**NCBI Hackathon, January 2016, Group 2:**

**MetaNetVar: Pipeline for applying network analysis tools for genomic variants analysis**

[Introduction](#h.r0m7fmtxf6kh)

[Methods](#h.helx2jw5ufc8)

[Tools used in the pipeline](#h.jheavj461vsu)

[Networks used](#h.x75nu2tnxmmk)

[Example data](#h.at6pr116b9h4)

[Results](#h.14btudcgde3f)

[Limitations](#h.ich0r7kzpnqv)

[Conclusions](#h.v1rp7vawtwtd)

[Data and Software Availability](#h.5rq4gsaqbjrp)

[Extra material (do not submit this)](#h.wnxnilim1gtg)

(as of now )\* all authors contributed equally (may change later)

Vijayaraj Nagarajan

Bioinformatics and Computational Biosciences Branch, National Institute of Allergy and Infectious Diseases, NIH, Bethesda, MD 20892

[nagarajanv@mail.nih.gov](mailto:nagarajanv@mail.nih.gov)

Vojtech Huser

Lister Hill National Center for Biomedical Communications (LHNCBC).

National Library of Medicine, NIH, Bethesda, MD

[vojtech.huser@nih.gov](mailto:vojtech.huser@nih.gov)

Eric Moyer

DbSnp Redesign Team (Contractor)

National Center for Biotechnology Information, Bethesda, MD 20892

[eric.moyer@nih.gov](mailto:eric.moyer@nih.gov)

Felix Francis

Bioinformatics and Systems Biology program, University of Delaware, Newark, DE 19716

[ffrancis@udel.edu](mailto:ffrancis@udel.edu)

Oscar Rodriguez

Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, New York 10029

[oscar.rodriguez@icahn.mssm.edu](mailto:oscar.rodriguez@icahn.mssm.edu)

Megan Hagenauer

Molecular, Behavioral Neuroscience Institute, University of Michigan, Ann Arbor, MI

hagenaue@umich.edu

Ben Busby

NCBI, NLM, Bethesda, MD

[ben.busby@nih.gov](mailto:ben.busby@nih.gov)

Keywords

network analysis, genetic variant, pipeline, next generation sequencing

# Introduction

Traditionally, the goal of Genomic 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 and it is 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.

One set of tools that has great potential for aiding in this problem are 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. Many of these tools are powerful, but somewhat inaccessible to users with weaker computational backgrounds. For example, 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 synthesized results from multiple network analyses.

This work was part of NCBI Jan 2016 hackathon. [reference to MASTER paper] The project was selected via the following process: Team leads were approached with rough ideas for a plan. Several (4-8) iterations of plan were discussed with team lead. Somewhere in the middle of that, abbreviated title came out with announcement

Applicants were amassed and team leads picked team members.

# Methods

The context of the Hackathon event allowed only 3 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 tool. 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.

## Tools used in the pipeline

The pipeline is using as much as possible 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. Tools vary in scope and some include additional analyses that precede network analysis.

**Tools used for gene annotation:**

**PLINK:**

Plink [PMID: 17701901, <http://pngu.mgh.harvard.edu/~purcell/plink/index.shtml>] is a widely-used whole genome association analysis toolset.

**VEGAS:**

Vegas [PMID: 20598278, http://gump.qimr.edu.au/VEGAS/] is an existing tool used to summarize SNP association p-values by gene.

**Tools used for network construction:**

**FunSeq2**

FunSeq2 [PMID: 24092746, http://funseq2.gersteinlab.org/] is an existing tool for prioritizing variants using several different approaches, including network based analysis. The limitation of FunSeq are the numerous dependencies that are needed to have it successfully running, the network based results have to be parsed out of all other annotations if the results are to be integrated with other data.

**NetworkX**

NetworkX [https://networkx.github.io/] 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.”

**HotNet2**

HotNet2 [PMID: 21385051, http://compbio.cs.brown.edu/projects/hotnet/] is an existing tool for detecting “significantly altered subnetworks in a large gene interaction network”. The limitation of HotNet2 are:

**dmGWAS**

dmGWAS [ref1: http://bioinfo.mc.vanderbilt.edu/dmGWAS/] 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 identify subnetworks that are particularly enriched for strong associations. The primary limitation that we observed for dmGWAS was computing time.

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

|  |  |  |  |
| --- | --- | --- | --- |
| Name | Avantages | Disadvantages | Note |
| 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 |
| 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 that better captures the local topology of the interaction network. | Challenging to run the scripts directly; poor documentation; Had to fix few bugs to get it working. |  |

Table1: Overview of tools

## 

## Networks used

Our goal was to enrich the user list of variants with data from external networks.

**GeneMania**

Protein protein interaction network

GeneMania VNTODO

**network 2**

**Multinet (for FunSeq)**

## Example data

As a sample input for our study, we searched 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 [PMID:14968411] with dbGAP identifier phs000182

As an additional input example, we used data from ClinVar [PMID:26582918]. ClinVar is a database hosted by the National Center for Biotechnology Information (NCBI) of interpretations of clinical significance of variants for reported conditions. It includes germline and somatic variants of any size, type or genomic location with interpretations from several sources (e.g., clinical testing laboratories, research laboratories or locus-specific databases). It includes a link of variants to phenotypes. We created ClinVar extracts for 20 conditions using variants from a major US Clinical Laboratory (LabCorp, Laboratory Corporation of America Holdings, Burlington, NC).

# Results

We implemented 4 programs into our pipeline. Our pipeline is utilizing 3 external networks as knowledge bases.

Figure 1 shows the overview of the pipeline.

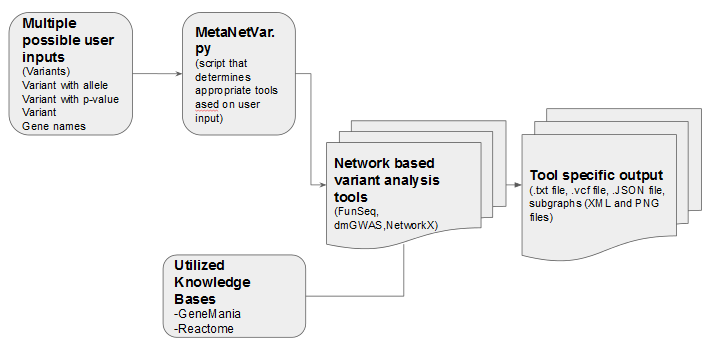


Figure1: Pipeline overview

To lower the adoption threshold for potential users, we offer Amazon Snapshot. The collection of tools and the pipeline script can be executed by executing stored amazon instance AMI # XXX-XXX-XXX. (VRTODO)

This work was a pilot project and we expect further modification of the pipeline.

## Limitations

Enriching a mere list of variants using external network (e.g., gene-gene) interaction network depends on the quality of curation and comprehensiveness of the network. If a network is biased to contain only cancer-related interactions, it may not be useful for non-cancer research projects.

Protein protein intera

# Conclusions

Our tool allows researchers with limited experience

Our work a limited technological advanceshows that

This platform is intended for use in a variety of future hackathons including work on cancer and evolutionary biology.

# 

# Data and Software Availability

The code for the pipeline is publically available on GitHub at <https://github.com/NCBI-Hackathons/Network_SNPs>

Amazon instance VRTODO AMI #