

2016

# MeSH-informed enrichment analysis and MeSH-guided semantic similarity among functional terms and gene products in chicken

Gota Morota

*University of Nebraska- Lincoln, morota@vt.edu*

Timothy M. Beissinger

*United States Department of Agriculture, Tim.Beissinger@ars.usda.gov*

Francisco Peñagaricano

*University of Florida, fpenagaricano@ufl.edu*

Follow this and additional works at: <http://digitalcommons.unl.edu/animalscifacpub>



Part of the [Genetics and Genomics Commons](#), and the [Meat Science Commons](#)

---

Morota, Gota; Beissinger, Timothy M.; and Peñagaricano, Francisco, "MeSH-informed enrichment analysis and MeSH-guided semantic similarity among functional terms and gene products in chicken" (2016). *Faculty Papers and Publications in Animal Science*. 902.

<http://digitalcommons.unl.edu/animalscifacpub/902>

This Article is brought to you for free and open access by the Animal Science Department at DigitalCommons@University of Nebraska - Lincoln. It has been accepted for inclusion in Faculty Papers and Publications in Animal Science by an authorized administrator of DigitalCommons@University of Nebraska - Lincoln.

# MeSH-informed enrichment analysis and MeSH-guided semantic similarity among functional terms and gene products in chicken

Gota Morota<sup>\*</sup>, Timothy M Beissinger<sup>† ‡</sup>, and Francisco Peñagaricano<sup>§ \*\*</sup>

<sup>\*</sup>Department of Animal Science, University of Nebraska-Lincoln, Lincoln,  
Nebraska

<sup>†</sup>United States Department of Agriculture, Agricultural Research Service,  
Columbia, Missouri

<sup>‡</sup>Division of Plant Sciences, University of Missouri, Columbia, Missouri

<sup>§</sup>Department of Animal Sciences, University of Florida, Gainesville, Florida

<sup>\*\*</sup>University of Florida Genetics Institute, University of Florida, Gainesville,  
Florida

13 Keywords: annotation, chicken, enrichment analysis, MeSH, semantic similarity

14

15 Running title: MeSH resources for chicken

16

17 Corresponding author:

18 Gota Morota

19 Department of Animal Science

20 University of Nebraska-Lincoln

21 PO Box 830908

22 Lincoln, NE 68583-0908, USA.

23 E-mail: morota@unl.edu

24

# Abstract

Biomedical vocabularies and ontologies aid in recapitulating biological knowledge. The annotation of gene products is mainly accelerated by Gene Ontology (GO) and more recently by Medical Subject Headings (MeSH). Here we report a suite of MeSH packages for chicken in Bioconductor and illustrate some features of different MeSH-based analyses, including MeSH-informed enrichment analysis and MeSH-guided semantic similarity among terms and gene products, using two lists of chicken genes available in public repositories. The two published datasets that were employed represent (i) differentially expressed genes and (ii) candidate genes under selective sweep or epistatic selection. The comparison of MeSH with GO overrepresentation analyses suggested not only that MeSH supports the findings obtained from GO analysis but also that MeSH is able to further enrich the representation of biological knowledge and often provide more interpretable results. Based on the hierarchical structures of MeSH and GO, we computed semantic similarities among vocabularies as well as semantic similarities among selected genes. These yielded the similarity levels between significant functional terms, and the annotation of each gene yielded the measures of gene similarity. Our findings show the benefits of using MeSH as an alternative choice of annotation in order to draw biological inferences from a list of genes of interest. We argue that the use of MeSH in conjunction with GO will be instrumental in facilitating the understanding of the genetic basis of complex traits.

# Introduction

Understanding the genetic basis of variation for complex traits remains a fundamental goal of biology. Different approaches, including whole-genome scans and genome-wide expression studies, have been used in order to identify individual genes underlying economically relevant traits in a wide spectrum of agricultural species. These studies usually generate lists of genes potentially involved in the phenotypes under study. The challenge is to translate these lists of candidate genes into a better understanding of the biological phenomena involved. It is increasingly accepted that overrepresentation or enrichment analysis (Drăghici et al., 2003) can provide further insights into the biological pathways and processes affecting complex traits.

Recently, the Medical Subject Headings (MeSH) vocabulary (Nelson et al., 2004) has been proposed for defining functional sets of genes in the context of enrichment analysis. MeSH is a controlled life and medical sciences vocabulary maintained by the National Library of Medicine to index documents in the MEDLINE database. Each bibliographic reference in the MEDLINE database is associated with a set of MeSH terms that describe the content of the publication. Importantly, MeSH contains a substantially more diverse and extensive range of categories than that of Gene Ontology (GO) (Ashburner et al., 2000), which is probably the most popular among the initiatives for defining functional classes of genes (Nakazato et al., 2008). Therein, GO terms are classified into three domains: biological processes, molecular functions, and cellular components. This ontology has been successfully used for dissecting relevant traits in livestock species (e.g, Peñagaricano et al., 2013; Gamba et al., 2013). Similarly, each MeSH term is clustered into 19 different categories; some MeSH categories, such as Diseases, are not included in GO, whereas other functional categories, such as Phenomena and Processes or Chemicals and Drugs, share similar concepts with those of GO. The recent availability of MeSH software packages has rendered agricultural species amenable to MeSH-based analysis (Tsuyuzaki et al., 2015). For instance, MeSH enrichment analysis has been successfully applied to mammals including dairy cattle, swine, and horse (Morota et al., 2015), and to maize (Beissinger and Morota, 2016). These studies showed the potential of MeSH for enhancing the biological interpretation of sets of genes in agricultural organisms.

69 The main objective of the current study was to report the availability of MeSH Bioconductor  
70 packages for chicken, and to illustrate the features of different MeSH-based analyses, including  
71 MeSH-informed enrichment analysis and MeSH-guided semantic similarity among terms and gene  
72 products. For this purpose, we used two lists of selected genes available in public repositories: (i)  
73 differentially expressed genes reported in a RNA-seq study (Zhuo et al., 2015) and (ii) candidate  
74 genes historically impacted by selection detected in a whole-genome scan using a broad spectrum  
75 of populations (Beissinger et al., 2015). The results of the MeSH-based enrichment analysis were  
76 contrasted with GO terms. The use of MeSH and GO terms in functional genomics studies can  
77 be further explored through computing the similarity between significant functional terms as well  
78 as the similarity between significant genes by leveraging the hierarchies of these two controlled  
79 vocabularies.

## 80 **Materials and Methods**

81 We used two datasets from previously published studies with the objective of demonstrate some  
82 capabilities of different MeSH-based analyses in chicken. The first dataset includes 263 genes that  
83 showed differential expression in abdominal fat tissue between high and low feed efficiency broiler  
84 chickens (Zhuo et al., 2015). The second dataset contains 352 genes identified by a whole-genome  
85 scan using Ohta’s between-population linkage disequilibrium measure,  $D_{IS}^2$ , in a panel that included  
86 72 different chicken breeds (Beissinger et al., 2015). In both datasets, the list of background genes  
87 was defined as all annotated genes in the chicken genome available in NCBI. Below we present the  
88 MeSH analyses coupled with several example code for illustration purposes.

89 The suite of MeSH (Tsuyuzaki et al., 2015) and the GOstats (Falcon and Gentleman, 2007)  
90 packages in Bioconductor were used for performing a hypergeometric test in the enrichment analysis.  
91 This test evaluates whether a given functional term or vocabulary is enriched or overrepresented  
92 with selected genes. In particular, the  $P$ -value of observing  $g$  significant genes in a functional term

93 (i.e. MeSH or GO term) was calculated by

$$Pvalue = 1 - \sum_{i=0}^{g-1} \frac{\binom{S}{i} \binom{N-S}{k-i}}{\binom{N}{k}}$$

94 where  $S$  is the total number of selected genes,  $N$  is the total number of analyzed genes, and  $k$  is  
95 the total number of genes in the functional term under study. The meshr package has a feature to  
96 perform a multiple testing correction by choosing from Benjamini-Hochberg, Q-value or empirical  
97 Bayes method. We used a lenient  $P$ -value 0.05 for the illustrative data in order to directly compare  
98 the results from MeSH enrichment analysis with the ones from the GOSTats package, which does  
99 not offer a multiple testing correction option. Although a multiple testing correction reduces false  
100 positives, if we view MeSH analysis as a tool to generate hypotheses or to obtain a big picture of  
101 selected genes for subsequent downstream analysis, we may want to know the top 10% of MeSH  
102 terms regardless of  $P$ -values.

103 The first step of MeSH analysis is to load the namespace of the packages.

```
104 library (MeSH.db)  
105 library (MeSH.Gga.eg.db)  
106 library (meshr)  
107  
108
```

109 The MeSH.db package contains the relationship between MeSH IDs and MeSH terms. The MeSH.Gga.eg.db  
110 is an annotation package that provides the correspondence between MeSH IDs and Entrez Gene  
111 IDs. This package was created based on gene2pubmed (<ftp://ftp.ncbi.nih.gov/gene/DATA/>) that  
112 maps Entrez Gene IDs and PubMed IDs. By using data licenced by PubMed  
113 (<http://www.nlm.nih.gov/databases/license/license.html>), we then associated PubMed IDs to MeSH  
114 terms. This was followed by merging MeSH terms with MeSH IDs via NLM MeSH (Tsuyuzaki et al.,  
115 2015). The meshr package performs a hypergeometric test and returns significantly enriched MeSH  
116 terms. Once the three packages are loaded, we proceed to create the object of a parameter class  
117 MeSHHyperGParams-class. This object contains all parameters required to run the hypergeometric  
118 test.

119

120

```

121 meshParams <- new("MeSHHyperGParams", geneIds = selectedGenes,
122                  universeGeneIds = universeGenes,
123                  annotation = "MeSH.Gga.eg.db", category = "D",
124                  database = "gene2pubmed",
125                  pvalueCutoff = 0.05, pAdjust = "none"
126                  )
127
128

```

Here geneIds and universeGeneIds are the vectors of Entrez Gene IDs for selected and background genes, respectively, category is one of the abbreviation codes for MeSH categories such as D (Chemicals and Drugs), C (Diseases), A (Anatomy), and G (Phenomena and Processes), pvalueCutoff is the numeric value for *P*-value cutoff, and pAdjust allows users to choose multiple testing methods from BH (Benjamini-Hochberg), QV (Q-value), lFDR (empirical Bayes), and none (unadjusted). Finally, the meshHyperGTest function accepts the MeSHHyperGParams-class object and perform a MeSH enrichment analysis.

```

136 meshR <- meshHyperGTest(meshParams)
137
138

```

The returned object is MeSHHyperGResult-class and we can access the results with the summary function.

```

141 summary(meshR)
142
143

```

The summary function returns a data.frame object with information about MeSH ID, *P*-value, MeSH term, Entrez Gene ID, and PubMed ID.

In addition, the hierarchical structures of MeSH and GO permitted us to compute semantic similarities between functional terms (Lord et al., 2003; Pesquita et al., 2009). This is a metric between two terms on the basis of their biological meanings of annotation: the closer two terms are in the hierarchy, the higher the similarity measure is between these terms. Figure 1 shows a MeSH hierarchy for illustrative purpose. In this example, the semantic similarity measure between Mesh Term 2 and Mesh Term 3 is greater than that of Mesh Term 1 and Mesh Term 2 because they are closer in the hierarchy. We employed the information content-based Jiang and Conrath's measure (Jiang and Conrath, 1998) to compute the pairwise similarities within GO ontologies and MeSH headings. The semantic similarity measure between two terms  $t_1$  and  $t_2$  is given by the information



155 content  $IC(t) = -\log p(t)$ , where  $p(t)$  is the probability of occurrence of the term  $t$  and its children  
156 terms in MeSH or GO hierarchy. The semantic distance metric is a function of

$$Dist = IC(t_1) + IC(t_2) - 2IC(MICA),$$

157 where MICA is the most informative common ancestor.

158 We further computed semantic similarity between selected genes by aggregating their MeSH  
159 or GO terms assigned. This is a similarity measure at the level of genes which is analogous to a  
160 similarity matrix among SNPs (Morota and Gianola, 2013). We calculated similarity scores over  
161 all pairs of terms between the two vocabulary sets of genes under consideration. All these GO and  
162 MeSH-guided semantic similarity analyses were carried out using the GOSemSim (Yu et al., 2010)  
163 and the MeSHSim (Zhou et al., 2015) Bioconductor packages, respectively. We selected exactly  
164 the same genes as were identified in GO categories when computing MeSH-based gene similarity to  
165 allow direct comparisons between these two functional vocabularies. Source code and reproducible  
166 output reports generated by R Markdown are available as Supporting Files.

## 167 **Data Availability**

168 The MeSH.db, MeSH.Gga.eg.db, and meshr packages are available for download at Bioconductor  
169 <https://www.bioconductor.org/>. The two datasets used in the current study have already been  
170 published. The gene expression data can be downloaded from  
171 <http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0135810#sec025>. Raw data for  
172 the selective sweep data are available from <http://dx.doi.org/10.6084/m9.figshare.1497961>, and  
173 selected genes can be found in Beissinger et al. (2015).

# Results

## Summary of MeSH and GO annotations

The organism and the biomaRt Bioconductor packages were queried to annotate genes by MeSH and GO terms. Table 1 shows the total number of genes (background and selected genes) annotated by MeSH and GO in each of the datasets under study. Both MeSH and GO terms had a similar number of annotated known genes (10,227 vs. 12,460), whereas the number of selected genes with MeSH terms assigned was about one-half of that of GO. For example, in the gene expression (selective sweep) data, 245 (333) genes are annotated by GO while only 110 (145) genes are annotated by MeSH. It is important to note that this difference could be because the majority of chicken genes are annotated by Inferred from Electronic Annotation (evidence code: IEA) in GO, whereas all MeSH terms are assigned by manual curation at NCBI. On the other hand, the advantage of using GO-IEA over MeSH is that MeSH does not include genes with no published literature in PubMed, while GO-IEA can still predict function for these genes. We expect that over time, MeSH will improve as new knowledge is created and published in the scientific literature.

## Enrichment analysis

Gene Expression Data: A subset of significant MeSH terms ( $P$ -value  $\leq 0.05$ ) enriched with differentially expressed genes detected in fat tissue between high and low feed efficiency chickens are highlighted in Table 2. The majority of the MeSH terms in the Chemicals and Drugs category are related to lipid deposition and lipid metabolism. For instance, *Lipoproteins* (MeSH:D008074), and *Apolipoproteins* (MeSH:D001053) are closely related to lipid transportation. Additionally, *Fatty Acid-Binding Proteins* (MeSH:D050556) regulates diverse lipid signals, while *PPAR alpha* (MeSH:D047493) controls lipid and lipoprotein metabolism. Interestingly, many GO terms related to lipid deposition and metabolism, such as *cholesterol metabolic process* (GO:0008203), *high-density lipoprotein particle assembly* (GO:0034380), *spherical high-density lipoprotein particle* (GO:0034366), and *high-density lipoprotein particle binding* (GO:0008035), were also significantly

enriched with differentially expressed genes (File S1). Similarly, MeSH terms related to Wnt proteins  
 and signalling pathways, such as *Wnt Proteins* (MeSH:D051153), *Wnt4 Protein* (MeSH: D060528),  
*Wnt1 Protein* (MeSH:D051155), and their counterparts in GO, such as *regulation of Wnt signal-*  
*ing pathway* (GO:0030111) and *Wnt signaling pathway* (GO:0016055), were found as significant.  
 The Wnt proteins are known to interact with lipids. We also found *Steroid 17-alpha-Hydroxylase*  
 (MeSH:D013254) and *steroid 17-alpha-monooxygenase activity* (GO:0004508) as significant terms;  
 these two categories are enriched in genes involved in the synthesis of lipids. Moreover, we detected  
 some MeSH terms related to the immune system regulation (e.g., *Interleukin-6* (MeSH:D015850)  
 and *Chemokines* (MeSH:D018925)). Lastly, *Glycoproteins* (MeSH:D006023), is produced from the  
 gene *AHSG* and plays a role in glucose metabolism and the regulation of insulin signaling. Taken  
 together, our findings confirm that MeSH enrichment analysis can either reinforce findings from  
 GO or even bring an additional biological insight. Figure 2 depicts the semantic similarity between  
 significant MeSH terms in the Chemicals and Drugs category. In general, this subset of MeSH terms  
 showed low to high levels of semantic similarity.

For the Diseases category, which is unique to MeSH-based analysis, a subset of significant  
 MeSH terms that deserves particular attention in the area of feed efficiency and lipid metabolism  
 in poultry is highlighted in Table 2. For instance, *Hyperplasia* (MeSH:D006965) is a potential  
 contributor to abdominal fat mass in broiler chickens; its relationship with *Diabetes Mellitus, Type 2*  
 (MeSH:D003924) is well-documented in humans. Some MeSH terms directly related to the immune  
 function, such as *Newcastle Disease* (MeSH:D009521) and *Inflammation* (MeSH:D007249), also  
 showed a significant enrichment with differentially expressed genes. Interestingly, *Hyperplasia* and  
*Inflammation* showed a moderate semantic similarity according to the MeSH hierarchy (File S1).

Selective Sweep Data: Table 2 shows the results of the MeSH-informed enrichment analysis  
 using genes putatively swept or under epistatic selection derived from a chicken diversity panel.  
 Most of these terms are related to insulin metabolism. For instance, resistance to insulin occurs  
 in birds due to high plasma glucose and fatty acid levels; this is supported by *Insulin Resistance*  
 (MeSH:D007333) in both the Diseases and Phenomena and Processes categories, as well as *Recep-*  
*tor, Insulin* (MeSH:D011972) and *Insulin* (MeSH:D007328) in the Chemicals and Drugs category.  
 Moreover, we identified MeSH terms involved in the circadian clock of chicken. These are *Period*

228 *Circadian Proteins* (MeSH:D056950), *CLOCK Proteins* (MeSH:D056926) and *ARNTL Transcrip-*  
 229 *tion Factors* (MeSH:D056930) in Chemicals and Drugs, as well as *E-Box Elements* (MeSH:D024721),  
 230 *Biological Clocks* (MeSH:D001683), and *Light* (MeSH:D008027) in Phenomena and Processes. Fig-  
 231 ure 3 shows the semantic similarities among MeSH terms in the Chemicals and Drugs category.  
 232 Biological clock-related annotations, such as *Period Circadian Proteins* and *CLOCK Proteins*, ex-  
 233 hibited moderate to high similarity. The results obtained from the other MeSH and GO categories  
 234 were shown in File S2.

## 235 Gene semantic similarity

236 Gene Expression Data: Comparison of gene semantic similarity between MeSH and GO Biological  
 237 Process for a subset of significant genes ( $n=49$ ) from the RNA-seq dataset is depicted in Figure 4.  
 238 MeSH-based gene semantic similarity analysis showed that genes related to energy reserve metabolic  
 239 process are highly related. For instance, genes that are involved in triacylglycerol and cholesterol  
 240 biosynthesis, such as methylsterol monooxygenase 1 (*MSMO1*), insulin induced gene 1 (*INSIG1*), 1-  
 241 acylglycerol-3-phosphate O-acyltransferase 9 (*AGPAT9*), and ADP ribosylation factor like GTPase  
 242 2 binding protein (*ARL2BP*), were highly similar to each other based on the MeSH hierarchy.  
 243 Interestingly, GO-based analysis produced slightly different results; for instance, the gene *MSMO1*  
 244 was highly similar to *INSIG1* but moderately similar to *AGPAT9* and *ARL2BP*. Additionally,  
 245 genes *MSMO1* and *INSIG1* were moderately or highly related to lecithin-cholesterol acyltransferase  
 246 (*LCAT*) and cytochrome b5 type A (microsomal) (*CYB5A*) based on the GO structure. These two  
 247 genes, involved in lipid metabolism, also showed high similarity to apolipoprotein A-I (*APOA1*)  
 248 and cytochrome P450, family 17, subfamily A, polypeptide 1 (*CYP17A1*). The relationship among  
 249 these genes were low to moderate based on the MeSH hierarchy. The results based on the GO  
 250 Molecular Function and Cellular Component categories were presented in File S3.

251 Selective Sweep Data: Gene semantic similarity based on both MeSH and GO Biological Pro-  
 252 cess among a subset of genes ( $n=45$ ) under selection is shown in Figure 5. Notably, a large group  
 253 of genes, including strawberry notch homolog 1 (*Drosophila*) (*SBNO1*), ARP5 actin-related pro-  
 254 tein 5 (*ACTR5*), SET domain containing 1B (*SETD1B*), Obg-like ATPase 1 (*OLA1*), and histone

deacetylase 9 (*HDAC9*) were highly related based on both MeSH and GO-guided semantic similarity analyses. All these genes are involved in chromatin organization and regulation of gene expression. Moreover, particular attention was paid to the top five candidates under epistatic selection reported by Beissinger et al. (2015). These genes are adenylate cyclase 5 (*ADCY5*), myosin light chain kinase (*MYLK*), phosphatidylinositol-4,5-bisphosphate 3-kinase, catalytic subunit beta (*PIK3CB*), calcium binding protein 39 (*CAG39*), and interleukin 1 receptor accessory protein (*IL1RAP*). Although none of these pair of genes appeared in a GO-based similarity matrix, *ADCY5* and *MYLK* presented a low to moderate gene semantic similarity based on the MeSH hierarchy (File S4).

## Discussion

This article reports the MeSH analysis for chicken using the newly developed Bioconductor packages. These new resources enabled us to carry out different MeSH-based analyses, including enrichment analysis and MeSH-guided semantic similarity among functional terms and gene products. We exemplified the potential usefulness of these MeSH-based approaches by using two different publicly available chicken data.

The adipose tissue is the major site for lipid deposition and lipid metabolism, and it plays a central role in energy homeostasis. Unsurprisingly, several MeSH terms closely related to fat metabolism, such as Lipoproteins, Apolipoproteins, Fatty Acid-Binding Proteins, and PPAR alpha, were significantly enriched with genes that showed differential expression in fat tissue between high and low feed efficiency broiler chickens. We found some genes were annotated by the same MeSH terms. For instance, gene overlap between Lipoproteins and Apolipoproteins was one-half and 66% of genes were shared between Fatty Acid-Binding Proteins and PPAR alpha. It is likely that this gene overlap is observed because each MeSH term inherits all annotations from its more specific child terms (Falcon and Gentleman, 2007). It is possible to address this issue by conducting a conditional analysis that is implemented in the GOstats package. Adding this feature in the meshr package might alleviate the overlap of genes. Also, adipose tissue is now recognized as a metabolically active tissue that has important endocrine and immune regulatory functions (Kershaw and Flier, 2004). Interestingly, we found many significant MeSH terms, such as Interleukin-6, Chemokines,

282 and Immunoglobulins, that are closely associated with the regulation of the immune function.  
283 Overall, our MeSH-based findings provide further insights into the biological mechanisms underlying  
284 differences in adiposity between high and low feed efficiency broiler chickens.

285 Included in our exemplary applications of MeSH annotations is a set of 352 genes previously iden-  
286 tified as putatively affected by selection. Genes identified through population-genetic approaches  
287 such as this can be elusive, because their identification does not rely on phenotypes. Therefore  
288 associating selection with any specific trait is often very difficult (Akey, 2009). As we demonstrate  
289 in this study, tools such as GO and now MeSH are useful for suggesting biological interpretations  
290 that can later be followed up on or drive future biological hypotheses. For instance, our results  
291 showed that insulin-related MeSH terms appeared unusually often in the set of genes impacted by  
292 selection. This implies that selection for insulin-related traits may have played an important role  
293 in differentiating chicken breeds. Furthermore, our analysis involved testing for semantic similarity  
294 between pairs of genes, which was particularly useful for evaluating the most promising gene-pairs  
295 highlighted by Beissinger et al. (2015) as candidates for epistatic selection. Our expectation was  
296 that these pairs of genes are likely to be related to each other, as they have been predicted to be  
297 involved in the same selected phenotype. Our finding that one pair showed at least a weak semantic  
298 similarity may be interpreted as evidence that these two genes, *ADCY5* and *MYLK* are the most  
299 likely among the set to truly be epistatic.

300 The recent advancement in cataloguing genes with MeSH and GO has made it possible to assess  
301 the role of selected genes and has opened new opportunities for genetic research. Enrichment  
302 analysis recapitulates a set of genes into higher-level biological features. We argue that obtaining  
303 a complete picture of genes of interest using MeSH and GO is an important initial step toward  
304 functional genomics studies in poultry as well as other agricultural species as it facilitates efforts to  
305 illuminate the genetic basis of phenotypic variation.

## 306 Acknowledgements

307 This work was supported in part by the University of Nebraska Layman Fund to G.M. Support  
308 for T.M.B. was provided by USDA-ARS. F.P. acknowledges financial support from the Florida

309 Agricultural Experiment Station and the Department of Animal Sciences, University of Florida.

# References

- Akey, J. M. (2009). Constructing genomic maps of positive selection in humans: Where do we go from here? *Genome Res.*, 19:711–722.
- Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., and Sherlock, G. (2000). Gene Ontology: tool for the unification of biology. *Nat Genet.*, 25:25–29.
- Beissinger, T. and Morota, G. (2016). Medical subject heading (MeSH) annotations illuminate maize genetics and evolution. *bioRxiv*.
- Beissinger, T. M., Gholami, M., Erbe, M., Weigend, S., Weigend, A., de Leon, N., Gianola, D., and Simianer, H. (2015). Using the variability of linkage disequilibrium between subpopulations to infer sweeps and epistatic selection in a diverse panel of chickens. *Heredity*, Advance online.
- Drăghici, S., Khatiria, P., Martinsb, R. P., Ostermeier, G. C., and Krawetz, S. A. (2003). Global functional profiling of gene expression. *Genomics*, 81:98–104.
- Falcon, S. and Gentleman, R. (2007). Using GOstats to test gene lists for GO term association. *Bioinformatics*, 23:257–258.
- Gambra, R., Peñagaricano, F., Kropp, J., Khateeb, K., Weigel, K. A., Lucey, J., and Khatib, H. (2013). Genomic architecture of bovine  $\kappa$ -casein and  $\beta$ -lactoglobulin. *J Dairy Sci.*, 96:5333–5343.
- Huang, D. W., Sherman, B. T., and Lempicki, R. A. (2009). Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.*, 37:1–13.
- Jiang, J. and Conrath, D. (1998). Semantic similarity based on corpus statistics and lexical taxonomy. In *Proceedings of the 10th International Conference on Research in Computational Linguistics*, Taiwan.



333 Kershaw, E. E. and Flier, J. S. (2004). Adipose tissue as an endocrine organ. *J Clin Endocrinol*  
334 *Metab.*, 89:2548–2556.

335 Lord, P. W., Stevens, R. D., Brass, A., and Goble, C. A. (2003). Investigating semantic simi-  
336 larity measures across the Gene Ontology: the relationship between sequence and annotation.  
337 *Bioinformatics*, 19:1275–1283.

338 Morota, G. and Gianola, D. (2013). Evaluation of linkage disequilibrium in wheat with an L1-  
339 regularized sparse Markov network. *Theor Appl Genet.*, 126:1991–2002.

340 Morota, G., Peñagaricano, F., Petersen, J. L., Ciobanu, D. C., Tsuyuzaki, K., and Nikaido, I.  
341 (2015). An application of MeSH enrichment analysis in livestock. *Anim Genet.*, 46:381–387.

342 Nakazato, T., Takinaka, T., Mizuguchi, H., Matsuda, H., Bono, H., and Asogawa, M. (2008).  
343 Biocompass: a novel functional inference tool that utilizes MeSH hierarchy to analyze groups of  
344 genes. *In Silico Biol.*, 8:53–61.

345 Nelson, S. J., Schopen, M., Savage, A. G., Schulman, J. L., and Arluk, N. (2004). The MeSH  
346 translation maintenance system: structure, interface design, and implementation. *Stud. Health*  
347 *Technol. Inform.*, 107:67–69.

348 Peñagaricano, F., Weigel, K. A., Rosa, G. J. M., and Khatib, H. (2013). Inferring quantitative  
349 trait pathways associated with bull fertility from a genome-wide association study. *Front Genet.*,  
350 3:307.

351 Pesquita, C., Faria, D., Falcão, A. O., Lord, P., and Couto, F. M. (2009). Semantic similarity in  
352 biomedical ontologies. *PLoS Comput. Biol.*, 5:e1000443.

353 Tsuyuzaki, K., Morota, G., Ishii, M., Nakazato, T., Miyazaki, S., and Nikaido, I. (2015). MeSH  
354 ORA framework: R/Bioconductor packages to support MeSH over-representation analysis. *BMC*  
355 *Bioinformatics*, 16:45.

356 Yu, G., Li, F., Qin, Y., Bo, X., Wu, Y., and Wang, S. (2010). GOSemSim: an R package for  
357 measuring semantic similarity among GO terms and gene products. *Bioinformatics*, 26:976–978.

- 358 Zhou, J., Shui, Y., Peng, S., Li, X., Mamitsuka, H., and Zhu, S. (2015). MeSHSim: An R/Biocon-  
359 ductor package for measuring semantic similarity over MeSH headings and MEDLINE documents.  
360 *J. Bioinform. Comput. Biol*, Online Ready.
- 361 Zhuo, Z., Lamont, S. J., Lee, W. R., and Abasht, B. (2015). RNA-seq analysis of abdominal fat  
362 reveals differences between modern commercial broiler chickens with high and low feed efficiencies.  
363 *PLoS ONE*, 10:e0135810.

364

## Supporting Information

365

- File S1: MeSH over-representation analysis (RNA-seq data)

366

- File S2: MeSH over-representation analysis (Selective sweep data)

367

- File S3: Gene Semantic Similarity (RNA-seq data)

368

- File S4: Gene Semantic Similarity (Selective sweep data)

# Tables

Table 1: Number of known and selected genes annotated by MeSH (Medical SubjectHeadings) and GO (Gene Ontology).

Data	Annotated Genes		Selected Genes		
	MeSH	GO	Total	MeSH	GO
RNA-seq	10227	12460	263	110	245
Selective Sweep			352	145	333

Table 2: A subset of statistically significant MeSH (Medical Subject Headings) terms. Background and Selected denote the number of background genes and selected genes annotated by the MeSH term, respectively. CD, D, and PP denote Chemicals and Drugs, Diseases, and Phenomena and Processes, respectively.

Data	Category	MeSH ID	Background	Selected	MeSH Term	<i>P</i> -value
RNA-seq	CD	D008074	14	4	<i>Lipoproteins</i>	0.0001
		D001054	7	2	<i>Apolipoproteins A</i>	0.0069
		D001053	5	2	<i>Apolipoproteins</i>	0.0034
		D050556	17	3	<i>Fatty Acid-Binding Proteins</i>	0.0037
		D047493	7	2	<i>PPAR alpha</i>	0.007
		D012177	6	2	<i>Retinol-Binding Proteins</i>	0.005
		D051153	91	8	<i>Wnt Proteins</i>	0.0003
		D060528	8	3	<i>Wnt4 Proteins</i>	0.0003
		D051155	19	2	<i>Wnt1 Proteins</i>	0.0488
		D015850	25	4	<i>Interleukin-6</i>	0.0078
		D018925	14	2	<i>Chemokines</i>	0.0276
		D007136	76	5	<i>Immunoglobulins</i>	0.0127
		D013254	1	1	<i>Steroid 17-alpha-Hydroxylase</i>	0.0188
		D006023	120	15	<i>Glycoproteins</i>	< 0.0001
	D	D006965	1	1	<i>Hyperplasia</i>	0.0188
		D003924	2	1	<i>Diabetes Mellitus, Type 2</i>	0.0373
		D009521	9	3	<i>Newcastle Disease</i>	0.0005
		D014802	5	2	<i>Vitamin A Deficiency</i>	0.0034
		D007249	12	2	<i>Inflammation</i>	0.0205
Sweeps	CD	D011972	2	8	<i>Receptor, Insulin</i>	0.0160
		D007328	26	3	<i>Insulin</i>	0.0268
		D056950	5	2	<i>Period Circadian Proteins</i>	0.0037
		D056926	8	2	<i>CLOCK Proteins</i>	0.0160
		D056930	6	2	<i>ARNTL Transcription Factors</i>	0.0122
	D	D007333	1	1	<i>Insulin Resistance</i>	0.0252
	PP	D007333	1	1	<i>Insulin Resistance</i>	0.0252
		D024721	8	2	<i>E-Box Elements</i>	0.0160
		D001683	13	2	<i>Biological Clocks</i>	0.0410
		D008027	28	3	<i>Light</i>	0.0325

# Figures

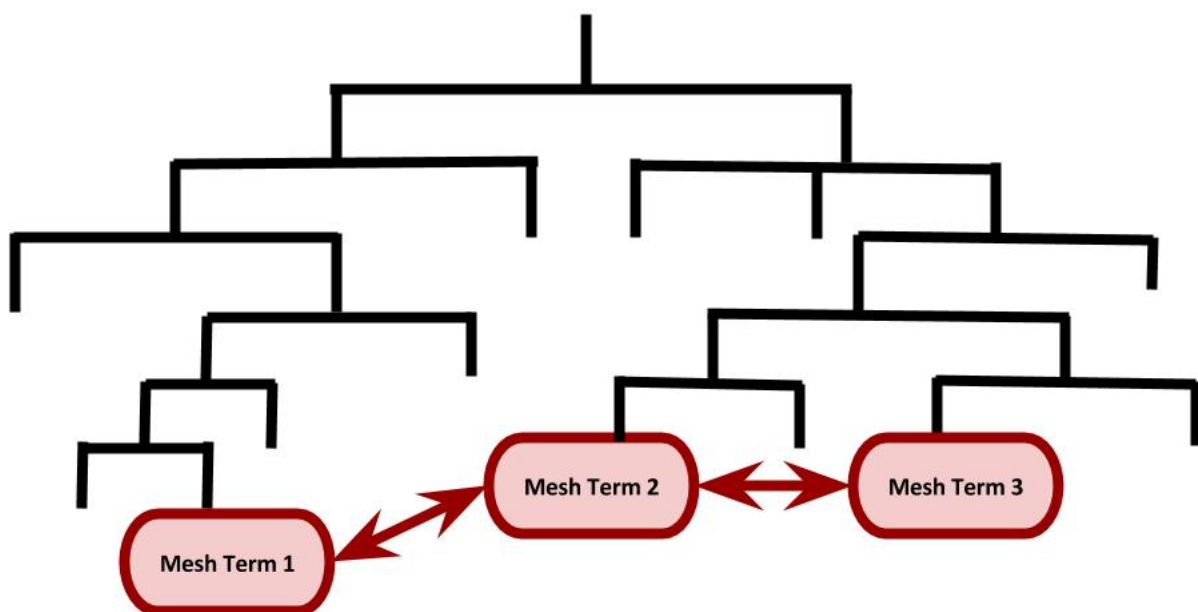


Figure 1: A cartoon illustrating semantic similarity among MeSH terms in the MeSH hierarchy. The semantic similarity measure between Mesh Term 2 and Mesh Term 3 is greater than that of Mesh Term 1 and Mesh Term 2 because they are closer in the hierarchy.

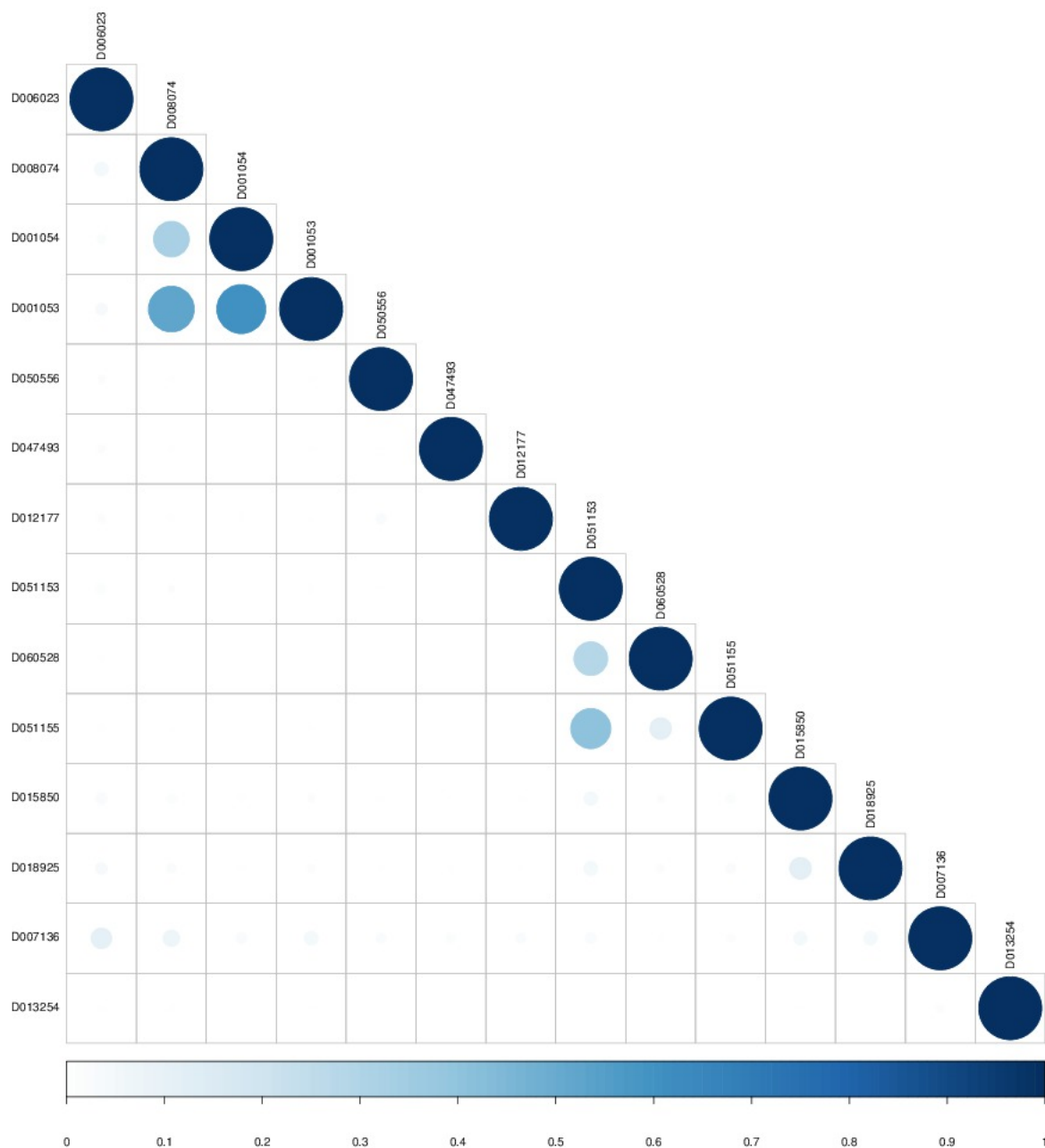


Figure 2: MeSH semantic similarity in the Chemicals and Drugs for the RNA-seq dataset. The higher the semantic similarity between MeSH terms, the bigger (darker) the circle. D006023 (Glycoproteins), D008074 (Lipoproteins), D001054 (Apolipoproteins A), D001053 (Apolipoproteins), D050556 (Fatty Acid-Binding Proteins), D047493 (PPAR alpha), D012177 (Retinol-Binding Proteins), D051153 (Wnt Proteins), D060528 (Wnt4 Proteins), D051155 (Wnt1 Proteins), D015850 (Interleukin-6), D018925 (Chemokines), D007136 (Immunoglobulins), and D013254 (Steroid 17-alpha-Hydroxylase).

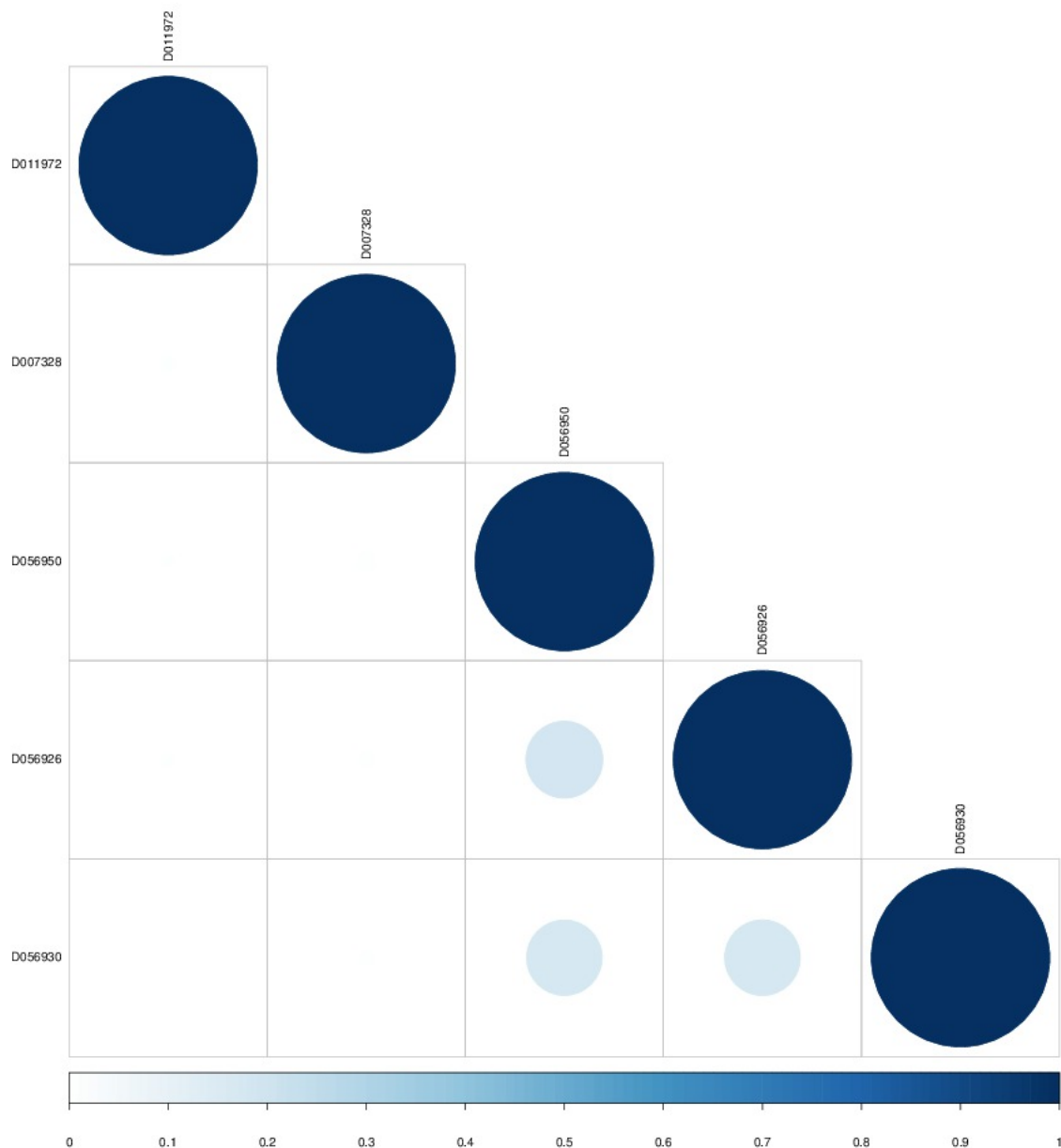


Figure 3: MeSH semantic similarity in the Chemicals and Drugs for the selective sweep dataset. The higher the semantic similarity between MeSH terms, the bigger (darker) the circle. D011972 (Receptor, Insulin), D007328 (Insulin), D056950 (Period Circadian Proteins), D056926 (CLOCK Proteins), and D056930 (ARNTL Transcription Factors).



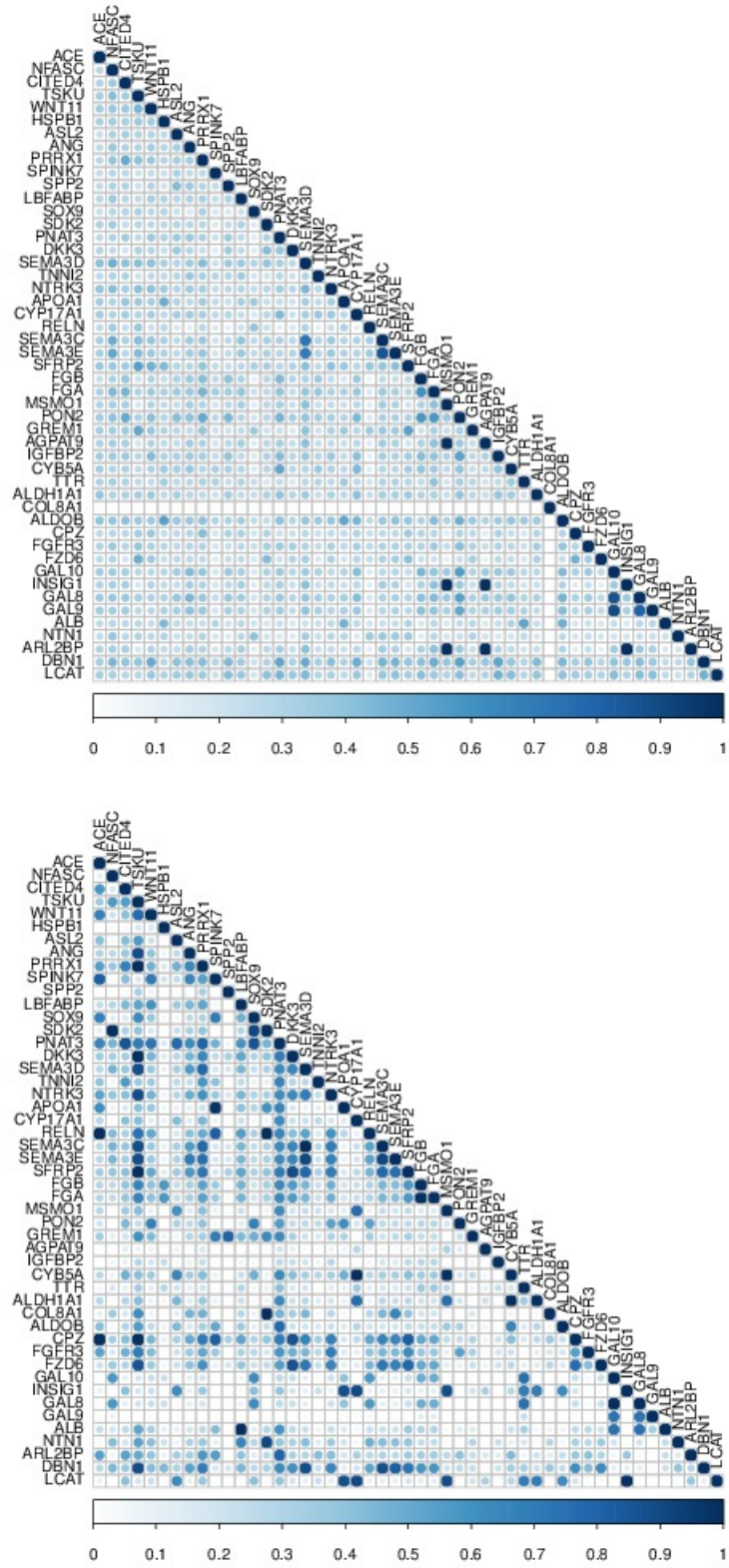


Figure 4: Gene semantic similarity for the RNA-seq dataset. The higher the semantic similarity between gene pairs, the bigger (darker) the circle. Top:MeSH, Bottom:GO.

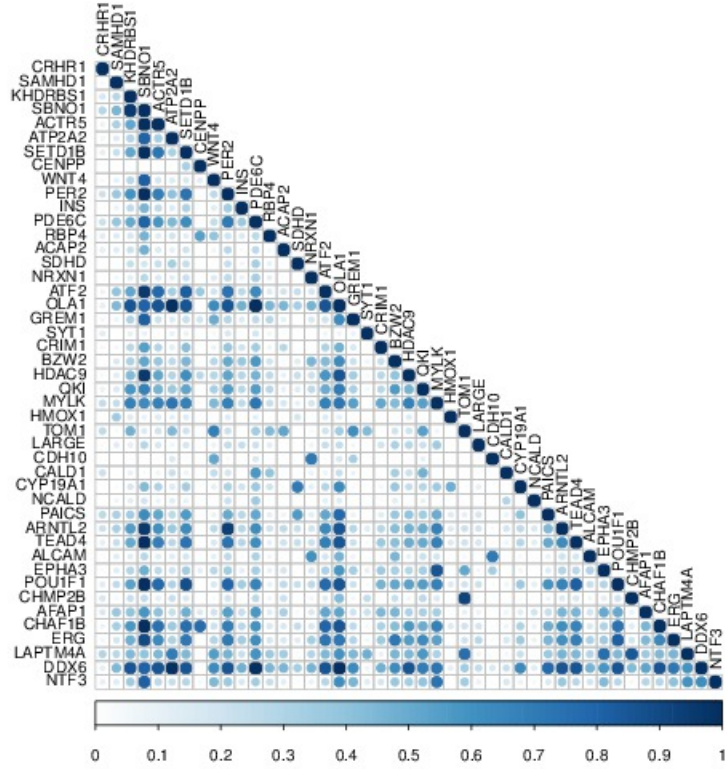
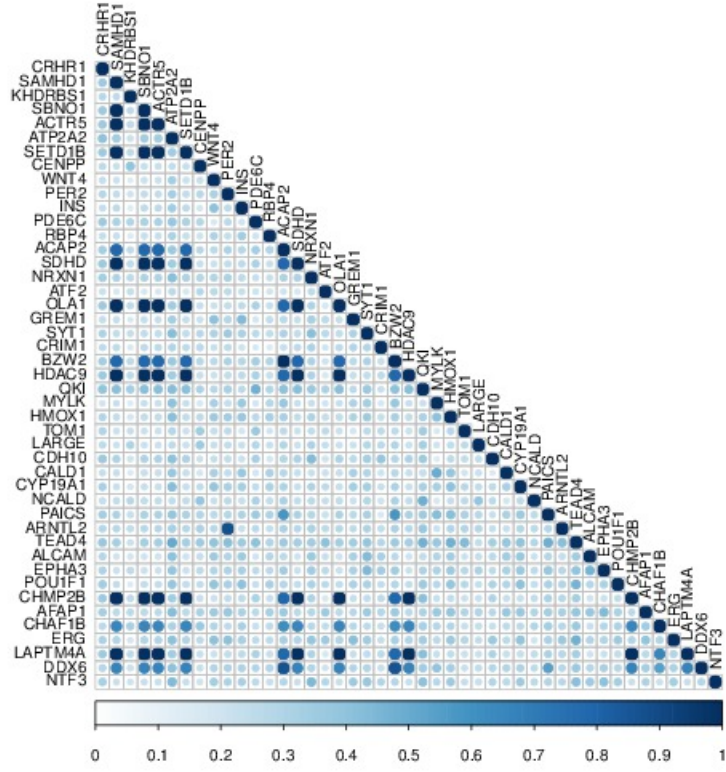


Figure 5: Gene semantic similarity for the selective sweep dataset. The higher the semantic similarity between gene pairs, the bigger (darker) the circle. Top:MeSH, Bottom:GO.