BMC Bioinformatics

Proceedings

Leveraging existing biological knowledge in the identification of candidate genes for facial dysmorphology

Hannah J Tipney[1], Sonia M Leach[1,2], Weiguo Feng[3], Richard Spritz[4], Trevor Williams[3] and Lawrence Hunter*[1]

Address: [1]Computational Pharmacology Department, University of Colorado at Denver and Health Sciences Center, Aurora, CO, USA, [2]ESAT, Research Division SCD, Katholieke Universiteit Leuven, B-3001 Leuven, Belgium, [3]Department of Craniofacial Biology, University of Colorado at Denver and Health Sciences Center, Aurora, CO, USA and [4]Human Medical Genetics Program, University of Colorado at Denver and Health Sciences Center, Aurora, CO, USA

Email: Hannah J Tipney - hannah.tipney@uchsc.edu; Sonia M Leach - sonia.leach@uschsc.edu; Weiguo Feng - weiguo.feng@uchsc.edu; Richard Spritz - richard.spritz@uchsc.edu; Trevor Williams - trevor.williams@uchsc.edu; Lawrence Hunter* - larry.hunter@uchsc.edu

* Corresponding author

from The First Summit on Translational Bioinformatics 2008
San Francisco, CA, USA. 10–12 March 2008

Published: 5 February 2009

BMC Bioinformatics 2009, **10**(Suppl 2):S12 doi:10.1186/1471-2105-10-S2-S12

This article is available from: http://www.biomedcentral.com/1471-2105/10/S2/S12

Abstract

Background: In response to the frequently overwhelming output of high-throughput microarray experiments, we propose a methodology to facilitate interpretation of biological data in the context of existing knowledge. Through the probabilistic integration of explicit and implicit data sources a functional interaction network can be constructed. Each edge connecting two proteins is weighted by a confidence value capturing the strength and reliability of support for that interaction given the combined data sources. The resulting network is examined in conjunction with expression data to identify groups of genes with significant temporal or tissue specific patterns. In contrast to unstructured gene lists, these networks often represent coherent functional groupings.

Results: By linking from shared functional categorizations to primary biological resources we apply this method to craniofacial microarray data, generating biologically testable hypotheses and identifying candidate genes for craniofacial development.

Conclusion: The novel methodology presented here illustrates how the effective integration of pre-existing biological knowledge and high-throughput experimental data drives biological discovery and hypothesis generation.

Background

The increased use of high-throughput analysis methods, such as microarrays, in mainstream biological research has led to a shift from studying small groups of reasonably well-characterized variables to exploring a complicated mire of thousands of inter-related variables simultaneously [1]. These methods are powerful, but their outputs are complicated and difficult to interpret due to the sheer

volume of data produced. Interpretation can be prohibitively time consuming in the absence of computational assistance.

The ultimate goal of any microarray experiment is to gain insight into the workings of cellular organisms by understanding the interactions of genes and proteins. For this to be accomplished, raw data must not only be converted into information, but this information must also be interpreted in context, to be transformed into timely biological discovery and knowledge [2]. Currently, the lack of a community-wide consensus on how best to integrate experimental data with information resources limits this knowledge acquisition [2]. The recent work of Saraiya *et al.* (2005) [1] highlighted a "critical need" for tools able to "connect numerical patterns to the underlying biological phenomena", as current techniques fail to adequately link microarray data to biological meaning, which limits researchers' biological insights [1].

One intuitive way to integrate biological knowledge and microarray data is through protein-interaction networks, where nodes represent proteins and edges symbolize relationships between proteins [3]. However, focusing solely on physical protein interactions, such network constructs neglect a wealth of knowledge currently distributed among hundreds of existing biological databases (over 1000 listed in this year's *Nucleic Acids Research* database issue alone [4]) that is directly applicable to proteins investigated via microarray experiments. Current protein network constructs typically focus on a small subset of this biological knowledge, producing incomplete and sparsely populated resources. This is a particular problem for higher eukaryotic organisms such as mice and humans, for which physical protein interaction data are limited.

In agreement with Lee *et al.* (2004) [5] and Leach *et al.* (2007) [3], we demonstrate by expanding the definition of 'interaction' to include functional information that a) there is enough publicly available biological information to produce biologically useful, well populated interaction networks for higher eukaryotic species, b) through the combination of expression data and functional information, it is possible to provide contextual insight into the network, and c) it is possible to effectively link to existing biological knowledge using current technology. Using a murine craniofacial developmental expression microarray dataset [6] and a recently published technique for weighting and integrating functional interaction information [3], we illustrate how the application of context sensitive methodology leverages the full force of current available biological knowledge, enabling the translation of complex high-throughput datasets into scientific insight and discovery.

Methods

Microarray expression data

A comprehensive murine craniofacial developmental expression dataset was used in this study [6]. Expression was analyzed through the microdissection of mandibular, maxillary and frontonasal prominences at time points E10.5–E12.5 at 0.5 day increments. Expression was measured using the Affymetrix MOE430_2A microarray system. 916 microarray probes, corresponding to 712 unique MGI identifiers, were clustered using the MANOVA test statistic in R [7]. Hierarchical complete clustering was undertaken on the resulting correlation coefficients. The resultant tree was cut to produce 36 clusters.

Explicit and implicit data sources

Traditionally an 'interaction' between two proteins is defined as a physical association. Here we expand the term to include functional relationships between pairs of proteins (encompassing any type of evidence, including physical, functional, genetic, biochemical, evolutionary, and computational evidence [3]). Functional interaction information was retrieved from a number of different resources falling into either of two categories: explicit and implicit. Explicit sources indicate a direct interaction between a pair of genes/proteins, and include experimentally measured physical, biochemical and genetic interactions, and computationally predicted gene neighborhoods, gene fusion events, or conserved phylogenetic profiles. Implicit sources provide information pertaining to an individual gene or protein attribute, which may be shared by any given pair of genes or proteins. Such attributes include literature references [8], sequence motifs (PReMod, InterPro) [9,10], protein categories (ChEBI) [11], protein complexes [12], phenotypes (as described by the Mouse Genome Database) [13], cellular location, molecular function, and biological processes (Gene Ontology) [14] and pathways (KEGG, Ingenuity) [3,15,16].

Network construction, weighting and visualization

Genes within any given cluster were defined to be the nodes in our network constructs; a network was produced for each cluster identified from the hierarchical complete clustering stage. Data from both implicit and explicit sources were used to define arc interactions between pairs of proteins. Applying the cons NoisyOR methodology of Leach *et al.* (2007) [3], the edges between each pair of nodes were assigned a combined reliability score (network component, σ^{NET}) based on the individual reliabilities of the sources asserting the edge [3]. Resultant networks were viewed using Cytoscape [17].

Network identification and interrogation

Based on significant tissue-restricted expression (expression limited to the mandibular prominence) and progres-

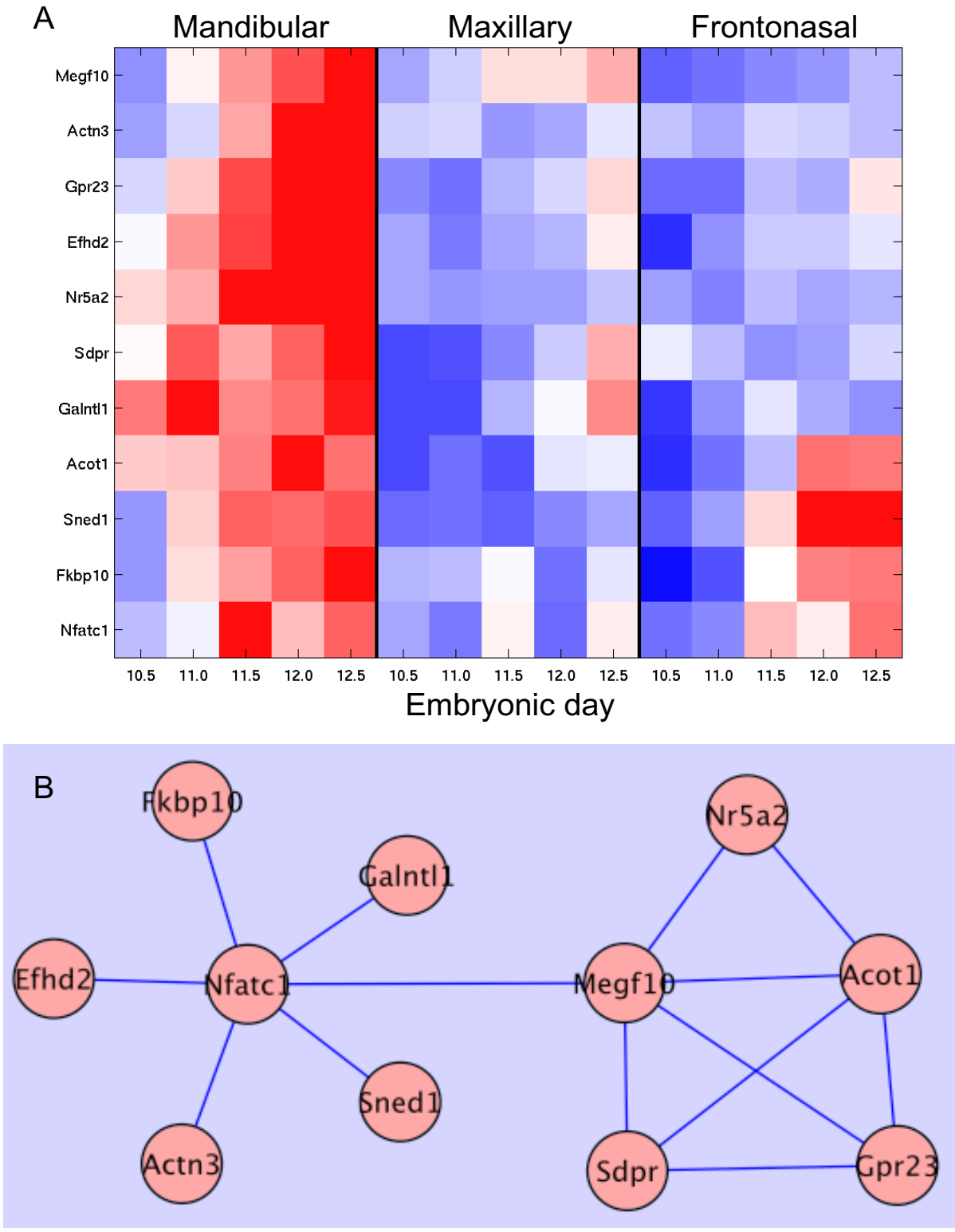

Figure 1
ChEBI informed network. A) Illustratates the mandibular-specific expression profile, B) the network of 11 nodes and their associated edges.

sive increase in expression from embryonic day (E) 11 to E12.5 (Figure 1a), a network of interest was identified comprising 52 nodes.

By using Cytoscape's ability to display edge attributes, rapid orientation within the sub-network was achieved. Recognizing implicit functional themes common to nodes and edges across the sub-network facilitated the identification of key information such as shared pathways, processes, locations and phenotypes, as well as over-represented protein families. This approach provides a high-level overview of the networks functional composition, which subsequently directs the user towards more focused analysis of individual nodes and their pairwise interactions.

Each pair of nodes and associated interactions were reviewed through the exploration of 'links' within Cytoscape to a number of biological resources (including EntrezGene, GenBank, OMIM, PubMed, MGI, iHOP and UniGene). Although we were working with murine developmental expression microarray data, our goal was to gain insight into human developmental processes. Pertinent information is distributed across both mouse and human resources, thus requiring all investigations be repeated to ensure that valuable data were not overlooked simply due to their residing in a species-specific entry or database. In addition to mining these associated database entries, a corpus of relevant literature was compiled by following links from each implicit source. This corpus was then manually assessed and interpreted in order to extract key information.

The network construction method applied here provides a maximum of 12 possible functional interactions per edge. The edge with the most support in this network was between *Myod1* [Myogenic differentiation 1; EntrezGene ID: 17927 (mouse) and 4654 (human)] and *Myog* [Myogenin; EntrezGene ID: 17928 (mouse) and 4656 (human)], and was asserted by nine sources (PubMed, PReMod, GO [BP, MF and CC], MGI Phenotype, InterPro, ZTransloc and ChEBI). Therefore, even with a structured interrogation strategy guided by both expression data and functional interactions, pursuing all informative leads was not a minor task. Considerable time was dedicated to the interpretation of the significance of each element. In this example, approximately 80 hours (10 days) of expert user time was required. The process is cyclical, where new information not only informs future discoveries, but previous work is frequently revisited to be viewed within new contexts.

Results and discussion

A mandibular-specific network: finding novelty

The group of 52 proteins was clustered on the basis of correlated mRNA expression, and with the construction of the network the goal was to develop biological hypotheses explaining the observed correlation. Those edges with the most support (in addition to correlated expression) frequently comprise well-documented relationships between the proteins encoded by these genes. Those edges with less support (fewer sources of shared characteristics) are therefore more likely to represent novel, as-yet uncharacterized relationships and so generate new hypotheses.

In this instance, a number of nodes (eleven) were linked by edges asserted by a single expert (ChEBI; Chemical Entities of Biological Interest). ChEBI is a dictionary of molecular entities focusing on 'small' chemical compounds of biological relevance and encompasses ontological classifications [11]. Interestingly, these nodes and edges also formed a discrete sub-network (Figure 1b). Although each of these nodes has reasonable amounts of associated biological knowledge, only the ChEBI data provided the shared functional interaction categorizations required for network construction (Table 1). This subgraph of 11 nodes (linked by edges solely asserted by shared ChEBI categorizations), exhibiting correlated and progressively up-regulated expression in mandibular tissue during mouse development, thus provided a unique opportunity to search for truly novel biological hypotheses.

The common theme: calcium and lipids

By observing the sub-graph as a group of 11 interacting proteins (as opposed to individual or pairs of proteins), the implicit functional information displayed through the presence of edges can be taken together to produce a strong and unified voice capable of highlighting themes which may have otherwise gone unnoticed if only single, unassociated implicit resources were used.

Table 1: Biological knowledge associated with each node.

ID	ChEBI	GO:BP	GO:CC	GO:MF	KEGG	PHENO
Acot1	✓	✓	✓	✓	-	-
Actn3	✓	✓	✓	✓	✓	-
Efhd2	✓	-	-	✓	-	-
Fkbp10	✓	✓	✓	✓	-	-
Galntl1	✓	-	✓	✓	✓	-
Gpr23	✓	✓	✓	✓	✓	-
Megf10	✓	✓	-	✓	-	-
Nfatc1	✓	✓	✓	✓	✓	-
Nr5a2	✓	✓	✓	✓	✓	✓
Sdpr	✓	-	✓	✓	-	-
Sned1	✓	✓	✓	✓	-	-

Data present represented by a ✓, absent by -.

In this instance, the ChEBI terms on which the network was constructed were "calcium(2+)" and "lipids". Cumulative evidence from GO and KEGG highlighted themes around muscle, acyl-CoA, lipids, signaling, and calcium signaling. Of the 11 nodes, only *Nr5a2* [nuclear receptor subfamily 5, group A, member 2; EntrezGene ID: 26424 (mouse) and 2494 (human)] had a MGI phenotype association. *Nr5a2* knockouts exhibit digestive, alimentary, and immune disruption, and are embryonic lethals.

Insights from primary literature

Primary databases also provided a source of literature. Publications attached to GO annotations, GeneRIFs, and phenotypes (for example) can subsequently be explored further. Published literature is the gold standard for classification and description of biological functions; however, much of the knowledge in this vast resource is difficult to assess in the absence of prior knowledge of what to query for. Searching for pertinent information regarding all 52 associated genes in the original network and craniofacial disorder constitutes a formidable challenge due to the extremely large number of primary

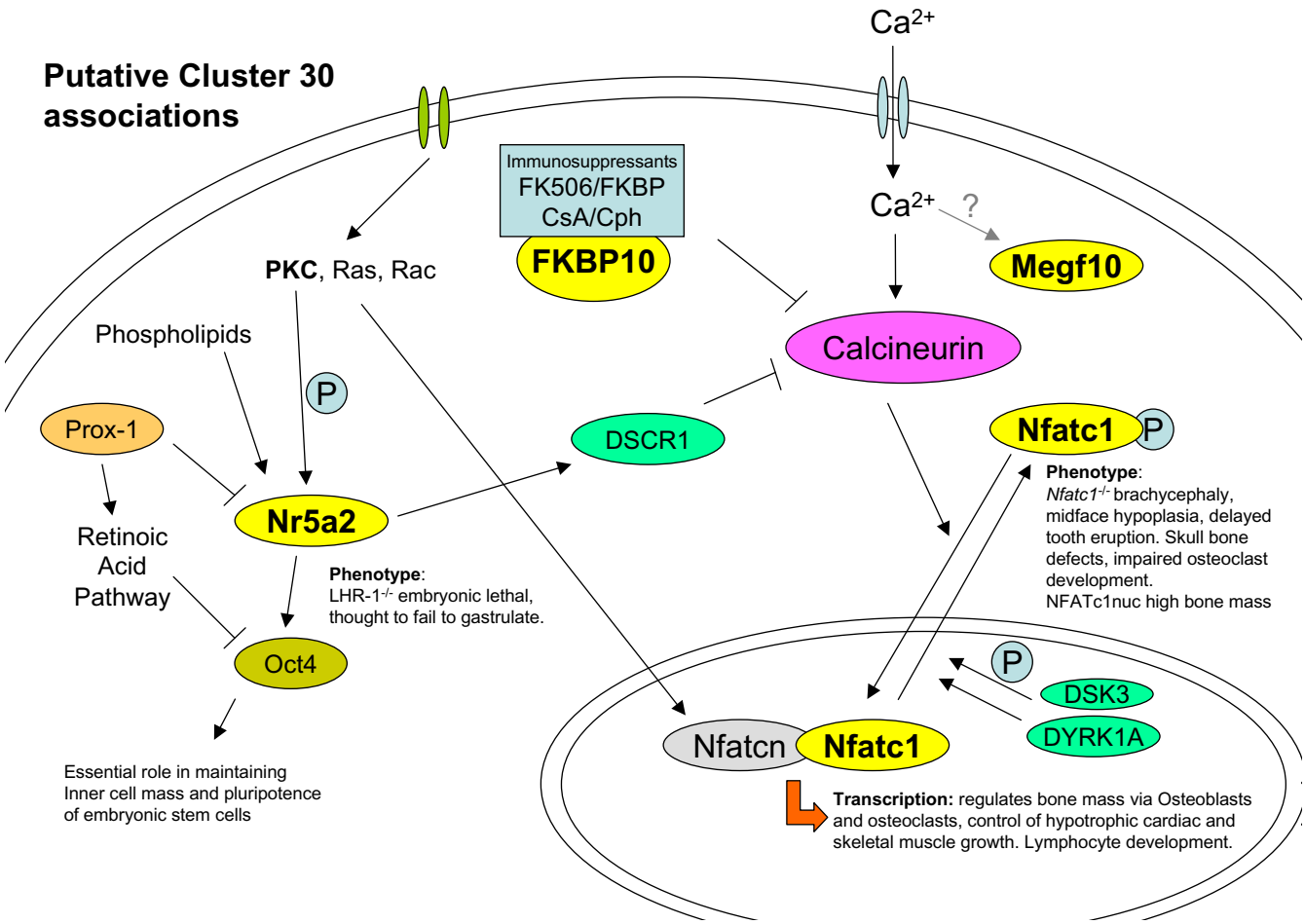

Figure 2

Putative functional associations of ChEBI identified proteins. From insights in the literature associations were found between Nfatc1, its phosphorylation status [19], calcineurin [20,21] and skeletal/craniofacial dysmorphology [19,22]; *Fkbp10*, its role in developing tissues [23] and negative regulation of calcineurin [20]; *Nr5a2*, its role in PKC-phosphorylation [24], DSCR1 expression [25](both PKC and DSCR1 are implicated in Nfatc-regulated transcription pathways [19,20]), and retinoic acid signalling [24,26](implicated in craniofacial development); α-actin (*Actn3*), its role in mediating calsarcin and calcineurin interactions [27]; and *Mef2*, its role in calcineurin-dependent gene-regulation [28]. *Megf10* has putative a calcium binding site (IPR001881), while *Sned1* and *Galntl1* are found at the sites of embryonic apoptosis and ossification [29,30] but little more was discovered about these poorly characterised proteins. Yellow ovals highlight those proteins in the ChEBI sub-network.

papers returned. However, the task becomes more manageable when the user has an insight into the relationships among a smaller subgroup of genes. In this instance, it was more productive to search over the ChEBI-specific sub-network for *Nfatc1*, a role in mandibular development, known associations with *Actn3*, *Sned1*, *Fkbp10*, *Galntl1*, *Efhd2*, and *Megf10*, and any associations with calcium signaling. By having a structured network rather than a gene list, mining the associated literature became a more targeted and thus more fruitful task.

As expected, the published literature provided a wealth of existing knowledge, and a putative biological pathway loosely centered on calcineurin was hypothesized to explain the correlated expression of the 11 genes (Figure 2). In the absence of a network to guide investigations, it would have been difficult, if not impossible, to associate these genes in a biologically meaningful way without domain-specific prior knowledge.

The importance of leveraging pre-existing biological knowledge

Although the availability and application of high-throughput methods such as expression microarray technology has been a key advance in biological research, the biological research community is still grappling with how best to harness the power of this technology. Biologists are dependent upon computational methodologies to help them navigate and interpret their raw data, and as such the ability of the biologist to generate new insights, hypotheses and discoveries is intimately associated to the methodology's ability to assist effective discovery of biologically meaningful information. Saraiya *et al.* (2005) [1] highlighted how failing to link microarray data to existing biological knowledge prevents biologists from leveraging their domain expertise to construct higher level, biologically relevant hypotheses. They argued that it was "imperative that users be able to access and link biological information to their data" [1]. In agreement with this point, the study presented here illustrates how effective access to pre-existing knowledge can drive biological discovery. Inadequate access to this wealth of information ultimately hinders scientific discovery. We have demonstrated that through the combination of both explicit and implicit data and a permissive visualization environment such as Cytoscape, it is possible to link large-scale microarray datasets to biological information in a manner which facilitates hypothesis generation.

Conclusion

The approach outlined here will be particularly useful when applied to analyses of large-scale datasets (such as from microarrays) to help understand the processes implicated in complex, multi-factorial disorders. In addition to the example presented here, application of this methodology to analysis of our craniofacial developmental expression microarray dataset [6] has led to identification and validation of four genes not previously implicated in craniofacial development [18]. We believe this methodology will be of significant use to the wider scientific community, and we are therefore also currently working towards explicitly capturing and automating this analysis protocol and developing a user interface to facilitate ease of investigation.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

HJT participated in the design of the system, undertook the analysis, and wrote the manuscript. SML designed and implemented the system. WF, TW and RS advised on and evaluated the biology, and critically appraised the manuscript. LH conceived of the system, supervised all aspects of its construction, and contributed to the manuscript.

Acknowledgements

H. Johnson and A. Gabow for comments. HT is funded by a Fulbright-AstraZeneca Research Fellowship. SL and LH are supported by NLM R01 LM008111/R01 LM009254. WF, RS and TW are supported by DE015191.

This article has been published as part of *BMC Bioinformatics* Volume 10 Supplement 2, 2009: Selected Proceedings of the First Summit on Translational Bioinformatics 2008. The full contents of the supplement are available online at http://www.biomedcentral.com/1471-2105/10?issue=S2.

References

1. Saraiya P, North C, Duca K: **An insight-based methodology for evaluating bioinformatics visualizations.** *IEEE Trans Vis Comput Graph* 2005, **11:**443-456.
2. Bassett D, Eisen M, Boguski M: **Gene expression informatics – it's all in your mine.** *Nat Genet* 1999, **21(1 Suppl):**51-55.
3. Leach S, Gabow A, Hunter L, Goldberg DS: **Assessing and combining reliability of protein interaction sources.** *Pac Symp Biocomput* 2007:433-444.
4. Galperin M: **The molecular biology database collection: 2007 update.** *Nucleic Acids Research* 2007, **35:**D3-4.
5. Lee I, Date S, Adai A, Marcotte E: **A probabilistic functional network of yeast genes.** *Science* 2004, **306:**1555-1558.
6. **GEO Dataset GSE7759 - Spatial and Temporal Analysis of Gene Expression During Growth and Fusion of the Mouse Facial Prominences.** .
7. **The Comprehensive R Archive Network** [http://cran.r-project.org/]
8. Wheeler D, Barrett T, Benson D, Bryant S, Canese K, Chetvemin V, Church D, DiCuccio M, Edgar R, Federhen S, *et al.*: **Database resources of the National Center for Biotechnology Information.** *Nucleic Acids Research* 2007, **35:**D5-12.
9. Blanchette M, Bataille A, Chen X, Poitras C, Laganiere J, Lefebvre C, Deblois G, Giguere V, Ferretti V, Bergeron D, *et al.*: **Genome-wide computational prediction of transcriptional regulatory modules reveals insights into human gene expression.** *Genome Research* 2006, **16:**656-668.
10. Mulder N, Apweiler R, Attwood T, Bairoch A, Bateman A, Binns D, Bork P, Buillard V, Cerutti L: **New developments in the InterPro database.** *Nucleic Acids Research* 2007, **35:**D224-228.
11. de Matos P, Ennis M, Zbinden M, McNaught A, Alcantara R, Darsow M, Guedj M, Ashburner M, Degtyarenko K: **ChEBI – Chemical Entities of Biological Interest.** *Nucleic Acids Research Database Collection* 2007, **646:**.

12. Lu Z: **Text Mining on GeneRIFs.** In *PhD Thesis Computational Bioscience Program, University of Colorado School of Medicine, CO, USA*; 2007.
13. Eppig J, Bult C, Kadin J, Richardson J, Blake J: **The Mouse Genome Database (MGD): from genes to mice – a community resource for mouse biology.** *Nucleic Acids Research* 2005, **33:**D471-475.
14. Consortium GO: **Gene Ontology: tool for the unification of biology.** *Nature Genetics* 2000, **25:**25-29.
15. **Ingenuity systems** [http://www.ingenuity.com]
16. Kanehisa M, Goto S, Hattori M, AokioKinoshita K, Itoh M, Kawashima S, Katayama T, Araki M, Hirakawa M: **From genomics to chemical genomics: new developments in KEGG.** .
17. Shannon P, Markiel A, Ozier O, Baliga N, Wang J, Ramage D, Amin N, Schwikowski B, Ideker T: **Cytoscape: A software environment for integrated models of biomolecular interaction networks.** *Genome Res* 2003, **13:**2498-2504.
18. Leach S, Tipney H, Feng W, Baumgartner WA, Kasliwal P, Schuyler R, Williams T, Spritz R, Hunter L: **3R Systems for biomedical discovery acceleration, with applications to craniofacial development.** *PLoS Computational Biology* 2008 in press.
19. Arron J, Winslow M, Polleri A, Chang C-P, Wu H, Gao X, Neilson J, Chen L, Heit J, Kim S, *et al.*: **NFAT dysregulation by increased dosage of DSCR1 and DYRK1A on chromosome 21.** *Nature* 2006, **441:**595-600.
20. Crabtree G: **Generic signals and specific outcomes: Signaling through Ca2+, calcineurin, and NF-AT.** *Cell* 1999, **96:**611-614.
21. Graef I, Chen F, Crabtree G: **NFAT signaling in vertebrate development.** *Current Opinion in Genetics & Development* 2001, **11:**505-512.
22. Winslow M, Pan M, Starbuck M, Gallo E, Deng L, Karsenty G, Crabtree G: **Calcineurin/NFAT signaling in osteoblasts regulates bone mass.** *Developmental Cell* 2006, **10:**771-782.
23. Patterson C, Schaub T, Coleman E, Davis E: **Developmental regulation of FKBP65: An ER-localized extracellular matrix binding-protein.** *Molecular Biology of the Cell* 2000, **11:**3925-3935.
24. Fayard E, Auwerx J, Schoonjans K: **LRH-1: An orphan nuclear receptor involved in development, metabolism and steroidogenesis.** *TRENDS in Cell Biology* 2004, **14:**250-260.
25. Paré J-F, Malenfant D, Courtmanche C, Jacob-Wagner M, Roy S, Allard D, Bélanger L: **The fetoprotein transcription factor (FTF) gene is essential to embryogenesis and cholesterol homeostasis and is regulated by a DR4 element.** *The Journal of Biological Chemistry* 2004, **279:**21206-21216.
26. Gu P, Goodwin B, Chung A, Xu X, Wheeler D, Price R, Galardi C, Peng L, Latour A, Koller B, *et al.*: **orphan nuclear receptorLRH-1 is required to maintain Oct4 expression at the epiblast stage of embryonic development.** *Molecular and Cellular Biology* 2005, **25:**3492-3505.
27. Frey N, Richardson J, Olson E: **Calsarcins, a novel family of sarcomeric calcineurin-binding proteins.** *PNAS* 2000, **97:**14632-14637.
28. Wu H, Naya F, McKinsey T, Mercer B, Shelton J, Chin E, AR , Michel R, Bassel-Duby R, Olson E, Williams R: **MEF2 responds to multiple calcium-regulated signals in the control of skeletal muscle fiber type.** *EMBO J* 2000, **19:**1963-1973.
29. Leimeister C, Schumacher N, Diez N, Gessler H: **Cloning and expression analysis of the mouse stroma marker Snep encoding a novel nidogen domain protein.** *Developmental Dynamics* 2004, **230:**371-377.
30. Tian E, Hagen K, Shum L, Hang H, Imbert Y, Young W, Bertozzi C, Tabak L: **An inhibitor of O-glycosylation induces apoptosis in NIH3T3 cells and developing mouse embryonic mandibular tissues.** *The Journal of Biological Chemistry* 2004, **279:**50382-50390.