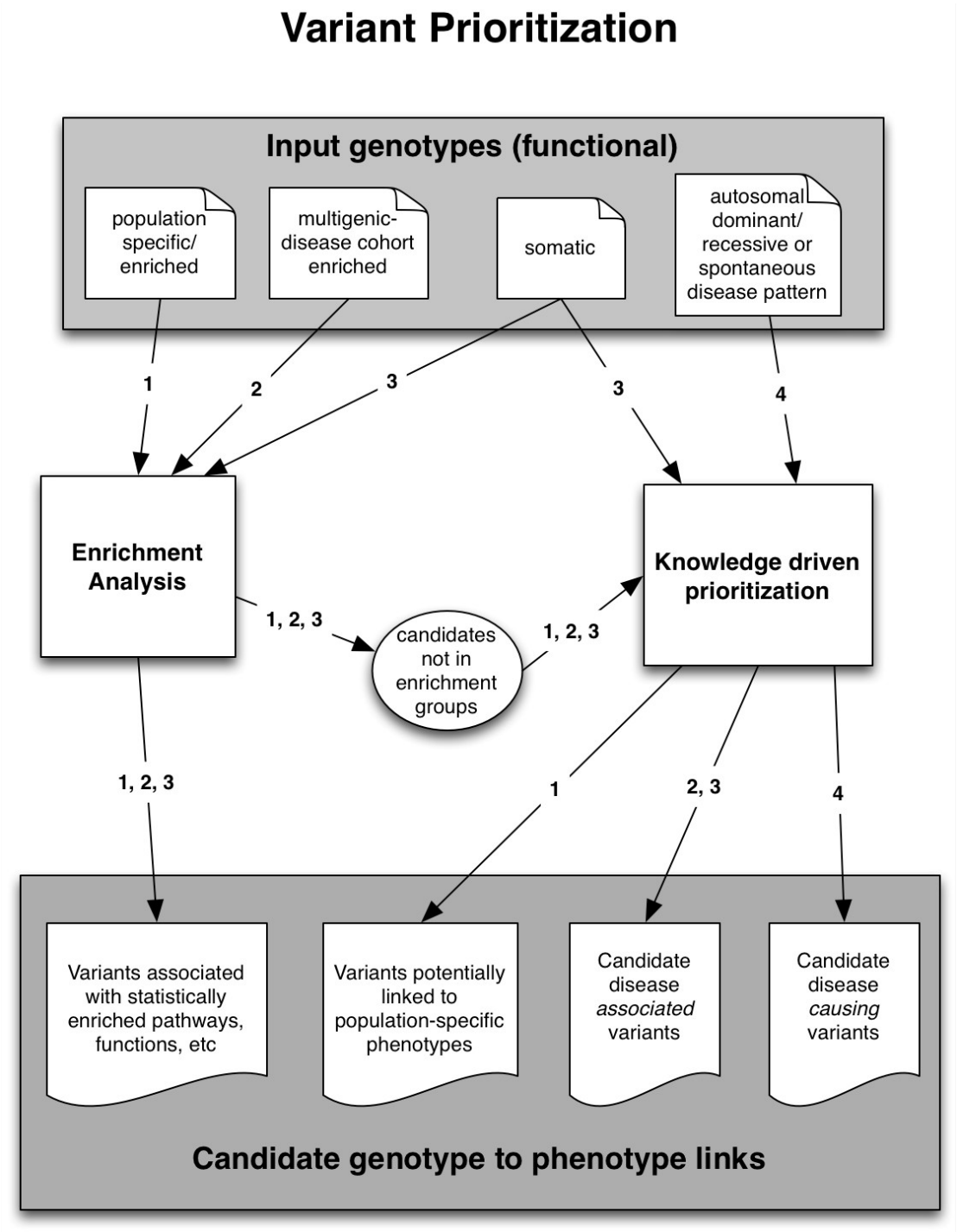
# Phase 4: Variant prioritisation




Even with sound experimental design and variant filtering protocols, whole exome sequencing studies often still produce many more candidates than can be verified experimentally. While possible to rank variants based on predicted impact on the protein, it is not always obvious to identify the strongest candidate(s) for involvement in a disease or phenotype of interest. For that reason, assessing candidate genes bearing functional

variants in the context of existing biomedical knowledge and their known biomolecular functions is an important step in producing a manageable set of variants for further validation or exploration.

Depending on the study, candidates can be evaluated individually or as a set. Typical questions that should be asked at this point (one or more):

1. **Is the variant *itself* known or predicted to be involved the disease of interest?**

   ANNOVAR, recommended in Phase 3 (**add link**) produces ClinVar (http://www.ncbi.nlm.nih.gov/clinvar/) annotations and identifiers that classify variants as disease causing.

2. **Is the variant in a *gene* known to be involved in the disease or in a related disease?**

   Online Mendelian Inheritance in Man (http://www.ncbi.nlm.nih.gov/omim) is an excellent resource for this information. Single genes can be searched using the HUGO gene symbol and selecting “Gene name” in the “Limits” menu. OMIM information for large gene sets can be programmatically obtained by accessing the API (see the ‘Tools” link at the aforementioned URL).

3. **Does the gene have a function that coincides with the pathology, etc?**

   For example, does a candidate gene identified in an inflammatory disorder have known biological roles and functions related to regulation of the inflammatory response? Gene Ontology annotations can be downloaded from: http://geneontology.org/page/download-annotations or a candidate gene’s GO annotations can be searched using AMIGO: http://amigo.geneontology.org/amigo/search/bioentity. To determine whether a set of functional variants contribute to a functions relating to a phenotype of interest, the genes they occur in can be assessed for enrichment of GO functions. The ‘gene set enricher’ web service by the Comparative Toxicogenomics Database (Davis et al, 2015) is an intuitive tool for this purpose, which also produces user friendly outputs that can be imported directly into a spreadsheet software for further filtering or manipulation: http://ctdbase.org/tools/analyzer.go

4. **Is the gene product in a pathway associated with the disease?**

   Genes can be mapped to KEGG pathways: http://www.genome.jp/kegg/ or REACTOME pathways: http://www.reactome.org/. Similarly, gene-to-pathway annotations

can be derived from the Pathway Ontology (Petri et al, 2014). As in (3), it is possible to determine whether a set of variants collectively impact a pathway(s) known to be associated with a disease or trait of interest by performing an enrichment analysis using the specific functionality in the CTD 'gene set enricher' web service.

5. **Does a mutation or animal knockout of the gene cause the disease or a hallmark phenotype of the disease?**

   The Mouse Genome Informatics database enables querying for human-mouse disease and phenotype connections using gene symbols as an input: http://www.informatics.jax.org/humanDisease.shtml (Figure 1) The Human Phenotype Ontology project also provides gene-to-phenotype mappings, which can be used in a similar manner: http://human-phenotype-ontology.org (see Figure 4). As in (3) and (4) phenotype statistical enrichment in gene sets can also be performed using the MamPhEA (Weng and Liao, 2010) webserver: http://evol.nhri.org.tw/phenome (Figure 2).

6. **Is the gene expressed in the tissue or organ of interest?**

   NCBI's Gene Expression Omnibus profiles (http://www.ncbi.nlm.nih.gov/geoprofiles) and the EBI's Expression Atlas (http://www.ebi.ac.uk/gxa) are excellent resources for this purpose.

7. **Does the gene product physically interact with a protein that is encoded by a known disease gene?**

   Much like genes that encode proteins that occur in the same pathway as a known disease gene product may cause the disease if mutated, so too may proteins that physically interact with known disease gene products. The STRING database and associated search tools (Szklarczyk et al, 2015 - http://string-db.org) are powerful resources for identifying interacting partners of a candidate gene's product, or to identify interactions between the products of a set of genes that bear functional variants.

   The relevance of this prioritization method is illustrated in Figure 3, which shows how several of the direct interaction partners of the product of the DMD muscular dystrophy gene are themselves known to cause the disease.

## Summary

Extant knowledge about the molecular and cellular mechanisms involved in a disease or phenotype of interest can be extremely useful to prioritise likely candidates from a list of genes that all bear functional variants. That said, and as mentioned before in the Phase 3 SOP, rarity/novelty along

with predicted deleteriousness and expected segregation with affected/cases and unaffected/control individuals in the study are the primary criteria for producing a candidate list. Therefore, variants should not be discarded as being irrelevant if the knowledge filter does not return disease, phenotype or function links. It is also important to note that all the abovementioned questions will yield useful information. However, in many cases, the insights gained from thoroughly interrogating one knowledge domain provide enough evidence to implicate a variant. For example, the ACTA1 gene in Figure 3 has not previously been implicated as a muscular dystrophy gene, but when following the abovementioned SOP, it is clear that it would be a strong candidate (Fig 4) if a deleterious variant were identified as using the Phase 3 SOP.

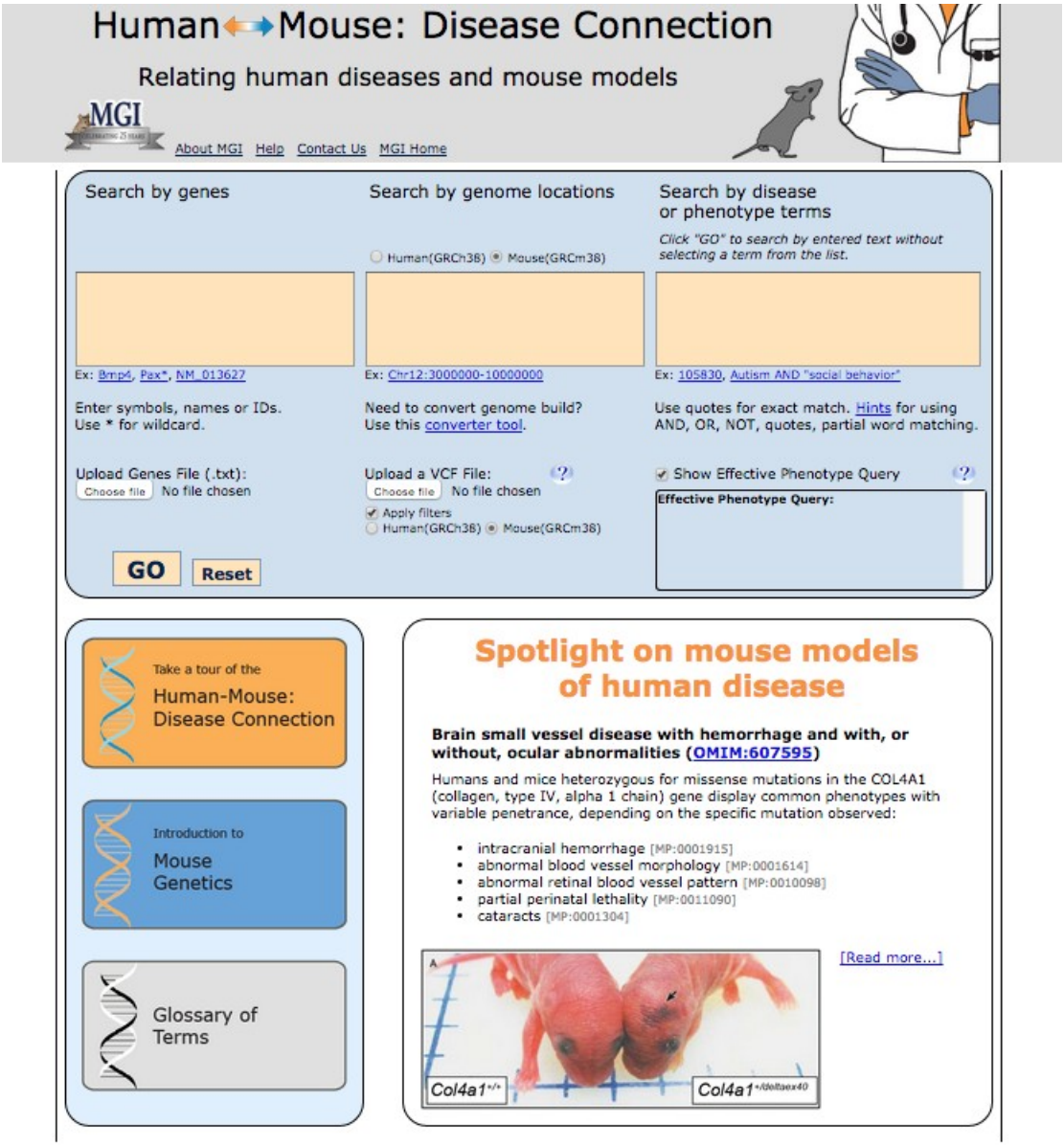



**Figure 1. The Jackson Lab's gene-to-phenotype search tool**

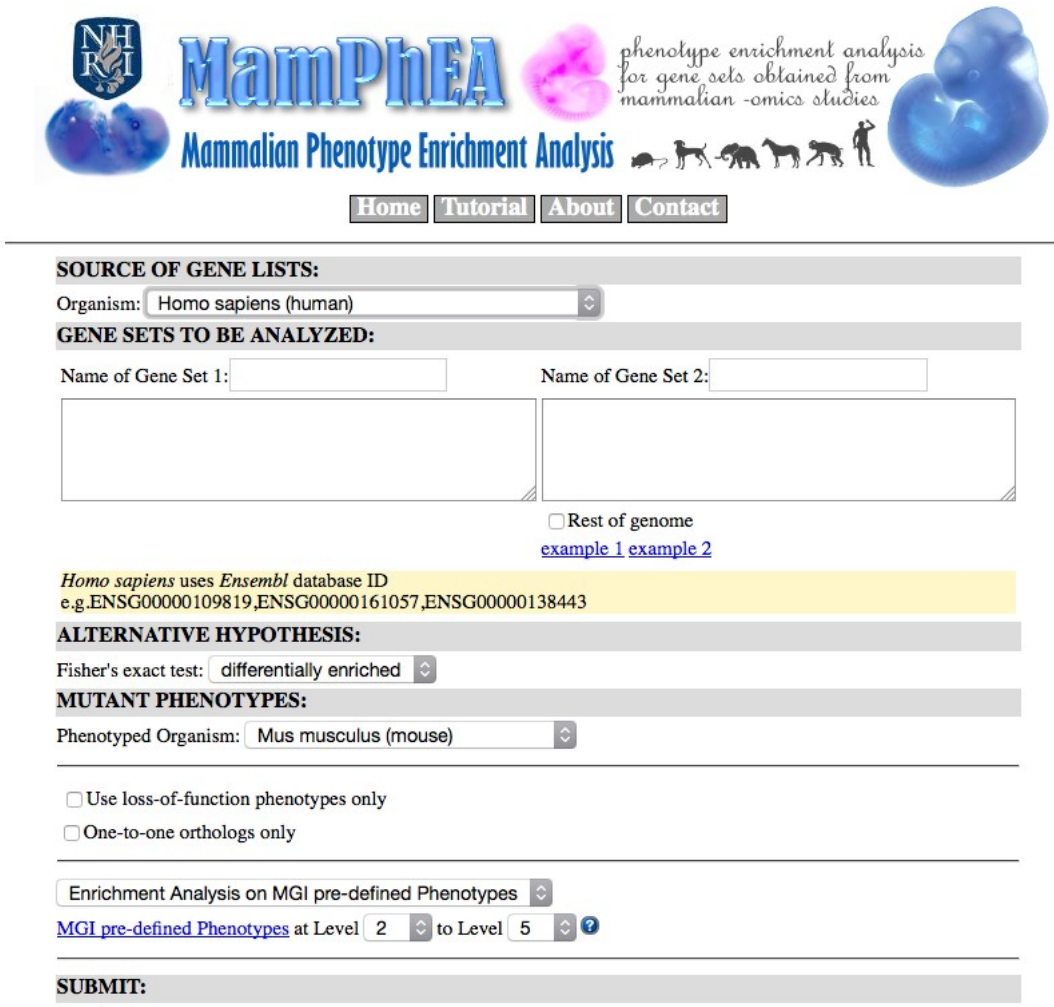



**Figure 2.** A gene set enrichment server based on mouse knockout phenotypes from the Jackson lab.

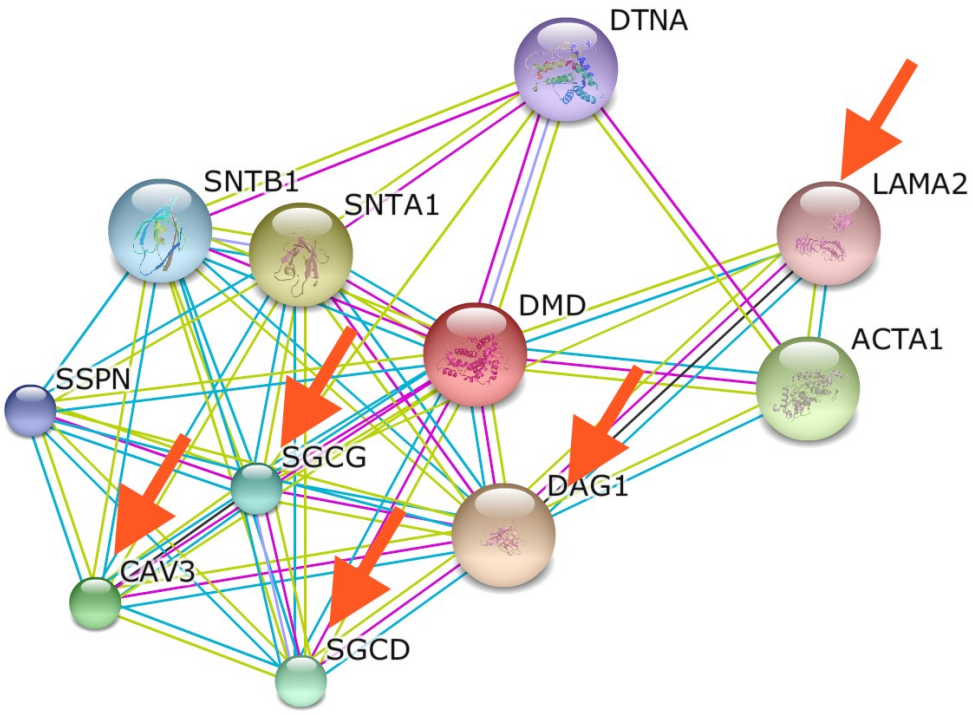



**Figure 3.** Interaction network of the DMD muscular dystrophy protein. Red arrows highlight other known muscular dystrophy proteins.

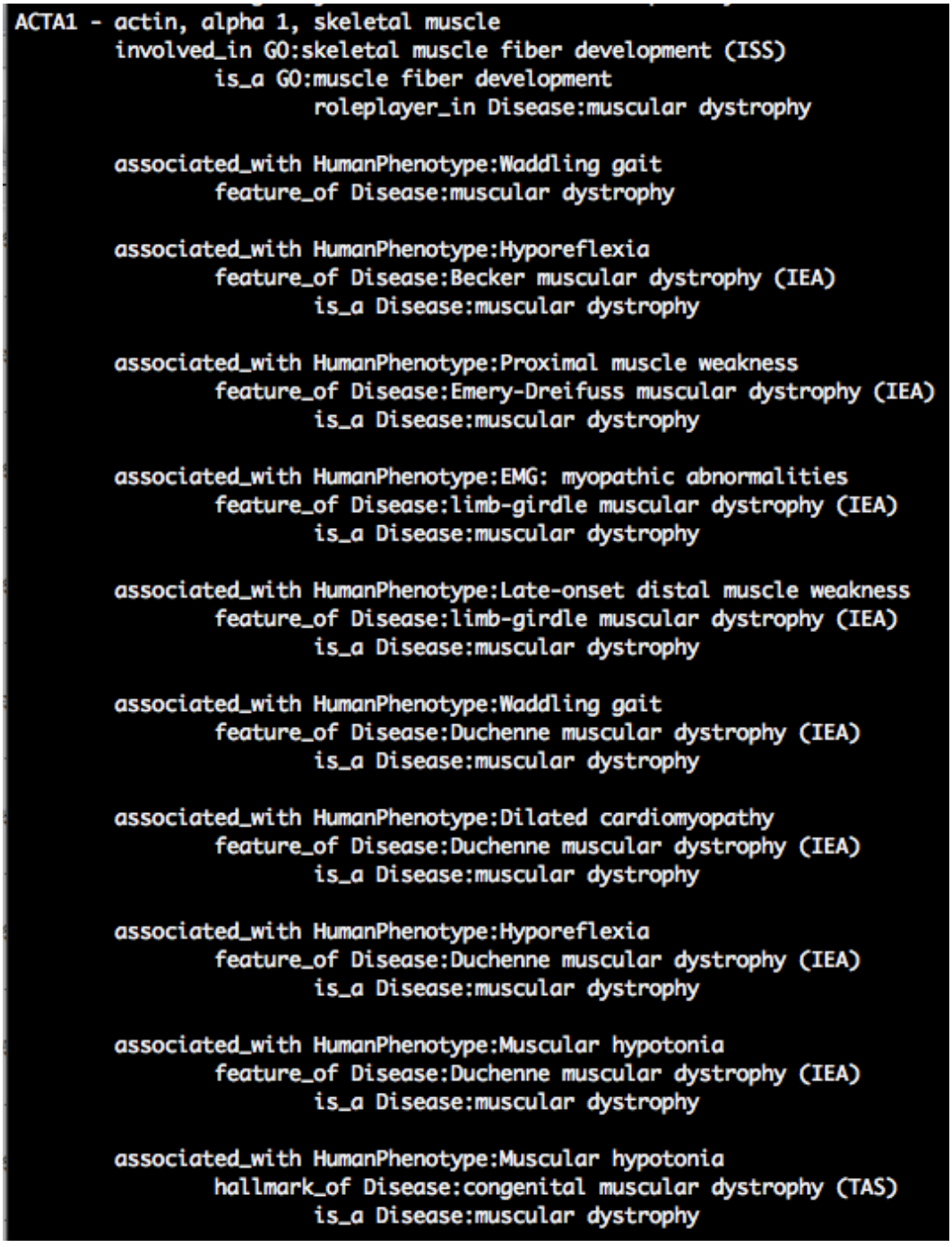



**Figure 4.** An illustration of how a new gene bearing a functional mutation segregating with affected individuals could be indirectly linked to muscular dystrophy based on existing knowledge of it's association to phenotypes relevant to the disease.

## References

Szklarczyk D, Franceschini A, Wyder S, Forslund K, Heller D, Huerta-Cepas J,
Simonovic M, Roth A, Santos A, Tsafou KP, Kuhn M, Bork P, Jensen LJ, von Mering

C. STRING v10: protein-protein interaction networks, integrated over the tree of
life. Nucleic Acids Res. 2015 Jan;43(Database issue):D447-52. doi:
10.1093/nar/gku1003. Epub 2014 Oct 28.

Davis AP, Grondin CJ, Lennon-Hopkins K, Saraceni-Richards C, Sciaky D, King
BL, Wiegers TC, Mattingly CJ. The Comparative Toxicogenomics Database's 10th year
anniversary: update 2015. Nucleic Acids Res. 2015 Jan;43(Database issue):D914-20.
doi: 10.1093/nar/gku935. Epub 2014 Oct 17.

Petri V, Jayaraman P, Tutaj M, Hayman GT, Smith JR, De Pons J, Laulederkind
SJ, Lowry TF, Nigam R, Wang SJ, Shimoyama M, Dwinell MR, Munzenmaier DH, Worthey
EA, Jacob HJ. The pathway ontology - updates and applications. J Biomed
Semantics. 2014 Feb 5;5(1):7. doi: 10.1186/2041-1480-5-7.

Weng MP, Liao BY. MamPhEA: a web tool for mammalian phenotype enrichment
analysis. Bioinformatics. 2010 Sep 1;26(17):2212-3. doi:
10.1093/bioinformatics/btq359. Epub 2010 Jul 6.