Designing Novel Queries for Analysing NoSQL Data of Gene-Disease Associations

Hira Yaseen
Department of Computer Science
COMSATS Institute of Information Technology
Sahiwal, Pakistan

Muhammad Atif Sarwar
Department of Computer Science
COMSATS Institute of Information Technology
Sahiwal, Pakistan

Javed Ferzund
Department of Computer Science
COMSATS Institute of Information Technology
Sahiwal, Pakistan

***Abstract*—To precisely identify gene associated diseases has been an open area of research for biological scientists to ensure clinical and psychological symptoms and treatment for human diseases. Because whole Human Genome is defined now it is the next step to find all necessary possible factors from such a complex data set that cause gene mutations and hence lead inherited and/or non-inherited diseases. So our research implementation combines all important factors from different biomolecular data sources to make one integrated data set and defines new relationships among these factors for gene associated disease/s that were not present in existing platforms. This paper presents a novel query model for NoSQL data storage that can help researchers to visualise relationships among gene factors and two new factors termed as "causative factors" and "drugs/treatment" for associated diseases. Since no data source applies graphical querying for gene associated diseases, our proposed novel cypher query model can help researchers to deeply analyse data set and get results in an efficient manner. The proposed query model writes novel cypher queries for this research domain on a graphical data model implemented in neo4j, which is a NoSQL (Not Only Structured) database. Use of NoSQL database and NoSQL query language has overcome certain limitations of relational databases, the existing data platforms had to cope up with. This paper gives a new suitable data storage format and effective data search queries for large, complex, semi-structured and multi-dimensional gene associated diseases data set to efficiently define new relationships among factors format to open new horizons of research.**

Keywords*—Cypher queries; NoSQL; data model; Gene-disease associations; Causative factors; Drugs/treatment

I. INTRODUCTION

The Human body Genome is made up of millions of cells that normally perform some pre-defined function for our daily survival. In each cell, a molecule named Deoxyribonucleic acid (DNA) is present that carries heredity information in all living organisms. This heredity information is called genetic code or gene structure that is a proper sequence of four nitrogenous data bases named as adenine (A), thymine (T), guanine (G), and cytosine (C). In each cell there exist 23 pairs of chromosomes where chromosomes are tightly enclosed DNA containing bit-level details of genetics. Out of 23 pairs, twenty two are termed as 'autosomes' and one pair is named as 'sex chromosome', responsible for transfer of gene code information from one generation to its offspring. A gene is a specific part of a chromosome that exists at some particular location and performs specified functions in all organisms. A gene can have so many alternative forms called 'allele' of a gene. Every human being can inherit only one allele of a gene. Allele of a gene can result in different physical traits such as eyes colour, hair colour and the shape of body parts etc.

Gene related diseases occur when any change in the gene code at chromosome level, gene level or allele level causes mutation/disorder of genetic code thus resulting in dysfunctional gene behaviour. These mutations are responsible for many inherited and non-inherited diseases in all living organisms. But particularly focusing diseases in humans associated with genes may involve a complex interaction of one or more genes with another gene, with single or combination of alleles and/or may be with some risk factors and causative factors. The risk of acquiring disease because of above mentioned causes is known as genetic susceptibility. Gene susceptibility can vary because of environmental factors for an existing life. Environmental factors such as exposure to radiations, chemicals, and sunlight can increase or decrease chances of gene mutations in a certain geographical area. Gene susceptibility conditions that can increase or lessen the potential for a disease are the latest research topics. A gene's code gives instructions to cell for making a specific protein and its production amount. Protein-protein interaction is also the latest research area to find gene-disease associations.

While working on gene-disease associations different data analytics techniques have been implemented by the researchers. As described in [1], G2D tool is web implementation used for finding gene associated diseases. This tool has worked on OMIM database and applies data mining algorithms to relate diseases with genes. In [2], microarray technology has been used to study gene expression profiles for Alzheimer's disease. In [3], sequence analysis of gene is used to study infectious disease. In [4], an analysis of amplified DNA sequences is used to study genetic diseases. However all of these techniques have used data sets from relational databases and apply different techniques on them.

In our research implementation we have introduced a novel way to relate gene-disease associations. We have made an effort to combine data from different research centres across the globe working on genomics and genes functions such as National Human Genome Research Institute (NHGRI), National centre for Biotechnology Information, and World Health Organisation (WHO).

Objectives of this research work are:

- To introduce graphical NoSQL unified data model that combines necessary factors from previous work implementations.
- To add some additional factors that can relate to gene-disease associations.
- To write effective cypher queries for finding new gene-disease associations.
- To relate 'causative factors' of a disease and suggest suitable 'drugs/treatment' to cure that disease.

Section 2 is a literature survey of some online available resources that store details of genes vs. diseases. This section covers all features provided by these publicly available data sources that can be used for research purposes. Section 3 includes data model that storage format of data set in NoSQL database using Neo4j. Section 4 describes novel queries for such a complex and large sized gene-disease association data set that has more than 100000 data entries to extract useful information from. This section describes fast and effective search queries that visually relate important factors for associations.

II. Related Work

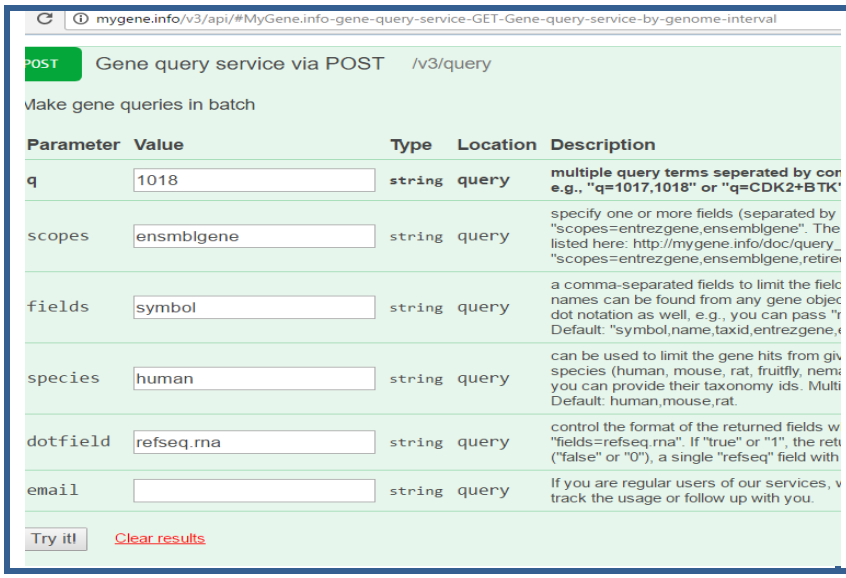
There have been many different platforms that has stored data sets relating to gene associated diseases in the form of relational databases and provided online as well as some offline tools. All biomolecular data was available in the form of large databases at some websites covering protein domains such as protein-protein interactions, genes ontology, tissues expressions, and gene expressions at different platforms. Wu, et al. 2012, has described in [5] that BioGPS is a centralized system built to aggregate distributed gene annotation resources user customisability options. However this system provides a publicly available web portal named 'MyGene.info' in which a gene query returns a list of canonical gene identifiers e.g. (NCBI Gene or Ensemble Gene IDs). This database helps users to discover gene centric resources only. Brown, et al. 2015, in [6] has provided insights of National Centre for Biotechnology Information NCBI's Entrez Gene Database for gene-specific information. This database keeps entries for sequence analysis of genomes, as it uses NCBI's Reference Sequence project (RefSeq). The data store includes nomenclature, genomic location, phenotypes and links to citations, sequences, variation details, maps, expression, homologs, and protein domains. Consortium, 2010, in [7] has provided a database named UniProt as a universal annotated protein sequences resource with querying facilities to help research community. UniProt is made up of four major parts. One is UniProtKB or UniProt Knowledgebase that has all protein information and a reference to all sources from which it is collected. Second is UniParc or UniProt Archive that contains history of all protein sequences. Third is UniRef or UniProt Reference Clusters that increase search speed for sequences by finding synonyms based upon sequence identity. Fourth part is UniMES or UniProt Metagenomic and Environmental Sequences database being updated for metagenomic data. Baker, et al. 2012, in [8] has integrated functional genomics in a web based system known as GeneWeaver. This web based system is powered by the Ontological Discovery Environment and this platform helps users to query different biological functions and their relations with genes. For example if a researcher wants to search a particular term the result includes all meta-data fields such as descriptions, publication information and NCBO Annotator [9] and Disease Ontology [10] terms. Liberzon, et al. 2011, in [11] has defined MSigDB that is another database for well-annotated gene sets showing all related biological processes. When user enters a query the result is a seven gene set collections. C1: for genes present in the same chromosome, C2: set of gene showing canonical pathways, C3: is for genes sets that share *cis*-regulatory motifs, C4: gives clusters of co-expressed modules for a large gene expression, C5: shows sets of genes relating to GO terms, C6: shows oncogenic signatures, and C7: lists immunologic signatures. Zambon, et al. 2012, in [12] has given a solution for pathway analysis of species, identifiers, gene sets and ontologies named as GO-Elite. GO-Elite takes benefits from the structured biological ontologies to show a minimum set of non-overlapping terms. This system provides enlists genes, phenotypes, diseases, pathways, and biomarkers with 50 IDs for more than 60 species. Barrett & Edgar, 2006, in [13] has introduced The Gene Expression Omnibus (GEO) repository at the National Centre for Biotechnology Information (NCBI) distributes gene expression data generated by DNA microarray technology. This web interface provides effective query searches and visualisation of data at individual gene levels. Kanehisa & Goto, 2000, in [14] has described KEGG (Kyoto Encyclopedia of Genes and Genomes) database that systematically analyse of relating genomic information with gene functions. A separate GENES database is introduced which keeps collection of indexed gene for all sequenced or partially sequenced genomes with annotation of gene functions. Rouillard, et al. 2016, in [15] has given a detailed description of database named "Harmonizome" which has gathered data from over 70 major online resources and mine gene based knowledge. However the datasets are stored in a relational database. In the tables of a relational data storage system the genes names are rows and their corresponding biological entities are columns. Huang, et al. 2009, in [16] has given a systematic analysis of gene lists using DAVID bio-informatic resources. This research work was aimed at finding biological semantics from large gene and/or protein lists using data sets and analytical tools on them. Data mining techniques has been used in DAVID to analyse genomic experiments. Bonifati, et al. 2003, in [17] has introduced that mutations in gene DJ-1 can associate to PARK7, which is a kind of human Parkinsonism. The authors have proven that loss of DJ-1 function results in neuro-degeneration. Moreau & Tranchevent, 2012, in [18] has described that statistical analysis of genes and proteins is required while integrating heterogeneous data sets. The authors have worked on expression data, sequence information, functional annotation and biomedical literature to rank genes and proteins because of limited resources. Lamb, et al. 2006, in [19] has introduced relation among genes, diseases and drugs. The authors have experimented cultured human cells along with pattern matching software to map molecules, genes and diseases. Teri

et al. 2008, in [20] has launched a project with the name of HapMap to enlist human genetic variations and their association studies to common diseases. This study has proven a great help in finding new research areas in pathophysiology of common diseases. Chen, et al. 2013, in [21] has integrated an open source, data store for long-non-coding RNA (lncRNA) and its associated diseases (LncRNADisease). This study mainly focuses on candidate lncRNA to find associated disease and its prognosis. The authors worked upon 480 experimentally supported lncRNA entries that associate to 166 diseases. Cookson, et. al. 2009, in [22], has worked on variations in gene expression and genome-wide gene expressions that can be mapped to understand complex diseases. It is concluded by the authors that gene mutation can help relating different gene factors that result in different quantitative level expressions. Clarke, et al. 2009, in [23] has found a relationship of genetic variations with lipoprotein level that can cause coronary disease. An increased level of lipoprotein is a high risk factor for heritable coronary artery disease. Özgür, et al. 2008, in [24] has gathered a data set that provides good candidate genes to get efficient experimentation on associated diseases using predictive analysis. The authors have implemented data mining based upon dependency parsing and support vector machines on a small available gene vs. disease data set to conclude genes most likely to cause associated disease. Little, et al. 2002, in [25], has proposed genotype study of genes for full human genome can be used to get gene-disease associations. Joshua et. al, 2010, in [26], has introduced PheWAS that determines phenome-wide scan can be better used for gene–disease associations.

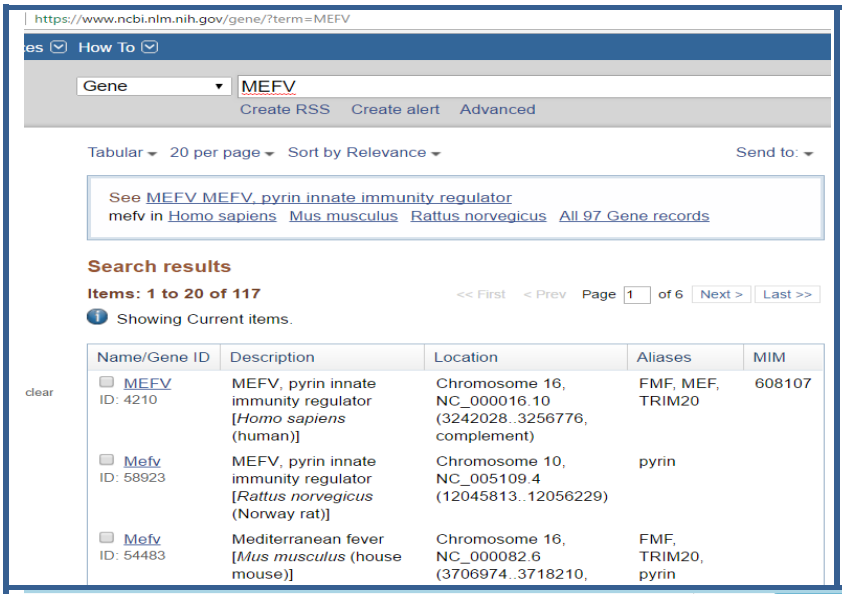
III. Proposed Query Model for NoSQL Data Storage Format in Neo4j

Based upon literature review it is observed that different publicly available data sources target different factors to determine diseases associated with any particular gene. So there is a need to get a unified data set containing all necessary factors to get all gene-disease associations. A comparative analysis of some known relational databases is done to get targeted factors of those data stores related to genes and diseases. Table 1 shows results of online databases when searched for a particular gene name or gene ID.

TABLE. I. Comparative Analysis of Available Gene Associated Diseases Databases

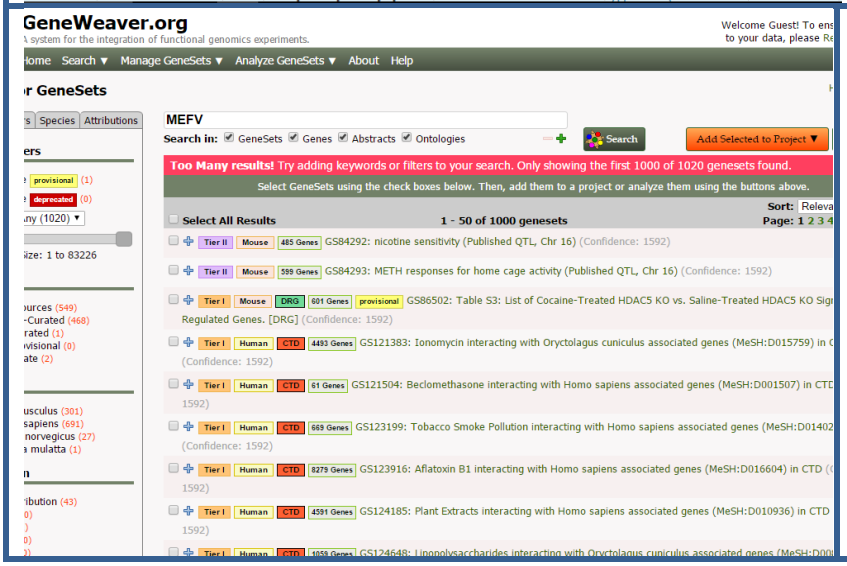


MyGeneinfo is an online data storage system that helps user to choose GeneID, scopes such as NCBI Gene database or Ensemble Gene IDs, one or more out of nine species, number of results, field terminator for files, ascending or descending order of fields returned, etc. and can email this matched gene results in the form of .csv file to a user email ID. This database is totally genes based for nine common species (human, mouse, rat, fruitfly, nematode, zebrafish, thale-cress, frog and pig).

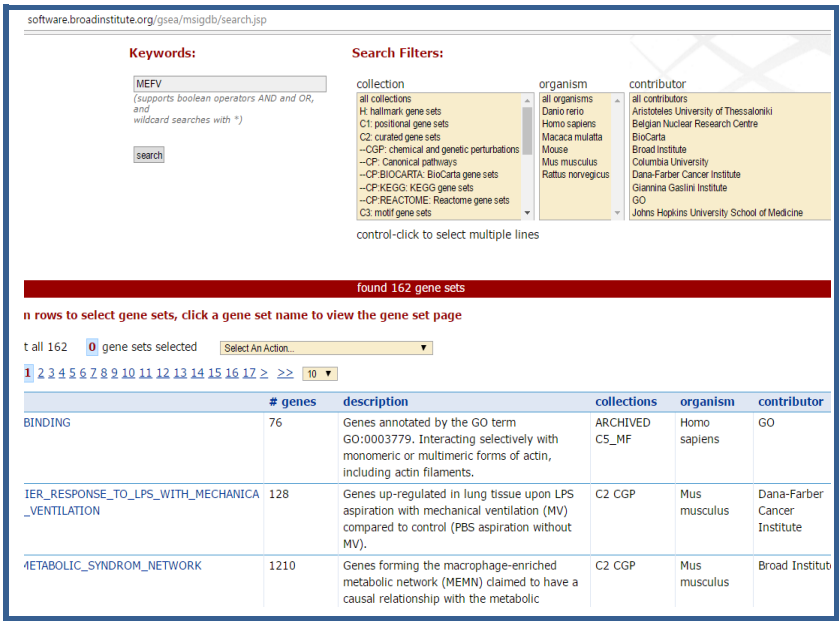


NCBI's Entrez Gene database provides gene associated diseases when searched for a particular gene name or a gene id. It also provides some additional content such as nomenclature, genomic location, phenotypes, links to citations, sequences, variation details, maps, expression, homologs, and protein domains. For example we searched for gene name = "MEFV" and it returned 117 results that include gene description, chromosome to which it belongs, aliases names and MIM database record ID.

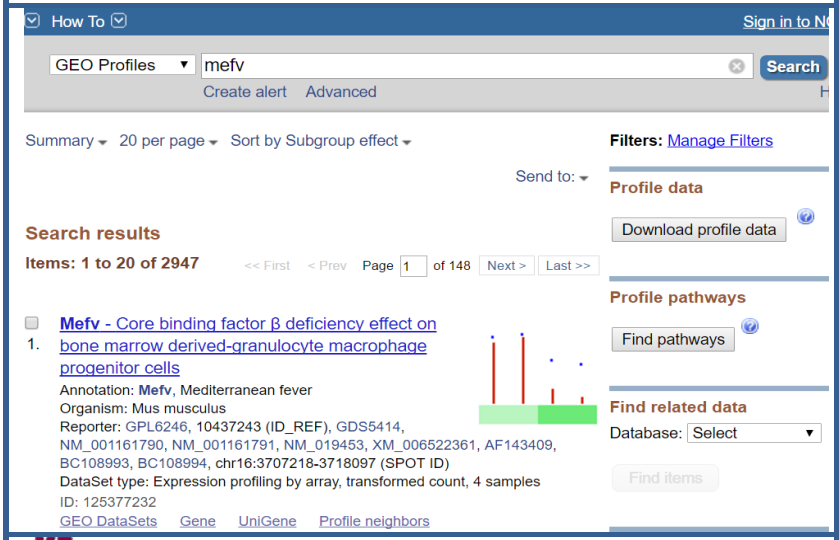
UniProt is a database for annotated protein sequences, that provides full protein name, gene names that make this protein, organism to which it belong, entry name, length etc. it provides user customized options to select fields that a user or researcher want to see and exclude the remaining fields. For example when searched for a protein "tubulin" it showed 86474 results for genes names in all organisms that exist.



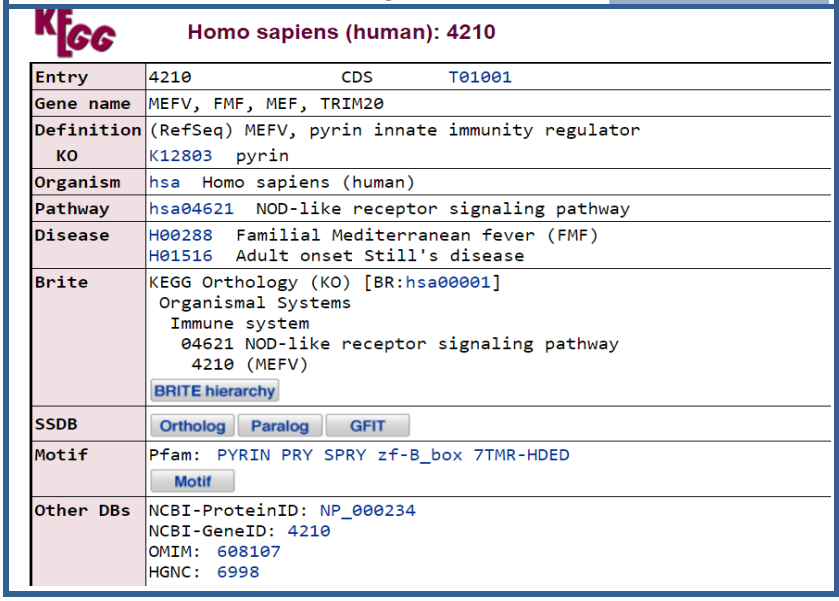
GeneWeaver is a data storage system that helps researchers to get relation of different biological functions with genes. For example if we search for a gene named "MEFV" it shows all meta-data fields such as descriptions for all alleles of the gene, publication information, NCBO annotator and associated diseases ontology terms.



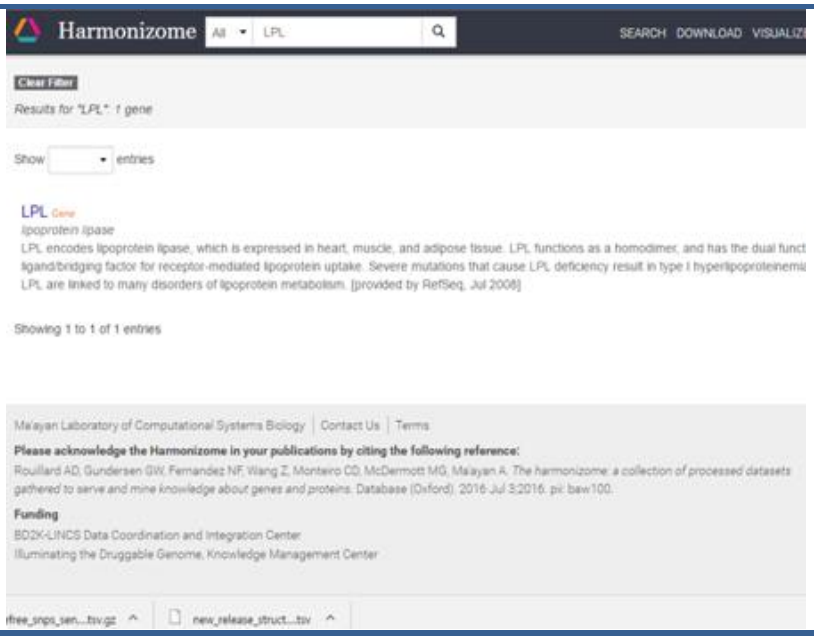
MSigDB that is another database management system for well-annotated gene sets. This storage system results in seven gene-sets for each query and displays all related biological processes. It provides information about gene name, gene id, description of the gene, collections, organism to which it belongs etc.



NCBI's The Gene Expression Omnibus (GEO) repository distributes gene expression data generated by DNA microarray technology. This data storage system shows gene annotation, organism, associated disease name, organism, reporter, data set type, etc. and visualisation of data at individual gene levels.



KEGG (Kyoto Encyclopedia of Genes and Genomes) database provides information at genomic level and analyses gene function relating to genomes. We tested the database by entering gene ID=4210 and it showed gene name, gene description, organism, diseases, and other databases references.



Harmonizome is considered as latest work done for gene associated diseases data set that has gathered data from over 70 major online resources and mine gene based knowledge. The data is stored in the form of relational database and it shows limitations as the data grown bigger and distributed. A user can only search using gene name as it is the row key. When we searched gene name= "LPL" it shows protein name the gene encodes, organ names where that protein is expressed and disease description that can be associated to mutation in LPL.

The comparative study of above mentioned web platforms shows that not only an integration of factors is required to get insights of gene-disease associations but also some important missing factors can be related while working for gene-disease associations. In our data set these missing factors are termed as 'causative factors' of a disease and 'drugs/treatment' to cure any diagnosed disease. Adding these two factors to get gene-disease associations in our data set opens up a new research area to find relation among causative factors themselves and to help in suggesting drugs for that particular causative factor. Relational databases on the other hand, for storing such a complex, multidimensional, huge sized, distributed data, show certain limitations that need to be addressed. Since no work has been done for storing this type of data sets in NoSQL databases we proposed data model for Neo4j to introduce new queries for this type of data sets. These queries can return fastest, comprehensive and effective results for multidimensional big data and can define relationships among different factors that have an association with another. Neo4j is the latest NoSQL graph based technology for data storage. It stored data in the form of entities and relationship between them. It is java based, highly scalable, reliable, network structured database service that uses object oriented Java API and property graph data model in which relations are class objects. CYPHER is the query language (CQL) used by Neo4j for user queries.

Our research implementation has two major parts. One is the data storage format for gene-diseases associations' data set and the other is writing novel queries for researchers to work upon these lines. This implementation provides researchers a conversion from natural language to cypher queries. First of all we integrated the required data for gene associated diseases from multiple resources available online. The data set includes gene name, gene identity, aliases of gene, description of the gene, gene category, number of SNPs (variations in diseases), disease id, disease name associated, description of the disease, chromosome of gene, position of gene in chromosome, alternative lengthening of tolemere (ALT) of a gene, causative factors of a disease and drug families that can be suggested. Using Neo4j we implement gene associated diseases data set in graphical form. Our proposed data model in Neo4j defines four entities 'Genes', 'Diseases', 'Causative Factors', 'Drugs' with their possibly defined attributes such as Gene ID (gid), Gene Name (gname), Gene category (category), Gene Description (g_description), Chromosome to which gene belongs, chromosomal position of a gene (pos), Alternative Length of tolemere for a gene (ALT), alternative form of a disease (NoOfSNPs), Disease ID (did), Disease Name (dname), and Disease Description (d_description). 'Gene' entity shows a relationship 'Associated With' towards 'Diseases' entity and the relationship type is many to many. Because one gene or allele of a gene can cause multiple diseases while on the other hand one disease may be generated because of one gene in one terrestrial are and because of another gene in another terrestrial area. Similarly one disease can have multiple causative factors and one causative factor can cause multiple diseases. And one drug/treatment can be used for multiple diseases or one disease may need to be treated by multiple drugs. So the relationship type between all entities is 'many to many'. The description of our data model to be implemented in neo4j is shown below in Figure 1.

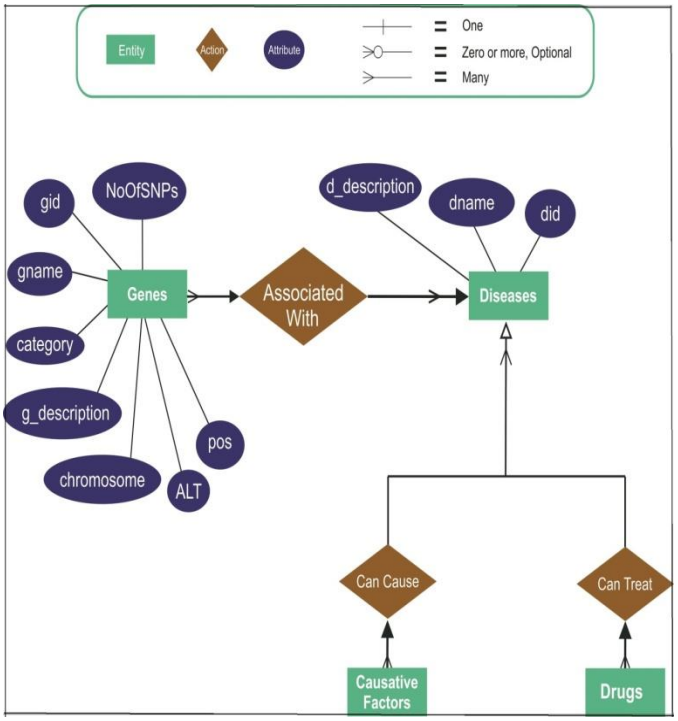


Fig. 1. NoSQL Data Model for gene-disease associations in Neo4j

IV. Results and Evaluation

To apply cypher queries it is necessary to load .csv data file from local file system path neo4j loads files by default from. For that purpose, a directory is created in 'C://Documents/Neo4j/default.grapghdb/import' (by default installation path of Neo4j in windows) and '.csv file' or data file is loaded in it. It is necessary to mention fieldterminator as comma, tab, semicolon, space or any particular character on which you are classifying indexing from .csv file. "Load csv" command along with path of file in local disk storage system is used to load file into neo4j cache. It is necessary to mention a variable for indexing (e.g. row in the query below) after "load csv file://path" and field terminator is applied with reference to this. "Create command" generates nodes for one or more column type/s and relationship between two different columns (entities) must be defined here. "Match command" is used to compare entity relationship, against a particular entity or value defined in the query. Since our graphical NoSQL data model shows inter relationship between entities we have written new cypher queries for our data set. For example cypher query to get associated disease name against gene name 'A4GALT' is shown below in Figure2. Column [0] is the first column of .csv file that contains all entries for gene names and accordingly column [6] contains all diseases names entries. The output of this query is to generate all disease names as nodes for which gene name = A4GALT from samplegene.csv file as shown below in Figure 3. At the end of Figure 3 it says 8 nodes and 0 relationships. Field terminator in the query is mentioned as comma for comma separated samplegene.csv file. Similarly if we want to see all genes belonging to a particular chromosome='12' then cypher query will be written as shown below in Figure 4.

The chromosome factor in the file is at column [10] and it is related to gene names (column [0]) with 'has_genes' relationship where return (keyword) contains both chromosome name nodes as well as gene names nodes. The output of this query is shown below in Figure 5 resulting in total 296 nodes with 145 gene names and others are diseases. Similarly cypher query can also be written to define multiple relationships between nodes in one query. For example for one gene id 'gid=29974' that has a relation of 'Associated_Diseases' with particular disease names (dname) and each disease names has a relation termed as 'due_to' with its related causative factors that may have caused this disease as defined in our samplegene.csv file (data set file). Cypher query to define these relationships for gene-disease associations in the data set is shown below in Figure 6. The output of this query is shown below in Figure 7 resulting in 102 nodes display having 68 relationships, where 34 node pairs have 'Associated_Diseases' relationship and 34 node pairs have 'due_to' relationship. A similar command can also be written that shows "can_be_treated_with" relationship to suggested drugs/treatment for each causative factor as defined in data set file.

```
$ load csv from "file:///D:/samplegene.csv" as row fieldterminator ","
  CREATE (gid{gname:row[0]})-[:Associated_Diseases]->
  (did{dname:row[6]})with gid match (did)<-[:Associated_Diseases]-
  (gid)where gid.gname='A4GALT' return did
```

Fig. 2. Cypher query to return disease names for a particular gene name= A4GALT

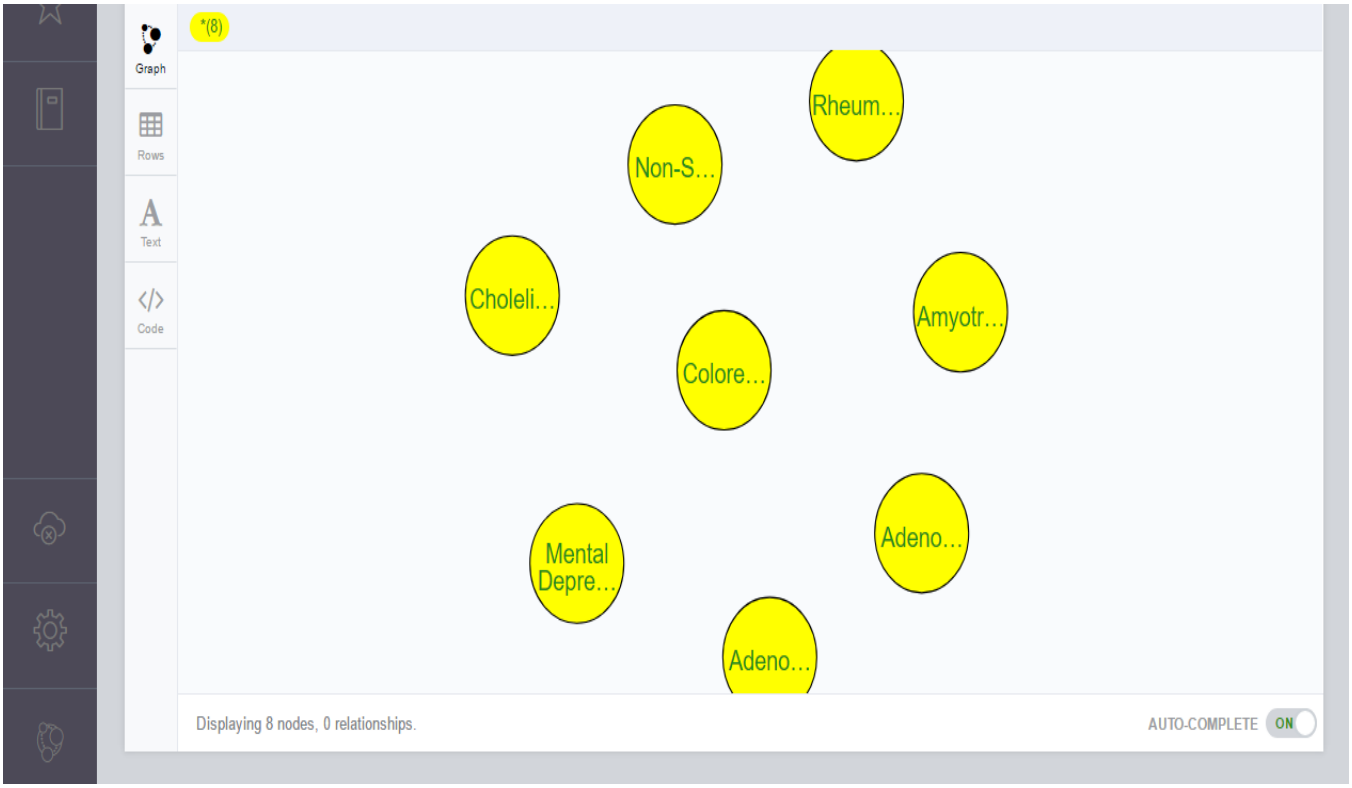


Fig. 3. All diseases names nodes against gene 'A4GALT' using 'Associated_Diseases' relation in neo4j

Fig. 4. Cypher query to return all gene names and chromosome name to which they belong

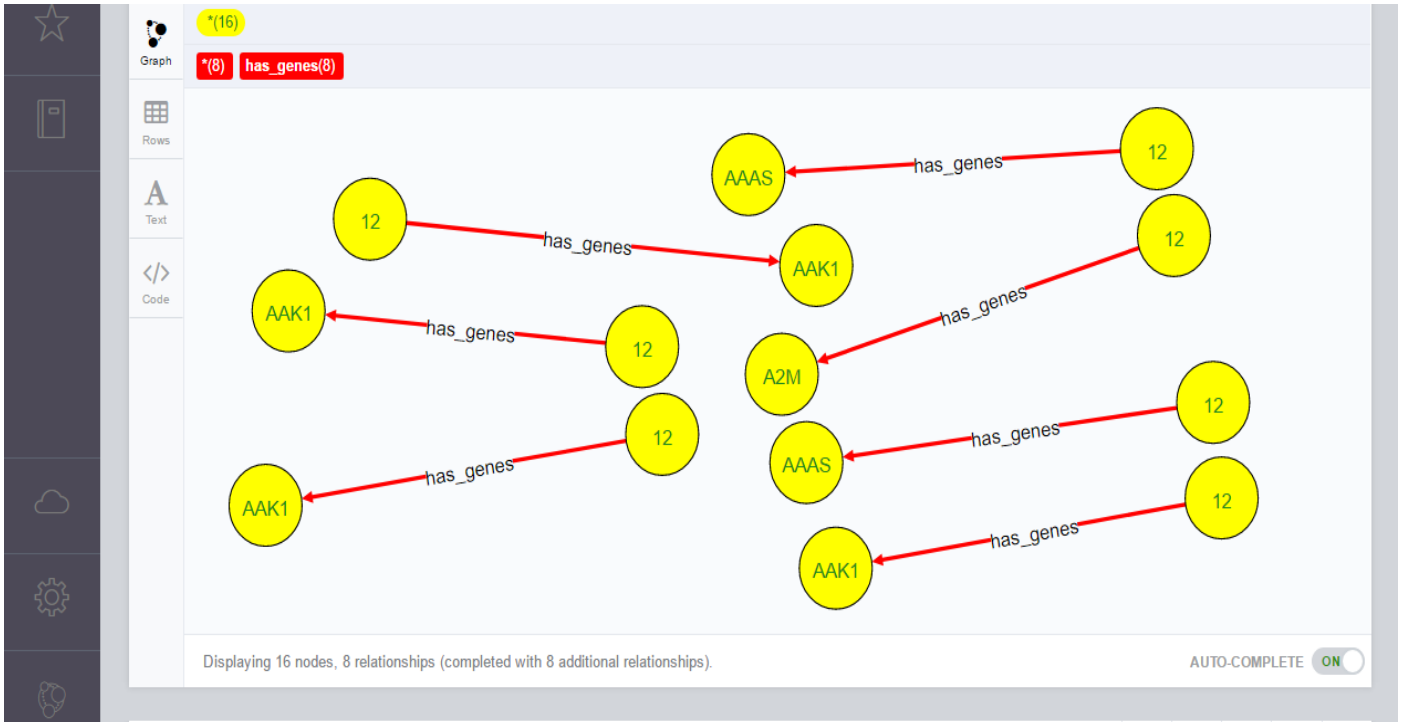


Fig. 5. Returning all genes names and chromosome name using 'has_genes' relation in neo4j

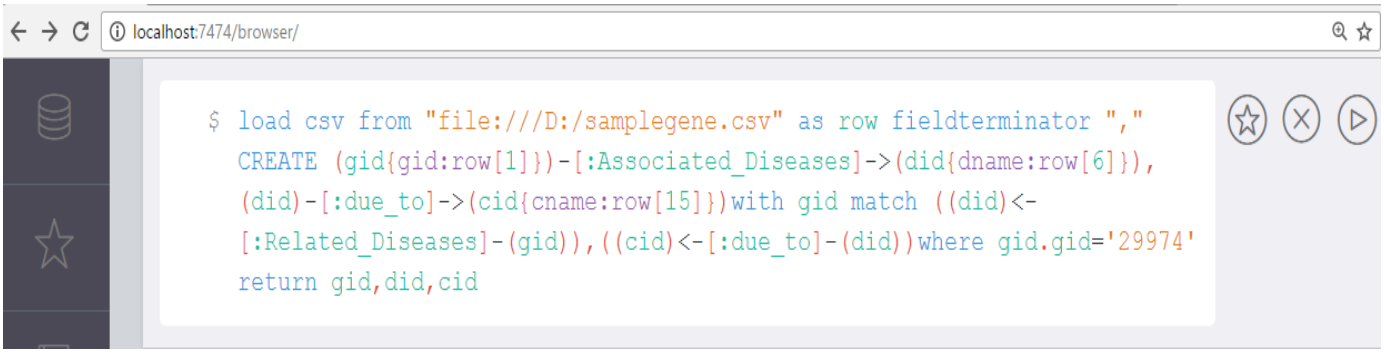


Fig. 6. Cypher query to return multiple relationships among different entities with respect to a particular gene ID

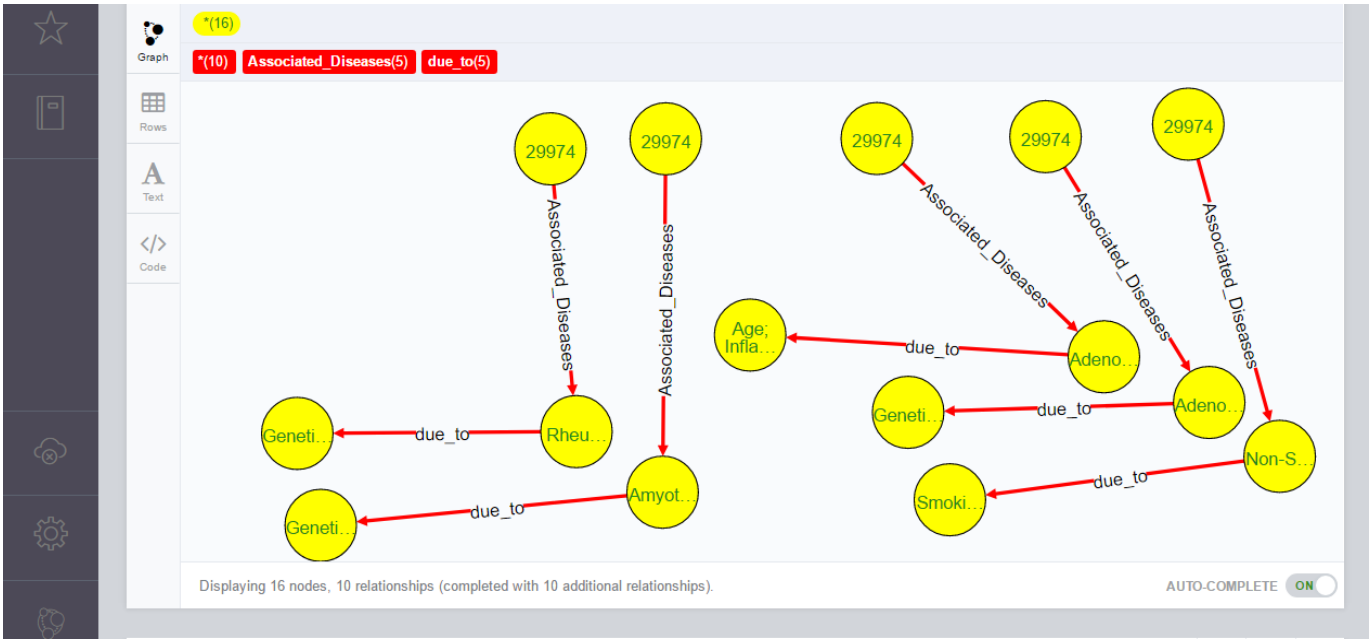


Fig. 7. Returning 'Associated_Diseases' relationship between gene id='29974'and relating disease names and 'due_to' relationship between disease name and its causative factors

This novel cypher query model can visualize relationships among different gene-disease factors, such as gene name and chromosomal position of that gene causing one or more associated diseases. This query model is a unified graphical representation of associations among gene and disease factors from all well-known data sources. This query model can find the following associations:

- Gene name or gene ID that cause one or more diseases
- One disease that may occur due to one or more genes
- Chromosome name where gene resides, position of a gene on chromosome, gene category and gene description to associate with linked diseases (for example nephritic syndrome must cause high blood pressure)
- All causative factors of a disease
- Possible drugs in case of clinical disease and treatment in case of psychological disease

It is concluded from the above research work implementations that gene-disease associations or any data set of this type can be better stored in graphical form of NoSQL databases. Graphical data storage format provides an easy to understand clear cut picture of all types of relations among entities. Novel cypher queries written for this data set can help researchers to relate gene name, gene ID, its chromosomal position, alternative length of gene tolemere, related diseases, disease description, disease variations, possible causative factors and drugs for clinical symptoms or treat for psychological disease symptoms with one another. By taking these queries into consideration, novel cypher queries for an extended gene-disease associations' data set and/or this type of data set can be defined. These queries are effective than most of the existing relational databases for showing special gene-disease associations.

Future work may include finding relationships among diseases and among causative factors to make better decisions for drugs/treatment to cure a disease. There could be different causative factors that may cause a genetic disease other than an inherited gene mutation and physicians can suggest preventive treatment/drugs or symptomatic treatment/drugs according to the found association for a particular disease. This representation of gene-disease associations can also help researchers to relate functional protein of a gene and associate protein-protein interaction to find candidate genes that can cause diseases.

REFERENCES

[1] C. Perez-Iratxeta, M. Wjst, P. Bork, and M.A. Andrade, "G2D: a tool for mining genes associated with disease" BMC genetics, vol. 6(1), pp. 45, 2005.

[2] K. Mirnics, F. A. Middleton, D. A. Lewis, and P. Levitt, "Analysis of complex brain disorders with gene expression microarrays: schizophrenia as a disease of the synapse" Trends in neurosciences, vol. 24(8), pp. 479-486, 2001.

[3] J. E. Clarridge, "Impact of 16S rRNA gene sequence analysis for identification of bacteria on clinical microbiology and infectious diseases" Clinical microbiology reviews, vol. 17(4), pp. 840-862, 2004.

[4] S. C. Kogan, M. Doherty, and J. Gitschier, "An improved method for prenatal diagnosis of genetic diseases by analysis of amplified DNA sequences" New England Journal of Medicine, vol. 317(16), pp. 985-990, 1987.

[5] C. Wu, I. MacLeod, and A.I. Su, "BioGPS and MyGene. info: organizing online, gene-centric information" Nucleic Acids Res., gks1114, 2012.

[6] G.R. Brown, V. Hem, and K.S. Katz, "Gene: a genecentered information resource at NCBI" Nucleic Acids Res., vol. 43, pp. D36–D42, 2015.

[7] U. Consortium, "The universal protein resource (UniProt)" Nucleic Acids Res., vol. 38, pp. D142–D148, 2010.

[8] E.J Baker., J.J. Jay, and J.A. Bubier, "GeneWeaver: a web-based system for integrative functional genomics" Nucleic Acids Res., vol. 40, pp. D1067–D1076, 2012.

[9] C. Jonquet, NH Shah, and MA Musen, "The open biomedical annotator" Summit on Translat. Bioinformat, San Francisco AMIA, pp. 56-60, 2009.

[10] JD Osborne, J Flatow, M Holko, SM Lin, WA Kibbe, LJ Zhu, MI Danila, G Feng, and RL. Chisholm, "Annotating the human genome with disease ontology" BMC Genomics, vol. 10, pp. S6, 2009.

[11] A. Liberzon., A. Subramanian, and R. Pinchback, "Molecular signatures database (MSigDB) 3.0." Bioinformatics, vol. 27, pp. 1739–1740, 2011.

[12] A.C Zambon, S. Gaj, and I. Ho, "GO-Elite: a flexible solution for pathway and ontology over-representation" Bioinformatics, vol. 28, pp. 2209–2210, 2012.

[13] T. Barrett, and R. Edgar, "Gene Expression Omnibus: Microarray Data Storage, Submission, Retrieval, and Analysis" Methods in enzymology, vol. 411, pp. 352-369, 2006.

[14] M. Kanehisa, and S. Goto, "KEGG: kyoto encyclopedia of genes and genomes" Nucleic acids research, vol. 28(1), pp. 27-30, 2000.

[15] A. D. Rouillard, G. W. Gundersen, N. F. Fernandez, Z. Wang, C. D. Monteiro, M. G. McDermott, and A. Ma'ayan, "The harmonizome: a collection of processed datasets gathered to serve and mine knowledge about genes and proteins" Database, 2016.

[16] D. W. Huang, B. T. Sherman, and R. A. Lempicki, "Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources" Nature protocols, vol. 4(1), pp. 44-57, 2009.

[17] V. Bonifati, P.Rizzuvan, M. J. Baren,O. Schaap, G. J. Breedveld, E. Krieger, and J. W.van Dongen, "Mutations in the DJ-1 gene associated with autosomal recessive early-onset parkinsonism" Science, vol. 299(5604), pp. 256-259, 2003.

[18] Y. Moreau, and L. C. Tranchevent, "Computational tools for prioritizing candidate genes: boosting disease gene discovery" Nature Reviews Genetics, vol. 13(8), pp. 523-536, 2012.

[19] J. Lamb, E.D. Crawford, D. Peck, J.W. Modell, I.C. Blat,M. J. Wrobel, and M. Reich, "The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease" science, vol. 313(5795), pp.1929-1935, 2006.

[20] T.A. Manolio, L. D. Brooks, and F. S. Collins, "A HapMap harvest of insights into the genetics of common disease" The Journal of clinical investigation, vol. 118(5), pp. 1590-1605, 2008.

[21] G. Chen, Z. Wang, D. Wang, C. Qiu, M. Liu,X. Chen, and Q. Cui, "LncRNADisease: a database for long-non-coding RNA-associated diseases" Nucleic acids research, vol. 41(D1), pp. D983-D986, 2013.

[22] W. Cookson, L. Liang,G. Abecasis, M. Moffatt, and M. Lathrop, "Mapping complex disease traits with global gene expression." Nature Reviews Genetics, vol. 10(3), pp. 184-194, 2009.

[23] R. Clarke, J. F. Peden, J. C. Hopewell, T. Kyriakou, A. Goel, S. C. Heath, and D. Bennett, "Genetic variants associated with Lp (a) lipoprotein level and coronary disease" New England Journal of Medicine, vol. 361(26), pp. 2518-2528, 2009.

[24] A. Özgür, T. Vu, G. Erkan, and D. R. Radev, "Identifying gene-disease associations using centrality on a literature mined gene-interaction network" Bioinformatics, vol. 24(13), pp. i277-i285, 2008

[25] J. Little, L. Bradley, M. S. Bray, M. Clyne, J. Dorman, D. L. Ellsworth, J. Hanson, M. Khoury, J. Lau, T. R. O'Brien, N. Rothman, D. Stroup, E. Taioli, D. Thomas, H. Vainio, S. Wacholder, and C. Weinberg, "Reporting, appraising, and integrating data on genotype prevalence and gene-disease associations" American journal of epidemiology, vol. 156(4), pp. 300-310, 2002.

[26] J. C. Denny, M. D. Ritchie, M. A. Basford, J. M. Pulley, L. Bastarache, K. Brown-Gentry, D. Wang, D. R. Masys, D. M. Roden, and D. C. Crawford, D.C., "PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene–disease associations" Bioinformatics, vol. 26(9), pp. 1205-1210, 2010.