Computation Based Disease Associations in Disease Inference

Preeti Kale[1] and Jagannath V Aghav[2]

[1] MIT College of Engineering Pune India
[2] College of Engineering Pune India

Abstract. With advances in sequencing technologies, large amounts of biological data is easily accessible for research. Advanced techniques to extract biologically relevant knowledge from this data have gained importance. In this paper, we discuss the *computation based disease associations* that infer human diseases. The biological knowledge of diseases dictates associations between diseases and biological molecules like genes, miRNAs and proteins as well as the associations between diseases and pathways and phenotypes. It is essential to model and represent this knowledge in a computational form with minimal loss of biological context. Based on biological assumptions and statistical analysis, samples of (disease) affected and normal individuals are compared to generate a hypothesis about the disease as depicted by disease associations. We survey computation based disease associations supported by internal interactions i.e interactions between various biological molecules as well as external interactions i.e interactions between the biological molecules and external factors like environment and drugs.

1 Introduction

Omics data is enormous. In this new age of next generation sequencing huge amounts of data is available which needs to be investigated to arrive at a meaningful logic. There are attempts to resolve the jigsaw puzzle of human disease and genome.

Disease is briefly defined as an abnormal condition of a person. But precise definition is difficult as social and environmental factors play a role in the manifestation of the disease. Generally diseases are classified as monogenic and polygenic diseases. Monogenic diseases are caused by a single mutation on a specific gene like Mendelian diseases. This single mutation has varied clinical phenotypes across patients. This can be a consequence of polymorphic or mutant disease-modifying genes and their interactions with environmental factors [1]. In contrast, polygenic diseases are caused by mutations in multiple genes and their phenotypic expression is often cumulative or cooperative [2]. Many complex diseases (e.g. diabetes, cancer and heart disease) are polygenic diseases.

The main contributions of this work include discussion on methods that identify human disease genes.This is described in the next section followed by a section on explanation of computation based disease associations.

2 Methods that Identify Human Disease Genes

Identification of disease genes is one of the primary problems in human genetics. To find disease genes, genetic linkage interval studies are used and candidate gene loci for a given phenotype is investigated. With advances in high-throughput genotyping technologies, genome-wide association studies (GWAS) [3] are used to study the entire human genome in thousands of unrelated individuals with respect to diseases. But GWAS generates huge amounts of data and finding causal disease genes is a challenging task. GWAS data is often represented by graphs (biological networks) whose nodes are biological molecules like proteins and genes and edges are the interactions between them. Strategies to classify disease gene identification methods are based on the properties of graphs and the underlying biological assumption. Barabasi et.al [4],have loosely classified disease gene identification methods into 3 classes as follows.

2.1 Linkage Methods

These methods assume that direct interaction partners of a disease protein are likely candidates to be associated with the same disease phenotype. Since direct interaction partners of a gene or protein are adjacent nodes in a graph, neighbourhood properties of networks are explored. In linkage methods the biological assumption is that if candidate disease genes interact then their products also interact. This directs us to find associations between diseases and biological molecules reviewed in the next section.

2.2 Modularity Based Methods

These methods are based on the observation that gene products belonging to the same topological, functional or disease module have a high likelihood of being involved in the same disease. They use clustering techniques to identify modules in the biological network. The gene products that do not belong to known disease module are potential novel members. Guilt by association is the underlying biological assumption in modularity based methods.

2.3 Diffusion Based Methods

These methods aim to identify the pathways that are closest to the known disease genes. These methods use network propagation and random walk algorithms discussed in section 5. They start from the seed proteins and iteratively find their way through the network based on probability. Proteins that interact with several disease proteins and those in close network proximity gain high probabilistic weight. The iterations in the network are continued until the network is stabilized. The steady state probability is used to rank the candidate genes as disease genes.

3 Computation Based Disease Associations

Human body as a system has biological molecules like genes, proteins and miRNAs interacting with each other across tissues, organs and cells. These interactions can be either internal interactions i.e interactions that take place between various biological molecules or external where the interactions take place between the biological molecules and external factors like environment and drugs.The study of each of these type of interactions enables us to understand systems biology in a better way. The following figure 1 shows computation based disease associations.

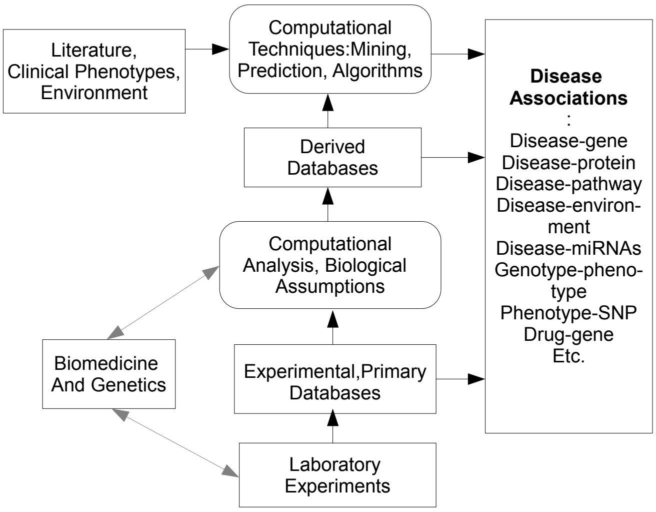


Fig. 1. Computation Based Disease association

As depicted in figure 1 databases for many disease associations are available [5–12] and are discussed in this section. Similarly literature and environment based computational techniques shown in figure 1 are described.

3.1 Disease Gene Association

The association links disease and genes that cause the disease. Disease gene association are widely studied and stored in databases like NIH-GAD Genetic Association database [6], UniProt [12] Online mendelian inheritance in man (OMIM) [7] etc. NIH GAD is an archive for human genetic association studies of complex diseases and disorders. The archive uses standardized molecular nomenclature and is gene centered. Phenotypic descriptions are captured at multiple levels enabling disease gene association. UniProt provides accurate annotated protein sequence knowledge base. Integration with other databases is facilitated by ID mapping and cross references and disease genes associations are identified. OMIM is a comprehensive knowledge base of human genes and diseases and is integrated with Entrez databases.

A cytoscape plug-in DisGeNet [13]integrates disease gene associations from UniProt [12], Online mendelian inheritance in man (OMIM) [7], Pharmacogenomics Knowledge Base (PHARMGKB) [14], Comparative Toxicogenomics Database(CTD) [15]and literature-derived human gene-disease network [16] covering different biomedical aspects of diseases. The functional analysis on integrated data reveals novel biological insights in human disease inference.

Disease gene associations are extracted by literature mining using MeSH annotations in Medline and PubMed as cited in [17]. In [18], authors present a new approach to predict disease gene associations based on text mining and network analysis. Network properties like degree,eigenvector,closeness and betweenness centrality are used to rank disease genes.

3.2 Disease Protein Association

The central dogma of molecular biology describes that DNA is transcribed into RNA which is translated into protein. These translated proteins play an important role in disease manifestation and disease phenotype.

The association between proteins and diseases can be extracted from many popular databases like OMIM [7],UniProt [12],GeneCards [19] by ID mapping and cross references. HDAPD [20] is a web tool for searching disease associated protein structures. In [21],disease protein association is extracted from GeneCards [19] and used for associating protein complexes and diseases.

3.3 Disease Environment Association

Diseases are caused by both genetic and non genetic factors like environment. Many diseases are caused by a combined influence of both genetic and environmental factors [22]. In [23] disease environment association is discussed. The paper describes the aspects of disease environment association like strength, consistency, coherence, temporality and biological gradient. These aspects determine causation of a disease. In [24] authors discuss disease environment association with modern day perspective and its significance. These studies show that finding disease environment association is a challenging task due to its noisy and confounding nature.

In-spite of the challenges, there are attempts to capture disease environment associations in a computational form. This is done by literature mining. In [25],the environmental etiological factors are found by extensive analysis of

Mesh (Medical Subject Heading) annotation in MEDLINE database. The authors semantically classify the environmental factors using UMLS(Unified Medical Language System).They integrate diseases with both genetic and environmental factors to get disease environment association. This modelling of gene environment interactions is simple and may be enhanced.

3.4 Disease Pathway Association

Biological pathways refer to a set of related genes which serially interact among each other. These interactions may lead to undesirable changes in a cell causing diseases. Hence analysis of disease pathway relation helps finding new factors causing diseases.

Examples of pathway databases include KEGG pathway [10],Reactome [26] and GeneMAPP [27].These databases are used to find disease pathway associations. In [17],authors collect biological pathways data from databases like BioCarta, GenMAPP, GeneGo etc. They evaluate the overlap between a disease and a pathway in terms of their constituent genes. They further construct a disease network based on shared pathways to get potential candidates for novel disease relationships.

3.5 Genotype Phenotype Association

Genotype phenotype association is a topic of intense research due to its direct relation to manifestation of diseases. Genotype codes for a phenotype. Phenotype describes the observable characteristics of an organism while genotype is the genetic make up [28]. Databases like OMIM [7] and dbGaP [8] archive and manage genotype phenotype associations. But finding the precise genotype phenotype relationship poses a challenge.

Genome-wide association studies (GWAS) produce huge amounts of data which has to be analysed to identify the genotype and the underlying phenotypic variation. Biobanks are formed to collect and store genetic information of individuals. These biobanks are linked to electronic health record systems to effectively retrieve phenotypic information. In [29], authors propose mining the human phenome via a reverse GWAS or a PheWAS (Phenome Wide Association Scans) to find genotype phenotype associations. PheWAS is possible by using semantic web technologies to link heterogeneous data by Resource Description Framework (RDF)data model discussed in [30].

Another challenge to identify genotype phenotype association is due to the univariate nature of GWAS. Univariate refers to the focus of GWAS on a single phenotype. Although phenotypic information is multivariate, it is converted into a single phenotypic composite score. Hence a computationally efficient multivariate procedure that performs well is required for GWAS. Examples of such multivariate techniques are TATES: Trait-based Association Test that uses Extended Simes procedure [31] and MultiPhen [32].

As discussed in this section genotype phenotype association is being evaluated and refined by various studies so that a better understanding of human diseases is possible. In [33],authors use genotype phenotype data to construct an assembled network. They further decompose the network into modules by identifying and prioritizing the candidate disease genes.

3.6 Disease MicroRNA Association

MicroRNAs (miRNAs) [34] are a class of small non-coding RNAs (22 nucleotides) which normally function as negative regulators of target mRNA expression at the post transcriptional level. MicroRNAs are critical in important biological functions like tissue development and embryonic development. The mutation of these miRNAs may result is dysfunction of miRNA leading to manifestation of diseases.

Due to significant association between miRNA and diseases, many ways of finding disease miRNA association are being explored. Human miRNA disease database (HMDD) [35] and miR2Disease [11] are manually curated disease miRNA association databases.

Efficient and feasible computational methods for predicting potential disease-miRNA associations have gained importance. In [36], authors apply random-walk on OMIM disease similarity network to predict potential OMIM disease-miRNA associations. The assumption is that functionally related miRNAs are often associated with phenotypically similar diseases.

In [37] potential miRNA-disease interactions are found by implementing random walk on the miRNA-miRNA functional similarity network. They adopt global network similarity measures in random walk with restart for miRNA-disease association (RWRMDA) technique. Similarity based measures are presented in [38] to predict new miRNA disease associations.

3.7 Phenotype SNP Association

SNPs i.e single nucleotide polymorphisms are mutations in genes leading to genetic variants. SNPs are associated with Mendelian diseases but how far they affect complex diseases is not completely known. GWAS has hugely enabled the study task of finding associations between phenotype(trait) and SNPs. As discussed in [39] , using phenotype SNP association leads to uncovering similarities between multiple traits suggesting alternative approaches to prioritising disease genes.

Finding phenotype SNP associations is a challenging task. With thousands of SNP markers being tested simultaneously in GWAS, it is important to filter out those SNPs that may lead to false associations as discussed in [40]. Refinements using Linkage Disequilibrium and four levels of analysis to encompass SNP, SNP block, gene, and pathway level comparisons is suggested by [41].Johnson et.al [9] collected available results from 118 GWAS articles into a database of 56,411 significant SNP-phenotype associations. This database is freely available.

3.8 Drug Gene Association

Understanding diseases leads to finding drugs that have the required therapeutic effect on genes. Hence, finding drug gene association is essential. Databases like DrugBank [42] and Drug2Gene store drug gene association information.

Computational methods are developed and used to find drug gene association.In [43] drug target interaction data is integrated to construct a heterogeneous network. A statistical model called Semantic Link Association Prediction (SLAP) is developed to assess the association of drug target pairs and to predict missing links.

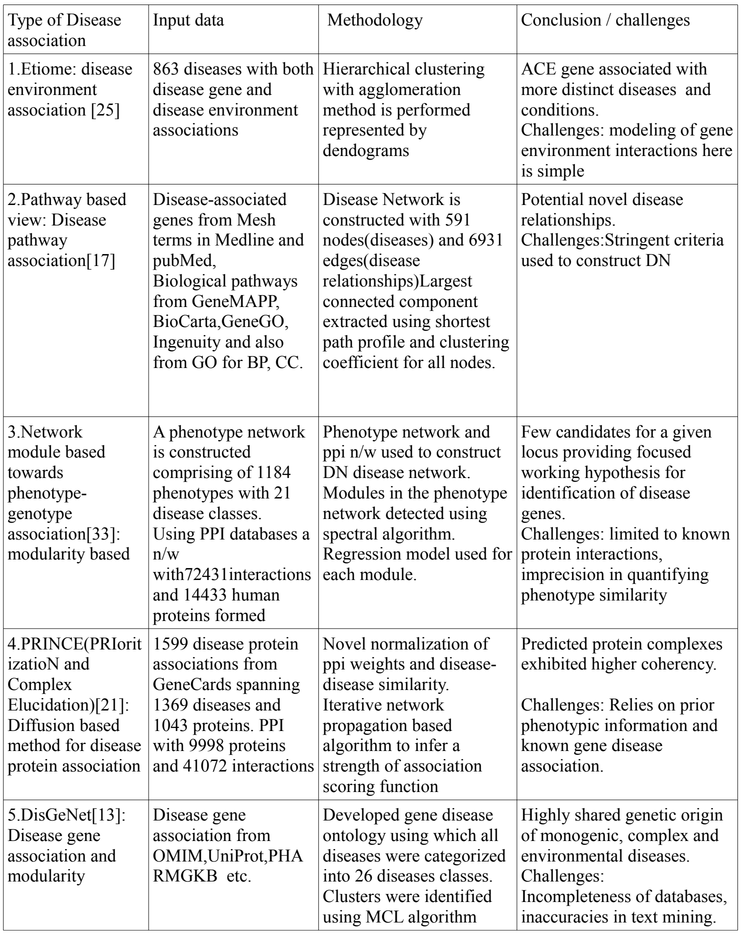
By incorporating multiple biological information sources the accuracy of drug target association and prediction can be improved. In [5] a novel framework, Similarity based Inference of drug TARgets(SITAR) is introduced. It incorporates multiple drug-drug and gene-gene similarity measures for drug target prediction. The assumption is that similar drugs tend to share similar targets.

4 Discussion and Conclusion

This work surveys computation based disease associations in the perspective of different biological molecules and external factors.Each disease association implicates biological molecules like genes, proteins and miRNAs as well as pathways, SNPs and environmental factors in the manifestation of the disease. Though we have included many computation based disease associations, the list can be extended with associations based on tissues,metabolism,functional genes etc. Combining disease associations may give useful insight in the process of disease manifestation.

Table 1 compares some computation based disease associations discussed in section 2 and section 3.Etiome [25] forms disease environment associations based on 863 diseases. It implicates ACE gene with most distinct diseases. Though modelling of gene environment interaction is simple, it shows the importance of linkage based analysis. Disease pathway association [17]is calculated using protein context and disease context indices. These are used to construct the disease network with 591 nodes and 6931 edges.Methods to relax the stringent criteria used for disease network construction may be investigated keeping in view its potential to find novel disease relationships. Third row of table 1 mentions a network phenotype genotype association [33] which may be categorized as modularity based approach for disease gene identification.Using spectral algorithm and phenotype network a locus for candidate disease genes is identified. Prince [21] is a diffusion based method and uses disease protein association. DisGeNet [13] uses Markov clustering to identify 26 disease classes.The results using DisGeNet show a shared genetic origin of monogenic, complex and environmental diseases. Table 1 summarises the disease association based methods, the methodologies used for disease inference and the challenges they pose. This understanding may enable us in making biological assumptions with higher confidence as we strive for minimal loss of biological context.

Table 1. Disease association for disease identification



Type of Disease association	Input data	Methodology	Conclusion / challenges
1.Etiome: disease environment association [25]	863 diseases with both disease gene and disease environment associations	Hierarchical clustering with agglomeration method is performed represented by dendograms	ACE gene associated with more distinct diseases and conditions. Challenges: modeling of gene environment interactions here is simple
2.Pathway based view: Disease pathway association[17]	Disease-associated genes from Mesh terms in Medline and pubMed, Biological pathways from GeneMAPP, BioCarta,GeneGO, Ingenuity and also from GO for BP, CC.	Disease Network is constructed with 591 nodes(diseases) and 6931 edges(disease relationships)Largest connected component extracted using shortest path profile and clustering coefficient for all nodes.	Potential novel disease relationships. Challenges:Stringent criteria used to construct DN
3.Network module based towards phenotype-genotype association[33]: modularity based	A phenotype network is constructed comprising of 1184 phenotypes with 21 disease classes. Using PPI databases a n/w with72431interactions and 14433 human proteins formed	Phenotype network and ppi n/w used to construct DN disease network. Modules in the phenotype network detected using spectral algorithm. Regression model used for each module.	Few candidates for a given locus providing focused working hypothesis for identification of disease genes. Challenges: limited to known protein interactions, imprecision in quantifying phenotype similarity
4.PRINCE(PRIoritizatioN and Complex Elucidation)[21]: Diffusion based method for disease protein association	1599 disease protein associations from GeneCards spanning 1369 diseases and 1043 proteins. PPI with 9998 proteins and 41072 interactions	Novel normalization of ppi weights and disease-disease similarity. Iterative network propagation based algorithm to infer a strength of association scoring function	Predicted protein complexes exhibited higher coherency. Challenges: Relies on prior phenotypic information and known gene disease association.
5.DisGeNet[13]: Disease gene association and modularity	Disease gene association from OMIM,UniProt,PHARMGKB etc.	Developed gene disease ontology using which all diseases were categorized into 26 diseases classes. Clusters were identified using MCL algorithm	Highly shared genetic origin of monogenic, complex and environmental diseases. Challenges: Incompleteness of databases, inaccuracies in text mining.

Early disease diagnosis may be possible by formalising the properties of each type of computation based disease association. Formalising rules for disease associations is a daunting task but it may assist in dynamic learning enabling disease identification.

References

1. Loscalzo, J., Kohane, I., Barabasi, A.L.: Human disease classification in the postgenomic era: a complex systems approach to human pathobiology. Molecular systems biology **3**(1) (2007)
2. Goh, K.I., Choi, I.G.: Exploring the human diseasome: the human disease network. Briefings in functional genomics **11**(6) (2012) 533–542
3. Hirschhorn, J.N.: Genomewide association studies–illuminating biologic pathways. (2009)
4. Barabási, A.L., Gulbahce, N., Loscalzo, J.: Network medicine: a network-based approach to human disease. Nature Reviews Genetics **12**(1) (2011) 56–68
5. LIAT PERLMAN, ASSAF GOTTLIEB, N.A.E.R., SHARAN, R.: Combining drug and gene similarity measures for drug-target elucidation. JOURNAL OF COMPUTATIONAL BIOLOGY **18**(2) (2011) 133145
6. Becker, K.G., Barnes, K.C., Bright, T.J., Wang, S.A.: The genetic association database. Nature genetics **36**(5) (2004) 431–432
7. Hamosh, A., Scott, A.F., Amberger, J.S., Bocchini, C.A., McKusick, V.A.: Online mendelian inheritance in man (omim), a knowledgebase of human genes and genetic disorders. Nucleic acids research **33**(suppl 1) (2005) D514–D517
8. Mailman, M.D., Feolo, M., Jin, Y., Kimura, M., Tryka, K., Bagoutdinov, R., Hao, L., Kiang, A., Paschall, J., Phan, L., et al.: The ncbi dbgap database of genotypes and phenotypes. Nature genetics **39**(10) (2007) 1181–1186
9. Johnson, A.D., O'Donnell, C.J.: An open access database of genome-wide association results. BMC medical genetics **10**(1) (2009) 6
10. Kanehisa, M., Goto, S.: Kegg: kyoto encyclopedia of genes and genomes. Nucleic acids research **28**(1) (2000) 27–30
11. Qinghua Jiang, Yadong Wang, Y.H.e.a.: mir2disease: a manually curated database for microrna deregulation in human disease. Nucleic Acids Research **37**(suppl 1) (2009)
12. Apweiler, R., Martin, M., O'Donovan, C., Magrane, M., Alam-Faruque, Y., Antunes, R., Barrell, D., Bely, B., Bingley, M., Binns, D., et al.: Ongoing and future developments at the universal protein resource. Nucleic Acids Research **39**(Database issue) (2011) D214–9
13. Bauer-Mehren, A., Bundschus, M., Rautschka, M., Mayer, M.A., Sanz, F., Furlong, L.I.: Gene-disease network analysis reveals functional modules in mendelian, complex and environmental diseases. PloS one **6**(6) (2011) e20284
14. Altman, R.B.: Pharmgkb: a logical home for knowledge relating genotype to drug response phenotype. Nature genetics **39**(4) (2007) 426
15. Mattingly, C.J., Rosenstein, M.C., Davis, A.P., Colby, G.T., Forrest, J.N., Boyer, J.L.: The comparative toxicogenomics database: a cross-species resource for building chemical-gene interaction networks. Toxicological sciences **92**(2) (2006) 587–595
16. Bundschus, M., Dejori, M., Stetter, M., Tresp, V., Kriegel, H.P.: Extraction of semantic biomedical relations from text using conditional random fields. BMC bioinformatics **9**(1) (2008) 207

17. Li, Y., Agarwal, P.: A pathway-based view of human diseases and disease relationships. PloS one **4**(2) (2009) e4346
18. Özgür, A., Vu, T., Erkan, G., Radev, D.R.: Identifying gene-disease associations using centrality on a literature mined gene-interaction network. Bioinformatics **24**(13) (2008) i277–i285
19. Rebhan, M., Chalifa-Caspi, V., Prilusky, J., Lancet, D.: Genecards: integrating information about genes, proteins and diseases. Trends in genetics: TIG **13**(4) (1997) 163
20. Lin, Y.R., Wei, H.Y., Tsai, T.L., Lin, T.H.: Hdapd: a web tool for searching the disease-associated protein structures. BMC bioinformatics **11**(1) (2010) 88
21. Vanunu, O., Magger, O., Ruppin, E., Shlomi, T., Sharan, R.: Associating genes and protein complexes with disease via network propagation. PLoS Comput Biol **6**(1) (01 2010) e1000641
22. Lichtenstein, P., Holm, N.V., Verkasalo, P.K., Iliadou, A., Kaprio, J., Koskenvuo, M., Pukkala, E., Skytthe, A., Hemminki, K.: Environmental and heritable factors in the causation of canceranalyses of cohorts of twins from sweden, denmark, and finland. New England Journal of Medicine **343**(2) (2000) 78–85
23. Hill, A.B.: The environment and disease: association or causation? Proceedings of the Royal Society of Medicine **58**(5) (1965) 295
24. Lucas, R.M., McMichael, A.J.: Association or causation: evaluating links between" environment and disease". Bulletin of the World Health Organization **83**(10) (2005) 792–795
25. Liu, Y.I., Wise, P.H., Butte, A.J., et al.: The etiome: identification and clustering of human disease etiological factors. BMC bioinformatics **10**(Suppl 2) (2009) S14
26. Matthews, L., Gopinath, G., Gillespie, M., Caudy, M., Croft, D., de Bono, B., Garapati, P., Hemish, J., Hermjakob, H., Jassal, B., et al.: Reactome knowledgebase of human biological pathways and processes. Nucleic acids research **37**(suppl 1) (2009) D619–D622
27. Dahlquist, K.D., Salomonis, N., Vranizan, K., Lawlor, S.C., Conklin, B.R.: Genmapp, a new tool for viewing and analyzing microarray data on biological pathways. Nature genetics **31**(1) (2002) 19–20
28. J, W.D.: Genotype-phenotype relationships. eLS nature publishing group (2001)
29. Denny, J.C., Ritchie, M.D., Basford, M.A., Pulley, J.M., Bastarache, L., Brown-Gentry, K., Wang, D., Masys, D.R., Roden, D.M., Crawford, D.C.: Phewas: demonstrating the feasibility of a phenome-wide scan to discover gene–disease associations. Bioinformatics **26**(9) (2010) 1205–1210
30. Pathak, J., Kiefer, R.C., Bielinski, S.J., Chute, C.G.: Mining the human phenome using semantic web technologies: A case study for type 2 diabetes. In: AMIA Annual Symposium Proceedings. Volume 2012., American Medical Informatics Association (2012) 699
31. van der Sluis S, Posthuma D, D.C.: Tates: Efficient multivariate genotype-phenotype analysis for genome-wide association studies. PLoS Genet **9** (2013)
32. Paul F. O'Reilly, Clive J. Hoggart, Y.P.e.a.: Multiphen: Joint model of multiple phenotypes can increase discovery in gwas. PloS one **7**(5) (2012)
33. Jeffrey Q Jiang1, A.W.M.D., Chen2, M.: Towards prediction and prioritization of disease genes by the modularity of human phenome-genome assembled network. Journal of Integrative Bioinformatics **7(2):149** (2010)
34. Bartel, D.P.: Micrornas: genomics, biogenesis, mechanism, and function. cell **116**(2) (2004) 281–297
35. Lu, M., Zhang, Q., Deng, M., Miao, J., Guo, Y., Gao, W., Cui, Q.: An analysis of human microrna and disease associations. PloS one **3**(10) (2008) e3420

36. Chen, H., Zhang, Z.: Prediction of associations between omim diseases and micrornas by random walk on omim disease similarity network. Hindawi Publishing Corporation,The ScientificWorld Journal (2013)
37. Chen, X., Liu, M.X., Yan, G.Y.: Rwrmda: predicting novel human microrna-disease associations. Mol. BioSyst. **8** (2012) 2792–2798
38. Chen, Zhang: Similarity-based methods for potential human microrna-disease association prediction. BMC Medical Genomics 2013 6:12 (2013)
39. et al., W.: ictnet: A cytoscape plugin to produce and analyze integrative complex traits networks. BMC Bioinformatics (2011)
40. Chan, E.K., Hawken, R., Reverter, A.: The combined effect of snp-marker and phenotype attributes in genome-wide association studies. Animal genetics **40**(2) (2009) 149–156
41. Lewis SN, Nsoesie E, W.C.Q.D.Z.L.: Prediction of disease and phenotype associations from genome-wide association studies. PLoSONE **6** (2011)
42. Wishart, D.S., Knox, C., Guo, A.C., Shrivastava, S., Hassanali, M., Stothard, P., Chang, Z., Woolsey, J.: Drugbank: a comprehensive resource for in silico drug discovery and exploration. Nucleic acids research **34**(suppl 1) (2006) D668–D672
43. Chen, B., Ding, Y., Wild, D.J.: Assessing drug target association using semantic linked data. PLoS Comput Biol **8**(7) (07 2012) e1002574