genes in a KEGG pathway map 38) Create a directory under PANOGA_protocol/out/ to store gene attribute files for each pathway, via typing the following command: >cd out/KeggPathwayMapGeneAttributeFiles >mkdir $DISEASE_NAME >cd ../.. Replace $DISEASE_NAME above with the disease name that you specified in Step 3. 39) Run the java script “createattributesforpathwaymap.jar” to create a gene attributes file for each identified pathway, which will be used in the next step to customize KEGG pathway maps. Each pathway specific file contain identified gene symbols and color specifications depending on the number of SNP targeted genes per base pair. Replace $DISEASE_NAME with the disease name that you specified in Step 3. >java –jar createattributesforpathwaymap.jar $DISEASE_NAME At the end of this step, gene attribute file for each of the identified sub-networks are saved under PANOGA_protocol/out/KeggPathwayMapGeneAttributeFiles/$DISEASE_NAME 40) Color SNP targeted genes for the pathway of interest using the KEGG Mapper – Color Pathway tool available at: http://www.genome.jp/kegg/tool/map_pathway3.html. 41) Type “hsa” followed by the KEGG Term ID for the pathway of interest to the “Select KEGG pathway map:” field. KEGG Term IDs of the pathways can be obtained from the first column of the $DISEASE_NAME_pathways_subnetwork_genes.csv file under PANOGA_procedure/out/$DISEASE_NAME/. 42) Browse gene attribute file created in Step 39 for the pathway of interest. 43) Hit “Execute” button. KEGG Mapper – Color Pathway tool 61 generates a customized pathway map, where the SNP targeted genes are colored based on the number of SNPs per base pair.

Timing

The time required to execute this protocol is most strongly related with the time required to run the jActiveModules plugin 59 of Cytoscape 68 to identify subnetworks and to obtain functional information of SNPs from F-SNP webserver 51. With a 1.5 Mbps ADSL connection under favorable operating conditions, it is reported to take approximately 9 min to download Cytoscape 69. For 30,000 SNPs, it takes approximately 40 secs to get functional information from SPOT webserver 49; 20 secs from SNPinfo webserver 50; 10 mins from SNPnexus webserver 73; and 3 hours from F-SNP webserver 51. On a PC with 500 GB of memory, loading a human PPI network of 10,000 nodes and 80,000 edges requires approximately 10 secs; importing a gene attribute file with 4,000 genes and 2 separate attributes for each gene requires 4secs. On such a network, the subnetwork identification step of PANOGA takes approximately 4 hours once jActiveModules plugin 59 is used with two attributes for each node. On a PC with the same configurations, ClueGO plugin requires approximately 45 mins for functional enrichment of ~500 subnetworks. The running times of each of the unspecified jar programs take less than 5 mins. Overall, an experienced user can execute the full protocol described within 9 hours.

Troubleshooting

Out-of-memory errors Subnetwork identification step \(Step 31) of PANOGA procedure is realized by jActiveModules plugin 59 of Cytoscape 68. While Cytoscape is working with big networks \(~10,000 nodes) memory problems might appear. If such a case happens in Step 31 of PANOGA procedure, Cytoscape will display an error message such as a Java Null Pointer Exception. You can address this problem by freeing memory via deleting big and unnecessary networks or via closing other applications that are running on the same computer

69. Alternatively, you can increase the memory reserved for Cytoscape via changing the following lines in the Cytoscape.vmoptions file, which is located under .cytoscape folder under your home directory: -Xmx20G -Xss30M Further instructions on how to increase memory space for Cytoscape are available at: http://www.cytoscape.org/cgi-bin/moin.cgi/How_to_increase_memory_for_Cytoscape.

Anticipated Results

PANOGA will create pathway and gene tables with several features in .csv format \(comma separated values) as shown in "Table 1":http://www.nature.com/protocolexchange/system/uploads/2151/original/PANOGA_NatProtExch_Table1.doc?1338222879 "Table 2":http://www.nature.com/protocolexchange/system/uploads/2152/original/PANOGA_NatProtExch_Table2.doc?1338222913 "Table 3":http://www.nature.com/protocolexchange/system/uploads/2153/original/PANOGA_NatProtExch_Table3.doc?1338222943 and "Table 4":http://www.nature.com/protocolexchange/system/uploads/2154/original/PANOGA_NatProtExch_Table4.doc?1338222973 , respectively. The files can be opened by Microsoft Excel or Open Office and displayed as spreadsheets. Each row of the pathway spreadsheet will correspond to the features of the identified pathway, i.e., KEGG term, KEGG term ID, p-value, rank, number of times found significant for different subnetworks, number of SNP targeted genes, number of typed SNPs in GWAS that are associated with the genes as part of the pathway under study, number of regulatory SNPs which are also found significant in GWAS, SNP targeted genes and their SNP counts. Gene table file, as shown in "Table 4":http://www.nature.com/protocolexchange/system/uploads/2154/original/PANOGA_NatProtExch_Table4.doc?1338222973 , will include different features of the genes that are found as part of the identified pathways. While some of these genes are SNP targeted genes, some others are identified as the neighbours of SNP targeted genes within the generated sub-networks. Each row of the gene spreadsheet will correspond to a gene symbol, entrez gene ID, number of times found in subnetwork, number of associated pathways, list of associated pathways, number of typed SNPs in GWAS, functional information of the typed SNPs in GWAS, SNP regulatory potential, number of regulatory SNPs. If the number of typed SNPs in GWAS is zero, that means this gene is identified through neighbour effect. PANOGA will also create customized KEGG pathway map for pathway under study, as shown in "Figure 3":http://www.nature.com/protocolexchange/system/uploads/2150/original/PANOGA_NatProtExch_Figure3.png?1338222828 . In this map, the shade of red color in genes map to the number of targeted SNPs \(typed in the GWAS of RA), per base pair of the gene. Hence, this pathway map will help the users to visualize affected genes along different routes within the pathway map.

References

Bakir-Gungor, B. & Sezerman, O. U. A New Methodology to Associate SNPs with Human Diseases According to Their Pathway Related Context. PLoS One 6, doi:ARTN e26277 DOI 10.1371/journal.pone.0026277 \(2011). Bindea, G. et al. ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. Bioinformatics 25, 1091-1093, doi:10.1093/bioinformatics/btp101 \(2009). Chelala, C., Khan, A. & Lemoine, N. R. SNPnexus: a web database for functional annotation of newly discovered and public domain single nucleotide polymorphisms. Bioinformatics 25, 655-661, doi:DOI 10.1093/bioinformatics/btn653 \

(2009). Cline, M. S. et al. Integration of biological networks and gene expression data using Cytoscape. Nat Protoc 2, 2366-2382, doi:DOI 10.1038/nprot.2007.324 \(2007) Lee, P. H. & Shatkay, H. F-SNP: computationally predicted functional SNPs for disease association studies. Nucleic Acids Res 36, D820-D824, doi:Doi 10.1093/Nar/Gkm904 \(2008). Saccone, S. F. et al. SPOT: a web-based tool for using biological databases to prioritize SNPs after a genome-wide association study. Nucleic Acids Res 38, W201-W209, doi:Doi 10.1093/Nar/Gkq513 \(2010). Shannon, P. et al. Cytoscape: A software environment for integrated models of biomolecular interaction networks. Genome Res 13, 2498-2504, doi:Doi 10.1101/Gr.1239303 \(2003). Xu, Z. L. & Taylor, J. A. SNPinfo: integrating GWAS and candidate gene information into functional SNP selection for genetic association studies. Nucleic Acids Res 37, W600-W605, doi:Doi 10.1093/Nar/Gkp290 \(2009).

Figures

Identification of SNP Targeted Pathways From Genome-wide Association Study (GWAS) Data

Burcu Bakir-Gungor[1-3,*], Osman Ugur Sezerman[3]

[1] Department of Genetics Bioinformatics, Faculty of Arts and Sciences, Bahcesehir University, Istanbul, TURKEY. [2] Department of Computer Engineering, Faculty of Engineering, Bahcesehir University, Istanbul, TURKEY. [3] Biological Sciences and Bioengineering, Faculty of Engineering and Natural Sciences, Sabanci University, Istanbul, TURKEY.

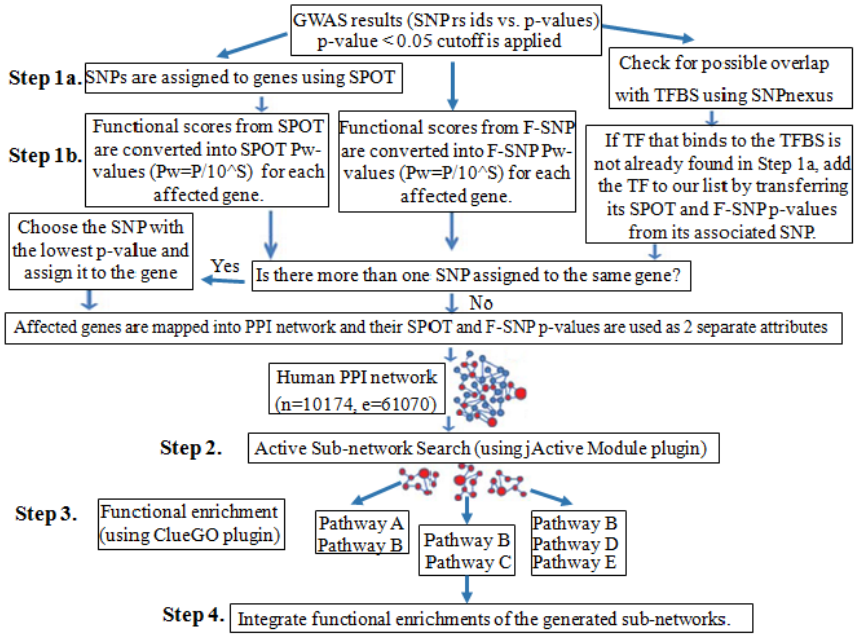
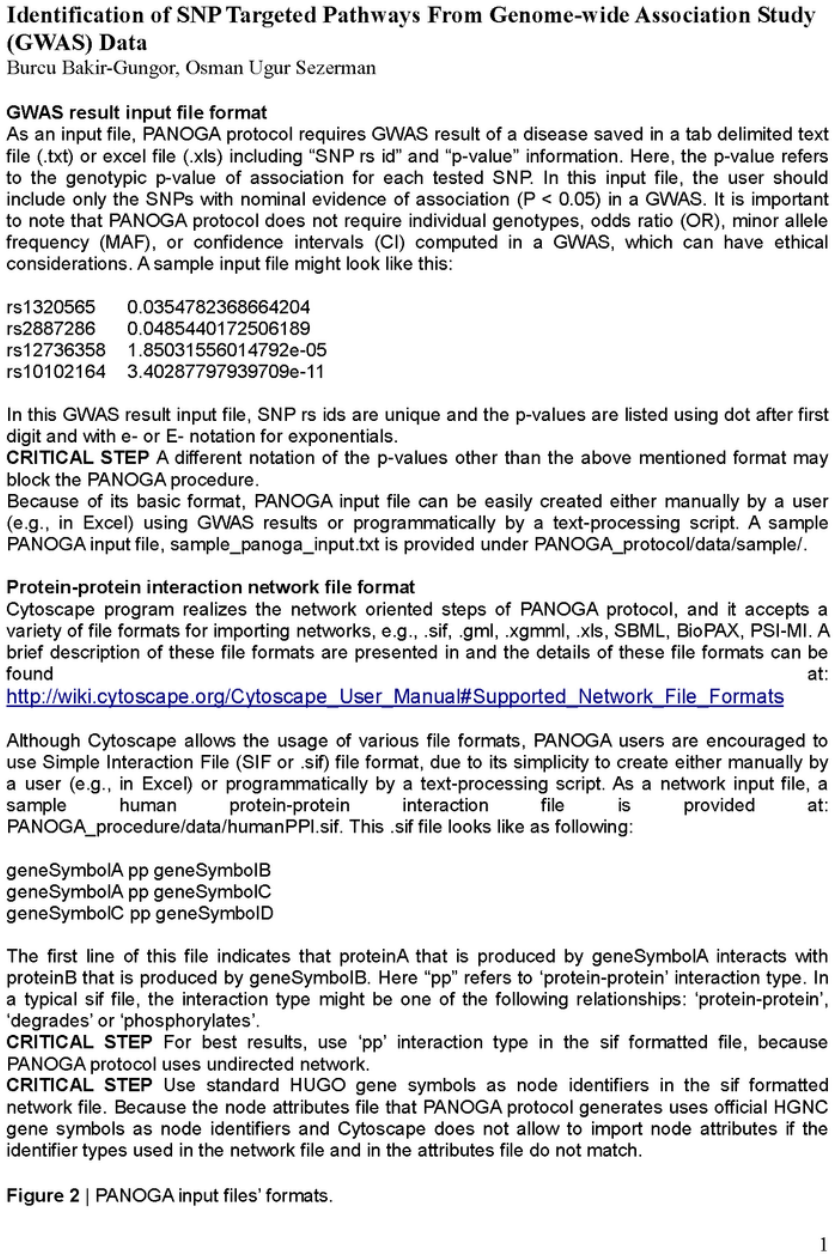


Figure 1: Outline of PANOGA protocol. In Step 1, a gene-wise Pw-value for association with disease was computed by integrating functional information. In Step 2, significant Pw-values were loaded as two separate attributes of the genes in a PPI network and visualized using Cytoscape. At this step, active sub-networks of interacting gene products that were also associated with the disease, are identified using jActive Modules plugin. In Step 3, genes in an identified active sub-network were tested whether they are part of functionally important KEGG pathways. Step 4 integrates the functional enrichments of the generated sub-networks.

1

Figure 1

Outline of PANOGA protocol. In Step 1, a gene-wise Pw-value for association with disease was computed by integrating functional information. In Step 2, significant Pw-values were loaded as two separate attributes of the genes in a PPI network and visualized using Cytoscape. At this step, active sub-networks of interacting gene products that were also associated with the disease, are identified using jActive Modules plugin. In Step 3, genes in an identified active sub-network were tested whether they are part of functionally important KEGG pathways. Step 4 integrates the functional enrichments of the generated sub-networks.



Identification of SNP Targeted Pathways From Genome-wide Association Study (GWAS) Data

Burcu Bakir-Gungor, Osman Ugur Sezerman

GWAS result input file format

As an input file, PANOGA protocol requires GWAS result of a disease saved in a tab delimited text file (.txt) or excel file (.xls) including "SNP rs id" and "p-value" information. Here, the p-value refers to the genotypic p-value of association for each tested SNP. In this input file, the user should include only the SNPs with nominal evidence of association ($P < 0.05$) in a GWAS. It is important to note that PANOGA protocol does not require individual genotypes, odds ratio (OR), minor allele frequency (MAF), or confidence intervals (CI) computed in a GWAS, which can have ethical considerations. A sample input file might look like this:

rs1320565 0.0354782368664204
rs2887286 0.0485440172506189
rs12736358 1.85031556014792e-05
rs10102164 3.40287797939709e-11

In this GWAS result input file, SNP rs ids are unique and the p-values are listed using dot after first digit and with e- or E- notation for exponentials.
CRITICAL STEP A different notation of the p-values other than the above mentioned format may block the PANOGA procedure.
Because of its basic format, PANOGA input file can be easily created either manually by a user (e.g., in Excel) using GWAS results or programmatically by a text-processing script. A sample PANOGA input file, sample_panoga_input.txt is provided under PANOGA_protocol/data/sample/.

Protein-protein interaction network file format

Cytoscape program realizes the network oriented steps of PANOGA protocol, and it accepts a variety of file formats for importing networks, e.g., .sif, .gml, .xgmml, .xls, SBML, BioPAX, PSI-MI. A brief description of these file formats are presented in and the details of these file formats can be found at: http://wiki.cytoscape.org/Cytoscape_User_Manual#Supported_Network_File_Formats

Although Cytoscape allows the usage of various file formats, PANOGA users are encouraged to use Simple Interaction File (SIF or .sif) file format, due to its simplicity to create either manually by a user (e.g., in Excel) or programmatically by a text-processing script. As a network input file, a sample human protein-protein interaction file is provided at: PANOGA_procedure/data/humanPPI.sif. This .sif file looks like as following:

geneSymbolA pp geneSymbolB
geneSymbolA pp geneSymbolC
geneSymbolC pp geneSymbolD

The first line of this file indicates that proteinA that is produced by geneSymbolA interacts with proteinB that is produced by geneSymbolB. Here "pp" refers to 'protein-protein' interaction type. In a typical sif file, the interaction type might be one of the following relationships: 'protein-protein', 'degrades' or 'phosphorylates'.
CRITICAL STEP For best results, use 'pp' interaction type in the sif formatted file, because PANOGA protocol uses undirected network.
CRITICAL STEP Use standard HUGO gene symbols as node identifiers in the sif formatted network file. Because the node attributes file that PANOGA protocol generates uses official HGNC gene symbols as node identifiers and Cytoscape does not allow to import node attributes if the identifier types used in the network file and in the attributes file do not match.

Figure 2 | PANOGA input files' formats.

1

Figure 2

PANOGA input files' formats GWAS result and Protein-protein interaction network file formats, required by PANOGA protocol.

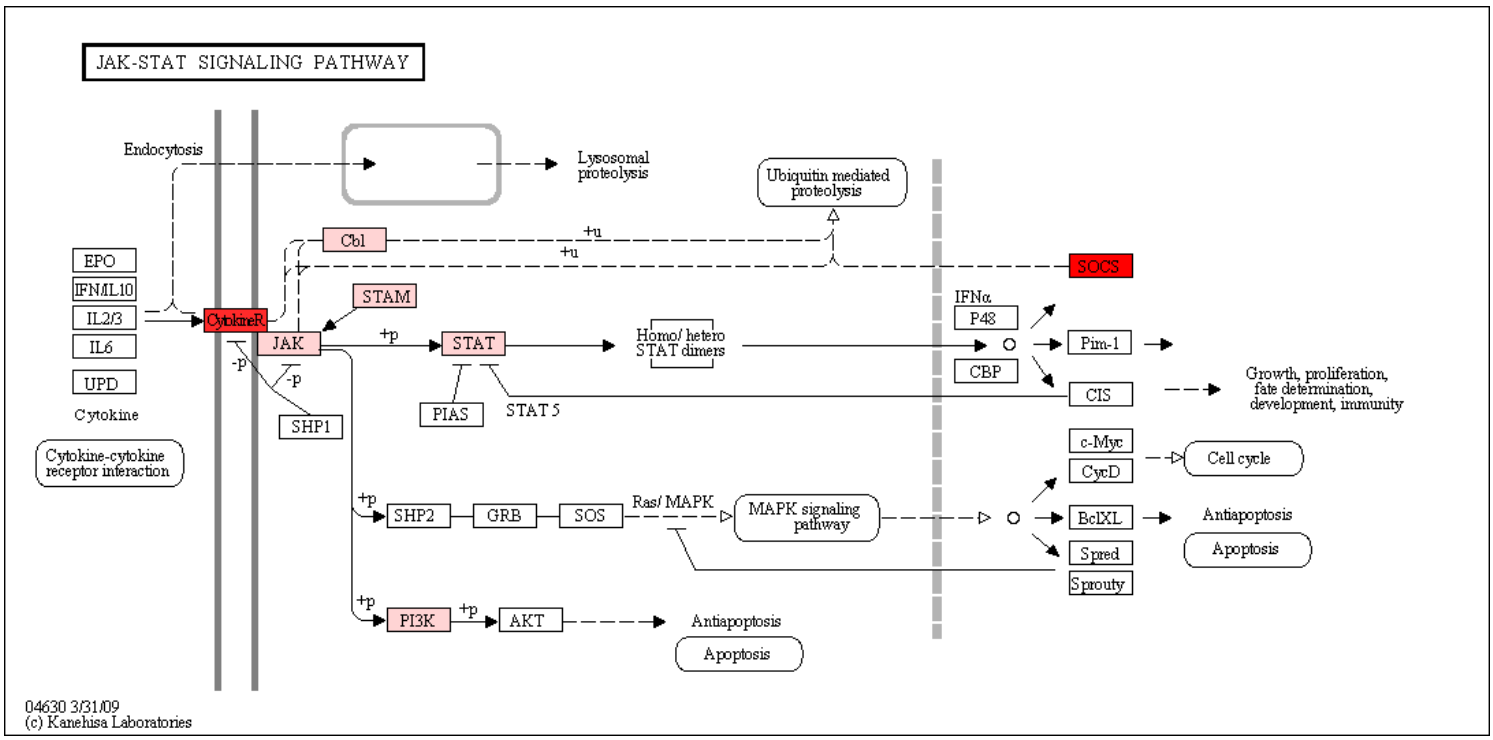


Figure 3

Customized KEGG pathway map of an identified SNP targeted pathway. Customized KEGG pathway map for JAK-STAT signaling pathway. The shade of red color in genes map to the number of targeted SNPs (typed in the GWAS), per base pair of the gene. Red refers to the highest targeted gene, whereas white refers to a gene product, not targeted by the SNPs.

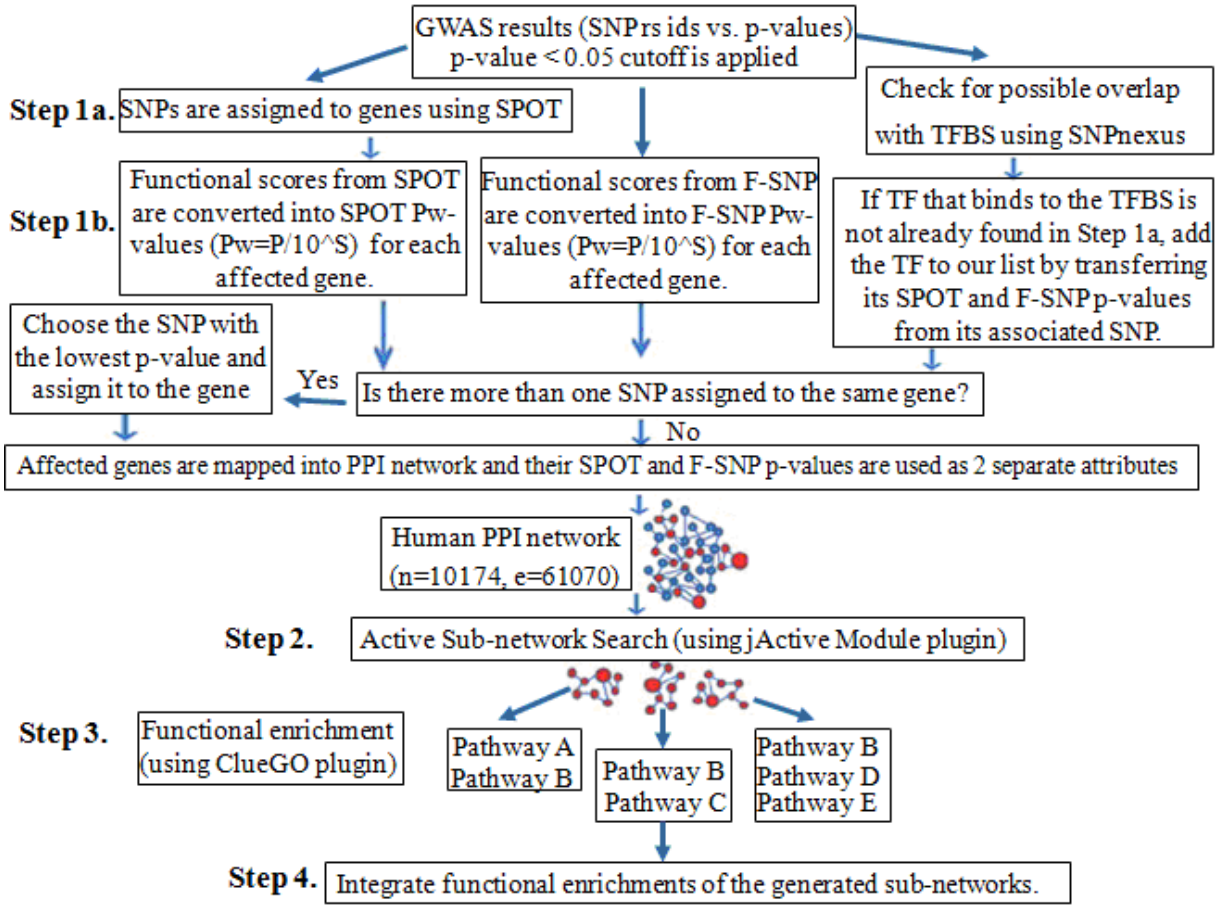


Figure 4

Figure 1 Outline of PANOGA protocol An image version of Figure.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- supplement0.doc
- supplement0.doc
- supplement0.doc
- supplement0.doc