# TITLE PAGE

## TITLE

Iron-related transcriptomic variations in Caco-2 cells: *in silico* perspectives.

## AUTHORS

Marc Aubry[1,*], Annabelle Monnier[2], Celine Chicault[3], Marie-Dominique Galibert[3], Anita Burgun[4], Jean Mosser[1,3]

## AFFILIATIONS

[1] Plateau Transcriptome OUEST-genopole® Rennes, Université de Rennes 1, 2 avenue du professeur Léon Bernard 35043 Rennes Cedex, France

[2] UMR 6553 Ecobio, Université de Rennes 1, Équipe Évolution des Génomes et Spéciation, 35042 Rennes Cedex, France

[3] CNRS UMR 6061 Génétique et Développement, Université de Rennes 1, Équipe Régulation Transcriptionelle et Oncogenèse, IFR140 GFAS, Faculté de Médecine, 2 avenue du professeur Léon Bernard, CS 34317, 35043 Rennes Cedex, France

[4] EA 3888 Modélisation Conceptuelle des Connaissances Biomédicales, Faculté de Médecine, Université de Rennes 1, 35043 Rennes Cedex, France

## CONTACT INFORMATION

Marc Aubry

Plateau Transcriptome OUEST-genopole® Rennes, Université de Rennes 1, 2 avenue du professeur Léon Bernard 35043 Rennes Cedex, France

marc.aubry @univ-rennes1.fr

Tel: +33 (0)2 23 23 45 76
Fax: +33 (0)2 23 23 44 78

# ABSTRACT PAGE

**ABSTRACT**

The iron absorption by duodenal enterocytes is a key step of its homeostasis. But the control of this absorption is complex and cannot be fully explicated with present knowledge. In a global transcriptome approach, we identified 60 genes over-expressed in hemin (iron) overload in Caco-2 cells, an *in vitro* model of duodenal enterocytes. The challenge from there was to identify the affected molecular mechanisms and achieve a biological interpretation for that cluster. In that purpose, we built up a functional annotation method combining evidence and literature. Our method identified four pathways in the Process hierarchy of the Gene Ontology (GO): lipid metabolism, amino acid and cofactor metabolism, response to stimulus and transport. The accuracy of this functional profile is supported by the identification of known pathways associated with the iron overload (response to oxidative stress, glutathione metabolism). But our method also suggests new hypotheses on the regulation of iron uptake in Caco-2 cells. It is hypothesized that plasma membrane remodeling and vesicular recycling could be a potential modulator of iron transport proteins activities. These assumptions yet require a biological validation and they will therefore direct further research. Our functional annotation method is a valuable tool designed to help the biologist understand the biological links between the genes of a cluster, elaborate working hypotheses and direct future work. This work is also a validation 'by hand' of a biomedical text-mining system.

**KEYWORDS**

Iron; Hemin; Caco-2 cells; Microarray Analysis of Gene Expression; Functional Annotation; Text-Mining

# MAIN TEXT

## INTRODUCTION

Iron (Fe) is essential for many key-life processes. In humans, anemia due to iron deficiency affects millions of people worldwide. Paradoxically, iron can also be toxic depending on its excessive accumulation. Indeed, pathogenic mutations in iron transporters and regulators participate in hereditary primary or secondary iron overload diseases, including hemochromatosis [1]. Iron overload promotes oxidative stress and can severely damage or abolish normal bactericidal mechanisms in mammals tissue fluids leading to overwhelming growth of bacteria or fungi. Because of its potential biological toxicity, and as part of defense against infection, iron distribution is meticulously regulated. In mammals, there is no active excretion mechanism for iron. Indeed, absorption is the most important regulation process of iron homeostasis [2]. It takes place in the small intestine (duodenum, upper jejunum) in enterocytes and its control is a complex process that cannot be fully explained with present knowledge [3]. Understanding the molecular regulations of iron absorption is therefore a key-step for clarifying iron homeostasis mechanisms.

Using a global (18.000 genes) transcriptome approach with Caco-2 cells as an *in vitro* model of intestinal absorptive cells, we identified 109 genes whose expression varied significantly according to the intracellular iron content [4]. These differentially expressed genes segregated in 5 expression clusters whether they were over- or under-expressed in response to high- or low-iron concentrations : 10 genes over-expressed in ferritin overload, 60 genes over-expressed in hemin overload, 6 genes under-expressed in hemin overload, 15 genes over-expressed in ferritin depletion, and 18 genes over-expressed in iron free condition. Most of these genes had never

been associated with iron metabolism. Based on the assumption that co-regulated genes share similar promoters and/or are involved in similar biological processes [5], these genes could be potentially involved in novel mechanisms of iron metabolism regulation. The challenge from there was thus to biologically characterize our five clusters. Given that heme is a well-absorbed form of dietary iron found in meat and that it provides one third of the dietary Fe in North America and Europe, it takes a prominent place in iron absorption [6-8]. We therefore focused our functional annotation on the 60 genes over-expressed in hemin overload condition (HO cluster) .

The functional annotation of a gene cluster deals with the identification of shared biological features among those genes. This step is essential for biological understanding of the cluster and can be achieved in many ways. Annotation databases like UniProt, OMIM or GeneBank can be used retrieving informations for each gene of a cluster to bring out their common points. This manual method is an effective solution for a couple of genes but it is time-consuming and fastidious for large clusters. Besides, in order to be consistent and coherent, this annotation must be standardized.

GO (http://www.geneontology.org/) is a controlled vocabulary of about 20,000 terms organized in three independent hierarchies for cellular components, molecular functions, and biological processes. It has transformed the functional annotation of gene products and has rapidly become a *de facto* standard in that field. Indeed, GO adresses the need for consistent descriptions of gene products and the representation of the functional informations related to them. Most of the model organisms and annotation databases use GO terms to describe the gene products or have been mapped to the Gene Ontology. Most of these annotations are peer-reviewed as they are made by trained biologists (annotators) or extracted from annotation databases. These informations are publicly available in the Gene Ontology Annotation (GOA) database

(http://www.ebi.ac.uk/GOA/) and many dedicated analysis tools are available on-line [9] and offer automated, practical and efficient solutions for retrieving statistically over-represented – 'enriched' – GO terms associated with a gene cluster. The major problem is that these enriched terms are often broad and thus less informative terms of the GO hierarchies [10]. Moreover, annotating genes with a controlled vocabulary is a tedious task needing an expert to inspect carefully the literature associated with each gene to determine the appropriate terms. As a consequence, the number of gene products and associated data are increasing faster than they can be annotated, and there are genes for which attributes are not yet well known and for which the literature has not yet been investigated by curators [11]. Annotation databases are therefore incomplete. Due to these information gaps, the functional annotation of our clusters and their biological interpretations remained complex and laborious.

Another obvious way to gather informations about gene functions is to read the scientific literature. This method is also unrealistic for high-throughput data – biologists currently spend a lot of time and effort in searching for all the available information about each small area of research - but many efforts were made in Natural Language Processing (NLP) and a few groups have demonstrated the ability for a computer to mine associations between gene symbols and GO terms in the records of the MEDLINE database [12-16]. It is therefore possible to build a functional annotation based on such associations. In a previous work [17], we showed evidence of the efficiency of combining informations queried from an annotation database (the Gene Ontology Annotation database) and informations extracted from the biomedical literature (the PubGene index) (http://www.pubgene.org/) in order to achieve a more robust and exhaustive annotation.

In the present study we use our annotation method to infer a functional profile for a gene cluster involved in a response to high-iron concentration. This cluster is made of 60 genes over-expressed in hemin overload (HO) situations.

We will compare and exemplify our results with what can be obtained with a GOA-based annotation tool for the HO cluster. We will also consider the benefits and limits of our method in terms of precision (in the information retrieval sense), biological accuracy, representativeness and comprehensiveness. Our methodology as been published elsewhere [17] and we will primarily discuss here the biological interpretation of the HO cluster and show to what extend this method helped us in this task: generating working hypotheses and directing future work. This study is also an infrequent validation 'by hand' of an automated extraction method of biological knowledge from scientific literature.

## MATERIALS AND METHODS

### HO cluster

The dataset used throughout this study is a cluster of 60 genes over-expressed in hemin overload (HO) situations. This cluster was identified in a microarray screening (oligo 22K expression array, Agilent technology) for variations in gene expression correlating with intracellular iron content [4]. See Additional file 1 for details on the genes of the HO cluster.

### GOTM

GOTM (GOTree Machine) (http://genereg.ornl.gov/gotm/) is a web-based platform for interpreting microarray data or any other gene sets using the Gene Ontology. It is part of the gene

set annotation module of the WebGestalt toolkit (WEB-based GEne SeT AnaLysis Toolkit) (http://genereg.ornl.gov/webgestalt/). GOTM is cross-referenced for many identifiers including LocusLink, Affymetrix (16 arrays), UniGene, Swiss-Prot, Ensembl and gene symbols (4 species). Queries were performed with the list of the official gene symbols approved by the HUGO Nomenclature Committee (HGNC) for the HO cluster. We chose GOTM for its tree-mode GO visualization, the adequacy of its statistical model, and the possibility to upload our own list of genes as the set of reference genes.

**'Enriched' terms**

A common problem in functional genomic studies is to detect significant enrichements of GO terms (GO terms with a number of associated genes significantly higher than expected) within a class of genes of interest, typically the class of significantly differentially expressed genes. This task is performed using various statistical tests refered to as : the binomial test, the chi-square test, the equality of two probabilities test, Fisher's exact test and the hypergeometric test. According to Rivals et al. [18]**,** we chose the hypergeometric statistical model. The reference list of genes was the whole Agilent array. Terms with a p-value $\leq 0.05$ where considered enriched. The analysis was restricted to the enriched terms associated with at least two genes in the HO cluster.

**GOA**

The GOA database aims to provide high-quality supplementary GO annotation to proteins in the UniProt (SWISS-PROT/TrEMBL) databases. Unlike many other annotation databases like UniProt, OMIM or GeneBank, GOA offers a consistent description of the gene products by describing them with controlled terms of the Gene Ontology. Most of the GOA content comes

from the manual curation of scientific literature (peer-review), with semi-automatic and electronic techniques being used to support the annotation process. The GOA files were downloaded on October 2005. The same statistical model was used to identify enriched GO terms: hypergeometric with the whole Agilent array as a reference. Terms with a p-value $\leq 0.05$ were considered enriched. The analysis was also restricted to the enriched terms associated with at least two genes in the HO cluster. This method was designed to offer a straightforward comparison with GOTM.

**PubGene**

PubGene (http://www.pubgene.org/) is a web-based database of gene-gene and gene-term associations based on co-occurrences in biomedical literature. It provides a full-scale literature network for 25,000 human genes extracted from the titles and abstracts of over 14 million article records from the MEDLINE citation database of the National Library of Medecine (NLM). The method assumes that if two genes are mentioned in the same MEDLINE record there should be an underlying biological relationship. Genes are linked to terms from the Gene Ontology and a probabilistic score (p-value) is computed that reflects the gene-term association strength which can be used to assess the relevance of each individual term. The computation of this probabilistic score assumes that occurrences of the gene and the term are independent. Therefore, a binomial formula can be used to estimate the probability of finding the gene and the term together in an article based on their respective frequencies in the whole database. Assuming a normal distribution, the expected number of articles mentioning the gene and the term is then compared to the number of times they actually occur together (see [19] for details). In clusters, the reliability of each term is a multiplication of its probabilistic scores. The literature annotation was

carried out with the 2.5 release of the PubGene database. Obsolete terms were replaced by updated ones if present in the term definition table of the GO database (i.e., specified in the term comment attribute) and obsolete terms with no updated term were discarded from the literature annotation. Terms being poorly associated with the HO cluster ($p > 0.01$) were also discarded. Enriched GO terms were identified with a hypergeometric model (whole Agilent array as a reference, p-value $\leq 0.05$). The analysis was restricted to the enriched terms associated with at leasttwo genes in the HO cluster.

**GO and other tools**

The version of GO used throughout this study is the February 2005 monthly release, available from the GO website. DAG graphical representations were achieved using dot v1.10 and Graphviz 1.13(v16). All other graphics and statistical analyses were done using the R language version 2.1.0.

**Effectiveness measures**

A standard procedure used for evaluating information retrieval systems is the measure of its precision:

$$\text{Precision} = TP /(TP + FP)$$

In our evaluation, the precision is the fraction of retrieved enriched terms that are relevant (true positives, TP) compared to all the retrieved enriched terms (true positives and false positives, FP). The recall is the fraction of relevant terms retrieved by the method:

$$\text{Recall} = TP /(TP + FN)$$

This measure cannot be computed without quantifying all the relevant enriched terms that can be

associated with our cluster, and particularly the relevant terms not retrieved by the method (false negatives, FN).

## RESULTS

### GOTM

Forty-nine genes out of the 60 genes of the HO cluster are annotated with GO Process terms. GOTM identified 15 enriched terms ($p \leq 0.05$) associated with at least two genes of the HO cluster (Figure 1A). These enriched terms annotate 38 (63.3%) genes of the HO cluster (Table 1). Coarse-grained terms "metabolism" and "cellular metabolism" are respectively associated with 26 and 21 genes of the HO cluster. Seven enriched terms (46.6%) are specific to two genes: GCLC and GCLM (Figure 2). These genes encode the two subunits of the human glutamate-cysteine ligase (also known as gamma-glutamylcysteine synthetase) and are clearly up-regulated in the hemin condition to compensate for the decrease of reduced glutathione observed with iron overload [20]. The GOTM annotation emphasized two high-level pathways: "cellular lipid metabolism" (5 genes) and "digestion" (2 genes). Nine genes (AKR1C3, AK1, ASL, FADS3, FBP1, GCLC, GCLM, HMOX1, SLC23A1) were related to 11 enriched terms in the lower level cysteine and glutathione metabolism pathways (Figure 2).

### GOTM/GOA

Queries in the GOA Human database retrieved 81 GO Process terms associated with the same genes. Among the 7 terms enriched ($p \leq 0.05$) in GOA, 4 were also enriched in GOTM:

"digestion", "metabolism", "cysteine metabolism", "glutathione biosynthesis" (Figure 1A). Three terms were associated with the HO cluster in GOA but not in GOTM: "signal transduction", "steroid metabolism" and "nucleobase, nucleoside, nucleotide and nucleic acid metabolism". Eleven enriched GO terms in GOTM were never associated with any genes of the HO cluster in the GOA Human database (Figure 1A). The differences between the two methods resulted from the use of the GO 'True Path Rule' (http://www.geneontology.org/GO.usage.shtml#truePathRule) in GOTM. This rule states that "the pathway from a child term all the way up to its top-level parent(s) must always be true". One of its implications is that one can expand a gene-term annotation to all its parent terms and increase the number of genes associated with low-depth (high granularity) terms of the Directed Acyclic Graphs (DAGs) of GO. This will in turn increase the enrichment probabilities of such terms. For example, all the genes annotated with "steroid metabolism" (AKR1C2, NR1I2 ) in GOA (Figure 2) are associated with parent terms ("metabolism" and "cellular metabolism") in GOTM . Conversely, 7 genes are associated with the process "organic acid metabolism" in GOTM but this term is not enriched in GOA, these genes being associated with children terms: AKR1C3 ("prostaglandine metabolism"), ASL ("arginine catabolism", "arginine biosynthesis", "amino acid metabolism"), FADS3 ("fatty acid biosynthesis", "fatty acid desaturation"), FBP1 ("gluconeogenesis"), GCLC ("cysteine metabolism", "glutamate metabolism"), GCLM ("cysteine metabolism") and SLC23A1 ("L-ascorbic acid metabolism"). These two methods, based on evidence and peer-review, are highly relevant: there were no false positives (precision 100 %) among the enriched terms. This delay reflects however the laborious nature of a peer-reviewed annotation: our knowledge of biology increases much faster than it is formalized and one of the consequences is that annotation databases are incomplete and in constant evolution. At the time we performed the annotations,

GSTA1 and GSTA3 were only associated with "metabolism" in GOA but the situation has evolved since, and GSTA1 is now annotated with "glutathione metabolism" in GOA.

The use of the "True Path Rule" in GOTM results in an aggregation of the information in the highest nodes of the DAG. This allows an easier identification of well annotated but high-level pathways. Coarse-grained terms like "metabolism" and "cellular metabolism" are both associated with more than 20 genes of the HO cluster but bears little informative knowledge in the perspective of its biological interpretation. Pruning those two terms leads to an annotation with only 12 genes (GOTM) and 9 genes (GOA) associated with the 16 enriched terms left (Table 1). The resulting representativeness of the cluster annotation – 15% in GOA and 20% in GOTM – is clearly questionable. Moreover, 7 out of 15 enriched terms (46.6%) in GOTM and 2 out of 7 enriched terms (28.6%) in GOA are specifically associated with two genes: GCLC and GCLM. These genes should not be over-representative in a cluster annotation perspective. If we prune away the two coarse-grained terms and those specifically associated with GCLC/GCLM, the annotation of the HO cluster drops to 6 enriched terms associated with 10 genes in GOTM and 3 enriched terms associated with 7 genes in GOA.

**GOA+PubGene**

Queries in the PubGene database retrieved 856 GO Process terms associated with 49 genes of the HO cluster. This list was shortened to 336 terms having a strong association to the whole cluster (probabilistic score $\leq 0.01$) [17]. Sixty-three terms were enriched and associated with at least two genes (Figure 1B). The PubGene and GOA annotations shared 9 terms (13.4%), 4 terms (6%) were specific to GOA and 54 (80.6%) to PubGene. With the latter method, 10 false positives terms (precision 84.1%) were discarded from the annotation. These terms were easily identified

after reading of the associated titles/abstracts. They often resulted from the ambiguity of the gene symbols (the official symbol of the glucosidase beta acid gene is GBA and it can be confused with the acronyms of “gibberellic acid”, ”4-guanidinobutyric acid” or ”gamma-band activity”), or from obvious text-mining shortcomings (“cellulose metabolism” associated to ”fat+cellulose diet” studies or ”cellulose chromatography” studies). The recall of the literature annotation was not evaluated because it was impossible to quantify the false negatives. The remaining 57 enriched GO Process terms found by at least one of the two methods were associated with 35 genes (58.3%) of the HO cluster (Table 1). The bias toward GCLC/GCLM was still important with 15 specific terms (26.3%) but this method saved 43 enriched terms to account for the remaining 33 genes. When only 7 GO terms were enriched with GOA alone (Figure 1A), 13 terms associated with an evidence code were enriched with the combination of the two methods (Figure 1B). Six annotated – but not enriched – terms in GOA are brought back by good associations in the literature index.

Combining GOA and PubGene annotations emphasized four sub-DAGs in the Process hierarchy of GO representing 41 enriched terms (71.9% of the enriched terms) and 22 genes (61.1% of the annotated genes and 36.6% of the HO cluster genes).

- Pathway #1: lipid metabolism (Figure 3). This pathway gathers 6 terms annotating 7 genes. Most of these terms specifically emphasize the “steroid metabolism” process. Three genes are associated with one related pathway: “digestion” and one of its child terms “regulation of cholesterol absorption”.

- Pathway #2: amino acid and cofactor metabolisms (Figure 4). Ten terms annotating 5 genes are pooled under the “amino acid and derivative metabolism” process. Five terms annotate 7 genes under “cofactor metabolism”. This pathway links the cysteine metabolism (and its derivatives like taurine) with the “glutathione metabolism” (associated with 6 genes of which the glutathione S-transferase A1 and A3). The ”heme metabolism” process which is associated with 4 genes (HMOX1, NR1I2, GCLC, GCLM) and its child node ”heme oxidation” is also associated with HMOX1.

- Pathway #3: response to stimulus (Figure 5). Nine genes are associated with 6 children terms of “response to stimulus”. This stimulus can be either biotic (3 genes are associated with “cellular defense response” and 3 terms of the “cytokine biosynthesis” process are enriched) or abiotic (5 genes associated with “response to hydrogen peroxide” and 6 genes with “xenobiotic metabolism”). Most of the proteins coded by these genes (GSTA1, GSTA2, GCLM, GCLC, ALDH1A1) are active in the xenobiotic detoxification processes and NR1I2 is a nuclear receptor which is a transcriptional regulator of some cytochrome P450 genes. This pathway is strengthen by two related processes: the “cell redox homeostasis” and the “induction of apoptosis by oxidative stress”.

- Pathway #4: transport (Figure 6). Eight genes are associated with 9 children terms of “transport”. Three terms emphasize the transport mechanisms mediated by vesicles. The EXOC7 gene encodes a protein of the exocyst complex, a cellular structure essential to exocytosis and that may also play a role in modulating microtubule dynamics [21]. The related enriched term “membrane fusion” is associated with 2 genes.

## DISCUSSION

### Annotation methods

In the annotations based on evidence (GOTM and GOA), the over-expressed genes in hemin overload situation are associated with the metabolisms of lipids and glutathione (Figure 2). As a matter of fact, glutathione is involved in many vital mechanisms including antioxidation, maintenance of the redox state, immune response modulation and xenobiotics detoxification [22,23]. From these annotations, it is obvious that the informations found in annotation databases are highly relevant but not well suited to statistical enrichment as they narrow to rather general biological pathways when related to a cluster of 60 genes.

Nevertheless, the combination of these informations with associations extracted from the biomedical literature appends a large number of true positives terms to the annotation of the HO cluster. The resulting functional profile of this cluster is therefore more comprehensive and its representativeness is better as the number of annotated genes and the overall number of genes associated with each terms are increased [17]. This term-enrichment brings comprehensiveness in two ways. The first is an improved accuracy of the annotation: in a given pathway, the granularity of the enriched GO terms is finer for literature compared to evidence (depths of enriched literature terms in the DAG are higher than that of enriched evidence terms). For example, the five genes annotated with ”cellular lipid metabolism” in GOTM (Figure 2) are annotated with “steroid metabolism” and three children nodes with our method (Figure 3). The second way to improve our cluster understanding is to provide supplementary pathways. The method’s recall cannot be easily calculated without knowing all the relevant terms for the HO cluster, but the literature brings forward relevant pathways not enriched and sometimes even not annotated at all in the evidence annotation. Despite its ‘natural’ relevance regarding the oxidative

stress caused by high concentration of hemin [24], the Pathway #3 (Figure 5) was only identified by PubGene. In Pathway #2 (Figure 4), the 'obvious' "heme metabolism" pathway was also neither identified by GOTM nor by GOA.

It is however demonstrated that our method is less precise (in the perspective of information retrieval systems) than those based only on the annotation databases and that many improvements need to be made to the NLP techniques used to mine associations in the scientific literature. For example, the rather simple 'occurrence assumption' of PubGene is particularly sensitive to artifactual associations. The term "protein transport" was for example associated with 17 genes but only one had a direct implication in the protein transport mechanisms. This shortcoming mostly concerns words (terms) commonly used in natural language in biology and therefore frequently employed in the titles/abstracts of scientific articles. These terms are however usually high-level (low depth) terms of the Gene Ontology, and thus bear less information.

Annotating genes using text-mining of the biomedical literature can reveal relatively indirect associations. But these associations can still make sense and be accurate in the perspective of associating biological processes to a gene cluster. However, these indirect annotations are less relevant for molecular functions and cellular components.

**Iron overload and oxidative stress**

From a biological point of view, the accuracy of our functional annotation method is supported by the identification of well known pathways associated with our cluster of over-expressed genes in iron overload and by the coherence between those pathways.

For example, in Pathway #3 (response to stimulus, Figure 5), the two subunits of the glutamate-cysteine ligase (gamma-glutamylcysteine synthetase) — the first rate limiting enzyme of

glutathione synthesis — are linked with the ”response to hydrogen peroxide” and directly associated with the cellular detoxification (”cellular defense response”). Glutathione is known to be an ubiquitous molecule found in all parts of the cell where it fulfils a range of functions from detoxification to protection from oxidative damage. In Pathway #2 (amino acid and cofactor metabolism, Figure 4), the enriched GO term ”glutathione metabolism” is associated with six genes including GCLC /GCLM and the glutathion S-transferases (GSTA1 /GSTA3). The alpha-class glutathione S-transferases (GSTs) protect various cell types from oxidative stress and lipid peroxidation. These enzymes are involved in cellular defense against toxic, carcinogenic, and pharmacologically active electrophilic compounds [25]. The biological link between GCL and GSTs highlighted by our annotation method is therefore coherent. Besides, this link is reported in a recent study by Wielandt et al. [26].

**Putative iron regulation mechanisms**

Our functional annotation method may also suggest putative iron regulation mechanisms. The Pathway #4 (transport, Figure 6) propose a putative link between hemin overloaded Caco-2 cells and vesicular trafficking. Indeed, this pathway includes the EXOC7 gene which encodes a protein of the exocyst, a conserved eight-subunit complex involved in the docking of exocytic vesicles [27]. This biological link between iron transport and the exocyst was also recently reported for genetic anemia in mice [28]. Furthermore the TSPAN8 gene (encoding a member of the tetraspanin family) is also included in the Pathway #4 and indirectly associated with the transcytosis biological process [29].

The underlying biological relations of these two processes (transport and vesicular trafficking) with iron metabolism in Caco-2 cells are presently not known. Nevetheless, we cannot exclude

the participation of such processes as they have been previously described for the cellular relocalization of iron transporters in response to iron stores [30-33].

**Conclusions**

Our functional annotation method applied to a cluster of 60 genes over-expressed in Hemin Overload (HO) situation in Caco-2 cells highlights four subDAGs in the Process hierarchy of GO: lipid metabolism, amino acid and cofactor metabolism, response to stimulus and transport. The accuracy of this functional profile is supported by the identification of well known or recently characterized pathways associated with the iron overload (response to oxidative stress, glutathione metabolism).

Our functional annotation method also allowed us to elaborate *in silico* perspectives on the regulation of iron uptake in Caco-2 cells. These assumptions yet require a biological validation. This validation is however a complex task as some actors of the iron homeostasis process are only putative or not clearly identified (see for example the discussions about the role of the human proton-coupled folate transporter/heme carrier protein 1 hPCFT/HCP1 [31,34,35]). Though this process cannot be fully explained with present knowledge our *in silico* hypotheses are supported by recent evidences and could therefore be an interesting starting point for the biologists working on the iron metabolism and absorption (*in vitro* and *in vivo*). Further research will therefore confirm, refine or cancel our annotation work.

It is also clear that many gene products of the HO cluster are not yet functionally characterized and could be new potential actors involved in one of the molecular pathways underlying the iron overload.

Our functional annotation method is a valuable tool designed to help the biologist understand the

biological links between the genes of a cluster, elaborate working hypotheses and direct future work.

## ACKNOWLEDGMENTS

We would like to thank Bertrand Toutain for his technical support and is helpful advices.

## GRANTS

This work is supported by grants from the Conseil Général de Bretagne, OUEST-genopole® and the Centre National de la Recherche Scientifique (CNRS).

# REFERENCES

[1] M.W. Hentze, M.U. Muckenthaler, N.C. Andrews, Balancing acts: molecular control of mammalian iron metabolism, Cell 117 (2004) 285–297.

[2] P.J. Sargent, S. Farnaud, R.W. Evans, Structure/function overview of proteins involved in iron storage and transport, Curr. Med. Chem. 12 (2005) 2683–2693.

[3] P. Mladenka, R. Hrdina, M. Hubl, T. Simunek, The fate of iron in the organism and its regulatory pathways, Acta Medica 48 (2005) 127–135.

[4] C. Chicault, B. Toutain, A. Monnier, M. Aubry, P. Fergelot, A. Le Treut, M.D. Galibert, J. Mosser, Iron-related transcriptomic variations in CaCo-2 cells, an in vitro model of intestinal absorptive cells, Physiol. Genomics 26 (2006) 55–67.

[5] A. Schulze, J. Downward, Navigating gene expression using microarrays: a technology review, Nat. Cell Biol. 3 (2001) 190–195.

[6] C.E. Carpenter, A.W. Mahoney, Contributions of heme and nonheme iron to human nutrition, Crit. Rev. Food Sci. Nutr. 31 (1992) 333–367.

[7] A. Uc, J.B. Stokes, B.E. Britigan, Heme transport exhibits polarity in Caco-2 cells: evidence for an active and membrane protein-mediated process, Am. J. Physiol. Gastrointest. Liver Physiol. 287 (2004) 1150–1157.

[8] S. Kalgaonkar, B. Lonnerdal, Effects of dietary factors on iron uptake from ferritin by Caco-2 cells, J. Nutr. Biochem. 19 (2008) 33–39

[9] P. Khatri, S. Draghici, Ontological analysis of gene expression data: current tools, limitations, and open problems, Bioinformatics 21 (2005) 3587–3595.

[10] E.B. Camon, D.G. Barrell, E.C. Dimmer, V. Lee, M. Magrane, J. Maslen, D. Binns, R. Apweiler, An evaluation of GO annotation retrieval for BioCreAtIvE and GOA, BMC Bioinformatics 6 (2005) 71-81.

[11] S. Raychaudhuri, H. Schutze, R.B. Altman, Using text analysis to identify functionally coherent gene groups, Genome Res. 12 (2002) 1582–1590.

[12] L. Tanabe, U. Scherf, L.H. Smith, J.K. Lee, L. Hunter, J.N. Weinstein, MedMiner: an Internet text-mining tool for biomedical information, with application to gene expression profiling, Biotechniques 27 (1999) 1210–1217.

[13] T.K. Jenssen, A. Laegreid, J. Komorowski, E. Hovig, A literature network of human genes for high-throughput analysis of gene expression, Nat. Genet. 28 (2001) 21–28.

[14] D. Chaussabel, A. Sher, Mining microarray expression data by literature profiling, Genome Biol. 3 (2002) 1011–1026.

[15] S. Raychaudhuri, J.T. Chang, P.D. Sutphin, R.B. Altman, Associating genes with gene ontology codes using a maximum entropy analysis of biomedical literature, Genome Res. 12 (2002) 203–214.

[16] A.J. Perez, C. Perez-Iratxeta, P. Bork, G. Thode, M.A. Andrade, Gene annotation from scientific literature using mappings between keyword systems, Bioinformatics 20 (2004) 2084–2091.

[17] M. Aubry, A. Monnier, C. Chicault, M. de Tayrac, M.D. Galibert, A. Burgun, J. Mosser, Combining evidence, biomedical literature and statistical dependence: new insights for functional annotation of

gene sets, BMC Bioinformatics 7 (2006) 241–258.

[18] I. Rivals, L. Personnaz, L. Taing, M.C. Potier, Enrichment or depletion of a GO category within a class of genes: which test?, Bioinformatics 23 (2007) 401–407.

[19] L.A. Adamic, D. Wilkinson, B.A. Huberman, E. Adar, A literature based method for identifying gene-disease connections, Proc. IEEE Comput. Soc. Bioinform. Conf. 1 (2002) 109–117.

[20] P. Cornejo, P. Varela, L.A. Videla, V. Fernandez, Chronic iron overload enhances inducible nitric oxide synthase expression in rat liver, Nitric Oxide 13 (2005) 54–61.

[21] S. Wang, Y. Liu, C.L. Adamson, G. Valdez, W. Guo, S.C. Hsu, The mammalian exocyst, a complex required for exocytosis, inhibits tubulin polymerization, J. Biol. Chem. 279 (2004) 35958–35966.

[22] G.K. Balendiran, R. Dabur, D. Fraser, The role of glutathione in cancer, Cell Biochem. Funct. 22 (2004) 343–352.

[23] U.E. Schaible, S.H. Kaufmann, Iron and microbial infection, Nat. Rev. Microbiol. 2 (2004) 946–953.

[24] V. Herbert, S. Shaw, E. Jayatilleke, T. Stopler-Kasdan, Most free-radical injury is iron-related: it is promoted by iron, hemin, holoferritin and vitamin C, and inhibited by desferoxamine and apoferritin, Stem Cells 12 (1994) 289–303.

[25] C. Frova, Glutathione transferases in the genomics era: new insights and perspectives, Biomol. Eng. 23 (2006) 149–169.

[26] A.M. Wielandt, V. Vollrath, M. Farias, J. Chianale, Bucillamine induces glutathione biosynthesis via activation of the transcription factor Nrf2, Biochem. Pharmacol. 72 (2006) 455–462.

[27] J.H. Lipschutz, K.E. Mostov, Exocytosis: the many masters of the exocyst, Curr. Biol. 12 (2002) 212–214.

[28] J.E. Lim, O. Jin, C. Bennett, K. Morgan, F. Wang, C.C. Trenor, M.D. Fleming, N. Andrews, A mutation in Sec15l1 causes anemia in hemoglobin deficit (hbd) mice, Nat. Genet. 37 (2005) 1270–1273.

[29] D. Lo, W. Tynan, J. Dickerson, M. Scharf, J. Cooper, D. Byrne, D. Brayden, L. Higgins, C. Evans, D.J. O'Mahony, Cell culture modeling of specialized tissue: identification of genes expressed specifically by follicle-associated epithelium of Peyer's patch by expression profiling of Caco-2/Raji co-cultures, Int. Immunol. 16 (2004) 91–99.

[30] Y. Ma, R.D. Specian, K.Y. Yeh, M. Yeh, J. Rodriguez-Paris, J. Glass, The transcytosis of divalent metal transporter 1 and apo-transferrin during iron uptake in intestinal epithelium, Am. J. Physiol. Gastrointest. Liver Physiol. 283 (2002) 965–974.

[31] M. Shayeghi, G.O. Latunde-Dada, J.S. Oakhill, A.H. Laftah, K. Takeuchi, N. Halliday, Y. Khan, A. Warley, F.E. McCann, R.C. Hider, D.M. Frazer, G.J. Anderson, C.D. Vulpe, R.J. Simpson, A.T. McKie, Identification of an intestinal heme transporter, Cell 122 (2005) 789–801.

[32] T.A. Rouault, The intestinal heme transporter revealed, Cell 122 (2005) 649–651.

[33] G.O. Latunde-Dada, R.J. Simpson, A.T. McKie, Recent advances in mammalian haem transport, Trends Biochem. Sci. 31 (2006) 182–188.

[34] N.C. Andrews, When is a heme transporter not a heme transporter? When it’s a folate transporter, Cell Metab. 5 (2007) 5–6.

[35] Y. Nakai, K. Inoue, N. Abe, M. Hatakeyama, K.Y. Ohta, M. Otagiri, Y. Hayashi, H. Yuasa, Functional characterization of human proton-coupled folate transporter/heme carrier protein 1 heterologously expressed in mammalian cells as a folate transporter, J. Pharmacol. Exp. Ther. (2007) 469–476.

# FIGURE LEGENDS

**Figure 1 - Overlaps between annotation methods**

Number of enriched GO terms associated with at least two genes in each annotation methods. (A) Overlap between GOTree Machine (GOTM) and Gene Ontology Annotation (GOA). (B) Overlap between GOA and PubGene. In brackets: number of true positives terms retrieved by PubGene.

**Figure 2 - GOTM/GOA annotation of the HO cluster**

Enriched GO terms associated with at least two genes of the HO cluster. GOTree Machine annotation terms in grey rectangles. GOA annotation terms in bold rectangles. Annotated genes are associated with all parent terms.

**Figure 3 - Pathway #1 (Cellular lipid metabolism and digestion)**

Enriched GO terms associated with at least two genes of the HO cluster: GOA (lightgrey), PubGene (darkgrey), GOA+PubGene (black).

**Figure 4 - Pathway #2 (Amino acid and cofactor metabolisms)**

Enriched GO terms associated with at least two genes of the HO cluster: GOA (lightgrey), PubGene (darkgrey), GOA+PubGene (black).

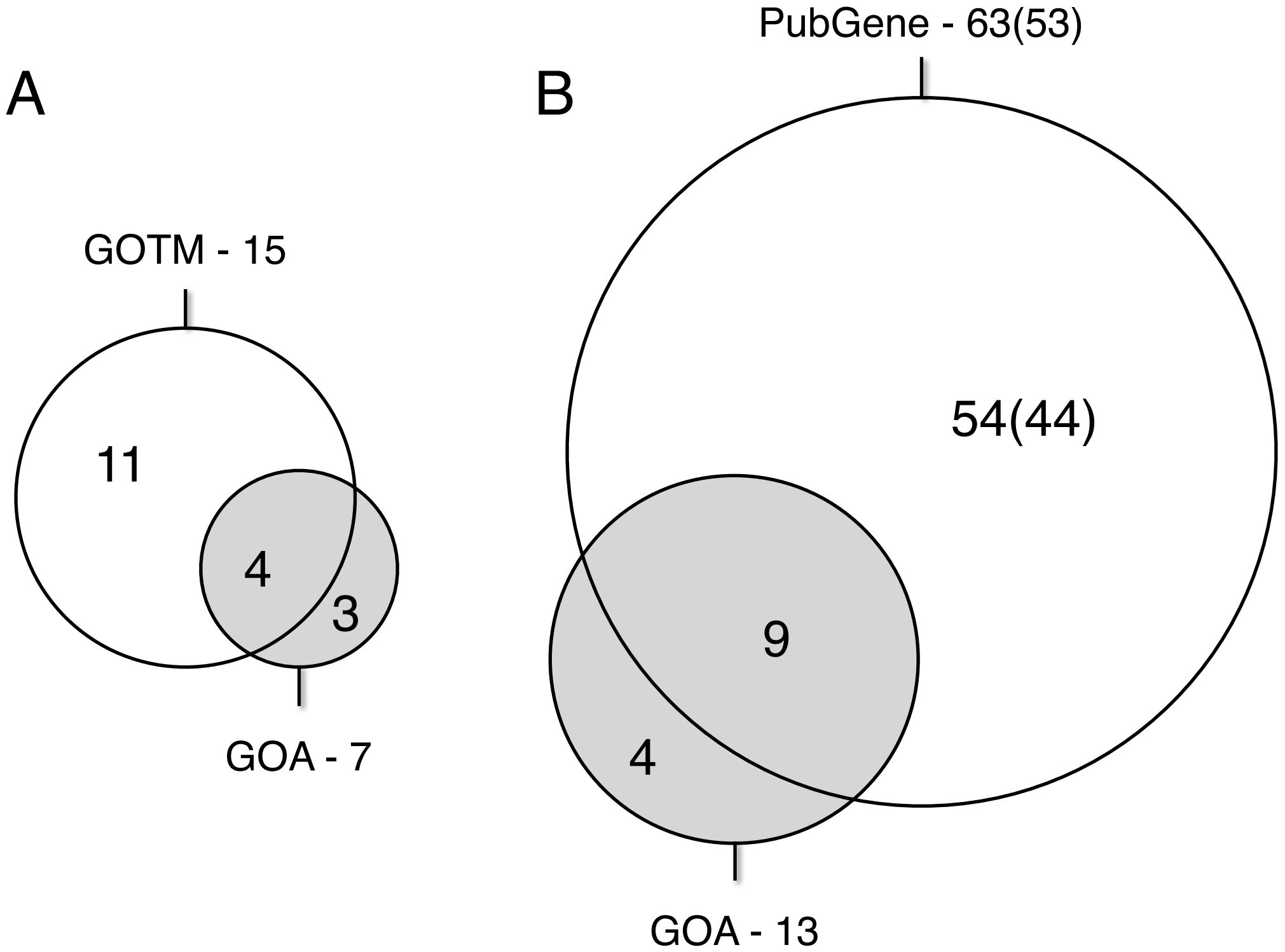
**Figure 5 - Pathway #3 (Response to stimulus)**

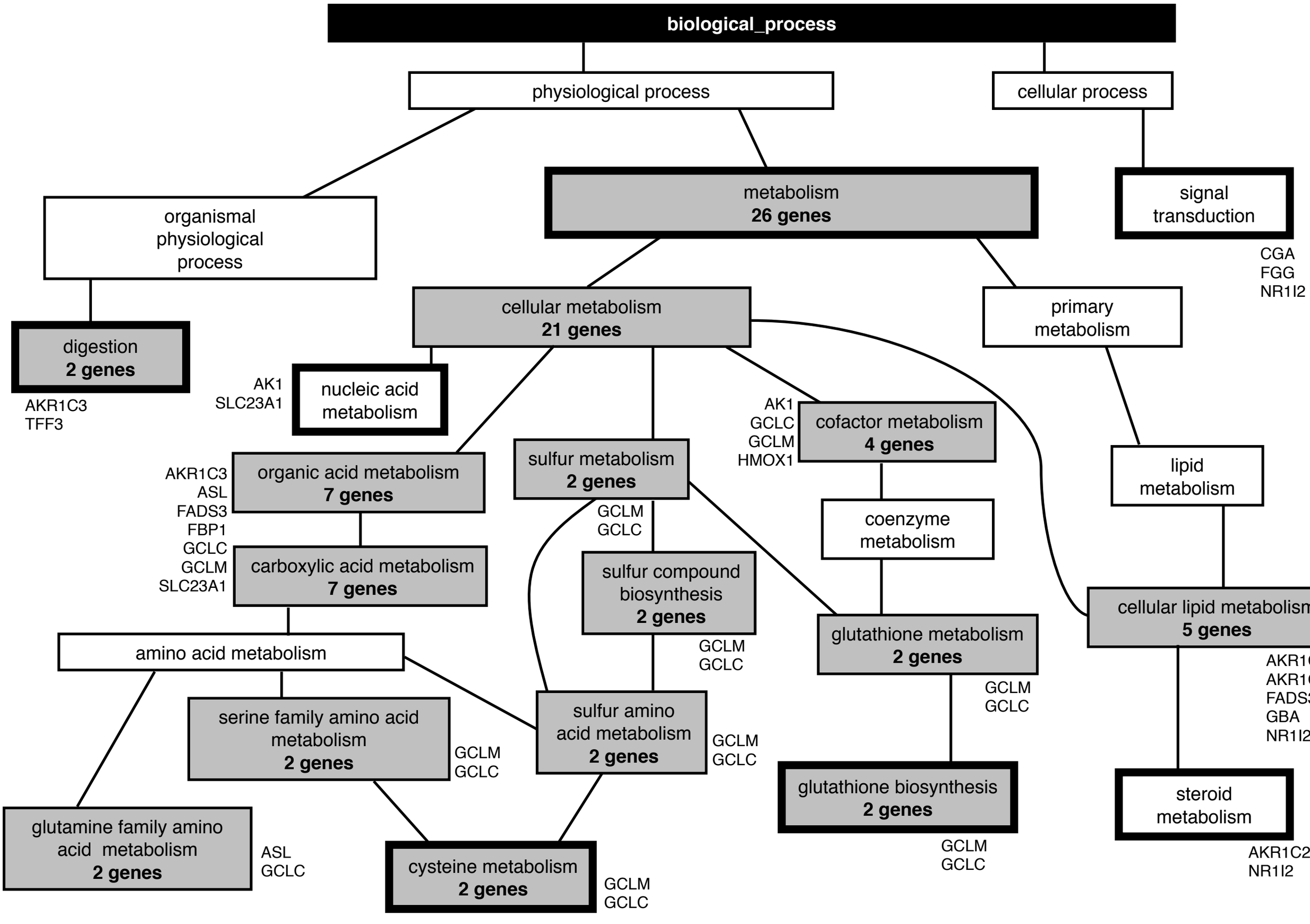
Enriched GO terms associated with at least two genes of the HO cluster: GOA (lightgrey), PubGene (darkgrey), GOA+PubGene (black). The terms ”defense response”, ”immune response”

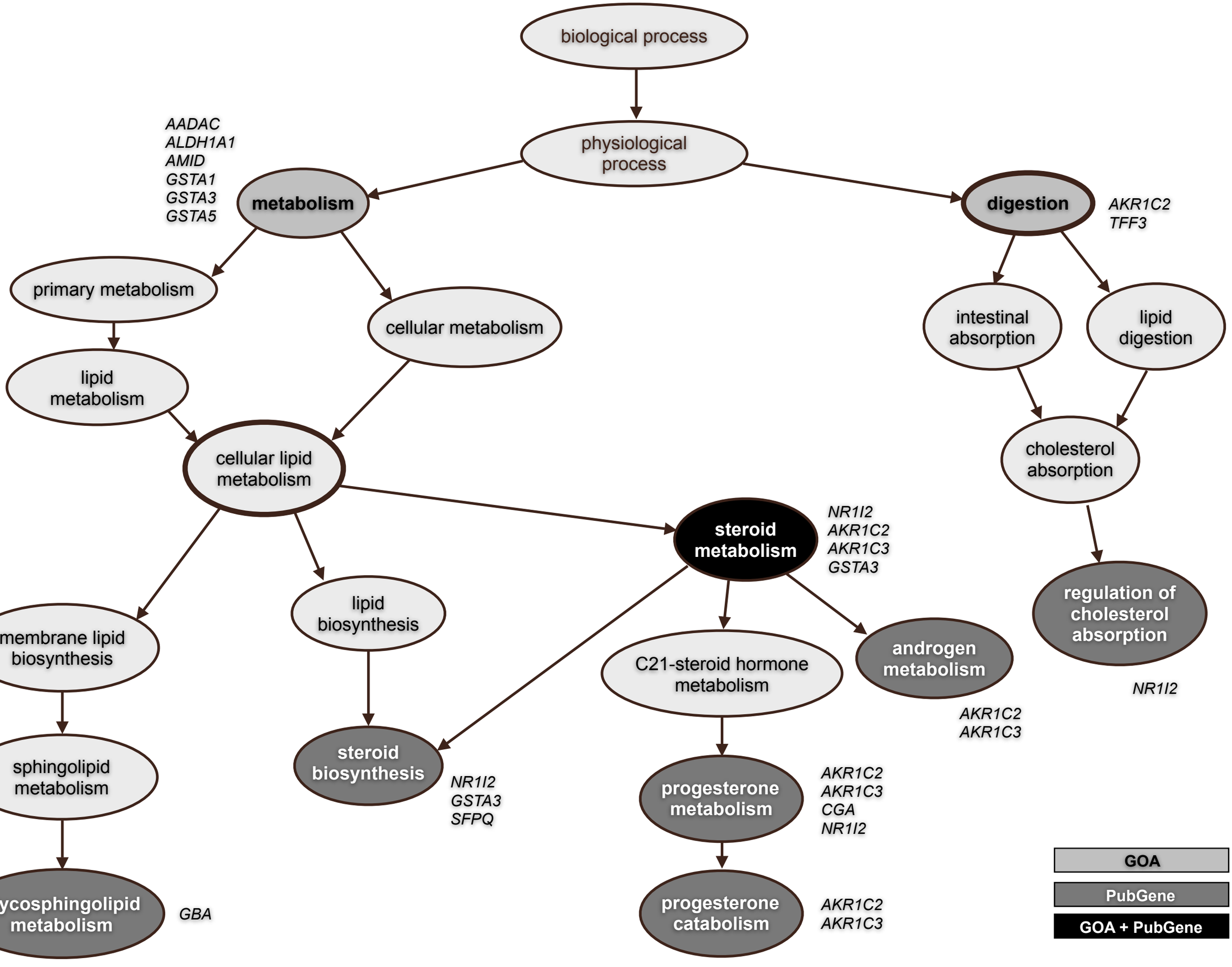
and ”response to stress” were found by GOA but were not enriched or only associated with one gene.

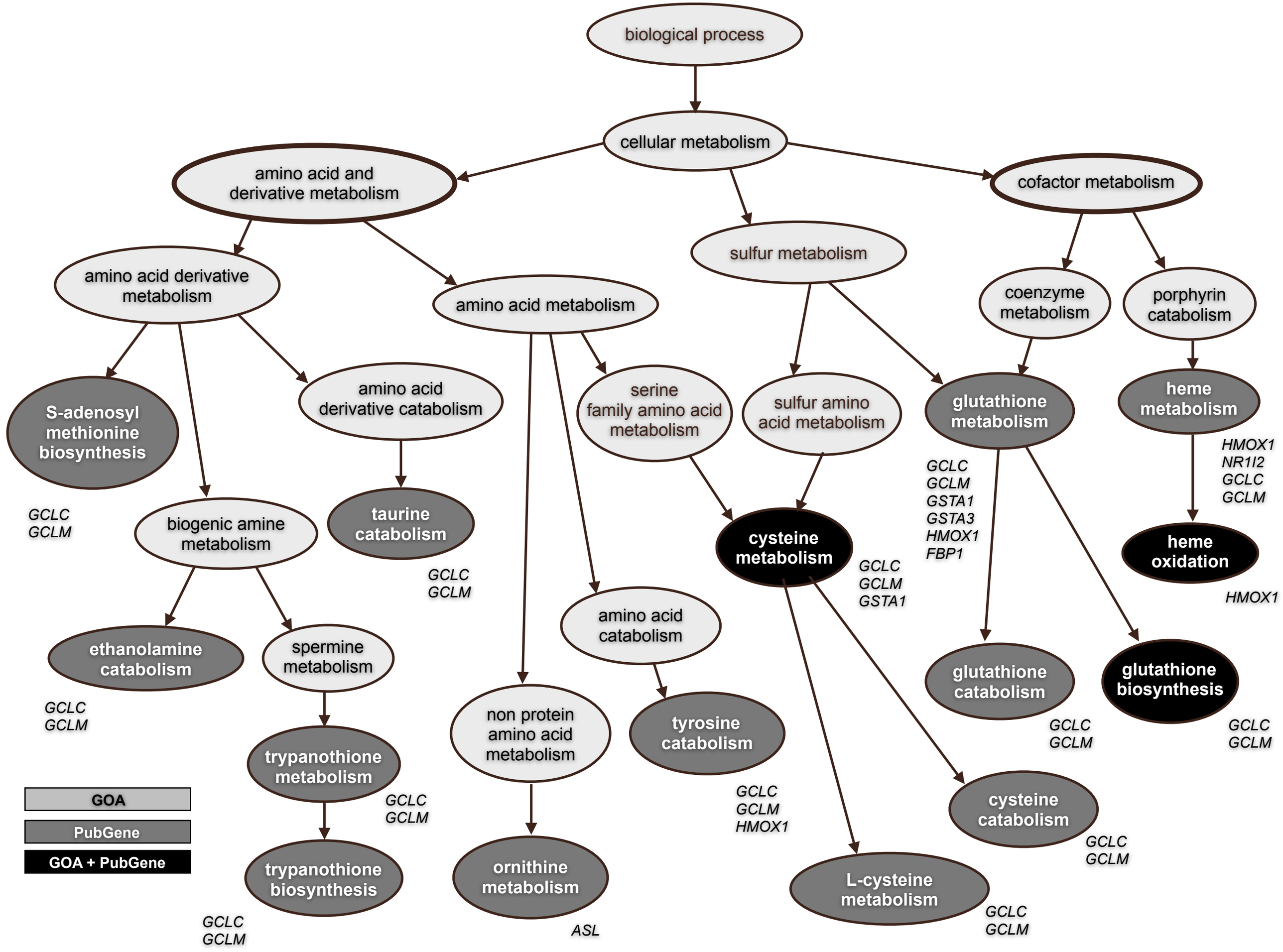
**Figure 6 - Pathway #4 (Transport)**

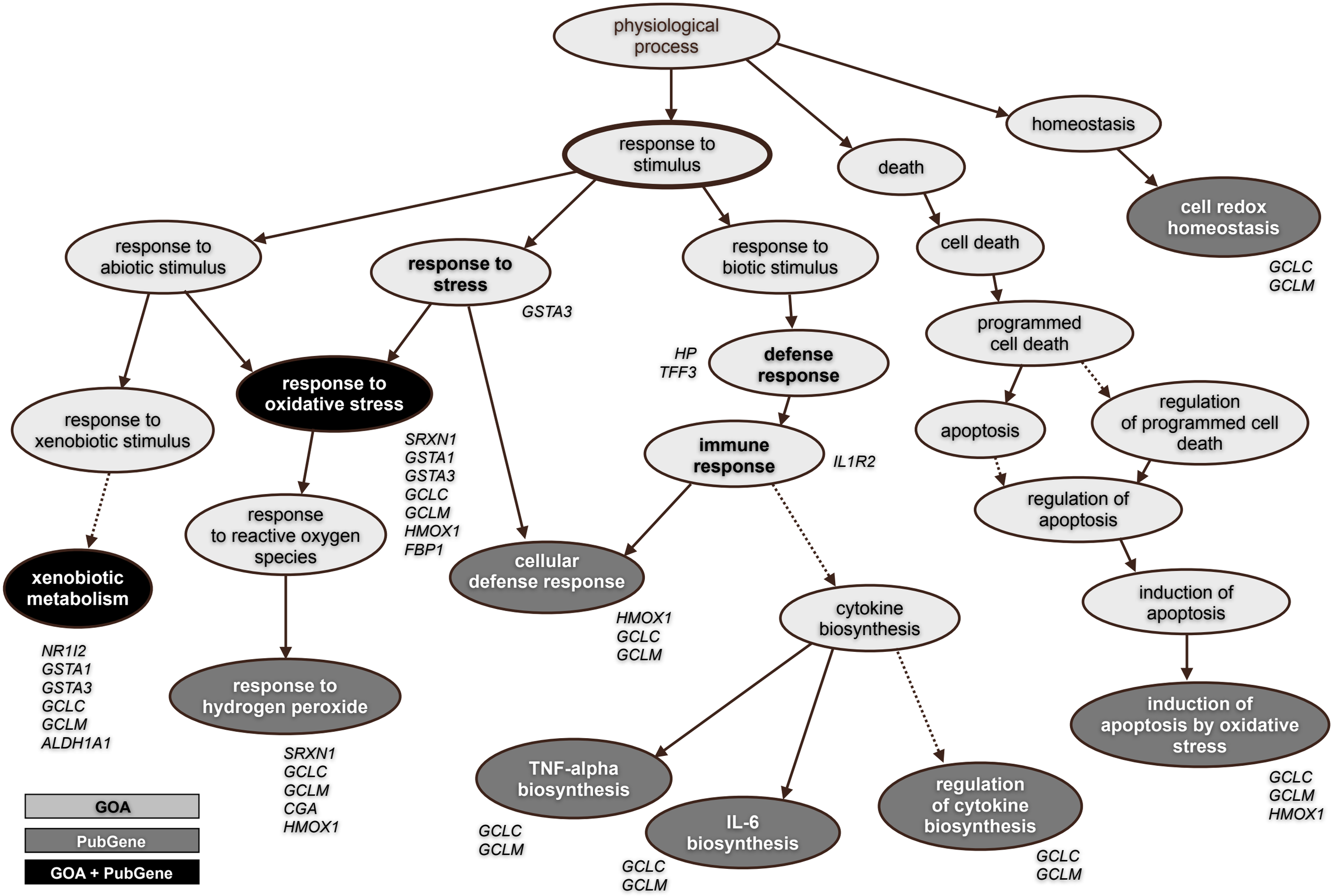
Enriched GO terms associated with at least two genes of the HO cluster: GOA (lightgrey), PubGene (darkgrey), GOA+PubGene (black).

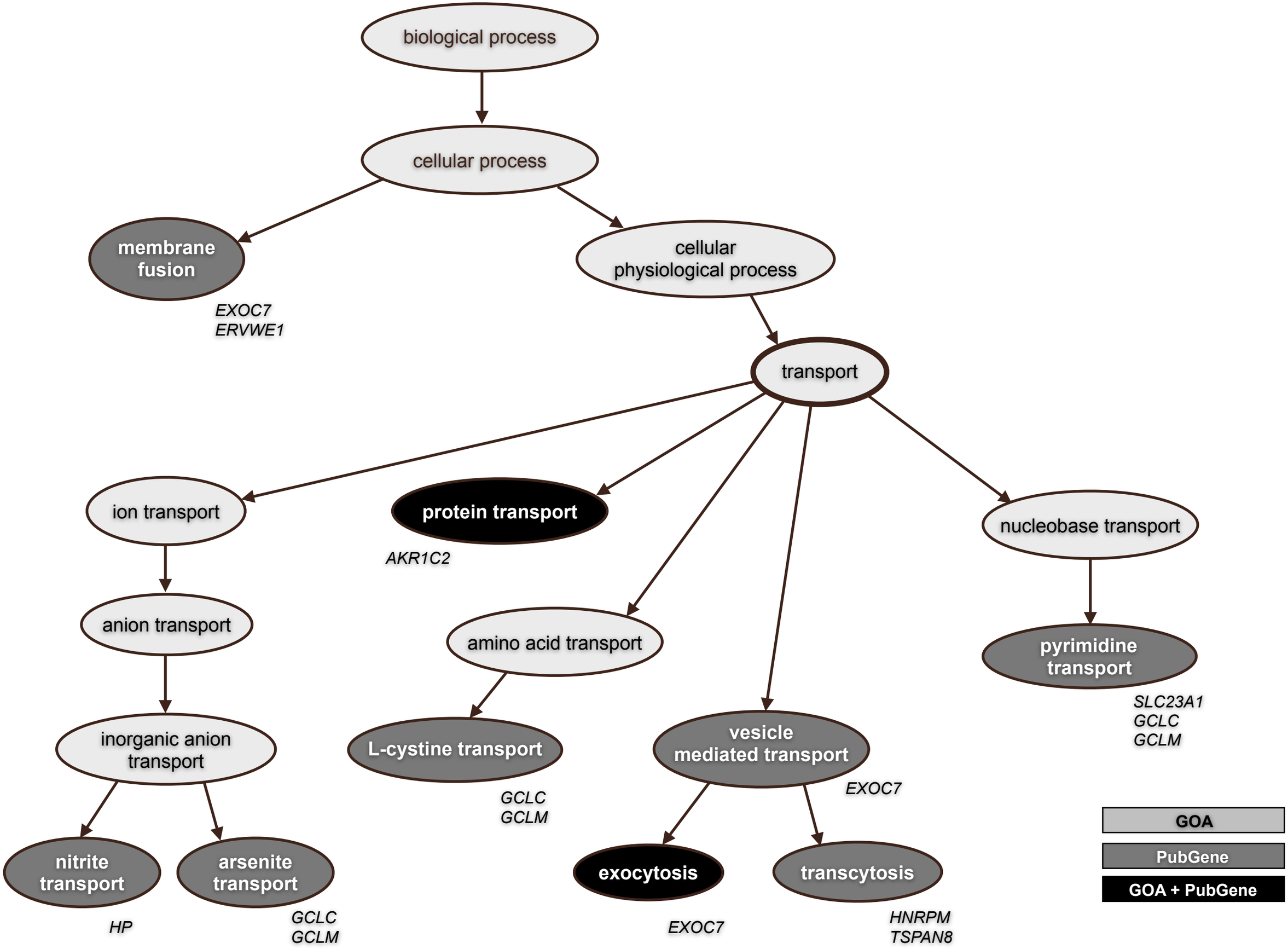

A
GOTM - 15
11
4
3
GOA - 7
B
PubGene - 63(53)
54(44)
9
4
GOA - 13


biological_process
physiological process
cellular process
organismal physiological process
metabolism
26 genes
signal transduction
CGA
FGG
NR1I2
digestion
2 genes
AKR1C3
TFF3
cellular metabolism
21 genes
primary metabolism
AK1
SLC23A1
nucleic acid metabolism
AK1
GCLC
GCLM
HMOX1
cofactor metabolism
4 genes
lipid metabolism
AKR1C3
ASL
FADS3
FBP1
GCLC
GCLM
SLC23A1
organic acid metabolism
7 genes
sulfur metabolism
2 genes
GCLM
GCLC
coenzyme metabolism
carboxylic acid metabolism
7 genes
sulfur compound biosynthesis
2 genes
GCLM
GCLC
cellular lipid metabolism
5 genes
AKR1C2
AKR1C3
FADS3
GBA
NR1I2
amino acid metabolism
glutathione metabolism
2 genes
GCLM
GCLC
serine family amino acid metabolism
2 genes
GCLM
GCLC
sulfur amino acid metabolism
2 genes
GCLM
GCLC
glutathione biosynthesis
2 genes
GCLM
GCLC
steroid metabolism
AKR1C2
NR1I2
glutamine family amino acid metabolism
2 genes
ASL
GCLC
cysteine metabolism
2 genes
GCLM
GCLC


biological process
physiological process
metabolism
AADAC
ALDH1A1
AMID
GSTA1
GSTA3
GSTA5
digestion
AKR1C2
TFF3
primary metabolism
cellular metabolism
intestinal absorption
lipid digestion
lipid metabolism
cellular lipid metabolism
cholesterol absorption
steroid metabolism
NR1I2
AKR1C2
AKR1C3
GSTA3
membrane lipid biosynthesis
lipid biosynthesis
regulation of cholesterol absorption
NR1I2
C21-steroid hormone metabolism
androgen metabolism
AKR1C2
AKR1C3
sphingolipid metabolism
steroid biosynthesis
NR1I2
GSTA3
SFPQ
progesterone metabolism
AKR1C2
AKR1C3
CGA
NR1I2
glycosphingolipid metabolism
GBA
progesterone catabolism
AKR1C2
AKR1C3
GOA
PubGene
GOA + PubGene


biological process
cellular metabolism
amino acid and derivative metabolism
cofactor metabolism
amino acid derivative metabolism
amino acid metabolism
sulfur metabolism
coenzyme metabolism
porphyrin catabolism
S-adenosyl methionine biosynthesis
amino acid derivative catabolism
serine family amino acid metabolism
sulfur amino acid metabolism
glutathione metabolism
heme metabolism
GCLC GCLM
biogenic amine metabolism
taurine catabolism
GCLC GCLM
cysteine metabolism
GCLC GCLM GSTA1
GCLC GCLM GSTA1 GSTA3 HMOX1 FBP1
HMOX1 NR1I2 GCLC GCLM
heme oxidation
HMOX1
ethanolamine catabolism
spermine metabolism
amino acid catabolism
glutathione catabolism
glutathione biosynthesis
GCLC GCLM
GCLC GCLM
GCLC GCLM
trypanothione metabolism
non protein amino acid metabolism
tyrosine catabolism
GCLC GCLM HMOX1
cysteine catabolism
GCLC GCLM
GOA
PubGene
GOA + PubGene
trypanothione biosynthesis
ornithine metabolism
L-cysteine metabolism
GCLC GCLM
GCLC GCLM
ASL
GCLC GCLM


physiological process
response to stimulus
homeostasis
death
response to abiotic stimulus
response to stress
response to biotic stimulus
cell death
cell redox homeostasis
GCLC
GCLM
GSTA3
HP
TFF3
defense response
programmed cell death
response to xenobiotic stimulus
response to oxidative stress
immune response
IL1R2
apoptosis
regulation of programmed cell death
SRXN1
GSTA1
GSTA3
GCLC
GCLM
HMOX1
FBP1
regulation of apoptosis
response to reactive oxygen species
cellular defense response
induction of apoptosis
xenobiotic metabolism
HMOX1
GCLC
GCLM
cytokine biosynthesis
NR1I2
GSTA1
GSTA3
GCLC
GCLM
ALDH1A1
response to hydrogen peroxide
induction of apoptosis by oxidative stress
SRXN1
GCLC
GCLM
CGA
HMOX1
TNF-alpha biosynthesis
GCLC
GCLM
IL-6 biosynthesis
GCLC
GCLM
regulation of cytokine biosynthesis
GCLC
GCLM
GCLC
GCLM
HMOX1
GOA
PubGene
GOA + PubGene


biological process
cellular process
membrane fusion
EXOC7
ERVWE1
cellular physiological process
transport
ion transport
anion transport
inorganic anion transport
nitrite transport
HP
arsenite transport
GCLC
GCLM
protein transport
AKR1C2
amino acid transport
L-cystine transport
GCLC
GCLM
vesicle mediated transport
EXOC7
exocytosis
EXOC7
transcytosis
HNRPM
TSPAN8
nucleobase transport
pyrimidine transport
SLC23A1
GCLC
GCLM
GOA
PubGene
GOA + PubGene