

# Linking Phenome, Genome and Toxicological Space

Daniel Edsgard

Technical University of Denmark, Center for Biological Sequence Analysis,  
Copenhagen, Denmark  
E-mail: edsgard@cbs.dtu.dk

Phenotypic and genomic data as well as chemogenomic approaches can be utilized to link phenotype, genome and toxicological space. An approach to pursue such a phenome-genome-chemical space linkage is outlined, with the application to male reproductive problems being the driving incentive. Due to the diversity of male reproductive problems in terms of the range of clinical manifestations, ontologies and controlled vocabularies is utilized to assist in systematic mining of biological and chemical databases as well as to structure findings.

A first draft of a phenome-genome-chemical space pipeline utilizes the Unified Medical Language System (UMLS) to extract candidate genes associated with a phenotype of interest, protein-protein interaction data to pull-down novel candidate genes, and chemical databases with text-mining based associations between genes and toxicants. The pipeline is also envisioned to be able to integrate other datatypes such as pathway and SNP data to corroborate findings.

## **Linking Phenome and Genome: Extraction of Phenotype Associated Genes by Utilization of Controlled Vocabularies**

The Unified Medical Language System (UMLS) is used to enable a mapping between indices of a number of controlled vocabularies. Such a vocabulary alignment can be utilized to extract a number of candidate genes associated with a disease phenotype. The International Classification of Disease (ICD) is used as a starting point onto which other vocabularies are mapped. Vocabularies of primary interest to be mapped onto ICD are Online Mendelian Inheritance Map (OMIM) and Medical Subject Headings (MeSH). A straightforward approach to extract candidate genes is to retrieve all ICD nodes of a branch of interest, such as the branch associated with male infertility phenotype, and compile a list of OMIM and MeSH identifiers mapped onto those ICD nodes. The list of identifiers is then used to query gene databases that utilizes the vocabulary from which the identifiers are derived.

## **Novel Phenotype Associated genes: Utilization of a Protein-Protein Interaction Network**

The list of candidate genes extracted by help of UMLS is expanded by utilization of a high-confidence interaction network of human proteins[1]. The protein-protein interaction data may introduce proteins that previously have had no known association with the phenotype of interest. The novel proteins found may be corroborated by SNP studies on a group with manifestations of the phenotype versus a control group. Data on localization, temporal expression and in which pathway a protein is found may also be used to further support the putative disease association of a gene.

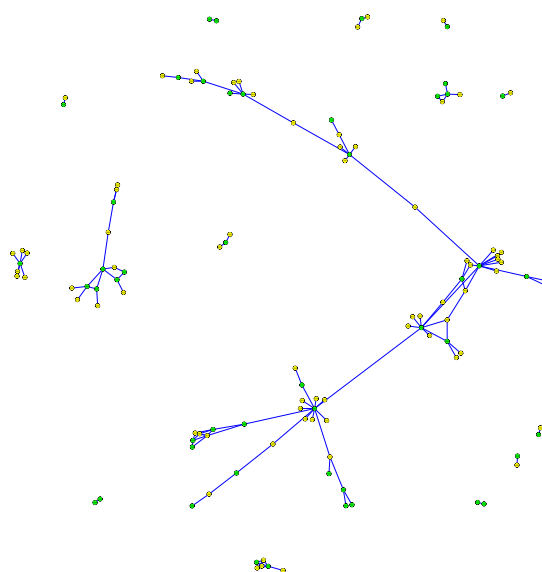


Figure 1: A protein-protein interaction network consisting of proteins known to be involved in male infertility (green nodes) and interaction partners (yellow nodes). Data on which proteins are known to be involved in male infertility were extracted from the OMIM database by making a robot that queried for records in which the term “male infertility” was mentioned.

### Linking Genome and Chemical Space

To identify small molecule compounds that may adversely modulate gene and protein networks, chemical databases containing associations to genes or proteins can be utilized. Querying such a database with the list of genes retrieved by the virtual pull-down from the interactome, a set of chemicals can be extracted. Hypothetically, found candidate toxicants can be validated by studies of food intake or other data on exposure to certain chemical environments and associating this to the phenotype of interest by mining patient registries.

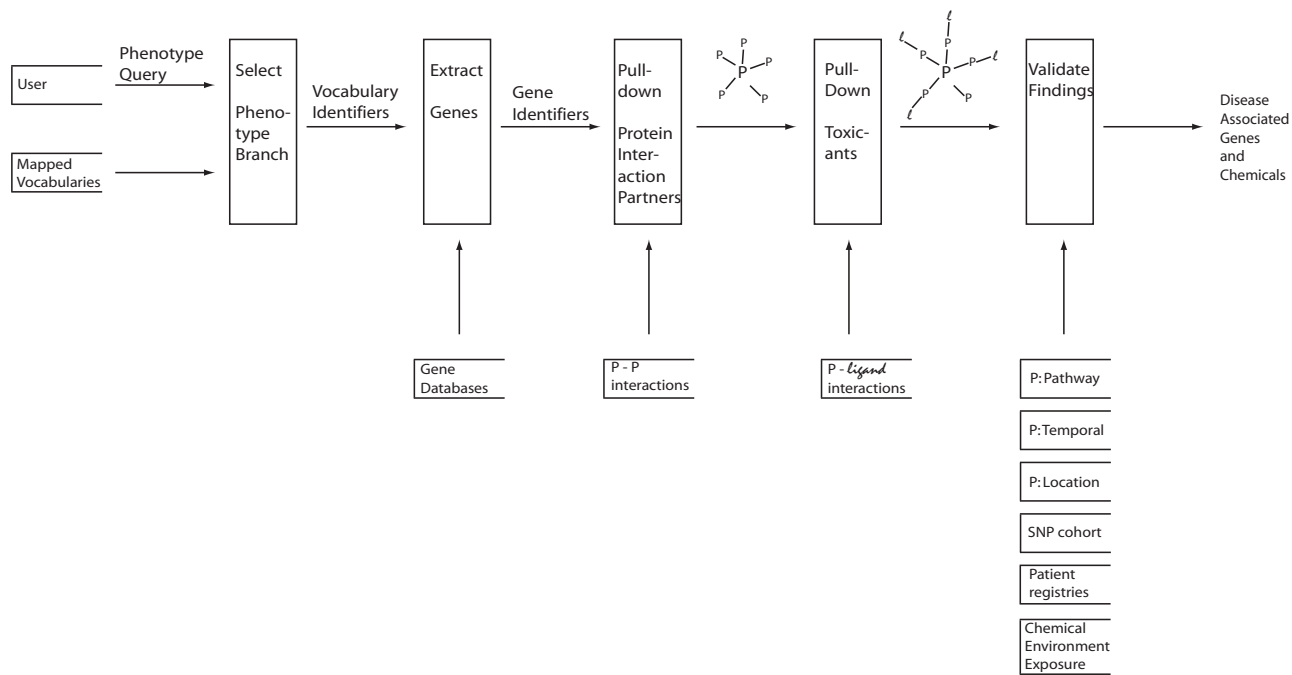


Figure 2: Pipeline for linking phenotype, genome and toxicological space. Abbreviations, P: protein, l: small molecule ligand.

## References

1. Lage, K. *et al.*, A human phenome-interactome network of protein complexes implicated in genetic disorders. *Nat. Biotechnol.* **25**, 309-316 (2007)