repiSalud
Repositorio Institucional en Salud

This is the peer reviewed version of the following article:

Fresno C, Llera AS, Girotti MR, Valacco MP, Lopez JA, Podhajcer OL, et al. The multi-reference contrast method: Facilitating set enrichment analysis. Comput Biol Med. 2012;42(2):188-94

which has been published in final form at:

https://doi.org/10.1016/j.compbiomed.2011.11.007

THE MULTI-REFERENCE CONTRAST METHOD: FACILITATING SET ENRICHMENT ANALYSIS

Cristóbal Fresno[1,*], Andrea S Llera[2,3] María R Girotti[2,3,a], María P Valacco[2, 3], Juan A López[5], Osvaldo Podhajcer[2,3], Mónica G Balzarini[2,4], Federico Prada[6] and Elmer A Fernández[1,2]

[1]BioScience Data Mining Group, Catholic University of Córdoba, Córdoba, Argentina.

[2]CONICET, Buenos Aires, Argentina.

[3]Laboratory of Molecular and Cellular Therapy, Leloir Institute, Buenos Aires, Argentina.

[4]Biometry Laboratory, National University of Córdoba, Córdoba, Argentina.

[5]National Center for Cardiovascular Research, Madrid, Spain

[6]Institute of Technology, School of Engineering and Sciences, UADE, Buenos Aires, Argentina.

[a] Present address: The Institute of Cancer Research, London, UK

* Corresponding author: Universidad Católica de Córdoba, Av. Armada Argentina 3555 (5017). Córdoba - Argentina. E-mail: cristobalfresno@gmail.com

ABSTRACT

Motivation: Set enrichment analysis (SEA) is used to identify enriched biological categories/terms within a high-throughput differential expression experiments. This is done by evaluating the proportion of differentially expressed genes against a background reference (BR): usually the whole genome or the chip-imprinted gene list. However it is seldom possible to evaluate the whole genome in some experimental settings. In addition different BRs may render different results thus, depending on the BR used, potentially relevant terms could be missing. To enhance SEA, we propose a visualization procedure to explore and identify relevant terms by means of simultaneous use of multiple BRs to enhance SEA. The proposed strategy is referred as multi-reference contrast method (MRCM).

Results: By means of the MRCM, it was able to find new biologically relevant information terms in various genomic/proteomic experiments. The use of the whole genome, the chip plus a user defined BR may provide new biological insights. Non-consensus terms were highly relevant and automatically highlighted by the MRCM, facilitating the exploration task evolved in ontology analysis on genomic/proteomic experiments. The method was evaluated in three microarrays

and one proteomic study where new relevant ontology categories/terms were identified and validated by literature.

Contact: cristobalfresno@ucc.edu.ar

Availability and Implementation: R source code available at https://sites.google.com/site/biologicaldatamininggroup/software.

Keywords: Gene ontology; Genomics; Proteomics; Data integration; Background reference

INTRODUCTION

Ontology analysis is currently one of the main steps in most of high-throughput genomic/proteomic experiments. It is usually carried out to relate differentially expressed genes/proteins (DEG/P) to biologically relevant terms, identifying enriched functions and/or pathways within the experiment. This task is achieved by querying large databases with controlled vocabulary (known as ontologies), where genes and their known functionality and location are stored. Gene Ontology (GO, www.geneontology.org) is the most popular ontology and it is structured as three direct acyclic graphs (DAGs) - molecular function (MF), biological process (BP) and cellular component (CC) [1]. Each DAG node represents a biological concept and has several genes associated to it. The nodes are organized in a hierarchical manner where the root is the most generic category/term and more specific ones are found downstream on the DAG.

Set enrichment analysis (SEA) is the traditionally used approach for ontology analysis, due to its trajectory and availability over commercial and public tools/websites [2-3]. Despite the different statistic variants for SEA calculations, each GO term/category is evaluated in a term by term fashion [4]. A term would result enriched if the observed proportion differed from the null distribution when compared with a background reference (BR). This reference (BR) could be:

i. The genome of the studied species (usually the default option in most bioinformatic tools and the only one in proteomic studies).
ii. The chip-specific gene list in microarray experiments.
iii. Specified by the user (“user defined”).

The Database for Annotation, Visualization and Integrated Discovery (DAVID, [5-6]) allows the selection of BR according to the needs of the experiment type and GoMiner [7], another well

known Ontology tool, requires the user to include the BR for the analysis. However, it was acknowledged that different BRs could yield different results and inappropriate BR selection could contradict statistical assumptions potentially biasing interpretation of results [2]. Currently there is no consensus about which the best option is in each experimental setting. For instance, when using the microarray gene set (or a subset of genes) the representation of all genes will be biased (representation bias). To avoid this, the full set of genes in the genome could be used, but another bias is introduced because there are several genes/proteins that cannot be detected [7].

In addition, a big effort is necessary by the user even for a single reference output interpretation (SEA outputs are usually long tabular lists or huge trees). To overcome this issue, some alternatives were developed such as, trimming of the DAG [8] or p-adjustment/filtering of the enrichment scores[3] which, in any case, valuable information could be lost.

Some visualization methods are provided by SEA tools. For instance, clustering visualization in DAVID to represent the relationship between genes and literature supported annotations (i.e. Gene Ontology, Kyoto Encyclopedia of Genes and Genomes, etc.) or DAG structures in GoMiner. However, they do not permit the comparison of results from different BRs.

In this work we present a method that integrates and visualizes results using several references at the same time. We do so by simultaneously querying DAVID with different BRs. Then, informative nodes/terms are automatically highlighted, by contrasting results, and displaying the ontology tree with a color code to visual integrate set enrichment results in a unique graph, for each main GO category (MF, BP and CC). The proposed strategy is named multi-reference contrast method (MRCM) and it is free available at https://sites.google.com/site/biologicaldataminginggroup/software.

We show that meaningful biological information is recovered by means of multi-reference comparison that is missed when only a single BR is used. The MRCM not only recovers enriched terms, but also improves the identification by highlighting informative terms using a color pattern. The method also allows an easy and fast identification of new informative terms, only seen with the user defined reference, enhancing the typical reference choice (genome or chip genes). The proposed strategy was tested on three microarray studies from the Gene

Expression Omnibus repository (GEO, www.ncbi.nlm.nih.gov/geo) and in a proteomic 2D-DIGE experiment. In all cases new relevant terms supported by evidence in the literature were identified.

METHODS

The following BRs were used:

- BR-I: the whole genome gene list, as provided by DAVID (default).
- BR-II: the complete gene list present in the analyzed chip (i.e. Affymetrix chips present in DAVID). This reference is only available for microarray experiments.
- BR- III: a smaller reference gene list, built from only those genes that were detectable in the experiment, such as using detection calls of flagged probesets in microarray chips[9-13] or identified spots in proteomic 2D-gels.

Affymetrix chips were processed using *affy* package [14-16]. Probesets having an expression call = "P" and unique anti-sense "_at" codification were used and differentially expressed genes identified by means of Bioconductor's *limma* package [17]. False discovery correction was applied in the detection procedure. The 2D-DIGE proteomic experiment was analyzed according to Fernandez *et al*. [18] recommendations.

A differentially expressed gene/protein list was uploaded to DAVID and enriched terms for each BR were identified using the full GO annotation (MF, BP and CC). Functional annotation charts were obtained without any filtering (EASE score = 1 and count threshold = 1) for each BR. In this way, all GO terms were retrieved from DAVID and locally stored. Then, EntreZ Gene ID and symbol names were obtained through DAVID to standardize gene identification as required by Bioconductor packages (Figure 1).

FIGURE 1 near here

A threshold EASE score value of 0.1 (as suggested in [19]) was used to identify enriched terms. These terms were then displayed using GO DAG structures provided by *GoStats*[20] and color-coded depending on if they were identified by one or more BRs. Finally, all the information was presented using a HTML front end report, which allows visual inspection of the contrasted DAGs.

The proposed method (MRCM) contrasts the enriched nodes provided by each BR evaluation and displays them, color-coded, in each GO DAG. The color scheme allows us to identify nodes found enriched by one or several BR evaluations. Enriched nodes were labeled as follows:

- Consensus node (CN): term enriched in all BRs.
- Non-consensus node (nCN): term enriched in at least one BR but not in all BRs.
- Not-enriched node (NEN): not enriched inner node of the DAG structure.

As CN nodes were able to be identified in any case, only nCNs were explored and validated. In particular, we focused our interest on leaves (nodes without child nodes), because they contain the most specific biological information and explain ancestor terms. Relevant selected candidates were validated by searching published papers in PubMed databases.

Example datasets for the MCRM

Human smoke dataset

We have analyzed data from Spira *et al*. [21] that compared the effects on bronchial epithelia from humans who were current smokers (CS), had never smoked (NS) or were former smokers. They concluded that cigarette smoking induces xenobiotic, redox-regulating and several oncogenes and decreases expression of several tumor suppressor and airway inflammation modulator genes. They also reported some potential oncogenes and tumor suppressors that failed to return to their normal expression level in former smokers.

Affymetrix HG-U133A sample chips (20 for each group, CS and NS) were pre-processed according to McClintick & Edenberg [12] (available at GEO repository with GSE994 accession series). However, we only include in the analysis those genes whose detection call was present in at least four chips in each group. This results in 4128 genes for differential expression analysis. One hundred and sixteen of these were differentially expressed (73 up and 43 down) using an adjusted p-value < 0.05 and a $|\log_2(\text{fold-change})| > 0.4$ (Supplementary Table 1).

Mouse smoke dataset

We also analyzed data from McGrath-Morrow *et al*. [22] that studied lung expression over a 14 day exposure to cigarette smoke (CgS) of neonatal *Mus musculus*. They showed that perinatal lungs were particularly susceptible to the damaging effects of CgS, inhibiting innate immunity and mildly impairing postnatal lung growth.

Gene expression profiling was carried out using Affymetrix Mouse Genome 430 2.0 microarray on lung samples collected from 6 mice exposed to CgS and 4 controls (available at GEO repository with GSE7310 accession series). In this case intensity was scaled for a target of 500 and genes present in at least 3 CgS chips and 2 controls were included (12905 genes). One hundred and eighteen differentially expressed genes were identified (10 up and 108 down) using an adjusted p-value < 0.05 (Supplementary Table 2).

Melanoma genomic dataset

Dataset from Packer *et al*. [23] was also analyzed. They carried out gene expression profiling to select novel downstream effectors of p14ARF in humans.

Affymetrix Human Genome U133 plus 2.0 microarrays were used for this study (12 wild-type and 23 mutant p14ARFs). This dataset is available at GEO repository with GSE7152 accession series. Intensity was scaled for a target of 500 and genes present in 6 wild-type and 12 mutant chips were included (11986 genes). One hundred and sixty five of these were found to be differentially expressed (68 up and 97 down) using an adjusted p-value < 0.05 (Supplementary Table 3).

Melanoma proteomic dataset

A data set of differential proteins was obtained from 2D-DIGE analysis of secretomes (i.e. extracellular proteins) of two melanoma cell lines that varied in the levels of expression of the protumorigenic protein SPARC ([24] and Girotti *et al*, unpublished). In the case of proteomic 2D-DIGE experiments, the nature of experimentation with proteins indicates that a BR-II is never available (there is no *a priori* list of proteins to be detected). Moreover, biological and technical constraints allow seeing only a subset of the proteins actually present in the proteome under study. In this particular case, for example, only extracellular proteins were analyzed. In addition, analysis of differential proteins by DIGE occurs before protein identification, and identification methods are usually applied only on differential spots. In order to build a BR-III for this experiment, we started by identifying all proteins present in the analyzed 2D-DIGE gels. The resulting list of proteins was then pooled with a list of secreted proteins of the same cell lines separated and identified by LC-MS/MS using an Orbitrap (Girotti *et al*, unpublished). This

integrated list resulted in 3154 genes, 72 of which were found to be DE (46 up and 26 down) (Supplementary Table 4).

FIGURE 2 near here

RESULTS

In Figure 2, overall results (from all data sets) are shown. We found that, merging the results from all the main GO categories (far left Venn diagram in Figure 2), most of the enriched terms (462) were shared by the different BRs. Chip reference results (BR-II) were completely contained into BR-I in each GO DAG, except for one enriched term in CC (far right Venn diagram in Figure 2). This term is found in the inner structure of the DAG; thus, it does not add new biological information (it could be explained by more terms further down the branch i.e. specific enriched leaves). The genome background (BR-I) produced much more enriched terms (125) than the other references. However even when BR-III only holds 43.4% of the genome genes (BR-I) (see Table 1), it provided 46 new enriched terms not recognized by any of the other two references. From them, an overall of 39% represented new remarkable biological information (see below sections for details).

TABLE 1 near here

FIGURE 3 near here

In Figure 3 the general DAG overlapping method is presented (in this case showing MF results from Packer dataset). For each main GO category the same DAG structure over each BR result was obtained and overlapped by the MCRM. Then a unique color-coded DAG is displayed for all the references. Enriched nodes were highlighted by different colors according to the overlap (Figure 3, Bottom DAG). The MCRM summarizes consensus nodes (in red) and highlights less consensus nodes/branches in orange (enriched in BR-I) and yellow (BR-I & BR-II). We found that by using the “user defined” reference (BR-III), new biological relevant branches terms emerged (see following sections). They were easily visualized by displaying them in green by the MRCM. These nodes were not identified by any of the other references.

TABLE 2 near here

Human smoke dataset

When analyzing MF ontology, we found that most enriched terms were consensus terms (see Table 2 and Supplementary Figure 1). The MRCM highlighted the nCN *electron carrier activity,* identified with BR-I and BR-II. This node was identified in the original work by Spira et al. [21]. The other nCN was *calcium ion bindin*, only enriched with BR-III. We found that genes in this term were reported to be involved in DNA damage mechanism, tumor cell migration and wound healing [25-27]. All these processes were also related to smoke by these authors.

In the BP DAG 55 nodes were found enriched. These nodes mainly belonged to metabolic, response-to-stimulus and cellular processes. Among them 27 were CNs. The MRCM highlighted a nCN branch (far right branch in Supplementary Figure 2), only enriched in both BR-I and BR-III, related to *angiogenesis*, a well-documented process related to smoke-derived injury [28]. In this branch, we also identified a new term (only detected by BR-III) named *skeletal system morphogenesis* which was found to adversely affect development when cigarette smoke interacts with cells [29].

Another new BP node (only enriched in BR-III) was identified by the MRCM. This node, *cell adhesion*, was reported to be induced by cigarette smoke extracts [30]. In addition, another 9 nCNs were highlighted by the MRCM, which were also literature validated (see Supplementary Table 5).

The CC cellular component DAG of this experiment has only 2 CNs from the 12 enriched ones. The MRCM identified 3 new nCNs (only detected by BR-III). One of them was the *proteinaceous extracellular matrix* node, whose genes were found to participate in the earliest stages of lung cancer development [31]. On the contrary, the enriched nodes found by BR-II (chip reference) were only CNs. This means that all other informative terms found by the genome reference or by the user defined reference with literature support would have been missed if only this reference had been used (see Supplementary Figure 3 and Table 6).

Mouse smoke dataset

The MCRM showed a major consensus over the main GO categories. All enriched terms were identified using BR-I (see Table 2).

However, in MF DAG four enriched terms were highlighted exclusively by the genome reference (Supplementary Figure 4) and three of them were found to be relevant such as *protein*, *RNA*

and *zinc ion* binding associated with smoke-related embryo deformity development during pregnancy [32, 21 and 11]. These terms would be missing if only chip reference analysis (BR-II) had been carried out.

In BP DAG eight enriched nodes were highlighted. Three of them were non consensus leaves (nCL) (Supplementary Figure 5). Two of these nCLs (*nutrient and cellular response to stimulus*) were related to oxidative stress induced by smoke in mice [33-34]. The third nCL, *cellular catabolic process,* proved to induce connective tissue breakdown by cigarette smoke by Dhami et al. [35] was automatically highlighted by MCRM.

In the CC GO category two terminal branches that ended in CNs, contained all the biological relevant information: *phosphoinositide 3-kinase complex*[36] and *PML body* which is related to viral infection, consistent with the McGrath-Morrow hypothesis (Supplementary Figure 6).

Melanoma genomic dataset

The analysis of the MF DAG showed 35 enriched terms distributed as shown in Table 2, where 16 of them represented CNs (Figure 3 and Supplementary Figure 7). In this case, the MRCM allowed us to identify three new enriched branches (only with BR-III) directly related to the experimental setting. The far left branch (Figure 3, A) ends in *transmembrane receptor activity* node, which holds many genes reported in the original work related to cell-surface receptor-mediated transduction pathways [23]. The *calcium ion binding* node was highlighted in the new central branch (Figure 3, B). This node was found to be a potential target for malignant melanoma therapy [37]. The last new branch (Figure 3, C) ended in the *carboxylic acid transmembrane transporter activity* node which contains genes of the SCL16 family. This family was reported fundamental for metabolism and pH regulation by Halestrap & Meredith [38] but not directly associated to melanoma.

The MRCM also highlighted two branches and one nCL present in BR-I and/or II. These relevant biological terms (*ATP binding*, *protein serine/threonine kinase activity* and *nucleoside-triphosphatase regulator activity*) were related to different p53 processes by different authors in melanoma studies [39-41].

In BP DAG 147 enriched terms were found, where 91 of these were consensus terms (Supplementary Figure 8). This dataset showed the greatest number of nCLs among the tested

gene expression datasets (see Table 2). By means of the MRCM, seven new terms (BR-III) relevant to the experiment emerged at different levels of the DAG. These were mainly related to three aspects: development, immunity and hemostasis. Relevant nodes in the first group *(cartilage development*, *organ morphogenesis* and *lipid transport)* were associated with different target genes of the p53 pathway [42-44]). Immunity related enriched nodes (*humoral immune response* and *positive regulation of alpha-beta T cell differentiation*) were also related to p53 activity where it could also affect T cell differentiation with other Wnt pathway genes [45-[46]. Validation of the last nodes group (*hemostasis* and *regulation of coagulation*) suggested that tissue factor protein plays an important role in blood coagulation and also at cytoplasm levels, where it is able to transduce a melanoma cell signal that promotes metastasis [47-48].

The remaining highlighted (by BR-I and BR-II) enriched nCLs in this DAG, suggested general biological aspects and others that had already been reported in the original work (*MAPKKK cascade*, *regulation of MAP kinase activity* and *regulation of Ras GTPase activity*). Our proposed method produces an easier visualization and identification of these terms than using only one reference.

In the CC DAG, 15 consensus terms were identified (Supplementary Figure 9). The MRCM could easily highlight membrane cellular parts related to the experimental setting. New enriched terms (BR-III) like *external side of plasma membrane* and *integral to membrane of membrane fraction* held genes related to tumor cell adhesion, invasion and metastatic activity [49-50].

Melanoma proteomic dataset

Proteomic studies, unlike microarray experiments, do not have a “chip” reference. The only two possible available references were the genome and user defined (BR-I and III, respectively). In this case, although BR-III only holds at most 17% of the BR-I, a major consensus on ontology was achieved as in the mouse dataset case.

Thirty nine enriched nodes present in MF DAG were found to be CNs. By means of the MRCM we were able to identify 15 terms (only present in the genome reference, BR-I) where 9 of them were leaves (Supplementary Figure 10). From these highlighted nodes, the *unfolded protein binding* node was found to be related to SPARC during embryonic development, when collagen IV deposition in basal lamina was studied, and when mentioned as a collagen molecular

chaperone in the endoplasmic reticulum [51-52]. The MRCM also highlighted a *voltage-gated chloride channel activity* nCN branch, where two CLIC family genes showed upregulated activity. These genes were found to be related to melanoma cell migration by Madeja et al. [53-54].

In this dataset, the highest number of enriched terms (228) and consensus (170) among all datasets was found for BP DAG (see Table 2 and Supplementary Figure 11). The MCRM highlighted 10 nodes (only in BR-I) related to SPARC. For instance *intermediate filament cytoskeleton organization* term is strongly affected by SPARC [55]. Likewise, *positive regulation of leukocyte migration and chemotaxis* are directly affected by SPARC expression as reported by Alvarez et al. [56] and Kelly et al. [57]. Response terms such as *cellular stress, axon injury* and *steroid hormone* were also highlighted and associated to this gene [58-63]. Matricellular proteins like SPARC are also involved in the highlighted terms *nervous system development* and *cell differentiation* [64-66]. Interestingly, by means of the MRCM we found highlighted terms in both MF and BP DAG where the CLIC family genes were present. A relation between SPARC and CLIC family on enriched *negative regulation of protein ubiquitination* node was recently suggested by Nakayama [67] and Bellei *et al.* [68].

One nCN, *amine transport*, was found to be highlighted in BP DAG by the MRCM and related only to BR-III. We have found no conclusive evidence of this process related to SPARC.

In CC DAG a consensus of 54 enriched terms was reached (Supplementary Figure 12). The MRCM highlighted (only in BR-I) three additional leaves, *basement membrane, ubiquitin conjugating enzyme complex* and *nuclear envelope* associated to SPARC expression by Fukunaga-Kalabis et al.[69], Anwar et al. [70] and Sacks-Wilner & Freddo [71].

DISCUSSIONS

Here we show that SEA results vary according to the used background gene reference list, potentially biasing or misleading the biological interpretation. In our tested datasets, a great consensus was reached regardless of the used BR, in agreement with Hedegaard [72], who suggested that if biological results (i.e. the differential genes) are reliable, results among reference should be comparable to some extent. Nevertheless, informative ontology terms

could be missed depending on the BR or the visualization scheme used (e.g. tabular format) making the discovery process of biologically relevant information difficult.

The contrast of several background reference gene lists, and in particular the inclusion of a specific user-defined (BR-III), showed that more relevant biological and experiment-specific information is made available. By using the user-defined BR we were able to find previously unseen enriched terms.

We also propose a color–code scheme to visualize ontology results. The color-code display facilitates the identification of informative biological terms, yielding a rapid overview of the experiment's results. The method automatically highlights nodes or DAG branches which, in our case, suggested being relevant to the experimental context. The proposed strategy facilitates the DAG inspection, avoiding looking into great detail and saving time in the analysis. Our results suggest that consensus nodes provide a global overview of the experiment and provide information about the expressed gene list reliability (they appear enriched no matter what BR is used). On the contrary, we found that non-consensus nodes provided interesting information, strongly related to the experimental setting and with published literature supporting their biological relevance. In some datasets, non-consensus terms also highlighted full new enriched branches of highly representative terms which were unseen or blurred when using the single reference strategy (see Supplementary Figures).

Unlike other tools, our method includes all the *a priori* (without trimming GO DAGs) and *a posteriori* (no further filtering criteria) information, in order to let the DAGs and the MCRM speak for themselves together. Our results (supported by literature validation) suggest that information is gained using DAVID and GO results without any constraints (no adjustments, no filtering). In this context the proposed MRCM for information retrieval (simultaneous data bases query) and visualization scheme, as a data-mining tool, helps us to easily visualize contextual information by means of highlighting potentially relevant nodes in the DAG (identifying novel information on the DAG picture). For instance, the *voltage-gated chloride channel activity* enriched branch, which proved to be highly significant in the melanoma proteomic experiment, would not be identified in the default DAVID strategy (it excludes terms with less than 3 genes by default). This is especially important in proteomic studies where, as in our case, a term holding two

genes (CLIC4 and CLIC1) was found to be enriched. These genes were represented by several differentially expressed isoforms in the gels of the melanoma proteomic dataset.

Conflict of interest: None declared

Funding: This work was supported by the National Agency for Promoting Science and Technology, Argentina [PICT00667/07 to A.S.L.], National Council of Science and Technology of Argentina [PIP2009 to M.B.], and Córdoba Ministry of Science and Technology, Argentina [PID2008 to E.A.F].

References

[1] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin,G. Sherlock, Gene Ontology: tool for the unification of biology, Nat. Genet., 25, 25-29 (2000)

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

[3] D. Wei Huang, B. T. Sherman, R. A. Lempicki, Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists, Nucleic Acids Res., 37, 1-13 (2009)

[4] 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, 401-407 (2007)

[5] G. Dennis, B. T. Sherman, D. A. Hosack, J. Yang, W. Gao, H. C. Lane, R. A. Lempicki, DAVID: database for annotation, visualization, and integrated discovery, Genome Biol., 4, P3 (2003)

[6] D. W. Huang, B. T. Sherman, Q. Tan, J. Kir, D. Liu, D. Bryant, Y. Guo, R. Stephens, M. W. Baseler, H. C. Lane, R. A. Lempicki, DAVID Bioinformatics Resources: expanded annotation database and novel algorithms to better extract biology from large gene lists. Nucleic Acids Res., 35, W169-W175 (2007)

[7] B. R. Zeeberg, W. Feng, G. Wang, M. D. Wang, A. T. Fojo, M. Sunshine, S. Narasimhan, D. W. Kane, W. C. Reinhold, S. Lababidi, K. J. Bussey, J. Riss, J. C. Barrett, J. N. Weinstein, GoMiner: a resource for biological interpretation of genomic and proteomic data, Genome Biol., 4, R28 (2003)

[8] F. Al-Shahrour, R. Díaz-Uriarte, J. Dopazo, FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes, Bioinformatics, 20,578-580 (2004)

[9] Affymetrix,Inc., GeneChip® Expression Analysis, Data Analysis Fundamentals, Part No. 701190 Rev. 4 (2004)

[10] K. J. Archer, S.E. Reese, Detection call algorithms for high-throughput gene expression microarray data, Brief Bioinform., 11, 244-252 (2010)

[11] A. J. Hackstadt, A. M. Hess, Filtering for increased power for microarray data analysis, BMC Bioinformatics, 10, 11 (2009)

[12] J. N. McClintick, H. J. Edenberg, Effects of filtering by Present call on analysis of microarray experiments, BMC Bioinformatics, 7, 49 (2006)

[13] R. Bourgon, R. Gentleman, W. Huber, Independent filtering increases detection power for high-throughput experiments, Proc. Natl. Acad. Sci. U. S. A., 107, 9546-9551 (2010)

[14] R Development Core Team, R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0 (2009).

[15] R. C. Gentleman, V. J. Carey, D. M. Bates, B. Bolstad, M. Dettling, S. Dudoit, B. Ellis, L. Gautier, Y. Ge, J. Gentry, K. Hornik, T. Hothorn, W. Huber, S. Iacus, R. Irizarry, F. Leisch, C. Li, M. Maechler, A. J. Rossini, G. Sawitzki, C. Smith, G. Smyth, L. Tierney, J. Y. Yang, J. Zhang, Bioconductor: Open software development for computational biology and bioinformatics, Genome Biol., 5, R80 (2004)

[16] L .Gautier, L. Cope, B. M. Bolstad, R. A. Irizarry, affy--analysis of Affymetrix GeneChip data at the probe level, Bioinformatics, 20, 307-315 (2004)

[17] G. K. Smyth, Linear models and empirical bayes methods for assessing differential expression in microarray experiments, Stat Appl Genet Mol Biol, 3, Article 3 (2004)

[18] E. A. Fernández, M. R. Girotti, J. A. López del Olmo, A. S. Llera, O. L. Podhajcer, R. J. Cantet, M. Balzarini, Improving 2D-DIGE protein expression analysis by two-stage linear mixed models: assessing experimental effects in a melanoma cell study, Bioinformatics, 24, 2706-2712 (2008)

[19] D. W. Huang, B. T. Sherman, R. A. Lempicki, Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources, Nat. Protoc., 4, 44-57 (2009)

[20] S. Falcon, R. Gentleman, Using GOstats to test gene lists for GO term association, Bioinformatics, 23, 257-258 (2007)

[21] A. Spira, J. Beane, V. Shah, G. Liu, F. Schembri, X. Yang, J. Palma, J. S. Brody, Effects of cigarette smoke on the human airway epithelial cell transcriptome. Proc. Natl. Acad. Sci. U. S. A., 101, 10143-10148 (2004)

[22] S. McGrath-Morrow, T. Rangasamy, C. Cho, T. Sussan, E. Neptune, R. Wise, R. M. Tuder, S. Biswal, Impaired lung homeostasis in neonatal mice exposed to cigarette smoke. Am. J. Respir. Cell. Mol. Biol., 38, 393-400 (2008)

[23] L. M. Packer, S. J. Pavey, G. M. Boyle, M. S. Stark, A. L. Ayub, H. Rizos, N. K. Hayward, Gene expression profiling in melanoma identifies novel downstream effectors of p14ARF, Int. J. Cancer, 121, 784-790 (2007)

[24] M. S. Sosa, M. R. Girotti, E. Salvatierra, F. Prada, J. A. de Olmo, S. J. Gallango, J. P. Albar, O. L. Podhajcer, A. S. Llera, Proteomics analysis identified N-cadherin, clusterin and HSP27 as mediators of SPARC activity in melanoma cells, Proteomics, 22, 4123–4134 (2007)

[25] P. Leanderson, C. Tagesson, Cigarette smoke-induced DNA damage in cultured human lung cells: role of hydroxyl radicals and endonuclease activation, Chem. Biol. Interact., 81, 197-208 (1992)

[26] J. Austermann, A. R. Nazmi, C. Müller-Tidow, V. Gerke, Characterization of the Ca2+ -regulated ezrin-S100P interaction and its role in tumor cell migration, J. Biol. Chem., 283, 29331-29340 (2008)

[27] R. Parsanejad, W. R. Fields, W. T. Morgan, B. R. Bombick, D. J. Doolittle, The time course of expression of genes involved in specific pathways in normal human bronchial epithelial cells following exposure to cigarette smoke, Exp. Lung Res., 34, 513-530 (2008)

[28] R. D. Egleton, K. C. Brown, P. Dasgupta, Angiogenic activity of nicotinic acetylcholine receptors: implications in tobacco-related vascular diseases, Pharmacol. Ther., 121, 205-223 (2009)

[29] J. D. Holz, Environmental agents affect skeletal growth and development, Birth Defects Res. C. Embryo. Today, 81, 41-50 (2007)

[30] H. W. Chen, C. K. Lii, H. J. Ku, T. S. Wang, Cigarette smoke extract induces expression of cell adhesion molecules in HUVEC via actin filament reorganization, Environ. Mol. Mutagen, 50, 96-104 (2009)

[31] J. J. Oh, E. O. Taschereau, A. K. Koegel, C. L. Ginther, J. K. Rotow, K. Z. Isfahani, D. J. Slamon, RBM5/H37 tumor suppressor, located at the lung cancer hot spot 3p21.3, alters expression of genes involved in metastasis, Lung Cancer, 70, 253-262 (2010)

[32] K. H. Horn, E. R. Esposito, R. M. Greene, M. M . Pisano, The effect of cigarette smoke exposure on developing folate binding protein-2 null mice, Reprod. Toxicol., 26, 203-209 (2008)

[33] S. S. Valenca, F. Silva Bezerra, A. A. Lopes, B. Romana-Souza, M. C. Marinho Cavalcante, A. B. Lima, V. L. Gonçalves Koatz, L. C. Porto, Oxidative stress in mouse plasma and lungs induced by cigarette smoke and lipopolysaccharide, Environ. Res., 108, 199-204 (2008)

[34] J. Turgeon, S. Dussault, P. Haddad, J. Groleau, C. Ménard, S. E. Michaud, F. Maingrette, A. Rivard, Probucol and antioxidant vitamins rescue ischemia-induced neovascularization in mice exposed to cigarette smoke: potential role of endothelial progenitor cells, Atherosclerosis, 208, 342-349 (2010)

[35] R. Dhami, B. Gilks, C. Xie, K. Zay, J. L. Wright, A. Churg, Acute cigarette smoke-induced connective tissue breakdown is mediated by neutrophils and prevented by alpha1-antitrypsin, Am. J. Respir. Cell. Mol. Biol., 22, 244-252 (2000)

[36] P. J. Barnes, Histone deacetylase-2 and airway disease, Ther. Adv. Respir. Dis., 3, 235-243 (2009)

[37] T. H. Charpentier, L. E. Thompson, M. A. Liriano, K. M. Varney, P. T. Wilder, E. Pozharski, E. A. Toth, D. J. Weber, The effects of CapZ peptide (TRTK-12) binding to S100B-Ca2+ as examined by NMR and X-ray crystallography, J. Mol. Biol., 396, 1227-1243 (2010)

[38] A. P. Halestrap, D. Meredith, The SLC16 gene family-from monocarboxylate transporters (MCTs) to aromatic amino acid transporters and beyond, Pflugers Arch., 447, 619-628 (2004)

[39] D. Walerych, M. Gutkowska, M. P. Klejman, B. Wawrzynow, Z. Tracz, M. Wiech, M. Zylicz, A. Zylicz, ATP binding to Hsp90 is sufficient for effective chaperoning of p53 protein, J. Biol. Chem., 285, 32020-32028 (2010)

[40] J. H. Chung, F. Bunz, Cdk2 is required for p53-independent G2/M checkpoint control, PLoS Genet., 26, e1000863 (2010)

[41] C. Miled, M. Pontoglio, S. Garbay, M. Yaniv, J. B. Weitzman, A genomic map of p53 binding sites identifies novel p53 targets involved in an apoptotic network, Cancer Res., 65, 5096-5104 (2005)

[42] J. C. Lougheed, J. M. Holton, T. Alber, J. F. Bazan, T. M. Handel, Structure of melanoma inhibitory activity protein, a member of a recently identified family of secreted proteins, Proc. Natl. Acad. Sci. U.S.A., 98, 5515-5520 (2001)

[43] M. Valente, F. Calabrese, Liver and apoptosis, Ital. J. Gastroenterol Hepatol., 31, 73-77 (1999)

[44] Y. Sasaki, H. Negishi, R. Koyama, N. Anbo, K. Ohori, M. Idogawa, H. Mita, M. Toyota, K. Imai, Y. Shinomura, T. Tokino, p53 family members regulate the expression of the apolipoprotein D gene, J. Biol. Chem., 284, 872-883 (2009)

[45] Y. Ichiki, M. Takenoyama, M. Mizukami, T. So, M. Sugaya, M. Yasuda, T. So, T. Hanagiri, K. Sugio, K. Yasumoto, Simultaneous cellular and humoral immune response against mutated p53 in a patient with lung cancer, J. Immunol., 172, 4844-4850 (2004)

[46] K. Okazuka, Y. Wakabayashi, M. Kashihara, J. Inoue, T. Sato, M. Yokoyama, S. Aizawa, Y. Aizawa, Y. Mishima, R. Kominami, p53 prevents maturation of T cell development to the immature CD4-CD8+ stage in Bcl11b-/- mice, Biochem. Biophys. Res. Commun., 328, 545-549 (2005)

[47] F. A. Siddiqui, A. Amirkhosravi, M. Amaya, H. Desai, T. Meyer, J. L. Francis, Purification and properties of human melanoma cell tissue factor, Clin. Appl. Thromb. Hemost., 7, 289-295 (2001)

[48] M. E. Bromberg, W. H. Konigsberg, J. F. Madison, A. Pawashe, A. Garen, Tissue factor promotes melanoma metastasis by a pathway independent of blood coagulation, Proc. Natl. Acad. Sci. U. S. A., 92, 8205-8209 (1995)

[49] J. Iida, D. Pei, T. Kang, M. A. Simpson, M. Herlyn, L. T. Furcht, J. B. McCarthy, Melanoma chondroitin sulfate proteoglycan regulates matrix metalloproteinase-dependent human melanoma invasion into type I collagen, J. Biol. Chem., 276, 18786-18794 (2001)

[50] F. Felicetti, I. Parolini, L. Bottero, K. Fecchi, M. C. Errico, C. Raggi, M. Biffoni, F. Spadaro, M. P. Lisanti, M. Sargiacomo, A. Carè, Caveolin-1 tumor-promoting role in human melanoma, Int. J. Cancer, 125, 1514-1522 (2009)

[51] N. Martinek, J. Shahab, J. Sodek, M. Ringuette, Is SPARC an evolutionarily conserved collagen chaperone?, J. Dent. Res., 86, 296-305 (2007)

[52] M. Pfaff, M. Aumailley, U. Specks, J. Knolle, H. G. Zerwes, R. Timpl, Integrin and Arg-Gly-Asp dependence of cell adhesion to the native and unfolded triple helix of collagen type VI, Exp Cell Res, 206,167-176 (1993)

[53] Z. Madeja, I. Szymkiewicz, A. Zaczek, J. Sroka, K. Miekus, W. Korohoda, Contact-activated migration of melanoma B16 and sarcoma XC cells. Biochem Cell Biol., 79, 425-440 (2001)

[54] Z. Madeja, A. Master, M. Michalik, J. Sroka, Contact-mediated acceleration of migration of melanoma B16 cells depends on extracellular calcium ions, Folia Biol (Krakow), 49, 113-124 (2001)

[55] M. J. Alvarez, Rol de la proteína de matriz extracelular SPARC en la progresión tumoral del melanoma humano. Ph.D. Thesis, Universidad de Buenos Aires - Instituto de Investigaciones Bioquímica: Argentina (2006).

[56] M. J. Alvarez, F. Prada, E. Salvatierra, A. I. Bravo, V. P. Lutzky, C. Carbone, F. J. Pitossi, H. E. Chuluyan, O. L. Podhajcer, Secreted protein acidic and rich in cysteine produced by human melanoma cells modulates polymorphonuclear leukocyte recruitment and antitumor cytotoxic capacity, Cancer Res., 65, 5123-5132 (2005)

[57] K. A. Kelly, J. R. Allport, A. M. Yu, S. Sinh, E. H. Sage, R. E. Gerszten, R. Weissleder, SPARC is a VCAM-1 counter-ligand that mediates leukocyte transmigration, J. Leukoc. Biol., 81, 748-756 (2007)

[58] S. V. Vadlamuri, SPARC affects glioma cell growth differently when grown on brain ECM proteins in vitro under standard versus reduced-serum stress conditions, Neuro. Oncol., 5, 244-254 (2003)

[59] M. W. Schellings, Y. M. Pinto, S. Heymans, Matricellular proteins in the heart: possible role during stress and remodeling, Cardiovasc. Res., 64, 24-31 (2004)

[60] C. Luna, G. Li, J. Qiu, D. L. Epstein, P. Gonzalez, Role of miR-29b on the regulation of the extracellular matrix in human trabecular meshwork cells under chronic oxidative stress, Mol. Vis., 15, 2488-2497 (2009)

[61] E. Au, M. W. Richter, A. J. Vincent, W. Tetzlaff, R. Aebersold, E. H. Sage, A. J. Roskams, SPARC from olfactory ensheathing cells stimulates Schwann cells to promote neurite outgrowth and enhances spinal cord repair, J Neurosci., 27, 7208-7221 (2007)

[62] S. C. Dieudonné, J. M. Kerr, T. Xu, B. Sommer, A. R. DeRubeis, S. A. Kuznetsov, I. S. Kim, P. Gehron Robey, M. F. Young, Differential display of human marrow stromal cells reveals unique mRNA expression patterns in response to dexamethasone, J. Cell. Biochem., 76, 231-243 (1999)

[63] R. S. Sawhney, Expression and regulation of SPARC, fibronectin, and collagen IV by dexamethasone in lens epithelial cells, Cell Biol. Int., 26, 971-983 (2002)

[64] C. Chavey, J. Boucher, M. N. Monthouël-Kartmann, E. H. Sage, I. Castan-Laurell, P. Valet, S. Tartare-Deckert, E. Van Obberghen, Regulation of secreted protein acidic and rich in cysteine during adipose conversion and adipose tissue hyperplasia, Obesity (Silver Spring), 14, 1890-1897 (2006)

[65] A. J. Vincent, P. W. Lau, A. J. Roskams, SPARC is expressed by macroglia and microglia in the developing and mature nervous system, Dev Dyn., 237, 1449-1462 (2008)

[66] C. Eroglu, The role of astrocyte-secreted matricellular proteins in central nervous system development and function, J. Cell Commun. Signal, 3, 167–176 (2009)

[67] K . Nakayama, Growth and progression of melanoma and non-melanoma skin cancers regulated by ubiquitination, Pigment. Cell Melanoma Res., 23, 338-351 (2010)

[68] B. Bellei, V. Maresca, E. Flori, A. Pitisci, L. Larue, M. Picardo, p38 regulates pigmentation via proteasomal degradation of tyrosinase, J Biol Chem., 285, 7288-7299 (2010)

[69] M. Fukunaga-Kalabis, A. Santiago-Walker, M. Herlyn, Matricellular proteins produced by melanocytes and melanomas: in search for functions, Cancer Microenviron, 1, 93-102 (2008)

[70] A. Anwar, D. A. Norris, M. Fujita, Ubiquitin proteasomal pathway mediated degradation of p53 in melanoma, Arch. Biochem. Biophys., Article in Press (2010)

[71] R. Sacks-Wilner, T. F. Freddo, Differences in nuclear pore density among human choroidal melanoma cell types, Ultrastruct. Pathol., 14, 311-9 (1990)

[72] J. Hedegaard, C. Arce, S. Bicciato, A. Bonnet, B. Buitenhuis, M. Collado-Romero, L. N. Conley, M. Sancristobal, F. Ferrari, J. J. Garrido, M. A. Groenen, H. Hornshøj, I. Hulsegge, L. Jiang, A. Jiménez-Marín, A. Kommadath, S. Lagarrigue, J. A. Leunissen, L. Liaubet, P. B. Neerincx, H. Nie, J. van der Poel, D. Prickett, M. Ramirez-Boo, J. M. Rebel, C. Robert-Granié, A. Skarman, M. A. Smits, P. Sørensen, G. Tosser-Klopp, M. Watson, Methods for interpreting lists of affected genes obtained in a DNA microarray experiment, BMC Proc., 3, S5 (2009)

FIGURE CAPTIONS

Figure 1: Schematic multi-reference contrast method flowchart.

Figure 2: Venn diagrams showing the distribution of the enriched terms found in all datasets for the different GO DAGs. Enriched terms in each subset are displayed in numbers (black). Corresponding nodes that fell onto leaves (most informative terms) are presented in percentages (gray).

Figure 3: Contrast schematic output for Packer dataset MF DAG. Enriched nodes are displayed in color for each BR used (top graphs). Combined results are summarized in a single graph according to Venn diagram color legend. MRCM highlights the central branches by BR-I and II, while A, B and C emerge only with BR-III.

Supplementary Figure 1: GO MF DAG for the human smoke dataset [21]. Node color legend is displayed on the Venn diagram for the three BR sets (see text). White nodes indicate no enrichment by any of the used BRs.

Supplementary Figure 2: GO BP DAG for the human smoke dataset [21]. Node color legend is displayed on the Venn diagram for the three BR sets (see text). White nodes indicate no enrichment by any of the used BRs.

Supplementary Figure 3: GO CC DAG for the human smoke dataset [21]. Node color legend is displayed on the Venn diagram for the three BR sets (see text). White nodes indicate no enrichment by any of the used BRs.

Supplementary Figure 4: GO MF DAG for the mouse smoke dataset [22]. Node color legend is displayed on the Venn diagram for the three BR sets (see text). White nodes indicate no enrichment by any of the used BRs.

Supplementary Figure 5: GO BP DAG for the mouse smoke dataset [22]. Node color legend is displayed on the Venn diagram for the three BR sets (see text). White nodes indicate no enrichment by any of the used BRs.

Supplementary Figure 6: GO CC DAG for the mouse smoke dataset [22]. Node color legend is displayed on the Venn diagram for the three BR sets (see text). White nodes indicate no enrichment by any of the used BRs.

Supplementary Figure 7: GO MF DAG for the melanoma genomic dataset [23]. Node color legend is displayed on the Venn diagram for the three BR sets (see text). White nodes indicate no enrichment by any of the used BRs.

Supplementary Figure 8: GO BP DAG for the melanoma genomic dataset [23]. Node color legend is displayed on the Venn diagram for the three BR sets (see text). White nodes indicate no enrichment by any of the used BRs.

Supplementary Figure 9: GO CC DAG for the melanoma genomic dataset [23]. Node color legend is displayed on the Venn diagram for the three BR sets (see text). White nodes indicate no enrichment by any of the used BRs.

Supplementary Figure 10: GO MF DAG for the melanoma proteomic dataset (Girotti et al. unpublished). Node color legend is displayed on the Venn diagram for the three BR sets (see text). White nodes indicate no enrichment by any of the used BRs.

Supplementary Figure 11: GO BP DAG for the melanoma proteomic dataset (Girotti et al. unpublished). Node color legend is displayed on the Venn diagram for the three BR sets (see text). White nodes indicate no enrichment by any of the used BRs.

Supplementary Figure 12: GO CC DAG for the melanoma proteomic dataset (Girotti et al. unpublished). Node color legend is displayed on the Venn diagram for the three BR sets (see text). White nodes indicate no enrichment by any of the used BRs.

Table 1: Gene population in each Gene Ontology main category according to the three background references used.

Example dataset	Molecular Function			Biological Process			Cellular Component		
	I	II	III	I	II	III	I	II	III
Human smoke	15143	10886	3212	14116	10391	3089	15908	11082	3299
	(*100*)	(*71.9*)	(*21.2*)	(*100*)	(*73.6*)	(*21.9*)	(*100*)	(*69.7*)	(*20.7*)
Mouse smoke	15404	12995	6549	14219	11944	6005	15855	13596	6888
	(*100*)	(*84.4*)	(*42.5*)	(*100*)	(*84.0*)	(*42.2*)	(*100*)	(*85.6*)	(*43.4*)
Melanoma genomic	15143	14128	6216	14116	13187	5798	15908	14741	6384
	(*100*)	(*93.3*)	(*41.0*)	(*100*)	(*93.4*)	(*41.1*)	(*100*)	(*92.7*)	(*40.1*)
Melanoma proteomic	15143	-	2561	14116	-	2381	15908	-	2583
	(*100*)	-	(*16.9*)	(*100*)	-	(*16.8*)	(*100*)	-	(*16.2*)

Total background reference (BR) population count for the different Gene Ontology categories and BR types (I genome, II chip or III user defined). In parenthesis, population percentage respects to BR-I members. Interestingly, BR-II is almost as complete as the genome (BR-I) while one would expect to have a tidier relationship as found in the melanoma genomic dataset. Filtering criteria on BR-III, has removed more than half of the total genome genes available in each Gene Ontology category.

Table 2

Table 2: Total enriched terms on Gene Ontology categories for the three background references used for the four example datasets

Example dataset	Node type	Molecular Function			Biological Process			Cellular Component		
		I	II	III	I	II	III	I	II	III
Human smoke	T	18	16	14	51	33	37	9	2	9
	nCL	1	1	1	10	3	3	4	0	3
Mouse smoke	T	22	18	18	24	17	16	10	6	2
	nCL	3	0	0	3	1	0	0	0	0
Melanoma genomic	T	26	24	25	127	114	116	38	33	24
	nCL	1	2	3	7	4	7	6	4	3
Melanoma proteomic	T	54	-	39	225	-	173	63	-	54
	nCL	9	-	0	23	-	1	3	-	0

T: total amount of enriched nodes for a given background reference (I genome, II chip or III user defined). nCL: non consensus nodes at the end of the branch (leaves), i.e. enriched nodes only detected by one or two BRs.

Figure 1

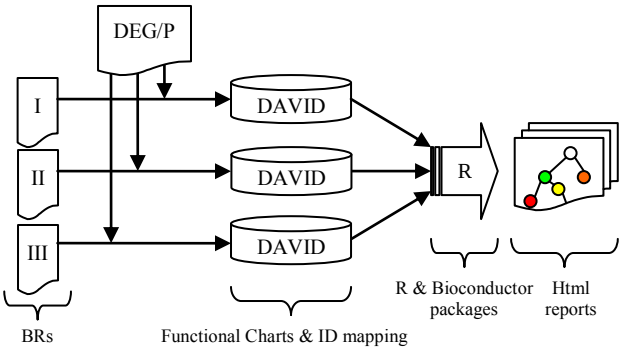

DEG/P
I
II
III
DAVID
DAVID
DAVID
R
R & Bioconductor packages
Html reports
BRs
Functional Charts & ID mapping

Figure 2

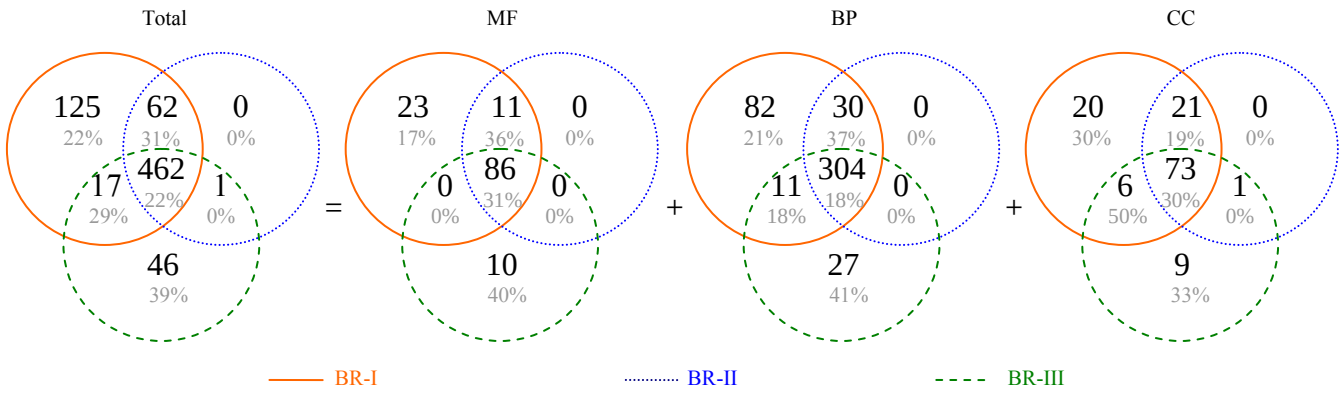
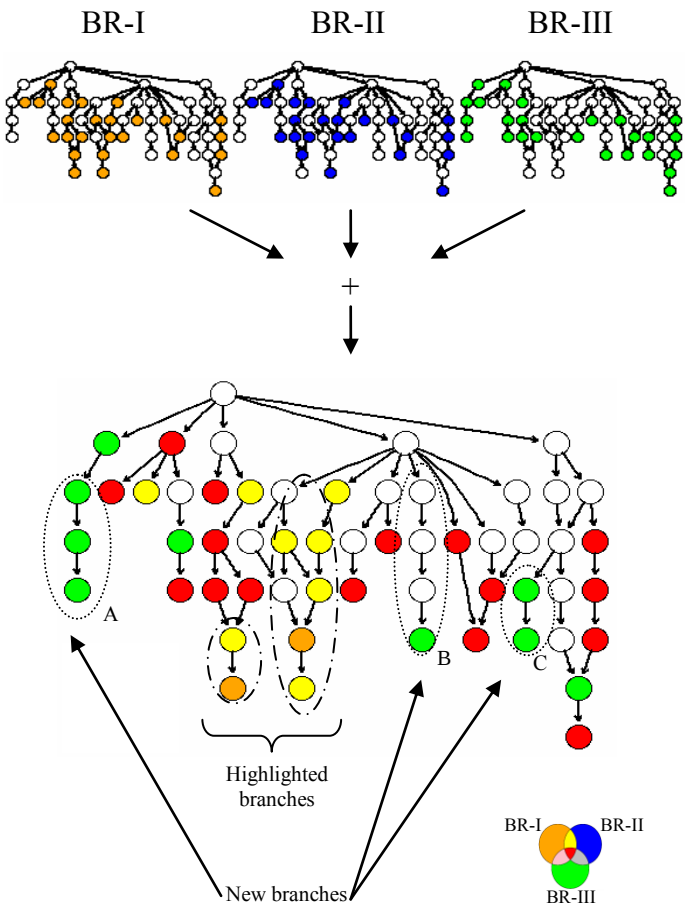

Total
125
22%
62
31%
0
0%
17
29%
462
22%
1
0%
46
39%
=
MF
23
17%
11
36%
0
0%
0
0%
86
31%
0
0%
10
40%
+
BP
82
21%
30
37%
0
0%
11
18%
304
18%
0
0%
27
41%
+
CC
20
30%
21
19%
0
0%
6
50%
73
30%
1
0%
9
33%
BR-I
BR-II
BR-III

Figure 3


BR-I
BR-II
BR-III
A
B
C
Highlighted branches
New branches
BR-I
BR-II
BR-III

Supplementary Table 1 to 4: microarray gene description for Spira et al.[21], McGrath-Morrow et al.[22], Packer et al.[23] and Girotti et al. (unpublished) respectively. The “Type” column descriptor stands for “Ref” (Reference), “Up” or “Down” gene expression according to the experimental setting described in the main article.

Supplementary Table 5: additional 9 biological process non-consensus nodes enriched by the Multi-Reference Contrast Method on Human smoke dataset.

Gene Ontology Biological Process term	Enrichment source	Literature reference
Cell redox homeostasis	BR-I & BR-II	Kaushik G et al., Cigarette smoke condensate promotes cell proliferation through disturbance in cellular redox homeostasis of transformed lung epithelial type-II cells. Cancer Lett., 270, 120-131 (2008).
Heterocycle metabolic process	BR-I	Cotgreave IA & Moldéus P, Lung protection by thiol-containing antioxidants., Bull. Eur. Physiopathol. Respir., 23, 275-277 (1987).
Positive regulation of cellular protein metabolic process	BR-I	Hu Q et al., The altertion and significance of surfactant protein A in rats chronically exposed to cigarette smoke. J Huazhong Univ Sci Technolog Med Sci., 28, 128-31 (2008).
Regulation of mitochondrial depolarization	BR-I & BR-II	Slebos DJ et al., Mitochondrial localization and function of heme oxygenase-1 in cigarette smoke-induced cell death, Am J Respir. Cell. Mol. Biol., 36, 409-417 (2006).
Regulation of synaptic plasticity	BR-I & BR-III	Qiao D et al., Oxidative mechanisms contributing to the developmental neurotoxicity of nicotine and chlorpyrifos. Toxicol. Appl. Pharmacol., 206, 17-26, (2205).
Response to insecticide	BR-I & BR-II	Bal R et al., Assessing the effects of the neonicotinoid insecticide imidacloprid in the cholinergic synapses of the stellate cells of the mouse cochlear nucleus using whole-cell patch-clamp recording. Neurotoxicology., 31, 113-120 (2010).
Response to	BR-I	Banerjee et al., Cellular and molecular mechanisms of

vitamin		cigarette smoke-induced lung damage and prevention by vitamin C. J. Inflamm. (Lond), 5, 21 (2008).
Vesicle targeting, to, from or within Golgi	BR-I	Mukhopadhyay S et al., Cytoplasmic provenance of STAT3 and PY-STAT3 in the endolysosomal compartments in pulmonary arterial endothelial and smooth muscle cells: implications in pulmonary arterial hypertension. Am. J. Physiol. Lung. Cell. Mol. Physiol., 294, L449-L468 (2008).
Vitamin metabolic process	BR-I	Bruno et al., Cigarette smoke alters human vitamin E requirements. J. Nutr., 135, 671-674 (2005).

BR-I, BR-II and BR-III correspond to genome, chip and user defined background references respectively.

Supplementary Table 6: additional five cellular component non-consensus nodes enriched by the Multi-Reference Contrast Method on Human smoke example dataset

Gene Ontology Cellular Component term	Enrichment Source	Literature reference
Cytosol	BR-I	Yasuda,S. et at., Cigarette smoke toxicants as substrates and inhibitors for human cytosolic SULTs, Toxicol. Appl. Pharmacol., 221, 13-20 (2007).
Extracellular space	BR-I & BR-III	Yin,L. et al., Alterations of extracellular matrix induced by tobacco smoke extract., Arch. Dermatol. Res. 292, 188-194 (2000).
Integral to plasma membrane	BR-I & BR-III	Rusznak,C. et al., Effect of cigarette smoke on the permeability and IL-1beta and sICAM-1 release from cultured human bronchial epithelial cells of never-smokers, smokers, and patients with chronic obstructive pulmonary disease, Am. J. Respir. Cell. Mol. Biol., 23, 530-536 (2000).
Proteinaceous extracellular matrix	BR-III	Oh,J.J. et al., RBM5/H37 tumor suppressor, located at the lung cancer hot spot 3p21.3, alters expression of genes involved in metastasis. Lung. Cancer, 70, 253-262 (2010).
Vacuole	BR-I	Cantin,A., Cellular response to cigarette smoke and oxidants: adapting to survive, Proc. Am. Thorac. Soc., 7, 368-375 (2010).

BR-I, BR-II and BR-III correspond to genome, chip and user defined background references respectively.

Supplementary Table 7: supplementary Gene Ontology direct acyclic graphs for each example dataset

Supplementary figure	Example dataset	Main GO category
1	Human Smoke, Spira et al. [21]	Molecular Function
2	Human Smoke, Spira et al. [21]	Biological Process
3	Human Smoke, Spira et al. [21]	Cellular Component
4	Mouse Smoke, McGrath-Morrow et al. [22]	Molecular Function
5	Mouse Smoke, McGrath-Morrow et al. [22]	Biological Process
6	Mouse Smoke, McGrath-Morrow et al. [22]	Cellular Component
7	Melanoma Genomic, Packer et al. [23]	Molecular Function
8	Melanoma Genomic, Packer et al. [23]	Biological Process
9	Melanoma Genomic, Packer et al. [23]	Cellular Component
10	Melanoma Proteomic, Girotti et al. unpublished	Molecular Function
11	Melanoma Proteomic, Girotti et al. unpublished	Biological Process
12	Melanoma Proteomic, Girotti et al. unpublished	Cellular Component

Node color legend is displayed on the Venn diagram for the three BR sets (see text). White nodes indicate no enrichment by any of the used background references.

MF

BR-I

BR-II

BR-III

catalytic activity

antioxidant activity

electron carrier activity

oxidoreductase activity

cofactor binding

enzyme inhibitor activity

oxidoreductase activity, acting on CH-OH group of donors

carbon-oxygen lyase activity

coenzyme binding

peptidase inhibitor activity

lipase inhibitor activity

oxidoreductase activity, acting on the CH-OH group of donors, NAD or NADP as acceptor

NADP or NADPH binding

phospholipase inhibitor activity

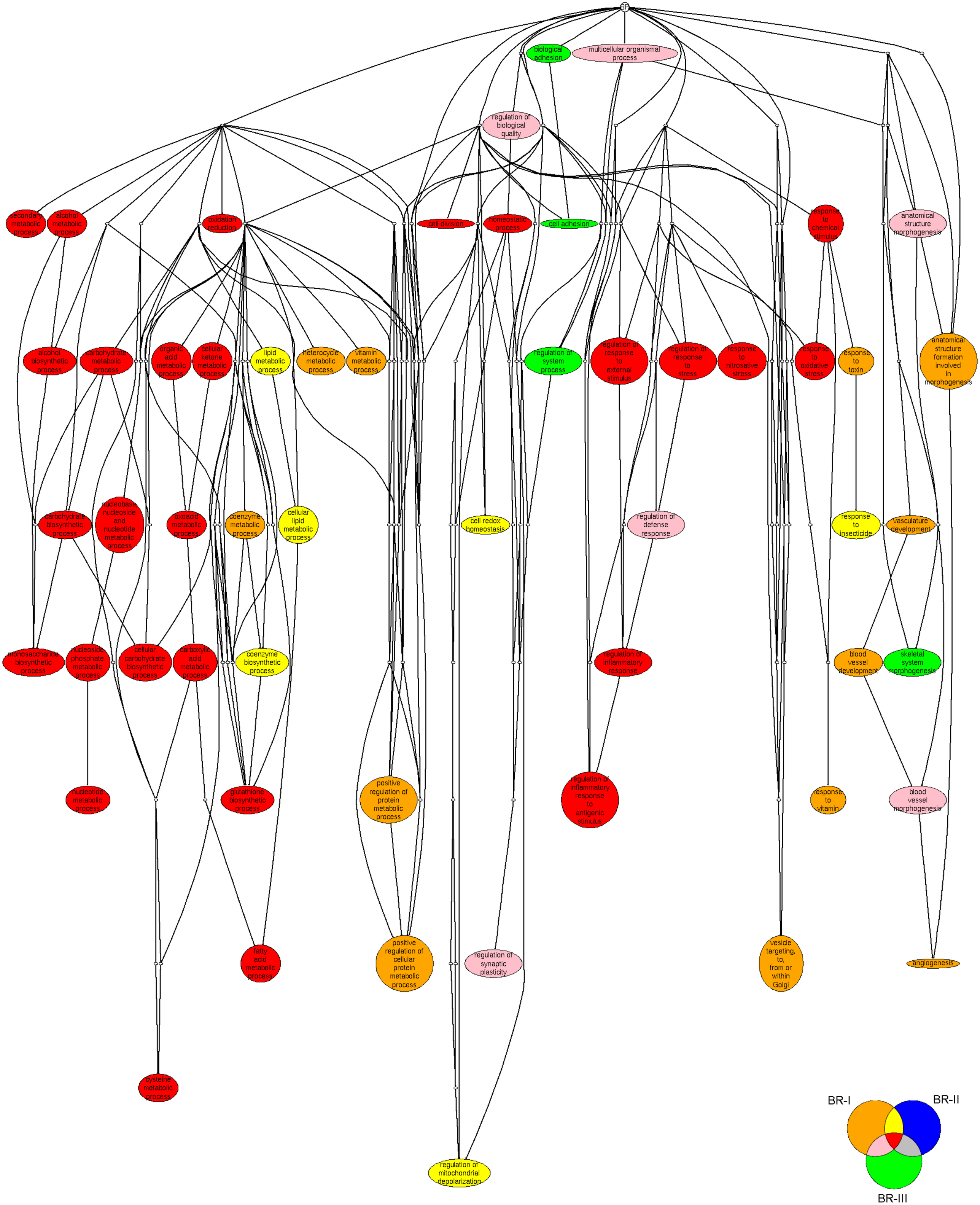
aldo-keto reductase activity

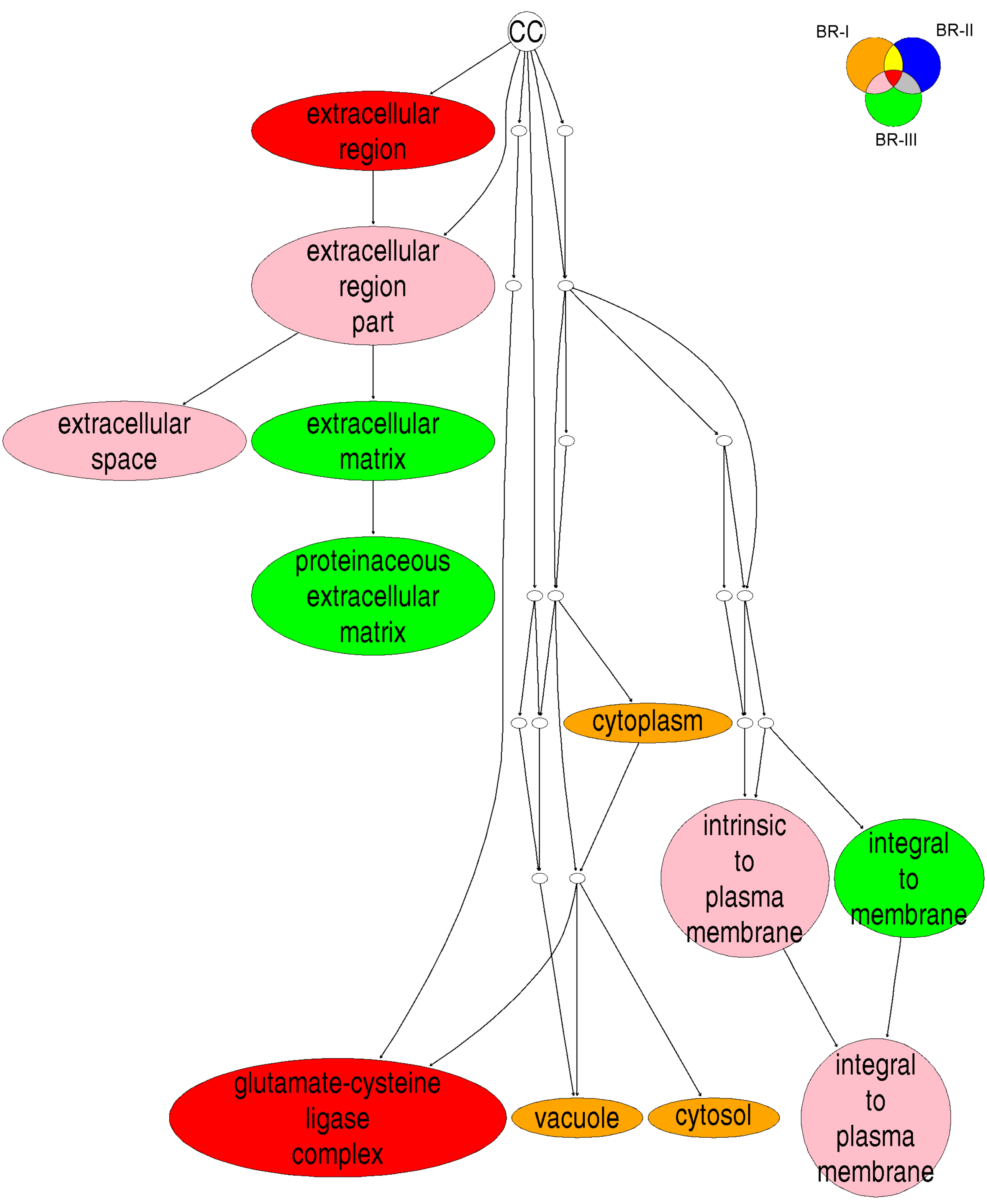
glutamate-cysteine ligase activity

calcium ion binding

phospholipase A2 inhibitor activity

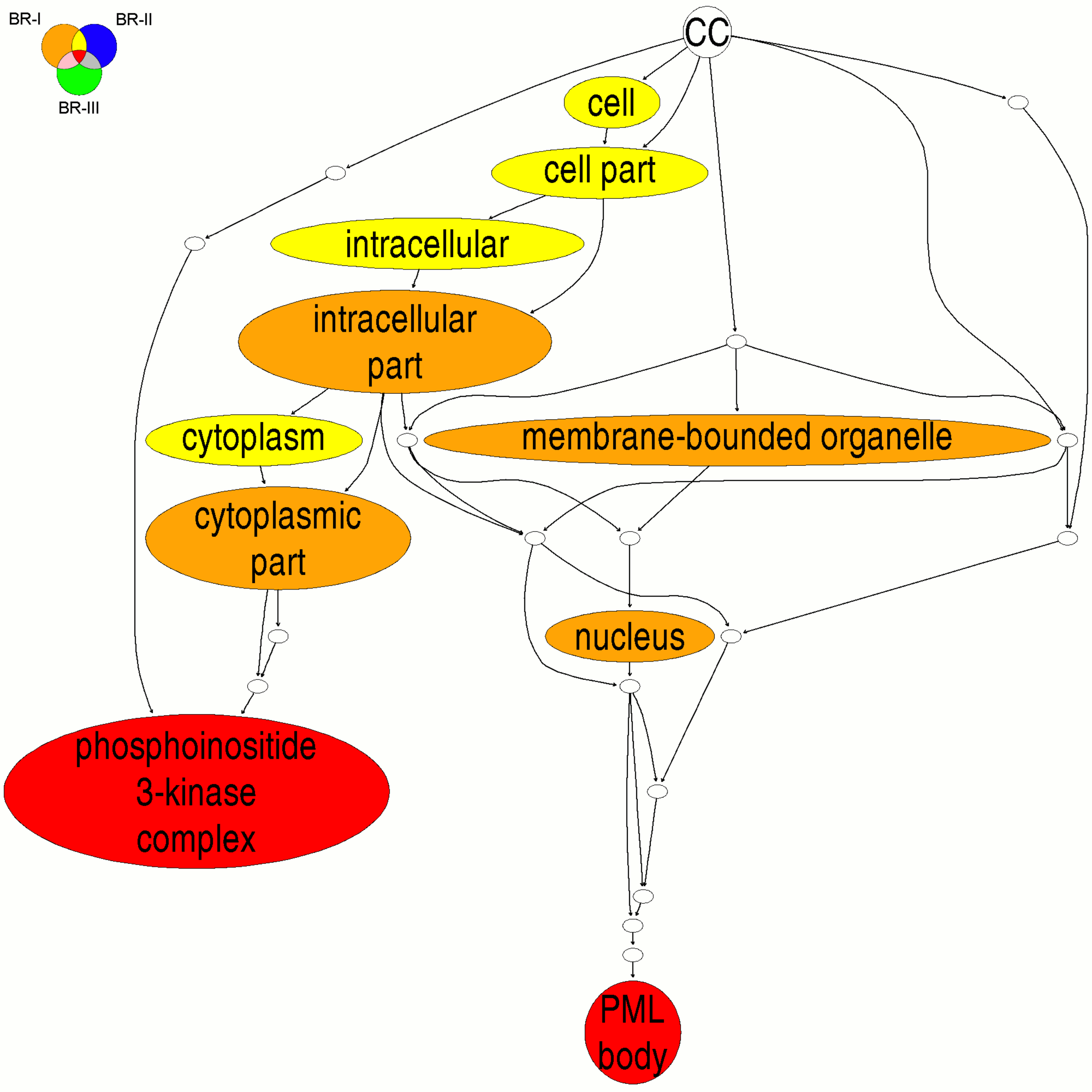
manganese ion binding

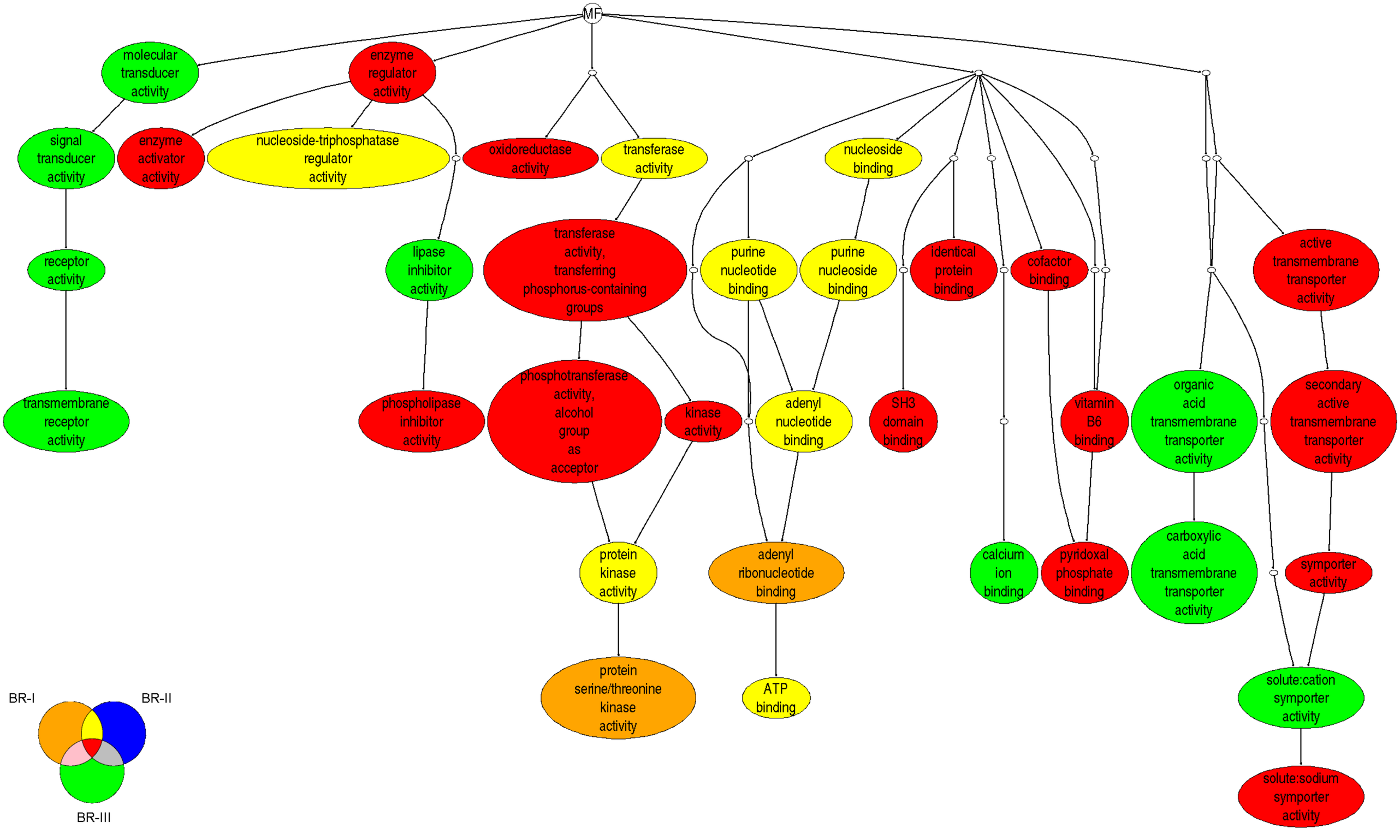

BP
biological adhesion
multicellular organismal process
regulation of biological quality
secondary metabolic process
alcohol metabolic process
oxidation reduction
cell division
homeostatic process
cell adhesion
response to chemical stimulus
anatomical structure morphogenesis
alcohol biosynthetic process
carbohydrate metabolic process
organic acid metabolic process
cellular ketone metabolic process
lipid metabolic process
heterocycle metabolic process
vitamin metabolic process
regulation of system process
regulation of response to external stimulus
regulation of response to stress
response to nitrosative stress
response to oxidative stress
response to toxin
anatomical structure formation involved in morphogenesis
carbohydrate biosynthetic process
nucleobase, nucleoside and nucleotide metabolic process
oxoacid metabolic process
coenzyme metabolic process
cellular lipid metabolic process
cell redox homeostasis
regulation of defense response
response to insecticide
vasculature development
monosaccharide biosynthetic process
nucleoside phosphate metabolic process
cellular carbohydrate biosynthetic process
carboxylic acid metabolic process
coenzyme biosynthetic process
regulation of inflammatory response
blood vessel development
skeletal system morphogenesis
nucleotide metabolic process
glutathione biosynthetic process
positive regulation of protein metabolic process
regulation of inflammatory response to antigenic stimulus
response to vitamin
blood vessel morphogenesis
fatty acid metabolic process
positive regulation of cellular protein metabolic process
regulation of synaptic plasticity
vesicle targeting, to, from or within Golgi
angiogenesis
cysteine metabolic process
regulation of mitochondrial depolarization
BR-I
BR-II
BR-III

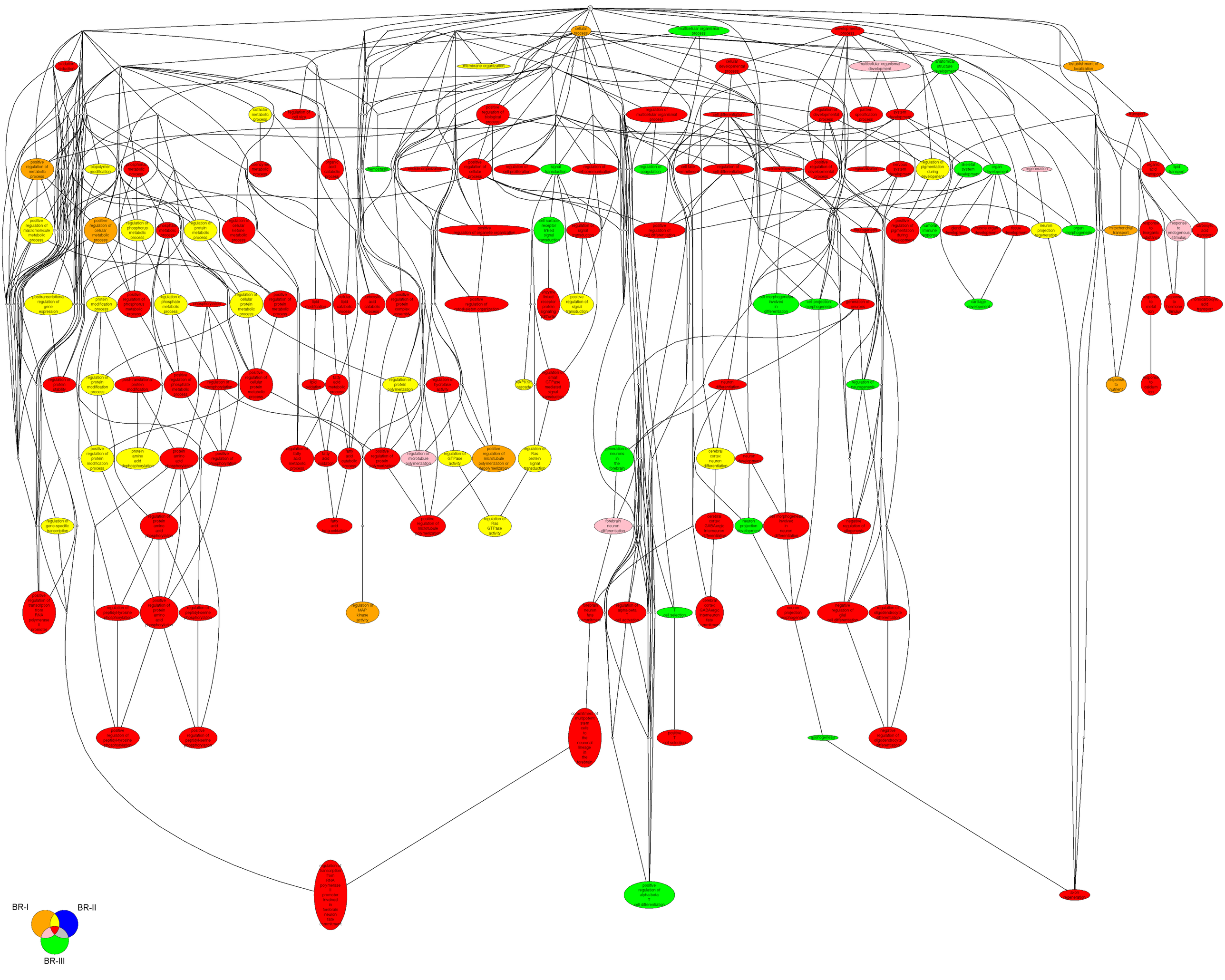

CC
BR-I
BR-II
BR-III
extracellular region
extracellular region part
extracellular space
extracellular matrix
proteinaceous extracellular matrix
cytoplasm
intrinsic to plasma membrane
integral to membrane
integral to plasma membrane
glutamate-cysteine ligase complex
vacuole
cytosol

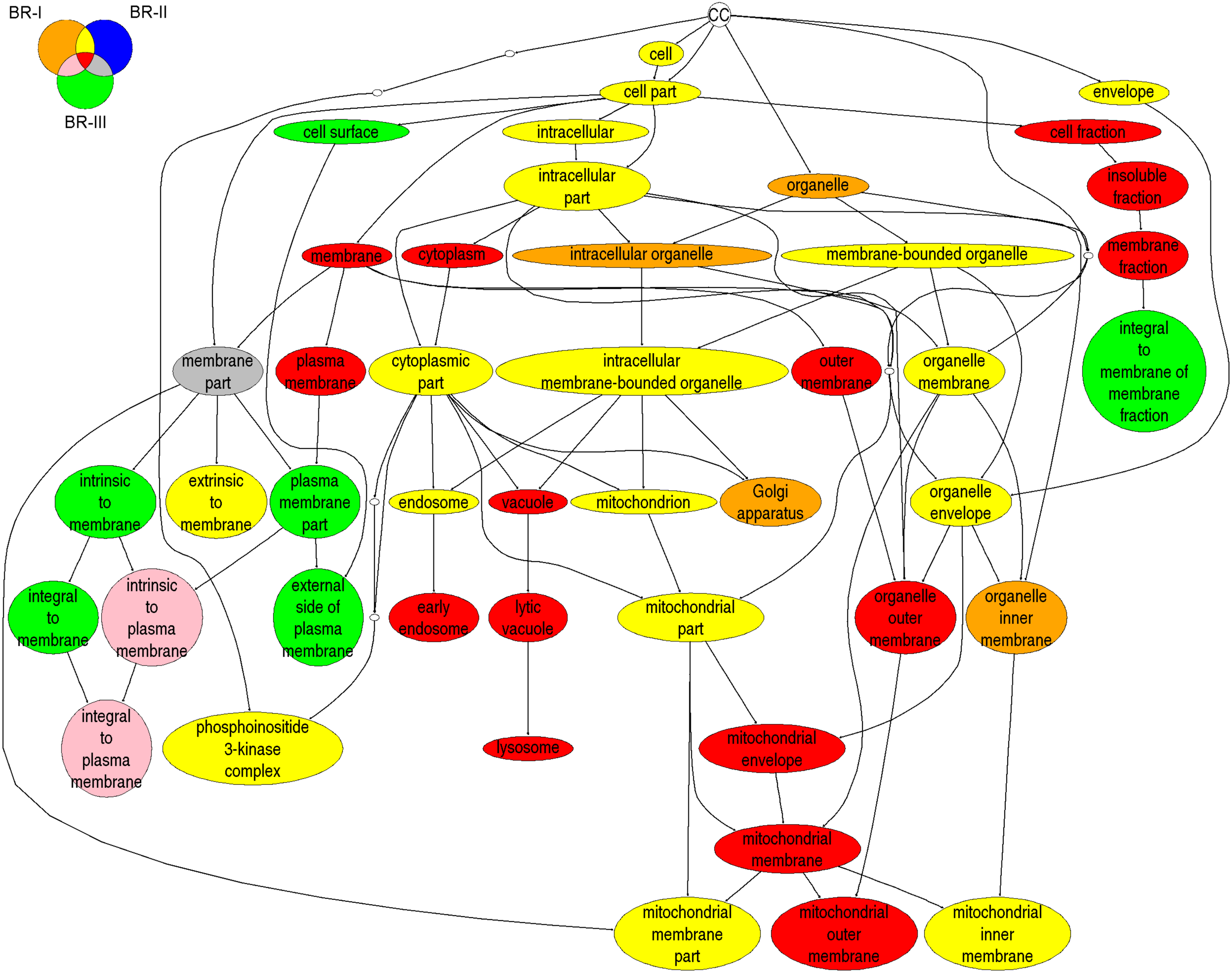
BR-I
BR-II
BR-III
MF
catalytic activity
binding
hydrolase activity
nucleotide binding
protein binding
hydrolase activity, acting on acid anhydrides
transferase activity, transferring glycosyl groups
ribonucleotide binding
purine nucleotide binding
RNA binding
phosphoinositide 3-kinase regulator activity
hydrolase activity, acting on acid anhydrides, in phosphorus-containing anhydrides
transferase activity, transferring pentosyl groups
purine ribonucleotide binding
guanyl nucleotide binding
pyrophosphatase activity
NAD+ ADP-ribosyltransferase activity
guanyl ribonucleotide binding
nucleoside-triphosphatase activity
GTP binding
zinc ion binding
GTPase activity

BP
immune system process
response to stimulus
multi-organism process
immune response
response to stress
response to chemical stimulus
response to biotic stimulus
aging
cellular response to stimulus
catabolic process
negative regulation of biological process
defense response
response to other organism
innate immune response
response to nutrient levels
response to virus
cellular catabolic process
negative regulation of cellular process
response to peptide hormone stimulus
negative regulation of cell proliferation
cytokine-mediated signaling pathway
biopolymer modification
post-translational protein modification
ISG15-protein conjugation
BR-I
BR-II
BR-III


BR-I
BR-II
BR-III
CC
cell
cell part
intracellular
intracellular part
cytoplasm
membrane-bounded organelle
cytoplasmic part
nucleus
phosphoinositide 3-kinase complex
PML body


MF
molecular transducer activity
enzyme regulator activity
signal transducer activity
enzyme activator activity
nucleoside-triphosphatase regulator activity
oxidoreductase activity
transferase activity
nucleoside binding
receptor activity
lipase inhibitor activity
transferase activity, transferring phosphorus-containing groups
purine nucleotide binding
purine nucleoside binding
identical protein binding
cofactor binding
active transmembrane transporter activity
transmembrane receptor activity
phospholipase inhibitor activity
phosphotransferase activity, alcohol group as acceptor
kinase activity
adenyl nucleotide binding
SH3 domain binding
vitamin B6 binding
organic acid transmembrane transporter activity
secondary active transmembrane transporter activity
protein kinase activity
adenyl ribonucleotide binding
calcium ion binding
pyridoxal phosphate binding
carboxylic acid transmembrane transporter activity
symporter activity
protein serine/threonine kinase activity
ATP binding
solute:cation symporter activity
solute:sodium symporter activity
BR-I
BR-II
BR-III


cellular process
multicellular organismal process
developmental process
oxidation reduction
membrane organization
cellular developmental process
multicellular organismal development
anatomical structure development
establishment of localization
cofactor metabolic process
regulation of cell size
positive regulation of biological process
regulation of multicellular organismal process
cell differentiation
regulation of developmental process
pattern specification process
system development
death
positive regulation of metabolic process
biopolymer modification
phosphorus metabolic process
carbohydrate metabolic process
organic acid catabolic process
homeostasis
vesicle organization
positive regulation of cellular process
regulation of cell proliferation
signal transduction
regulation of cell communication
regulation of coagulation
cell fate commitment
regulation of cell differentiation
cell development
positive regulation of developmental process
hemostasis
nervous system development
regulation of pigmentation during development
skeletal system development
organ development
regeneration
organic acid transport
lipid transport
positive regulation of macromolecule metabolic process
positive regulation of cellular metabolic process
regulation of phosphorus metabolic process
alcohol metabolic process
regulation of protein metabolic process
regulation of cellular ketone metabolic process
positive regulation of organelle organization
cell surface receptor linked signal transduction
regulation of signal transduction
positive regulation of cell differentiation
endochondral ossification
positive regulation of pigmentation during development
tumor necrosis immune response
gland development
muscle organ development
tissue development
neuron projection regeneration
organ morphogenesis
mitochondrial transport
response to inorganic substance
response to endogenous stimulus
carboxylic acid transport
posttranscriptional regulation of gene expression
protein modification process
positive regulation of phosphorus metabolic process
regulation of phosphate metabolic process
phosphorylation
regulation of cellular protein metabolic process
positive regulation of protein metabolic process
lipid modification
cellular lipid catabolic process
carboxylic acid catabolic process
positive regulation of protein complex assembly
positive regulation of cytoskeleton organization
activin-like receptor protein signaling pathway
positive regulation of signal transduction
cell morphogenesis involved in differentiation
cell projection morphogenesis
generation of neurons
cartilage development
response to metal ion
response to hormone stimulus
monocarboxylic acid transport
regulation of protein stability
regulation of protein modification process
post-translational protein modification
positive regulation of phosphate metabolic process
regulation of phosphorylation
positive regulation of cellular protein metabolic process
lipid oxidation
fatty acid metabolic process
regulation of protein polymerization
regulation of hydrolase activity
MAPKKK cascade
regulation of small GTPase mediated signal transduction
neuron differentiation
regulation of neurogenesis
response to nutrient
response to calcium ion
positive regulation of protein modification process
protein amino acid dephosphorylation
protein amino acid phosphorylation
positive regulation of phosphorylation
regulation of fatty acid metabolic process
fatty acid oxidation
fatty acid catabolic process
positive regulation of protein polymerization
regulation of microtubule polymerization
regulation of GTPase activity
positive regulation of microtubule polymerization or depolymerization
regulation of Ras protein signal transduction
generation of neurons in the forebrain
cerebral cortex neuron differentiation
neuron development
regulation of gene-specific transcription
regulation of protein amino acid phosphorylation
fatty acid beta-oxidation
positive regulation of microtubule polymerization
forebrain neuron differentiation
cerebral cortex GABAergic interneuron differentiation
neuron projection development
cell morphogenesis involved in neuron differentiation
negative regulation of oligodendrocyte differentiation
positive regulation of transcription from RNA polymerase II promoter
regulation of peptidyl-tyrosine phosphorylation
positive regulation of protein amino acid phosphorylation
regulation of peptidyl-serine phosphorylation
regulation of MAP kinase activity
forebrain neuron fate commitment
regulation of alpha-beta T cell activation
T cell selection
cerebral cortex GABAergic interneuron fate commitment
neuron projection morphogenesis
negative regulation of gliogenesis
regulation of oligodendrocyte differentiation
positive regulation of peptidyl-tyrosine phosphorylation
positive regulation of peptidyl-serine phosphorylation
commitment of multipotent stem cells to the neuronal lineage in the forebrain
positive T cell selection
gliogenesis
regulation of transcription from RNA polymerase II promoter involved in forebrain neuron fate commitment
positive regulation of alpha-beta T cell differentiation
axon ogeneration
BR-I
BR-II
BR-III


BR-I
BR-II
BR-III
CC
cell
cell part
envelope
cell surface
intracellular
cell fraction
intracellular part
organelle
insoluble fraction
membrane
cytoplasm
intracellular organelle
membrane-bounded organelle
membrane fraction
membrane part
plasma membrane
cytoplasmic part
intracellular membrane-bounded organelle
outer membrane
organelle membrane
integral to membrane of membrane fraction
intrinsic to membrane
extrinsic to membrane
plasma membrane part
endosome
vacuole
mitochondrion
Golgi apparatus
organelle envelope
integral to membrane
intrinsic to plasma membrane
external side of plasma membrane
early endosome
lytic vacuole
mitochondrial part
organelle outer membrane
organelle inner membrane
integral to plasma membrane
phosphoinositide 3-kinase complex
lysosome
mitochondrial envelope
mitochondrial membrane
mitochondrial membrane part
mitochondrial outer membrane
mitochondrial inner membrane

MF

binding

antioxidant activity

catalytic activity

enzyme regulator activity

translation factor activity, nucleic acid binding

ribosome binding

carbohydrate binding

sugar binding

protein binding

glycoprotein binding

kininogen binding

protein N-terminus binding

growth factor binding

identical protein binding

unfolded protein binding

receptor binding

protein complex binding

enzyme binding

collagen binding

oxidoreductase activity

oxidoreductase activity, acting on superoxide radicals as acceptor

peptidase activity

enzyme inhibitor activity

metalloenzyme regulator activity

peptidase inhibitor activity

metalloenzyme inhibitor activity

proteoglycan binding

platelet-derived growth factor binding

protein phosphorylated amino acid binding

integrin binding

kinase binding

superoxide dismutase activity

oxidoreductase activity, acting on the CH-OH group of donors, NAD or NADP as acceptor

peptidase activity, acting on L-amino acid peptides

intramolecular oxidoreductase activity, transposing S-S bonds

intramolecular oxidoreductase activity, interconverting keto- and enol-groups

endopeptidase inhibitor activity

magnesium ion binding

heparin binding

phosphotyrosine binding

protein kinase binding

peroxiredoxin activity

cysteine-type peptidase activity

endopeptidase activity

threonine-type peptidase activity

protein disulfide isomerase activity

serine-type endopeptidase inhibitor activity

metalloendopeptidase inhibitor activity

GDP-dissociation inhibitor activity

thioredoxin peroxidase activity

cysteine-type endopeptidase activity

threonine-type endopeptidase activity

voltage-gated anion channel activity

voltage-gated chloride channel activity

BR-I

BR-II

BR-III


BR-I
BR-II
BR-III
response to stimulus
localization of cell
cellular component organization
macromolecule localization
interspecies interaction between organisms
response to activity
oxidation reduction
regulation of response to stimulus
locomotion
cell motion
generation of precursor metabolites and energy
cell motility
regulation of organelle organization
plasma membrane organization
extracellular matrix organization
cellular macromolecule localization
establishment of protein localization
regulation of cellular response to stress
response to axon injury
positive regulation of locomotion
cell migration
cell development
nervous system development
organ development
positive regulation of cell motion
intermediate filament cytoskeleton organization
positive regulation of organelle organization
negative regulation of ligase activity
intracellular transport
protein transport
cellular protein localization
hexose metabolic process
response to steroid hormone stimulus
hydrogen peroxide biosynthetic process
superoxide anion generation
negative regulation of cell differentiation
regulation of ubiquitin-protein ligase activity during mitotic cell cycle
negative regulation of ubiquitin-protein ligase activity
regulation of protein complex disassembly
glucose metabolic process
regulation of leukocyte migration
positive regulation of cell migration
MAPKKK cascade
regulation of membrane potential
negative regulation of ubiquitin-protein ligase activity during mitotic cell cycle
positive regulation of protein complex disassembly
nucleoside diphosphate metabolic process
double-strand break repair
malate metabolic process
regulation of leukocyte chemotaxis
positive regulation of leukocyte migration
proteasomal protein catabolic process
proteasomal ubiquitin-dependent protein catabolic process
positive regulation of leukocyte chemotaxis
anaphase-promoting complex-dependent proteasomal ubiquitin-dependent protein catabolic process
pentose-phosphate shunt
negative regulation of protein ubiquitination

CC
envelope
extracellular region
membrane-enclosed lumen
macromolecular complex
extracellular region part
intracellular
cell surface
cell soma
cell fraction
cell projection
extracellular space
extracellular matrix
organelle
intracellular part
soluble fraction
protein complex
neuron projection
proteinaceous extracellular matrix
extracellular matrix part
non-membrane-bounded organelle
intracellular organelle
membrane-bounded organelle
cytoplasm
proteasome complex
ubiquitin conjugating enzyme complex
axon
basement membrane
vesicle
organelle part
intracellular non-membrane-bounded organelle
intracellular membrane-bounded organelle
cytoplasmic part
proteasome core complex
fibrillar collagen
membrane-bounded vesicle
organelle lumen
cytoplasmic vesicle
cytoskeleton
intracellular organelle part
endoplasmic reticulum
vacuole
mitochondrion
cytosol
ER-Golgi intermediate compartment
cell projection cytoplasm
vesicle lumen
cytoplasmic membrane-bounded vesicle
intermediate filament cytoskeleton
intracellular organelle lumen
lytic vacuole
organelle envelope
dendrite cytoplasm
pigment granule
cytoplasmic membrane-bounded vesicle lumen
intermediate filament
endoplasmic reticulum lumen
lysosome
mitochondrial lumen
nuclear envelope
melanosome
platelet alpha granule
mitochondrial matrix
platelet alpha granule lumen
BR-I
BR-II
BR-III