Published in final edited form as:
Int J Bioinform Res Appl. 2010 ; 6(2): 163–178.

Strategies for enhanced annotation of a microarray probe set

TuShun R. Powers, **Selene M. Virk**, and **Elba E. Serrano**
Department of Biology, New Mexico State University, Las Cruces, NM 88003, USA

TuShun R. Powers: tupowers@nmsu.edu; Selene M. Virk: selene@nmsu.edu; Elba E. Serrano: serrano@nmsu.edu

Abstract

We aim to determine the biological relevance of genes identified through microarray-mediated transcriptional profiling of *Xenopus* sensory organs and brain. Difficulties with genetic data analysis arise because of limitations in probe set annotation and the lack of a universal gene nomenclature. To overcome these impediments, we used sequence based and semantic linking methods in combination with computational approaches to augment probe set annotation on a commercially available microarray. Our curation efforts enabled linkage of probe sets and expression data to public databases, increased the biological significance of our microarray data, and assisted with the tentative identification of unidentified probe sets.

Keywords

bioinformatics; data mining; gene identification; inner ear; mechanosensory hair cell; enhanced annotation; microarray probe set

1 Introduction

Xenopus is a well established model organism for cellular and genetic investigations of complex biological processes during development (Nieuwkoop and Faber, 1967; Pollet et al., 2005; Wullimann et al., 2005). Microarrays increasingly are implemented as powerful tools for large scale assessment of gene expression patterns during essential features of *Xenopus* life such as embryonic tissue specification, organogenesis, neural induction, and hormonal signalling (Altmann et al., 2001; Munoz-Sanjuan et al., 2002; Baldessari et al., 2005; Das et al., 2006). In our laboratory, *Xenopus* is implemented as a model for identification of genes that are implicated in inner ear function and sensorineural organogenesis (Varela-Ramirez et al., 1998; Serrano et al., 2001; Quick and Serrano, 2007). To this end we are using the *Affymetrix* GeneChip® *Xenopus laevis* Genome Array for transcriptional profiling of inner ear organs as a method for large scale identification of genes specific to the inner ear, especially those involved in inner ear development (Powers et al., 2007). Our comparative approach analyses *Xenopus* inner ear transcriptional profiles together with those of brain and other *Xenopus* organs. These experiments are undertaken with the long term goal of uncovering genes essential for the process of mechanotransduction and the maintenance and regeneration of the mechanosensory hair cell phenotype.

During our analysis of *Xenopus* microarray expression data, we observed that the vendor supplied annotation information for the *Xenopus laevis* GeneChip® probe set IDs (Xl-PSIDs) was a key limiting factor in our ability to interpret results. In part, this is because genetics research with the tetraploid *Xenopus* species, *X. laevis*, can be inherently difficult. The *X.*

Correspondence to: Elba E. Serrano, `serrano@nmsu.edu`.

laevis genome has not been sequenced and the majority of *X. laevis* gene annotations arise from ESTs and cDNA libraries. Consequently, gene annotations for *X. laevis* may not be as comprehensive as those for *X. tropicalis*, a diploid sister species that is viewed as a superior alternative for molecular genetics experiments due to its smaller genome and shorter generation time (Amaya, 2005). The *X. tropicalis* genome has been sequenced and is available through the University of Santa Cruz (UCSC) genome browser (Kent et al., 2002).

Thus, a central theme that emerged in our genetic analysis is that variations in gene nomenclature and functional descriptions can cause ambiguities when we attempt to extract biological significance from Xl-PSID annotations. What is the remedy for this problem, which is a challenge shared by many investigators? This issue is especially problematic for researchers who work with organisms other than human or mouse, or with organisms having an unsequenced or poorly annotated genome. Some laboratories have solved this problem by developing methods to match genes to multiple gene identifiers and integrate this information into a queriable database (Glasner et al., 2003). For example, Dai et al. (2004) developed the GeneView system to provide more extensive gene annotation information for microarray chips than their vendors.

In this paper we summarise our efforts to enhance the annotations of the Xl-PSIDs using curation procedures that linked them to more informative gene identifiers such as the HGNC symbols (HUGO Gene Nomenclature Committee) and UniProt (Universal Protein Resource) IDs (Table 1). HGNC symbols are derived from the official gene name of a human protein approved by the HUGO Gene Nomenclature Committee (http://www.genenames.org/). As in the case of HGNC symbols, a consortium oversees the curation of protein sequences with accurate and functional annotations for UniProt IDs (The UniProt Consortium, 2007). Thus, linkage of Xl-PSIDs to these public databases increases the value added from the enhanced annotation because a large scientific community regularly oversees and updates the gene identifiers, thereby minimising the impact of the lack of standardisation of gene nomenclature. We developed three computational strategies that relied either on sequence similarity (Figure 1(A)) or semantic similarity (Figure 1(B)). Our results demonstrate how manual and automated annotation identification techniques can be used to enrich the available information that can be mined from a transcriptional profiling experiment in a manner that is cost-effective and easily achievable for a small laboratory operation.

2 Datasets

The 15,611 PSIDs on the *Affymetrix* GeneChip® *Xenopus laevis* Genome Array comprise 15,503 probe sets representing 14,400 transcripts (16 probe pairs per transcript) that are a combination of mRNAs and ESTs with a bias towards 3′ UTRs. Within this total, 135 PSIDs are *Affymetrix* controls. For the purpose of this research paper we will only discuss the 15,476 PSIDs that begin with the ‘Xl.’ label because the transcripts used to design these probe sets can be identified in the UniGene clusters by their *X. laevis* UniGene ID. Annotation information for Xl-PSIDs is provided online at *The NetAffx Analysis Center* (http://www.affymetrix.com/analysis/index.affx) or in the vendor supplied annotation file, Xenopus_laevis.na25.annot.csv (http://www.affymetrix.com/support/technical/byproduct.affx?product=xenopus). The annotation provided for the Xl-PSIDs includes information about gene identifiers such as a one line gene title, a gene symbol, the archival UniGene cluster, the UniGene ID, the UniGene cluster type, the Entrez gene ID, the RefSeq protein ID, the RefSeq transcript ID, Gene Ontology (GO) terms with GO ID number, the Swissprot ID, and the GenBank Accession number.

After inspecting the annotation file from *Affymetrix*, we determined that the one line '*Gene Title*' column contained information that could be used to group 98% of Xl-PSIDs into nine different categories (Table 1): Category X, those without a designated '*Gene Title*'; Category O, so named for 'other', contains one line descriptions consisting of gene names or anything else that does not fall into the other eight categories; Category C has designations with cDNA IMAGE numbers (integrated molecular analysis of genomes and their expression, Miller et al., 1997) or numbered cDNA clones; Category H contains hypothetical proteins (in which 20 have additional modifiers such as a putative gene name for the hypothetical protein); Category M contain the MGCXXXX protein numbers from the mammalian gene collection (http://mgc.nci.nih.gov/), and four categories are designated as Transcribed Locus (TL), with TLS, TLM, and TLW abbreviated for transcribed loci that are strongly (>90%), moderately (70–90%) or weakly (<70%) similar to another protein in an aligned region (percentages of similarity are defined by UniGene (http://www.ncbi.nlm.nih.gov/UniGene/FAQ.shtml)). As can be seen in Figure 2, most of the Xl-PSIDs were allocated into the following categories: O (other, 30%), TL (transcribed locus, 26%), and H (hypothetical proteins, 21%). The quality of annotation information provided for the nine different categories varies, with the encircled groups (TL, M, H, and C) in Figure 2 having the most ambiguous gene descriptions. These groupings were used to analyse the outcomes of our enhancement of the Xl-PSID annotations before and after linkage to HGNC symbols or UniProt IDs.

3 Computational approaches

3.1 Sequence based computational approaches for linking Xl-PSIDs to HGNCs

3.1.1 Annotation with implementation of manual primary literature searches, open source software, and public database queries—Interesting gene expression patterns for inner ear function are not lacking in the literature. Due to the importance of ion channels for the process of mechanotransduction, as well as their role in hereditary disorders of hearing and balance (Gabashvili et al., 2007; Serrano et al., 2001), we are especially interested in enhancing Xl-PSIDs with annotations for this gene family. However, making connections between experimental data found in the literature and microarray data can be challenging because of variable nomenclature as well as new understanding of gene function for different species. The type of gene identifiers referenced in the literature greatly depends on when the paper was published. If the nomenclature has changed since publication, there can be difficulties in establishing connections to Xl-PSIDs. Additionally, the name of proteins in publications can vary for similar genes between species.

Through literature searches we identified a publication by Gabashvili et al. (2007) that contained a list of 262 genes for Ion Channel Activity (ICA), which are expressed in inner ear tissue. The ICA gene list was compiled based on data gleaned from cDNA libraries, microarray analysis, and RT-PCR experiments. Because the Gabashvili et al. (2007) reference already had HGNCs as the gene symbol of choice, we did not need to search for these. However, it should be noted, that if HGNC symbols are not known, GenBank accession numbers, gene names or symbols can be used to find the HGNC symbols in the UCSC genome browser (Kent et al., 2002). This process of identifying target genes was the first step in the curation process outlined in Figure 1(A1).

We began the process of identifying putative ICA Xl-PSIDs by making a text file of all HGNCs from the list in the paper and using this to collect *Homo sapiens* protein sequences with BioMart (http://www.ensembl.org/biomart/martview/4ad4b6a9d10e301741f0d1e2755fe0f0), a mining tool that can be used to retrieve information from the Ensembl database (Figure 1(A1)). When the Ensembl database (release 49) was filtered with the list of 262 ICA HGNC symbols, a total of 241 protein sequences were recovered from the *H. sapiens* gene dataset (NCBI36). We used the TBLASTN algorithm software (Altschul et al., 1990) in a local search with the *H.*

sapiens ICA proteins as the protein query, and the 15,476 consensus nucleotide sequences for Xl-PSIDs that begin with the ‘Xl.’ label (http://www.affymetrix.com/support/technical/byproduct.affx?product=xenopus) as the nucleotide database. Results from the TBLASTN search were evaluated by the returned e-value and a list was compiled of the best Xl-PSID matches to ICA HGNC symbols (e-values less than 10^{-14}). This labour-intensive method limits the number of Xl-PSIDs that can be processed at once, but provides the highest quality annotation for Xl-PSIDs.

3.1.2 Large scale batch annotation with implementation of open source software and public database queries—This approach (Figure 1(A2)), like the manual annotation identification, used sequence similarity to link Xl-PSIDs to UniProt IDs by mining the UniProt database. For this automated strategy, the BLASTX program (Altschul et al., 1990) was used in a large scale batch search. All known *H. sapiens* (human), *Mus musculus* (mouse), *Caenorhabditis elegans* (worm) and *Drosophila melanogaster* (fly) protein sequences were collected from UniProt (The UniProt Consortium, 2007), then compared against nucleotide queries consisting of the 15,476 consensus sequences for the Xl-PSIDs that begin with the ‘Xl.’ label. With this method the annotation of an entire chip can be readily enhanced.

3.2 Semantic based computational approaches for linking Xl-PSIDs to HGNCs

To address the limitations of vendor-supplied annotation data, a local instance of an annotation database was created that incorporated internal expression data and publicly available gene annotation information (Figure 1(B)). The vendor-supplied annotation for Xl-PSIDs is in a format that is not readily searchable in a high-throughput fashion. In an attempt to overcome this limitation, a MySQL® database, *XenEnhance*, was created to store the Xl-PSID annotation data. We opted to store data in a relational database rather than in another format (such as a flat file) because a relational database can accommodate expansion of *Xenopus* resource annotation efforts in the future beyond microarray curation. *XenEnhance* permits linkage of Xl-PSIDs with the UniGene cluster IDs. This approach relies on the fact that Xl-PSIDs are derived from UniGene cluster IDs; therefore, their linkage to UniGene cluster IDs is straightforward. Although the vendor provides the UniGene cluster IDs as part of the Xl-PSID annotation, vendor UniGene cluster ID updates are provided less frequently than those on the UniGene public database which are updated monthly (http://www.ncbi.nlm.nih.gov/UniGene/FAQ.shtml).

Specifically, *XenEnhance* consists of seven database tables (Figure 3) created using data retrieved in flat file format from four data sources: the *X. laevis* GeneChip® annotation file, UniGene, RefSeq, and HGNC symbol data. Two tables were created from the UniGene data. The table ‘ug_clusters_82’ stores descriptive information for each UniGene cluster in the release, while the table ‘sim_proteins_ug’ stores descriptive information for the human proteins that are similar to the UniGene clusters in ‘ug_clusters_82’. One table, ‘hgnc_051908’, was populated with HGNC data that was downloaded on May 19, 2008. This table contains descriptive information for each current official gene symbol. Another table, ‘hs_prot2rna_29’, was created using data from the RefSeq file, release29.accession2geneid (http://www.ncbi.nlm.nih.gov/RefSeq/). This table contains human RefSeq protein accession, the corresponding human RefSeq RNA accession, and the human Entrez gene ID for the corresponding gene. Two additional tables were created using data from the vendor-supplied annotation file for the *X. laevis* GeneChip®. The Xl-PSID links the data in these two tables. One of these tables, ‘xl_probeset_031808’, contains descriptive information whereas the second table, ‘xl_probe_031808’, contains the individual probe ID and the sequences that comprise a probe set. Finally, the linking table, ‘ug_82_hgnc_051908’, was created to maintain the association between a *X. laevis* UniGene cluster and HGNC official gene symbols (see Figure 4(A)–(C)).

To format the data from the various sources into a form that could be inserted into *XenEnhance*, text parsers and table population software were written in the Java™ language (SDK version 1.5.0_13) to extract the required data from text files and load data into the tables.

The vendor-supplied annotation file is parsed in such a way as to create a separate row in the corresponding database table for each Xl-PSID (a separate table was created for the individual probes, which are linked to the Xl-PSIDs). The UniGene Xl.data file was parsed to extract cluster ID, cluster description, the gene represented by the cluster, and the gene ID. The parsed data were loaded into the table 'ug_clusters_82'. The presence of the gene ID permits direct linking to NCBI's Entrez Gene without the need to create a local instance of that data set. As mentioned previously, the Xl-PSIDs are derived from UniGene cluster IDs and the two can therefore be readily linked without the creation of a distinct linking table.

To make the link between the Xl-PSIDs and the HGNC official gene symbols, it was necessary to combine data from other sources to that obtained from UniGene. The first step toward this linkage relied upon the UniGene file, Xl.data. Using the contents of this file, we were able to link *X. laevis* UniGene clusters with proteins from different species, including human. When a UniGene cluster has a human protein with which it shares a degree of similarity, the protein accession ID is used to locate the corresponding human Entrez gene ID from the RefSeq file, release29.accession2geneid. In addition to Entrez gene ID, the release29.accession2geneid file provides RefSeq RNA accessions (when available), the taxon ID, and the RefSeq protein accessions. Using the Entrez gene ID, the HGNC database table could be searched to locate the corresponding official gene symbol. The Entrez gene ID is used as the direct link between Xl-PSIDs and HGNC official gene symbols. It is also possible to use the RefSeq protein accession to link to the HGNC data, which would be useful in cases where the HGNC symbol could not be linked to Xl-PSID by means of Entrez gene ID. Because we were using a more recent release of UniGene than was used when the chip was created, it was necessary for us to use current cluster assignments of the Xl-PSIDs to make the linkage to HGNC symbols.

4 Results and discussion

4.1 All three curation approaches successfully link Xl-PSIDs to HGNCs or UniProt IDs

The HGNC symbols for the 241 ICA genes that were manually identified, matched to 134 Xl-PSIDs (Table 1A1). Seventy-three percent of the HGNC symbols were linked to Xl-PSIDs in the O category, showing that even this labour intensive technique can enhance one line gene titles. The majority of enhanced annotations were identified using the large scale sequence similarity based approach (Table 1A2). UniProt IDs from *H. sapiens* and *M. musculus* had the largest counts for matches, linking 7,259 (47%) and 6,952 (45%) of the Xl-PSIDs, respectively, to a new annotation (Table 1A2). Using the semantic approach, we were able to associate 4031 (26%) of the 15,476 Xl-PSIDs that were derived from UniGene clusters to HGNC symbols based on various degrees of similarity between *X. laevis* and human proteins.

All three procedures had the highest matches to the other (O) and hypothetical protein (H) categories. Over half of the matches obtained with the sequence similarity approach enhanced annotations within the O category (Table 1A); 1/3 of the matches obtained with the semantic based approach enhanced annotations within the O and the H categories (Table 1B).

4.2 Automated large scale data curation procedures enhance 60% of the Xl-PSIDs

Figure 5(A) illustrates that together, the two large scale approaches (sequence and semantic based) matched a majority of the Xl-PSIDs to an additional gene identifier such as a UniProt ID or HGNC symbol (9108 matches, Table 2A). More than half of all the Xl-PSIDs were enhanced with an additional annotation, with over half of these matching 4–6 annotations.

Within the enhanced group (Figure 5(B)), over a quarter of Xl-PSIDs were matched to five new annotations.

When we compared the category distribution of the enhanced (Figure 6(A)) and non-enhanced (Figure 6(B)) Xl-PSIDs, we observed that the majority of the enhanced Xl-PSIDs were in the O (other, 45%) and H (hypothetical proteins, 27%) categories (Figure 6(A)). The third largest group with enhanced annotations, the M category (MGCXXXX proteins from the mammalian gene collection) included 10% of the Xl-PSIDs with enhanced annotations (Figure 6(A)). Our curation strategies identified 929 additional annotations within the M category (74%, Table 2). The TL category remained elusive, with 3,733 (94.7%) of the Xl-PSIDs having only *Affymetrix* annotations (Table 2) after curation.

5 Conclusion

We have attacked the problem of enhancing microarray chip annotation with multiple approaches that were either sequence similarity based, or relied on semantic relationships between publicly available annotation data from *X. laevis* and other species. HGNC and UniProt databases were selected as purveyors of target enhancement annotations because they are collaboratively maintained by the research community and they are frequently updated (http://www.genenames.org/; The UniProt Consortium, 2007). Using the semantic approach, we increased our flexibility in mining the microarray data by creating a relational database that linked the vendor supplied annotation data for Xl-PSIDs with publicly available annotation for *X. laevis* and other species. In so doing, we were able to enhance the existing annotation for 60% of the Xl-PSIDs and to associate gene function properties with Xl-PSIDs that lacked this information in the annotation provided by the microarray vendor (Figures 5 and 6). In the future, we intend to extend our curation efforts by augmenting Xl-PSID annotations with information from online public resources dedicated to *Xenopus* genetics. For example, websites such as *Xenbase* (http://www.xenbase.org) and the *Xenopus* Gene Collection (XGC) (http://xgc.nci.nih.gov/) are repositories for *Xenopus* gene annotation data that are especially useful because they provide lists of full length clones. XGC comprises 10,291 full open reading frame clones and 9,138 non-redundant genes and *Xenbase* is currently in the process of linking human homologs, along with other species, to *Xenopus* genes. In addition, software tools developed as a result of our annotation efforts will be made freely available to the *Xenopus* community.

Ideally, we would prefer to work with universal gene symbols. However, presently there is no universally accepted consensus gene nomenclature for all species and there is variability in the degree of completion of publicly available annotation data for various species. Consequently, multifaceted approaches for annotation such as those presented here will continue to provide value for researchers by enhancing vendor-supplied microarray annotation and tentatively identifying previously unidentified Xl-PSIDs.

Acknowledgments

The authors express their gratitude to Dr. Charles Whittaker for insights and assistance with the mapping of *X. laevis* Xl-PSIDs to sequenced genomes. This work was supported in part by NIH (GM008136; DC003292; P50GM068762) awards to E.E.S. and an NSF IGERT Fellowship to S.M.V. (DGE-0504304).

References

Altmann CR, Bell E, Sczyrba A, Pun J, Bekiranov S, Gaasterland T, Brivanlou AH. 'Microarray-based analysis of early development in *Xenopus laevis*'. Developmental Biology 2001;236(1):64–75. [PubMed: 11456444]

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. ‘Basic local alignment search tool’. Journal of Molecular Biology 1990;215(3):403–410. [PubMed: 2231712]

Amaya E. ‘Xenomics’. Genome Research 2005;15(12):1683–1691. [PubMed: 16339366]

Baldessari D, Shin Y, Krebs O, König R, Koide T, Vinayagam A, Fenger U, Mochii M, Terasaka C, Kitayama A, Peiffer D, Ueno N, Eils R, Cho KW, Niehrs C. ‘Global gene expression profiling and cluster analysis in *Xenopus laevis’*. Mechanisms of Development 2005;122(3):441–475. [PubMed: 15763214]

Dai H, Tian B, Zhao WD, Leung A, Smith SR, Wan JS, Yao X. ‘Dynamic integration of gene annotation and its application to microarray analysis’. Journal of Bioinformatics and Computational Biology 2004;1(4):627–645. [PubMed: 15290757]

Das B, Cai L, Carter MG, Piao YL, Sharov AA, Ko MS, Brown DD. ‘Gene expression changes at metamorphosis induced by thyroid hormone in *Xenopus laevis* tadpoles’. Developmental Biology 2006;291(2):342–355. [PubMed: 16458881]

Gabashvili IS, Sokolowski BH, Morton CC, Giersch AB. ‘Ion channel gene expression in the inner ear’. Journal of the Association for Research in Otolaryngology 2007;8(3):305–328. [PubMed: 17541769]

Glasner JD, Liss P, Plunkett G III, Darling A, Prasad T, Rusch M, Byrnes A, Gilson M, Biehl B, Blattner FR, Perna NT. ‘ASAP, a systematic annotation package for community analysis of genomes’. Nucleic Acids Research 2003;31(1):147–151. [PubMed: 12519969]

Kent, WJ.; Sugnet, CW.; Furey, TS.; Roskin, KM.; Pringle, TH.; Zahler, AM.; Haussler, D. The human genome browser at UCSC; Genome Research. 2002. p. 996-1006.Website: http://genome.ucsc.edu/

Miller G, Fuchs R, Lai E. ‘IMAGE cDNA clones, UniGene clustering, and ACeDB: an integrated resource for expressed sequence information’. Genome Research 1997;7(10):1027–1032. [PubMed: 9331373]

Munoz-Sanjuan I, Bell E, Altmann CR, Vonica A, Brivanlou AH. ‘Gene profiling during neural induction in *Xenopus laevis*: regulation of BMP signalling by post-transcriptional mechanisms and TAB3, a novel TAK1-binding protein’. Development 2002;129(23):5529–5540. [PubMed: 12403722]

Nieuwkoop, PD.; Faber, J. Normal Table of Xenopus laevis (Daudin): A Systematical and Chronological Survey of the Development from the Fertilized Egg Till the End of Metamorphosis. 2. North-Holland Publishing Co; Amsterdam: 1967.

Pollet N, Muncke N, Verbeek B, Li Y, Fenger U, Delius H, Nierhrs C. ‘An atlas of differential gene expression during early *Xenopus* embryogenesis’. Mechanisms of Development 2005;122(3):365–439. [PubMed: 15763213]

Powers T, Trujillo-Provencio C, Whittaker C, Serrano EE. ‘Gene expression profiling of xenopus organs yields insight into the *Xenopus* inner ear transcriptome’. Association for Research in Otolaryngology. Abstr 2007;30(741):254.

Quick QA, Serrano EE. ‘Cell proliferation during the early compartmentalization of the *Xenopus laevis* inner ear’. International Journal of Developmental Biology 2007;51(3):201–210. [PubMed: 17486540]

Serrano EE, Trujillo-Provencio C, Sultemeier D, Bullock WM, Quick QA. ‘Identification of genes expressed in the *Xenopus* inner ear’. Cellular and Molecular Biology (Noisy-le-grand) 2001;47(7): 1229–1239.

The UniProt Consortium. The Universal Protein Resource (UniProt); Nucleic Acids Res. 2007. p. D193-D197.UniProt website: http://www.pir.uniprot.org/

Varela-Ramirez A, Trujillo-Provencio C, Serrano EE. ‘Detection of transcripts for delayed rectifier potassium channels in the *Xenopus laevis* inner ear’. Hearing Research 1998;119(1–2):125–134. [PubMed: 9641325]

Wullimann MF, Rink E, Vernier P, Schlosser G. ‘Secondary neurogenesis in the brain of the African clawed frog, *Xenopus laevis*, as revealed by PCNA, *Delta-1*, *Neurogenin-related-1*, and *NeuroD* expression’. The Journal of Comparative Neurology 2005;489(3):387–402. [PubMed: 16025451]

Websites

Affymetrix website, Obtained through the internethttp://www.affymetrix.com/support/technical/byproduct.affx?product=xenopus

Biomart via Ensembl, Obtained through the internethttp://www.ensembl.org/biomart/martview/4ad4b6a9d10e301741f0d1e2755fe0f0

HUGO Gene Nomenclature Committee, Obtained through the internethttp://www.genenames.org/

NetAffx website, Obtained through the internethttp://www.affymetrix.com/analysis/index.affx

The mammalian gene collection website, Obtained through the internethttp://mgc.nci.nih.gov/

The Reference Sequence (RefSeq) Collection website, Obtained through the internethttp://www.ncbi.nlm.nih.gov/RefSeq/

The Xenopus Gene Collection (XGC) website, Obtained through the internet: http://xgc.nci.nih.gov/

UniGene website, Obtained through the internet: http://www.ncbi.nlm.nih.gov/UniGene/FAQ.shtml

Xenbase website, Obtained through the internet: http://www.xenbase.org

Biographies

TuShun R. Powers is a Postdoctoral Researcher in the Department of Biology at New Mexico State University. She received her PhD in Molecular Biology with a special interest in ruminant microbiology from New Mexico State University in 2005. She is intrigued by interdisciplinary research approaches and has interests in microbiology, molecular biology, bioinformatics, and functional genomics.

Selene M. Virk is a PhD student in the Department of Biology at New Mexico State University, where she received her Bachelor and Masters of Science degrees. Her current research interests include developing software tools to facilitate functional gene annotation, nervous system development, mechanotransduction, and modelling of complex biological systems. She has ten years experience as a Software Engineer in the biotech industry.

Elba E. Serrano is a Regent’s Professor of Biology at New Mexico State University and a member of the MIT Systems Biology Cell Decision Processes Center. She received her undergraduate degree in physics with distinction from the University of Rochester and her PhD in Biological Sciences from Stanford University with an emphasis in neuroscience and biophysics. Her research interests include neural regeneration, mechanotransduction, and disorders of hearing and balance. She has a special interest in promoting interdisciplinary education that bridges the life and quantitative/physical sciences, and in programmes that encourage students to pursue advanced degrees in Science, Technology, Engineering, and Mathematics (STEM) disciplines.

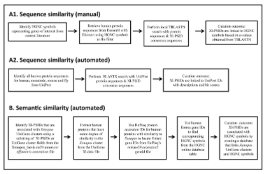

A1. Sequence similarity (manual)
Identify HGNC symbols representing genes of interest from current literature
Retrieve human protein sequences from Ensembl with Biomart using HGNC symbols as the filter
Perform local TBLASTN search with protein sequences & Xl-PSID consensus sequences
Curation outcome: Xl-PSIDs are linked to HGNC symbols based on e-values obtained from TBLASTN
A2. Sequence similarity (automated)
Identify all known protein sequences for human, nematode, mouse and fly from UniProt
Perform BLASTX search with UniProt protein sequences & Xl-PSID consensus sequences
Curation outcome: Xl-PSIDs are linked to UniProt IDs with descriptions and bit scores
B. Semantic similarity (automated)
Identify Xl-PSIDs that are associated with Xenopus UniGene clusters using a substring of Xl-PSIDs or UniGene cluster fields from the Xenopus_laevis.na24.annot.csv Affymetrix annotation file
Extract human proteins that have some degree of similarity to the Xenopus cluster from the UniGene Xl data file
Use RefSeq protein accession IDs for human proteins with similarity to Xenopus to locate Entrez gene IDs from RefSeq's release29.accession2geneid file
Use human Entrez gene IDs to find corresponding HGNC symbols from the HGNC online database table
Curation outcome: Xl-PSIDs are associated with HGNC symbols by creating a database that links Xenopus UniGene clusters and HGNC symbols

Figure 1.

Overview of strategies used to curate Xl-PSIDs with enhanced gene identification through either sequence similarity (A) or semantic similarity (B) computational methods. A1. HGNCs (official gene symbols) linked to Xl-PSIDs for genes of interest manually identified in the literature using sequence similarity as the linkage parameter. A2. Automated large scale linkage of Xl-PSIDs to UniProt IDs (for *H. sapiens*, *C. elegans*, *M. musculus*, and *D. melanogaster* protein sequences) using sequence similarity matching. B. Semantic-based data flow process to link Xl-PSIDs to HGNCs

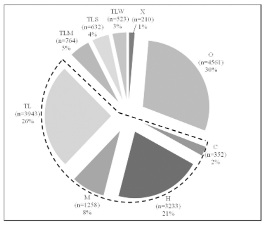


Figure 2.

Pie chart of *Affymetrix* annotated gene titles (*n* = 15,476) divided into 9 different categories based on the 'Gene Title' information provided by the vendor. X, no label; O, other; C, cDNA clones/IMAGE; H, hypothetical protein; M, MGCXXXX protein; TL, transcribed locus; TLS, TLM and TLW are strongly, moderately and weakly similar to annotated proteins, respectively. The encircled TL, M, H and C categories contain Xl-PSIDs that lack descriptive gene titles

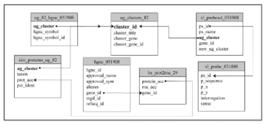
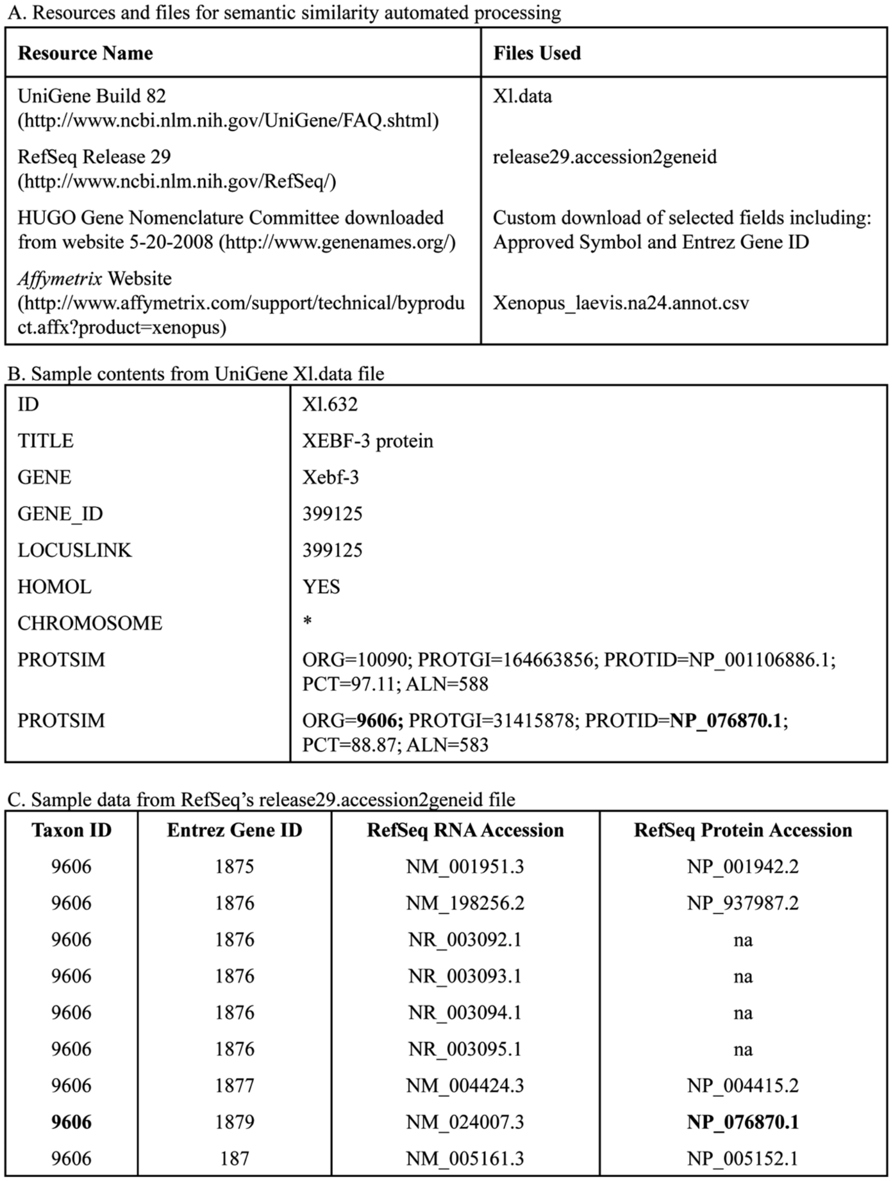


Figure 3.

Schematic view of XenEnhance depicting the relationships between tables. Note, arrows do not imply unidirectional linkage, but demonstrate how HGNCs were linked to Xl-PSIDs



A. Resources and files for semantic similarity automated processing

Resource Name	Files Used
UniGene Build 82 (http://www.ncbi.nlm.nih.gov/UniGene/FAQ.shtml)	Xl.data
RefSeq Release 29 (http://www.ncbi.nlm.nih.gov/RefSeq/)	release29.accession2geneid
HUGO Gene Nomenclature Committee downloaded from website 5-20-2008 (http://www.genenames.org/)	Custom download of selected fields including: Approved Symbol and Entrez Gene ID
Affymetrix Website (http://www.affymetrix.com/support/technical/byproduct.affx?product=xenopus)	Xenopus_laevis.na24.annot.csv

B. Sample contents from UniGene Xl.data file

ID	Xl.632
TITLE	XEBF-3 protein
GENE	Xebf-3
GENE_ID	399125
LOCUSLINK	399125
HOMOL	YES
CHROMOSOME	*
PROTSIM	ORG=10090; PROTGI=164663856; PROTID=NP_001106886.1; PCT=97.11; ALN=588
PROTSIM	ORG=**9606;** PROTGI=31415878; PROTID=**NP_076870.1**; PCT=88.87; ALN=583

C. Sample data from RefSeq's release29.accession2geneid file

Taxon ID	Entrez Gene ID	RefSeq RNA Accession	RefSeq Protein Accession
9606	1875	NM_001951.3	NP_001942.2
9606	1876	NM_198256.2	NP_937987.2
9606	1876	NR_003092.1	na
9606	1876	NR_003093.1	na
9606	1876	NR_003094.1	na
9606	1876	NR_003095.1	na
9606	1877	NM_004424.3	NP_004415.2
9606	1879	NM_024007.3	**NP_076870.1**
9606	187	NM_005161.3	NP_005152.1

Figure 4.

Data sources for automated processing. A. List of the sources used to retrieve information for the relational database, XenEnhance. B. Sample contents from UniGene Xl.data file. Xl-PSIDs were associated to HGNC symbols using UniGene cluster ID and the PROTSIM line. Note: Entrez Gene ID is the gene index for the X. laevis gene associated with the cluster. C. Sample data from the RefSeq data file (release29. accession2geneid) retrieved using the PROTID from the PROTSIM line in 4(B)

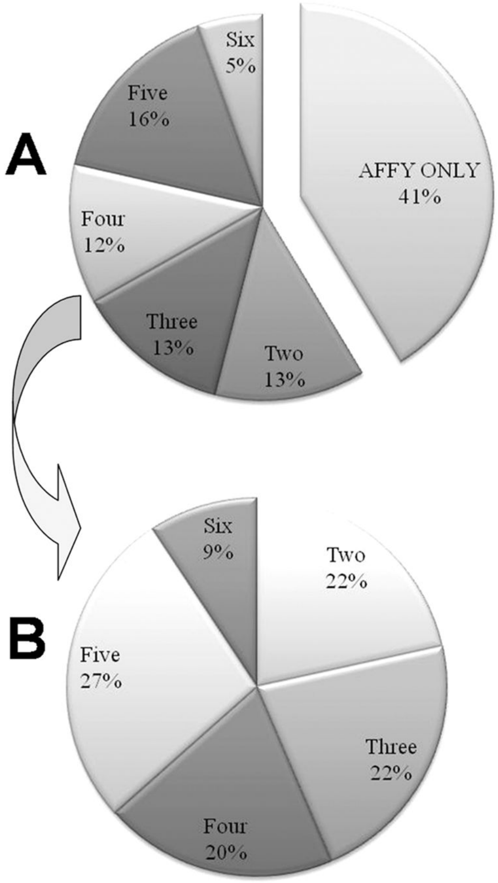


Figure 5.

Total number of annotations associated with each Xl-PSID after automated large scale curation of the data with sequenced based and semantic based approaches. A. Pie chart showing the relative percentage of Xl-PSIDs with 1–6 annotations. The ‘AFFY ONLY’ percentage represents the number of Xl-PSIDs that were not enhanced by the curation process and therefore have only 1 annotation. B. Pie chart showing relative percentage for the subset of Xl-PSIDs from 5A with multiple (2–6) annotation matches

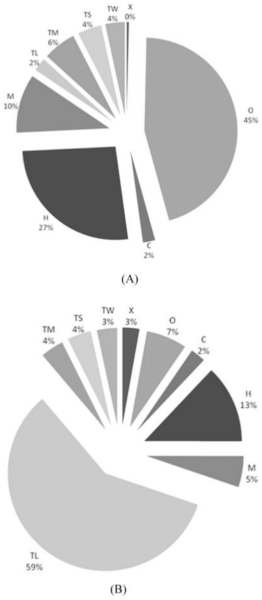


Figure 6.

Summary of the enhancement of Xl-PSID annotations with the addition of HGNC symbols and UniProt IDs by automated large scale data curation. (A) PSIDs with enhanced annotation, 2–6 each (*n* = 9,108) and (B) PSIDs without enhanced annotation (*n* = 6,368)

Table 1

Number of Xl-PSIDs with enhanced annotations after data curations with sequence (A) and semantic (B) based procedures

		A. Sequence based curation					B. Semantic based curation: automated large scale
Affymetrix annotations		1. Manual	2. Automated large scale				
Nine categories of gene titles	# Affy Xl-PSIDs	Ion Channel HGNCs	H. sapiens UniProt IDs	C. elegans UniProt IDs	M. musculus UniProt IDs	D. melanogaster UniProt IDs	HGNCs
X = no label	210	0	25	14	24	12	5
O = other	4561	98	3872	2218	3814	2462	1332
C = clone/IMAGE	352	4	130	49	114	50	106
H = hypothetical protein	3233	14	1647	660	1563	652	1286
M = MGCXXXX protein	1258	7	611	267	581	263	523
TL = transcribed locus	3943	3	89	21	78	16	113
TLS = trans locus strongly sim. to	632	4	295	125	269	125	198
TLM = trans locus moderately sim. to	764	3	380	109	323	131	313
TLW = trans locus weakly sim. to	523	1	210	63	186	72	155
Total probe sets	15476	134	7259	3526	6952	3783	4031

Table 2

Number of Xl-PSIDs with enhanced annotations after automated large scale data curation procedures

			B. # Xl-PSIDs with 1–6 annotations after curation procedures					
Nine categories of gene titles	**# Affy Xl-PSIDs**	**A. Total # Xl-PSIDs with enhanced annotations**	**1-Affy Only**[*]	**2 Only**	**3 Only**	**4 Only**	**5 Only**	**6 Only**
X = no label	210	32	178	9	9	3	11	0
O = other	4561	4131	430	286	836	835	1632	542
C = clone/IMAGE	352	193	159	67	55	24	34	13
H = hypothetical protein	3233	2406	827	784	557	491	429	145
M = MGCXXXX protein	1258	929	329	323	182	198	163	63
TL = transcribed locus	3943	210	3733	134	56	11	7	2
TLS = trans locus strongly sim. to	632	379	253	93	85	84	86	31
TLM = trans locus moderately sim. to	764	520	244	156	139	110	78	37
TLW = trans locus weakly sim. to	523	308	215	111	84	55	47	11
Total probe sets	15476	9108	6368	1963	2003	1811	2487	844

[*] The number of Xl-PSIDs that were not enhanced by the curation process and retained only the *Affymetrix* gene identifiers are listed under ‘1-Affy only’.