for several different thresholds and then used the receiver operating characteristic (ROC) curve to determine if changing the threshold affected CTen's performance in terms of the true positive rate versus the false positive rate. As seen in Additional file 2 and Additional file 3, CTen's performance is robust to the precise threshold used for developing the HECS gene databases.

The HECS genes are highly unique to each cell type

We also determined the percentage of HECS genes shared by any two cell types within the human and mouse databases. As seen in Figure 3, the vast majority of cell types have highly distinct sets of HECS genes, with two mouse cell types sharing an average of only 16.1% HECS genes, while human cell types share an average of 11.6%. The two groups of cell types which share the most HECS genes in both mouse and human datasets belong to the nervous and reproductive systems (denoted by red and purple ticks beside the heatmap in Figure 3). Immune cells in different cell states also share the majority of their HECS genes (e.g., human CD8+ T-cells and CD4+ T-cells share 90.4% of their HECS genes) but the number of HECS genes shared between two different immune cells (e.g., B-cells versus T-cells) is generally less than 50% (Additional file 4 provides a more detailed heatmap). In all, the strategy behind the development of the HECS database ensures that HECS genes are limited to a few cell types - characterizing a signature for each tissue. Therefore, the HECS database provides a powerful means of identifying cell/tissue specific enrichment in user gene lists.

Data preprocessing and calculating the enrichment score

A minimal amount of preprocessing is applied to the user supplied gene list to ensure that, first, the list is properly parsed, and second, the user supplied genes are found in the HECS database. The workflow of the CTen website is shown in Figure 4A. At the upload screen (Figure 4B), users can upload a list of either gene symbols or Entrez gene IDs, and optionally upload multiple lists at once by choosing the appropriate format (the CTen webpage provides a single and multi-list example). The gene list is processed to determine the number of unique user genes found in the database and if the list does not appear to be one of the two gene identifiers stated above or the inappropriate format was selected, the website shows a parsing error screen and asks the user to ensure that the proper identifier is selected. If there are no parsing errors, CTen produces a table showing the number of unique user genes mapped in the CTen database for each uploaded list (Figure 4C). If no user genes are found in the database, CTen produces another error, "*No genes found in the database*" and the user is asked to reevaluate the uploaded gene list. Should CTen not detect either of these errors, the option to continue to enrichment appears and the user can complete their analysis.

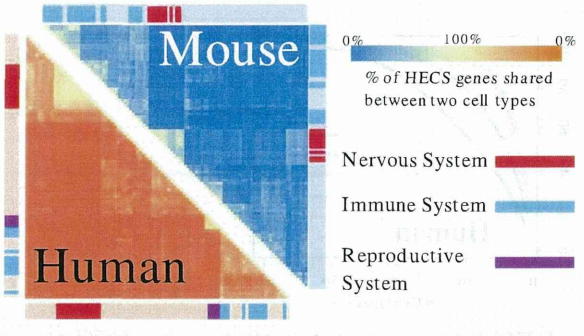

Figure 3 HECS genes shared between different cell types. A heatmap of the percentage of HECS genes shared between any two cell types in the mouse (upper triangle) and human (lower triangle) databases. Ticks adjacent to the heatmap denote cell types belonging to the nervous (red), immune (blue), or reproductive systems (purple).

Using the one-sided, Fisher's Exact test for enrichment, the enrichment score returned from CTen is the -log10 of the Benjamini-Hochberg (BH) adjusted P-values (all calculations are performed in R [17]). Although the enrichment score is a statistic in origin (indeed the enrichment scores could be used to control the false discovery rate), we advise users to consider the enrichment score to be a ranking and to not apply a strictly statistical understanding of the number. This is due to the sensitivity of the score to the size of the gene list being analyzed, and we show in detail in the Results and Discussion that ranking the results allows for easier interpretation. The appropriate contingency tables are constructed using the intersection of the user list and the HECS genes for each cell type. The gene universe (or gene background) against which the enrichment is calculated is currently defined to be all of the genes annotated in the human or mouse arrays defined above. Importantly, the enrichment scores for each gene list are calculated separately.

When only a single list is processed, a radar map of the enrichment scores is produced but in the case of multiple gene lists being supplied, P-values between each list cannot be compared since the length of the gene lists differ. So we developed a "weighted-ranking" strategy in which the enrichment scores for the 10 most enriched cell types in each list are scaled by the maximum enrichment score for that list. The enrichment scores of cell types either not present in the top 10 or present in the top 10 but with enrichment score of less than 2 are excluded. This procedure selects only the most enriched cell types for each list and allows us to visualize whether the enrichment scores of the top cell types were similar

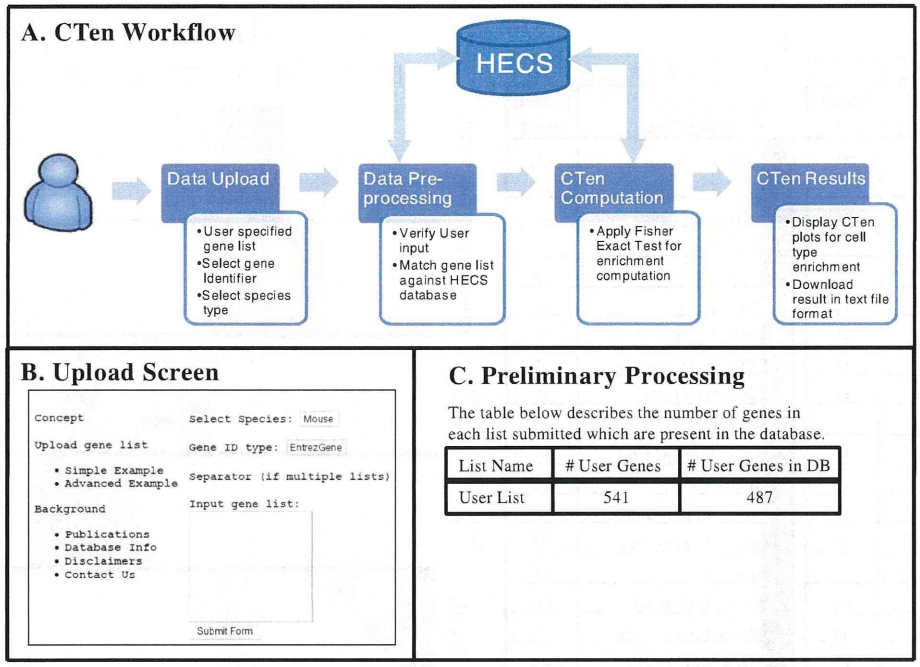

Figure 4 Overview of a CTen session. CTen has been designed with a user-friendly interface to allow for rapid analysis of several gene lists, simultaneously. Panel (**A**) illustrates the workflow between the user, CTen interface and the HECS database. (**B**) At the upload screen, users can copy and paste their gene lists straight from spreadsheet software (e.g., Excel) and select the appropriate parameters (species, gene identifier and the separator used if multiple lists are being uploaded). (**C**) The data goes through preliminary processing to ensure the gene list(s) is parsed properly and that the supplied genes are found in the HECS database. Once the user list passes preliminary processing, the enrichment of all cell types in the HECS database is calculated.

or if one cell type's enrichment score was dominant. The influenza-infected lung tissue example and the advanced use-case in the Results and Discussion illustrate CTen's output for single and multi-list analyses.

Finally, for both single- and multi-lists analyses, the final enrichment scores for all cell types can be downloaded for further processing by the users.

Results and discussion

CTen correctly identifies cell types

To assess CTen's ability to identify the correct cell type associated with gene expression data, we used an independent database of cell-specific gene expression (GNF1M_plus_macrophage_small dataset from BioGPS; abbreviated GNFM1) to develop several lists of genes which were highly expressed in select cell types. This data set is an interesting test case for CTen because the differences in the experimental protocol tests CTen's performance when using different microarray technologies and biological conditions. In the GNF1M experiment, they used mice which were ~2 weeks older (compared to the mice used to develop the Mouse MOE430 Gene Atlas data set), used a different ratio of male and female mice, and employed custom microarray slides (GPL1037) [3]. For several cell types (5 tissues and 3 lymphocytes; 2 lymphocytes in different cellular states), we selected the top 5% of the most highly expressed probes. Entrez Gene IDs were mapped using the annotation files available from BioGPS, and the resulting lists analyzed in CTen.

We found that CTen consistently ranked the correct cell type the highest for each tissue tested (Figure 5A) and, with the exception of bone marrow, there was a large difference in the scores between the first and second most enriched cell types. Not surprisingly, bone marrow was identified as being highly enriched for bone. For the lymphocyte gene lists (Figure 5B), CTen not only identified the correct lymphocyte but most often identified the correct cellular state of the lymphocyte as being the top ranked cell type. Only for the unstimulated macrophages did CTen rank the inappropriate cellular states the highest. Thus, from independent, cell-specific gene expression data, we confirmed that CTen provides clear guidance in relating gene expression data to the appropriate cell type.

Ranking of the enrichment scores are robust

As with any analysis, small changes in experiment parameters should not greatly change the outcome. P-values from the Fisher Exact test are very sensitive to changes in the size of the gene list, but for many enrichment analyses, it has been observed that the rankings of the

A

	CTen Results	
	Cell-type/Tissue	Enrich ment
Lung	Lung	88
	Ovary	22
	Uterus	21
	Heart	21
	Bladder	17
Bone Marrow	Bone Marrow	45
	Bone	43
	Spleen	15
	Follicular B Cells	14
	B Cells (GL7 pos; Alum)	9
Spleen	Spleen	34
	Follicular B Cells	23
	Lymph Nodes	19
	B Cells (GL7 neg; Alum)	16
	B Cells (GL7 pos; Alum)	15
Liver	Liver	261
	Kidney	44
	Adipose Brown	17
	Lung	15
	Adrenal Gland	11
Kidney	Kidney	143
	Liver	38
	Adipose White	9
	Adipose Brown	8
	Adrenal Gland	7

B

	CTen Results	
	Cell-type/Tisse	Enrich ment
CD4+ T Cell	CD4+ T cells	42
	CD8+ T cells	40
	Follicular B Cells	40
	CD4+ SP Thymoctyes	37
	Foxp3+ Tcells	34
CD8+ T Cell	CD8+ T cells	47
	CD4+ T cells	43
	CD4+ SP Thymoctyes	37
	Follicular B Cells	37
	NK Cells	34
Unstim. Macro.	Macro. BM Lps 6Hrs	82
	Macro. BM	80
	Macro. BM Lps 24 Hrs	80
	Macro. BM Lps 2Hrs	75
	Microglia	61
Macro. (LPS; 7h)	Macro. BM Lps 6Hrs	159
	Macro. Peri Lps 7Hrs	116
	Macro. BM Lps 2Hrs	105
	Macro. BM Lps 24 Hrs	101
	Macro. BM	71
B Cell	Follicular B Cells	110
	B Cells (GL7 pos; KLH)	83
	B Cells (GL7 pos; Alum)	82
	B Cells (GL7 neg; Alum)	78
	B Cells (GL7 neg; KLH)	76

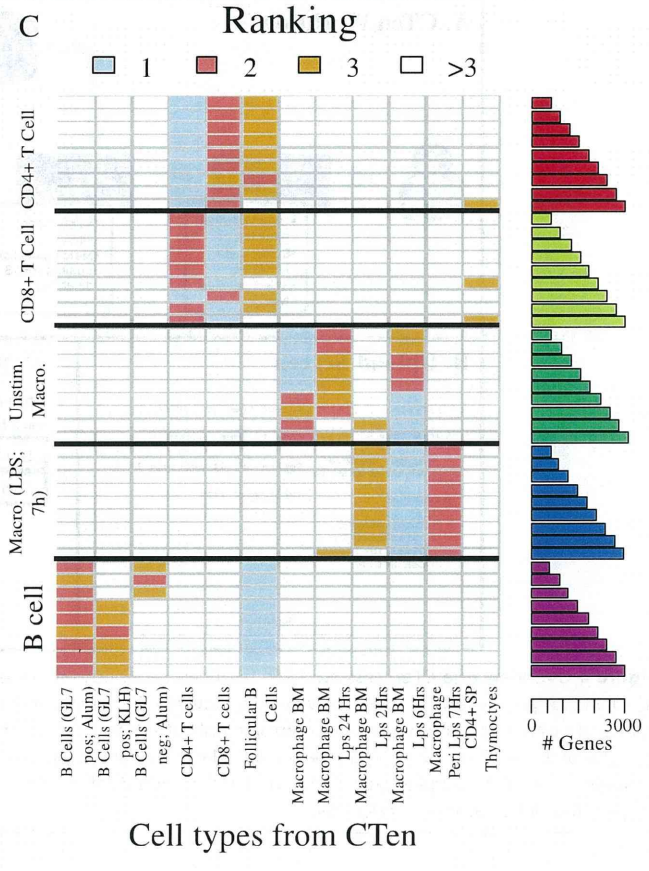

Figure 5 CTen validation and robustness of the enrichment score rankings. From an additional cell-specific gene expression database, we took the top 5% most highly expressed genes for five kinds of tissue (lung, bone marrow, kidney, spleen and liver) and three lymphocytes in different cellular states ($CD4^+$ and $CD8^+$ T cells, B cells and macrophages that were unstimulated or collected 7 hours after simulation with LPS). Each list was analyzed in CTen and we show the top 5 enrichments scores for (**A**) the tissue (**B**) lymphocyte test lists. To evaluate the robustness of these results, we repeated this procedure for the top 2-10% most highly expressed genes in each cell type. The enrichment scores from CTen were ranked from highest to lowest, and (**C**) the heatmap illustrates the top 3 most enriched cell types identified by CTen (columns) for each lymphocyte data tested (row labels). The bar plot on the right hand side summarizes the number of genes per test list. The heatmap for the tissue data is available as Additional file 5.

enrichment scores are very robust [7,18]. Here, we asked if CTen could robustly rank the correct cell types by repeating the procedure described above - now using a list of the top 2, 3,. . ., 10% most highly expressed genes for the selected tissues and lymphocytes, resulting in 90 test lists. The different sizes of the lists simulate different differential expression criteria during gene expression analysis. As shown in Figure 5C, although the sizes of the gene lists (and the underlying enrichment scores) vary considerably, CTen most often ranks the appropriate cell type the most highly. CTen was also able to identify the proper cell state of the lymphocytes as well although unstimulated macrophage data was assigned to *bone marrow macrophages collected 6 h after exposure to lipopolysaccharide (LPS)* 4 out of 9 times. CTen performed even better for the tissue data, always ranking the appropriate tissue the highest (Additional file 5). In all, CTen can accurately identify a broad range of cell types and very often identify the cellular state as well. The results are very robust to changes in the length of the test data, which can be equated to changes in the cutoff criteria used during microarray analysis.

Minimizing the false positive rate

While CTen accurately identified the appropriate cell type as having the highest enrichment score, we think it's important to provide a comprehensive analysis of CTen's accuracy for select cutoff values of the enrichment score. Using the same test lists developed above for Figure 5C, we used the receiver operating characteristic (ROC) curve to identify what level of enrichment was necessary to maximize the sensitivity (true positive rate, TPR) while minimizing the false positive rate (FPR) (Figure 6). Demanding a minimal enrichment score of 2

provides a low FPR and, indeed, we found that for randomly generated lists of genes, CTen rarely assigned scores above 2 (Additional file 6). But we see here, raising the enrichment score cutoff from 2 to 25 greatly minimizes the FPR without sacrificing the TPR. Requiring enrichment scores above 25 only reduces the sensitivity of the analysis. A similar analysis to this was performed using the two databases from which CTen was constructed resulting in nearly identical ROC curves (Additional file 2 and Additional file 3). These curves also suggest enrichment scores of 20–25 to optimally minimize the FPR for mouse data, but slightly lower enrichment scores (15 to 20) offer optimal performance for human data. It should be noted that these performance results are dependent on the size of the gene list. Thus, for gene lists which are hundreds to thousands of genes in number, a minimum enrichment score of 2 is recommended, but scores of 20–25 appear to offer optimal performance.

CTen versus GO analysis of influenza infected lung tissue

Using a list of genes found to be upregulated in lung tissue collected from mice infected with influenza virus (microarray data unpublished; the gene list is available on the CTen website under the "Simple Example" tab), we compared the results of a CTen analysis to a GO analysis using DAVID [7]. Using the CTen website, we find a very high enrichment of bone marrow derived and peritoneal macrophages (Figure 7A), both of which have been exposed to lipopolysaccharide (LPS) and collected at different time points. Macrophage migration to the site of infection is one of the first steps in coordinating the innate immune response [19]. Both LPS exposure [20] and influenza infection [21] induces the activation of the Toll-like receptor pathways, and macrophages are often susceptible to influenza infection themselves [22]. Thus, an increase in macrophage numbers is consistent with previously published studies [23] and the observation of the resulting cell type as "*macrophage exposed to LPS*", indicates that the macrophages have possibly become infected with the influenza virus as well.

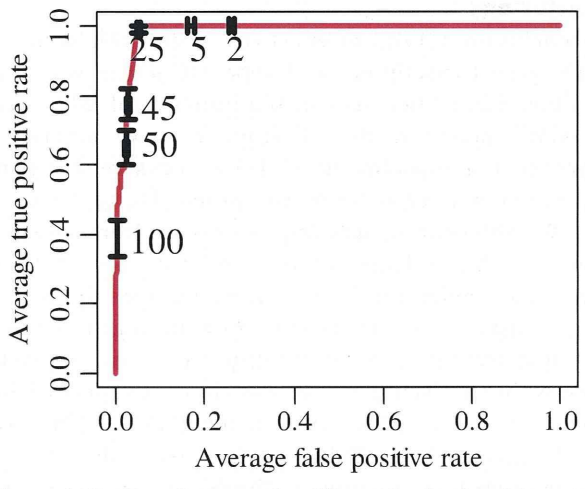

Figure 6 CTen's performance for different levels of enrichment. Using the same test lists behind the results shown in Figure 5C, we constructed an ROC curve to evaluate CTen's classification performance for different levels of the enrichment score. The error bars depict the 95% confidence interval of the ROC curve for the enrichment scores shown.

DAVID uses modules of related biological terms to interpret large gene lists into a meaningful biological context and reports the scores of each module as the -log10 of the average P-value for each term within the module [24]. Using the default settings, DAVID identifies the Toll-like receptor pathway (Figure 7B, Cluster #1) as the most significant cluster of annotations (Enrichment score: 12.62; full results available in Additional file 7). However, the clusters indicating enhanced macrophage presence have a low significance (Cluster #29; enrichment score: 1.74) and are very closely followed by a T-cell related cluster (Enrichment score: 1.68). Taken together, these results indicate that although both analyses can identify aspects of the cellular state of the sample, CTen is better suited to identify the known changes in the cellular demographics of the RNA samples.

Advanced use-case: distinguishing changes in lymphocyte cell count from gene transcription

The most exciting potential of CTen is that, when applied to clustering studies, cell type enrichment analysis can be used to approximate the evolution of local cellular demographics. Our laboratory's research is primarily focused on reconstructing the host response during an influenza infection [25]; a goal which requires us to be able to integrate local intracellular signaling (Toll-like receptor/RIG-I/NFκB pathways) with the coordinated migration, infiltration, and activity of macrophages, T-cells, B-cells, and other immune related cell-types. Being able to resolve the various cell types present in a sample from microarray data would greatly facilitate discovery in a broad range of *in vivo* studies.

Figure 8 illustrates the proposed strategy for identifying cellular signatures in *in vivo* data and its implications for *in vivo* microarray based studies. In this illustration, microarray data was assembled over a span of 5 days from the lungs of mice infected with influenza virus on day 0 (lung tissues are illustrated in Figure 8A). After normalizing and differential expression testing, four gene clusters (Figure 8B) were identified using the user's preferred clustering tool.

In this case, we are illustrating potential results from using the WGCNA package [15], which applies color labels to each cluster. The genes for each cluster can be uploaded and analyzed in one session to identify the

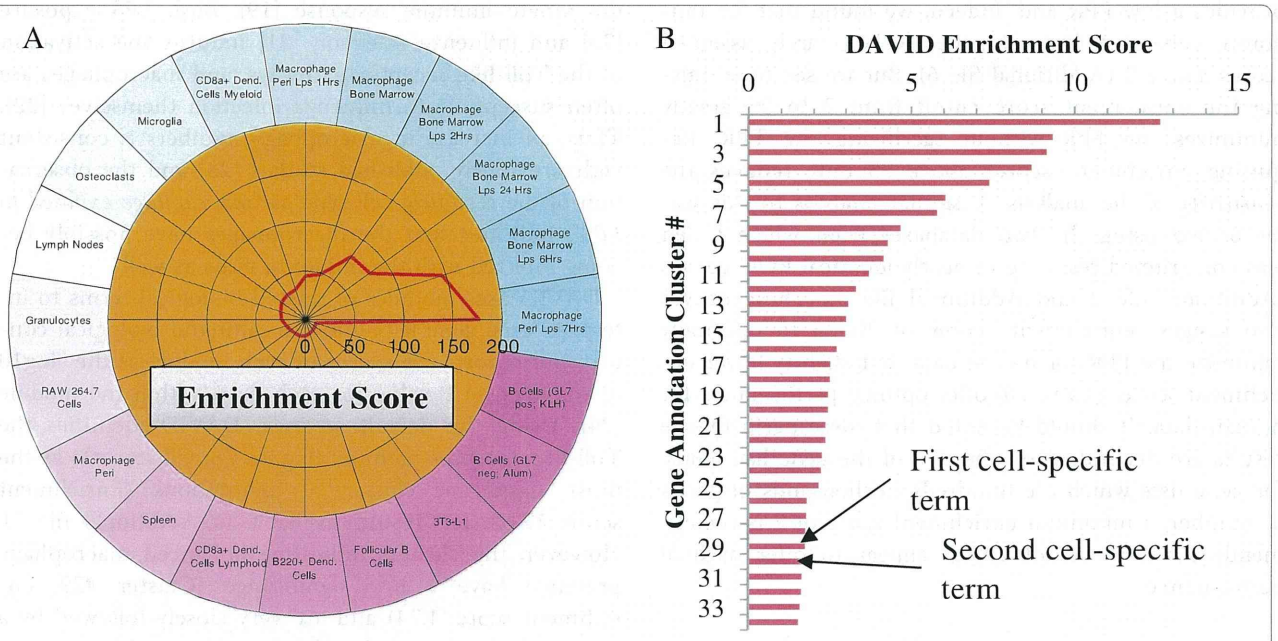

Figure 7 CTen versus GO analysis. A list of upregulated genes in lung tissue collected from mice infected with the influenza virus is analyzed in (**A**) CTen and (**B**) DAVID. The first cluster to have a cell specific term is ranked 29th in the DAVID analysis. A complete list of the terms belonging to each cluster is available in Additional file 7.

most enriched cell types in each cluster. In Figure 8C, we find that macrophages are highly enriched in the dark red cluster while several categories of B- and T-cells (CD8+ T-cells) are the most enriched cell types in the green and black clusters, respectively. Interestingly, the orange cluster is not enriched for any cell type, and we would conclude that transcripts in the orange cluster represent differential gene expression due to transcriptional differences between the samples (as opposed to difference in the cellular makeup of the samples) and are suitable for further analysis using traditional approaches. The dark red, green and black clusters can be further analyzed for pathway or functional enrichment to identify processes that may be coordinated with cell migration. This result may also help researchers decide the appropriateness of additional analyses. Some analyses, such as gene network inference, will have to carefully consider how to remove the effects of cell migration prior to network construction. Furthermore, the green, black and dark red clusters' gene expression is highly correlated to the corresponding lymphocyte's cell count change. Thus, we may be able to infer the relative changes in the B cell, T cell and macrophage count in the infected tissue.

In all, this example illustrates how CTen has been designed to facilitate the understanding of clustering results by identifying conserved expression patterns that are the result of changes in the numbers of a particular cell type, providing critical guidance for selecting additional analyses for each gene set and allowing users to infer changes in cellular demographics between samples. Based on the CTen enrichment platform, we propose a novel analytical workflow for *in vivo* microarray, as illustrated schematically in Figure 8D, which ensures that enriched biological pathways and processes identified in a set of differentially expressed genes can be interpreted in the proper cellular context.

Conclusions

In conclusion, CTen can effectively distinguish between active gene transcription and apparent gene expression resulting from differences in the numbers of select cell types in microarray data. Furthermore, we provide a novel research workflow which helps to ensure that gene expression is interpreted in the proper biological context. We will continuously improve the enrichment algorithm so that a larger number of gene lists can be processed simultaneously (currently, users are limited to 20 gene lists in a single session). Recently, a gene set enrichment analysis based on the degree of pairwise correlation within a given gene set was shown to successfully relate samples to their corresponding tissue [26]. No simple interface is available yet for researchers, but it will be interesting to compare the performance between these two approaches in the near future. Additionally, we plan to introduce additional cell specific gene expression datasets so users can compare the results from different databases. And finally, while the examples focused on lymphocyte migration, CTen can be used in several

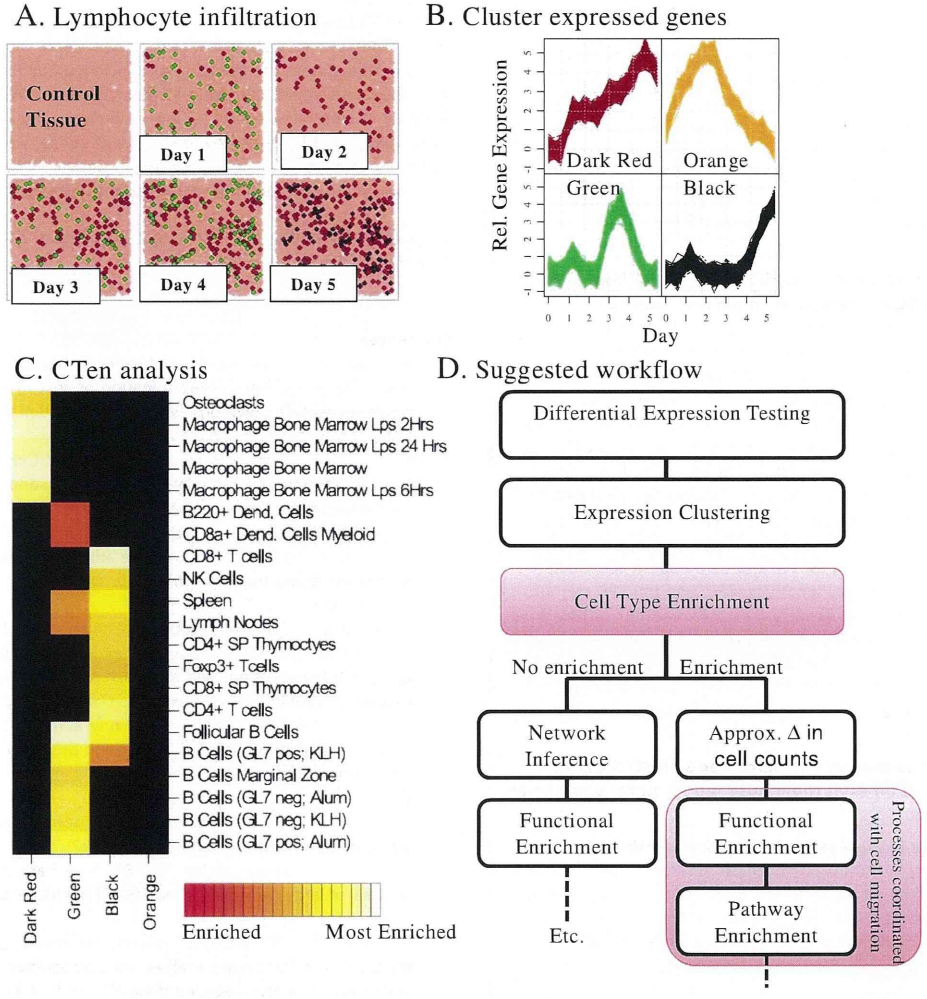

Figure 8 Detecting changes in lymphocyte cell counts and the appropriateness of follow-up analyses using CTen. (**A**) In infected tissue, as B cells (green), T cells (black), and macrophages (dark red) migrate into the sample area, they bring with them their own unique quantities of RNA, resulting in conserved expression patterns proportional to the cell type's increased (or decreased) presence. Here, we illustrate the potential results from a clustering study using WGCNA [15], resulting in (**B**) 4 clusters with unique temporal expression profiles; 3 of which are highly correlated to the changes in the number of lymphocytes shown in (A). (**C**) Analyzing the genes present in each cluster, CTen can distinguish which clusters represent gene expression and which represent cell migration, and determine the cell type responsible for the observed gene regulation. Using a "weighted-ranking" strategy (see text), CTen produces a heatmap showing the most enriched cell types for each cluster. Based on the CTen result, only the orange cluster is appropriate for further analysis using traditional bioinformatic techniques while the green, black, and dark red clusters reflect the relative changes in the number of lymphocytes during the infection. (**D**) We propose a new analytical workflow strategy which ensures that continued analysis of *in vivo* microarray data properly identifies events which may be coordinated with (or coordinating) cell migration.

other scenarios; for example, comparing excised tissue to ensure homogeneity between tissue samples.

Availability and requirements

Project name: CTen

Project home page: http://www.influenza-x.org/~jshoemaker/cten/

Operating system: Platform independent

Programming Language: PHP and R

Other requirements: None

License: EULA

Additional files

Additional file 1: A list of the cell types currently available in CTen.

Additional file 2: The enrichment performance of the mouse HECS database for select HECS criteria and enrichment scores. We evaluated (1) does the precise cutoff for defining a HECS gene affect the enrichment performance and (2) for each cutoff, what values of the enrichment score seems to best minimize the false positive rate (FPR) without impacting the true positive rate (TPR). We reconstructed the HECS database by defining the HECS assignment threshold as (A) 5, (B) 10, (C) 15, and (D) 20 times the median. Then, from the Mouse MOE430 Gene Atlas dataset, we took the top 10% of the most highly expressed

genes for each cell type. From this 10%, we randomly sampled between 500 to 4000 genes 3 times to create 288 gene lists. Using the same procedures described in the CTen implementation, these lists were analyzed for cell type enrichment for each HECS database constructed. The ROC curve illustrates the that sensitivity (TPR) and the FPR are not greatly affected by the HECS assignment threshold selected. Furthermore, on each figure, we show the performance expected for selected values of the enrichment score. We see that selecting enrichment scores of 2 or higher results in a reasonably low FPR but this can be significantly improved by demanding enrichments scores of ~25 before the TPR is affected.

Additional file 3: The enrichment performance of the human HECS database for select HECS criteria and enrichment scores. We evaluated (1) does the precise cutoff for defining a HECS gene affect the enrichment performance and (2) for each cutoff, what values of the enrichment score seems to best minimize the false positive rate (FPR) without impacting the true positive rate (TPR). We reconstructed the HECS database by defining the HECS assignment threshold as (A) 5, (B) 10, and (C) 15 times the median. Then, from the Human U133A/GNF1H Gene Atlas dataset, we took the top 10% of the most highly expressed genes for each cell type. From this 10%, we randomly sampled between 500 to 4000 genes 3 times to create 252 gene lists. Using the same procedures described in the CTen implementation, these lists were analyzed for cell type enrichment for each HECS database constructed. The ROC curve illustrates the that sensitivity (TPR) and the FPR are not greatly affected by the HECS assignment threshold selected. Furthermore, on each figure, we show the performance expected for selected values of the enrichment score. We see that selecting enrichment scores of 2 or higher results in a reasonably low FPR but this can be significantly improved by demanding enrichments scores of ~20 before the TPR is affected.

Additional file 4: A heatmap of the percentage of HECS genes shared by any two cell types in the mouse (upper right) and human (lower left) databases.

Additional file 5: The highest ranked cell types identified by CTen. Using the GNF1M_plus_macrophage_small dataset from BioGPS, the top 2-10% most highly expressed genes for the tissues shown were analyzed in CTen. The enrichment scores from CTen were ranked from highest to lowest, and the heatmap illustrates the top 3 most enriched cell types (columns) for each lymphocyte data tested (row labels). BM = bone marrow.

Additional file 6: Expected enrichment scores for random gene lists. We analyzed in CTen 150 lists of 100–400 randomly selected IDs for (A) mouse and (B) human Entrez Gene IDs - this resulted in a distribution of enrichment scores. The distributions were fit to a gamma distribution using the MASS package in R. Here, we show the density histogram and fitted gamma function (left hand axis) and the probability distribution function (right hand axis). The red bar highlights the enrichment score which is 95% confidently above 0 ($\alpha = 0.95$ at enrichment scores of 1.66 and 1.67 in the mouse and human data, respectively).

Additional file 7: A list of genes upregulated in mouse lungs which have been infected with influenza virus and the full results of analyzing this list in DAVID.

Competing interests
No competing interests to declare.

Authors' contributions
JES designed the project, built the database and wrote the manuscript. TJL, SG, YK and HK revised the manuscript. JES, TJL, SG, YM, and HK implemented the website and YM maintains the public server. All authors have read and approved the final manuscript.

Acknowledgements
This work was supported by the Japanese Science and Technology Agency's ERATO influenza induced host responses project.

Author details
[1]JST ERATO KAWAOKA Infection-induced Host Responses Project, Tokyo, Japan. [2]The Systems Biology Institute, Tokyo, Japan. [3]Influenza Research Institute, Department of Pathobiological Sciences, School of Veterinary Medicine, University of Wisconsin-Madison, Madison, Wisconsin, USA. [4]Institute of Medical Science, Division of Virology, Department of Microbiology and Immunology, University of Tokyo, Tokyo, Japan. [5]Sony Computer Science Laboratories, Inc, Tokyo, Japan. [6]Open Biology Unit, Okinawa Institute of Science and Technology, Okinawa, Japan.

Received: 9 December 2011 Accepted: 31 August 2012
Published: 6 September 2012

References
1. Hacia JG, Fan JB, Ryder O, Jin L, Edgemon K, Ghandour G, Mayer RA, Sun B, Hsie L, Robbins CM, *et al*: Determination of ancestral alleles for human single-nucleotide polymorphisms using high-density oligonucleotide arrays. *Nat Genet* 1999, 22(2):164–167.
2. Shao L, Fan X, Cheng N, Wu L, Xiong H, Fang H, Ding D, Shi L, Cheng Y, Tong W: Shifting from population-wide to personalized cancer prognosis with microarrays. *PLoS One* 2012, 7(1):e29534.
3. Su AI, Wiltshire T, Batalov S, Lapp H, Ching KA, Block D, Zhang J, Soden R, Hayakawa M, Kreiman G, *et al*: A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc Natl Acad Sci USA* 2004, 101 (16):6062–6067.
4. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, *et al*: Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci USA* 2005, 102(43):15545–15550.
5. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, *et al*: Gene ontology: tool for the unification of biology. *The Gene Ontology Consortium. Nat Genet* 2000, 25(1):25–29.
6. Chang JT, Nevins JR: GATHER: a systems approach to interpreting genomic signatures. *Bioinformatics* 2006, 22(23):2926–2933.
7. da Huang W, Sherman BT, Lempicki RA: Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* 2009, 4(1):44–57.
8. Kaimal V, Bardes EE, Tabar SC, Jegga AG, Aronow BJ: ToppCluster: a multiple gene list feature analyzer for comparative enrichment clustering and network-based dissection of biological systems. *Nucleic Acids Res* 2010, 38:W96–102.
9. Baas T, Baskin CR, Diamond DL, Garcia-Sastre A, Bielefeldt-Ohmann H, Tumpey TM, Thomas MJ, Carter VS, Teal TH, Van Hoeven N, *et al*: Integrated molecular signature of disease: analysis of influenza virus-infected macaques through functional genomics and proteomics. *J Virol* 2006, 80(21):10813–10828.
10. Schwartz R, Shackney SE: Applying unmixing to gene expression data for tumor phylogeny inference. *BMC Bioinforma* 2010, 11:42.
11. Shen-Orr SS, Tibshirani R, Khatri P, Bodian DL, Staedtler F, Perry NM, Hastie T, Sarwal MM, Davis MM, Butte AJ: Cell type-specific gene expression differences in complex tissues. *Nat Methods* 2010, 7(4):287–289.
12. McCall MN, Uppal K, Jaffee HA, Zilliox MJ, Irizarry RA: The Gene Expression Barcode: leveraging public data repositories to begin cataloging the human and murine transcriptomes. *Nucleic Acids Res* 2011, 39:D1011–1015. Database issue.
13. Wu C, Orozco C, Boyer J, Leglise M, Goodale J, Batalov S, Hodge CL, Haase J, Janes J, Huss JW 3rd, *et al*: BioGPS: an extensible and customizable portal for querying and organizing gene annotation resources. *Genome Biol* 2009, 10(11):R130.
14. Eisenberg E, Levanon EY: Human housekeeping genes are compact. *Trends Genet* 2003, 19(7):362–365.
15. Langfelder P, Horvath S: WGCNA: an R package for weighted correlation network analysis. *BMC Bioinforma* 2008, 9:559.
16. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, *et al*: Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol* 2004, 5(10): R80.
17. Team RDC: *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing; 2010.

18. da Huang W, Sherman BT, Lempicki RA: **Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists.** *Nucleic Acids Res* 2009, **37**(1):1–13.
19. Medzhitov R: **Recognition of microorganisms and activation of the immune response.** *Nature* 2007, **449**(7164):819–826.
20. Aderem A, Ulevitch RJ: **Toll-like receptors in the induction of the innate immune response.** *Nature* 2000, **406**(6797):782–787.
21. Sun L, Liu S, Chen ZJ: **SnapShot: pathways of antiviral innate immunity.** *Cell* 2010, **140**(3):436–436. e432.
22. Yu WC, Chan RW, Wang J, Travanty EA, Nicholls JM, Peiris JS, Mason RJ, Chan MC: **Viral replication and innate host responses in primary human alveolar epithelial cells and alveolar macrophages infected with influenza H5N1 and H1N1 viruses.** *J Virol* 2011, **85**(14):6844–6855.
23. Reading PC, Whitney PG, Pickett DL, Tate MD, Brooks AG: **Influenza viruses differ in ability to infect macrophages and to induce a local inflammatory response following intraperitoneal injection of mice.** *Immunol Cell Biol* 2010, **88**(6):641–650.
24. da Huang W, Sherman BT, Tan Q, Collins JR, Alvord WG, Roayaei J, Stephens R, Baseler MW, Lane HC, Lempicki RA: **The DAVID Gene Functional Classification Tool: a novel biological module-centric algorithm to functionally analyze large gene lists.** *Genome Biol* 2007, **8**(9):R183.
25. Fukuyama S, Kawaoka Y: **The pathogenesis of influenza virus infections: the contributions of virus and host factors.** *Curr Opin Immunol* 2011, **23**(4):481–486.
26. Chang JT: **Deriving transcriptional programs and functional processes from gene expression databases.** *Bioinformatics* 2012, **28**(8):1122–1129.

doi:10.1186/1471-2164-13-460
Cite this article as: Shoemaker *et al.*: CTen: a web-based platform for identifying enriched cell types from heterogeneous microarray data. *BMC Genomics* 2012 **13**:460.

BIOINFORMATICS ORIGINAL PAPER Vol. 28 no. 15 2012, pages 2016–2021
doi:10.1093/bioinformatics/bts270

Systems Biology Advance Access publication May 10, 2012

Software support for SBGN maps: SBGN-ML and LibSBGN

Martijn P. van Iersel[1,2,3,*], Alice C. Villéger[4], Tobias Czauderna[5], Sarah E. Boyd[6], Frank T. Bergmann[7], Augustin Luna[8,9], Emek Demir[10], Anatoly Sorokin[11], Ugur Dogrusoz[12], Yukiko Matsuoka[13], Akira Funahashi[14], Mirit I. Aladjem[15], Huaiyu Mi[16], Stuart L. Moodie[1], Hiroaki Kitano[13,16], Nicolas Le Novère[1] and Falk Schreiber[5,17]

[1]EMBL European Bioinformatics Institute, Hinxton, UK, [2]Netherlands Consortium for Systems Biology (NCSB), Amsterdam, [3]Department of Bioinformatics - BiGCaT, University of Maastricht, Maastricht, The Netherlands, [4]School of Computer Science, Faculty of Engineering and Physical Sciences, University of Manchester, Manchester, UK, [5]Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), Gatersleben, Germany, [6]School of Mathematical Sciences, Faculty of Science, Monash University, Melbourne, Australia, [7]Control and Dynamical Systems, California Institute of Technology, Pasadena, CA, [8]National Cancer Institute, Bethesda, MD, [9]Bioinformatics Program, Boston University, Boston, MA, [10]Computational Biology, Memorial Sloan Kettering Cancer Center, New York, NY, USA, [11]Institute of Cell Biophysics RAS, Puschino, Russia, [12]Computer Engineering Department, Bilkent University, Ankara, Turkey, [13]The Systems Biology Institute, Tokyo, [14]Department of Biosciences and Informatics, Keio University, Yokohama, Japan, [15]Depatment of Preventive Medicine, Keck School of Medicine, University of Southern California, Los Angeles, CA, USA, [16]Okinawa Institute of Science and Technology, Okinawa, Japan and [17]Institute of Computer Sciences, Faculty of Natural Sciences III, University of Halle, Halle, Germany

Associate Editor: Trey Ideker

ABSTRACT

Motivation: LibSBGN is a software library for reading, writing and manipulating Systems Biology Graphical Notation (SBGN) maps stored using the recently developed SBGN-ML file format. The library (available in C++ and Java) makes it easy for developers to add SBGN support to their tools, whereas the file format facilitates the exchange of maps between compatible software applications. The library also supports validation of maps, which simplifies the task of ensuring compliance with the detailed SBGN specifications. With this effort we hope to increase the adoption of SBGN in bioinformatics tools, ultimately enabling more researchers to visualize biological knowledge in a precise and unambiguous manner.

Availability and implementation: Milestone 2 was released in December 2011. Source code, example files and binaries are freely available under the terms of either the LGPL v2.1+ or Apache v2.0 open source licenses from http://libsbgn.sourceforge.net.

Contact: sbgn-libsbgn@lists.sourceforge.net

Received on December 13, 2011; revised on April 24, 2012; accepted on May 1, 2012

1 INTRODUCTION

The Systems Biology Graphical Notation (SBGN, Le Novère *et al.*, 2009) facilitates the representation and exchange of complex biological knowledge in a concise and unambiguous manner: as standardized pathway maps. It has been developed and supported by a vibrant community of biologists, biochemists, software developers, bioinformaticians and pathway databases experts.

SBGN is described in detail in the online specifications (see http://sbgn.org/Documents/Specifications). Here we summarize its concepts only briefly. SBGN defines three orthogonal visual languages: Process Description (PD), Entity Relationship (ER) and Activity Flow (AF). SBGN maps must follow the visual vocabulary, syntax and layout rules of one of these languages. The choice of language depends on the type of pathway or process being depicted and the amount of available information. The PD language, which originates from Kitano's Process Diagrams (Kitano *et al.*, 2005) and the related CellDesigner tool (Funahashi *et al.*, 2008), is equivalent to a bipartite graph (with a few exceptions) with one type of nodes representing pools of biological entities, and a second type of nodes representing biological processes such as biochemical reactions, transport, binding and degradation. Arcs represent consumption, production or control, and can only connect nodes of differing types. The PD language is very suitable for metabolic pathways, but struggles to concisely depict the combinatorial complexity of certain proteins with many phosphorylation states. The ER language, on the other hand, is inspired by Kohn's Molecular Interaction Maps (Kohn *et al.*, 2006), and describes relations between biomolecules. In ER, two entities can be linked with an interaction arc. The outcome of an interaction (for example, a protein complex), is considered an entity in itself, represented by a black dot, which can engage in further interactions. Thus ER represents dependencies between interactions, or putting it differently, it can represent which interaction is necessary for another one to take place. Interactions are possible between two or more entities, which make ER maps roughly equivalent to a hypergraph in which an arc can connect more than two nodes. ER is more concise than PD when it comes to representing protein modifications and protein interactions, although it is less capable when it comes to presenting biochemical reactions. Finally, the third language in the SBGN family is AF, which

*To whom correspondence should be addressed.

represents the activities of biomolecules at a higher conceptual level. AF is suitable to represent the flow of causality between biomolecules even when detailed knowledge on biological processes is missing.

Efficient integration of the SBGN standard into the research cycle requires adoption by visualization and modeling software. Encouragingly, a growing number of pathway tools (see http://sbgn.org/SBGN_Software) offer some form of SBGN compatibility. However, current software implementations of SBGN are often incomplete and sometimes incorrect. This is not surprising: as SBGN covers a broad spectrum of biological phenomena, complete and accurate implementation of the full SBGN specifications represents a complex, error-prone and time-consuming task for individual tool developers. This development step could be simplified, and redundant implementation efforts avoided, by accurately translating the full SBGN specifications into a single software library, available freely for any tool developer to reuse in their own project. Moreover, the maps produced by any given tool usually cannot be reused in another tool, because SBGN only defines how biological information should be visualized, but not how the maps should be stored electronically. Related community standards for exchanging pathway knowledge, namely BioPAX (Demir *et al.*, 2010) and SBML (Hucka *et al.*, 2003), have proved insufficient for this role (more on this topic in Section 4). Therefore, we observed a second need, for a dedicated, standardized SBGN file format.

Following these observations, we started a community effort with two goals: to encourage the adoption of SBGN by facilitating its implementation in pathway tools, and to increase interoperability between SBGN-compatible software. This has resulted in a file format called SBGN-ML and a software library called LibSBGN. Each of these two components will be explained separately in the next sections.

2 THE SBGN-ML FILE FORMAT

SBGN-ML is a dedicated lightweight XML-based file format describing the overall geometry of SBGN maps, while also preserving their underlying biological meaning. SBGN-ML is designed to fulfill two basic requirements:

(1) easy to draw (as a machine) and read (as a human) and

(2) easy to interpret (as a machine).

The first set of requirement deals with the graphical aspect of SBGN. It means it should be easy to render a SBGN-ML file to the screen. Therefore, the format stores all necessary information, such as coordinates, to draw the map faithfully, so that rendering tools do not have to perform any complex calculations. Incidentally, this implies the layout of the whole SBGN map has to be expressed explicitly: the size and position of each graphical object and the path of each arc. Various efforts have shown that generating a layout for heterogeneous biological pathways is a computationally hard problem, so a good layout is always worth preserving, if only from a computational perspective. Besides, the choice of a specific layout by the author of a map is often driven by concerns related to aesthetics, readability or to reinforce ideas of chronology or proximity. This information might be lost with automated layouts. Layout conventions predate SBGN, and are not part of any standard,

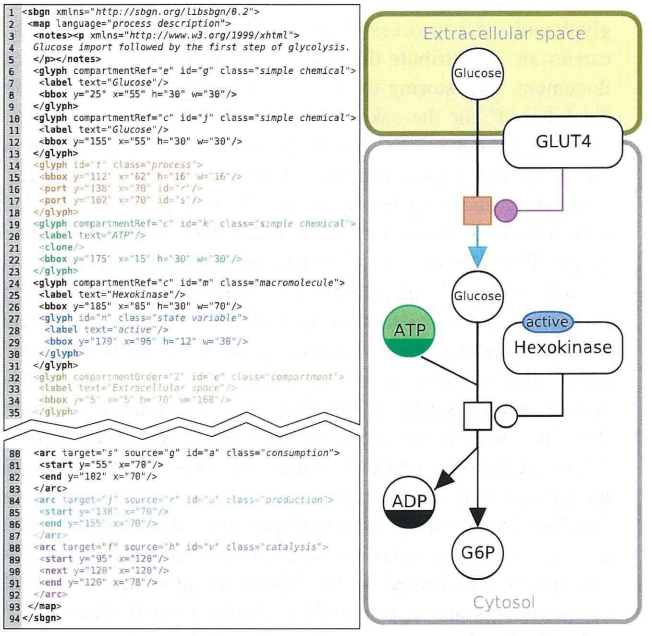

Fig. 1. An example PD map (right) with the corresponding SBGN-ML code (left). This example shows the import of glucose followed by the first step of glycolysis. The colors used have no special meaning in SBGN, here they merely indicate the relation between each SBGN glyph and its SBGN-ML representation; a process node in orange, a simple chemical (ATP) in green, a production arc in cyan, a catalysis arc in purple, a compartment in yellow and a state variable in blue

but they nonetheless play a large role in making it easier for other human beings to understand the biological system being described.

The second requirement encompasses two perpendicular characteristics of SBGN as a language: semantics and syntax. Beyond the picture itself, the format should capture the biological meaning of an SBGN map. Therefore, SBGN-ML specifies the nature of graphical elements (glyphs), following the SBGN terminology (e.g., macromolecule, process, etc.). For example, we can distinguish between a 'logic arc' and a 'consumption arc' even though they have the same visual appearance. Supporting tools refer to this terminology and draw the glyph according to the SBGN specifications. In terms of syntax, SBGN-ML encodes information on relationships between the various SBGN objects: the glyphs at both ends of an arc, the components of a complex, the members of a compartment and the 'decorations' (such as unit of information and state variable) belonging to specific glyphs and arcs. This semantic and syntactic information is essential to a number of automated tasks, such as map validation, or network analysis (as the topology of the underlying biological network can be inferred from the various relationships encoded by the format).

To explain the syntax of SBGN-ML in more detail, consider the example in Figure 1. This figure shows a PD map describing the import of glucose by GLUT4, followed by the first step of the glycolysis. The root element is named 'sbgn' (line 1). Below that, there is a 'map' element with an attribute indicating that the PD language is used. Below the map element, one finds a series of glyph and arc elements. Each glyph carries a 'class' attribute to denote the meaning in SBGN terms. In this example, there is a