bioRxiv preprint doi: https://doi.org/10.1101/112201; this version posted March 2, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

An extensible ontology for inference of emergent whole cell function from relationships between subcellular processes

Jens Hansen[1,2], David Meretzky[1,2], Simeneh Woldesenbet[1,2,3], Gustavo Stolovitzky[4,5], and Ravi Iyengar[1,2]

[1] Department of Pharmacological Sciences and [2] Systems Biology Center New York, Icahn School of Medicine at Mount Sinai, New York NY 10029

[3] Department of Life Science, IMC University of Applied Sciences Krems, Krems an der Donau, Austria

[4] Thomas J. Watson Research Center, IBM, Yorktown Heights, NY USA and [5] Department of Genetics and Genomics Sciences, Icahn School of Medicine at Mount Sinai, New York NY

Address Correspondence to

Ravi Iyengar

Department of Pharmacological Sciences

Icahn School of Medicine at Mount Sinai

1425 Madison Room 12-70

New York NY 10029

Phone: 212-659-1707

e-mail Ravi.Iyengar@mssm.edu

bioRxiv preprint doi: https://doi.org/10.1101/112201; this version posted March 2, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Abstract:

Whole cell responses arise from coordinated interactions between diverse human gene products that make up various pathways for sub-cellular processes (SCP). Lower level SCPs interact to form higher level SCPs, often in a context specific manner. We sought to determine if capturing such relationships enables us to describe the emergence of whole cell functions from interacting SCPs. We developed the "Molecular Biology of the Cell" ontology based on standard cell biology and biochemistry textbooks. Currently, our ontology contains 754 SCPs and 18,989 expertly curated gene-SCP associations. In contrast with other ontologies, our algorithm to populate the SCPs with genes is flexible and enables extension of the ontology on demand. Since cell biological knowledge is constantly growing we developed a dynamic enrichment algorithm for the prediction of SCP relationships beyond the current taxonomy. This algorithm enables us to identify interactions between SCPs as a basis for higher order function in a context dependent manner, allowing us to provide a detailed description of how these SCPs together give rise to whole cell functions. We conclude that this ontology can, from omics data sets, enable the development of detailed SCP networks for predictive modeling of human cell biological functions.

bioRxiv preprint doi: https://doi.org/10.1101/112201; this version posted March 2, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Introduction

A major goal in systems biology at the cellular level is to understand how different subcellular processes (SCPs) that are organized as individual pathways or small networks [1] are integrated to give rise to whole cell functions. This is critical for understanding how interactions between proteins (i.e. gene products) are required for cell, tissue, and organ level physiological responses such as energy generation, action potential generation, contractility, cell movement, and secretion. Whole cell functions arise from coordinated activities of SCPs that are made up of multiple gene products, typically organized as pathways or small networks. SCPs interact with each other in a context specific manner; depending on the whole cell function, an SCP can interact with different sets of other SCPs. Delineating interactions between SCPs for cellular functions of interest can help us build coarse-grained whole cell models of mammalian cells. These models can be used to interpret and predict both cellular and tissue level behaviors in response to a variety of perturbations.

High-throughput experimental methods, such as transcriptomics and proteomics, enable us to identify many genes that are associated with whole cell function or change in cellular state. Computational tools like gene set enrichment analysis [2, 3] allow for the identification of ranked lists of SCPs that are involved in the cellular physiological functions being studied. For such analyses, there are a number of biological ontologies that are available. Gene Ontology [4, 5] is arguably the most well-known ontology and is useful in associating gene lists obtained by high-throughput technologies with biological function. There are other valuable pathway based ontologies as well, such as KEGG (http://www.genome.jp/kegg/),Wikipathways (http://www.wikipathways.org) or Reactome (http://www.reactome.org/). Recently, a data-driven approach has been used to build a network-extracted ontology–NeXO [6]- where molecular components and their hierarchical interrelationships have been defined and populated with gene products based on the network analysis of yeast protein-protein interactions.

Current ontologies provide an excellent base for the development of a new relational ontology that captures how interactions between SCPs can give rise to whole-cell functions. Such an ontology should not only contain relationships that are currently established, but also have the capability to enumerate interrelationships between SCPs that lie outside of the annotated hierarchical organization. If an ontology can be extensible in a facile manner, the relationships identified by the ontology can enable systems level discovery of how higher level functions can rise from lower level SCPs. Physiological function often depends on the context specific relationships between SCPs that are not grouped together based on canonical taxonomy. A critical step in going from genotype to phenotype is the unambiguous delineation of interactions between SCPs that are organized to function coordinately to yield whole cell functions. In this study we sought to develop such a cell biology ontology that enables discovery of relationships between SCPs for whole cell function. Research in cell biology and biochemistry over the past fifty years has largely focused on identifying the molecular basis of cell physiology at various levels, and thus provides detailed descriptions of SCPs and their relationships. Such SCPs are typically organized in text books as pathways or small networks. These pathways or small networks can be used to describe diverse functions such as biosynthesis or degradation of amino acids or sugars, mRNA surveillance, and actin cytoskeleton dynamics. We reasoned that using a well-recognized cell biology textbook could be a good starting point to develop a cellular ontology that reflects prior knowledge of how SCPs interact at multiple levels to give rise to whole cell functions. We used Molecular Biology of the Cell [7] as our

bioRxiv preprint doi: https://doi.org/10.1101/112201; this version posted March 2, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

primary source to develop a cell biological ontology and call it the MBC ontology. Additionally, we used biochemistry text books [8, 9] and review articles (Suppl. table 1) to develop an integrated backbone for the ontology. We populated the ontology with genes by combining text mining and statistical enrichment approaches of research articles available on PubMed. Currently, the MBC ontology contains 19,412 gene-SCP relationships arising from 754 SCPs and 5350 genes. To make our MBC ontology extensible, we developed an algorithm that allows for the identification of relationships between SCPs irrespective of whether they lie within or outside the initial taxonomy generated from the textbooks and review articles. This algorithm allows for enrichment analyses by integrating SCPs in a context specific manner. Such integration enables for the identification of those SCPs that form the basis for the whole cell function of interest.

Results

Generation and population of the MBC Ontology

We used the Molecular Biology of the Cell [7], additional biochemistry textbooks [8, 9], and review articles (suppl. table 1) to define SCPs and arrange them in hierarchical parent-child relationships that typically span three or four levels. For example, the level-1 SCP Cytoskeleton dynamics is a child of the level-0 overall cell function process Molecular Biology of the Cell and has five level-2 children SCPs, among which is the SCP Actin filament dynamics (Fig. 1A). The latter has 6 level-3 children SCPs and 6 level-4 grand children SCPs. In total our ontology consists of 27 level-1 SCPs, 124 level-2 SCPs, 486 level-3 SCPs and 115 level-4 SCPs (Figure 1B).SCPs of the same level describe cell biological/biochemical functions of similar detail. Higher levels, i.e. levels that are closer to the level-0 overall cell function, contain SCPs that describe more general functions. Lower levels contain SCPs that describe more detailed sub-cellular functions. SCPs that have the same parent are siblings and belong to the same children set.

To populate our ontology with genes we used a combination of text mining and enrichment analysis. We downloaded SCP-specific abstracts from the PubMed website using SCP-specific PubMed queries (supp. table 1). To identify genes and other biological entities in these abstracts, such as metabolites, we generated a dictionary for text terms of biological entities(suppl. Fig. 1-14, suppl. tables 2 - 31).We used the dictionary to count the number of articles within each SCP-specific article set that mention a particular gene at least once (step 1 in Fig. 2 and suppl. Fig. 15). The identified gene lists contained many false positive gene-SCP associations since SCP specific PubMed abstracts do not only contain gene products that belong to that particular SCP; they may also contain gene products that interact with that SCP, or are processed by it, but are part of another SCP. Biologists often select proteins to study a particular SCP that in reality belong to a different SCP, so the article counts of these proteins are normally higher in the abstract set of the SCP within which it has primary function. For example, transferrin, along with its receptor, is internalized by clathrin-mediated endocytosis and transferrin is often used to investigate the endocytosis process. In reality, the transferrin receptor belongs to the SCP, Cellular iron uptake, as transferrin is the major carrier of iron in blood. Similarly, many proteins that are glycosylated in the Golgi are mentioned as substrates in abstracts that describe the glycosylation machinery; however, these proteins may have different functions and hence belong to different SCPs, distinct from the SCPs describing the glycosylation pathway. To remove such false positive gene associations, we subjected the abstract counts to two rounds of statistical enrichment analysis, each followed by the automatic removal

bioRxiv preprint doi: https://doi.org/10.1101/112201; this version posted March 2, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

of non-selective genes, i.e. genes that belonged to a different SCP. Both times, we used Fisher's exact test to calculate a p-value for each gene-SCP association (suppl. Fig. 16). Afterwards, our algorithm compared the p-value distribution that was obtained for each gene and kept only those SCP associations of that gene which were associated with the most significant p-values (suppl. Fig. 17, see methods for details). The first enrichment calculated, for each gene, the selectivity of SCP association in comparison to all other SCPs of the same level (step 2a). The second enrichment calculated the selectivity of each gene-SCP association in comparison to all other SCPs of the same children set (step 3). The second enrichment step was necessary to correctly distribute the genes between the different children set processes. In many cases a gene that belonged to a particular SCP was also associated with one or more of its sibling SCPs after the first enrichment step. SCPs of the same children set often contained very high abstract counts for that gene in comparison to the other SCPs of the same level, causing the calculation of very low p-values for the association of that gene with all sibling SCPs. The direct comparison of the abstract counts between the sibling SCPs in the second enrichment step increased the accuracy of our algorithm to correctly associate the gene with that (sibling) SCP it belongs to. Gene level-2 and level-3 SCP associations were manually validated(step 4)(Suppl. table 33), and the ontology was re-populated after integration of the manual validation results (for details see suppl. methods and Fig. 2). For the population of level-1 and level-4 SCPs we used a modified protocol. Since the first enrichment step depends on a sufficiently high number of SCPs (>100) within the same level that was not the case for level-1 SCPs, as well as a broad coverage of known biology that was not the case for level-4 SCPs, we replaced the first enrichment step for these levels by an inheritance procedure (step2b). We kept only those genes in level-1 SCPs that were at least part of one of its level-2 children or level-3 grand children SCPs. Similarly, we kept only those genes of each level-4 SCP that were also part of its level-3 parent. The second enrichment step (step 3) was applied to level-1 and level-4 SCPs as described above. We then added all genes of level-3 SCPs to their level-2 parent and level-1 grandparent SCPs if they were not already associated with them (step 5). Similarly, we added all genes of level-2 SCPs to their level-1 parent SCPs.

Internal consistency of the populated MBC Ontology

Our populated ontology contained 754 SCPs and 5350 genes organized in 19,412gene-SCP associations. Level-1 SCPs consisted on average of 233 +/- 218 genes, level-2 SCPs of 55 +/- 46 genes, level-3 SCPs of 11 +/- 11 genes and level-4 SCPs of 6 +/- 7 genes (Fig. 3). We were interested in whether our population algorithm was able to generate consistent results between level-2 parent and level-3 children SCPs, i.e. if our ontology is internally consistent. If the hierarchical SCP relationships reflect common biological understanding, there should be a significant overlap between the genes of parent and children SCPs. For this analysis, we focused on level-2 and level-3 SCPs that were populated based on both enrichment steps, and considered their gene composition before the addition of the genes of children SCP to their parent SCPs. We determined for each level-3 child SCP how many of its genes are also associated with its level-2 parent SCP (suppl. Fig.18b, suppl. table 34). The average percentage of child SCP genes that were also associated with its annotated parent was 45.9% +/- 36.9%. To document the accuracy of our annotation, we identified for each level-3 child SCP that level-2 SCP that contained the highest number of the genes of the level-3 child SCP. Often we identified more than one level-2 SCP during this analysis. Nevertheless, in about 74% of the children SCPs the annotated parent was either the best matching or among the best matchinglevel-2 SCPs.

bioRxiv preprint doi: https://doi.org/10.1101/112201; this version posted March 2, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

To ascertain a (nearly) complete coverage of the sub-functions of a level-2 parent SCP by its level-3 children SCPs, we determined how many genes of a parent SCP were also part of at least one of its children SCPs (suppl. Fig.18c). For most of the parent SCPs the children SCPs contained on average 76% of the parental genes, indicating that we did not miss major sub-functions of the parent SCPs. To analyze if the children SCPs describe mutually exclusive functions, we documented the overlap between the parent genes and the union of all its children genes after the removal of one or more children SCPs. If every child SCP describes a unique sub-function of the parental SCP, the removal of every child SCP should decrease the number of overlapping children set genes that are part of the parent SCP. For one of the level-2 parent SCPs at a time, we sequentially removed one to all of its level-3 children SCPs, re-populated the remaining level-3 SCPs using the entire pipeline for population of SCPs (see Fig.2, excluding step 4 (manual validation)). We re-calculated the overlap between the genes of the parent and remaining children SCPs(suppl. Fig.18c), and observed a continuous decrease in the number of overlapping genes. This further documents the internal consistency of our ontology and the biological correctness of our annotated hierarchy.

Genes within an SCP show correlated expression

Genes that are involved in the same sub-cellular function act in concert with each other and could correlate in their gene expression over different conditions or tissues[10].We analyzed if our algorithm populates SCPs with genes that show high co-expression and can therefore be assumed to be functionally related. Using genome-wide expression data, measured in different human tissues (http://www.gtexportal.org)[11], we calculated the Pearson correlation between all pairs of genes. For each level, we then distributed the correlations into two groups: the first group contained all correlations between any two genes that were associated with at least one common SCP of that level, the second group contained all correlations between any two genes that did not belong to any common SCP. In this analysis, we focused on level-3 SCPs since they describe clearly defined sub-cellular functions and do not summarize multiple different or opposite sub-cellular functions as is the case for many level-1 and level-2 SCPs. The correlations between the genes of both groups belong to two significantly different populations (Fig.4), providing evidence that genes of the same level-3 SCP show more correlated expression. As expected, the two populations for level-1 and level-2 SCPs were less different (suppl. Fig. 19).We conclude that the gene co-expression analysis further supports the ability of our algorithm to populate SCPs with genes that are most likely a part of the function of the SCP.

Inference of relationships between SCPs at the same level

The competitive nature of our algorithm to populate SCPs allowed us to infer new SCP-SCP relationships between level-3 SCPs independent of their annotated family hierarchical relationships. These inferred relationships can be a guide for biological understanding, as well as form the basis for a dynamic enrichment analysis that considers context specific SCP interactions. SCPs of the same level compete with each other for genes during the population approach. The degree of competition between any pair of two SCPs can be used to define weighted horizontal SCP-SCP relationships: the more that two SCPs compete with each other for genes, the more that they are related to each other. To identify the degree of competition between any two SCPs of the same level, we performed leave-one-out analysis. We iteratively removed one SCP at a time from the level-3 SCP set, and re-populated the remaining level-3 SCPs as described above (Fig. 2). For each of the remaining SCPs, we compared the alternative gene sets

bioRxiv preprint doi: https://doi.org/10.1101/112201; this version posted March 2, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

with the original gene sets to identify those remaining processes that were influenced by the removal of the current SCP. We developed an additional algorithm to quantify these effects. This algorithm considers new gene additions to the remaining SCPs as well as rank increases of original genes within the remaining SCPs. This way we could generate a level-3 SCP-SCP interaction network where edges link the removed SCPs to the remaining SCPs and the edge width is proportional to the distribution of genes between these two SCPs. We predicted 2484 interactions between level-3 SCPs that connect SCPs within and beyond the annotated families (see Fig. 5).

Dynamic enrichment analysis enables the emergence of whole cell functions from interactions between SCPs

To test if our ontology can be used to resolve whole cell functions as sets of interacting SCPs, we used data from three published studies, using different high-throughput approaches to characterize a well-defined whole cell phenotype. We wanted to determine if we could predict the phenotype described in the publication from the results of our enrichment analysis without any prior assumptions. We first analyzed the data using standard enrichment analysis, and then compared this approach to a new type of enrichment analysis that considers context specific interactions between SCPs, which we call dynamic enrichment analysis. Standard enrichment analysis uses statistical tests such as Fisher's exact test to identify SCPs that are significantly associated with the experimental gene list. SCPs are then ranked by significance, followed by manual inspection and selection of those processes that can be related to the whole cell function of interest. Often the researcher focuses on the top ranked SCPs and ignores the rest. Whole cell functions are based on various SCPs which are distributed over different families and therefore are described in different sub-chapters of text books, but can function in an integrated manner. In such a scenario, an SCP with low statistical significance could still make an important contribution to whole cell function if it is related to another SCP with high statistical significance. We used the inferred relationships between level-3 SCPs to address this question by use of the dynamic enrichment approach. For each dataset, we generated new context-specific higher level SCPs by merging two or three level-3 SCPs that were connected by inferred relationships and contained at least one experimentally determined gene (e.g. a differentially expressed gene or a differential protein binding partner). We added these context specific level-3 SCP-units to the standard SCPs and repeated the enrichment analysis. The SCP networks obtained by the dynamic enrichment analysis described the emergence of the whole cell functions that were analyzed in the published papers used as case studies. This could be done without any prior assumptions or manual selection. Three examples are described below.

In the first case study, we tested our ontology on proteomic data by analyzing the mutant cystic fibrosis transmembrane conductance regulator (dF508 CFTR) interactome [12]. The mutated anion channel is the cause of cystic fibrosis. Impaired dF508 CFTR folding leads to its premature intracellular degradation and consequently reduces CFTR activity at the plasma membrane of bronchial epithelial cells, characterizing cystic fibrosis as a protein folding disease. We sought to determine if we could identify the SCPs that enable protein proofreading and degradation in cells from the proteomic data without any prior assumptions. The authors investigated the different protein interaction partners of WT CFTR vs dF508 CFTR via co-immunoprecipitation and proteomic analysis. Standard enrichment analysis of the identified different interaction partners revealed SCPs that were distributed over a wide range of cellular functions: actin cytoskeleton, vesicular traffic, glucose and lipid metabolism, ribonucleoprotein biogenesis, protein folding, and quality control in the ER as well as proteasomal degradation (suppl. figure 20A and B). All

bioRxiv preprint doi: https://doi.org/10.1101/112201; this version posted March 2, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

these processes were described by the authors after manual inspection of the genes and were thought to contribute to the reduction of plasma membrane dF508 CFTR. Based on the statistical significance of the SCPs the results of the standard enrichment analysis would propose potential changes in ribosome metabolism, sugar metabolism and vesicular traffic as major disease mechanisms. In contrast, the dynamic enrichment analysis predicted protein folding defects and proteasomal degradation, as the major pathogenic mechanism (figure 5a, suppl. figure 20c and d). We summarized these SCPs under the new inferred context-specific parent SCP called "Protein proofreading and degradation", allowing us to provide a logic for the hierarchical organization of SCPs that give rise to the observed cellular phenotype. It should be noted that the SCP Protein proof reading and degradation which is not in the starting taxonomy can be related to the two level-1 SCPs Intracellular degradation and Post-translational protein modification (dashed lines in figure. 6a).

In the second case study, we tested our ontology using a list of genes that were identified in a genome-wide RNAi screening as stimulators of the secretory pathway [13]. We sought to determine if we could identify secretory pathways as the whole cell functions from the lists of knocked out genes. The authors identified these genes by documenting the effect of their knockdown on the surface arrival of the fluorescently labeled viral protein tsO45G. Standard enrichment analysis identified hierarchically connected level-1 to level-4 SCPs that describe vesicular transport, nucleotide metabolism and gene expression (suppl. figure 21a and b). To distinguish which of these are the primary SCPs involved we conducted dynamic enrichment analysis. The results of the dynamic enrichment analysis showed a cluster of SCPs that described vesicle fusion and membrane recycling processes at the plasma membrane (figure 5b), indicating that the investigated phenotype is the two major parts of the secretory pathways. Thus dynamic enrichment analysis with the MBC Ontology allows us to correctly identify the SCPs driving the studied whole cell function.

In the third case study, we used gene expression data as the test dataset. The pretreatment of the breast cancer cell line BT-20 with the EGFR inhibitor erlotinib for 4 to 24h dramatically sensitizes the cell line towards doxirubicin induced apoptosis[14]. To identify possible mechanisms for this observation, the authors analyzed differentially expressed genes (DEGs) after 6h and 24h erlotinib treatment and identified significant changes in 16 out of 34 GeneGo cellular networks that include pathways related to apoptosis and DNA damage response. They also demonstrated that treatment with erlotinib diminishes the ability of the cells to form colonies in soft agar, a standard test for metastatic potential. We sought to determine if the MBC Ontology, using the transcriptomic data could identify the SCPs by which erlotinib pretreatment sensitizes the cells to doxyrubicin. Using our ontology, we re-analyzed the DEGs. Using standard enrichment analysis we identified multiple different SCPs based on the DEGs after 6h erlotinib treatment (e.g. Actin filament dynamics, TGF-beta superfamily signaling, RNA surveillance and degradation or Pre-mRNA 3-end-cleavage and polyadenylation) (suppl. figure 23a and b), as well as after 24h treatment (e.g. Apoptosis, JAK/STAT signaling pathway or Hemidesmosome organization) (suppl. figure 23e and f). None of these results clearly points to a mechanism that is triggered by erlotinib to sensitize the cells to doxyrubicin toxicity. Dynamic enrichment analysis of the DEGs after 6h identified a cluster of SCPs that describes transcription associated DNA re-arrangements (Fig. 6c), an underrecognized cellular process [15]. This SCP indicates he mechanism of doxyrubicin action involving Topoisomerase II poisoning, DNA adduct formation, and DNA intercalation that can influence DNA topology and nucleosome dynamics [16]. The results of the dynamic enrichment analysis therefore indicate that erlotinib pre-treatment re-wires the cellular machinery at doxyrubicin's site of action, suggesting a mechanism for increased sensitivity

bioRxiv preprint doi: https://doi.org/10.1101/112201; this version posted March 2, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

towards doxyrubicin that was not discussed by the authors. The dynamic enrichment results of the DEGs after 24h erlotinib treatment offer an explanation for the diminished colony formation activity induced by erlotinib. We identified a cluster of SCPs describing cell cortical actin filament dynamics (Fig. 6). Actin filament remodeling has been documented as a major cause for significant decrease in migration and colony formation [17].

Discussion

To understand how genomic and proteomic determinants are integrated into cell, tissue, and organ level responses, we need to understand how these determinants are organized to execute selective whole cell functions. Many individual pathways and small networks within cells combine to produce cellular phenotypes. Without a flexible ontology that focuses on cell biological pathways and networks it will be impossible to relate genomic and proteomic characteristics to cell physiology as the basis for tissue and organ level function. The Molecular Biology of the Cell Ontology is designed to cover sub-cellular activities that together give rise to whole cell function. We used standard cell biological literature to design an ontology that matches the common approaches used by cell biologists and biochemists to organize knowledge. Currently, the ontology contains 753 SCPs that cover a wide range of cell biological mechanisms, although the current list of SCPs is not comprehensive. Testing our ontology on varied published data sets obtained from diverse high-throughput technologies, we find that the MBC Ontology provides a reasonable organizational map that allows for intuitive understanding of hierarchical relationships for the emergence of whole cell functions.

Different whole cell functions may engage sets of similar and different pathways (i.e. SCPs), in different combinations. We developed an algorithm that enables an extensible ontology for the prediction of relationships between level-3 SCPs that are within as well as beyond the relationships between annotated sibling SCPs. We used these predictions to introduce - to our knowledge for the first time - the concept of dynamic enrichment analysis that considers process dependencies in a context specific manner. The lack of consideration of such process dependencies is one of the major criticisms of standard enrichment analysis [18]. By applying dynamic enrichment analysis to the test datasets, we were able to accurately predict the whole cell response being studied. Additionally, dynamic enrichment analysis made us consider SCPs that well-matched the investigated phenotype in contrast to the results of standard enrichment analysis, where they ranked lower, such that they could have been overlooked when considering only the top 5 or 10 processes.

Currently, our ontology only accounts for about a quarter of the human genome. However, our population and validation approach allows for extension of the ontology as needed. As the gene assignment to an SCP is flexible, as are the horizontal and vertical relationships between SCPs, the MBC Ontology can be easily extended beyond what we described here for additional cellular, tissue, and organ level functions. As biological validation of a function of a gene within a sub-cellular process is likely to come from detailed studies, capturing such validation in an ontology will require extensive literature study. This is essentially a big data problem wherein one requires a relatively facile approach. As part of developing the MBC Ontology, we have devised an automated text-mining algorithm that allows for a relatively rapid manual validation of functional associations. Thus we are able to combine the speed of computer searches with expert human knowledge. As cell biology, biochemistry, and physiology are vast fields with different domains of expertise, this combination of computer searches and human validation can be used

bioRxiv preprint doi: https://doi.org/10.1101/112201; this version posted March 2, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

in the longer term to produce a definitive ontology of all sub-cellular processes that account for all of the gene products encoded by the human genome.

Methods

Methods and associated references are available in the online version of the paper.

Author contributions

J.H. and R.I. developed the overall concepts for the project. J.H. designed the ontology, developed the algorithms, and tested it. D.M. analyzed co-expression of genes within SCPs. S.W. tested the ontology using standard enrichment analysis. J.H., D.M., G.S. and R.I. analyzed the results and wrote the manuscript. R.I. has overall responsibility for this study.

Acknowledgements

This study was supported by NIH grants GM54508 andP50-GM071558.

References

1. Jordan, J.D., Landau, E.M. & Iyengar, R. Signaling networks: the origins of cellular multitasking. *Cell* **103**, 193-200 (2000).
2. Mootha, V.K. et al. PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nature genetics* **34**, 267-273 (2003).
3. Subramanian, A. et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 15545-15550 (2005).
4. Ashburner, M. et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature genetics* **25**, 25-29 (2000).
5. Gene Ontology, C. Gene Ontology Consortium: going forward. *Nucleic acids research* **43**, D1049-1056 (2015).
6. Dutkowski, J. et al. A gene ontology inferred from molecular networks. *Nature biotechnology* **31**, 38-45 (2013).
7. Alberts, B., Johnson, A., Lewis, J., Morgan, D. & Raff, M. Molecular Biology of The Cell. *Taylor & Francis Group* (2015).
8. Berg, J.M., Tymoczko, J.L. & Stryer, L. Biochemistry. *New York: W. H. Freeman* **5th ed.** (2002).
9. Rosenthal, M.D. & Glew, R.H. Medical Biochemistry. *John Wiley & Sons, Inc.* (2009).
10. Han, J.D. et al. Evidence for dynamically organized modularity in the yeast protein-protein interaction network. *Nature* **430**, 88-93 (2004).
11. Mele, M. et al. Human genomics. The human transcriptome across tissues and individuals. *Science* **348**, 660-665 (2015).

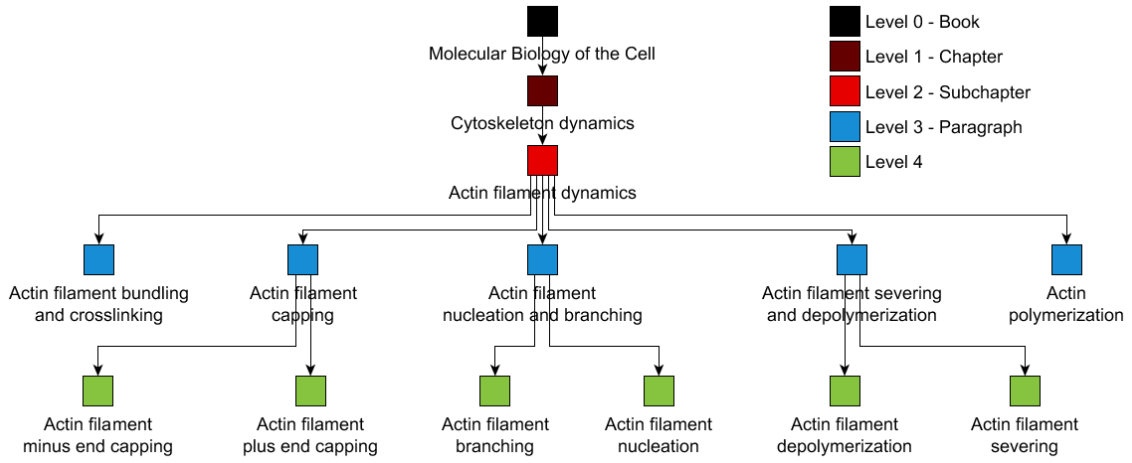
bioRxiv preprint doi: https://doi.org/10.1101/112201; this version posted March 2, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

12. Pankow, S. et al. F508 CFTR interactome remodelling promotes rescue of cystic fibrosis. *Nature* **528**, 510-516 (2015).
13. Simpson, J.C. et al. Genome-wide RNAi screening identifies human proteins with a regulatory function in the early secretory pathway. *Nature cell biology* **14**, 764-774 (2012).
14. Lee, M.J. et al. Sequential application of anticancer drugs enhances cell death by rewiring apoptotic signaling networks. *Cell* **149**, 780-794 (2012).
15. Kim, N. & Jinks-Robertson, S. Transcription as a source of genome instability. *Nature reviews. Genetics* **13**, 204-214 (2012).
16. Yang, F., Teves, S.S., Kemp, C.J. & Henikoff, S. Doxorubicin, DNA torsion, and chromatin dynamics. *Biochimica et biophysica acta* **1845**, 84-89 (2014).
17. Engel, N. et al. Actin cytoskeleton reconstitution in MCF-7 breast cancer cells initiated by a native flax root extract. *Advancement in Medicinal Plant Research* **3**, 92-105 (2015).
18. Khatri, P., Sirota, M. & Butte, A.J. Ten years of pathway analysis: current approaches and outstanding challenges. *PLoS computational biology* **8**, e1002375 (2012).

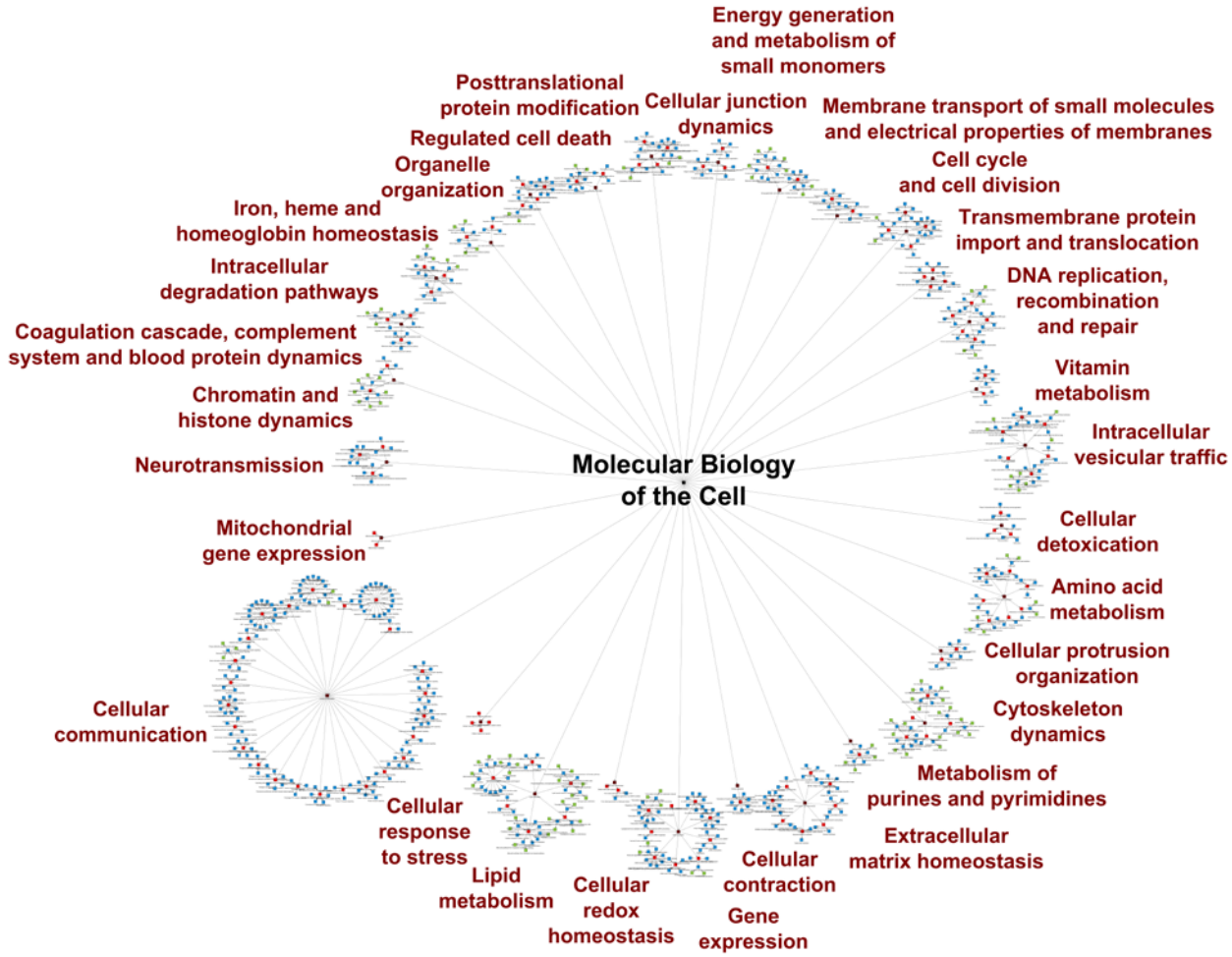
bioRxiv preprint doi: https://doi.org/10.1101/112201; this version posted March 2, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Figure 1

a



b



bioRxiv preprint doi: https://doi.org/10.1101/112201; this version posted March 2, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Fig. 1: **Organization of the Molecular Biology of the Cell (MBC) Ontology**. (a) The hierarchical organization of the MBC Ontology is demonstrated using the example of level-2 SCP Actin filament dynamics. It is the child of the level-1 SCP Cytoskeleton dynamics and has 6 level-3 children SCPs and 6 level-4 grandchildren SCPs. The biological detail that is described by the SCPs increases with increasing level number and corresponds to the units of the MBC textbook, where indicated. Each SCP is made of several genes/gene products that are organized as pathways or small networks. (b) The MBC Ontology currently consists of 754SCPs that are hierarchically organized from level-1 to level 4 as shown in figure 1a. Processes are colored as indicated in figure 1a. Names describe all level-1 SCPs.

bioRxiv preprint doi: https://doi.org/10.1101/112201; this version posted March 2, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Figure 2

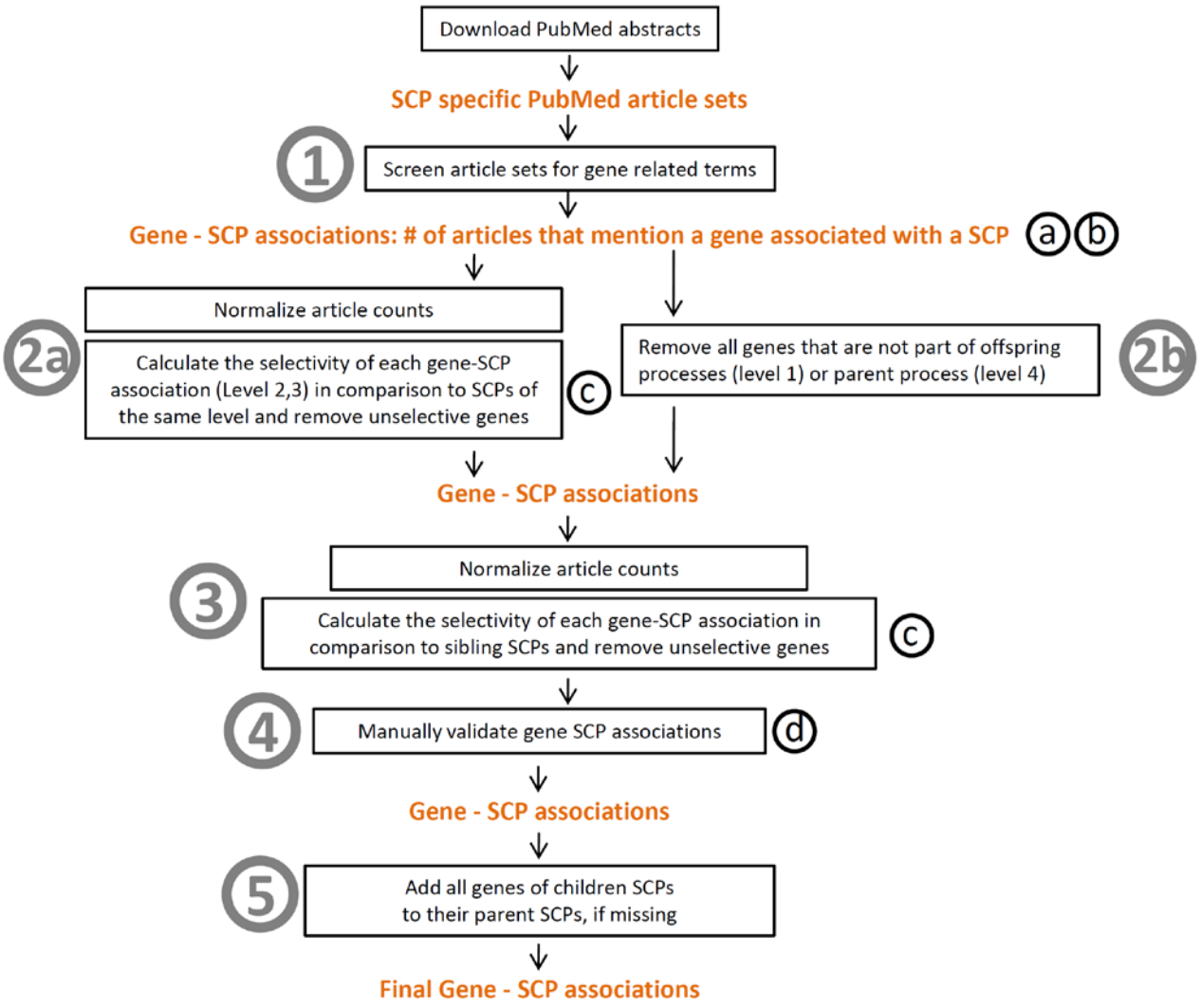


Fig. 2: Computational pipeline to populate SCPs with genes/gene products. Circled numbers indicate steps used to populate the SCPs with genes/gene products. Circled small letters indicate the following operational modifications that were incorporated based on the manual validation of the gene - SCP associations. (a) Remove genes that are the results of misinterpreted terms during the text mining of the PubMed article sets. (b) Decrease the abstract counts of all genes that were assigned to a SCP when they belong to a sibling SCP. (c) Never remove a gene that was manually identified as a true positive in an earlier version of the MBC Ontology. (d) Remove all genes that were manually identified as false positives. See main text and suppl. methods for details.

bioRxiv preprint doi: https://doi.org/10.1101/112201; this version posted March 2, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Figure 3

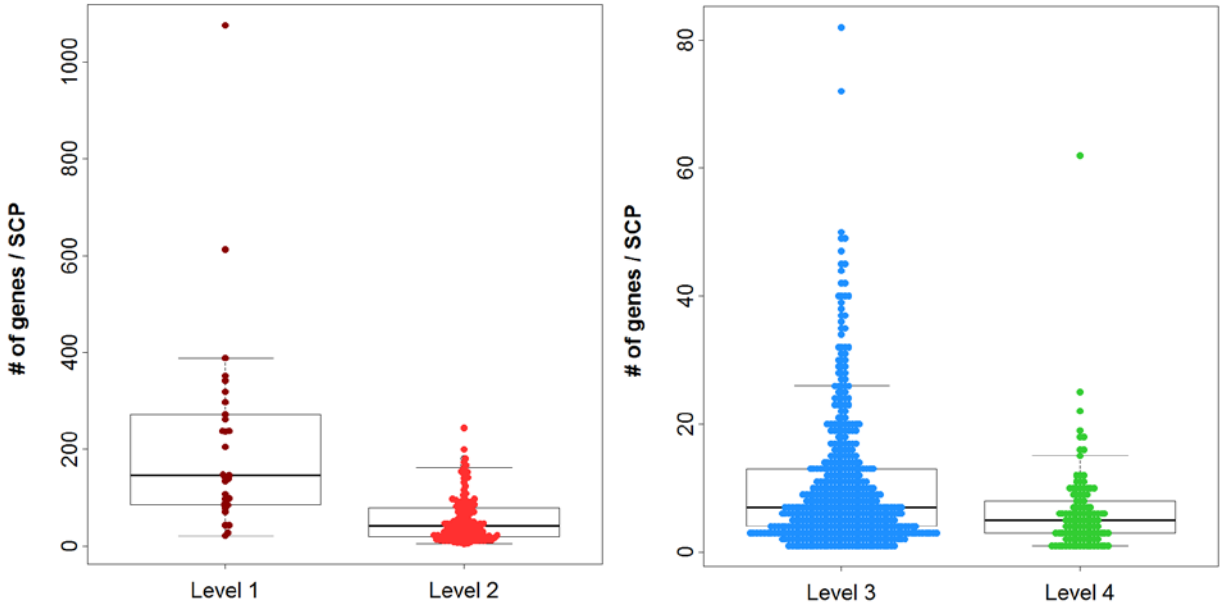


Fig. 3: Numbers of genes per SCPs at various levels. The number of genes that were associated with each level 1 SCP and level 2 SCP as well with each level 3 and level 4 SCP was counted and visualized. Each colored dot corresponds to the number of genes of one SCP of the indicated level.

bioRxiv preprint doi: https://doi.org/10.1101/112201; this version posted March 2, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Figure 4

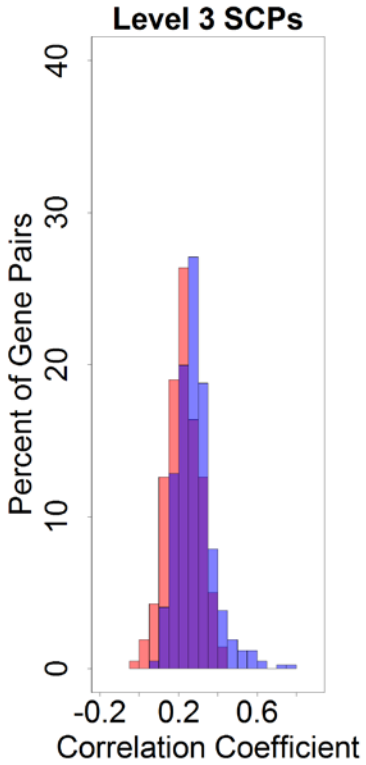


Fig.4: Higher levels of co-expression in various tissues for genes that are part of the same level-3 SCPs. We determined the Pearson correlation coefficient between any two genes that belong to level-3 SCPs for their expression in 30 different tissues using the GTEx data. We distributed the correlations into two groups: The first group (blue) contained all correlated gene pairs that were part of at least one level-3 SCP (15,024 pairs), the second group (red) all other correlated gene pairs (6,719,761 pairs).For ease of visualization both groups were normalized to 100 and percentage of pairs for the various correlation coefficients are shown. The Kolmogorov-Smirnov test shows that gene expression correlations significantly differ between the two groups (p-value = 4.885e-15).

bioRxiv preprint doi: https://doi.org/10.1101/112201; this version posted March 2, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Figure 5

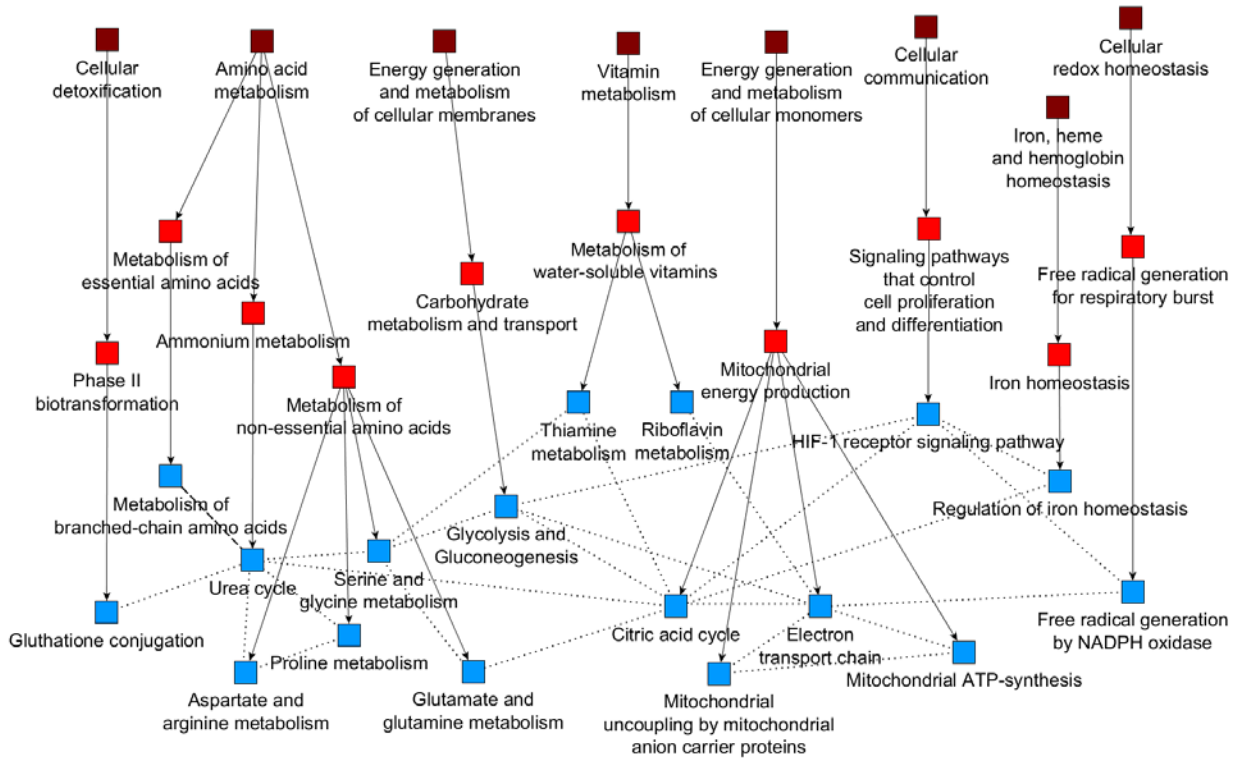
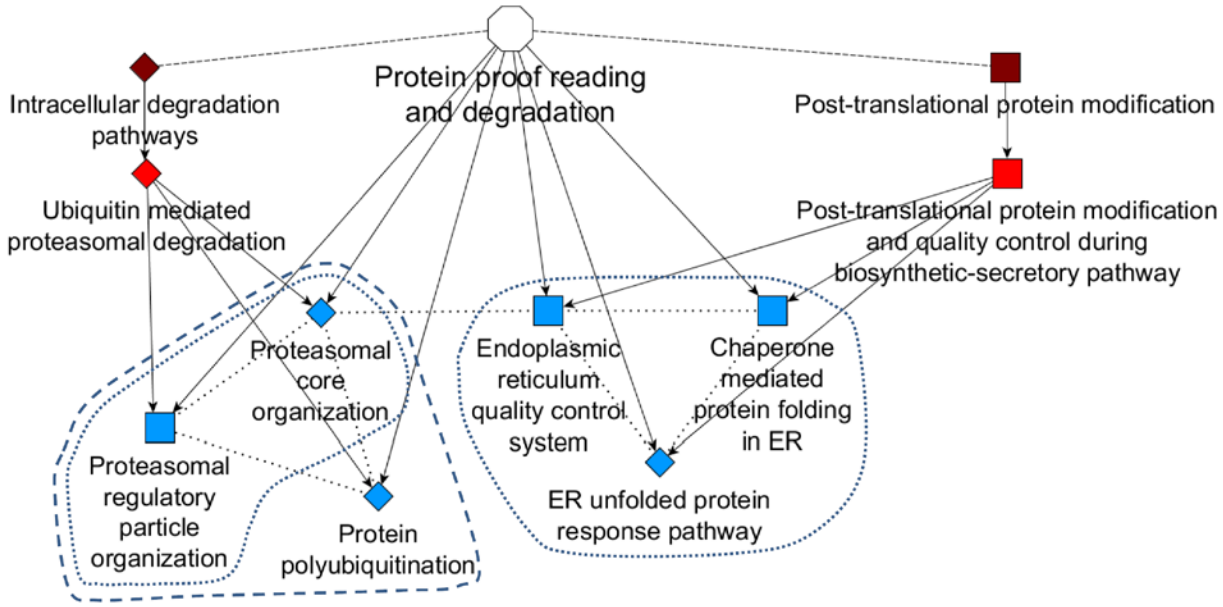


Fig.5: Algorithm to connect level-3 SCPs beyond family relationships defined by the initial taxonomy. One level-3 SCP at a time was removed from the ontology and the remaining level-3 SCPs were re-populated. Each of the remaining SCPs was analyzed, if it contained genes of the removed SCP, either as a new member or as a member that showed increase in rank. Changes were quantified and results associated with SCP pairs. The top 25% predicted SCP-SCP interactions were considered. The newly identified SCP interactions are shown by dotted lines while the original taxonomy is shown as solid arrows. All predicted interaction partners of three SCPs, Urea cycle, Citric acid cycle, and Electron transport chain are shown. To illustrate the hierarchical relationships all level-2 parent SCPs (light red) and level-1 grandparent SCPs (dark red) are also shown.

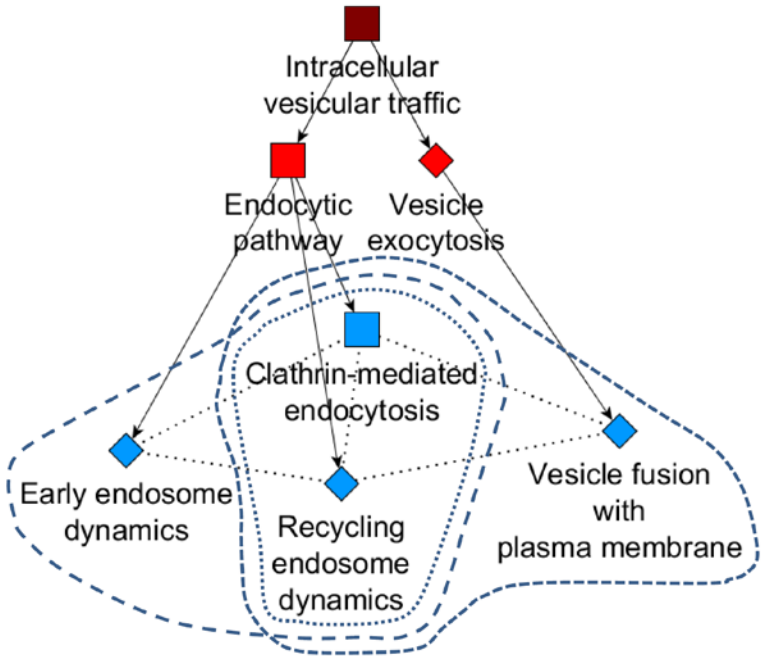
bioRxiv preprint doi: https://doi.org/10.1101/112201; this version posted March 2, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Figure 6

a

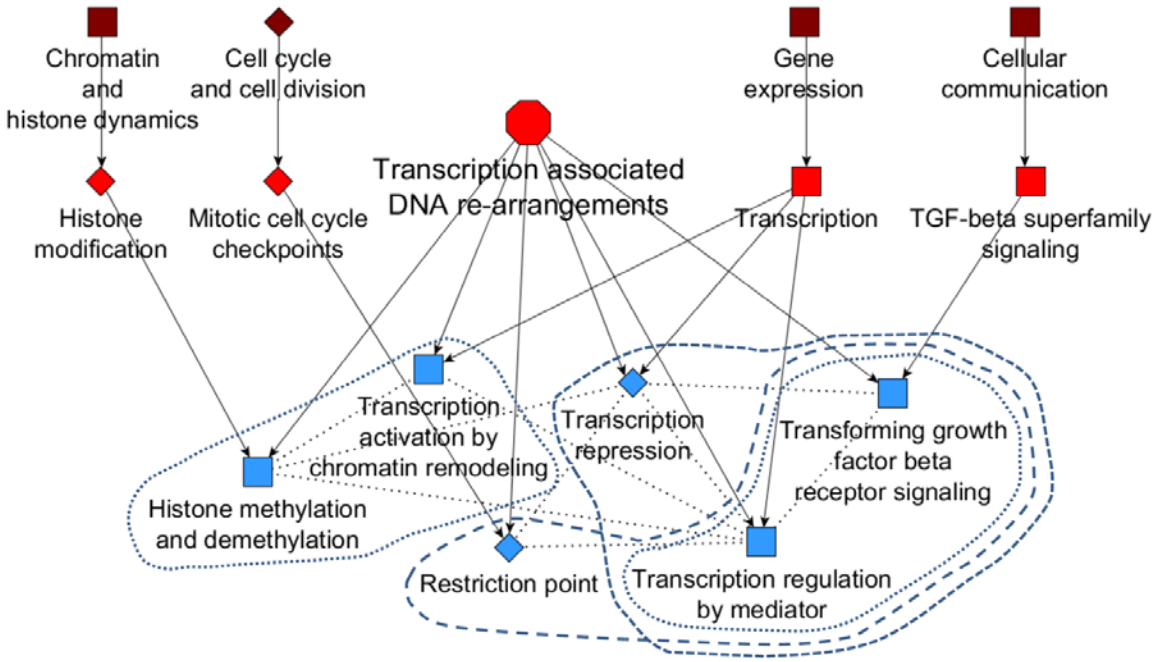


b

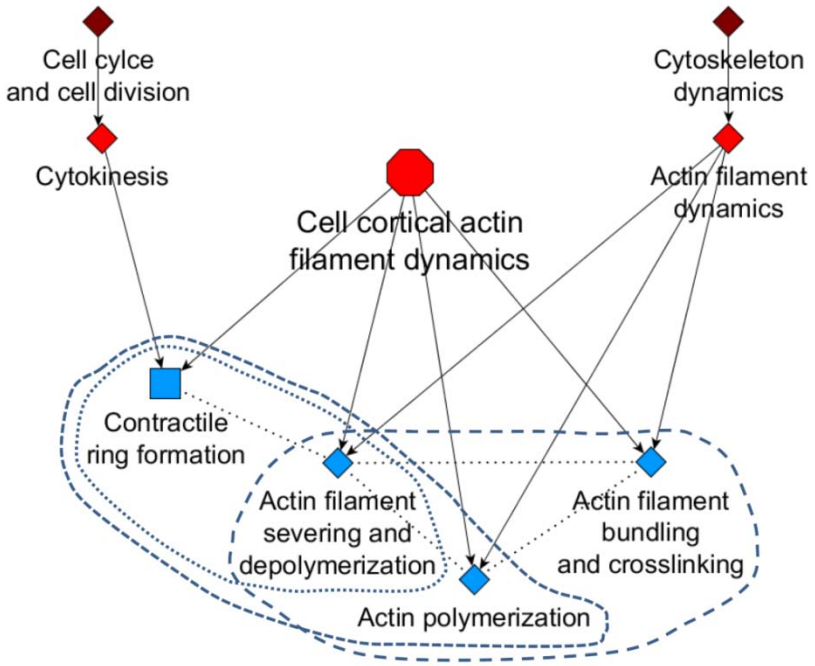


bioRxiv preprint doi: https://doi.org/10.1101/112201; this version posted March 2, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

c



d



bioRxiv preprint doi: https://doi.org/10.1101/112201; this version posted March 2, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Fig. 6: Dynamic enrichment analysis identifies context relevant higher level SCPs emerging from interactions of level-3 SCPs. Three case studies using experimental high throughput datasets from literature are used to demonstrate the ability of the MBC Ontology to identify SCPs relationships that give rise to the whole cell function that was studied in each case. SCP units were generated by merging any two or three level-3 SCPs that were related to each other based on the top 25% predicted SCP relationships and contained at least one experimentally perturbed gene. The merged SCPs were added to the original level-3 SCPs and the experimental data was analyzed for enrichment analysis using the new group of level-3 SCPs by the Fisher's exact test. All level-3 SCPs (blue rectangles and diamonds) that were among the top 5 predictions (either as a single SCP or as a SCP unit) were connected with each other based on the predicted SCP-SCP relationships (dotted black lines). The largest SCP network was kept and assigned to a context specific higher level SCP (open [unspecified level] or light red [level-2] octagon). Annotated level-1 grandparents (dark red rectangles) and level-2 parents (light red rectangles) were added to demonstrate that the combined level-3 SCPs belong to different SCP families. Arrows connect parent SCPs with their children SCPs. Blue lines encircle level-3 SCPs that were identified as part of a new merged SCP unit. Squares indicate level-1 to level-3 SCPs that were identified by the original taxonomy. Diamonds indicate level-1 to level-3 SCPs that were only identified by the dynamic enrichment approach that enables extension of the original taxonomy. (A) Dynamic enrichment analysis identifies the inferred context relevant higher-level SCP Protein proofreading and degradation as the major disease mechanism that reduces CFTR activity at the plasma membrane in cystic fibrosis. (B) The context-specific SCP Vesicle Recycling was identified as a major requirement for secretion. (C) Transcription associated DNA re-arrangements were identified as a major mechanism by which erlotinib increases sensitivity of breast cancer cells to doxycyclin. (D) Cell cortical actin filament dynamics were predicted to be responsible for reduction in colony formation activity of erlotinib treated breast cancer cells. See also suppl. figures 20 c/d, 21 c/d and 22 c/d/g/h.