SOFTWARE TOOL ARTICLE

MetaNetVar: Pipeline for applying network analysis tools for genomic variants analysis [version 1; referees: 3 approved]

Eric Moyer[1], Megan Hagenauer[2], Matthew Lesko[3], Felix Francis[4], Oscar Rodriguez[5], Vijayaraj Nagarajan[6], Vojtech Huser[7], Ben Busby[3]

[1]National Center for Biotechnology Information, Bethesda, USA
[2]Molecular, Behavioral Neuroscience Institute, University of Michigan, Ann Arbor, USA
[3]National Center for Biotechnology Information, National Library of Medicine, Bethesda, USA
[4]Bioinformatics and Systems Biology program, University of Delaware, Newark, USA
[5]Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, USA
[6]Bioinformatics and Computational Biosciences Branch, National Institute of Allergy and Infectious Diseases, National Institute of Mental Health, Bethesda, USA
[7]Lister Hill National Center for Biomedical Communications, National Library of Medicine, National Institute of Mental Health, Bethesda, USA

v1 **First published:** 13 Apr 2016, **5**:674 (doi: 10.12688/f1000research.8288.1)
Latest published: 13 Apr 2016, **5**:674 (doi: 10.12688/f1000research.8288.1)

Abstract

Network analysis can make variant analysis better. There are existing tools like HotNet2 and dmGWAS that can provide various analytical methods. We developed a prototype of a pipeline called MetaNetVar that allows execution of multiple tools. The code is published at https://github.com/NCBI-Hackathons/Network_SNPs. A working prototype is published as an Amazon Machine Image - ami-4510312f .

This article is included in the Hackathons channel.

Open Peer Review

Referee Status: ✓ ✓ ✓

	Invited Referees 1	2	3
version 1 published 13 Apr 2016	✓ report	✓ report	✓ report

1 **John Didion**, National Institutes of Health USA

2 **Sahar Al Seesi**, University of Connecticut USA

3 **Tomasz Adamusiak**, Thomson Reuters USA

Discuss this article

Comments (0)

Corresponding author: Ben Busby (ben.busby@nih.gov)

How to cite this article: Moyer E, Hagenauer M, Lesko M *et al.* **MetaNetVar: Pipeline for applying network analysis tools for genomic variants analysis [version 1; referees: 3 approved]** *F1000Research* 2016, **5**:674 (doi: 10.12688/f1000research.8288.1)

Grant information: The work on this project by Vojtech Huser, Eric Moyer and Ben Busby was supported by the Intramural Research Program of the National Institutes of Health (NIH)/ National Library of Medicine (NLM)/ Lister Hill Center (VH) and NCBI (EM and BB). Megan Hagenauer's work on this project was supported by the Pritzker Neuropsychiatric Disorders Research Consortium.
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: No competing interests were disclosed.

First published: 13 Apr 2016, **5**:674 (doi: 10.12688/f1000research.8288.1)

Introduction

Traditionally, the goal of genome-wide association studies (GWAS) has been to associate single nucleotide polymorphisms (SNPs) and their respective haplotype blocks with disease status, allowing the eventual identification of particular genes responsible for disease phenotype. Unfortunately, only a small subset of diseases arise from variants within a single gene. For most complex diseases, it is likely that the disease arises due to the interactive effects of multiple genetic variants, and different collections of these variants may be present in different patients. Within a GWAS study, these variants individually will exhibit low predictive power making it difficult for researchers to obtain a sufficient sample size to identify them with high confidence. Therefore, tools that can help detect groups of interacting genetic variants are needed[1–3].

One set of tools that has great potential for aiding in this problem is network analyses. Within these tools, the results from GWAS studies are overlaid on networks constructed from curated molecular interaction data, such as databases documenting protein-protein interactions (PPIs), protein-DNA interactions, metabolite interactions, and gene-gene co-expression[1,2]. Many of these tools are powerful, but somewhat inaccessible to users with weaker computational backgrounds. For example, installing, configuring, running, and comparing the output of multiple network analysis tools could require a working knowledge of command-line scripting, Python, R, and Perl. Therefore, the goal of our hackathon team was to create a single command-line pipeline within which a user could input the results of a GWAS study, execute existing network analysis tools, and then access results from multiple network analyses. This work was conducted as part of the NCBI January 2016 Hackathon.

Methods

The context of the hackathon event allowed only three development days to create the pipeline which impacted the scope and design of the tool. The focus was on allowing one input file to be directed towards multiple tools; consolidation of results from individual tools was out of scope. Similarly, each tool output was not post-processed for unified output. We envision that future improvement to the pipeline may offer advanced visualisation options; however, this was not part of this pilot implementation. A working instance of the pipeline is also published as an Amazon Machine Image ami-4510312f.

Tools used in the pipeline

As much as possible, the MetaNetVar pipeline uses existing tools for network analysis. We only considered tools that are freely available, with no license restrictions. We describe briefly each tool that is integrated into the MetaNetVar pipeline. Tools vary in scope, and some include additional functions that include network analysis.

FunSeq2 (Version 2.1.2)

FunSeq2 is an existing tool for prioritizing variants using several different approaches, including network-based analysis[4,5]. FunSeq2 identifies hub genes and provides the measure of centrality for those hub genes. Inference of the network analysis results requires further processing of the program's output. We chose to include FunSeq2 in our pipeline because of its capability to identify functionally important, non-coding variants in the context of biological networks.

NetworkX (Version 1.10)

NetworkX is a network analysis framework available in a Python language software package. It allows for "the creation, manipulation, and study of the structure, dynamics, and functions of complex networks"[6]. It contains many standard graph algorithms and accepts and outputs 13 file formats, where nodes can be anything and edges can hold any type of data. NetworkX was used to calculate the degrees and betweenness centrality of nodes (genes) and to create XML format and static PNG figures of subnetworks containing the input genes. The degrees and the betweenness centrality gives you a measurement of how important the gene is in the network. A network analysis framework similar to NetworkX is CytoscapeJS[7]. We chose to include NetworkX in our pipeline because of our experience with Python.

HotNet2 (Version 1.0.1)

HotNet2 is an algorithm for detecting "significantly altered subnetworks in a large gene interaction network"[8–10]. The algorithm uses heat diffusion kernel to capture the local topology of the interaction network. The subnetworks in genome scale interactions that have non-random mutations are identified using this approach. The limitation of HotNet2 are the challenges in getting the scripts running straight out of the box, along with the long computational time involved in the preliminary influence matrix creation process. We chose to include HotNet2 in our pipeline because of our experience with Python.

dmGWAS (Version 3.0)

dmGWAS_3.0 is an existing tool for overlaying gene-level summaries of case-control association p-values onto an existing network (in this case, we use the network extracted from GeneMania detailed below) and then identifying subnetworks that are particularly enriched for strong associations using a greedy algorithm[11,12]. Unlike the previous version of dmGWAS (2.0), dmGWAS_3.0 also has the ability to incorporate differential gene coexpression data (in other words, the difference in gene co-expression between cases and controls) as weights for the edges in the network, but for the sake of simplicity we did not make use of this new functionality in our pipeline. Due to this choice, we discovered that the dense-module search output (ResList.RData) took the format of the previous version dmGWAS_2.0[13] and could not be manipulated using the tools referenced in the current documentation. Therefore, we created our own short script to extract out the basic statistics and subnetwork nodes associated with each input gene present in the network (see below: ModuleStrengthSummaryByGene.txt and Top1000ModuleScores.txt). We later discovered that some of the old tools capable of manipulating dmGWAS_2.0 output (ResList.RData) were preserved in the current code package and could be used for further data exploration by a motivated user by loading the output file (ResList.RData) into R and installing the requisite packages (dmGWAS, igraph), although some of the tools did not appear to be fully functional anymore (such as the subnetwork plotting capability in simpleChoose()).

Overall, the primary limitation that we observed for dmGWAS was computing time, so we adapted the existing code to make use of parallel computing using the BiocParallel package in R[14].

To utilize dmGWAS_3.0, it is first necessary to convert the input file containing the case-control association p-values for each SNP to a gene-level summary. Within our pipeline, we complete this conversion using VEGAS, an existing command line (Linux/Unix) based tool recommended within the dmGWAS documentation[15,16]. It should be noted that by default, VEGAS uses the HapMap2 CEU (Central Europeans, Utah) population to estimate patterns of linkage disequilibrium for each gene.

VEGAS is written in Perl but also makes use of two R packages (corpcor and mvtnorm) and depends on functions provided by PLINK, a commonly-used whole genome analysis toolset[17,18]. The output of VEGAS requires further processing before input into dmGWAS. We found that several of the VEGAS gene-level p-values were rounded to either 0 or 1, which was incompatible with dmGWAS, so we substituted the minimum p-value present in our test file (1e-06) for 0 and 0.999 for 1.

Table 1 provides a summary of the tools used in this pipeline, including notes about their advantages and disadvantages.

Networks used

Network construction based on a user-provided list of variants required accessing molecular interaction network data from external databases. We describe the network databases utilized by each tool. Some networks (such as Multinet) are used by multiple tools (FunSeq2 and HotNet2).

FunSeq2

FunSeq2 utilizes multinet[19], which is an integrated network consisting of regulatory interactions from the ENCODE regulatory network[20,21], phosphorylation interactions from the SignaLink database[22–24], protein-protein interactions from BioGRID (release 3.1.83)[25], and metabolic interactions from KEGG[26]. While there are options for users to bring in their own network for use with FunSeq2 analysis, this pipeline prototype uses the pre-packaged multinet.

NetworkX & dmGWAS

The NetworkX and dmGWAS are libraries and do not include particular network data.

We paired GeneMania with NetworkX and dmGWAS. The GeneMania network is a protein-protein interaction network[27,28]. Two genes are connected if they are found to interact in a protein-protein interaction study. The network was created from various protein-protein interaction databases, including BioGRID and Pathway Commons[29,30]. We used version 2014-10-15 of Homo_sapiens.COMBINED network.

HotNet2

HotNet2 uses mutation data to prioritize subnetworks by identifying significantly mutated subnetworks in genome scale interaction networks. In our pipeline, we have used the 2012 version of HINT (High-quality INTeractomes) a database of high-quality protein-protein interactions[31].

Example data

As a sample input for our pilot, we searched NCBI dbGaP for a sample study that provided a real world list of variants. We used data from a clinical study of age-related macular degeneration[32] with dbGAP identifier phs000182.v3.p1.

As an additional input example, we used data from ClinVar[33]. ClinVar is a database of interpretations of clinical significance of variants for reported conditions, hosted by the National Center for Biotechnology Information (NCBI). It includes germline

Table 1. An overview of the tools used in our pipeline.

Name	Advantages	Disadvantages	Platform
FunSeq2	Uses ENCODE Regulatory Network data to identify hubs	Output needs to be parsed to better understand the network related results; make sure the correct reference build and the correct coordinate system (inclusive or exclusive) is used	Perl program
NetworkX	Ease-of-use, rapid development, open-source, flexible graph implementations	Cannot use for large-scale problems with more than 100 million nodes	Python library
HotNet2	HotNet2 algorithm uses heat diffusion kernel analogous to random walk with restart to better capture the local topology of the interaction network.	Challenging to run the scripts directly; poor documentation; had to fix some bugs to get it working; the one time influence matrix creation for a new network may take several hours.	Python
dmGWAS	Predicts molecular subnetworks that are enriched for disease-associations using the full results from a GWAS study (no thresholding of input by p-value or rank!).	Computationally intensive: may take days to produce results. Our updated version of dmGWAS uses parallel computing to speed up processing time, but still may take several hours even on a large cloud server. Some syntax provided in the documentation does not function, and output contents follow an older format. However, only minimal tweaking was required for dmGWAS to be integrated into the pipeline.	R package (but dependent on command-line VEGAS and PLINK toolsets)

and somatic variants of any size, type, or genomic location with interpretations from several sources (such as clinical testing laboratories, research laboratories, or locus-specific databases). It includes a link of variants to phenotypes. For this example, we identified variations submitted by LabCorp and extracted disease-variant pairs, for diseases with 30+ variants. The example dataset is provided on the MetaNetVar GitHub page.

Results

We implemented four network analysis programs or platforms into our pipeline (FunSeq2, NetworkX, HotNet2, dmGWAS), utilizing molecular interaction data from several external knowledge databases (listed above). Figure 1 provides an overview of the pipeline.

To lower the adoption threshold for potential users, we offer the snapshot of our working instance as an Amazon Machine Image. The collection of tools and the pipeline script can be executed using an instance of our publicly available Amazon Machine Image: ami-4510312f. The accompanying supplementary file describes the step-by-step procedure for running our pipeline using the published Amazon Machine Image.

We discuss below the results from individual tools integrated into our pipeline. Results of all of these tools, using the example dataset, is also provided on the MetaNetVar GitHub page.

FunSeq2

The parsed input file required for the FunSeq2 analysis, the PHP script that generates this parsed input file from the original dbGaP association data, and the output files (using default parameters) are provided in the GitHub project page. An example file generated from a filtered list of SNPs from the ClinVar database, for the Cardiomyopathy phenotype is also provided for testing purposes and can be found at http://github.com/NCBI-Hackathons/Network_SNPs/blob/master/test/sample_output/funseq2/cardiomyopathyfunseqoutput.

NetworkX

We took the NetworkX library and created a script that we refer to as SNPsNet. This script generates one output file containing the degrees and betweenness centrality measure of genes that are input into the pipeline, as well as creating two directories (see Figure 2). The two directories contain figures of subnetworks with the input genes and the XML format of the subnetworks. With these results, the user can prioritize the input variants or genes by sorting how important each gene is based on degreeness or centrality, as well as visualizing the subnetwork. Since NetworkX is not primarily a visualization tool, the XML file can be input into several other tools to better visualize the graph.

HotNet2

The influence matrix for HINT was pre-computed and then used in the current version of our pipeline. Influence matrix creation is a one-time process for a given network and, if required, advanced users may use custom influence matrices with MetaNetVar by modifying the path to the input influence matrix file and corresponding gene index file. For evaluation of MetaNetVar, we generated heat scores from a test mutation file. The .json file containing heat scores on each gene, which was used in subsequent steps, may be accessed at https://github.com/NCBI-Hackathons/Network_SNPs/blob/master/heat_score.json.

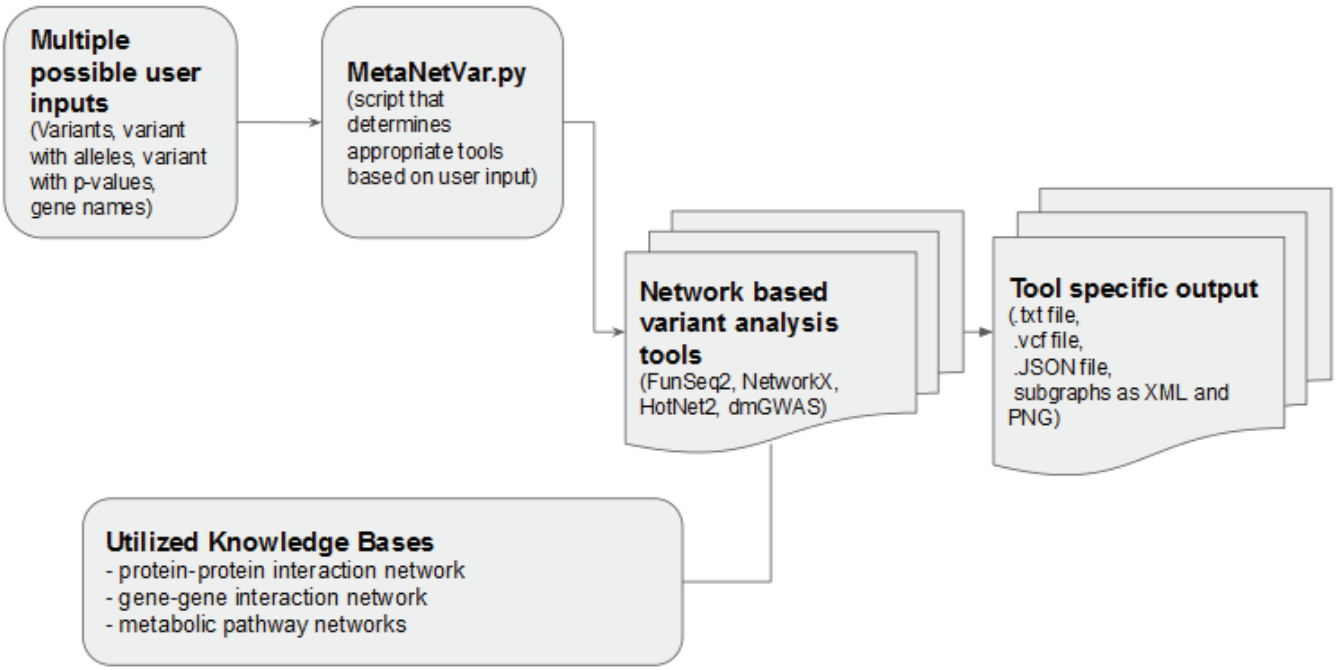

Figure 1. An overview of the pipeline.

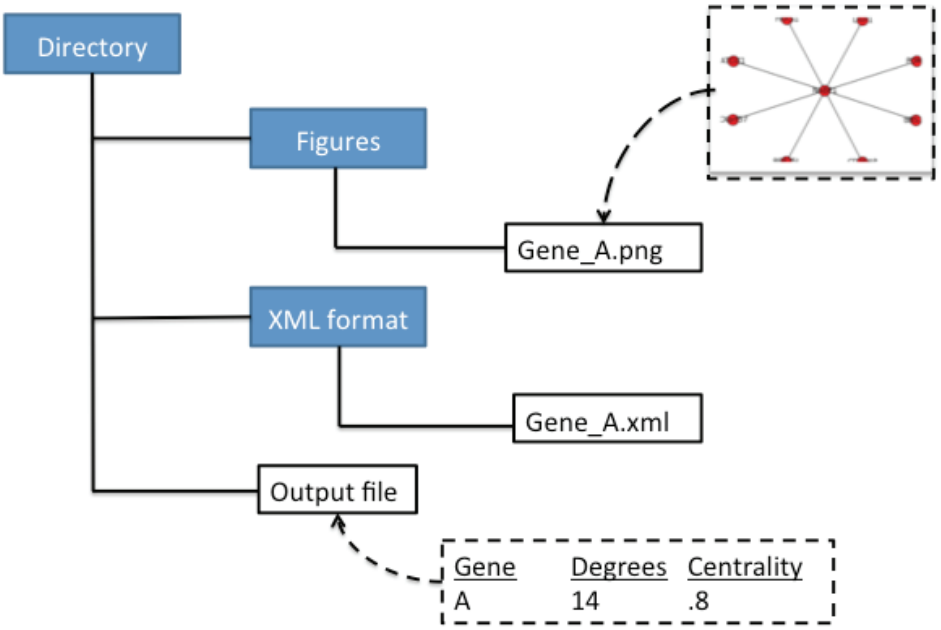

Figure 2. NetworkX outputs a file containing the degreeness and centrality of each gene, as well as two directories containing subnetwork graph figures for each input gene (.png) and its XML format.

The final step of weighted graph generation uses the influence matrix for HINT, the HINT index file, and the heat score .json file, to remove edges with weight less than the delta value, and extract the resulting connected components. Two output files were generated: components.txt (available at https://github.com/NCBI-Hackathons/Network_SNPs/blob/master/components.txt) and results.json (available at https://github.com/NCBI-Hackathons/Network_SNPs/blob/master/results.json)

dmGWAS

The sample association file from the age-related macular degeneration dataset (phs000182) was parsed down to a two-column text file containing only SNP identifiers ("rs numbers") and case-control association p-values. This file was fed into VEGAS and a gene-level summary file was created, which was further parsed into a simple two-column text file containing gene identifiers (gene symbol) and "weight" (an integrated p-value for the gene). Within dmGWAS, this input was overlaid onto a network provided by GeneMania to produce a network of weighted nodes from which particularly "dense" subnetworks are identified (full output: ResList.RData). Finally, our program summarizes the data into two easily navigable tab-delimited .txt files which can be viewed within accessible programs such as Microsoft Excel (ModuleStrengthSummaryByGene.txt, Top1000ModuleScores.txt). Figure 3 and Figure 4 demonstrate example output files.

Limitations

The current version of the pipeline is set to use data from dbGaP and ClinVar out-of-the-box. However, advanced users could tweak the provided scripts to make it run using other input formats. Some of the components of the pipeline use processes that are parallel and compute-intensive in nature. Using the provided working implementation of the pipeline through Amazon Web Services requires some computing skills.

	Gene	Zn_NormalizedModuleScore	Percentile Rank	OriginalAssociationPvalue
1	GPR1	18.9024816	6.46E-05	0.09762
2	TNFRSF1A	18.8442386	0.00016161	0.148
3	CD3EAP	18.8442386	0.00016161	0.148
4	DNASE1	18.8140478	0.00025858	0.18
5	COL18A1	18.7791487	0.00032323	0.222
6	TIE1	18.7543278	0.00038787	0.255
7	BMP4	18.7352549	0.00045252	0.282
8	SPINK1	18.6605316	0.00051716	0.399
9	PSG11	18.6231576	0.00058181	0.462
10	FUT1	18.6062016	0.00064645	0.491

Figure 3. An example of the first output summary file produced by dmGWAS in our pipeline: ModuleStrengthSummaryByGene.txt. This file provides the Normalized Module Score for each gene included in the network ("Zn", where a larger value indicates the gene is more enriched for significant case-control associations), and the gene-level summary case-control association p-value provided by VEGAS. It is ordered by percentile rank to allow comparison across different network analysis programs.

Gene	Zn_NormalizedModuleScore	PercentileRank	OriginalAssociationPvalue	SubgraphGeneSets
GPR1	18.90248159	6.46E-05	0.09762	CFH, CFHR2, PLEKHA1, CFHR3, CFHR4, CFHR5, CRB1, BTBD16, F13B, KCNT2, HTRA1, C2, GPR1, SKIV2L, EHMT2, CFB, PRRT1, CFHR1
TNFRSF1A	18.84423858	0.0001616	0.148	CFH, TNFRSF1A, CFHR2, PLEKHA1, CFHR3, CFHR4, CFHR5, CASP6, F13B, KCNT2, HTRA1, C2, FKBPL, SKIV2L, EHMT2, CFB, PRRT1, CFHR1
CD3EAP	18.84423858	0.0001616	0.148	CFH, ASPM, CFHR2, CFHR3, CD3EAP, CFHR4, CFHR5, CRB1, BTBD16, F13B, KCNT2, HTRA1, C2, TNXB, ZBTB41, CFB, ZBTB12, CFHR1
DNASE1	18.81404777	0.0002586	0.18	CFH, CFHR2, CFHR3, CFHR4, CFHR5, CRB1, BTBD16, F13B, KCNT2, HTRA1, C2, TNXB, FKBPL, SKIV2L, EHMT2, CFB, DNASE1, CFHR1
COL18A1	18.7791487	0.0003232	0.222	CFH, CFHR2, PLEKHA1, CFHR3, CFHR4, CFHR5, CRB1, F13B, KCNT2, HTRA1, C2, TNXB, COL18A1, SKIV2L, EHMT2, CFB, PRRT1, CFHR1
TIE1	18.75432775	0.0003879	0.255	CFH, TIE1, CFHR2, PLEKHA1, CFHR3, CFHR4, CFHR5, CRB1, F13B, HTRA1, C2, TNXB, FKBPL, SKIV2L, EHMT2, CFB, PRRT1, CFHR1
BMP4	18.73525493	0.0004525	0.282	CFH, CFHR2, PLEKHA1, CFHR3, BMP4, CFHR4, CFHR5, CASP6, F13B, KCNT2, HTRA1, C2, FKBPL, SKIV2L, EHMT2, CFB, PRRT1, CFHR1
SPINK1	18.66053162	0.0005172	0.399	CFH, CFHR2, PLEKHA1, CFHR3, CFHR4, CFHR5, BTBD16, F13B, KCNT2, SPINK1, HTRA1, C2, SKIV2L, EHMT2, CFB, SLC44A4, PRRT1, CFHR1
PSG11	18.62315763	0.0005818	0.462	CFH, CFHR2, PLEKHA1, CFHR3, CFHR4, CFHR5, CRB1, BTBD16, F13B, KCNT2, HTRA1, C2, ZBTB41, CFB, PRRT1, ZBTB12, PSG11, CFHR1
FUT1	18.60620157	0.0006465	0.491	CFH, CFHR2, PLEKHA1, CFHR3, CFHR4, CFHR5, BTBD16, F13B, KCNT2, HTRA1, C2, FUT1, FKBPL, SKIV2L, EHMT2, CFB, PRRT1, CFHR1
ODF1	18.60503424	0.0007111	0.493	CFH, CFHR2, CFHR3, CFHR4, CFHR5, CRB1, BTBD16, F13B, ODF1, KCNT2, HTRA1, C2, TNXB, ZBTB41, CFB, PRRT1, ZBTB12, CFHR1

Figure 4. An example of the second output summary file produced by dmGWAS in our pipeline: Top1000ModuleScores.txt. This second output provides similar information as the first output file, but expands it to include the list of genes (nodes) present in each gene of interest subnetwork. Only subnetwork output for the top 1000 seed genes is provided (as determined by percentile rank).

Conclusions

Our tool, MetaNetVar, allows researchers with limited computational experience to access a host of powerful network analysis tools for application to genomic datasets. This platform is intended for use in a variety of future hackathons, including work on cancer and evolutionary biology, but will most likely also be used by participants from the current hackathon, as well as other interested individuals. Since this work was a pilot project, we expect further modification of the pipeline as new users provide feedback. Ideally, the future pipeline would include a unified output summary, better network visualization tools, and the ability to integrate known disease-related variants into the analysis, such as from ClinVar[33], from PheGeni[34], or from the output of epistasis analyses[35].

Data and software availability

Latest source code: https://github.com/NCBI-Hackathons/Network_SNPs

Archived source code as at time of publication: http://dx.doi.org/10.5281/zenodo.48202[36]

Amazon instance ID: ami-4510312f

Amazon instance name: NCBI-Hackathon-20160122-Network-SNPs

License: CC0 1.0 Universal

Author contributions

All of the authors participated in designing the study, carrying out the research, and preparing the manuscript. All authors were involved in the revision of the draft manuscript and have agreed to the final content.

Competing interests

No competing interests were disclosed.

Grant information

The work on this project by Vojtech Huser, Eric Moyer and Ben Busby was supported by the Intramural Research Program of the National Institutes of Health (NIH)/National Library of Medicine (NLM)/Lister Hill Center (VH) and NCBI (EM and BB). Megan Hagenauer's work on this project was supported by the Pritzker Neuropsychiatric Disorders Research Consortium.

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Acknowledgements

The authors thank Lisa Federer, NIH Library Writing Center, for manuscript editing assistance.

Supplementary material

Software manual. This document provides instructions on how to start and run the NCBI-Hackathon-20160122-Network-SNPs instance in AWS using a Mac computer.

References

1. Leiserson MD, Eldridge JV, Ramachandran S, *et al.*: **Network analysis of GWAS data.** *Curr Opin Genet Dev.* 2013; **23**(6): 602–10.
PubMed Abstract | Publisher Full Text | Free Full Text
2. Bolouri H: **Modeling genomic regulatory networks with big data.** *Trends Genet.* 2014; **30**(5): 182–91.
PubMed Abstract | Publisher Full Text
3. Halldórsson BV, Sharan R: **Network-based interpretation of genomic variation data.** *J Mol Biol.* 2013; **425**(21): 3964–9.
PubMed Abstract | Publisher Full Text
4. Khurana E, Fu Y, Colonna V, *et al.*: **Integrative annotation of variants from 1092 humans: application to cancer genomics.** *Science.* 2013; **342**(6154): 1235587.
PubMed Abstract | Publisher Full Text | Free Full Text
5. GersteinLab@Yale: **Funseq2 [Internet].** [cited 2016 Feb 24].
Reference Source
6. NetworkX developer team: **Overview — NetworkX [Internet].** [cited 2016 Feb 24].
Reference Source
7. Franz M, Lopes CT, Huck G, *et al.*: **Cytoscape.js: a graph theory library for visualisation and analysis.** *Bioinformatics.* 2016; **32**(2): 309–11.
PubMed Abstract | Publisher Full Text | Free Full Text
8. Raphael Lab: **HotNet [Internet].** [cited 2016 Feb 24].
Reference Source
9. Raphael Group: **GitHub - hotnet2 [Internet].** [cited 2016 Feb 24].
Reference Source
10. Vandin F, Upfal E, Raphael BJ: **Algorithms for detecting significantly mutated pathways in cancer.** *J Comput Biol.* 2011; **18**(3): 507–22.
PubMed Abstract | Publisher Full Text
11. Vanderbilt University Bioinformatics, Systems Medicine Laboratory: **dmGWAS 3.0 [Internet].** [cited 2016 Feb 24].
Reference Source
12. Jia P, Zheng S, Long J, *et al.*: **dmGWAS: dense module searching for genome-wide association studies in protein-protein interaction networks.** *Bioinformatics.* 2011; **27**(1): 95–102.
PubMed Abstract | Publisher Full Text | Free Full Text
13. Jia P, Zheng S, Zhao Z: **dmGWAS 2.0: dense module searching for genome-wide association studies in protein-protein interaction network [Internet].** [cited 2016 Feb 24].
Reference Source
14. Carey V, Lawrence M, Morgan M: **Introduction to BiocParallel [Internet].** [cited 2016 Feb 24].
Reference Source
15. Liu JZ, McRae AF, Nyholt DR, *et al.*: **A versatile gene-based test for genome-wide association studies.** *Am J Hum Genet.* 2010; **87**(1): 139–45.
PubMed Abstract | Publisher Full Text | Free Full Text
16. Liu J, MacGregor S: **VEGAS: Versatile Gene-based Association Study [Internet].** [cited 2016 Feb 24].
Reference Source
17. Purcell S: **PLINK: Whole genome data analysis toolset [Internet].** [cited 2016 Feb 24].
Reference Source
18. Purcell S, Neale B, Todd-Brown K, *et al.*: **PLINK: a tool set for whole-genome association and population-based linkage analyses.** *Am J Hum Genet.* 2007; **81**(3): 559–75.
PubMed Abstract | Publisher Full Text | Free Full Text
19. Kellis M, Wold B, Snyder MP, *et al.*: **Defining functional DNA elements in the human genome.** *Proc Natl Acad Sci U S A.* 2014; **111**(17): 6131–8.
PubMed Abstract | Publisher Full Text | Free Full Text
20. National Human Genome Research Institute: **The ENCODE Project: ENCyclopedia Of DNA Elements [Internet].** [cited 2016 Feb 24].
Reference Source
21. Khurana E, Fu Y, Chen J, *et al.*: **Interpretation of genomic variants using a unified biological network approach.** *PLoS Comput Biol.* 2013; **9**(3): e1002886.
PubMed Abstract | Publisher Full Text | Free Full Text
22. SignaLink: **SignaLink 2.0 [Internet].** [cited 2016 Feb 24].
Reference Source
23. Fazekas D, Koltai M, Türei D, *et al.*: **SignaLink 2 - a signaling pathway resource with multi-layered regulatory networks.** *BMC Syst Biol.* 2013; **7**: 7.
PubMed Abstract | Publisher Full Text | Free Full Text
24. Korcsmáros T, Farkas IJ, Szalay MS, *et al.*: **Uniformly curated signaling pathways reveal tissue-specific cross-talks and support drug target discovery.** *Bioinformatics.* 2010; **26**(16): 2042–50.
PubMed Abstract | Publisher Full Text
25. TyersLab.com: **BioGrid [Internet].** [cited 2016 Feb 24].
Reference Source
26. Ogata H, Goto S, Sato K, *et al.*: **KEGG: Kyoto Encyclopedia of Genes and Genomes.** *Nucleic Acids Res.* 1999; **27**(1): 29–34.
PubMed Abstract | Publisher Full Text | Free Full Text
27. Warde-Farley D, Donaldson SL, Comes O, *et al.*: **The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function.** *Nucleic Acids Res.* 2010; **38**(Web Server issue): W214–W220.
PubMed Abstract | Publisher Full Text | Free Full Text
28. University of Toronto: **GeneMANIA [Internet].** [cited 2016 Feb 25].
Reference Source
29. Cerami EG, Gross BE, Demir E, *et al.*: **Pathway Commons, a web resource for biological pathway data.** *Nucleic Acids Res.* 2011; **39**(Database issue): D685–D690.
PubMed Abstract | Publisher Full Text | Free Full Text
30. Memorial Sloan-Kettering Cancer Center, University of Toronto: **Pathway Commons [Internet].** [cited 2016 Feb 25].
Reference Source
31. Leiserson MD, Vandin F, Wu HT, *et al.*: **Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes.** *Nat Genet.* 2015; **47**(2): 106–14.
PubMed Abstract | Publisher Full Text | Free Full Text
32. Abecasis GR, Yashar BM, Zhao Y, *et al.*: **Age-related macular degeneration: a high-resolution genome scan for susceptibility loci in a population enriched for late-stage disease.** *Am J Hum Genet.* 2004; **74**(3): 482–94.
PubMed Abstract | Publisher Full Text | Free Full Text
33. Landrum MJ, Lee JM, Benson M, *et al.*: **ClinVar: public archive of interpretations of clinically relevant variants.** *Nucleic Acids Res.* 2016; **44**(D1): D862–D868.
PubMed Abstract | Publisher Full Text | Free Full Text
34. National Center for Biotechnology Information: **PheGenI: Phenotype-Genotype Integrator [Internet].** [cited 2016 Feb 25].
Reference Source
35. Upton A, Trelles O, Cornejo-García JA, *et al.*: **Review: High-performance computing to detect epistasis in genome scale data sets.** *Brief Bioinform.* 2015; pii: bbv058.
PubMed Abstract | Publisher Full Text
36. John G, TriLe965, Hsu J, *et al.*: **Structural_Variant_Comparison: Initial Post-Hackathon Release.** *Zenodo.* 2016.
Data Source

Open Peer Review

Current Referee Status: ✓ ✓ ✓

Version 1

Referee Report 29 April 2016

doi:10.5256/f1000research.8914.r13366

Tomasz Adamusiak
Thomson Reuters, Boston, MA, USA

Excellent work given the limited hackathon time frame. I commend the authors for providing an AMI image and source code for the project.

Minor comments:

- *aiding in this problem is network analyses.*
  Should be analysis
- Would change AWS deployment manual format to pdf and provide instructions how to stop the EC2 instance so that the user doesn't not incur unnecessary costs.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Competing Interests: No competing interests were disclosed.

Referee Report 26 April 2016

doi:10.5256/f1000research.8914.r13368

Sahar Al Seesi
Department of Computer Science and Engineering, University of Connecticut, Storrs, CT, USA

The authors describe a pilot version of an integrated pipeline of network analysis tools for genomic variants. It includes four existing tools. The pipeline analyzes the input files and run the tools applicable to the input files. The value of this contribution would greatly increase if the pipeline consolidated the output of the different tools. The authors acknowledge this fact and plan to include that in future versions of the pipeline

The manuscript is well written, and the functionality of tools included is clearly described.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Competing Interests: No competing interests were disclosed.

Referee Report 21 April 2016

doi:10.5256/f1000research.8914.r13367

John Didion
National Human Genome Research Institute, National Institutes of Health, Bethesda, MD, USA

Minor points:

- "...somewhat inaccessible to users with weaker computational backgrounds" - I have a strong computational background, and dealing with poor build processes and user interfaces is frustrating to me also. Maybe rephrase this to say that different tools have different levels of usability, which can be ameliorated by providing a single, well-designed interface to multiple tools.

- In Figure 1, "Network-based variant analysis tools" is represented as a single box, but there are multiple steps encapsulated there. It would be more informative to show the steps involved for each of the four tools and the output of each tool.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Competing Interests: No competing interests were disclosed.